RTIA'18

22nd Nov-24th Nov 2018

International conference on

Recent Trends in IoT and Its Application

Conference Proceeding

ARYAN INSTITUTE OF ENGINEERING & TECHNOLOGY

AIET Bhubaneswar

Aryan Institute of Engineering and Technology Bhubaneswar

Organized By-

DEPARTMENT OF

COMPUTER SCIENCE ENGINEERING

AIET, Bhubaneswar, 752050

RTIA-2018

Recent Trends in IOT and Its Application

22nd Nov. – 24th Nov. 2018

CONFERENCE PROCEEDING

Organized by

Department of Computer Science Engineering
Aryan Institute of Engineering and Technology
Bhubaneswar – 752050

List of Sponsors

Aryan Infra Projects

Citicon Engineers Pvt. Ltd

Oltron Technology Pvt. Ltd.

Citiz Essential Services Pvt. Ltd.

ABOUT THE CONFERENCE

Department of Computer Science Engineering, AIET, Bhubaneswar is organizing an Online International Conference on Recent Trends in IOT and Its Application (RITA-18) on 22-24 Nov 2018.This conference aims to bring together leading academic scientists, researchers and research scholars to exchange and share their experiences and research results on all aspects of IOT, Machine Learning and Artificial Intelligence. It also provides a premier interdisciplinary platform for researchers, practitioners and educators to present and discuss the most recent innovations, trends, and concerns as well as practical challenges encountered and solutions adopted in the fields of IOT, Machine Learning and Artificial Intelligence. Participants will develop a deep understanding of latest tools and technologies pertaining to latest topics and will get a chance to connect with other professionals in the field.

ABOUT THE DEPARTMENT

The Computer Science & Engineering Department became operational in the year 2009 since inception of AIET with 60 students admitted through OJEE as per guide line of AICTE. The department has taken several initiatives to bridge the gap between industry and academia by conducting several training and certification programs such as DB2, Android, Oracle, .NET, SAP to name a few. It has its own Software Development Centre which has significantly contributed for the development of CMS (College Management System) and is effectively operational now. Currently Department of CSE is enriched with scholar faculties. They not only contribute their knowledge to build vibrant academics, but also utilize their skills to apply the science of computing to solve real life research problems. The department lays stress on the application-based aspects through laboratories, seminars, group discussions, viva-voce and project work, keeping pace with the growth in Computer Technology.

ABOUT THE INSTITUTE

Established in the year 2009 , Aryan Institute of Engineering and Technology(AIET) is one of the premier engineering colleges in the self-financing category of Engineering education in eastern India. It is situated at temple city Bhubaneswar, Odisha and is a constituent member of Aryan Educational Trust. This reputed engineering college is accredited by NAAC, UGC and is affiliated to BPUT, Odisha. AIET aims to create disciplined and trained young citizens in the field of engineering and technology for holistic and national growth.

The college is committed towards enabling secure employment for its students at the end of their four year engineering degree course. (The NAAC accreditation in the year 2018 vouches for the college's determination and dedication for a sustainable learning environment). The academic fraternity of AIET is a unique blend of faculty with industry and academic experience. This group of facilitators work with a purpose of importing quality education in the field of technical education to the aspiring students. Affordable fee structure along with approachable location in the smart city of Bhubaneswar, makes it a preferred destination for aspiring students and parents.

Organizing Committee Members

PATRON:

Dr. Madhumita Parida
Chairperson
Aryan Institute of Engineering and Technology

Director
Prof. Sasmita Parida

Jt. Organizing Secy.
Prof. Jui Pattanayak

Convener
Assist. Prof. A. K. Sahoo

Treasurer
Prof. Sanjay Kumar Padhi

Organizing Secy.
Prof. Laxmi

Jt. Treasurer
Asst. Prof. Sushree Sangita Jena

INTERNATIONAL ADVISORY COMMITTEE

NATIONAL ADVISORY COMMITTEE

Dr. Astha Chauhan
Professor
Department of Electronics & Telecommunication Engineering
Indian Institute of Technology, Delhi

Dr. Deepak Kumar
Professor
Department of Computer Science Engineering
Indian Institute of Technology, Madras

Dr. Dharmendra Singh Raghav
Professor
Department of Computer Science Engineering
Indian Institute of Technology, Bombey

Dr. Bharati Chowdhury
Professor
Department of Computer Science Engineering
Indian Institute of Technology, Roorkee

Dr. Debashish Gountia
Associate Professor
Department of Electronics & Telecommunication Engineering
Indian Institute of Technology, Roorkee

Dr. Manish Kumar
Associate Professor
Department of Electrical Engineering
Indian Institute of Technology,
Guwahati

Dr. Nidhi Pal
Associate Professor
Department of Computer Science Engineering
Indian Institute of Technology, Kharagpur

Dr. Pankaj Kumar
Associate Professor
Department of Electrical Engineering
National Institute of Technology, Tirchi

Dr. Prem Prakash
Assistant Professor
Department of Electronics & Telecommunication Engineering
National Institute of Technology, Raipur

Dr. Sarang Kapoor
Assistant Professor
Department of Electrical Engineering
National Institute of Technology, Durgapur

LOCAL COMMITTEE MEMBERS

Assist. Prof. Prakash Dehury
Department of Computer Science Engineering

Assist. Prof. Vidya Mohanty
Department of Computer Science Engineering

Assist. Prof. Pravt Kumar Routray
Department of Computer Science Engineering

Assist. Prof. Radha Mohan Acharya
Department of Computer Science Engineering

Assist. Prof. T. R. Baitharu
Department of Computer Science Engineering

Prof. P. K. Subudhi
Department of Electronics and Communication Engineering

Prof. Sangita Pal
Department of Electronics and Communication Engineering

Assist. Prof. P. C. Satapathy
Department of Electronics and Communication Engineering

Assist. Prof. Ajit Kumar Panda
Department of Electrical Engineering

Assist. Prof. Prasanta Kumar Sahoo
Department of Electronics and Communication Engineering

Assist. Prof. Ipsita Samal
Department of Electronics and Communication Engineering

Assist. Prof. Ankita Panda
Department of Electronics and Communication Engineering

Assist. Prof. D. K. Sahoo
Department of Electrical Engineering

Assist. Prof. Subashish Mohanty
Department of Electrical Engineering

Assist. Prof. S. M. Samal
Department of Electrical Engineering

Assist. Prof. S. K. Tripathy
Department of Electrical Engineering

Assist. Prof. Abhishek Mohanty
Department of Electrical Engineering

Assist. Prof. S. K. Mishra
Department of Electrical Engineering

Conference Committee Management

1. **Reception Management**
   - Rasmi Rekha Mahato
   - Rajkumari Lopamudra
   - Bidyabati Samal
   - Urmila Sahoo
   - Smita Digal
   - Sunita Priyadarsini

2. **Transit/Accommodation Management**
   - Krishna Naidu
   - Biswanath Majhi
   - Kishore Chandra Pradhan
   - Duga Prasad Padhy
   - Jyoti Prakash Khunti
   - Deepen Kumar Jena
   - Manoranjan Mahunta

3. **Seminar Hall Management**
   - Prasannjeet Pattanaik
   - Pravat Ranjan Mishra
   - Priti Ranjan Jena
   - Samrat Kharabela Senapati
   - Bidyadhar Jena
   - Chinmaya Mohapatra

4. **Catering Management**
   - Pradyumna Badajena
   - Pravat Kumar Sahoo
   - Hara Prasad Mishra
   - Asutosh Bal

5. **Printing/Stationary Management**
   - Sumit Ghosh
   - Akhilesh Singh
   - Vikas Meher
   - Vishal Gupta

6. **Design Team**
   - Sridhar Jena
   - Mahesh Nayak
   - Nitesh Samal

7. **Anchoring In Inauguration Ceremony**
   - Rudra Prasad Nanda
   - Nilimashree Niharika

Conference Sub-Committee Management

- Dillip Kumar Pradhan
- Junaid Mohammad
- Chinmaya Mohapatra
- Naresh Sharma
- Pradeep Ghadei
- Dinesh Shah
- Narendra Mallick

Aryan Institute of Engineering & Technology
Bhubaneswar – 752050, Khurda, India

TEL: 0674-259-5051
e-mail: chairman@aryan.ac.in

Dr. Madhumita Parida, Chairperson

Chairperson's Message

On behalf of the Aryan Institute of Engineering & Technology (AIET), I extend a very warm welcome to all the delegates and participants present today for the International Conference on the subject "Recent Trends in IOT and Its Application (RTIA'18)" AIET has borne the mantle of excellence, committed to ensuring the students their own space to learn, grow and broaden their horizon of knowledge by indulging into diverse spheres of learning. In our endeavor to raise the standards of discourse, we continue to remain aware to meet the changing needs of our stakeholders.

Last but not the least; we would also like to thank the staff, faculty members, the organizers, and the students for their contribution in successfully organizing and managing this event. This event wouldn't have been possible without their guidance and constant support.

We welcome all of you to AIET and hope that, this conference will act as a medium for all to ponder upon the topic of discussion, challenge us to strive towards it, and inspiring us to go ahead.

Thank you!

Dr. Madhumita Parida

Dr. Madhumita Parida

Aryan Institute of Engineering & Technology
Bhubaneswar – 752050, Khurda, India

TEL: 0674-2377-101
e-mail: director@aryan.ac.in

Prof. Sasmita Parida, Director

Director's Message

It is my great pleasure to present the proceedings of the International Conference on "Recent Trends in IOT and Its Application", RTIA'2018. This conference is organized by the Department of Computer Science Engineering of Aryan Institute of Engineering & Technology, Bhubaneswar. We have received 63 research papers and after review 40 papers were accepted for publication in the proceeding. These review papers were presented in different categories of oral sessions, poster sessions, and invited talks. RTIA'2018 covered various research areas of Computer Science, Electrical and Electronics Engineering.

I welcome on behalf of AIET to all the delegates and speakers for their participation in this International Conference. I am sure that this proceeding will useful to researchers in their fields.

I wish all the success to the conference.

Sasmita Parida

Sasmita Parida

Aryan Institute of Engineering & Technology
Bhubaneswar – 752050, Khurda, India

TEL: 0674-2377-102
e-mail: principal@aryan.ac.in

Dr. S. S. Khuntia, Principal

Principal's Message

The conferences are necessary to bring at the culture of information exchange and feedback on developing trends in technologies. I am delighted to note that the Department of Computer Science Engineering is organizing the International Conference entitled "Recent Trends in IOT and Its Application", RTIA'2018. Certainly, this type of conference not only brings all the researchers, students in one platform, but it also inculcates the research culture among the entire fraternity of Education in the country, thereby, contributing to the development of the nation.

I hope that this conference would certainly induce innovative ideas among the participants paving way for new inventions and technologies in Electronics, Computers, and Communications.

I Congratulate, Mr. Amiya Kumar Sahoo, HOD, Computer Science Engineering, and his team for initiating the conduction of such a conference in our Institute.

I wish the conference a grand success.

Dr. S. S. Khuntia

Aryan Institute of Engineering & Technology
Bhubaneswar – 752050, Khurda, India

TEL: 0674-2377-101
e-mail: hodmechanical@aryan.ac.in

Mr. Amiya Kumar Sahoo, Head of the Dept. of Computer Science Engineering

Convenor's Message

It gives us immense pleasure to invite you at Aryan Institute of Engineering & Technology which is going to organize an International Conference on "Recent Trends in IOT and Its Application", RTIA'18.

New Technologies are introducing every day that will radically transform the future of this fields. Vision of this conference is to promote excellence in scientific knowledge and innovations in the field of computer science, electronics and electrical engineering and other related disciplines to motivate young researchers. The aim of the conference is to provide a forum to researchers around the globe to explore and discuss on various aspects of computer science. The conference consists of various sessions Each session will be addressed by outstanding experts and key note speakers who will highlight the recent advances in various facets of computer science engineering.

I would like to thank technical program committee, local organizing committee, volunteers and the staff members of AIET for their dedicated support. Finally, I would like to thank all the authors, coauthors and volunteers who directly or indirectly contributed to the conference.

I wish the conference for a great success.

Mr. Amiya Kumar Sahoo

ORAL		Session 1	Lunch	Session 2	Tea Break	Session 3	Poster(18:00 - 19:00)	
22-Nov	Room 1	Classification and Regression Trees		Classification and Regression Trees		Sensor Applications and Deployments	22-Nov	Embedded Hardware
	Room 2	Sensor Networks		Sensor Networks		Embedded Hardware		Network Reliability
	Room 3	Sensor Applications and Deployments		Sensor Applications and Deployments		Human Centered Computing		Sensor Applications and Deployments
	Room 4	Human Centered Computing		Human Centered Computing		Security Protocols		Human Centered Computing
23-Nov	Room 1	Security Protocols		Machine Learning		Machine Learning		Machine Learning
	Room 2	Network Management		Network Management		Security Protocols		Security Protocols
	Room 3	Distributed Architectures		Distributed Architectures		World Wide Web	23-Nov	Network Management
	Room 4	Embedded Hardware		Embedded Hardware		Mobile and Wireless Security		Sensor Networks
24-Nov	Room 1	Machine Learning						Classification and Regression Trees
	Room 2	Mobile and Wireless Security						Mobile and Wireless Security
	Room 3	Supervised Learning by Classification						Supervised Learning by Classification

Contents

Robust fuzzy factorization machine with noise clustering-based membership function estimation

Sanjay Kumar Padhi, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sk.padhi2@gmail.com*

Sachikanta Pati, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, sachikantapati98@outlook.com*

Prativa Barik, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, p.barik213@gmail.com*

Romeo Jena, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, romeo_jena2@hotmail.com*

Keywords:
Collaborative filtering
Factorization machines
Membership function
Noise clustering

A B S T R A C T

Factorization machine (FM) is a promising model-based algorithm for collaborative filtering (CF), but can bring inferior performances if datasets include users having low confidence. In this paper, a robust FM model is proposed by introducing the noise clustering-based noise rejection mechanism into Fuzzy FM, which utilizes fuzzy memberships of users for considering the responsibility of each user in FM modeling. By automatically updating fuzzy memberships with user-wise criteria of prediction errors, the FM model is better fitted to reliable users and is expected to improve the generalization ability for predicting the preference degrees of unknown items. The characteristics of the proposed method are demonstrated through numerical experiments with MovieLens movie evaluation data such that the prediction ability for not only the training ratings but also the test ratings of reliable users can be improved by carefully tuning the noise sensitivity weight.

1. Introduction

Collaborative filtering (CF) [1] is a basic technique for tackling with information overload, for which there are a number of memory-based algorithms and model-based algorithms [2,3]. Factorization machine (FM) [4] is a promising model-based algorithm, which has been demonstrated to be efficient in handling sparse datasets. For implementation to real world tasks, several efficient algorithms have been recently developed [5,6]. Besides useful applications of these CF algorithms, however, FM can bring inferior performances if datasets include users having low confidence. In order to reduce the influences of element-wise outliers and noise, several challenges of introducing robust criteria such as bounded set variability [7] and ℓ_1-norm loss [8] have been considered. In the following parts of this paper, another concept of user-wise responsibility is mainly focused.

Fuzzy Factorization machine (Fuzzy FM) [9] is an extension of FM, which considers the responsibility weight of each user in FM modeling. If we have a priori knowledge on when each user lastly posted his/her rating, we can estimate fuzzy membership for evaluating the activity level of each user based on the period from the last rating and construct an FM model, which reflects current trends well. Here, the confidence of users can also be estimated by the non-noise degree in real world tasks, where low-confidence or rare characteristic users, i.e., noise users, may have quite unique preference tendencies and are not suitable for collaborative model construction. In general, however, noise users can be identified only after we reveal the intrinsic preference tendency because we have no a priori knowledge on who are noise users. Then, a promising approach is to construct the preference prediction model in conjunction with noise user identification, simultaneously.

In this paper, automatic estimation of fuzzy memberships is considered for Fuzzy FM under the principle of noise rejection. Fuzzy clustering [10,11] is an unsupervised classification method for extracting intrinsic cluster structures varied in multivariate datasets. Noise fuzzy clustering [12] achieved robust fuzzy clustering by introducing an additional noise cluster, which is designed to have equal distances from all objects. By damping all noise objects into the noise cluster, the cluster prototypes of normal clusters are robustly estimated without the influence of noise. Because the noise clustering concept is reduced to robust estimation of the least square-type modeling [12], the simple noise rejection mechanism was also implemented to fuzzy robust principal component analysis [13] and robust non-negative matrix factorization [14].

In this paper, a robust Fuzzy FM model is proposed by introducing the noise clustering-based noise rejection mechanism into Fuzzy FM. In the iterative procedure of the proposed method, the fuzzy membership of each user is automatically updated in each iteration considering the non-noise degree under the current Fuzzy FM model. Then, the Fuzzy FM model is updated following the responsibility weights of users.

The remaining parts of this paper are organized as follows: Section 2 presents a brief review on FM, Fuzzy FM and noise fuzzy clustering, and then, Section 3 proposes a novel robust Fuzzy FM model. The characteristics of the proposed model are demonstrated in Section 4

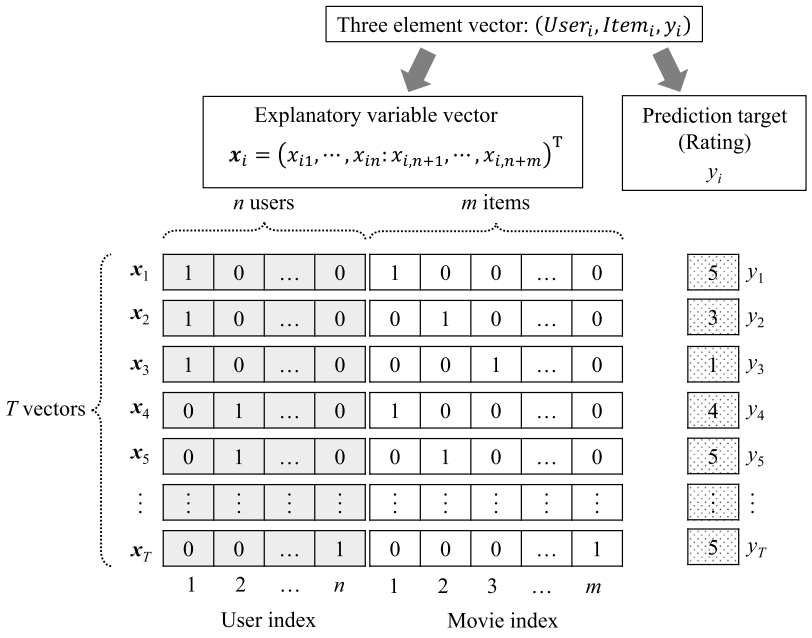


Fig. 1. A sample data representation in FM.

through a CF experiment with a real-world dataset downloaded from MovieLens web site [15], and the summary conclusion is given in Section 5.

2. Factorization machine, Fuzzy FM and noise fuzzy clustering

2.1. Model equation for FM in CF tasks

CF is a basic technique for implementing personalized recommendation, which reduces information overload considering user preferences [1,2]. The problem space of CF is often designed with an evaluation matrix among n users and m items, and the goal of CF is to recommend promising items to the target user such that the (unknown) items are expected to be preferred by the user. In general, we have very many items and each user evaluated only a small portion of them, i.e., the evaluation matrix is very sparse and has many missing elements. Then, the CF task is reduced to missing value estimation in the evaluation matrix [16], where the promising items having higher predicted values are recommended to the user.

Although the memory-based algorithms and their extensions have been utilized in many real applications such as Amazon.com [17], the model-based algorithms are expected to be more computationally efficient in the prediction phase and have higher generalization capability. FM is a promising model-based algorithm, which has been demonstrated to be efficient in handling sparse datasets [4].

Assume that we have evaluation rating data for m types of contents given by n users. In general CF tasks, each user evaluates only a small portion of all contents, and so, the $n \times m$ evaluation matrix is often very sparse such that it contains only T ratings in total, where $T \ll n \cdot m$. In the following, the total T rating scores are assumed to be stored in the form of a three-element vector $(User_i, Item_i, y_i)$ = (person, content, rating), $i = 1, \ldots, T$.[1] Following Rendle's literature [4], the feature vectors utilized in FM modeling are represented as shown in Fig. 1.

Here, the goal of FM modeling is to predict the response variable y_i from the explanatory variable vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in}: x_{i,n+1}, \ldots, x_{i,n+m})^{\top}$, where input information $User_i$ and $Item_i$ are handled as categorical variables and are jointed into a single high-dimensional sparse vector. The model equation for FM of degree $d = 2$ is defined as:

$$\begin{aligned}\hat{y}(\boldsymbol{x}) &= w_0 + \sum_{j=1}^{n+m} w_j x_j + \sum_{j=1}^{n+m-1} \sum_{\ell=j+1}^{n+m} \langle \boldsymbol{v}_j, \boldsymbol{v}_\ell \rangle x_j x_\ell \\ &= w_0 + \sum_{j=1}^{n+m} w_j x_j \\ &\quad + \sum_{j=1}^{n+m-1} \sum_{\ell=j+1}^{n+m} (v_{j1} v_{\ell 1} + v_{j2} v_{\ell 2}) x_j x_\ell, \end{aligned} \tag{1}$$

where w_0 is the constant term, w_j $(j = 1, \ldots, n+m)$ are the coefficients of linear term x_j and $\boldsymbol{v}_j$ $(j = 1, \ldots, n+m)$ are the coefficients of the interaction term, where the dimension k of the latent term v was set to 2 in the followings. The larger the dimension k, the greater the nonlinearity of the regression model and the greater the complexity of the prediction, but the more likely it results in the overtraining situation. On the other hand, the smaller the dimension k, the greater the generalization ability and the better it is suitable for analyzing sparse data. The prediction model is further extended to a higher degree case with $d > 2$, where higher order interaction among three or more variables can be considered.

The objective function to be minimized based on the least-squares error criterion is defined as Eq. (2), in which regularization terms are added to prevent overfitting.

$$\begin{aligned} J_{FM} &= \frac{1}{2} \sum_{i=1}^{T} (y_i - \hat{y}(\boldsymbol{x}_i))^2 + \frac{\lambda_0}{2} \cdot (w_0)^2 \\ &\quad + \frac{\lambda_1}{2} \sum_{j=1}^{n+m} (w_j)^2 + \frac{\lambda_2}{2} \sum_{j=1}^{n+m} \sum_{f=1}^{k} (v_{jf})^2. \end{aligned} \tag{2}$$

Based on the Stochastic Gradient Descent (SGD), each parameter is iteratively updated.

Here, the calculation order of the interaction term $\mathcal{O}(k(n+m)^2)$ can be reduced to a linear order $\mathcal{O}(k(n+m))$ as Eq. (3), where we can ignore the model parameters that directly depends on two variables (j, ℓ).

$$\sum_{j=1}^{n+m} \sum_{\ell=j+1}^{n+m} \langle \boldsymbol{v}_j, \boldsymbol{v}_\ell \rangle x_j x_\ell$$

[1] In [4], other input information such as the other rating items and the rating time in a certain period can be also jointed including both categorical and numerical values. However, for simplicity, $user \times item$ information is only considered in this paper.

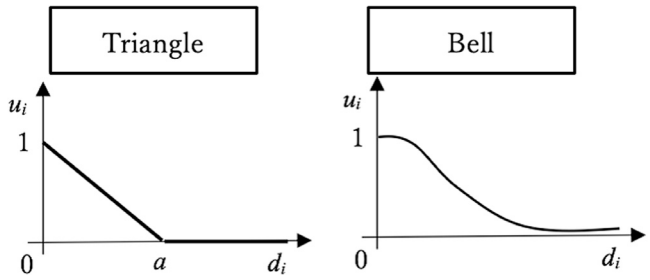


Fig. 2. Pre-fixed fuzzy memberships in Fuzzy FM.

$$
\begin{aligned}
&= \frac{1}{2}\sum_{j=1}^{n+m}\sum_{\ell=1}^{n+m}\langle \boldsymbol{v}_j, \boldsymbol{v}_\ell\rangle x_j x_\ell - \frac{1}{2}\sum_{j=1}^{n+m}\langle \boldsymbol{v}_j, \boldsymbol{v}_j\rangle x_j x_j \\
&= \frac{1}{2}\left(\sum_{j=1}^{n+m}\sum_{\ell=1}^{n+m}\sum_{f=1}^{k} v_{jf}v_{\ell f}x_j x_\ell - \sum_{j=1}^{n+m}\sum_{f=1}^{k} v_{jf}v_{jf}x_j x_j\right) \\
&= \frac{1}{2}\sum_{f=1}^{k}\left(\left(\sum_{j=1}^{n+m} v_{jf}x_j\right)\left(\sum_{\ell=1}^{n+m} v_{\ell f}x_\ell\right) - \sum_{j=1}^{n+m} v_{jf}^2 x_j^2\right) \\
&= \frac{1}{2}\sum_{f=1}^{k}\left(\left(\sum_{j=1}^{n+m} v_{jf}x_j\right)^2 - \sum_{j=1}^{n+m} v_{jf}^2 x_j^2\right). \qquad (3)
\end{aligned}
$$

2.2. Fuzzy Factorization Machine

When creating a model for FM, if the data of each user in the training dataset have a variation from the reliability viewpoint, and the reliability can be available as a prior knowledge with fuzzy memberships, then, the accuracy of the prediction model can be improved by a weighted objective function of the fuzzy memberships. For example, the ratings of users, who have reported their ratings until recently, are considered to be highly reliable because they represent the latest preferences, while the ratings of users, who have not reported their ratings for a long period of time, are considered to be less reliable because they may have different trends from the latest preferences.

Zhou proposed Fuzzy FM [9], which incorporates the reliability of a prior knowledge as fuzzy memberships in the FM objective function. When d_i is the number of days since the evaluator of the ith observation last reported the evaluation value, the fuzzy membership u_i to the model representing *the latest preference* can be defined by a function such as Fig. 2.

The weighted objective function with fuzzy memberships is defined as follows:

$$
\begin{aligned}
J_{FFM} &= \frac{1}{2}\sum_{i=1}^{T} u_i(y_i - \hat{y}(\boldsymbol{x}_i))^2 + \frac{\lambda_0}{2}\cdot(w_0)^2 \\
&\quad + \frac{\lambda_1}{2}\sum_{j=1}^{n+m}(w_j)^2 + \frac{\lambda_2}{2}\sum_{j=1}^{n+m}\sum_{f=1}^{k}(v_{jf})^2. \qquad (4)
\end{aligned}
$$

This will allow us to estimate a model that can accurately predict values of users having large fuzzy memberships u_i.

2.3. Noise fuzzy clustering

The goal of k-Means-type clustering [18,19] is to partition n objects $\boldsymbol{x}_i, i = 1, \ldots, n$ into C clusters such that intra-cluster objects are mutually similar but inter-cluster objects are not. Fuzzy c-Means (FCM) [10] is a basic fuzzy clustering model, where each cluster c is represented by its cluster centroid $\boldsymbol{b}_c$ and each object i is assigned to multiple clusters with fuzzy memberships u_{ci} of the degree of belongingness to cluster c. Under the probabilistic constraint of $\sum_{c=1}^{C} u_{ci} = 1$, the FCM objective function to be minimized is defined as:

$$
J_{FCM} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2, \qquad (5)
$$

where θ ($\theta > 1$) is the weighting exponent for membership fuzzification such that the larger the value of θ, the fuzzier the cluster boundaries. If $\theta \to 1$, the model is reduced to the crisp k-Means. In general, $\theta = 2$ is often used as the standard setting.

In order to reduce the noise sensitivity of the least square-type objective function, Davé proposed noise fuzzy clustering [12], where an additional noise cluster was introduced for damping all noise objects into it. When we use the $(C+1)$-th cluster as the noise cluster, to which all objects have equal distances, the FCM objective function is modified as:

$$
J_{NFCM} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2 + \sum_{i=1}^{n} u_{C+1,i}^{\theta}\gamma, \qquad (6)
$$

where γ is the adjustable weight for tuning noise sensitivity, which is called the *noise sensitivity weight* in this paper, and is identified with the fixed distance between each object and the noise cluster. Here, the probabilistic constraint is also modified as $\sum_{c=1}^{C+1} u_{ci} = 1$. If an object is more distant than γ from all normal clusters, the object is damped into the noise cluster having large membership $u_{C+1,i}$.

When the number of normal cluster is $C = 1$, the noise fuzzy clustering model is reduced to the robust centroid estimation [20] and the concept can be utilized in robust least square modeling by replacing the FCM criterion with other analytical criteria such as robust principal component analysis [13] and robust non-negative matrix factorization [14].

In the next section, the above noise clustering concept is applied to robust FM modeling by utilizing it for fuzzy membership estimation in Fuzzy FM.

3. Robust Fuzzy FM based on noise fuzzy clustering concept

3.1. Automated fuzzy membership estimation in Fuzzy FM

This paper proposes a novel Robust Fuzzy FM model, which introduces the noise rejection mechanism of noise fuzzy clustering into Fuzzy FM. Instead of relying on a prior knowledge to fix fuzzy memberships in Fuzzy FM, the proposed model considers updating the fuzzy memberships automatically by degrading the fuzzy membership of noise users, who are unsuitable in model estimation, and increasing the contribution of non-noise users.

In order to express the non-noise degree of each user with a fuzzy membership, the membership of user a, $a = 1, \ldots, n$ in model estimation is designed as $u_a (u_a \in [0, 1])$, where the degree of belongingness to the noise cluster corresponds to $1 - u_a$. Here, user-wise prediction errors of $e_a = \sum_{i \in I_a}(y_i - \hat{y}(\boldsymbol{x}_i))^2$ are used as the clustering criterion in the noise clustering context, where I_a represents the set of items evaluated by user a such that $\sum_{a=1}^{n} |I_a| = T$. Then, the objective function to be minimized is defined as:

$$
\begin{aligned}
J_{RFFM} &= \sum_{a=1}^{n}\left(u_a^{\theta}\sum_{i\in I_a}(y_i - \hat{y}(\boldsymbol{x}_i))^2 + (1-u_a)^{\theta}\sum_{i\in I_a}\gamma\right) \\
&\quad + \frac{\lambda_0}{2}\cdot(w_0)^2 + \frac{\lambda_1}{2}\sum_{j=1}^{n+m}(w_j)^2 \\
&\quad + \frac{\lambda_2}{2}\sum_{j=1}^{n+m}\sum_{f=1}^{k}(v_{jf})^2 \\
&= \sum_{a=1}^{n}\left(u_a^{\theta}e_a + (1-u_a)^{\theta}\sum_{i\in I_a}\gamma\right) \\
&\quad + \frac{\lambda_0}{2}\cdot(w_0)^2 + \frac{\lambda_1}{2}\sum_{j=1}^{n+m}(w_j)^2 \\
&\quad + \frac{\lambda_2}{2}\sum_{j=1}^{n+m}\sum_{f=1}^{k}(v_{jf})^2. \qquad (7)
\end{aligned}
$$

γ is the noise weight, which implies that all objects have common weights to the noise cluster. If the Fuzzy FM model can estimate the

rating of object a with an average error less than γ, then the non-noise weight u_a will be a large value, while if the approximation error is greater than γ, then $1-u_a$ will become larger and the object will be assigned to the noise cluster being rejected from Fuzzy FM model estimation. In other words, the smaller the value of γ is, the more objects are absorbed as noise. The θ is the weighting exponent for fuzzy memberships, and the larger the value is, the fuzzier the memberships are. By the way, the proposed model is reduced to the conventional (non-fuzzy) FM when $u_a = 1, \forall a$.

3.2. Parameter updating formulas and sample algorithm

The model parameters are updated by iterative hybrid implemen-tation of the alternative optimization for fuzzy membership estimation and the Stochastic Gradient Descent (SGD) for FM model estimation.

Considering the optimality condition $\partial J_{RFFM}/\partial u_a = 0$, the updating formula for u_a is derived as follows:

$$u_a = \left(1 + \left(\frac{e_a}{\gamma |I_a|}\right)^{\frac{1}{\theta-1}}\right)^{-1}, \tag{8}$$

$$e_a = \sum_{i \in I_a} (y_i - \hat{y}(\boldsymbol{x}_i))^2. \tag{9}$$

Next, FM parameters are calculated by SGD as follows:

$$\begin{aligned} w_0 &\leftarrow w_0 - \eta \frac{\partial J_{RFFM}}{\partial w_0} \\ &= w_0 - \eta \left(\sum_{i=1}^{T} \frac{\partial J_{RFFM}}{\partial \hat{y}(\boldsymbol{x}_i)} \cdot \frac{\partial \hat{y}(\boldsymbol{x}_i)}{\partial w_0} + \lambda_0 \cdot w_0 \right), \end{aligned} \tag{10}$$

$$\begin{aligned} w_j &\leftarrow w_j - \eta \frac{\partial J_{RFFM}}{\partial w_j} \\ &= w_j - \eta \left(\sum_{i=1}^{T} \frac{\partial J_{RFFM}}{\partial \hat{y}(\boldsymbol{x}_i)} \cdot \frac{\partial \hat{y}(\boldsymbol{x}_i)}{\partial w_j} + \lambda_1 \cdot w_j \right), \end{aligned} \tag{11}$$

$$\begin{aligned} v_{jf} &\leftarrow v_{jf} - \eta \frac{\partial J_{RFFM}}{\partial v_{jf}} \\ &= v_{jf} - \eta \left(\sum_{i=1}^{T} \frac{\partial J_{RFFM}}{\partial \hat{y}(\boldsymbol{x}_i)} \cdot \frac{\partial \hat{y}(\boldsymbol{x}_i)}{\partial v_{jf}} + \lambda_2 \cdot v_{jf} \right). \end{aligned} \tag{12}$$

Here, each gradient value is derived as follows:

$$\frac{\partial J_{RFFM}}{\partial \hat{y}(\boldsymbol{x}_i)} = -2u^{\theta}_{\phi(i)}(y_i - \hat{y}(\boldsymbol{x}_i)), \tag{13}$$

where the function $\phi(i)$ outputs user index a, who evaluated the ith rating such as $i \in I_a$.

$$\frac{\partial \hat{y}(\boldsymbol{x}_i)}{\partial w_0} = 1, \tag{14}$$

$$\frac{\partial \hat{y}(\boldsymbol{x}_i)}{\partial w_j} = x_{ij}, \tag{15}$$

$$\frac{\partial \hat{y}(\boldsymbol{x}_i)}{\partial v_{jf}} = x_{ij} \cdot \sum_{\ell=1}^{n+m} v_{\ell f} x_{i\ell} - v_{jf} x_{ij}^2. \tag{16}$$

Utilizing the above updating formulas, a sample algorithm for the proposed Robust Fuzzy Factorization Machine (Robust Fuzzy FM) is described as follows:

Step 1 Initialize FM parameters w_0, w_j and v_{jf} by implementing the conventional FM and set all fuzzy memberships as $u_a = 1, \forall a$. (This process is equivalent to implementation of Fuzzy FM with fixed fuzzy memberships of $u_a = 1, \forall a$.)

Step 2 Update fuzzy memberships u_a by Eq. (8).

Step 3 Update FM parameters w_0, w_j and v_{jf} by Eqs. (10)–(12).

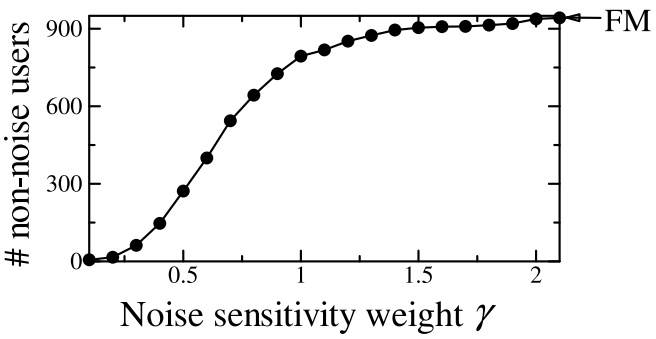


Fig. 3. Number of non-noise users.

Step 4 If all FM parameters were convergent, stop. Otherwise, return to **Step 2**.

In the proposed Robust Fuzzy FM algorithm, **Step 2** of updating fuzzy memberships u_a is responsible for the noise rejection mechanism, which is newly added into the conventional Fuzzy FM. The calculation of Eq. (8) just needs the iterative calculation of Eq. (9) for $i \in I_a$, where $|I_a|$ is generally much lesser than the number of contents such that $|I_a| \ll m$. So, the additional calculation cost for all $a \in \{1, 2, \ldots, n\}$ is $\mathcal{O}(T)$, $T \ll n \cdot m$ and is comparative with other FM calculations such as Eqs. (10)–(12).

4. Numerical experiments

This section demonstrates the characteristics of the proposed method using a benchmark dataset of MovieLens 100 K dataset.

4.1. Dataset

MovieLens100k dataset [15] is a benchmark dataset available at MovieLens web site of the University of Minnesota. The *MovieLens100k* dataset contains 100,000 ratings ($y_i \in \{1, 2, 3, 4, 5\}$) for 1682 movies by 943 users, where a larger y_i implies a better rating. Each user rated at least 20 movies. Then, each rating score is given in a 3-D vector information as $(User_i, Movie_i, y_i)$ = (user ID, movie ID, rating score), $i = 1, \ldots, T$. The sparse feature vectors for FM modeling are composed of the total $2625 = 943 + 1682$ elements as Fig. 1, where each categorical information has $x_{ij} \in \{0, 1\}$.

Here, prediction performance is evaluated with the pre-partitioned subsets of 'ua.base' and 'ua.test', which are designed in *ml-100k* for training and testing, respectively. The size of the dataset is 90570×2625 for training set and 9430×2625 for test set, where the test set includes exactly 10 ratings per user.

4.2. Effects of noise sensitivity weight in fuzzy membership estimation

The proposed Robust Fuzzy FM was implemented for the training dataset with various noise sensitivity weight γ drawn from the interval $\gamma \in [0.1, 2.0]$ at intervals of 0.1, and their results are compared with that of the conventional FM, which corresponds to the proposed method with $\gamma \to \infty$, for demonstrating the characteristics of the proposed method. The other model parameters were set as $\lambda_0 = \lambda_1 = \lambda_2 = 0.0001$, $\eta = 0.00001$ and $\theta = 2.0$.

First, the derived fuzzy memberships are compared for studying the effect of the noise rejection mechanism. Fig. 3 shows the number of non-noise users who were not rejected by the noise clustering mechanism, where defuzzification into the noise/non-noise crisp classes were performed such that users with $u_a > 0.5$ are assigned to the non-noise class while users with $u_a \leq 0.5$ are to the noise class. When the noise sensitivity γ is 0.1, the proposed method is very sensitive to noise and

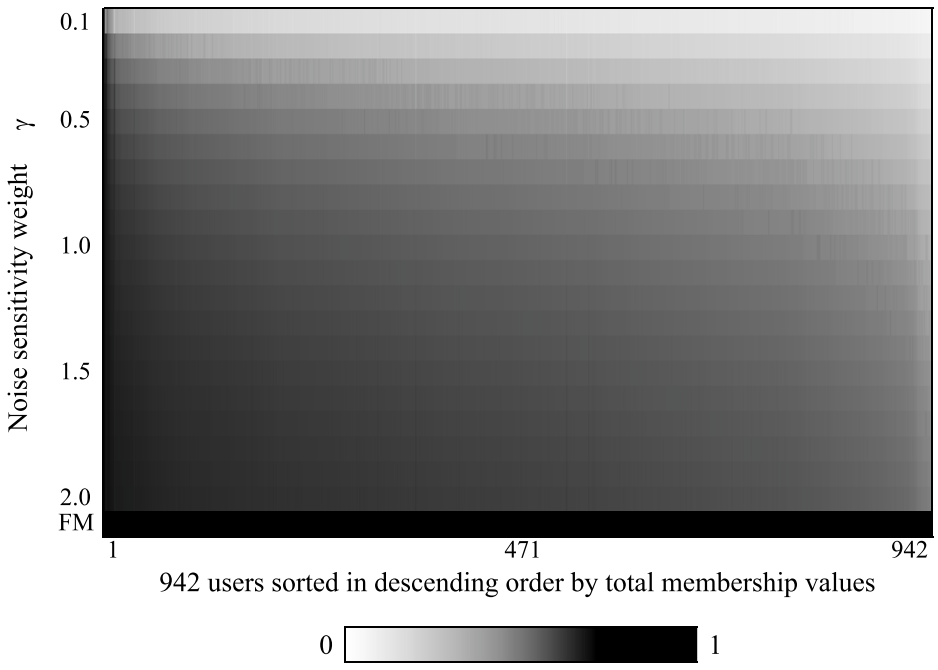


Fig. 4. Gradual change of fuzzy memberships.

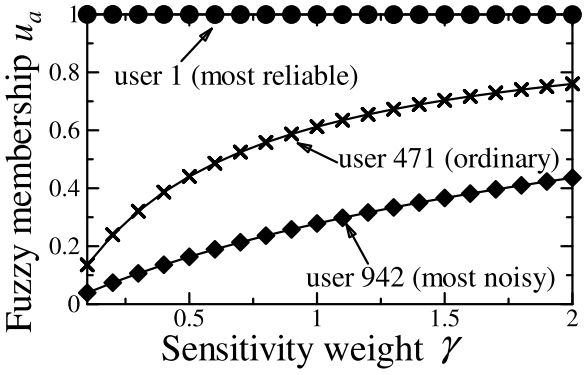


Fig. 5. Gradual change of fuzzy memberships of three representatives.

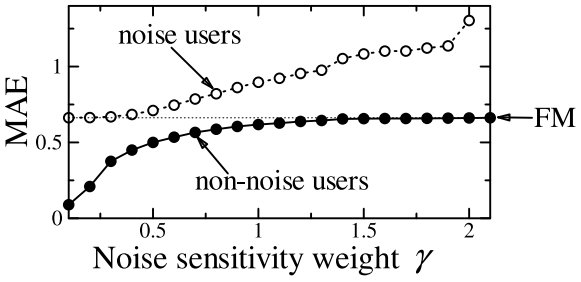


Fig. 6. MAE in rating value estimation for training dataset.

almost all users were assigned to the noise class. As γ becomes larger, the number of non-noise users are gradually increased, and when $\gamma >$ 2.0, almost all users were assigned to the non-noise cluster.

The gradual change of fuzzy memberships is further investigated in Fig. 4, which depicts each membership value u_a given with each γ value in grayscale representation. A brighter color implies more noisy characteristic. For easy comparison, the user IDs listed on the horizontal axis are sorted in descending order of the sum of membership values, i.e., the most left users are mostly non-noise users while the most right users are mostly noisy users. Additionally, in order to support intuitive understanding, three representative users, i.e., most reliable (non-noise) user (*user* 1), ordinary user (*user* 471) and most noisy user (*user* 942), were excerpted and their memberships are depicted again in Fig. 5. The figures demonstrate that the non-noise membership is kept almost $u_a = 1$ for reliable users while it is gradually degraded for ordinary users. Contrarily, noisy users have only a small membership and were rejected from FM modeling even when γ was relatively large. Then, the proposed automatic fuzzy membership updating scheme is useful in gradually tuning the noise rejection level in FM model training.

4.3. Evaluation of prediction ability for test dataset

Second, the effect of noise user rejection in FM modeling is studied from the viewpoint of missing value prediction in the CF task. As demonstrated in the previous subsection, in the FM training phase, we have a smaller number of non-noise users as the noise sensitivity weight γ becomes smaller. Fig. 6 shows the mean absolute errors (MAE) in the rating value estimation for the training dataset, where the MAE values are compared between noise/non-noise classes after defuzzification of fuzzy memberships. Comparing with the performance of the conventional FM depicted in the horizontal dotted line, the prediction performance for non-noise class having smaller numbers of users were gradually improved as γ becomes smaller in $\gamma < 1.0$ while MAE for noise class users becomes closer to the conventional FM. So, the derived FM model with smaller γ achieved better matching to the rating scores given by a few reliable users, where majority users were absorbed into the noise cluster.

Next, the prediction performance for the test dataset is studied. Fig. 7 shows MAE for the test dataset, where the performance of noise/non-noise classes are compared with that of the conventional FM for both training and test datasets. Generally, because the FM model is learned for well-fitting to the training dataset, the prediction errors for the test dataset, which were not used in the training phase, become worse than the training one. However, Fig. 7 indicates that the generalization capability of the proposed robust FM model was

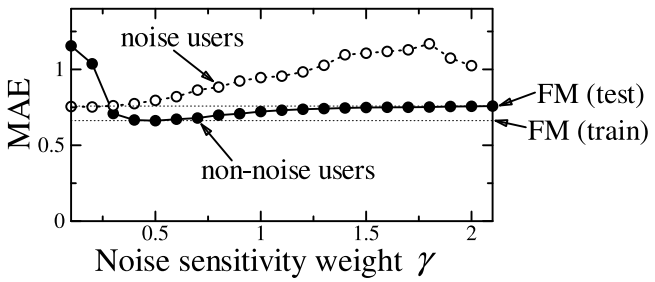


Fig. 7. MAE in rating value estimation for testing dataset.

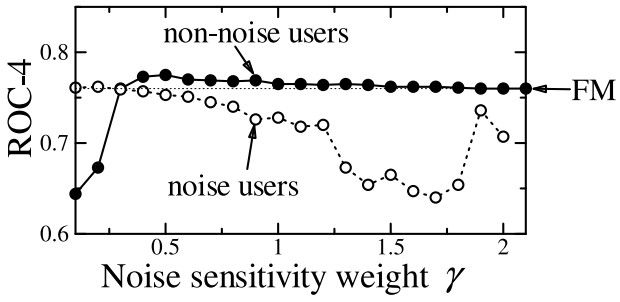


Fig. 8. ROC-4 of recommendation ability for testing dataset.

improved for non-noise users such that MAE became the smallest in around $\gamma = 0.5$. That is, the conventional FM model was severely overfitted to the noise users and caused significant deterioration in test prediction. On the other hand, the proposed method improves the fitting to the general features of reliable users, whose rating tendencies are consistent with others.

Especially, the test prediction for the reliable (non-noise) users achieved almost a comparative performance with the training prediction of the conventional FM model in around $\gamma \in [0.4, 0.6]$. Then, the proposed method demonstrated the high generalization capability in CF tasks.

The similar feature can be found from the viewpoint of recommendation ability. Fig. 8 compares the receiver operating characteristic (ROC) sensitivity criteria [21,22] for the test dataset. The ROC curve is a true positive rate vs. false positive rate plots drawn by changing the threshold of the applicability level in recommendation, and the lower area of the curve becomes large as the recommendation ability is higher. In this paper, the movie whose rating score is $y_i = 4$ or larger were to be recommended and the criterion is called *ROC-4* [16].

In Fig. 8, the best *ROC-4* was again achieved at $\gamma = 0.5$ for the non-noise class, where MAE was also minimum. Then, we can see that the proposed robust FM model simultaneously achieved both the minimum error prediction and the best recommendation ability by carefully tuning the sensitivity weight γ.

4.4. Comparison with pre-fixed fuzzy memberships in conventional Fuzzy FM

Third, the performance of the proposed method is compared with the conventional Fuzzy FM, in which fuzzy memberships are pre-fixed considering a priori knowledge. In order to estimate the reliability degree of each user a, this experiment utilizes the following two membership functions:

$$u_a^{act} = \exp\left(-\frac{(time_{last} - time_a)^2}{\gamma(time_{last} - time_{start})^2}\right), \tag{17}$$

$$u_a^{freq} = \exp\left(-\frac{(count_{max} - count_a)^2}{\gamma(count_{max} - count_{min})^2}\right), \tag{18}$$

where $time_a$ and $count_a$ are the time stamp (Unix seconds since 1/1/1970 UTC) of last movie evaluation by user a and the number of movies evaluation by user a, respectively. u_a^{act} is designed for weakening the responsibility of non-active users, who did not post their movie evaluation recently. On the other hand, u_a^{freq} is designed for weakening the responsibility of non-frequent users, who rated a relatively small number of movies only. The sensitivity weight γ tunes the shapes of membership functions such that a smaller γ causes rejection of a larger number of users. Here, in the conventional Fuzzy FM, these fuzzy memberships are pre-fixed before application and are not updated.

Fig. 9 shows the numbers of active/frequent users to have high responsibility in FM modeling with various γ values. The figures form similar trajectories to Fig. 3 and the pre-fixed u_a^{act} and u_a^{freq} seem to play similar roles to the proposed non-noise fuzzy memberships of Eq. (8) by tuning γ.

Next, the prediction ability is considered. Comparing Figs. 10 and 11 with Figs. 6 and 7, the pre-fixed activity/frequency memberships cannot control the noise rejection level by tuning the sensitivity weights and are not available for robust FM modeling since the two partitioned classes, i.e., active/non-active or frequent/non-frequent, do not have significant differences in MAE.

Therefore, robust FM modeling should be implemented with adaptive fuzzy memberships utilizing criteria of prediction errors like the proposed Robust Fuzzy FM method.

5. Conclusions

In this paper, a novel robust FM model was proposed by introducing the noise clustering-based robust modeling scheme into Fuzzy FM. By automatically updating the non-noise fuzzy membership of each object, the FM prediction model is constructed by rejecting the influence of noise objects. The characteristics of the proposed method were demonstrated through numerical experiments with *MovieLens100k* benchmark dataset such that the robust training model can also improve the test prediction for the reliable (non-noise) users.

In order to improve the usability of the proposed method, the following issues can be considered in our future works. First, how to select the optimal sensitivity weight γ should be investigated. In this paper, the basic characteristics of the proposed method were demonstrated by comparing the prediction results with various γ values but their optimality was only intuitively discussed. More subjective discussions would be needed.

Second, the applicability to much more larger datasets should be investigated. Compared with the conventional Fuzzy FM with pre-fixed memberships, the proposed algorithm includes an additional membership updating step. The reduction of its computational cost can be considered.

Third, the convergence characteristics can be investigated. Although the fuzzy membership calculation based on iterative optimization principle is useful in finding local optimal solutions, its hybridization with SGD may not be an effective idea. It is a possible future work to study the condition for ensuring the convergence characteristics.

Fourth, the extension to a collaborative system [23] can be considered. Although FM is useful in utilizing multiple information in the virtual joint vector like Fig. 1, multi-source information can be jointed under privacy preservation in many real applications. Introduction of collaborative analysis scheme is a possible practical issue.

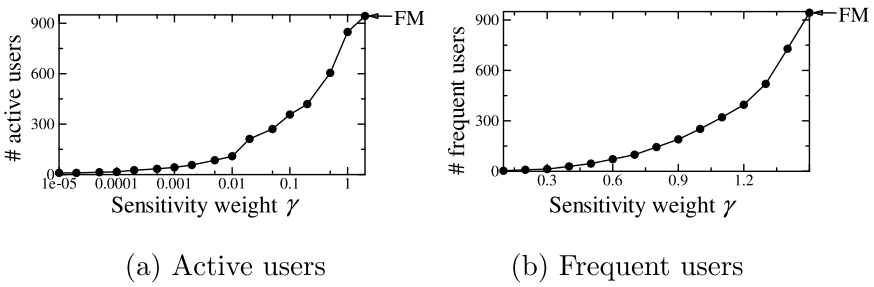


Fig. 9. Number of active/frequent users considering u_a^{act} and u_a^{freq}.

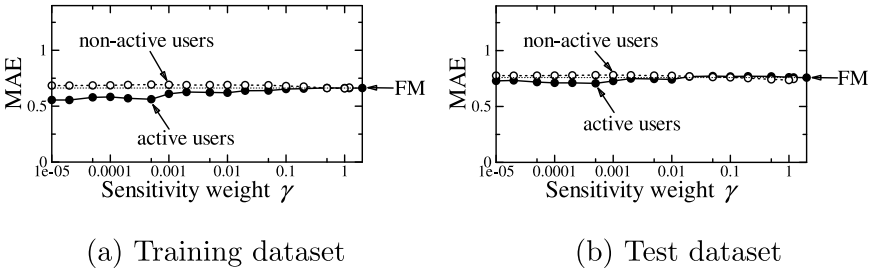


Fig. 10. MAE in rating value estimation with activity membership.

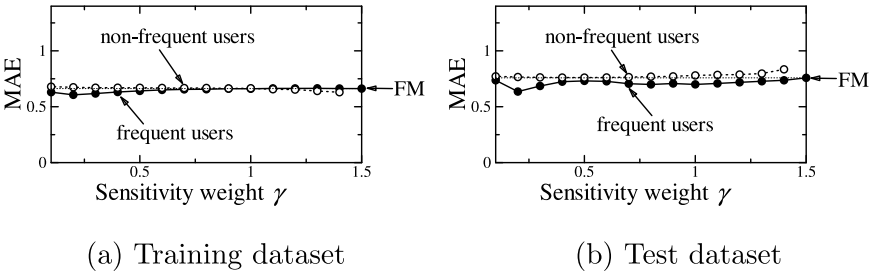


Fig. 11. MAE in rating value estimation with frequency membership.

References

[1] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gardon, J. Riedl, Grouplens: Applying collaborative filtering to usenet news, Commun. ACM 40 (1999) 77–87.
[2] X. Su, T.M. Khoshgoftaar, A survey of collaborative filtering techniques, Adv. Artif. Intell. 2009 (2009) 421425.
[3] C.C. Aggarwal, Recommender Systems, Springer International Publishing, 2016.
[4] S. Rendle, Factorization machines, in: Proc. the 2010 IEEE International Conference on Data Mining, 2010, pp. 995–1000.
[5] I. Bayer, FastFM: A library for factorization machines, J. Mach. Learn. Res. 17 (2016) 1–5.
[6] G. Jiang, H. Wang, J. Chen, H. Wang, D. Lian, E. Chen, xLightFM: Extremely memory-efficient factorization machine, in: Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 337–346.
[7] S. Punjabi, P. Bhatt, Robust factorization machines for user response prediction, in: Proc. 2018 World Wide Web Conference, 2018, pp. 669–678.
[8] C.H. Liu, T. Zhang, J.D. Li, J.W. Yin, P.L. Zhao, J.L. Sun, S.C.H. Hoi, Robust factorization machine: A doubly capped norms minimization, in: Proc. 2019 SIAM International Conference on Data Mining, 2019, pp. 738–746.
[9] J. Zhou, Fuzzy factorization machine, Inform. Sci. 546 (2021) 1135–1147.
[10] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, NY, 1981.
[11] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering, Springer-Verlag, Berlin Heidelberg, 2008.
[12] R.N. Davé, Characterization and detection of noise in clustering, Pattern Recognit. Lett. 12 (1991) 657–664.
[13] K. Honda, A. Notsu, H. Ichihashi, Fuzzy PCA-guided robust k-means clustering, IEEE Trans. Fuzzy Syst. 18 (2010) 67–79.
[14] M. Ueno, K. Honda, S. Ubukata, A. Notsu, Robust non-negative matrix factorization based on noise fuzzy clustering mechanism, in: Proc. 2019 2nd Artificial Intelligence and Cloud Computing Conference and 2019 Asia Digital Image Processing Conference, 2019, pp. 1–5.
[15] F.M. Harper, J.A. Konstan, The MovieLens datasets: history and context, ACM Trans. Interact. Intell. Syst. 5 (2015) 19.
[16] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 230–237.
[17] B. Smith, G. Linden, Two decades of recommender systems at Amazon.com, IEEE Internet Comput. 21 (2017) 12–18.
[18] J.B. MacQueen, Some methods of classification and analysis of multivariate observations, in: proc. 5th berkeley symposium on math. stat. and prob, 1967, pp. 281–297, 1967.

[19] J.-J. Wu, Advances in K-Means Clustering, Springer, Berlin, Heidelberg, 2012.
[20] R.N. Davé, R. Krishnapuram, Robust clustering methods: A unified view, IEEE Trans. Fuzzy Syst. 5 (1997) 270–293.
[21] J.A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1289.
[22] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874.
[23] T.-C.T. Chen, K. Honda, Fuzzy Collaborative Forecasting and Clustering, Springer International Publishing, Switzerland, 2020.

Optical solitons with Biswas–Milovic equation in magneto-optic waveguide having Kudryashov's law of refractive index

Sangita Pal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sangitapal2@outlook.com*

Madhusudan Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, madhusudandas55@hotmail.com*

Prangya Paramita Padhi, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, prangya.p.padhi@gmail.com*

Subhendu Sahoo, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, s.sahoo95@outlook.com*

ARTICLE INFO

Keywords:
Kudryashov
Biswas–Milovic
Solitons

ABSTRACT

In the present study, we introduce the enhanced Kudryashov's algorithm which takes full advantage of the well-known Kudryashov's method and the new Kudryashov's method to extract optical solitons for the Biswas–Milovic equation in magneto-optic waveguide coupled system having Kudryashov's law of refractive index. Bright, dark and singular soliton solutions are retrieved. The obtained solitons appear with appropriate constraints to guarantee the existence of these solitons.

1. Introduction

Optical solitons is a core area of research in the field of nonlinear optics and telecommunication industry. The main point on the study of propagation of these solitons is with optical fibers, magneto-optic waveguides, metamaterials and metasurfaces, DWDM systems, FBGs and several other devices. So, the study of theses solitons has gained a great attention in the last years and many researchers have developed several integration schemes to study this dynamics of soliton propagation [1–15]. There is a wide variety of integration algorithms such as the modified simple equation method [16–19], (G'/G)-expansion method [20], trial function approach [21], extended trial function method [22], the improved modified extended tanh-function technique [23], F-expansion method [24–26], Kudryashov's method [27,28], the new Kudryashov's approach [29] and so on. In the present work, we seek optical solitons of the Biswas–Milovic equation in magneto-optic waveguide coupled system having Kudryashov's law of refractive index. The model is given by [30]

$$i\frac{\partial q^m}{\partial t} + a_1\frac{\partial^2 q^m}{\partial x^2} + \left(\frac{b_1}{|q|^{2n}} + \frac{c_1}{|q|^n} + d_1|q|^n + e_1|q|^{2n} + \frac{f_1}{|r|^{2n}} + \frac{g_1}{|r|^n} + h_1|r|^n + k_1|r|^{2n}\right) q^m$$
$$= Q_1 r^m + i\left\{\zeta_1\frac{\partial q^m}{\partial x} + \alpha_1\frac{\partial}{\partial x}\left(q^m|q|^{2n}\right) + \theta_1 q^m\frac{\partial}{\partial x}\left(|q|^{2n}\right) + \mu_1|q|^{2n}\frac{\partial q^m}{\partial x}\right\}. \quad (1)$$

$$i\frac{\partial r^m}{\partial t} + a_2\frac{\partial^2 r^m}{\partial x^2} + \left(\frac{b_2}{|r|^{2n}} + \frac{c_2}{|r|^n} + d_2|r|^n + e_2|r|^{2n} + \frac{f_2}{|q|^{2n}} + \frac{g_2}{|q|^n} + h_2|q|^n + k_2|q|^{2n}\right) r^m$$
$$= Q_2 q^m + i\left\{\zeta_2\frac{\partial r^m}{\partial x} + \alpha_2\frac{\partial}{\partial x}\left(r^m|r|^{2n}\right) + \theta_2 r^m\frac{\partial}{\partial x}\left(|r|^{2n}\right) + \mu_2|r|^{2n}\frac{\partial r^m}{\partial x}\right\}. \quad (2)$$

In Eqs. (1) and (2), $a_l(l = 1, 2)$ represents the coefficients of the chromatic dispersion. The parameters b_l, c_l, d_l, and e_l represent the self-phase modulation coefficients, while f_l, g_l, h_l, and k_l are cross-phase modulation coefficients. On the right-hand side, Q_l represents the magnetic field effect that avoids the formation of soliton clutter. From the perturbation terms, ζ_l are the coefficients of intermodal dispersion. Also, α_l represents the coefficients of self-steepening terms in order to avoid shockwave formation, θ_l, μ_l

are the coefficients of nonlinear dispersion, while n, m give the power nonlinearity and the maximum intensity. The enhanced Kudryashov's algorithm will be implemented to get solitons of the governing model. This algorithm takes full advantage of the well-known Kudryashov's method and the new Kudryashov's method. In the next section, we give an overview of the new approach.

2. The enhanced Kudryashov's method

Considering the nonlinear evolution equation (NLEE) as follows:

$$F(u, u_x, u_t, u_{xt}, u_{xx}, \ldots) = 0. \tag{3}$$

where $u = u(x, t)$ is an unknown function, F is a polynomial in u and its temporal and spatial independent variables.

The central proceedings of the enhanced Kudryashov's method as follows:

Step–1: By using the following transformation:

$$u(x, t) = U(\xi), \quad \xi = k(x - vt), \tag{4}$$

where k, v are constant to be determined later. Then Eq. (3) is reduced to a nonlinear ordinary differential equation of the form

$$P(U, -kvU', kU', k^2U'', \ldots) = 0. \tag{5}$$

Step–2: Assuming that the solution of Eq. (5) can be expressed in the form

$$U(\xi) = \lambda_0 + \sum_{l=1}^{N} \sum_{i+j=l} \lambda_{ij} Q^i(\xi) R^i(\xi), \tag{6}$$

where $\lambda_0, \lambda_{ij}(i, j = 0, 1, \ldots, N)$ are constants to be determined and the functions $R(\xi)$ and $Q(\xi)$ satisfy the following ODEs:

$$R'(\xi)^2 = R(\xi)^2(1 - \chi R(\xi)^2), \tag{7}$$

$$Q'(\xi) = Q(\xi)(\eta Q(\xi) - 1). \tag{8}$$

The solutions of (7) and (8) are respectively

$$R(\xi) = \frac{4a}{4a^2 e^{\xi} + \chi e^{-\xi}}, \tag{9}$$

$$Q(\xi) = \frac{1}{\eta + be^{\xi}}, \tag{10}$$

where a, b, η and χ are arbitrary constants.

Step–3: Determining the positive integer number N in Eq. (6) by balancing the highest order derivatives and the nonlinear term in Eq. (5).

Step–4: Substituting (6) into (5) along with (7) and (8). As a result of this substitution, we get a polynomial of $Q(\xi)$, $R(\xi)$ and $R'(\xi)$. In this polynomial we gather all terms of same powers and equating them to be zero, we get an over-determined system of algebraic equations which can be solved by the Maple or Mathematica to get the unknown parameters $k, v, a, b, \eta, \chi, \lambda_0, \lambda_{ij}(i, j = 0, 1, \ldots, N)$. Consequently, we obtain the exact solutions of (3).

3. Application to Biswas-Milovic equation

The application of the enhanced Kudryashov's method to the Biswas–Milovic equation in magneto-optic waveguide having generalized Kudryashov's nonlinear refractive index structure, will be discussed in the present section.

In order to solve the system, the following solution structure is selected.

$$q(x, t) = P_1(\xi) e^{i(\phi(x,t))}, \tag{11}$$

$$r(x, t) = P_2(\xi) e^{i(\phi(x,t))}, \tag{12}$$

where, the wave variable ξ is given by

$$\xi = k(x - vt). \tag{13}$$

Here, $P_l(\xi)(l = 1, 2)$ represents the amplitude component of the soliton solution and v is the speed of the soliton, while the phase component $\phi(x, t)$ is defined as

$$\phi(x, t) = -\kappa x + \omega t + \theta. \tag{14}$$

where κ is the frequency of the solitons, while ω represents the wave number, and θ is the phase constant. Substituting (11) and (12) into (1) and (2) and then decomposing into real and imaginary parts gives

$$f_l P_l^m P_{\tilde{l}}^{-2n} + g_l P_l^m P_{\tilde{l}}^{-n} + h_l P_l^m P_{\tilde{l}}^n + k_1 P_l^m P_{\tilde{l}}^{2n} - Q_l P_{\tilde{l}}^m + k^2(m-1)ma_l P_l^{m-2}\left(P_l'\right)^2 - \kappa m\zeta_l P_l^m + \\ k^2 m a_l P_l^{m-1} P_l'' - \kappa^2 m^2 a_l P_l^m + b_l P_l^{m-2n} + c_l P_l^{m-n} + d_l P_l^{m+n} + P_l^{m+2n}\left(e_l - \kappa m\left(\alpha_l + \mu_l\right)\right) - m\omega P_l^m = 0, \tag{15}$$

and

$$-km P_l^{m-1} P_l' \left(2\kappa m a_l + \zeta_l + \upsilon\right) - k P_l' P_l^{m+2n-1}\left(m\mu_l + \alpha_l(m+2n) + 2n\theta_l\right) = 0. \tag{16}$$

Here, $\tilde{l} = 3 - l$. Using the balancing principle leads to $P_{\tilde{l}} = \gamma P_l$, then Eq. (15) reduces to

$$-\left(a_l\kappa^2 m^2 + \kappa m\zeta_l + m\omega + \gamma^m Q_l\right) P_l^m + \left(b_l + f_l\gamma^{-2n}\right) P_l^{m-2n} + \left(c_l + g_l\gamma^{-n}\right) P_l^{m-n} + \\ \left(d_1 + h_1\gamma^n\right) P_l^{m+n} + a_l k^2(m-1)m P_l^{m-2} P_l'^2 + a_l k^2 m P_l^{m-1} P_l'' + \left(e_l + k_l\gamma^{2n} - \kappa m\alpha_l - \kappa m\mu_l\right) P_l^{m+2n} = 0. \tag{17}$$

Also, we have the following constraints

$$a_1 = a_2\gamma^m, \quad \left(e_l - \kappa m\left(\alpha_l + \mu_l\right)\right) + k_l\gamma^{2n} = \gamma^m\left(\gamma^{2n}\left(e_2 - \kappa m\left(\alpha_2 + \mu_2\right)\right) + k_2\right), \\ d_l + h_l\gamma^n = \gamma^m\left(d_2\gamma^n + h_2\right), \quad c_l + g_l\gamma^{-n} = \gamma^m\left(c_2\gamma^{-n} + g_2\right), \\ a_1\kappa^2 m^2 + Q_l\gamma^m + \zeta_1\kappa m + m\omega = \gamma^m\left(a_2\kappa^2 m^2 + \zeta_2\kappa m + m\omega\right) + Q_2, \\ b_l + f_l\gamma^{-2n} = \gamma^m\left(b_2\gamma^{-2n} + f_2\right). \tag{18}$$

From the imaginary part equation, it is possible to obtain υ as

$$\upsilon = -(2\kappa m a_l + \zeta_l), \tag{19}$$

and

$$m\mu_l + \alpha_l(m+2n) + 2n\theta_l = 0. \tag{20}$$

Using the transformation

$$P_l(\xi) = U(\xi)^{\frac{1}{n}}. \tag{21}$$

So that, Eq. (17) transforms to

$$-n^2U^2\left(a_l\kappa^2 m^2 + \zeta_l\kappa m + m\omega + \gamma^m Q_l\right) + n^2\left(b_l + f_l\gamma^{-2n}\right) + n^2U\left(c_l + g_l\gamma^{-n}\right) + \\ n^2U^3\left(d_l + h_l\gamma^n\right) + a_l k^2 mnUU'' + a_l k^2 m(m-n)U'^2 + n^2U^4\left(e_l + k_l\gamma^{2n} - \kappa m\left(\alpha_l + \mu_l\right)\right) = 0. \tag{22}$$

Eq. (22) can be rewritten as

$$n^2A_5U^4 + n^2A_4U^3 + n^2A_3U^2 + n^2A_2U + n^2A_1 + n^2k^2\left(nUU'' + (m-n)U'^2\right) = 0, \tag{23}$$

where

$$A_1 = \frac{b_l + f_l\gamma^{-2n}}{a_l m}, \quad A_2 = \frac{c_l + g_l\gamma^{-n}}{a_l m}, \quad A_3 = -\frac{a_l\kappa^2 m^2 + Q_l\gamma^m + \zeta_l\kappa m + m\omega}{a_l m}, \\ A_4 = \frac{d_l + h_l\gamma^n}{a_l m}, \quad A_5 = \frac{e_l + k_l\gamma^{2n} - \kappa m\left(\alpha_l + \mu_l\right)}{a_l m}. \tag{24}$$

Balancing UU'' with U^4 in Eq. (23) gives $N = 1$. Consequently, we reach

$$U(\xi) = \lambda_0 + \lambda_{01}R(\xi) + \lambda_{10}Q(\xi). \tag{25}$$

Substituting (25) into (23) along with (7) and (8). As a result of this substitution, we get a polynomial of $Q(\xi), R(\xi)$ and $R'(\xi)$. In this polynomial we gather all terms of same powers and equating them to be zero, we get an over-determined system of algebraic equations as follows:

$$A_5\lambda_0^4 n^2 + A_4\lambda_0^3 n^2 + A_3\lambda_0^2 n^2 + A_2\lambda_0 n^2 + A_1 n^2 = 0, \tag{26}$$

$$A_5\lambda_{01}^4 n^2 - k^2\lambda_{01}^2 m\chi - k^2\lambda_{01}^2 n\chi = 0, \tag{27}$$

$$4A_5\lambda_{10}\lambda_0^3 n^2 + 3A_4\lambda_{10}\lambda_0^2 n^2 + 2A_3\lambda_{10}\lambda_0 n^2 + A_2\lambda_{10}n^2 + k^2\lambda_{10}\lambda_0 n = 0, \tag{28}$$

$$6A_5\lambda_0^2\lambda_{10}^2 n^2 + A_3\lambda_{10}^2 n^2 + 3A_4\lambda_0\lambda_{10}^2 n^2 + \lambda_{10}^2 k^2 m - 3\eta\lambda_0\lambda_{10}k^2 n = 0, \tag{29}$$

$$A_4\lambda_{10}^3 n^2 + 4A_5\lambda_0\lambda_{10}^3 n^2 - 2\eta k^2\lambda_{10}^2 m + 2\eta^2 k^2\lambda_0\lambda_{10}n - \eta k^2\lambda_{10}^2 n = 0, \tag{30}$$

$$A_5\lambda_{10}^4n^2+\eta^2k^2\lambda_{10}^2m+\eta^2k^2\lambda_{10}^2n=0, \tag{31}$$

$$6A_5\lambda_0^2\lambda_{01}^2n^2+A_3\lambda_{01}^2n^2+3A_4\lambda_0\lambda_{01}^2n^2+k^2\lambda_{01}^2m=0, \tag{32}$$

$$A_5\lambda_{01}^2\lambda_{10}^2n^2=0, \tag{33}$$

$$3A_4\lambda_{10}\lambda_{01}^2n^2+12A_5\lambda_0\lambda_{10}\lambda_{01}^2n^2=0, \tag{34}$$

$$A_4\lambda_{01}^3n^2+4A_5\lambda_0\lambda_{01}^3n^2-2k^2\lambda_0\lambda_{01}n\chi=0, \tag{35}$$

$$4A_5\lambda_{01}^3\lambda_{10}n^2-2k^2\lambda_{01}\lambda_{10}n\chi=0, \tag{36}$$

$$4A_5\lambda_{01}\lambda_0^3n^2+3A_4\lambda_{01}\lambda_0^2n^2+2A_3\lambda_{01}\lambda_0n^2+A_2\lambda_{01}n^2+k^2\lambda_{01}\lambda_0n=0, \tag{37}$$

$$12A_5\lambda_0^2\lambda_{01}\lambda_{10}n^2+2A_3\lambda_{01}\lambda_{10}n^2+6A_4\lambda_0\lambda_{01}\lambda_{10}n^2+2\lambda_{01}\lambda_{10}k^2n=0, \tag{38}$$

$$3A_4\lambda_{01}\lambda_{10}^2n^2+12A_5\lambda_0\lambda_{01}\lambda_{10}^2n^2-3\eta\lambda_{01}\lambda_{10}k^2n=0, \tag{39}$$

$$4A_5\lambda_{01}\lambda_{10}^3n^2+2\eta^2k^2\lambda_{01}\lambda_{10}n=0, \tag{40}$$

$$2k^2\lambda_{01}\lambda_{10}n-2k^2\lambda_{01}\lambda_{10}m=0, \tag{41}$$

$$2\eta k^2\lambda_{01}\lambda_{10}m-2\eta k^2\lambda_{01}\lambda_{10}n=0. \tag{42}$$

This system of equations can be solved by the Maple or Mathematica to get the following results:

Result–1

$$\begin{aligned}
&\lambda_0=-\frac{A_4(m+n)}{2A_5(2m+n)},\quad \lambda_{01}=\pm\sqrt{\frac{\chi(m+n)\left(3A_4^2m(m+n)-2A_3A_5(2m+n)^2\right)}{2A_5^2m(2m+n)^2}},\\
&k=\pm n\sqrt{\frac{3A_4^2m(m+n)-2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}},\quad \lambda_{10}=0,\\
&A_1=-\frac{A_4^2(m-n)(m+n)^2\left(5A_4^2m(m+n)-4A_3A_5(2m+n)^2\right)}{16A_5^3m(2m+n)^4},\\
&A_2=-\frac{A_4\left(2m^2+mn-n^2\right)\left(A_4^2m(m+n)-A_3A_5(2m+n)^2\right)}{2A_5^2m(2m+n)^3}.
\end{aligned} \tag{43}$$

Consequently, we obtain the exact solutions of Eqs. (1) and (2) as follows

$$q(x,t)=\left\{-\frac{A_4(m+n)}{2A_5(2m+n)}\pm\sqrt{\frac{\chi(m+n)\left(3A_4^2m(m+n)-2A_3A_5(2m+n)^2\right)}{2A_5^2m(2m+n)^2}}\times\right.$$
$$\left.\frac{4a\exp\left[\pm n\sqrt{\frac{3A_4^2m(m+n)-2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x-vt)\right]}{4a^2\exp\left[\pm 2n\sqrt{\frac{3A_4^2m(m+n)-2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x-vt)\right]+\chi}\right\}^{\frac{1}{n}}e^{i(-\kappa x+\omega t+\theta)} \tag{44}$$

$$r(x,t)=\gamma\left\{-\frac{A_4(m+n)}{2A_5(2m+n)}\pm\sqrt{\frac{\chi(m+n)\left(3A_4^2m(m+n)-2A_3A_5(2m+n)^2\right)}{2A_5^2m(2m+n)^2}}\times\right.$$
$$\left.\frac{4a\exp\left[\pm n\sqrt{\frac{3A_4^2m(m+n)-2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x-vt)\right]}{4a^2\exp\left[\pm 2n\sqrt{\frac{3A_4^2m(m+n)-2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x-vt)\right]+\chi}\right\}^{\frac{1}{n}}e^{i(-\kappa x+\omega t+\theta)} \tag{45}$$

Setting $\chi = \pm 4a^2$ in solutions (44) and (45).

Consequently, we have bright soliton and singular soliton solutions

$$q(x,t) = \left\{ -\frac{A_4(m+n)}{2A_5(2m+n)} \pm \sqrt{\frac{2(m+n)\left(3A_4^2m(m+n) - 2A_3A_5(2m+n)^2\right)}{A_5^2m(2m+n)^2}} \times \right.$$
$$\left. \operatorname{sech}\left[n\sqrt{\frac{3A_4^2m(m+n) - 2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x - vt) \right] \right\}^{\frac{1}{n}} e^{i(-\kappa x + \omega t + \theta)}, \tag{46}$$

$$r(x,t) = \gamma \left\{ -\frac{A_4(m+n)}{2A_5(2m+n)} \pm \sqrt{\frac{2(m+n)\left(3A_4^2m(m+n) - 2A_3A_5(2m+n)^2\right)}{A_5^2m(2m+n)^2}} \times \right.$$
$$\left. \operatorname{sech}\left[n\sqrt{\frac{3A_4^2m(m+n) - 2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x - vt) \right] \right\}^{\frac{1}{n}} e^{i(-\kappa x + \omega t + \theta)}, \tag{47}$$

and

$$q(x,t) = \left\{ -\frac{A_4(m+n)}{2A_5(2m+n)} \pm \sqrt{-\frac{2(m+n)\left(3A_4^2m(m+n) - 2A_3A_5(2m+n)^2\right)}{A_5^2m(2m+n)^2}} \times \right.$$
$$\left. \operatorname{csch}\left[n\sqrt{\frac{3A_4^2m(m+n) - 2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x - vt) \right] \right\}^{\frac{1}{n}} e^{i(-\kappa x + \omega t + \theta)}, \tag{48}$$

$$r(x,t) = \gamma \left\{ -\frac{A_4(m+n)}{2A_5(2m+n)} \pm \sqrt{-\frac{2(m+n)\left(3A_4^2m(m+n) - 2A_3A_5(2m+n)^2\right)}{A_5^2m(2m+n)^2}} \times \right.$$
$$\left. \operatorname{csch}\left[n\sqrt{\frac{3A_4^2m(m+n) - 2A_3A_5(2m+n)^2}{2A_5^2m(2m+n)^2}}(x - vt) \right] \right\}^{\frac{1}{n}} e^{i(-\kappa x + \omega t + \theta)}. \tag{49}$$

Provided that

$$3A_4^2m^2(m+n) - 2A_3A_5m(2m+n)^2 > 0.$$

Result–2

$$\lambda_0 = -\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)} - \frac{\lambda_{10}}{2\eta}, \quad \lambda_{10} = \mp\frac{\eta\sqrt{-m(m+n)\left(2A_3A_5(2m+n)^2 - 3A_4^2m(m+n)\right)}}{A_5m(2m+n)}, \tag{50}$$
$$\lambda_{01} = 0, \quad k = \pm n\sqrt{\frac{2A_3A_5(2m+n)^2 - 3A_4^2m(m+n)}{A_5m(2m+n)^2}},$$
$$A_1 = \frac{\left(m^2-n^2\right)\left(A_4^2m(m+n) - A_3A_5(2m+n)^2\right)^2}{4A_5^3m^2(2m+n)^4},$$
$$A_2 = -\frac{A_4\left(2m^2+mn-n^2\right)\left(A_4^2m(m+n) - A_3A_5(2m+n)^2\right)}{2A_5^2m(2m+n)^3}.$$

Consequently, we obtain the exact solutions of Eqs. (1) and (2) as follows

$$q(x,t) = \left\{ -\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)} - \frac{\lambda_{10}}{2\eta} + \frac{\lambda_{10}}{b\exp\left[\pm n\sqrt{\frac{2A_3A_5(2m+n)^2-3A_4^2m(m+n)}{A_5m(2m+n)^2}}(x-vt)\right] + \eta} \right\}^{\frac{1}{n}} e^{i(-\kappa x+\omega t+\theta)} \tag{51}$$

$$r(x,t) = \gamma\left\{ -\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)} - \frac{\lambda_{10}}{2\eta} + \frac{\lambda_{10}}{b\exp\left[\pm n\sqrt{\frac{2A_3A_5(2m+n)^2-3A_4^2m(m+n)}{A_5m(2m+n)^2}}(x-vt)\right] + \eta} \right\}^{\frac{1}{n}} e^{i(-\kappa x+\omega t+\theta)} \tag{52}$$

Setting $\eta = \pm b$ in solutions (51) and (52).

Consequently, we have dark soliton and singular soliton solutions

$$q(x,t)=\left\{-\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)}\pm\frac{\sqrt{-m(m+n)\left(2A_3A_5(2m+n)^2-3A_4^2m(m+n)\right)}}{2A_5m(2m+n)}\times\right.$$
$$\left.\tanh\left[n\sqrt{\frac{2A_3A_5(2m+n)^2-3A_4^2m(m+n)}{4A_5m(2m+n)^2}}(x-vt)\right]\right\}^{\frac{1}{n}}e^{i(-\kappa x+\omega t+\theta)}, \tag{53}$$

$$r(x,t)=\gamma\left\{-\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)}\pm\frac{\sqrt{-m(m+n)\left(2A_3A_5(2m+n)^2-3A_4^2m(m+n)\right)}}{2A_5m(2m+n)}\times\right.$$
$$\left.\tanh\left[n\sqrt{\frac{2A_3A_5(2m+n)^2-3A_4^2m(m+n)}{4A_5m(2m+n)^2}}(x-vt)\right]\right\}^{\frac{1}{n}}e^{i(-\kappa x+\omega t+\theta)}, \tag{54}$$

and

$$q(x,t)=\left\{-\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)}\pm\frac{\sqrt{-m(m+n)\left(2A_3A_5(2m+n)^2-3A_4^2m(m+n)\right)}}{2A_5m(2m+n)}\times\right.$$
$$\left.\coth\left[n\sqrt{\frac{2A_3A_5(2m+n)^2-3A_4^2m(m+n)}{4A_5m(2m+n)^2}}(x-vt)\right]\right\}^{\frac{1}{n}}e^{i(-\kappa x+\omega t+\theta)}, \tag{55}$$

$$r(x,t)=\gamma\left\{-\frac{A_4A_5m(m+n)}{2A_5^2m(2m+n)}\pm\frac{\sqrt{-m(m+n)\left(2A_3A_5(2m+n)^2-3A_4^2m(m+n)\right)}}{2A_5m(2m+n)}\times\right.$$
$$\left.\coth\left[n\sqrt{\frac{2A_3A_5(2m+n)^2-3A_4^2m(m+n)}{4A_5m(2m+n)^2}}(x-vt)\right]\right\}^{\frac{1}{n}}e^{i(-\kappa x+\omega t+\theta)}. \tag{56}$$

Provided that

$$3A_4^2m^2(m+n)-2A_3A_5m(2m+n)^2<0.$$

4. Conclusions

The enhanced Kudryashov's method is considered in the paper to extract optical solitons for the Biswas–Milovic equation in magneto-optic waveguide coupled system having Kudryashov's law of refractive index. Successfully, we have obtained bright, dark and singular soliton solutions with constraint conditions that emerged to guarantee the existence of these solitons. This algorithm takes full advantage of the well-known Kudryashov's method and the new Kudryashov's method. Therefore, it is quite simple, comprehensive and effective to obtain optical solitons for other types of equations which will be investigated in our future work.

References

[1] A. Biswas, Solitons in magneto-optic waveguides, Appl. Math. Comput. 153 (2004) 387–393.
[2] A. Biswas, D. Milovic, Bright and dark solitons of the generalized nonlinear Schrödinger's equation, Commun. Nonlinear Sci. Numer. Simul. 15 (2010) 1473–1484.
[3] N.A. Kudryashov, Mathematical model of propagation pulse in optical fiber with power nonlinearities, Optik 212 (2020) 164750.
[4] N.A. Kudryashov, Almost general solution of the reduced higher-order nonlinear Schrödinger equation, Optik 230 (2021) 166347.
[5] N.A. Kudryashov, Solitary waves of the non-local Schrödinger equation with arbitrary refractive index, Optik 231 (2021) 166443.
[6] Q. Zhou, L. Liu, H. Zhang, C. Wei, J. Lu, H. Yu, A. Biswas, Analytical study of thirring optical solitons with parabolic law nonlinearity and spatio-temporal dispersion, Eur. Phys. J. Plus 130 (2015) 138.
[7] M. Ekici, Q. Zhou, A. Sonmezoglu, S.P. Moshokoa, M. Zaka Ullah, A. Biswas, M. Belic, Solitons in magneto-optic waveguides by extended trial function scheme, Superlattices Microstruct. 107 (2017) 197–218.
[8] H.O. Bakodah, A.A. Al Qarni, M.A. Banaja, Q. Zhou, S.P. Moshokoa, A. Biswas, Bright and dark thirring optical solitons with improved adomian decomposition method, Optik 130 (2017) 1115–1123.
[9] S. Liu, Q. Zhou, A. Biswas, W. Liu, Phase-shift controlling of three solitons in dispersion-decreasing fibers, Nonlinear Dyn. 98 (2019) 395–401.
[10] A. Biswas, 1-soliton solution of the generalized Radhakrishnan, Kundu, Lakshmanan equation, Phys. Lett. A 373 (2009) 2546–2548.

[11] A. Biswas, Saima Arshed, Optical solitons in presence of higher order dispersions and absence of self-phase modulation, Optik 174 (2018) 452–459.
[12] A. Biswas, M. Mirzazadeh, M. Eslami, Q. Zhou, A. Bhrawy, M. Belic, Optical solitons in nano-fibers with spatio-temporal dispersion by trial solution method, Optik 127 (2016) 7250–7257.
[13] X. Liu, H. Triki, Q. Zhou, W. Liu, Anjan Biswas, Analytic study on interactions between periodic solitons with controllable parameters, Nonlinear Dyn. 94 (2018) 703–709.
[14] Q. Zhou, M. Mirzazadeh, E. Zerrad, A. Biswas, M. Belic, Bright, dark and singular solitons in optical fibers with spatio-temporal dispersion and spatially dependent coefficients, J. Modern Opt. 630 (2016) 950–954.
[15] Q. Zhou, Q. Zhu, Y. Liu, H. Yu, P. Yao, A. Biswas, Thirring optical solitons in birefringent fibers with spatio-temporal dispersion and Kerr law nonlinearity, Laser Phys. 25 (2015) 015402.
[16] A. Biswas, Y. Yildirim, E. Yasar, H. Triki, A.S. Alshomrani, M.Z. Ullah, Q. Zhou, S.P. Moshokoa, M. Belic, Optical soliton perturbation with full nonlinearity for Kundu–Eckhaus equation by modified simple equation method, Optik 157 (2018) 1376–1380.
[17] A.H. Arnous, A.R. Seadawy, R.T. Alqahtani, Anjan Biswas, Optical solitons with complex Ginzburg–Landau equation by modified simple equation method, Optik 144 (2017) 475–480.
[18] A.H. Arnous, R.T. Alqahtani, M.Z. Ullah, A. Biswas, Dispersive optical solitons with DWDM technology by modified simple equation method, Optoelectron. Adv. Mater. Rapid Commun. 12 (2018) 431–435.
[19] A. Darwish, E. Abo El-Dahab, H. Ahmed, A.H. Arnous, M.S. Ahmed, A. Biswas, P. Guggilla, Y.Y. Yıldırım, F. Mallawi, M.R. Belic, Optical solitons in fiber bragg gratings via modified simple equation, Optik 203 (2020) 163886.
[20] M. Mirzazadeh, M. Eslami, A.H. Arnous, Dark optical solitons of biswas–milovic equation with dual-power law nonlinearity, Eur. Phys. J. Plus 130 (2015) 4.
[21] A.H. Arnous, M.Z. Ullah, S.P. Moshokoa, Q. Zhou, H. Triki, M. Mirzazadeh, A. Biswas, Optical solitons in nonlinear directional couplers with trial function scheme, Nonlinear Dyn. 88 (2017) 1891–1915.
[22] A. Biswas, M. Ekici, A. Sonmezoglu, A.H. Kara, Optical solitons and conservation law in birefringent fibers with Kundu–Eckhaus equation by extended trial function method, Optik 179 (2019) 471–478.
[23] A.H. Arnous, A. Biswas, M. Ekici, A.K. Alzahrani, M.R. Belic, Optical solitons and conservation laws of kudryashov's equation with improved modified extended tanh–function, Optik 225 (2021) 165406.
[24] A. Biswas, M. Ekici, A. Sonmezoglu, M.R. Belic, Highly dispersive optical solitons with quadratic–cubic law by F–expansion, Optik 182 (2019) 930–943.
[25] A. Biswas, M. Ekici, A. Sonmezoglu, M.R. Belic, Highly dispersive optical solitons with cubic–quintic–septic law by F–expansion, Optik 182 (2019) 897–906.
[26] A. Biswas, M. Ekici, A. Sonmezoglu, M.R. Belic, Highly dispersive optical solitons with non–local nonlinearity by F–expansion, Optik 183 (2019) 1140–1150.
[27] N.A. Kudryashov, One method for finding exact solutions of nonlinear differential equations, Commun. Non. Sci. Numer. Simulat. 17 (2012) 2248–2253.
[28] E.M.E. Zayed, A.H. Arnous, DNA dynamics studied using the homogeneous balance method, Chin. Phys. Lett. 29 (8) (2012) 80203–80205.
[29] N.A. Kudryashov, Method for finding highly dispersive optical solitons of nonlinear differential equations, Optik 206 (2020) 163550.
[30] E.M.E. Zayed, K.A. Gepreel, R.M.A. Shohib, M.E.M. Alngar, Solitons in magneto-optics waveguides for the nonlinear Biswas–Milovic equation with Kudryashov's law of refractive index using the unified auxiliary equation method, Optik 235 (2021) 166602.

An analysis of Harmony Search for solving Sudoku puzzles

Jui Pattanaik, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, juipattanayak@gmail.com*

Ashis Acharya, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, ashisacharya12@gmail.com*

Rajesh Tripathy, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, rajeshtripathy1@outlook.com*

Madhusmita Mohanty, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, madhusmitamohanty@gmail.com*

Keywords:
Sudoku puzzle
Harmony search
Metaheuristics
Local search

A B S T R A C T

The Harmony Search metaheuristic has been used to solve many different optimization problems. Several papers examined its effectiveness for solving Sudoku puzzles. Another paper claims that it is ineffective for solving Sudoku puzzles and further that the method itself lacks novelty compared to other evolutionary algorithms. Our paper analyzes the search process in harmony search when applied to a specific Sudoku puzzle examined in earlier research. The basic harmony search procedure is re-implemented and tested to evaluate its performance and verify its applicability to the specific example. We found that the while the criticisms of the method for this problem are valid, that the performance can be improved with a rather simple modification. First, we propose a new objective function for the search procedure. This proposed objective function facilitates the search method to find a proper solution. Second, the modified version of the harmony search, where harmony search is combined with local search is introduced and analyzed for its contribution of 'improvisation' in harmony search procedure by comparing the performance of local search and the modified search. For a specific problem, the modified version of harmony search generates a unique solution with new objective function in favorable time. Then extended experiments were performed for various Sudoku problems. We find that while the modified search procedure produces solutions more quickly, that it suffers the same issue that the original method has in that it sometimes fails to find a feasible solution.

1. Introduction

In recent years, a tremendous amount of research has been conducted related to the application of metaheuristics to combinatorial optimization problems. While some of these efforts have gained recognition and respect, others face criticism due to unpredictable performance and lack of theoretical foundations. The Harmony Search (HS) algorithm based on jazz music was proposed by Geem [9]. Since its introduction, HS has been applied to problems in areas such as scheduling optimization [6], reliability problem [27] and facility layout design [11] because of its simple structure and easiness to be applied. HS is even capable of derivative for discrete variables [7] and the result can be independent from parameter setting [10]. However, Weyland [25] raised the issue of its novelty, and also its limitations, and Weyland [26] presented criticism of its application to Sudoku, which was proposed in Geem [8]. The result provided in Geem [8] could not be verified by Weyland [26] and the question about the applicability of the HS algorithm came to the front. Our study is focused on the analysis of HS for its usability as a Sudoku solver, the introduction of a modified search method, and a comparison between the original search method and our modified one. Note that we are not addressing the arguments surrounding the novelty of Harmony search that were mentioned in Weyland [25] and then rebutted several times, for example in Kim [12] and Saka et al. [20]. We are simply demonstrating that some simple improvements can improve the performance of the method on Sudoku problems.

A Sudoku puzzle assigns a single-digit number between 1 through 9 to each location in a 9×9 matrix. A number must not be repeated in each row, column, and within each of the nine 3×3 blocks. As shown in Fig. 1, each puzzle has some cells that have already been filled with numbers. This puzzle, which is well known for its addictive nature [13], has been further popularized by the many versions which have been developed and released as mobile applications. The level of difficulty of a Sudoku puzzle depends not only on the number of pre-filled cells but also on the techniques required to find the proper values for each cell [19]. An empty Sudoku grid has 6.67×10^{21} possible combinations [5, 19], but the pre-filled cells serve as constraints and reduce the number of possible combinations.

	5		3		6			7
				8	5		2	4
	9	8	4	2		6		3
9		1			3	2		6
	3						1	
5		7	2	6		9		8
4		5		9		3	8	
	1		5	7				2
8			1		4		7	

Fig. 1. Example of Sudoku Puzzle [8].

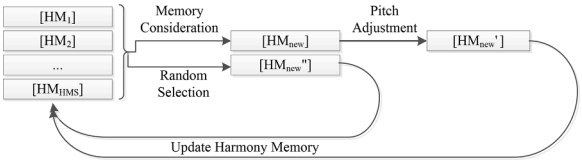


Fig. 2. Procedure of Harmony Search.

In recent years, many research efforts involving Suduku have applied evolutionary search algorithms such as genetic algorithms [14], particle swarm optimization [17], artificial bee colony algorithms [18]. A detailed discussion of meta-heuristic approaches can be found in Lewis [13] and Mishra et al. [15]. In Mishra et al. [15], more than 10 meta-heuristics were introduced and analyzed on their performance on Sudoku, but the HS was not included in this discussion. As previously mentioned, our paper analyzes the performance of HS and its applicability to Sudoku and it is not intended to support or contradict the previous research presented in Weyland [26] or Geem [8]. Heuristics that are designed specifically for Sudoku puzzles and have excellent solution times were introduced in Coelho and Laporte [3]. Thus, comparing the performance of HS to other heuristics that are specifically designed-to-solve Sudoku puzzles would not be a new contribution. Our study is mainly interested in exploring the performance of HS and in proposing a modification which will improve it for these puzzles. We propose the HS algorithm with some modification, embedding local search to solve Sudoku puzzle. Furthermore, we analyze the effect of harmony search procedure by comparing its result to the variant containing the local search.

In the following section, we explored the mathematical form of Sudoku prior to discussing the harmony search. Section 3 describes a basic harmony search used in Geem [8], and we propose a possible objective function to solve Sudoku more accurately. An extended form of HS algorithm is introduced and detailed in this section. Section 4 presents a comparison of results of previous research to this study. Additional analysis for the proposed algorithm and further applications of the proposed algorithm are presented. Finally, we draw our conclusion in Section 5.

2. Sudoku problem formulation

The Sudoku puzzle found in Fig. 1 can be modeled as a linear program [2]. Specifically, it can be modeled as a binary integer program (BIP) for general $n \times n$ puzzles. The decision variables are defined as follows:

$$x_{ij}^{k} = \begin{cases} 1, & \textit{if element } (i,j) \textit{ of the } n \times n \textit{ matrix contains the integer } k \\ 0, & \textit{otherwise} \end{cases}$$

This problem is a special case of a linear program because it only considers the constraints and can be modeled as a constraint program [1, 2]. Thus, the objective function is just set to zero as in Eq. (1) and the constraints are set to work for its satisfiability.

$$\text{Min} 0 \tag{1}$$

$$\text{s.t.} \sum_{i=1}^{9} x_{ij}^{k} = 1,\ j = 1..9,\ k = 1, \ldots\ldots, 9 \tag{2}$$

$$\sum_{j=1}^{9} x_{ij}^{k} = 1,\ i = 1..9,\ k = 1, \ldots\ldots, 9 \tag{3}$$

$$\sum_{j=3q-2}^{3q} \sum_{i=3p-2}^{3p} x_{ij}^{k} = 1,\ p = 1..3, q = 1..3,\ k = 1, \ldots\ldots, 9 \tag{4}$$

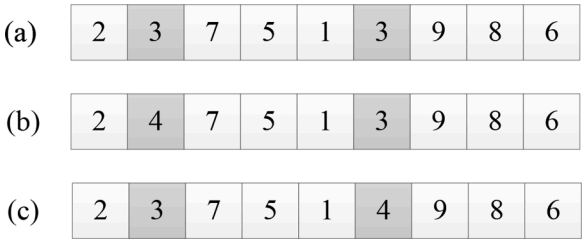


Fig. 4. re-adjustment procedure.

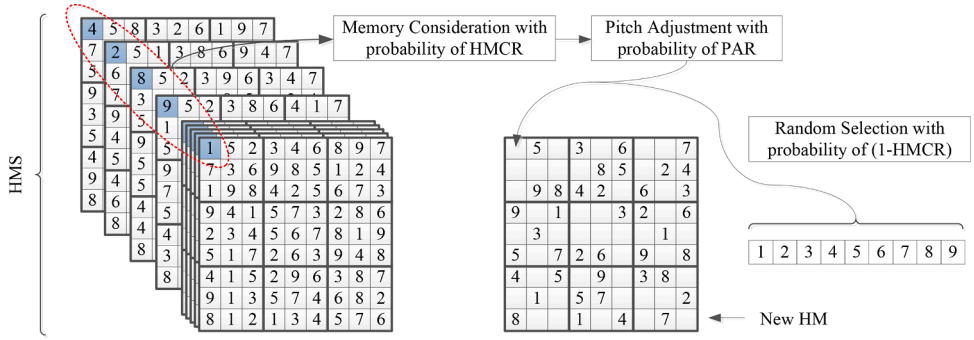


Fig. 3. HM construction process.

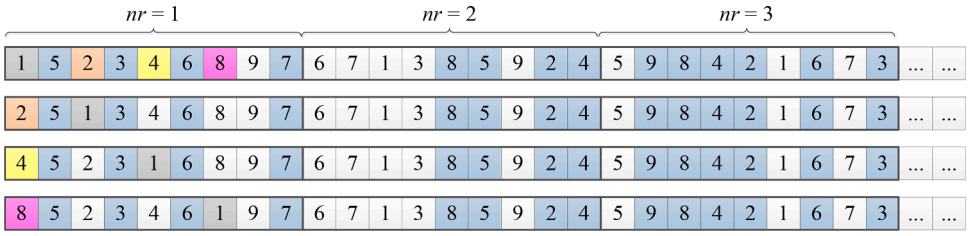
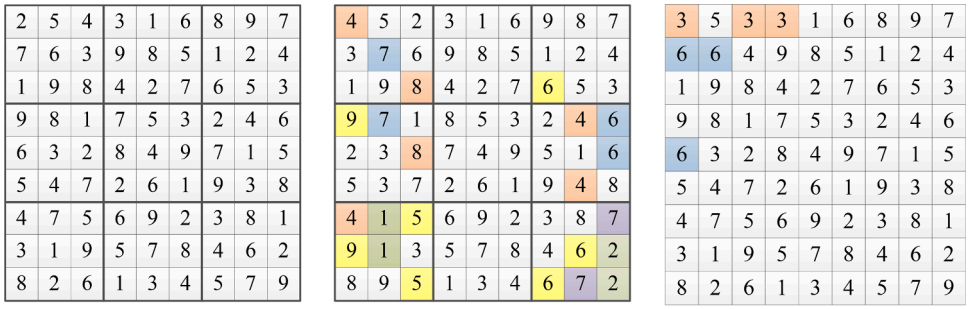


Fig. 5. Example of two-way exchange in a *2-opt* algorithm.

Table 1
the result of the experiments.

HMS	HMCR	PAR	Runs finding the unique solution in 10^4 iterations in Weyland [26] and this study	10^6 iterations in Weyland [26] and this study	Iterations to obtain the optimal solution in Geem [8]	by HS with Re-adjustment		by HS with embedded local search (HS2E)
1	0.5	0.01	0	0	66	395167	(0.05)	2
		0.1	0	0	337	655541	(0.1)	15
		0.5	0	0	422	n/a	(0)	14
	0.7	0.01	0	0	287	3978	(0.35)	1
		0.1	0	0	3413	297840	(0.15)	7
		0.5	0	0	56	n/a	(0)	3
	0.9	0.01	0	0	260	828969	(0.05)	136
		0.1	0	0	n/a	90892	(0.1)	61
		0.5	0	0	1003	88848	(0.25)	19
2	0.5	0.01	0	0	31	180210	(0.05)	12
		0.1	0	0	94	502616	(0.1)	12
		0.5	0	0	175	n/a	(0)	14
	0.7	0.01	0	0	102	15930	(0.3)	2
		0.1	0	0	77	364327	(0.1)	16
		0.5	0	0	99	n/a	(0)	11
	0.9	0.01	0	0	n/a	203425	(0.15)	11
		0.1	0	0	n/a	54147	(0.2)	19
		0.5	0	0	1325	126627	(0.15)	3
10	0.5	0.01	0	0	49	450860	(0.35)	25
		0.1	0	0	280	825597	(0.05)	7
		0.5	0	0	188	n/a	(0)	18
	0.7	0.01	0	0	56	546484	(0.05)	2
		0.1	0	0	146	33106	(0.15)	11
		0.5	0	0	259	n/a	(0)	33
	0.9	0.01	0	0	180	8199	(0.15)	4
		0.1	0	0	217	1798	(0.25)	7
		0.5	0	0	350	198292	(0.15)	10
50	0.5	0.01	0	0	147	n/a	(0)	2
		0.1	0	0	372	n/a	(0)	2
		0.5	0	0	649	n/a	(0)	7
	0.7	0.01	0	0	165	55735	(0.05)	5
		0.1	0	0	285	133542	(0.1)	23
		0.5	0	0	453	n/a	(0)	19
	0.9	0.01	0	0	87	n/a	(0)	2
		0.1	0	0	329	10781	(0.05)	26
		0.5	0	0	352	638784	(0.15)	30



(a) The unique solution

(b) An optimal solution with Eq. (12) in this study

(c) An optimal solution with Eq. (12) in Weyland (2015)

Fig. 6. The example of solutions that generate the optimal objective value of 0 in Eq. (12).

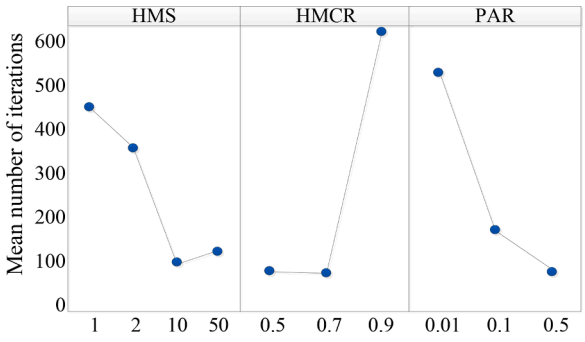


Fig. 7. Main effects plot for the number of iterations.

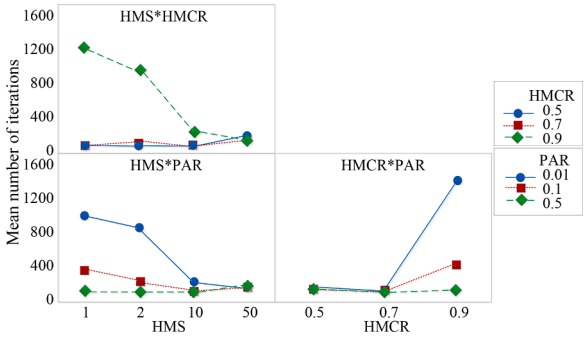


Fig. 8. Interaction effects plots for the number of iterations.

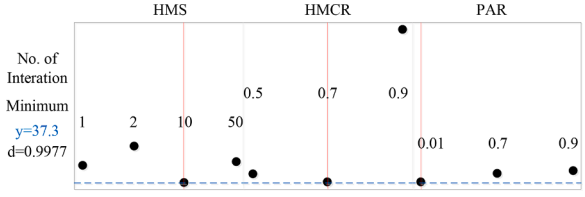


Fig. 9. Response optimization.

$$\sum_{k=1}^{9} x_{ij}^{k} = 1,\ i = 1..9,\ j = 1,\ldots..,9 \tag{5}$$

$$x_{ij}^{k} = 1,\ \forall (i,j,k) \in G \tag{6}$$

$$x_{ij}^{k} \in \{0,1\},\ \forall i,\ j \tag{7}$$

Eq. (2) through Eq. (4) indicates that the single number should be assigned to each row, column, and block, where m indicates the dimension of submatrix. Eq. (5) ensures that every cell must have a number. The given number is set to 1 at Eq. (6) where G indicates the set of cell location and the number specifically given at that cell. Eq. (7) restricts the variable uses to only binary.

This constraint program can be formulated another way as follows:

$$\text{Min} \left| \sum_{i=1}^{9} x_{ij}^{k} - 1 \right| + \left| \sum_{j=1}^{9} x_{ij}^{k} - 1 \right| + \left| \sum_{j=3q-2}^{3q} \sum_{i=3p-2}^{3p} x_{ij}^{k} - 1 \right| \tag{8}$$

$$\text{s.t. } i = 1,\ldots,9,\ j = 1,\ldots,9,\ k = 1,\ldots,9,\ p = 1,\ldots,3,\ q = 1,\ldots,3$$
$$\sum_{k=1}^{9} x_{ij}^{k} = 1,\ i = 1,\ldots,9,\ j = 1,\ldots,9 \tag{9}$$

$$x_{ij}^{k} = 1,\ \forall (i,j,k) \in G \tag{10}$$

$$x_{ij}^{k} \in \{0,1\},\ \forall i,\ j \tag{11}$$

Due to differences in objective function, the method of construction in harmony search could not be compared with that of a search procedure. However, we were able to compare the unique solution and the solving time of a specific Sudoku problem in the two algorithms. The formulation was coded and executed using CPLEX 12.6 on an Intel Core i5 CPU, 8G memory computer, which generated the optimal solution within 0.01 second for the example problem given in Geem [8]. Additional tests using extended examples (Section 4.3) were examined through the CPLEX application. The additional tests generated the optimal solutions within 0.02 s for even the hardest problem. Thus, rating the performance of each algorithm in terms of solving time was ineffective. Instead, the search procedure itself is tested for its effectiveness.

3. Harmony Search

Harmony Search's approach is inspired by the improvisation process used by jazz musicians. There are three phases of this approach: initialization, improvisation, and memory updates. The harmony memory size (HMS) is defined as the number of solution vectors called harmony memory (HM). The HMS, harmony memory consideration rate (HMCR) and pitch adjustment rate (PAR) should be determined at the initialization phase. Two parameters, HMCR and PAR, are used for constructing a solution for the next generation in improvisation phase with the use of random selection for a new HM. After generating new HM, the new group of HM is updated based on its solution quality; if a newly generated HM is better than the HM in the previous group, then the new HM is included to a new generation of HM where the worst value is removed. Fig. 2 illustrates the procedure of HM.

3.1. Basic HS Model for Sudoku

As it was mentioned, Sudoku is solved by filling in the cells with numbers 1 through 9 while abiding to the 'single number' rule [2] in the matrix. Thus, the objective function of BIP is not the focus of the problem—the constraints are. However, one of the characteristics of

Table 3
Number of iterations needed to solve the Sudoku puzzle using HS2E.

	Easy (36)	Medium (29)	Hard (23)	SD1 (24)	SD2 (23)	SD3(22)
Min (run time)	24 (0.6s)	196 (6.8s)	9765 (354s)	16626 (891s)	26450 (1523s)	20708 (1098s)
Median	882	22645	256780	46365	191397	216998.5
Probability of Success	100%	77%	63%	87%	67%	80%

Table 2
The comparison of HS2E and *2-opt* algorithm.

	Number of iterations				Time to Solve (in s)			
	Min	Median	Mean	Max	Min	Median	Mean	Max
HS2E	1	39.5	54.2	384	0.07	0.75	0.91	4.95
2-opt	20	379.5	584.1	3442	0.48	9.33	14.22	83.53

improvement type heuristics such as HS is that they have a guided approach to a given problem. In other words, the algorithm decides whether to adapt the new solution or to keep the previous solution in each iteration depending on the resulting value. Thus, setting the objective function would be very important to develop a proper solution procedure. Mishra et al. [15] introduced various objective functions used to solve Sudoku using evolutionary metaheuristics. In Geem [8], the objective function is based on the sum of each row, column, and block as follows:

$$\text{Minimize } Z = \sum_{i=1}^{9}\left|\sum_{j=1}^{9} x_{ij} - 45\right| + \sum_{j=1}^{9}\left|\sum_{i=1}^{9} x_{ij} - 45\right| + \sum_{r=1}^{9}\left|\sum_{(s,t)\in B_r} x_{st} - 45\right| \tag{12}$$

where x_{ij} = cell at row i and column j, and B_r = set of coordinates for block r. The characteristic of the Sudoku matrix requires that sum of each row, column, and block should be 45. Therefore, the given objective function should be zero when the allocation of single number satisfies its constraints. It has been mentioned that the given objective function does not guarantee that the numbers 1 through 9 are shown exactly once in a row, column, and block [8]. However, there is no specific method introduced to avoid this violation in Geem [8], and it is expected that the iteration itself would drive the process to generate the intended solutions. The violation of constraints with zero objective function value (OFV) has been reported in Weyland [26]. Therefore, we propose a modified objective function as follows.

$$\begin{aligned}\text{Minimize } Z = &\sum_{i=1}^{9}\left(\sum_{j=1}^{9} x_{ij} - 45\right)^2 + \sum_{j=1}^{9}\left(\sum_{i=1}^{9} x_{ij} - 45\right)^2 \\ &+ \sum_{r=1}^{9}\left(\sum_{(s,t)\in B_r} x_{st} - 45\right)^2 + \sum_{i=1}^{9}\left(\sum_{j=1}^{9} S_{ij} - 60\right)^2 + \sum_{j=1}^{9}\left(\sum_{i=1}^{9} S_{ij} - 60\right)^2 \\ &+ \sum_{r=1}^{9}\left(\sum_{(s,t)\in B_r} S_{st} - 60\right)^2\end{aligned} \tag{13}$$

where $S_{ij} = (x_{ij} - \overline{x})^2$ and $\overline{x} = 5$, which are the deviation and mean value of 1 through 9, respectively. The summation of the deviation for each row is expected to be 60 if the number 1 through 9 are evenly distributed. The use of a square instead of an absolute sign places increased emphasis on the non-negativity condition. Furthermore, the term including S_{ij}, a deviation, ensures that numbers 1 through 9 appears only once in a row, column, and block.

For HS to solve Sudoku in this study, the process follows exactly as indicated in Weyland [26] and Geem [8]. The whole process is summarized in Fig. 2. The solution was generated randomly in accordance with the HMS. New solutions are continuously generated until the maximum number of iterations is met. Each new solution is modified using one of the three methods: memory consideration, pitch adjustment, and random selection. In each step, memory consideration is used to choose a value from one of the existing HM with the probability of HMCR and the random selection was used to choose a value from 1 through 9 in uniformly random with the probability of 1-HMCR. This means that every number in each cell in Sudoku matrix is assigned by either memory consideration or random selection. An additional method with the probability of PAR, called pitch adjustment, is applied to the one already assigned by memory consideration. According to Geem [8], the pitch adjustment adds or subtracts 1 from the originally assigned number with a chance of 1/2 except when the number 1 and 9 has a lower and upper limit. The HM constriction process is presented in Fig. 3.

3.2. Modification: re-adjustment

The HS in Geem [8] does not explicitly consider avoiding duplication of each number in each row, column, and block. However, many heuristic approaches adopt certain methods to avoid such duplication. In Pacurib et al. [18], a penalty function is used to avoid duplication, and many other researchers have used a blocking mechanism in methods [14, 16, 22, 23]. The objective function we present in Eq. (9) is much tighter and therefore *encourages* solutions without any duplication in each row, column, and bock, but it does not *require* these solutions. Therefore, at the final stage of each iteration, a number that appeared more than once in each row is replaced with another number that is not present in the current row.

Fig. 4 shows the process of re-adjustment in the final stage of each iteration. Since number 3 appears twice in (a), the first 3 is replaced with a 4 as shown in (b). 4 is an acceptable substitute because it was not present in the row before substitution. If the second cell in (a) is a given number that cannot be replaced, then the second 3 is replaced with a 4. In the case in which there are more than three identical numbers in a row, the substitution process is the same as explained in Fig. 4 except there are additional numbers for substitution. The process finds the position for replacement from left to right. Once the process finds the specific position for the number to replace, a randomly generated number which has not been shown in a row are chosen for the substation. Once done, the next position requiring a substitute is found. It proceeds until all the numbers, 1 to 9, are shown a row. This process increases the chances of finding a solution.

3.3. Modification: embedding local search

The performance of local search is in general not as effective as well developed metaheuristics, but it can be embedded into a heuristic to improve the search process of the original method. The *2-opt* algorithm is an improvement technique used for a variety of combinatorial problems, and it can be easily adapted due to its simplicity and easy implementation. The algorithm was first introduced by Croes [4] to solve a traveling salesmen problem, and it has been adapted to many other combinatorial problems since then. To apply the procedure to Sudoku, a single two-way exchange is performed in each row. This is because the final HM of each iteration has a feasible arrangement of numbers in each row after the re-adjustment procedure, and the feasibility of each column is not considered at this stage. However, the number exchange in each row forces the algorithm to search for the best objective function possible by placing each cell's 'single number' in each column as well as each block. Shown below is a *2-opt* algorithm adapted for this study. Fig. 5 presents an example of number exchange in the beginning of an algorithm.

Step 1. Let C be the initial solution provided by the HM with re-adjustment and z its OFV. Set $C^* = c$, $Z^* = z$, $i = 1, j = i+1$, $nr = 1$, and $i, j \notin G$, where G is set of fixed cell as in Eq. (6)
Step 2. Exchange the numbers in cell i and j in the solution C. If the result of this exchange, C', improve the OFV, $z' < z^*$, then set $z^* = z'$ and $C^* = C'$. If $j < g_{nr}$, where g_{nr} is max number in G^c in each row, then $j = j+1$; otherwise $i = i+1$ and $j = i+1$. If $i < g_{nr}$ then repeat step 2; otherwise set $nr = nr+1$. If $nr \leq 9$, then repeat step 2; otherwise go to step 3.
Step 3. If $C \neq C^*$, set $C = C^*, z = z^*, i = 1, j = i+1, nr = 1$ and go to step 2. Otherwise, stop the process and return C^* as the best solution.

4. Experimental results

As mentioned in Weyland [26], the issue of the objective function and the result in Geem [8] demonstrates the need to repeat the experiments before pursuing further analysis for the extendibility of HS. Table 1 shows the result of the problem instance used throughout this study. To compare the specific value, we use the result of HS run in Weyland [26] and Geem [8]. This study verifies that the classic (or original) HS could not find a unique solution for Sudoku puzzle within

10^6 iterations, the number cited in Weyland [26]. Hence, it is logical that the original HS could not find a unique solution within 10^4 times of iterations, the maximum number of iterations cited in Geem [8]. Each set of parameters were tested 20 times for a total of 720 runs. The HS with an objective function in Eq. (12) generates an optimal solution, which is not a feasible solution as shown in Fig. 6. The number place-ments that violate the single number rule are highlighted.

The proposed objective function, Eq. (13), was used for the rest of the experiments and the classic HS did not generate a unique solution in the given number of iterations. However, as mentioned before, it was reported that HS has been successfully applied to other type of optimization problems and that there are many hybridized HS methods that are capable of generating favorable solutions for a certain problem. Therefore, we modified the procedure as explained in 3.2; the re-adjustment process was added to make each row feasible at the final stage of HM in each iteration. The result of the modification is shown on the second and third column in the right side of Table 1. The value in parenthesis indicates the probability of finding the solution. As shown in the table, the modification seems to be unsuccessful. Many iterations are necessary to increase the probability of finding the optimal solution. However, the re-adjustment modification would be embedded in the beginning of the local search *2-opt* because this two-way exchange algorithm requires a feasible configuration to improve its performance. After adapting the *2-opt* to HS, the result is dramatic as shown in the last column of Table 1. The number indicates the minimum number of iterations to find the optimal solution among the 20 trials in each parameter set.

Through this experiment, it was verified that the HS with embedded *2-opt* — which we refer to as HS2E — generated favorable results for the specific Sudoku problem. However, it was uncertain whether the *2-opt* algorithm was capable of replicating HS2E's performance by itself. This uncertainty questioned the contribution of HS in HS2E. To verify the contribution of 'HS' in HS2E, a randomly generated initial solution was given to a *2-opt* procedure instead of HM in HS2E, and the random solution was compared to the improvised solution in HS. For this evaluation, the parameters were set for HS2E and the test for parameter selection was done.

4.1. Parameter Selection

To select the parameter set, a general full factorial analysis with three factors, multiple levels, and 20 replications generated to provide for Table 1 was conducted. Since there is no dependable OFV for each factor, the number of iterations to find zero OFV serves as the response for this analysis. It was concluded that each of the factors and their interaction—HMS, HMCR, and PAR—are influential to determine the number of iterations necessary.

The main effects graphed in Fig. 7 indicates that the parameter set would be the most effective with (HMS, HMCR, PAR) = (10, 0.7, 0.5). However, the interaction effect does not provide a clear distinction of each parameter as shown in Fig. 8, with the exception of a few sets that must not be combined as a parameter: HMS = 1, HMCR = 0.9, and PAR=0.01. Finally, the tool named 'Response Optimizer', provided in Minitab® and used as a parameter tuning tool, was used to select the best combination of parameters to minimize the number of iterations needed to find the optimal solution; the tool provided the parameter set (HMS, HMCR, PAR) =(10, 0.7, 0.01) as shown in Fig. 9.

The number of iterations is expected to be 37.3 with use of 'the parameter set'. d = 0.9977 indicates that the setting is well fit to overall response (d = 1 represents ideal case).

4.2. HS2E vs. 2-opt algorithm

With a given parameter set, additional 50 runs were tested. The procedure for HS2E and *2-opt* algorithm were implemented in C++, and executed on an Intel Core i5 class computer with 8GB of memory. Table 2 summarizes the result of these two algorithms.

As shown in Table 2, the proposed HS2E is superior to *2-opt* algorithm because HS2E requires fewer iterations and less solution time than the *2-opt* algorithm. We noticed that the median value of the number of iterations for HS2E is very close to the one generated by a response optimizer (y = 37.3). Overall performance indicates that the application of HS as a Sudoku solver is effective for this specific problem when combined with the local search, *2-opt*. We expect that statistical analysis would provide the same result.

4.3. Additional experiments

As mentioned in Section 1, the level of difficulty of a Sudoku puzzle depends on the numbers that are given to the grid in the theoretical count. The example problem in Weyland [26] and Geem [8] has 40 given numbers. The Sudoku puzzles with more than 30 given numbers fall into the 'easy' categories based on the example used in Mantere and Koljonen [14], Pacurib et al. [18], Sato and Inoue [22], and Wang et al. [24]. Thus, we performed additional experiments to test the method's applicability in Sudoku puzzles of varying difficulty. Three instances used in Pacurib et al. [18] were tested for HS2E: easy, medium, and hard. Additional instances in Sato et al. [21] were used for extremely difficult cases: SD1, SD2, and SD3. Table 3 shows the result of the experiment for these six instances. The numbers in parenthesis indicate the number of givens in Sudoku grid. The experiment was repeated 30 times for each problem within 10^6 iterations.

The average iterations are not significant in this experiment because some trials did not find an optimal solution within 10^6 iterations and because we do not have information regarding how far the iterations might proceed. However, the median can be calculated since the majority of trials reached the solution. Through these experiments, it can be inferred that when fewer numbers are given to a Sudoku grid, HS needs more iterations to find a unique solution. This is expected since the given number in Sudoku grid is related to the number of combinations the Sudoku puzzle can have. However, the probability of success is not directly related to the number of givens in Sudoku. The method generated a solution for SD2, named 'Al Escargot' known as the one of the hardest Sudoku puzzles, with 67% of probability of success, which is less than that of SD3. A medium difficulty problem with 29 givens had lower probability of success than that of SD1 with only 24 givens. The data structure inside of a Sudoku grid could affect the chance of finding a solution. We observed that it takes around 8 h to have 10^6 iterations for the medium and hard problems and around 12 h for the very difficult SD1 through SD3.

5. Conclusion and discussion

In this paper, we analyzed the effectiveness of basic harmony search for solving Sudoku puzzles. The effectiveness of harmony search in Sudoku has been a subject of debate. In this research, the contribution of the harmony search was evaluated and the effect of harmony memory construction to local search algorithm was shown by comparing the results of harmony search with that of the local search with a randomly generated initial solution. The improvised procedure in harmony search facilitated the local search to find the optimal solution with a proposed objective function, which was much more specific to the optimality condition.

Even though HS2E was proposed to explore the HS effect as a search procedure to a specific given problem, the HS2E was applied to solve general Sudoku instances other than the one given in the previous research. The experimental result showed that HS2E generated the optimal solution satisfactorily and it is capable of solving extremely difficult problems. However, its probability of success for medium to extremely hard problem is not 100%. Thus, identifying other search methods to embed which will increase the probability of success on

References

[1] K Apt, Principles of Constraint Programming, Cambridge University Press, Cambridge, 2003.
[2] AC Bartlett, TP Chartier, AN Langville, TD Rankin, Integer programming model for the Sudoku problem, J Online Math. Appl 8 (1) (2008).
[3] LC Coelho, G Laporte, A comparison of several enumerative algorithms for Sudoku, J. Oper. Res. Soc. 65 (2014) 1602–1610, https://doi.org/10.1057/jors.2013.114.
[4] GA Croes, A method for solving traveling-salesman problems, Oper. Res. 6 (1958) 791–812, https://doi.org/10.1287/opre.6.6.791.
[5] B Felgenhouer, F Jarvis, Mathematics of Sudoku I, Math Spectr 39 (2006) 15–22.
[6] a ZW Geem, Optimal scheduling of multiple dam system using Harmony Search Algorithm, in: F Sandoval, A Prieto, J Cabestany, M Graña (Eds.), Computational and Ambient Intelligence, Springer, Berlin, Heidelberg, 2007, pp. 316–323.
[7] ZW Geem, Novel derivative of harmony search algorithm for discrete design variables, Appl. Math. Comput. 199 (2008) 223–230, https://doi.org/10.1016/j.amc.2007.09.049.
[8] b ZW Geem, Harmony search algorithm for solving Sudoku, in: B Apolloni, RJ Howlett, L Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Springer, Berlin, Heidelberg, 2007, pp. 371–378.
[9] ZW Geem, JH Kim, G V Loganathan, A new heuristic optimization algorithm: Harmony Search, Simulation 76 (2001) 60–68, https://doi.org/10.1177/003754970107600201.
[10] ZW Geem, KB Sim, Parameter-setting-free harmony search algorithm, Appl. Math. Comput. 217 (2010) 3881–3889, https://doi.org/10.1016/j.amc.2010.09.049.
[11] S Kang, J Chae, Harmony search for the layout design of an unequal area facility, Expert Syst. Appl. 79 (2017) 269–281, https://doi.org/10.1016/j.eswa.2017.02.047.
[12] JH Kim, Harmony Search algorithm: a unique music-inspired algorithm, Procedia Eng 154 (2016) 1401–1405, https://doi.org/10.1016/j.proeng.2016.07.510.
[13] R Lewis, Metaheuristics can solve sudoku puzzles, J. Heuristics. 13 (2007) 387–401, https://doi.org/10.1007/s10732-007-9012-8.
[14] T Mantere, J Koljonen, Solving, rating and generating sudoku puzzles with GA, in: Proceedings of the 2007 IEEE Congress on Evolutionary Computation, CEC 2007, 2007, pp. 1382–1389, https://doi.org/10.1109/CEC.2007.4424632.
[15] DB Mishra, R Mishra, KN Das, AA Acharya, et al., Solving Sudoku puzzles using evolutionary techniques—a systematic survey, in: M Pant, K Ray, TK Sharma, et al. (Eds.), Soft Computing: Theories and Applications, Springer, Singapore, 2018, pp. 791–802.
[16] TK Moon, JH Gunther, JJ Kupin, Sinkhorn solves Sudoku, IEEE Trans. Inf. Theory 55 (2009) 1741–1746, https://doi.org/10.1109/TIT.2009.2013004.
[17] A Moraglio, J Togelius, Geometric particle swarm optimization for the sudoku puzzle, in: Proceedings of the Ninth annual conference on Genetic and evolutionary computation - GECCO '07, ACM Press, New York, 2007, p. 118.
[18] JA Pacurib, GMM Seno, JPT Yusiong, Solving Sudoku puzzles using improved Artificial Bee Colony Algorithm, in: Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC), IEEE, 2009, pp. 885–888.
[19] R Pelánek, Difficulty rating of Sudoku puzzles by a computational model, in: Proceedings of the FLAIRS Conference, 2011.
[20] MP Saka, O Hasançebi, ZW Geem, Metaheuristics in structural optimization and discussions on harmony search algorithm, Swarm Evol. Comput. 28 (2016) 88–97, https://doi.org/10.1016/j.swevo.2016.01.005.
[21] Y Sato, N Hasegawa, M Sato, GPU acceleration for Sudoku solution with genetic operations, in: Proceedings of the 2011 IEEE Congress of Evolutionary Computation, CEC 2011, 2011, pp. 296–303.
[22] Y Sato, H Inoue, Solving Sudoku with genetic operations that preserve building blocks, in: Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, CIG2010, 2010, pp. 23–29.
[23] H Simonis, Sudoku as a constraint problem, in: B Hnich, P Prosser, B Smith (Eds.), Proceedings of the Fourth International Workshop on Modelling and Reformulating Constraint Satisfaction Problems, 2005, pp. 13–27.
[24] Z Wang, T Yasuda, K Ohkura, An evolutionary approach to sudoku puzzles with filtered mutations, in: Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC), 2015, pp. 1732–1737.
[25] D Weyland, A Rigorous Analysis of the Harmony Search Algorithm: How the research community can be misled by a "Novel" methodology. Modeling, Analysis, and Applications in Metaheuristic Computing, IGI Global, 2010, pp. 72–83.
[26] D Weyland, A critical analysis of the harmony search algorithm-How not to solve sudoku, Oper. Res. Perspect. 2 (2015) 97–105, https://doi.org/10.1016/j.orp.2015.04.001.
[27] D Zou, L Gao, J Wu, et al., A novel global harmony search algorithm for reliability problems, Comput. Ind. Eng. 58 (2010) 307–316, https://doi.org/10.1016/j.cie.2009.11.003.

Analysis of French phonetic idiosyncrasies for accent recognition

Rakhi Jha, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, rakhijha91@yahoo.co.in*

Subrat Dash, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, subratdash43@gmail.com*

Laxmi, *Department of Computer Science Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, laxmi3349@gmail.com*

Rudra Prasad Nanda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, rudraprasad858@gmail.com*

Keywords:
Accent recognition
French accent classification

A B S T R A C T

Speech recognition systems have made tremendous progress since the last few decades. They have developed significantly in identifying the speech of the speaker. However, there is a scope of improvement in speech recognition systems in identifying the nuances and accents of a speaker. It is known that any specific natural language may possess at least one accent. Despite the identical word phonemic composition, if it is pronounced in different accents, we will have sound waves, which are different from each other. Differences in pronunciation, in accent and intonation of speech in general, create one of the most common problems of speech recognition. If there are a lot of accents in language we should create the acoustic model for each separately. We carry out a systematic analysis of the problem in the accurate classification of accents. We use traditional machine learning techniques and convolutional neural networks, and show that the classical techniques are not sufficiently efficient to solve this problem. Using spectrograms of speech signals, we propose a multi-class classification framework for accent recognition. In this paper, we focus our attention on the French accent. We also identify its limitation by understanding the impact of French idiosyncrasies on its spectrograms.

1. Introduction

Accent recognition is one of the most important topics in automatic speaker and speaker-independent speech recognition (SI-ASR) systems in recent years. The growth of voice-controlled technologies has becoming part of our daily life, nevertheless variability in speech makes these spoken language technologies relatively difficult. One of the profound variability in a speech signal is the accent. Different models could be developed to handle SI-ASR by accurately classifying the various accent types [1]. Such a successful accent recognition module can be integrated into a natural language processor, leading to its wide ranging impact in finance [2], medical science [3], and sustainable environment [4].

Dialect/accent refers to the different ways of pronouncing/speaking a language within a community. Some illustrative examples could be American English versus British English speakers or the Spanish speakers in Spain versus Caribbean. During the past few years, there have been significant attempt to automatically recognize the dialect or accent of a speaker given his or her speech utterance. Recognition of dialects or accents of speakers prior to automatic speech recognition (ASR) helps in improving performance of the ASR systems by adapting the ASR acoustic and/or language models appropriately. Moreover, in applications such as smart assistants as the ones used in smartphones, by recognizing the accent of the caller and then connecting the caller to agent with similar dialect or accent will produce more user-friendly environment for the users of the application.

Most of the existing techniques do not possess good accuracy in identifying the various accents. One of the reasons we are having trouble to have a good accuracy in the accent recognition problem is the lack of knowledge we have of English syllabic structure. In order to approximate English phonology, we have to understand the native language similarities of articulation, intonation, and rhythm. In the past, the research has focused on phone inventories and sequences, acoustic realizations, and intonation patterns. Therefore, it is important to study the English syllable structure. The main problem behind word recognition is the understanding of the syllable. It usually consists of an obligatory vowel with optional initial and final consonants. One familiar way of subdividing a syllable is into *onset* and *rhyme*. All syllables in all languages phonetically at least consist of onset and rhyme. However, these categories alone do not indicate where the syllable is placed within the word. In order to capture foreign accents in English, we want to highlight those constituents of the syllable that are most likely to prove difficult for speakers of languages in which they are not contained [5].

In this paper, we focus on the specifications of the French language. We are interested in identifying the idiosyncrasies [6] of French people that lead a model into predicting the wrong accent.

1.1. Related work

Berkling et al. [5] discussed the tonal and non-tonal languages and their treatment in speech recognition systems. In Kardava et al. [7], they have developed an approach to solve the above mentioned problems and create more effective, improved speech recognition system of Georgian language and of languages, that are similar to Georgian language. Katarina et al. proposed [8], an automatic method of detection of the degree of foreign accent and the results are compared with accent labeling carried out by an expert phonetician. In [9], they give a new approach for modeling allophones in a speech recognition system based on hidden Markov models.

In [10], they studied mutual influences between native and non-native vowel production during learning, *i.e.*, before and after short-term visual articulatory feedback training with non-native sounds. To obtain a speaker's pronunciation characteristics, [11] gave a method based on an idea from bionics, which uses spectrogram statistics to achieve a characteristic spectrogram to give a stable representation of the speaker's pronunciation from a linear superposition of short-time spectrograms. Hossari et al. in [12] used a two-stage cascading model using Facebook's fastTex implementation [13] to learn the word embeddings. Davies et al. presented advanced computer vision methods, emphasizing machine and deep learning techniques that have emerged during the past 5–10 years [14]. The book provides clear explanations of principles and algorithms supported with applications. In [15], Farris present the Gini index and several measures of integrity.

1.2. Contributions of the paper

The main contributions of this paper[1] can be summarized as follows:

- Highlighting the problem of the limit in the context of the study of accent recognition. In this paper, we will show there exists a "natural" limit of the accuracy when it comes to accent classification. The main aim of this work will be to address that limit and give a solution to that problem.
- Highlighting French idiosyncrasies restricting the accuracy values of deep learning models. In this paper, we focused our work on the French speakers. We decided to study the language habits of French speakers that could explain the decrease in precision. Indeed, the English language is an Indo-European Germanic language while the French is a Latin language, which means that their structure is very different. Thus, we will find strongly similar words between the two languages, but the way of pronouncing them will often vary a lot. Thus, the study of these Latin habits is particularly interesting in the context of our work: understanding which aspects of the French language reduce the effectiveness of our models will allow us to better recognize a French accent later on.
- Highlighting the incidence of these idiosyncrasies in the spectrograms, and therefore the models in question. Once we have isolated more clearly the responsible French idiosyncrasies, we determine their real impact on the models used (CNN in our case) by the precise study of spectrograms of vocal samples used. In this case, we will compare different spectrograms for the same sentence and determine the differences between a "French" and "English" spectrogram, for a specific idiosyncrasy.

Table 1
Highlighting of French main mispronunciations of the English language.

Usual English pronunciations and French pronunciation
short A, as in fat `French Accent` : pronounced "ah" as in father
long A followed by a consonant, as in gate `French Accent` : pronounced like the short e in get
ER at the end of a word, as in water `French Accent` : pronounced air
short I, as in sip `French Accent` : pronounced "ee" as in seep
long I, as in kite `French Accent` : elongated and almost turned into two syllables: [ka it]
short O, as in cot `French Accent` : pronounced either "uh" as in cut, or "oh" as in coat
U in words like full `French Accent` : pronounced "oo" as in fool

The rest of the paper is structured as follows. Section 2 discusses the data and the methods we used in our preliminary study (dataset and neural networks) and Section 3 discusses results we obtained with these methods. In Sections 4 and 5, we analyzed the French speakers idiosyncrasies and their consequences on spectrograms. Finally, Section 6 concludes the work and discusses our future works.

2. A primer on French speakers idiosyncrasies

In this section, we provide a primer to the readers on the various types of speech idiosyncrasies exhibited by French speakers.

2.1. French-infused vowels

Nearly every English vowel is affected by the French accent [10]. French has no diphthongs, so vowels are always shorter than their English counterparts. The long A, O, and U sounds in English, as in say, so, and Sue, are pronounced by French speakers like their similar but un-diphthonged French equivalents, as in the French words sais, seau, and sou. For example, English speakers pronounce say as [seI], with a diphthong made up of a long "a" sound followed by a sort of "y" sound. But French speakers will say [se] - no diphthong, no "y" sound. English vowel sounds which do not have close French equivalents are systematically replaced by other sounds, as it is showed in Table 1.

2.2. Dropped vowels, syllabification, and word stress

French people pronounce all schwas (unstressed vowels). Native English speakers tend toward "r'mind'r", but French speakers say "ree-ma-een-dair". They will pronounce amazes "ah-may-zez", with the final e fully stressed, unlike native speakers who will gloss over it: "amaz's". And the French often emphasize the -ed at the end of a verb, even if that means adding a syllable: amazed becomes "ah-may-zed".

Short words that native English speakers tend to skim over or swallow will always be carefully pronounced by French speakers. The latter will say "peanoot boo-tair and jelly", whereas native English speakers opt for pean't butt'r 'n' jelly.

Because French has no word stress (all syllables are pronounced with the same emphasis), French speakers have a hard time with stressed syllables in English, and will usually pronounce everything at the same stress, like actually, which becomes "ahk chew ah lee". Or they might stress the last syllable — particularly in words with more than two: computer is often said "com-pu-TAIR".

2.3. French-accented consonants

H is always silent in French, so the French will pronounce happy as "appy". Once in a while, they might make a particular effort, usually resulting in an overly forceful H sound — even with words like hour and honest, in which the H is silent in English. J is likely to be pronounced "zh" like the G in massage. R will be pronounced either as in French or as a tricky sound somewhere between W and L. Interestingly, if a word starting with a vowel has an R in the middle, some French speakers will mistakenly add an (overly forceful) English H in front of it. For example, arm might be pronounced "hahrm".

TH's pronunciation will vary, depending on how it is supposed to be pronounced in English:

- voiced TH [ð] is pronounced Z or DZ: "this" becomes "zees" or "dzees"
- unvoiced TH is pronounced S or T: "thin" turns into "seen" or "teen"

Letters that should be silent at the beginning and end of words (psychology, lamb) are often pronounced.

3. Accent recognition system

3.1. Features for detecting accents

Spectrograms are pictorial representation of sound we can use for speech recognition [11]. The x-axis represents time in seconds while the y-axis represents frequency in Hertz. Different colors represent the different magnitude of frequency at a particular time. We can think of the spectrogram as an image. Fig. 1 represents a sample speech single and its corresponding spectogram. Once the audio file is converted to an image, the problem reduces to an image classification task. Based on the number of images, algorithms like Support Vector Machines (SVM), *etc.* are used to classify sound, validate the speaker.

3.2. Our proposed framework for detecting accents

We used different Machine Learning and Deep Learning models, and the first one is a two convolutional layers neural network with 5 different accents as shown in Fig. 2. This neural network is a 2-layer Convolutional Neural Network: one with 32 filters and a ReLu activation function, and another one with 64 filters and a ReLu activation function.

We will focus on this 2-layer CNN for the rest of our work.

4. Results and discussion

4.1. Dataset

Everyone who speaks a language, speaks it with an accent. A particular accent essentially reflects a person's linguistic background. When people listen to someone speak with a different accent from their own, they notice the difference, and they may even make certain biased social judgments about the speaker. In this paper, we used the Speech Accent Archive [16]. It has been established to uniformly exhibit a large set of speech accents from a variety of language backgrounds. The distribution of speech signals across the five languages is represented in Fig. 3. Native and non-native speakers of English all read the same English paragraph and are carefully recorded.

This dataset allows us to compare the demographic and linguistic backgrounds of the speakers in order to determine which variables are key predictors of each accent. The speech accent archive demonstrates that accents are systematic rather than merely mistaken speech. It contains 2140 speech samples, each from a different talker reading the same reading passage. Talkers come from 177 countries and have 214 different native languages. Each talker is speaking in English. The samples were collected by many individuals under the supervision of Steven H. Weinberger, the most up-to-date version of the archive is hosted by George Mason University and can be found here: https://www.kaggle.com/rtatman/speech-accent-archive. [16]

Table 2
Average accent classification accuracy across the different languages using various benchmarking models.

Comparison of SVM and CNNs				
Model	Overall ACC	F1 Macro	F1 Micro	Hamming Loss
SVM	0.3518	0.33458	0.33458	0.38043
2-layer CNN	0.70652	0.405	0.70652	0.29348
4-layer CNN	0.6529	0.52	0.73913	0.26087

4.2. Accent recognition metric

In order to provide an objective evaluation of the accent recognition task, we compute the overall accuracy, F1-macro, F1-micro and hamming loss [17]. These metrics are defined as:

$$ACC = \frac{tp + tn}{(tp + fp) + (tn + fn)}$$

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

$$F1_{micro} = 2\frac{Micro - precision * Micro - recall}{\text{Micro-precision} + \text{Micro-recall}}$$

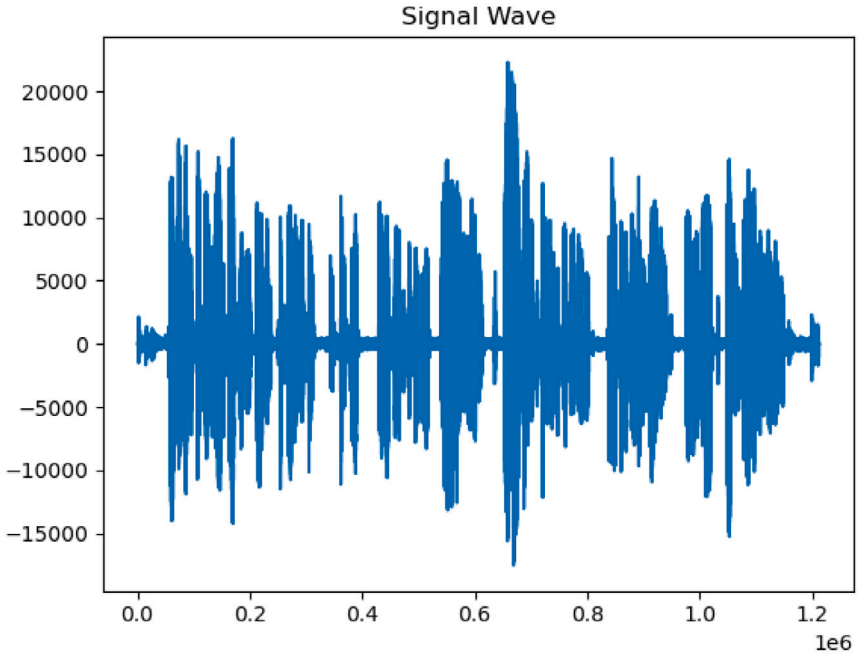
$$HL = \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} Y_{i,l} \oplus X_{i,l}$$

In the overall accuracy formula, tp, tn, fp, fn stand respectively for true positive, true negative, false positive and false negative. In the Hamming loss formula, $\oplus$ denotes exlusive-or, $X_{i,l}$ ($Y_{i,l}$) stands for boolean that the ith datum (ith prediction) contains the lth label

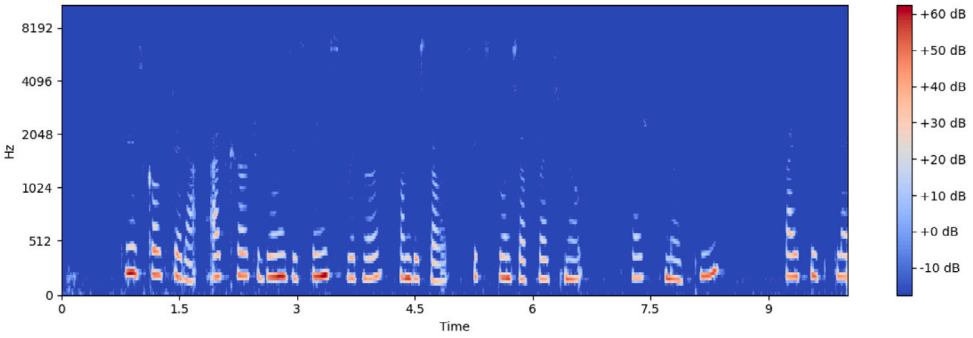
Table 2 demonstrates the evaluation metric obtained via SVM technique and two variants of CNN model.

With regular machine learning methods as SVM, we obtained low accuracy of 0.35. As expected, the impact of Deep Learning methods [14] is quite clear here. We observe from Table 2, that the Convolutional Neural Networks achieves an accuracy of 0.65. However, we observe that we do not obtain an optimal score if we use too many layers in our model. Depending upon how large our dataset is, the CNN architecture is implemented. Adding layers unnecessarily to any CNN will increase our number of parameters only for the smaller dataset. It is true for some reasons that on adding more hidden layers, it will give a better accuracy. That is true for larger datasets, as more layers with less stride factor will extract more features for the input data. In CNN, how we play with the architecture is completely dependent on what our requirement is and how our data is. Increasing unnecessary parameters will only overfit your network, and that is the reason why our CNN with 2 layers has better results than with 4.

A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average will aggregate the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if we suspect there might be class imbalance issue (*i.e.* we may have many more examples of one class, as compared to other classes). Table 2 explains this scenario clearly. We observe that neural networks show better F1-score values in the context of multi-class classification. In such situation, Hamming Loss is a good measure of model performance. The lower the Hamming loss, the better is the model performance. In our case, Hamming loss ranges from 0.26 till 0.39, which is considered as good results, especially in the context of 5-class multi-class classification problem.



(a) Signal



(b) Spectrogram

Fig. 1. Signal and spectrogram of a french accent sample.

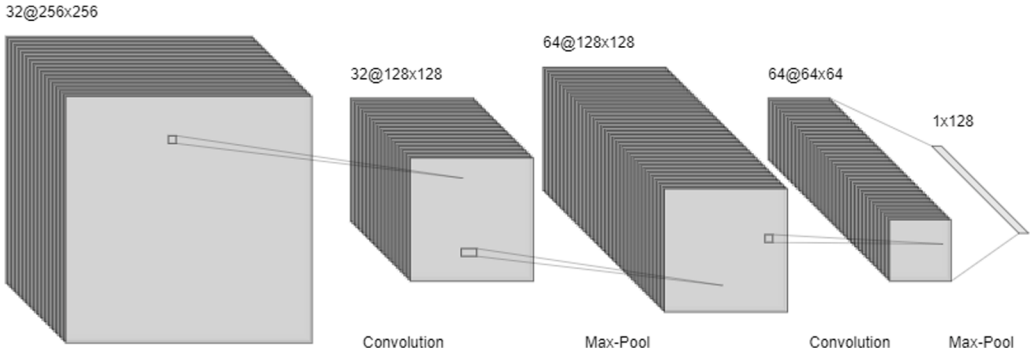


Fig. 2. CNN with 2 layers with ReLu activation function.

Table 3
Multi-class classification metric values using the SVM model.

SVM				
Classes	ACC	AGF	AUC	GI
English	0.42391	0.21774	0.36781	−0.26437
Arabic	0.71739	0.0	0.5	0.0
French	0.34783	0.51315	0.49171	−0.01658
German	0.92391	0.0	0.5	0.0
Hindi	0.95652	0.0	0.5	−0.01124

Table 4
Multi-class classification metric values using our proposed 2-layer CNN model.

2-layer CNN				
Classes	ACC	AGF	AUC	GI
English	1.0	1.0	1.0	1.0
Arabic	0.95	0.71	0.74	0.48
French	0.85	0.84	0.84	0.69
German	0.84	0.80	0.80	0.61
Hindi	0.87	0.32	0.53	0.06

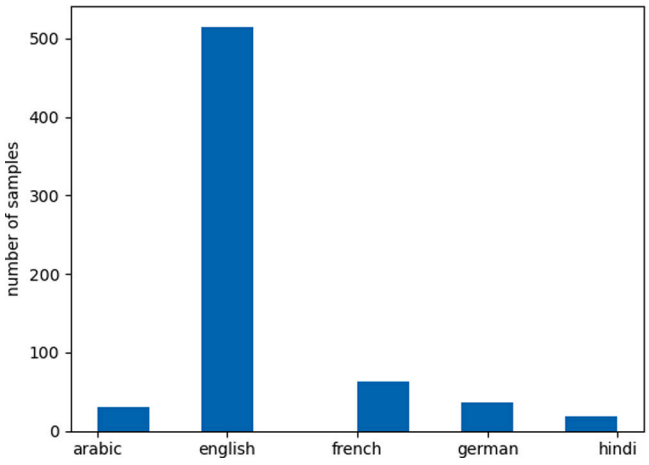


Fig. 3. The distribution of the samples across the five languages in the dataset.

4.3. Multi-class accent recognition metric

In this case of multi-class classification, we are considering ACC, AGF, AUC and GI.

$$ACC = \frac{tp + tn}{tp + tn + tp + tn}$$

$$AGF = (1 + \beta^2)\frac{precision * recall}{(\beta^2 * precision) + recall}$$

$$AUC = \frac{recall + sensibility}{2}$$

$$GI = 1 - \sum_{j=1}^{n} p_j^2 = 1$$

We obtained these results in the confusion matrices with the 2-layer CNN and the SVM method:

Table 3 indicates that the results for Arabic, Hindi and German accents are better. This can easily be explained by the size of the data sets corresponding to each accent. This is a result that shows fairly well the limit of *classical* machine learning algorithms. This limit in the evaluation scores for classical machine learning models are also observed in the broad areas of network security [18] and computer vision [19]. In this specific application of accent recognition, we observe that an increase in the number of vocal samples do not lead to an increased accuracy values. This difference is due to the lack of capacity of the SVM which has difficulty processing information as complex as images.

Table 4 indicates that the results are much more harmonized between the different accents. We still do not have a perfect match between the size of the dataset and the performance of the model, but the disparities between accents disappear.

We can observe from Tables 3 and 4 that the *classical* machine learning methods are quite ineffective and that the deep learning methods stand out clearly in accent recognition; that is why we will use the 2-layer CNNs as a reference for the rest of the paper. In most case, the SVM method is not powerful enough for us to have a good accuracy. That can be explained with the results we obtained on the Gini Index [15]. The values obtained by the index are quite low (negative values are considered quite low positive values), which means that in the case of SVM, the spectrograms are similar in nature. Such SVM methods are not selective enough to clearly determine the accent (which is also shown by the AGF values). However, the SVM method is not totally to be excluded: in the context of the Hindi accent or the German accent, the SVM turns out to be more effective than all the deep learning methods used.

The total computing time is 1 min and 23 s when our proposed model is executed on Google Colab using GPU.

5. Impact of idiosyncrasies on speech spectrograms

We will now study the idiosyncrasies of the French language and how it impacts the corresponding spectograms of the speech signals.

The spectrogram is a representation allowing to observe the whole of the decomposition spectral voice and speech on the same graphic representation. This tool is precise, informative and reliable to analyze the characteristics of sound production. In a first-cut analysis, we associate the spectrogram with the temporal pace, the power profile and segmentation. More extensively, there are a significant number of indicators, metrics and tools. This includes the fundamental frequency and its derivatives, the alteration of voice and speech, and more generally the assessment of intelligibility. It is its ability to measure vocal alteration that will interest us here. We will focus on primarily two pieces of information given by the spectrogram: amplitude and frequency in our study.

5.0.1. The un-diphthonged "y"

Firstly, we will analyze differences on the spectrograms for the word "Wednesday", where the French speaker is not supposed to use the "y" sound, like it was explained in French-infused Vowels. Here are the spectrograms of an English speaker and a French speaker of the sentence "and we will go meet her Wednesday at the train station" in Fig. 4 and Fig. 5.

We can see, as expected, that at the end of the word (1.3-1.4 for English and 1.05-1.1 for French), the "y" is almost not even pronounced by the French speaker, while the English speaker pronounced it clearly. Indeed, the frequencies used are relatively similar on the whole of the audio sample, but certain syllables are *pressed* with a much higher frequency by a French speaker. Consequently, the corresponding amplitude will be low in magnitude. This explains a clear difference between the perception of a word between a French speaker and an English speaker: the non-native will tend to pronounce English less loudly, but will support certain syllables much more than an English speaker.

5.0.2. Voiced TH [ð] is pronounced Z or DZ

French people tend to say "zees" instead of "these". That is what we can see in the sentence "Please call Stella, ask her to bring these things from the store.".

It is quite complicated to delimit the word "these" in this sentence because it is quite quick, so we will delimit "bring these", as the word "bring" does not represent a major problem for French speakers.

Here, we see that French speakers tend to diminish the importance of the word "bring" but accentuate the word "these", whereas English speakers seem to pronounce the sequence "bring these" at the same

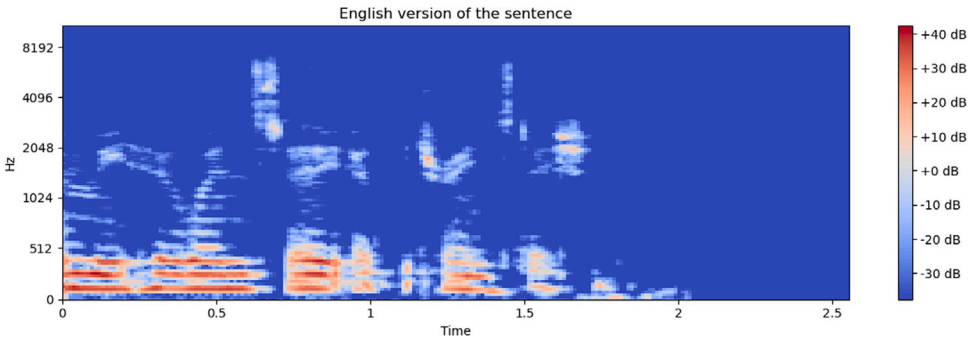


Fig. 4. "Wednesday" in English version: 0.8s-1.4s.

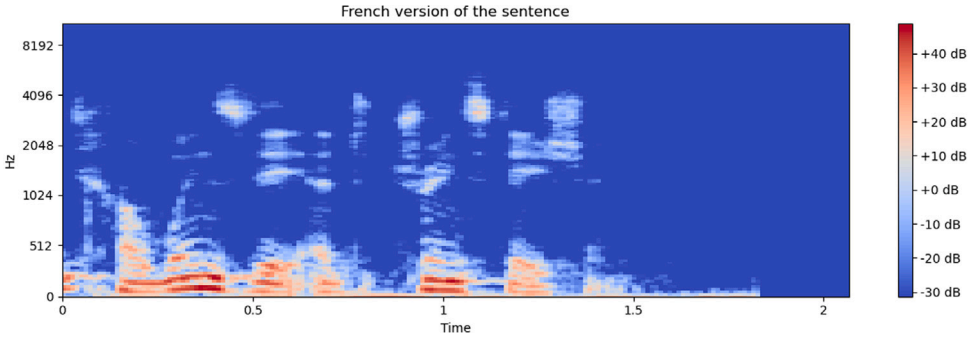


Fig. 5. "Wednesday" in French version: 0.7s-1.10s.

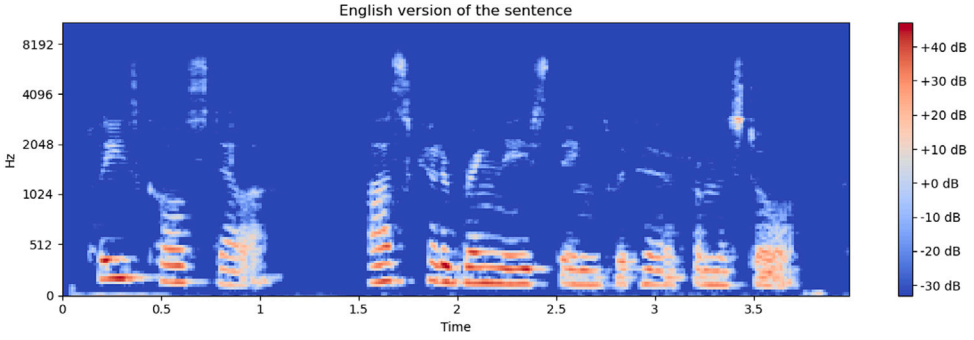


Fig. 6. "Bring these" in English version: 2.5s-3s.

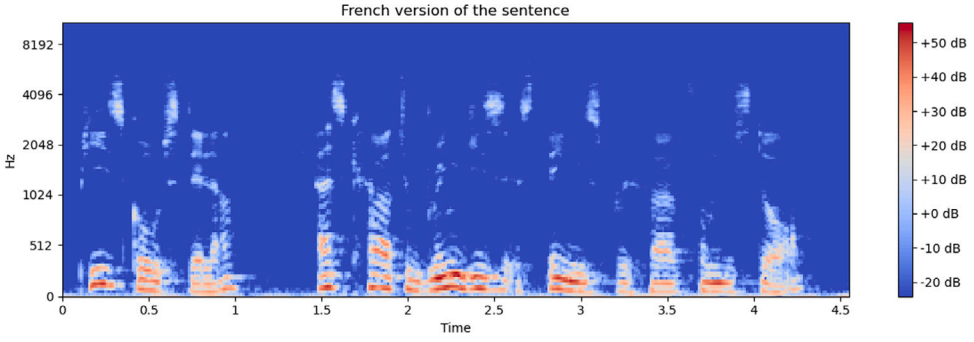


Fig. 7. "Bring these" in French version: 2.6s-3.2s.

frequency (see Fig. 6 and Fig. 7). We remark that is why, for French speakers, the "th" sounds like "z". Indeed, the closest sound to "th" is "z" in the French language, so it is only natural for us to use it. Nevertheless, we believe the reason why they accentuate it (because we could just use the sound "z" more discreetly) is because of the role of words like "these", "the", "this"... They are articles, and in the French language, they tend to accentuate the most important parts of the sentence, which made this French speaker diminish "bring", and accentuate "these".

Thus, French speakers idiosyncrasies have a direct impact on audio samples spectrograms. Then, we can easily understand why these idiosyncrasies have a direct impact on the results of deep learning models:

the first reason why we use spectrograms in order to develop Speech Recognition Systems is to turn an audio classification problem into an image classification problem. Then, if the idiosyncrasies of a specific language have that much effect on spectrograms, that means that the different languages have different spectrograms and this should help the deep learning models to get a better classification between English and French.

6. Conclusions and future work

In this paper, we have concluded that the classical deep learning models are not powerful enough to accurately predict the accent of an user. Therefore, we decided to study the differences between tonal and non-tonal languages, in order to clearly identify the obstacles that prevent us from achieving better results in accent recognition. To fulfill this purpose, we decided to devote our analysis on the French accent, which is a non-tonal language. In this paper, we studied the idiosyncrasies of French speakers: the characteristics of the spoken French language that have a direct impact on the pronunciation of English words by French speakers. In addition, we determined the consequences these idiosyncrasies have on spectrograms, and consequently on the accuracy of deep learning models. In the future, we would like to work further on the subject of French idiosyncrasies, by building a model which determines if an idiosyncrasy is present in an audio sample or not. This would allow us to more easily determine the presence of a French accent in an audio sample. Such accurate recognition of accents in a speech signal will lead to better automatic speech recognition systems.

References

[1] O. Omidvar, J. Dayhoff, Neural Networks and Pattern Recognition, Oxford University Press, 1996, pp. 23–26.
[2] M.S. Islam, E. Hossain, Foreign exchange currency rate prediction using a GRU-LSTM hybrid network, Soft Comput. Lett. (2020) 100009.
[3] H.Y. Raji-Lawal, A.T. Akinwale, O. Folorunsho, A.O. Mustapha, Decision support system for dementia patients using intuitionistic fuzzy similarity measure, Soft Comput. Lett. 2 (2020) 100005.
[4] J. Wu, F. Orlandi, T. Alskaif, D. O'Sullivan, S. Dev, Ontology modeling for decentralized household energy systems, in: 2021 International Conference on Smart Energy Systems and Technologies, SEST, IEEE, 2021.
[5] K. Berkling, J. Vonwiller, C. Cleirigh, Scope, syllable core and periphery evaluation: Automatic syllabification and application to foreign accent identification, Speech Commun. 35 (2001) 125–138.
[6] S. Awang, M. Maros, N. Ibrahim, Language idiosyncrasies in second language learners' use of communication strategies, Asian Soc. Sci. 11 (2015) 55–70.
[7] I. Kardava, J. Antidze, N. Gulua, Solving the problem of the accents for speech recognition systems, Int. J. Signal Process. Syst. 4 (2016) 235–238.
[8] B. Katarina, D. Jouvet, Automatic detection of foreign accent for automatic speech recognition, in: Proc. 16th Int. Con. Phon. Sc., 2007, pp. 2185–2188.
[9] D. Jouvet, K. Bartkova, J. Monné, On the modelization of allophones in an HMM based speech recognition system, in: Second European Conference on Speech Communication and Technology, 1991.
[10] N. Kartushina, A. Hervais-Adelman, U.H. Frauenfelder, N. Golestani, Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training, J. Phonetics 57 (2016) 21–39.
[11] J. Yanjie, X. Chen, J. Yu, L. Wang, Y. Xu, S. Liu, Y. Wang, Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network, Complex Intell. Syst. (2020).
[12] M. Hossari, S. Dev, J.D. Kelleher, TEST: A terminology extraction system for technology related terms, in: Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, 2019, pp. 78–81.
[13] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, 2016, arXiv preprint arXiv:1607.01759.
[14] E.R. Davies, O. Camps, M. Turk, Advanced Methods and Deep Learning in Computer Vision, Academic Press, 2021, pp. 441–452.
[15] F.A. Farris, The gini index and measures of inequality, Amer. Math. Monthly 117 (2010) 851–864.
[16] S. Weinberger, Speech accent archive. George Mason University. This dataset is distributed under a ccby-nc-sa 2.0 license, 2013.
[17] S. Dev, H. Javidnia, M. Hossari, M. Nicholson, K. McCabe, A. Nautiyal, C. Conran, J. Tang, W. Xu, F. Pitié, Identifying candidate spaces for advert implantation, 2019 IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT, IEEE, 2019, pp. 503–507.
[18] M.S. Elsayed, N.-A. Le-Khac, S. Dev, A.D. Jurcut, Machine-learning techniques for detecting attacks in SDN, in: 2019 IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT, IEEE, 2019.
[19] M. Jain, C. Meegan, S. Dev, Using GANs to augment data for cloud image segmentation task, in: 2021 IEEE International Geoscience and Remote Sensing Symposium, IGARSS, IEEE, 2021.

A fuzzy optimization model for methane gas production from municipal solid waste

Rashmita Panigrahi, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, rashmitapanigrahi116@gmail.com*

Niladri Bhusan Biswal, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, niladribiswal26@gmail.com*

Manoj Mohanta, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, manoj.mohanta62@outlook.com*

Priya Chandan Satpathy, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, priyachandan.satpathy57@outlook.com*

Keywords:
Municipal solid waste (MSW)
Optimization
Multi-period
Greenhouse gas (GHG) emission
Transportation
Waste
Waste management

A B S T R A C T

The availability of non-renewable fossil fuels in Jordan continues to decrease, which increases reliance on energy sources, such as, methane gas produced from municipal solid waste (MSW). Furthermore, during the COVID-19 pandemic, solid wastes were significantly increased, especially in lockdown periods and this increase requires an immediate response to this global emergency by improving MSW management system. Unfortunately, little previous research efforts have been directed to propose optimization models that optimize concurrently economic and environmental aspects with the utilization of the available resources from transportation trucks of different types and capacities. This research, therefore, develops an optimization model for efficient MSW management system to increase the percentage of waste transported from multiple depots to anaerobic digestion plants (ADP) or recycling centers. The objective function of the optimization model is two-fold; maximizing quantities of transported waste and minimizing both transportation costs and greenhouse gas (GHG) emissions generated from different types of transport trucks over a six-day period. A case study was presented, where the optimization results showed that on average 1236.36 mega Watt-hour (MWh) of energy potential at a minimal average processing cost of 165.22 $/ton could be generated from transported 3540 tons of waste over six days. Such energy can be utilized to promote sustainability and develop an eco-city powered by renewable energy. In conclusion, the proposed model is found efficient in enhancing the performance of the existing MSW and results in significant reductions in environmental impacts and transportation costs and maximizing trucks and facilities utilizations.

1. Introduction

For low-income countries, Municipal Solid Waste (MSW) management is a critical issue, which impacts the environment, socio-economic, health, aesthetics, and infrastructure, due to the generated volume of wastes, treatment, and disposal methods. Typically, the MSW system deals with wastes from its source of generation to final disposal, including all the operations and transformation of this waste [1]. Typically, the MSW has several sources, i.e., residential, commercial, and municipal services [2], which makes the management of MSW a persistent challenge for many developing countries [3].

Practically, MSW landfills cause very harmful effects on the environment and hence many treatment methods, including recycling, incineration, and mechanical biological treatment, can be implemented to reduce negative environmental impacts [4]. The decomposition of organic wastes via anaerobic processes produced methane gas. Although, the upgrading process of emitted gases to methane gas is complex and expensive, the generated methane gas can be used as energy to promote sustainability.

Over the years, the net amount of MSW generated and its impact on the environment in Jordan have increased significantly [5]. Studies showed [6] that around 70–90% of waste was collected, but only about 5–10% of solid waste has been recycled. A major problem is that this system does not include separation of the recyclable and non-recyclable solid waste. Ideally, waste separation requires strong individual's commitment, long-term information, and educational campaigns [6], however weak initiatives have been directed towards these aspects in Jordan. Further, the current MSW management lacks critical dedicated facilities, such as, recycling centers, and aerobic digestion plants, which are necessary to increase the ability of controlling and utilizing dumped quantities of generated solid waste. As a result, large amounts of waste are accumulated in landfills, which alerts decision makers for the necessity of establishing of recycling centers, and aerobic digestion plant and developing cost-effective and environmentally friendly

transportation system. Moreover, the existing waste transportation system between facilities suffers poor planning of transportation trucks, which may lead to negative impacts on environment due to excessive number of trips between facilities and incurs high transportation costs. In order to improve the current MSW management system and enhance its efficiency, this research analyzes the effectiveness of the current MSW and then develops an optimization model to maximize methane gas collection in MSW management system through maximizing trans-ported waste and minimizing both transportation costs and greenhouse gas emissions while considering truck types and capacities. The remaining of this paper including the introduction is outlined as follows. Section 2 presents literature review. Section 3 develops the optimization model. Section 4 illustrates the optimization model and discusses research results. Section 5 summarizes research conclusions. The results of this research can provide great assistance to decision makers in Jordanian municipalities on how to establish and develop a sustainable MSW management system.

2. Literature review

Recently, the MSW management has received significant research attention. For example, Badran and El-Haggar [1] proposed a mixed integer programming model for a municipal solid waste management system in Port Said, Egypt. It included the use of the concept of collection stations, which have not yet been used in Egypt. The results showed that the best model would include 27 collection stations of 15-ton daily capacity and 2 collection stations of 10-ton daily capacity. Any transfer of waste between the collection station and the landfill should not occur. Mora et al. [7] presented a mixed integer linear model to reduce the economic and environmental impacts by minimizing different costs at 'kerbside' waste collection system in a municipality in Italy. A heuristic procedure was applied to obtain some admissible solutions of the real problem. Five alternative configurations of kerbside system, diverging in number of sub-areas, synchrony of vehicles and directionality of the arcs, were compared in an economic point of view. Finally, Life-Cycle Assessment was used as a tool to compare the overall potential environmental impacts of the alternative of kerbside collection systems and to compare the kerbside system with the traditional bring one. Põldnurk [8] assessed the environmental and economic feasibility of source sorting paper and bio-waste in rural municipalities, improvement of administrative efficiency, and economic cost-effectiveness resulting in reorganization of waste management administration, and optimization options of the municipal waste collection logistics through inter-municipal waste collection districts. The results showed that in rural areas central collection of sources sorted bio-waste is not economically and environmentally feasible, however the central collection of source-sorted paper waste may be considered environmentally beneficial if applied through inter-municipal cooperation. Zhao and Zhu [9] developed a multi-depot vehicle-routing model with the minimization of total cost and total risk through two-commodity flow formulation, and simultaneous of planning tours and vehicle acquisitions for the explosive waste collection and designing the return-trips between collection centers and recycling centers. A case study in Nanchuan of South-west China and related test instances were presented to elucidate their developed approach. Habibi et al. [3] proposed a multi-objective robust optimization model for a MSW management system and addressed the economic, environmental, and social perspectives simultaneously by minimizing the total cost, the greenhouse gas emission, and the resulting visual pollution. Their model was validated using real data for long-term planning of Tehran's MSW management system by examining five candidate sites for the construction of new facilities. Trochu et al. [10] addressed the reverse logistics network design problem under environmental policies targeting recycled wood materials from the construction, renovation, and demolition (CRD) industry. The main objective was to determine the location and the capacities of the sorting facilities to ensure compliance with the new regulation and prevent the wood from being massively landfilled using a mixed-integer linear programming model (MILP) to minimize the total cost of the wood recycling process collected from CRD sites. The proposed MILP model was applied for a case study in the CRD industry within the province of Quebec, Canada. Tsai et al. [11] applied exploratory factor analysis to test the validity and reliability of MSW attributes of cities in Vietnam under uncertainty. Fuzzy set theory was used to translate the linguistic references into the qualitative attributes of MSW management. The decision-making trial and evaluation laboratory was used to address the interrelationships among the attributes. The causal interrelationships among 14 attributes were identified. The results showed that technical integration and social acceptability were the aspects that drive MSW management, while treatment innovations, safety and health, economic benefits, and technology functionality and appropriateness were determined to be the linkage criteria. Finally, the distinctions between cities were presented. Tsai et al. [12] presented a systematic data-driven bibliometric analysis on MSW management as a foundation in a circular economy and applied the entropy weight method to convert the frequencies to weights and performed regional comparisons based on a database. A bibliographic coupling analysis was conducted and revealed that Africa and North America have less studies than other regions. Xiao et al. [13] proposed system dynamic model, which simulates the entire process of MSW production, sorting, collection, and final treatment and then analyzed policy impacts on MSW management from a dynamic and complex perspective in Shanghai. Seven scenarios were set to simulate the impacts of these policies. Results showed that economic policy has the largest impact on future MSW management, where MSW generation in 2035 will decline by 3.25 million tons if Gross Domestic Production growth rate decrease by 1%. Istrate et al. [14] fulfilled the review on published life-cycle assessment studies on MSW management systems with the aim of identifying waste-to-energy solutions and their impact on the system's environmental performance. Discrepancies were observed with respect to the environmental consequences of both the diversion of organic waste from incineration to anaerobic digestion and the diversion of waste from incineration to mechanical-biological treatment plants. Deus et al. [15] developed an aggregate indicator for environmental impact assessment of MSW management in the small municipalities of the state of Sao Paulo (Brazil). Additionally, the study aimed at creating a classification of the municipalities considered to identify the best management practices. The results showed that the average waste generation was 223.89 kg, the average carbon dioxide equivalent (CO2e) emissions was 0.166 tons (inhabitant−1 year−1) and the average amount of energy savings was 51.37 kWh. Tong et al. [16] employed a system thinking approach to analyze the crucial roles of the informal sector in SWM system in Vietnam. The analysis was built on the field survey including elements and key driving forces of the systems with 36 scrap dealers, 127 scrap buyers, and 760 households and in-depth interviews with experts in the Mekong Delta region, Vietnam. Results stated that informal systems should be integrated into the SWM process. Batur et al. [17] formulated a mixed integer linear programming model for long-term planning of municipal solid waste management system taking into consideration different process, capacity, and location possibilities that may occur in complex waste management processes at the same time. The results that the developed model provides significant convenience for the multi-objective optimization of financial-environmental-social costs and the solution of some uncertainty problems of decision-making tools, such as, life cycle assessment. Iyamu et al. [18] reviewed the common themes limiting MSW management sustainability in the BRIC (Brazil, Russia, India and China) countries, as well as the historical transition of MSW management to a sustainable level in some high-income countries such as United States, Japan, Denmark, and Australia. They focused on the interaction of MSW management with technology systems, socio-economic factors, related environmental issues, influence on policy and decision making. The key MSWM findings was used to develop a thematic framework, underpinned by the different interacting factors of

policy; environmental; socio-economic; and technology. Pinha and Sagawa [19] presented a system dynamics model for MSW management which involved resources, destinations of waste and cost structure of service/system. As a case study, the context of a Brazilian city of 230, 000 inhabitants was modelled and scenarios for 10 years were proposed. The scenario that presented better results with feasible investments prescribes an increase from 8.5% to 15% in the public collection of dry waste together with a productivity improvement of the sorting process. The simulations showed that the revenues from the recyclables do not cover the expenditures of the service provider and allowed pointing out scenarios that make the provider less dependent of governmental sub-sidy. Paul and Bussemaker [20] developed a web-based decision support system that can be used in planning and management of MSW for assessing the suitability of waste valorization in a particular location, such as, waste types, waste quantities and related waste contractors in England. Waste market opportunities and circular economy partners were also identified through the web application and these results were presented in context of waste-derived supply chain decisions. Hajar et al. [21] examined the development of the MSW management sector in Jordan from sustainability standpoint and developed potential scenarios to attain Jordan Vision 2025 target and gradually place this sector on a green growth path. The Sustainability Window analysis tool was used to assess the sustainability of the studied sector over the 2010–2015 period. Three scenarios were proposed and compared using the Sustainability Window tool: Mechanical biological treatment-anaerobic digestion, mechanical biological treatment-composting, and incineration. It was concluded from the Sustainability Window analysis that the 2010–2015 Jordanian municipal solid waste sector growth did not fulfill all sustainability criteria. Pinupolu and raja Kommineni [22] suggested a method of MSW management through public-private partnership (PPP) for Vijayawada city, which faced the problem of disposal and handling of municipal solid waste. Installation cost, land required for the proposed solid waste treatment and population were assessed by the geometrical progression method for the anticipated year 2051. Results indicated that the total quantity of evaluated solid waste created in the year 2051 is 2788 tons/day. Sarbassov et al. [23] performed compositional analysis of the municipal solid waste produced at the Astana International Airport and evaluated different waste management scenarios in terms of greenhouse gas emissions. Recyclable and combustible fractions were found to be the major fractions (over 50%) of the total municipal solid waste generated in the Astana International Airport. Four base greenhouse gas emissions scenarios were proposed and discussed. Viau et al. [24] aimed to critically evaluate the modelling of substitution in life cycle cost of recovered material from MSW management systems. They performed a systematic analysis of 51 life cycle assessment studies on MSW management systems published in the peer-reviewed literature and found that 22% of the substitution ratios are only implicitly expressed. Finally, guidance for the documentation of substitution ratios, with the aim of reaching more credible and robust analyses were developed. Kulkarni and Anantharama [25] presented a global backdrop of MSW management during COVID-19 outbreak and examined various aspects of MSW management. The data and information were collected from several scientific research papers from different disciplines, publications from governments and multilateral agencies and media reports. They presented challenges and opportunities in the aftermath of the ongoing pandemic and recommended alternatives approaches for MSW treatment and disposal and outlines the future scope of work to achieve sustainable waste management during and aftermath of the pandemics. Lately during the COVID-19 crisis, the transition from fossil fuel energy sources to green energy sources is urgent and crucial issue for globe to address the emergency pandemics and secure sustainable economies. Thus, new studies for generating energy were considered from different green sources such as Mostafaeipour et al. [26] who studied the feasibility of a new power generation system from wind for urban applications. Also, Rezaei et al. [27] evaluated the production of hydrogen by establishing hybrid wind and solar power plants. With the same manner, Wang et al. [28] considered the solar energy by identifying optimal sites for constructing the solar photovoltaic panels. Wang et al. [29] offered an assessment approach for cleaner energy sources using data envelopment analysis and fuzzy model.

Recently in the MSW management system, little research efforts were directed to develop optimization models with multiple objective functions that integrate the concurrent economic and environmental aspects with the utilization of the available resources from transportation trucks. In addition, most of the proposed MSW management system ignored truck types with different capacity and GHG emission for each type. Therefore, this paper proposes optimization model with multiple objective functions, including, minimizing total transportation, minimizing GHG emissions and unfilled trucks' capacities, maximizing total transported quantities, and maximizing satisfaction levels on utilization of collection stations and recycling centers, while considering various truck types and capacities for transporting collected and processed waste quantities.

3. Optimization model development

The key elements of solid waste chain are shown in Fig. 1, which includes I depots; i= [1, …, I], J collection stations; j= [1, …, J], R recycling plants; r= [1, …, R], L landfills, l= [1, …, L], and M anaerobic digestion plants; m= [1, …, M]. In any selected area, the waste bins will be divided into clusters which are assigned to depots. The trucks will collect wastes from each cluster's bins and accumulate wastes in the assigned depot of this cluster [30]. Then, waste will be transported from depot i to collection station j for processing. Because waste separation at source is not applicable in many areas, the waste separation is performed in collection stations to sort the recyclable and non-recyclable wastes. After processing, recyclable waste is transported from collection station j to recycling center r, whereas non-recyclable waste is transported to the L landfill. Finally, the organic recyclable waste is converted to a special compost amendment, which is transported from recycling center r to anaerobic digestion plant m. For tth period, the amount of waste presented in each stage from previous period, is defined as beginning inventories, $E^{b}_{(t-1)}$, while the amount of waste that will be remained in each stage for the present period, is defined as ending inventories, E^{f}_{t}, and both inventories are considered. Let NT^{u}_{ijt}, NT^{u}_{jlt}, and NT^{u}_{rmt} denote the number of trips taken by truck u at day t for transporting waste from the ith depot to the jth collection station; the jth collection station to the rth recycling center; the jth collection station to the lth landfill; and the rth recycling center to the mth anaerobic digestion plant, respectively. Let GH^{I}, GH^{J}, GH^{R} and GH^{L} denote the amount of GHG emitted from processing one ton of waste (g/ton) at the depots, collection stations, recycling centers and landfills, respectively. Let GH^{u} denotes the amount of GHG emitted (g/km) by truck type u. Each stage has its own associated GHG emissions resulting from processing of waste, transportation of waste in trucks, or both as shown in Fig. 2.

It is assumed that: *(i)* waste is separated and sorted at collection stations, *(ii)* fixed cost daily rates for depots, collection stations, recycling centers, landfills, and anaerobic digestion plants are calculated as total fixed cost divided by expected economic lifespan in days, *(iii)* variable rates ($/ton) of depots, collection stations, recycling centers, landfills, and anaerobic digestion plants are calculated as operational costs per ton divided by the daily capacity of the collection stations, *(iv)* fuel and maintenance costs are proportional to traveled distance, *(v)* distances are measured from the centroids of the destinations, and *(vi)* the beginning inventory at the first period is zero for all stages.

3.1. Model description

There are several decision variables and parameters are shown in Appendix A (Nomenclature). Let FC_i, FC_j, FC_r, FC_l and FC_m denote the fixed costs per day t for depots, collection stations, landfills, recycling centers and anaerobic digestion plants, respectively. Then, the total

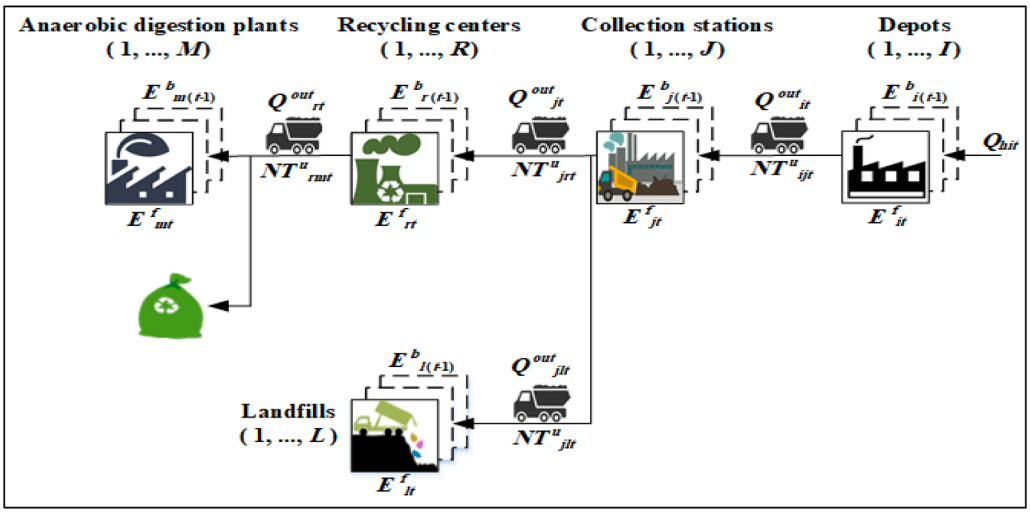


Fig. 1. Illustration of stages for the optimization model.

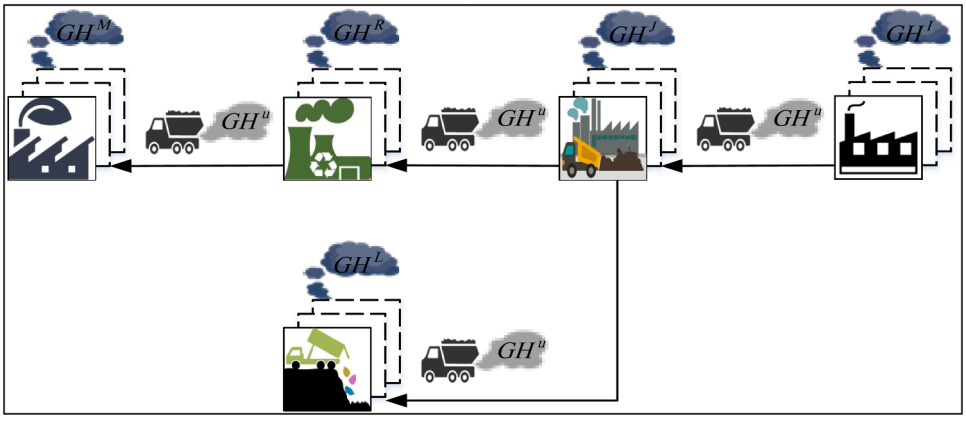


Fig. 2. Illustration of stages for the optimization model.

daily fixed cost, TFC, incurred in the system is calculated as given in Eq. (1).

$$TFC = \sum_{i=1}^{I} FC_i + \sum_{j=1}^{J} FC_j + \sum_{r=1}^{R} FC_r + \sum_{l=1}^{L} FL_l + \sum_{m=1}^{M} FC_m \quad (1)$$

Variable costs are incurred due to processing waste in each stage on day t. Let v_j, v_r, v_l, and v_m denote the variable costs per ton of the processing for the collection stations, recycling centers, landfills, and anaerobic digestion plants, respectively. Also, let Q_{jt}, Q_{rt}, Q_{lt} and Q_{mt} denote the total waste quantities at collection station j, recycling center r, landfills l, and anaerobic digestion plant m on day t, respectively. Then, the total daily variable cost, TVC, is calculated as stated in Eq. (2):

$$TVC = \sum_{t=1}^{T}\sum_{j=1}^{J} v_j \times Q_{jt} + \sum_{t=1}^{T}\sum_{r=1}^{R} v_r \times Q_{rt} + \sum_{t=1}^{T}\sum_{l=1}^{L} v_l \times Q_{lt} + \sum_{t=1}^{T} \times \sum_{m=1}^{M} v_m \times Q_{mt} \quad (2)$$

The total transportation cost of waste quantities is calculated by multiplying the transportation cost rate ($/ton.km$) by the distance travelled and quantity carried by truck type u. Let α-denotes the cost rate of transportation ($/ton.km$). Also, let d_{ij}, d_{jr}, d_{jl} and d_{rm} denote the distance travelled from depot i to collection station j, from collection station j to recycling center r and to landfill l, and from recycling center r to anaerobic digestion plant m, respectively. Finally, let Q_{ijt}, Q_{jrt}, Q_{jlt} and Q_{rmt} denote the quantity of waste transported on day t from depot I to collection station j, from collection station j to recycling center r, and to landfill l, and from recycling center r to anaerobic digestion plant m, respectively. The total cost of transporting waste quantities, TQC, is estimated using Eq. (3).

$$TQC = \alpha \times \left(\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J} d_{ij} \times Q_{ijt} + \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{r=1}^{R} d_{jr} \times Q_{jrt} + \sum_{t=1}^{T}\sum_{j=1}^{J} \times \sum_{l=1}^{L} d_{jl} \times Q_{jlt} + \sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{m=1}^{M} d_{rm} \times Q_{rmt} \right) \quad (3)$$

The total cost of fuel consumption is calculated by multiplying the unit cost of fuel, τ ($/L), by fuel consumed ($Liter/km$) by U truck types, travelled distance between any two stations, and number of trips over T days. Let TC_u denotes the fuel consumption of truck type u. Then, the total cost, TTC, of fuel consumption by all U trucks between all stages over T days is obtained as given in Eq. (4).

$$TTC = \tau \times \left(\sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} TC_u \times d_{ij} \times NT_{uij} + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{l=1}^{L} TC_u \times d_{jl} \times NT_{ujl} + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{r=1}^{R} TC_u \times d_{jr} \times NT_{ujr} + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{m=1}^{M} TC_u \times d_{rm} \times NT_{urm} \right) \quad (4)$$

Note that the transportation cost depends on a cost rate of the transported quantities. Consequently, if the quantities transported increased then the transportation cost will increase proportionally. While the fuel consumption cost depends on the travelled distance by each operated truck. Realistically, the cost rate for transported quantities is different from the fuel consumption cost.

Let GH^I, GH^J, GH^R, GH^L and GH^M denote the amount of GHG emitted from processing one ton of waste (g/ton) at I depots, J collection stations, R recycling centers, L landfills and M anaerobic digestion plants, respectively. Then, the total amount of GHG emissions, GHE, in the system is calculated as stated in Eq. (5).

$$GHE = GH^I \times \sum_{t=1}^{T} \sum_{i=1}^{I} Q_{it} + GH^J \times \sum_{t=1}^{T} \sum_{j=1}^{J} Q_{jt} + GH^R \times \sum_{t=1}^{T} \sum_{r=1}^{R} Q_{rt} + GH^L \times \sum_{t=1}^{T} \sum_{l=1}^{L} Q_{lt} + GH^M \times \sum_{t=1}^{T} \sum_{m=1}^{M} Q_{mt} \quad (5)$$

Let NT^u_{ijt}, NT^u_{jlt}, and NT^u_{rmt} denote the number of trips taken by truck type u to transport waste on day t from depot i to collection station j, the collection station j to recycling center r; from collection station j to landfill l, and from recycling center r to anaerobic digestion plant m, respectively. The GHG emitted from U truck types over T days is calculated by multiplying the amount of GHG emissions by both the total distance travelled between any pair of stages and number of trips. Then, the total amount, GHT, of GHG emitted due to transporting waste from depot i to anaerobic digestion plant m, is estimated as in Eq. (6):

$$GHT = GH^u \times \left(\sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij} \times NT_{uijt} + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{j=1}^{J} \times \sum_{r=1}^{R} d_{jr} \times NT_{ujrt} + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{l=1}^{L} d_{jl} \times NT_{ujlt} + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{r=1}^{R} \times \sum_{m=1}^{M} d_{rm} \times NT_{urmt} \right) \quad (6)$$

The unfilled capacity in transportation truck type u is calculated by subtracting the total transported quantity between two stages by truck type u on day t from its capacity. Let R_u denotes the capacity of truck type u. Then, the total unfilled capacities, Q^{TOT}, for U truck types over T days between all pairs of stages is calculated using Eq. (7).

$$Q^{TOT} = \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\left(NT^u_{ijt} \times R_u \right) - Q_{ijt} \right) + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{l=1}^{L} \left(\left(NT^u_{jlt} \times R_u \right) - Q_{jlt} \right) + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{r=1}^{R} \left(\left(NT^u_{jrt} \times R_u \right) - Q_{jrt} \right) + \sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{m=1}^{M} \left(\left(NT^u_{rmt} \times R_u \right) - Q_{rmt} \right) \quad (7)$$

Utilizing formulas 1 to 7, two objective functions will be developed; the first objective function, Z_1, aims at minimizing the total costs and environmental impacts of the waste management system as shown in Formula (8). The second objective function, Z_2, seeks maximizing the total of transported quantities between all pairs of stages and thereby optimizing methane production as stated in Formula (9).

$$Z_1 = \left(TFC + TVC + TQC + TTC + GHE + GHT + Q^{TOT} \right) \quad Min \ \ Z_1 \quad (8)$$

$$Z_2 = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} Q_{ijt} + \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{r=1}^{R} Q_{jrt} + \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{m=1}^{M} Q_{rmt} \quad Max Z_2 \quad (9)$$

The main constraints are as follows:

1 The waste quantity, Q_{it}, of at depot i on day t is equal to the total quantity, Q_{hit}, of waste transported to depot i from different areas on day t and the beginning inventory at depot i, $E^b_{i(t-1)}$, from previous day (t-1). That is in Eq. (10).

$$\sum_{h=1}^{H} Q_{hit} + E^b_{i(t-1)} = Q_{it} \quad (10)$$

2 The total quantity, Q_{it}, of waste at depot i on day t shall not exceed its capacity, C_i, as stated in Inequality (11).

$$Q_{it} \leq C_i, \forall i, t \quad (11)$$

3 Let NTA_{uijt} denotes the number of trucks that travel from depot i to collection station j on day t. The total quantity transported from depot i to collection station j on day t shall not exceed capacity of truck type u as stated in Inequality (12).

$$\sum_{j=1}^{J} Q_{ijt} \leq \sum_{u=1}^{U} NTA_{uijt} \times R_u, \forall i, t \quad (12)$$

4 The total quantity of waste, Q_{ijt}, transported from depot i to collection station j on day $t$is equal to the waste quantity, $Q^{out}{}_{it}$, leaving from depot i on day t as stated in Eq. (13).

$$\sum_{j=1}^{J} Q_{ijt} = \sum_{j=1}^{J} Q_{it}^{out}, \forall i, t \tag{13}$$

5 The number of trips to J collection stations on day t shall not exceed the number of available trips on the same day as given by Inequality (14).

$$\sum_{j=1}^{J} NT_{uijt} \leq NTA_{uijt}, \forall i, t, u \tag{14}$$

6 The ending inventory, $E^{f}{}_{it}$, at depot i on day t is equal to the total quantity of waste in depot i minus the quantity of waste leaving this depot on the same day, or Eq. (15)

$$Q_{it} - Q_{it}^{out} = E_{it}^{f}, \forall i, t \tag{15}$$

7 The ending inventory, $E^{f}{}_{it}$, at depot i should not exceed the percentage, λ, of the total waste which enters this depot as given in Inequality (16).

$$E_{it}^{f} \leq \lambda \times \sum_{h=1}^{H} Q_{hit}, \forall i, t \tag{16}$$

8 The total quantity, Q_{jt}, of waste at collection station j on day t cannot exceed its capacity, C_j, as shown by Inequality (17).

$$Q_{jt} \leq C_j, \forall j, t \tag{17}$$

9 The quantity of waste, Q_{ijt}, at collection station j on day t is equal to the waste transported quantity to collection station j from depot i on day t plus the beginning inventory, $E^{b}{}_{j(t\text{-}1)}$, at collection station j from previous day (t-1) as shown in Eq. (18).

$$\sum_{i=1}^{I} Q_{ijt} + E_{j(t-1)}^{b} = Q_{jt}, \forall j, t \tag{18}$$

10 A certain proportion, R_L, of the waste transported from depot i to collection station j goes to landfills as given in Eq. (19).

$$R_L \times \sum_{i=1}^{I} Q_{ijt} = \sum_{l=1}^{L} Q_{jlt}, \forall j, t \tag{19}$$

11 Let NTA_{ujlt} denotes the number of available trips by truck type u on day t from collection station j to landfill l. The quantity transported from collection station j to landfill l on day t must not exceed the capacity of the available trips as stated in Formula (20).

$$\sum_{l=1}^{L} Q_{jlt} \leq \sum_{u=1}^{U} NTA_{ujlt} \times R_u, \forall j, t \tag{20}$$

12 The waste quantity enters landfill l from collection station j on day t is equal to the waste quantity, $Q^{out}{}_{jlt}$, that leaves collection station j towards landfill l as stated in Eq. (21).

$$\sum_{l=1}^{L} Q_{jlt} = \sum_{l=1}^{L} Q_{jlt}^{out}, \forall j, t \tag{21}$$

13 Let NTA_{ujrt} denotes the number of available trips on day t from collection station to recycling center r by truck u. The quantity transported from collection station j to recycling center r on day t cannot exceed the capacity of the available trips as given in Eq. (22).

$$\sum_{r=1}^{R} Q_{jrt} \leq \sum_{u=1}^{U} NTA_{ujrt} \times R_u, \forall j, t \tag{22}$$

14 The quantity of waste that enters recycling center r from collection station j on day t, is equal to the quantity of waste, $Q^{out}{}_{jrt}$, that leaves collection station j towards R recycling centers on day t as stated in Eq. (23).

$$\sum_{r=1}^{R} Q_{jrt} = \sum_{r=1}^{R} Q_{jrt}^{out}, \forall j, t \tag{23}$$

15 The ending inventory, $E^{f}{}_{jt}$, at collection station j on day t is equal to the total quantity of waste in J collection stations minus the quantity of waste leaving the collection stations to R recycling centers and L landfills on the same day t as given in Eq. (24).

$$Q_{jt} - \sum_{r=1}^{R} Q_{jrt}^{out} - \sum_{l=1}^{L} Q_{jlt}^{out} = E_{jt}^{f}, \forall j, t \tag{24}$$

16 The number of trips to L landfills and R recycling centers on day t does not exceed the number of available trips on the same day as given in inequalities (25a) and (25b), respectively.

$$\sum_{l=1}^{L} NT_{ujlt} \leq NTA_{ujlt}, \forall u, j, l, t \tag{25a}$$

$$\sum_{r=1}^{R} NT_{ujrt} \leq NTA_{ujrt}, \forall u, j, r, t \tag{25b}$$

17 The ending inventory, $E^{f}{}_{jt}$, at collection station j cannot exceed the ratio λ of the total waste which enters that collection station j from all I depots as stated in Eq. (26).

$$E_{jt}^{f} \leq \lambda \times \sum_{i=1}^{I} Q_{ijt}, \forall j, t \tag{26}$$

18 The quantity of waste, Q_{lt}, at landfill l on a given day t is equal to the sum of total quantity of waste, Q_{jlt}, transported from all J collection stations to landfill l plus the beginning inventory, $E^{b}{}_{l\ (t\text{-}1)}$ at landfill l from previous day as given in Eq. (27).

$$\sum_{j=1}^{J} Q_{jlt} + E_{l(t-1)}^{b} = Q_{lt}, \forall l, t \tag{27}$$

19 The total waste quantity, Q_{lt}, at landfill l at any given day t, cannot exceed the landfill's capacity, C_l, as stated in Formula (28).

$$Q_{lt} \leq C_l, \forall l, t \tag{28}$$

20 The quantity of waste, Q_{jrt}, at recycling center r on day t is equal to the sum of total waste quantity transported to recycling center r from all J collection stations and the beginning inventory, $E^b_{r(t-1)}$, at recycling center r from previous day t-1 as given in Eq. (29).

$$\sum_{j=1}^{J} Q_{jrt} + E^b_{r(t-1)} = Q_{rt}, \forall r, t \tag{29}$$

21 The total waste quantity, Q_{rt}, at recycling center r on day t cannot exceed its capacity, C_r, as stated in Inequality (30).

$$Q_{rt} \leq C_r, \forall r, t \tag{30}$$

22 Let NTA_{urmt} denotes the number of available trips by truck u on day t from recycling center r to anaerobic digestion plant m. The quantity transported from recycling center r to M anaerobic digestion plants on day t by U truck types cannot exceed the capacity of the available trips as stated in Inequality (31).

$$\sum_{m=1}^{M} Q_{rmt} \leq \sum_{u=1}^{U} NTA_{urmt} \times R_u, \forall r, m, t \tag{31}$$

23 To ensure efficiency, the total waste quantity transported from recycling center r to anaerobic digestion plant m on day t, is equal to the waste quantity, Q^{out}_{rmt}, leaving from recycling center r on the same day as stated in Eq. (32).

$$\sum_{m=1}^{M} Q_{rmt} = \sum_{m=1}^{M} Q^{out}_{rmt}, \forall r, t \tag{32}$$

24 A certain proportion R_R of the quantity of the waste transported from J collection stations to recycling center r on day t generates revenue, REV_t, as given in Eq. (33).

$$R_R \times \sum_{j=1}^{J} Q_{jrt} = REV_t, \forall r, t \tag{33}$$

25 The ending inventory, E^f_{rt}, at recycling center r on day t is equal to the total quantity of waste at recycling center, r, minus both the quantity of waste leaving that recycling center to anaerobic digestion plant m and the quantity used to generate revenues on the same day. Mathematically as stated in Eq. (34).

$$Q_{rt} - \sum_{m=1}^{M} Q^{out}_{rmt} - REV_t = E^f_{rt}, \forall r, t \tag{34}$$

26 The number of trips to anaerobic digestion plant on any given day t, does not exceed the number of available trips on the same day as stated in Eq. (35).

$$\sum_{m=1}^{M} NT_{urmt} \leq NTA_{urmt}, \forall r, m, t \tag{35}$$

27 The ending inventory, E^f_{rt}, at recycling center r on day t cannot exceed the ratio λ of the total waste which enters the recycling center as mentioned in Formula (36).

$$E^f_{rt} \leq \lambda \times \sum_{j=1}^{J} Q_{jrt}, \forall r, t \tag{36}$$

28 The quantity of waste at anaerobic digestion plant m on day t is equal to the sum of total quantity of waste transported to anaerobic digestion plant m from R recycling centers and the beginning inventory at anaerobic digestion plant m, $E^b_{m(t-1)}$, from previous day t-1 as expressed in Eq. (37).

$$\sum_{r=1}^{R} Q_{rmt} + E^b_{m(t-1)} = Q_{mt}, \forall m, t \tag{37}$$

29 The quantity of waste at anaerobic digestion plant m cannot exceed the capacity of the m^{th} anaerobic digestion plant, C_m, as explained in Inequality (38).

$$Q_{mt} \leq C_m, \forall m, t \tag{38}$$

30 On any given day t, transported waste from depot i to collection station j must not exceed the capacity of the available trips by U truck types from all I depots to collection station j on the same day, as shown in Inequality (39).

$$\sum_{u=1}^{U} \sum_{i=1}^{I} NT_{uijt} \times R_u \geq \sum_{i=1}^{I} Q_{ijt}, \forall j, t \tag{39}$$

31 On day t, the transported waste from J collection stations to the landfill l must not exceed the capacity of the available trips from all collection station to landfill l by U truck types on the same day, as shown in Formula (40).

$$\sum_{u=1}^{U} \sum_{j=1}^{J} NT_{ujlt} \times R_u \geq \sum_{j=1}^{J} Q_{jlt}, \forall l, t \tag{40}$$

32 On day t, the capacity of the number of trucks which shall transport the waste from J collection stations to recycling center r on day t by U trucks must be greater than or equal to the quantity of waste transported from J collection stations to recycling center r by U truck types on the same day, as given in Inequality (41).

$$\sum_{u=1}^{U} \sum_{j=1}^{J} NT_{ujrt} \times R_u \geq \sum_{j=1}^{J} Q_{jrt}, \forall r, t \tag{41}$$

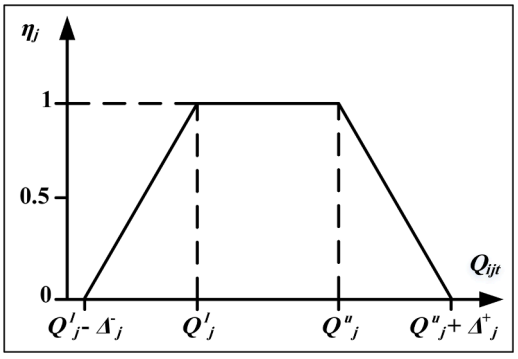


Fig. 3. Trapezoidal membership function for utilization of collection stations.

33 The capacity of the number of trucks, NT_{urmt}, used to transport waste from R recycling centers to anaerobic digestion plant m on day t must be greater than or equal to the quantity of waste transported from R recycling centers to digestion plant m by u truck types, as shown in the Inequality (42).

$$\sum_{u=1}^{U}\sum_{r=1}^{R} NT_{urmt} \times R_u \geq \sum_{r=1}^{R} Q_{rmt}, \forall m, t \tag{42}$$

34 Some variables should be integers and always positive as stated Inequality (43).

$$NT_{uijt},\ NT_{ujrt}, NT_{ujlt}, NT_{urmt}\ \geq 0\ \&\ Integer, \forall u, i, j, l, r, m, t \tag{43}$$

3.2. Satisfaction on utilization of collection stations

The aim of this satisfaction model is to maximize the daily utilization of collection stations while processing waste quantities. Let the upper and lower limits of the preferable target of quantities be denoted by Q^u_j and Q^l_j, respectively. Let Δ^-_j and Δ^+_j denote the maximum negative and positive allowable deviation from the preferable quantity target at collection station j, respectively. Also, let δ^-_j and δ^+_j denote any negative or positive deviation from the preferable quantity target at collection station j, respectively. Considering the capacity and costs issues at collection stations, the utilization of any collection station should range between averages of daily quantity of 150 to 200 ton and hence results in 100% satisfaction on utilization. However, quantities fall beyond Q^u_j or below Q^l_j will incur overtime or under time costs, respectively. The trapezoidal membership function for collection station j, η_j, shown in Figure 3 is, therefore, found appropriate for measuring the satisfaction level on utilization of collection stations. If the transported quantity, Q_{ijt}, to collection station j on day t falls within the preferable limits, then the satisfaction on utilization of collection station j will be 100%.

The objective function is then to maximize the sum of the membership functions of utilization at J collection stations as formulated in Formula (44).

$$Max \sum_{j=1}^{J} \eta_j \tag{44}$$

The objective function in Formula (44) is subjected to the following constraints:

a The amount of any negative deviation, δ^-_j and is how far is the transported quantity from the lower limit as shown in Inequality (45).

$$\sum_{i=1}^{I} Q_{ijt} + \delta^-_j \geq \Delta^-_j, \forall j, t \tag{45}$$

b The amount of any positive deviation, δ^+_j, is how far is the transported quantity from the upper limit as given in Formula (46).

$$\sum_{i=1}^{I} Q_{ijt} - \delta^+_j \geq \Delta^+_j, \forall j, t \tag{46}$$

c The value of the utilization membership function is calculated using Eq. (47).

$$\eta_j + \frac{\delta^-_j}{\Delta^-_j} + \frac{\delta^+_j}{\Delta^+_j} = 1, \forall j \tag{47}$$

d The value of the membership function should not be lower than the minimum required utilization, θ_j, of collection station j as stated in Inequality (48).

$$\eta_j \geq \theta_j, \forall j \tag{48}$$

e The ranges of the negative and positive deviations are decided as given in inequalities (49) and (50), respectively.

$$0 \leq \delta^-_j \leq \Delta^-_j, \forall j \tag{49}$$

$$0 \leq \delta^+_j \leq \Delta^+_j, \forall j \tag{50}$$

3.3. Satisfaction on utilization of recycling centers

The goal of this model is to maximize the membership function of the utilization of the recycling centers while processing waste quantities on day t. If the quantities transported, Q_{jrt}, to recycling center r on day t are within the preferable target, then the utilization membership function, η_r, of the recycling centers will be 100%. Let the upper and lower limits of the preferable target of quantities be denoted by Q^u_r and Q^l_r, respectively. Let Δ^-_r and Δ^+_r denote the maximum negative and positive allowable deviation from the preferable quantity target at recycling center r, respectively. Also, let δ^-_r and δ^+_r denote any negative or positive deviation from the preferable quantity target at recycling center r. The objective function is to maximize the satisfaction on the utilization of the R recycling centers for processing waste as stated in Formula (51).

$$Max \sum_{r=1}^{R} \eta_r \tag{51}$$

The objective function is subjected to the following constraints:

i The amount of any negative deviation, δ^-_r, at recycling center r is how far is the transported quantity from J collection stations from the lower limit as in Inequality (52).

$$\sum_{j=1}^{J} Q_{jrt} + \delta^-_r \geq \Delta^-_r, \forall r, t \tag{52}$$

ii The amount of any positive deviation, δ^+_r, at recycling center r for is how far is the transported quantity J collection stations from the upper limit. That is stated in Eq. (53).

$$\sum_{j=1}^{J} Q_{jrt} - \delta^+_r \geq \Delta^+_r, \forall r, t \tag{53}$$

iii The value of the utilization membership function at recycling center r is calculated using Eq. (54).

$$\eta_r + \frac{\delta^-_r}{\Delta^-_r} + \frac{\delta^+_r}{\Delta^+_r} = 1, \forall r \tag{54}$$

iv The membership function value should not be lower than the minimum required utilization, θ_r, of recycling center r as stated in Eq. (55).

$$\eta_r \geq \theta_r, \forall r \tag{55}$$

v The range of the negative and positive deviations are shown in inequalities (56) and (57), respectively.

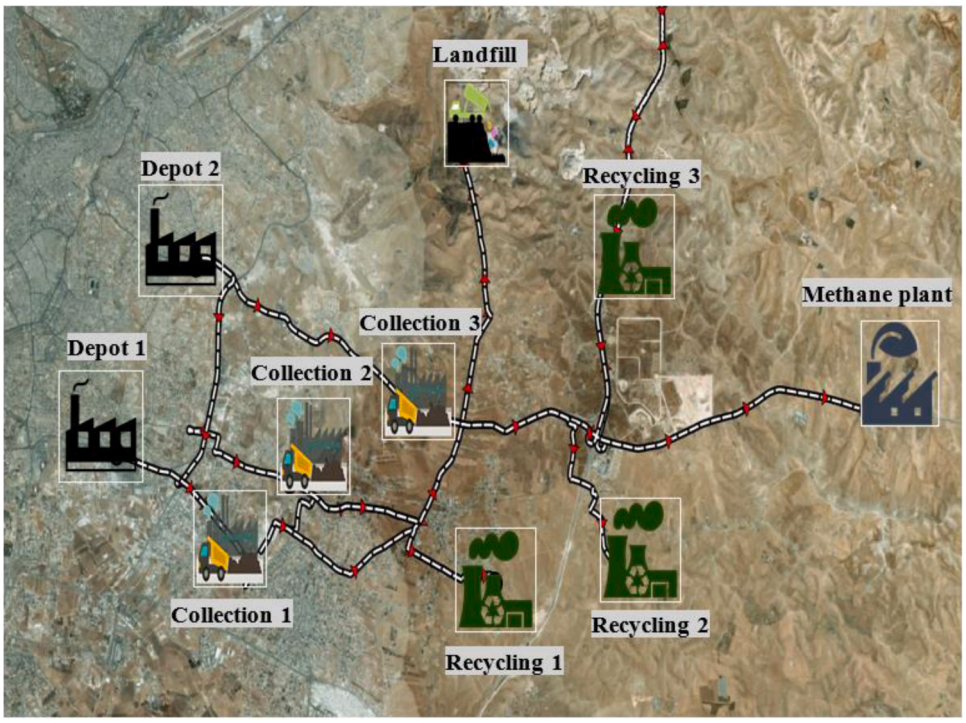


Fig. 4. Representation of case study on QGIS map.

Table 1
Values of model parameters.

Depot		Anaerobic digestion plant (ADP)	
Parameter	**Value**	**Parameter**	**Value**
Fixed cost, FC_i	75 ($)	Fixed cost, FC_m	250 ($)
Variable cost	-	Variable cost, v_m	8 ($/ton)
Capacity, C_i	32000 (tons)	Capacity, C_m	10000 (tons)
GH^I	0.9 (g/km)	GH^M	6.9 (g/km)
Landfill (L)		**Recycling center (RC)**	
Fixed cost, FC_l	100 ($)	Fixed cost, FC_r	200 ($)
Variable cost, v_l	4.5 ($/ton)	Variable cost, v_r	7 ($/ton)
Capacity, C_l	1000000 (tons)	Capacity, C_r	20000 (tons)
GH^L	0.6 (g/km)	GH^R	6.8 (g/km)
Collection station (CS)		**Distances**	
Fixed cost, FC_j	150 ($)	d_{ij}	30 (km)
Variable cost, v_j	5 ($/ton)	d_{jl}	29 (km)
Capacity, C_j	32000 (tons)	d_{jr}	25 (km)
GH^J	5.2 (g/km)	d_{rm}	35 (km)
R_1	10 (tons)	GH_1	1.30 (g/km)
R_2	8 (tons)	GH_2	1.26 (g/km)
R_3	4 (tons)	GH_3	1.18 (g/km)
τ	0.61 ($/L)	$\propto$	0.3 ($/km × g)
Δ^-_j	250 (tons)	Δ^+_j	100 (tons)
Δ^-_r	20 (tons)	Δ^+_r	150 (tons)
θ_j	0.8	θ_r	0.9
Q^u_j	200 (tons)	Q^l_j	150 (tons)
Q^u_r	70 (tons)	Q^l_r	100 (tons)
R_R	0.25	R_L	0.2

$$0 \le \delta^-_r \le \Delta^-_r, \forall r \quad (56)$$

$$0 \le \delta^+_r \le \Delta^+_r, \forall r \quad (57)$$

The complete optimization model is formulated by combining the presented three models: minimizing both total cost and GHG emission and maximizing satisfaction membership functions for the utilization of collection stations and recycling centers.

4. Analysis and results

A selected area in Amman, the capital of Jordan, was mainly considered to test the validity of the modelling contribution developed in this study. The population of the selected area is 250 thousand and the waste production per capita is 1.4 kg/day. On average, the collected waste per year is 120 thousand tons. This case study considers two depots (I =2), three collection stations (J=3), three recycling centers (R =3), one landfill (L =1), and a single anaerobic digestion plant (M =1) as represented on the QGIS map in Fig. 4.

Three types of trucks (U = 3) will be used to transport the waste quantities. Table 1 displays the general model parameters [1, 3], including fixed and variable costs for all stages with their capacities, the environmental parameters, and distances measured using QGIS software. Due to insufficient database about GHG emissions and costs for MSW management system in Jordan, the parameters values, such as, the operational costs (fixed and variable costs) and GHG emissions from trucks and facilities were adopted from some related studies [3].

The waste is tracked over a period of 6 days (T =6). The beginning inventories are zeros for all stages. The quantities that entered the depots on days (t= 1) to (t= 6) are 650, 440, 440, 710, 750, and 550 tons, respectively. Solving the proposed model using Lingo 18.0 software (Processor: Intel (R) Core (TM) i7-7700U; CPU @ 3.60 GHz, 3.60 GHz), the obtained optimal results for transported quantities and ending inventories for all stages were calculated by the model as displayed in Table B-1 shown in Appendix B. Given a certain number of available trips, *NTA*, the model calculates the optimal number of trips needed by each truck types (R_1, R_2, R_3) to transport all the waste. The obtained optimal trip numbers between stages are displayed in Table B-2 in Appendix B.

4.1. Results of optimization model

The optimal total cost of the system over six-day period was calculated for each stage and then displayed in Table 2. It is found that the average processing cost was 165.22 $/ton. From Table 2, collection center 1 on average over 6 days incurred variable costs, *TVC,* 2528.84

Table 2
Costs of the system.

	Day	Depot		Collection station			Landfill	Recyclingcenter			Aerobic Digestion Plant
		1	2	1	2	3		1	2	3	ADP
***TVC* ($)**	1	-	-	2150.00	1200.00	1200.00	409.50	1919.00	1197.00	1206.50	1535.63
	2	-	-	2690.00	1560.00	1560.00	809.55	2665.70	2449.10	2220.15	4397.40
	3	-	-	2807.00	1681.50	1668.00	206.82	2889.71	2789.01	1863.05	7511.02
	4	-	-	2806.02	1670.51	1558.00	306.52	2889.71	2789.01	1863.05	7511.02
	5	-	-	2150.00	1355.80	1106.60	417.50	1939.00	1197.00	1206.50	1535.63
	6	-	-	2570.00	1560.00	1560.00	709.55	2665.70	2449.10	2220.15	4397.40
On average ($/day)		-	-	2528.84	1504.63	1442.10	476.57	2494.80	2145.04	1763.23	4481.35
***TQC* ($)**	1	1753.50	1449.00	1280.55	792.00	986.40	-	472.50	342.90	285.12	-
	2	1744.05	1392.30	1656.03	1018.80	1255.44	-	742.46	752.76	567.72	-
	3	1681.05	1429.16	348.00	254.84	302.40	-	1105.65	689.58	857.76	-
	4	1744.05	1392.30	1656.03	1018.80	1255.44	-	742.46	752.76	567.72	-
	5	1681.05	1429.16	348.00	254.84	302.40	-	1105.65	689.58	857.76	-
	6	1753.50	1449.00	1280.55	792.00	986.40	-	472.50	342.90	285.12	-
On average ($/day)		1726.20	1423.49	1094.86	688.55	848.08	-	773.54	595.08	570.12	-
***TTC* ($)**	1	389.39	320.43	279.29	172.63	216.00	-	108.89	80.52	66.43	-
	2	385.31	309.91	372.92	236.68	279.08	-	166.53	164.70	132.86	-
	3	374.75	315.25	335.50	226.55	261.26	-	187.88	148.84	159.03	-
	4	374.75	315.25	335.50	226.55	261.26	-	166.53	164.70	132.86	-
	5	385.31	309.91	279.29	172.63	216.00	-	187.88	148.84	159.03	-
	6	385.31	309.91	372.92	236.68	279.08	-	108.89	80.52	66.43	-
On average ($/day)		382.47	313.44	329.24	211.95	252.11	-	154.43	131.35	119.44	-

Table 3
Total emissions resulting from system.

Stage	Component	Sources	Total (g)
Depots	Depot 1	Emissions (g)	1851.30
		Emissions from NT_{ijt} (g)	5223.04
	Depot 2	Emissions (g)	1596.60
		Emissions from NT_{ijt} (g)	3978.20
	Total = 12649.20		
Collection Station (CS)	CS1	Emissions (g)	7952.88
		Emissions from NT_{jlt} and NT_{jrt} (g)	4452.70
	CS2	Emissions (g)	4619.16
		Emissions from NT_{jlt} and NT_{jrt} (g)	2876.58
	CS3	Emissions (g)	4605.12
		Emissions from NT_{jlt} and NT_{jrt} (g)	3251.80
	Total= 27758.20		
Landfill	L1	Emissions	323.44
Recycling Center (RC)	RC1	Emissions (g)	5350.10
		Emissions from NT_{rmt} (g)	1974.00
	RC2	Emissions (g)	4606.18
		Emissions from NT_{rmt} (g)	1915.60
	RC3	Emissions (g)	3786.30
		Emissions from NT_{rmt} (g)	1525.92
	Total = 60211.60		
Anaerobic digestion plant	ADP	Emissions (g)	3024.70
The average amount of emissions from the system over 6 days			16820 g/day

Table 4
Utilization results for collection stations and recycling centers.

Utilization in Collection Stations							
Day (*t*)	**1**	**2**	**3**	**4**	**5**	**6**	**On average**
CS1	93%	98%	100%	84%	88%	89%	92%
CS2	80%	80%	81%	89%	83%	87%	83%
CS3	80%	80%	80%	90%	90%	81%	84%
Utilization in Recycling Centers							
Day (*t*)	**1**	**2**	**3**	**4**	**5**	**6**	**On average**
RC1	99%	90%	90%	95%	96%	96%	94%
RC2	90%	94%	92%	91%	95%	93%	93%
RC3	91%	98%	90%	94%	92%	92%	93%

($/day); thus, collection center 1 required higher variable costs with respect to collection station 2 and 3. Conversely, collection center 3 on average required lower variable costs, 1442.10 ($/day). However, the highest variable cost, *TVC*, in the system was incurred in aerobic digestion plant on average with 4481.35 ($/day); because of high-tech operations are needed to process waste at such plants. Moreover, on average landfills required the lowest variable costs in the system 476.57 ($/day).

On the other hand, the total transportation costs, *TQC*, were slightly different on average over six-day period for all stages. The close *TQC* results was incurred due to insignificant differences between distances of the trips. However, depot 1 incurred the highest *TQC* over the six days with average of 1726.20 ($/day). While recycling centers incurred on average the lower *TQC*, especially recycling center 3 which required 570.12 ($/day). In addition, it is noticed that the fuel consumption costs for trucks, *TTC*, will not seriously change over time except for unconditional incidents. On average, the *TTC* costs were changed from 119.44 ($/day) in recycling center 3 to 382.47 ($/day) in depot 1.

Finally, the emissions resulting from processing and transporting waste in the waste management system were estimated and displayed in Table 3, where it is noted that the average emission from depots, collection stations, and recycling centers is 16.82 kg/day. The higher emissions were emitted from recycling centers 60211.60 g; because of the advanced processes applied in recycling centers. In contrast, at the depots the waste does not require advance processes, so the emissions were 12649.20 g. These values can provide valuable information to transportation planning engineering on the effect of system emission on environment sustainability.

The utilization results for all collection stations and recycling centers over six-day period are displayed in Table 4. It is noted that the smallest satisfaction values for collection stations and recycling centers are 80% and 90%, respectively, which indicates acceptable utilization. The transported quantities for collection station 1 and recycling center 1 were the highest because the lower transportation costs. Thus, the highest average utilization for collection stations was found in collection station 1. As well, recycling center 1 achieved the highest utilization. The differences in utilization percentages were occurred because of the transported quantities.

The generated energy from processing wastes is different due to the operational capacities and waste types. Currently, the ADP are not operated in Jordan and there is lack of data about wastes. In similar studies, it was reported that the energy potential of 584 tons/day MSW is

Table 5
The expected generated energy in ADP.

Day (t)	Entering Q, Q_{rmt}	Energy potential (MWh)	Electrical power (MW)	Power to grid (MW)
1	102.38	568.70	7.19	4.73
2	190.79	1059.80	13.39	8.82
3	207.57	1153.01	14.57	9.60
4	250.62	1392.14	17.59	11.59
5	294.9	1638.11	20.70	13.63
6	289.19	1606.39	20.30	13.37
On average	222.58	1236.36	15.63	10.29

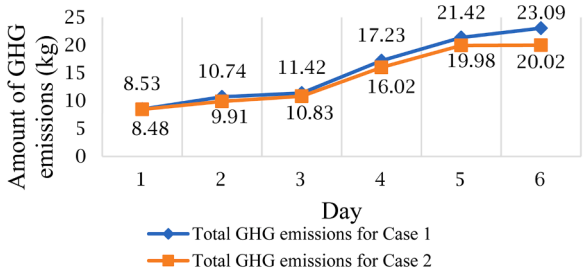


Fig. 5. Comparison between total emissions for Case 1 and Case 2.

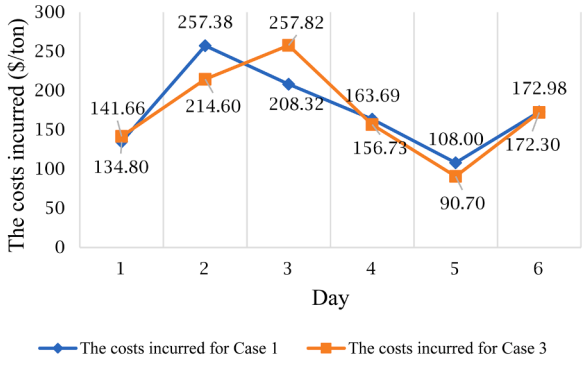


Fig. 6. Comparison between costs for Case 1 and Case 3.

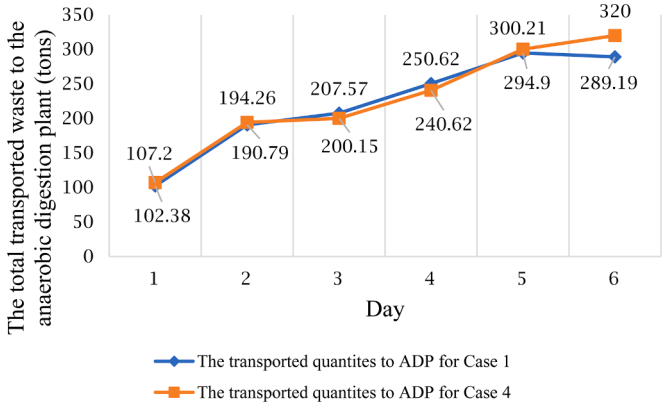


Fig. 7. Comparison between the quantities for Case 1 and Case 4.

about 3244 MWh, and the electrical power of the same quantity is about 41 MW whereas the power to grid is about 27 MW [31]. Table 5 summarized the expected generated energy from the transported quantities to ADP.

Table 6
Comparison between all applied case studies.

Case study	Total costs ($/ton)	Total GHG emissions (kg/day)	The transported quantities (ton/day)
Case 1 (All objective functions were considered)	165.22	16.82	222.58
Case 2 (Minimizing GHG emissions only)	138.78	14.21	282.52
Case 3 (Minimizing incurred costs only)	155.91	19.4	298.76
Case 4 (Maximizing the transported quantities only)	236.73	20.64	451.22

4.2. Sensitivity analysis

Further analysis was conducted on each objective function over six-day period. The beginning inventories are zero for all stages on the first day. The quantities that entered the depot are 650, 440, 440, 710, 750, and 550 tons on days 1 to 6, respectively. When the objective function only considered minimizing the GHG emissions (case 2) from waste transportation, the average GHG emissions per day are 14.21 kg/day. However, the average processing cost incurred is 138.78 $/ton and the average quantity transferred to the anaerobic digestion plant is 282.52 ton/day. A comparison between the total emissions is displayed in Fig. 5. In this case 2, the optimization model only aims to minimize the GHG emissions. In all days, the total emissions were slightly reduced. The results showed a reduction of GHG emissions by 0.05, 0.83, 0.59, 1.21, 1.44, and 3.07 kg over six-day period, respectively.

However, when only minimizing total costs (case 3) was considered as the objective function, the average processing cost over six-day period is 155.91 $/ton. While the average GHG emissions per day are 19.40 kg/day and the average quantity transferred to the anaerobic digestion plant is 298.76 ton/day. A comparison between the optimal costs is displayed in Fig. 6, where it is noticed that the incurred costs fluctuated between 257.82 ($/ton) on day 3 and 90.70 ($/ton) on day 5. The largest cost reduction was achieved in day 3 by 49.5 ($/ton). On the other hand, the smallest reduction was found to be 0.68 ($/ton) in day 6. The rates of cost reduction over the 6 days change in response to the transported quantities, and transportation.

Finally, when solving the model to maximize the total quantities only (case 4), and the average quantity transferred to the anaerobic digestion plant over the six-day period is 451.22 ton/day, while the average processing cost is 236.73 $/ton and the average GHG emissions are 20.64 kg/day. A comparison between the total quantities transported is displayed in Fig. 7. The objective function in case 4 was aimed to maximize the total transported quantities to ADP. Mostly, over the six-day period the transported quantities to ADP were increased. For example, an increasing by 4.82, 5.31, and 30.81 ton were achieved in days 1, 5, and 6, respectively.

Table 6 shows a comparison between all case studies with respect to the objective functions (Total costs, GHG emissions, and Transported quantities). In conclusion, the introduced optimization model is not

Table B-1

Optimal waste quantities to be transported (tons).

Day (t)	**Day 1**			**Day 2**			**Day 3**		
Depot	**Depot 1**	**Depot 2**		**Depot 1**	**Depot 2**		**Depot 1**	**Depot 2**	
Beginning Inv., $E^b_{i(t-1)}$	0.00	0.00		105.00	90.00		103.50	87.00	
Entering Q, Q_{hit}	350.00	300.00		240.00	200.00		230.00	210.00	
Q in depot, Q_{it}	350.00	300.00		345.00	290.00		333.50	297.00	
Exiting Q, Q_{it}^{out}	245.00	210.00		241.50	203.00		233.45	207.90	
Ending Inv., E^f_{it}	105.00	90.00		103.50	87.00		100.05	89.10	
Day (t)	**Day 4**			**Day 5**			**Day 6**		
Beginning Inv., $E^b_{i(t-1)}$	100.05	89.10		123.02	146.73		141.90	164.02	
Entering Q, Q_{hit}	310.00	400.00		350.00	400.00		250.00	300.00	
Q in depot, Q_{it}	410.05	489.10		473.02	546.73		391.90	464.02	
Exiting Q, Q_{it}^{out}	287.04	342.37		331.11	382.71		274.33	324.81	
Ending Inv., E^f_{it}	123.02	146.73		141.90	164.02		117.57	139.21	
Day (t)	**Day 1**			**Day 2**			**Day 3**		
Collection Center	**CS1**	**CS2**	**CS3**	**CS1**	**CS2**	**CS3**	**CS1**	**CS2**	**CS3**
Beginning Inv., $E^b_{j(t-1)}$	0.00	0.00	0.00	64.50	36.00	36.00	80.70	46.80	46.80
Entering Q, Q_{ijt}	215.00	120.00	120.00	204.50	120.00	120.00	199.45	121.90	120.00
Q in collection, Q_{jt}	215.00	120.00	120.00	269.00	156.00	156.00	280.15	168.70	166.80
Q to landfill, Q_{jlt}	43.00	24.00	24.00	40.90	24.00	24.00	39.89	24.38	24.00
Q to recycling, Q_{jrt}	107.50	60.00	60.00	147.40	85.20	85.20	128.20	76.84	76.08
Ending Inv., E^f_{jt}	64.50	36.00	36.00	80.70	46.80	46.80	112.06	67.48	66.72
Day (t)	**Day 4**			**Day 5**			**Day 6**		
Beginning Inv., $E^b_{j(t-1)}$	112.06	67.48	66.72	93.62	80.37	88.72	89.23	102.11	101.61
Entering Q, Q_{ijt}	200.00	200.41	229.00	203.82	260.00	250.00	200.15	200.00	199.00
Q in collection, Q_{jt}	312.06	267.89	295.72	297.44	340.37	338.72	289.38	302.11	300.61
Q to landfill, Q_{jlt}	62.41	53.58	59.14	59.49	68.07	67.74	57.88	60.42	60.12
Q to recycling, Q_{jrt}	156.03	133.94	147.86	148.72	170.18	169.36	144.69	151.05	150.31
Ending Inv., E^f_{jt}	93.62	80.37	88.72	89.23	102.11	101.61	86.81	90.63	90.18
Day (t)	**Day 1**			**Day 2**			**Day 3**		
Landfill	**L1**			**L1**			**L1**		
Beginning Inv., $E^b_{l(t-1)}$	0.00			91.00			179.90		
Entering Q, Q_{jlt}	91.00			88.90			88.27		
Q in landfill, Q_{lt}	91.00			179.90			268.17		
Ending Inv., E^f_{lt}	91.00			179.90			268.17		
Day (t)	**Day 4**			**Day 5**			**Day 6**		
Beginning Inv., $E^b_{l(t-1)}$	268.17			443.30			638.61		
Entering Q, Q_{jlt}	175.13			195.30			178.42		
Q in landfill, Q_{lt}	443.30			638.61			817.03		
Ending Inv., E^f_{lt}	443.30			638.61			817.03		
Day (t)	**Day 1**			**Day 2**			**Day 3**		
Recycling Center	**RC1**	**RC2**	**RC3**	**RC1**	**RC2**	**RC3**	**RC1**	**RC2**	**RC3**
Beginning Inv., $E^b_{r(t-1)}$	0.00	0.00	0.00	30.00	19.05	19.20	42.00	37.46	36.36
Entering Q, Q_{jrt}	100.00	63.50	64.00	110.00	110.00	97.80	109.92	63.00	108.20
Q in recycling, Q_{rt}	100.00	63.50	64.00	140.00	124.85	121.20	151.92	100.46	144.56
REV_t	25.00	15.88	16.00	27.50	26.45	25.50	27.48	15.75	27.05
Q to plant, Q_{rmt}	45.00	28.58	28.80	70.50	60.95	59.34	78.86	54.57	74.14
Ending Inv., E^f_{rt}	30.00	19.05	19.20	42.00	37.46	36.36	45.58	30.14	43.37
Day (t)	**Day 4**			**Day 5**			**Day 6**		
Beginning Inv., $E^b_{r(t-1)}$	45.58	30.14	43.37	43.67	50.39	73.01	73.10	45.12	78.38
Entering Q, Q_{jrt}	100.00	137.83	200.00	200.00	100.00	188.26	106.00	200.05	140.00
Q in recycling, Q_{rt}	145.58	167.97	243.37	243.67	150.39	261.27	179.10	245.17	218.38
REV_t	36.40	41.99	60.84	60.92	37.60	65.32	44.78	61.29	54.60
Q to plant, Q_{rmt}	65.51	75.59	109.52	109.65	67.68	117.57	80.60	110.33	98.27
Ending Inv., E^f_{rt}	43.67	50.39	73.01	73.10	45.12	78.38	53.73	73.55	65.51
Day (t)	**Day 1**			**Day 2**			**Day 3**		
Anaerobic digestion plants	**M1**			**M1**			**M1**		
Beginning Inv., $E^b_{m(t-1)}$	0.00			102.38			293.16		
Entering Q, Q_{rmt}	102.38			190.79			207.57		
Q in plant, Q_{mt}	102.38			293.16			500.73		
Ending Inv., E^f_{mt}	102.38			293.16			500.73		
Day (t)	**Day 4**			**Day 5**			**Day 6**		
Beginning Inv., $E^b_{m(t-1)}$	500.73			751.35			1046.25		
Entering Q, Q_{rmt}	250.62			294.90			289.19		
Q in plant, Q_{mt}	751.35			1046.25			1335.44		
Ending Inv., E^f_{mt}	751.35			1046.25			1335.44		

Table B-2
Optimal number of trips between stages.

Depot to collection station						
Day (*t*)	**Day 1**		**Day 2**		**Day 3**	
Depot *i*	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)
Depot 1	(12,15,4)	(12,15,2)	(12,12,10)	(12,12,7)	(14,12,6)	(14,12,0)
Depot 2	(16,15,4)	(13,10,0)	(16,9,4)	(14,8,0)	(14,12,4)	(12,11,0)
Day (*t*)	**Day 4**		**Day 5**		**Day 6**	
Depot 1	(14,12,6)	(14,12,0)	(16,15,4)	(13,10,0)	(16,9,4)	(14,8,0)
Depot 2	(12,15,4)	(12,15,2)	(14,12,4)	(12,11,0)	(12,12,10)	(12,12,7)
Collection station (CS) to Landfill						
Day (*t*)	**Day 1**		**Day 2**		**Day 3**	
CS *j*	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)
CS1	(4,3,0)	(2,3,0)	(2,3,0)	(2,3,0)	(4,3,4)	(4,0,0)
CS2	(4,2,4)	(2,0,1)	(2,0,4)	(2,0,1)	(6,0,0)	(3,0,0)
CS3	(4,3,0)	(0,3,0)	(4,6,0)	(0,3,0)	(2,6,4)	(0,3,0)
Day (*t*)	**Day 4**		**Day 5**		**Day 6**	
CS1	(6,0,0)	(3,0,0)	(2,6,4)	(0,3,0)	(4,6,0)	(0,3,0)
CS2	(4,3,0)	(0,3,0)	(4,2,4)	(2,0,1)	(2,3,0)	(2,3,0)
CS3	(2,0,4)	(2,0,1)	(4,3,4)	(4,0,0)	(4,3,0)	(2,3,0)
Collection station (CS) to Recycling Centers (RCs)						
Day (*t*)	**Day 1**		**Day 2**		**Day 3**	
CS *j*	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)
CS1	(4,3,12)	(4,3,11)	(10,6,4)	(10,6,0)	(14,3,0)	(13,0,0)
CS2	(4,0,12)	(4,0,5)	(10,0,0)	(9,0,0)	(6,6,9)	(5,3,1)
CS3	(4,3,8)	(4,2,1)	(8,3,0)	(7,2,0)	(8,0,4)	(8,0,0)
Day (*t*)	**Day 4**		**Day 5**		**Day 6**	
CS1	(10,0,0)	(9,0,0)	(4,3,12)	(4,3,11)	(8,3,0)	(7,2,0)
CS2	(6,6,9)	(5,3,1)	(4,3,8)	(4,2,1)	(4,0,12)	(4,0,5)
CS3	(14,3,0)	(13,0,0)	(8,0,4)	(8,0,0)	(10,6,4)	(10,6,0)
Recycling centers to Anaerobic digestion plants						
Day (*t*)	**Day 1**		**Day 2**		**Day 3**	
RC *r*	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)	**Available Trips** (R_1, R_2, R_3)	**Actual Trips** (R_1, R_2, R_3)
RC1	(0,6,9)	(0,6,0)	(4,4,3)	(4,4,0)	(8,6,0)	(8,0,0)
RC2	(0,2,4)	(0,2,4)	(8,2,2)	(5,1,1)	(8,2,2)	(4,2,0)
RC3	(4,4,6)	(3,0,0)	(6,0,0)	(6,0,0)	(3,10,4)	(2,7,0)
Day (*t*)	**Day 4**		**Day 5**		**Day 6**	
RC1	(8,2,2)	(4,2,0)	(0,6,9)	(0,6,0)	(3,10,4)	(2,7,0)
RC2	(8,6,0)	(8,0,0)	(6,0,0)	(6,0,0)	(0,2,4)	(0,2,4)
RC3	(4,4,3)	(4,4,0)	(4,4,6)	(3,0,0)	(8,2,2)	(5,1,1)

biased for specific objective function. Moreover, the formulated constraints should control the absence of any eliminated objective function. Case 1 (all objective functions were considered) showed more moderate results. Although case 2 (minimizing GHG emissions only) presented more optimal results in costs, GHG emissions, and transported quantities; but because the higher transported quantities that exceeded the preferable limits the utilization of collection stations and recycling centers were decreased. In case 3 (minimizing incurred costs only), with enforcing the optimization model to minimize the incurred costs only; it showed a good response but with higher GHG emissions. For case 4 (maximizing the transported quantities only), in response to the objective function high quantities were transported without restrictions on costs and GHG emissions.

5. Conclusions

MSW management system is a crucial public health service, which requires instant and powerful attention from decision makers during and aftermath of the COVID-19 crisis. However, the COVID-19 crisis has affected the industries on many levels. The volume of recyclable wastes was significantly increased during COVID-19 outbreak especially in the lockdown periods. Then, the decision makers should respond to this public health emergency by improving the MSW management system. Consequently, the objective of this article was to establish a more efficient MSW management system which maximizes the percentage of waste transported from depots to anaerobic digestion plants for recovery and recycling using mathematical optimization model. The model was built to maximize waste transported while minimizing both transportation costs and greenhouse gas emissions using different types of available trucks and ensuring sufficient utilization of the system. Results showed that, a quantity of 3540 tons of waste was assumed to enter the depot over a period of six-days. Accordingly, the model determined that an optimal quantity of waste could be transported from all depots to anaerobic digestion plants using 686 trips out of 1026 available trips with minimal environmental impacts. Revenues could be generated using 670.35 tons of non-organic recyclable waste.

In addition, the average optimal quantity which entered the anaerobic digestion plant was 222.58 ton/day of waste and this could potentially generate approximately 1236.36 MWh of energy potential, 15.63 MW of electrical power, and 10.29 MW power to grid. The minimal average system cost of this would be 165.22 $/ton. With the aid of the recommended waste management system and its optimization model, the public municipalities could utilize research results in developing effects plans that lead to reduce SWM transportations costs and enhance facilities utilization. In addition, the proposed optimization models supports building a sustainable MSW system that is valuable under normal and unexpected conditions; such as, COVID-19 outbreak. More importantly, 1 MW of electricity can be generated for each 50 tons of waste thus contributing to the major goal of the "green city" concept, which are found consistent with the obtained results by relevant studies on the same field. Future research considers analyzing the MSW system over longer periods with separation process at the source and stochastic collected quantities.

Appendix A. Nomenclature

(a) Decision variables

Variable	Description
Qit	Waste quantity at depot i on day t.
Qjt	Waste quantity at collection station j on day t.
Qrt	Waste quantity at recycling center r on day t.
Qlt	Waste quantity at the landfill l on day t.
Qmt	Waste quantity at anaerobic digestion plant m on day t.
$Qijt$	Waste quantity transported from depot i to collection station j on day t.
$Qjrt$	Waste quantity which is transported from collection station j to recycling center r on day t.
$Qjlt$	Waste quantity which is transported from collection station j to landfill l on day t.
$Qrmt$	Waste quantity that is transported from recycling center r to anaerobic digestion plant m on day t.
$NTuijt$	Number of trips by truck type u from depot i to collection station j on day t.
$NTujrt$	Number of trips by truck type u from collection station j to recycling center r on day t.
$NTujlt$	Number of trips by truck type u from collection station j to landfill l on day t.
$NTurmt$	Number of trips by truck type u from recycling center r to anaerobic digestion plant m on day t.
$REVt$	Generated revenues on day t.
E^f_{it}	Ending inventory at depot i on day t.
E^f_{jt}	Ending inventory at collection station j on day t.
E^f_{lt}	Ending inventory at landfill l at on day t.
E^f_{rt}	Ending inventory at recycling center r on day t.
E^f_{mt}	Ending inventory at anaerobic digestion plant m on day t.
δ_j^-	Negative deviation from the preferable target for collection station j.
δ_j^+	Positive deviation from the preferable target for collection station j.
δ_r^-	Negative deviation from the preferable target for recycling center r.
δ_r^+	Positive deviation from the preferable target for recycling center r.
η_j	Utilization membership function of collection station j.
η_r	Utilization membership function of recycling center r.

(a) Model parameters

Parameter	Description
Q_{hit}	Quantity of waste transported to depot i from cluster h on day t.
$E^b_{i(t-1)}$	Beginning inventory at depot i on day (t-1).
$E^b_{j(t-1)}$	Beginning inventory at collection station j from previous day (t-1).
$E^b_{l(t-1)}$	Beginning inventory at landfill l from previous day (t-1).
$E^b_{r(t-1)}$	Beginning inventory at recycling center r from previous day (t-1).
$E^b_{m(t-1)}$	Beginning inventory at anaerobic digestion plant m from previous day (t-1).
Λ	Ratio of ending inventory.
R_L	A proportion of waste that is moved to L landfill.
R_R	A proportion of waste that generates R revenues.
C_i	Capacity of depot i.
C_j	Capacity of collection station j.
C_l	Capacity of landfill l.
C_r	Capacity of recycling center r.
C_m	Capacity of anaerobic digestion plant m.
NTA_{uijt}	Available trips of trucks type u on day t to ship waste from depot i to collection station j.
NTA_{ujrt}	Available trips of trucks type u on day t to ship waste from collection station j to recycling center r.
NTA_{ujlt}	Available trips of trucks type u on day t to ship waste from collection station j to landfill l.
NTA_{urmt}	Available trips of trucks type u on day t to ship waste from recycling center r to digestion plant m.
GH^I	Amount of of GHG emitted from processing one ton of waste (g/ton) at I depots.

(*continued on next page*)

(*continued*)

Parameter	Description
GH^J	Amount of of GHG emitted from processing one ton of waste (g/ton) at J collection stations.
GH^R	Amount of GHG emitted from processing one ton of waste (g/ton) at R recycling centers.
GH^L	Amount of of GHG emitted from processing one ton of waste (g/ton) at L landfills.
GH^M	Amount of of GHG emitted from processing one ton of waste (g/ton) at M digestion plants.
GH^u	Amount of of GHG emitted (g/km) from truck type u.
FC_i	Fixed cost per day for depot i.
FC_j	Fixed cost per day for collection station j.
FC_r	Fixed cost per day for recycling center r.
FC_l	Fixed cost per day for landfill l.
FC_m	Fixed cost per day for anaerobic digestion plant m.
VC_j	Variable cost per day at collection station j.
VC_r	Variable cost per day at recycling center r.
VC_l	Variable cost per day at landfill l.
VC_m	Variable cost per day at digestion plant m.
α	Transportation cost (\$/ton × km).
d_{ij}	Distance travelled from depot i to collection station j.
d_{jr}	Distance travelled from collection station j to recycling center, r.
d_{jl}	Distance travelled from collection station j to landfill l.
d_{rm}	Distance travelled from recycling center r to anaerobic digestion plant m.
T	Fuel cost per Litre (\$/L).
TC^u	Fuel consumption of truck type u.
R_u	Capacity of transportation truck type u.
Q_j^u	Upper limit of the preferable quantity target at collection station j.
Q_j^l	Lower limit of the preferable quantity target at collection station.
Q_r^u	Upper limit of the preferable quantity target of recycling center r.
Q_r^l	Lower limit of the preferable quantity target of recycling center r.
Δ_j^-	Maximum negative allowable deviation from the lower preferable target Q^l_j at collection station j.
Δ_j^+	The maximum positive allowable deviation from the upper preferable target Q^u_j at collection station j.
Δ_r^-	The maximum negative allowable deviation from the lower preferable target Q^l_r at recycling center r.
Δ_r^+	The maximum positive allowable deviation from the upper preferable target Q^u_r at recycling center r.
θ_j	The minimum required utilization of collection station j.
θ_r	Minimum acceptable utilization of recycling center r.

Appendix B. Table

Table

References

[1] M.F. Badran, S.M. El-Haggar, Optimization of municipal solid waste management in port said – Egypt, Waste Manag. 26 (5) (2006) 534–545, https://doi.org/10.1016/j.wasman.2005.05.005.
[2] G. Tchobanoglous, H. Theisen, S. Vigil, Integrated Solid Waste Management, McGraw-Hill, New York, 1993.
[3] F. Habibi, E. Asadi, S.J. Sadjadi, F. Barzinpour, A multi-objective robust optimization model for site-selection and capacity allocation of municipal solid waste facilities: a case study in Tehran, J. Clean. Prod. 166 (2017) 816–834, https://doi.org/10.1016/j.jclepro.2017.08.063.
[4] S.C. Paolo, G. Manuela, L. Claudio, European trends in greenhouse gases emissions from integrated solid waste management, Environ. Technol. 2125 (2015), 1022230, https://doi.org/10.1080/09593330.2015.1022230.
[5] Jordan Green Building Council, Your Guide to Waste Management in Jordan Waste Sorting Informative Booklet, Jordan Green Building Council, 2016.
[6] P.S. Calabro, A. Satira, Recent advancements towards resilient and sustainable municipal solid waste collection systems, Curr. Opin. Green Sustain. Chem. (2020), 100375, https://doi.org/10.1016/j.cogsc.2020.100375.
[7] C. Mora, R. Manzini, M. Gamberi, A. Cascini, Environmental and economic assessment for the optimal configuration of a sustainable solid waste collection system: a 'kerbside' case study, Prod. Plan. Control 25 (9) (2013) 737–761, https://doi.org/10.1080/09537287.2012.750386.
[8] J. Põldnurk, Optimisation of the economic, environmental, and administrative efficiency of the municipal waste management model in rural areas, Resour. Conserv. Recycl. 97 (2015) 55–65, https://doi.org/10.1016/j.resconrec.2015.02.003.
[9] J. Zhao, F. Zhu, A multi-depot vehicle-routing model for the explosive waste recycling, Int. J. Prod. Res. 54 (2) (2016) 550–563, https://doi.org/10.1080/00207543.2015.1111533.
[10] J. Trochu, A. Chaabane, M. Ouhimmou, Reverse logistics network redesign under uncertainty for wood waste in the CRD industry, Resour. Conserv. Recycl. 128 (2018) 32–47, https://doi.org/10.1016/j.resconrec.2017.09.011.
[11] F.M. Tsai, T.D. Bui, M.L. Tseng, K.J. Wu, A causal municipal solid waste management model for sustainable cities in Vietnam under uncertainty: a comparison, Resour. Conserv. Recycl. 154 (2020), 104599, https://doi.org/10.1016/j.resconrec.2019.104599.
[12] F.M. Tsai, T.D. Bui, M.L. Tseng, M.K. Lim, J. Hu, Municipal solid waste management in a circular economy: a data-driven bibliometric analysis, J. Clean. Prod. 275 (2020), 124132, https://doi.org/10.1016/j.jclepro.2020.124132.
[13] S. Xiao, H. Dong, Y. Geng, X. Tian, C. Liu, H. Li, Policy impacts on municipal solid waste management in Shanghai: a system dynamics model analysis, J. Clean. Prod. 262 (2020), 121366, https://doi.org/10.1016/j.jclepro.2020.121366.
[14] I.R. Istrate, D. Iribarren, J.L. Gálvez-Martos, J. Dufour, Review of life-cycle environmental consequences of waste-to-energy solutions on the municipal solid waste management system, Resour. Conserv. Recycl. 157 (2020), 104778, https://doi.org/10.1016/j.resconrec.2020.104778.
[15] R.M. Deus, F.D. Mele, B.S. Bezerra, R.A.G. Battistelle, A municipal solid waste indicator for environmental impact: assessment and identification of best management practices, J. Clean. Prod. 242 (2020), 118433, https://doi.org/10.1016/j.jclepro.2019.118433.
[16] Y.D. Tong, T.D.X. Huynh, T.D. Khong, Understanding the role of informal sector for sustainable development of municipal solid waste management system: a case study in Vietnam, Waste Manag. 124 (2021) 118–127, https://doi.org/10.1016/j.wasman.2021.01.033.
[17] M.E. Batur, A. Cihan, M.K. Korucu, N. Bektaş, B. Keskinler, A mixed integer linear programming model for long-term planning of municipal solid waste management systems: against restricted mass balances, Waste Manag. 105 (2020) 211–222, https://doi.org/10.1016/j.wasman.2020.02.003.
[18] H.O. Iyamu, M. Anda, G. Ho, A review of municipal solid waste management in the BRIC and high-income countries: a thematic framework for low-income countries, Habitat Int. 95 (2020), 102097, https://doi.org/10.1016/j.habitatint.2019.102097.
[19] A.C.H. Pinha, J.K. Sagawa, A system dynamic modelling approach for municipal solid waste management and financial analysis, J. Clean. Prod. 269 (2020), 122350, https://doi.org/10.1016/j.jclepro.2020.122350.
[20] M. Paul, M.J. Bussemaker, A web-based geographic interface system to support decision making for municipal solid waste management in England, J. Clean. Prod. 263 (2020), 121461, https://doi.org/10.1016/j.jclepro.2020.121461.
[21] ... H.A.A. Hajar, A. Tweissi, Y.A.A. Hajar, R. Al-Weshah, K.M. Shatanawi, R. Imam, M.A.A. Hajer, Assessment of the municipal solid waste management sector development in Jordan towards green growth by sustainability window analysis

J. Clean. Prod. 258 (2020), 120539 https://doi.org/10.1016/j.jclepro.2020.120539.
A. P. Pinupolu, H. raja Kommineni, Best method of municipal solid waste management through public-private partnership for Vijayawada city, in: Proceedings of the Materials Today 33, 2020, pp. 217–222, https://doi.org/ 10.1016/j.matpr.2020.03.816.
B. ... & Y. Sarbassov, C. Venetis, B. Aiymbetov, B. Abylkhani, A. Yagofarova, D. Tokmurzin, V.J. Inglezakis, Municipal solid waste management and greenhouse gas emissions at international airports: a case study of Astana international airport, J. Air Transp. Manag. 85 (2020), 101789 https://doi.org/10.1016/j.jairtraman.2020.101789.
[24] S. Viau, G. Majeau-Bettez, L. Spreutels, R. Legros, M. Margni, R. Samson, Substitution modelling in life cycle assessment of municipal solid waste management, Waste Manag. 102 (2020) 795–803, https://doi.org/10.1016/j.wasman.2019.11.042.
[25] B.N. Kulkarni, V. Anantharama, Repercussions of COVID-19 pandemic on municipal solid waste management: challenges and opportunities, Sci. Total Environ. 743 (2020), 140693, https://doi.org/10.1016/j.scitotenv.2020.140693.
[26] A. Mostafaeipour, M. Rezaei, M. Jahangiri, M. Qolipour, Feasibility analysis of a new tree-shaped wind turbine for urban application: a case study, Energy Environ. 31 (7) (2020) 1230–1256, https://doi.org/10.1177/2F0958305X19888878.
[27] M. Rezaei, K.R. Khalilpour, M. Jahangiri, Multi-criteria location identification for wind/solar based hydrogen generation: the case of capital cities of a developing country, Int. J. Hydrog. Energy 45 (58) (2020) 33151–33168, https://doi.org/10.1016/j.ijhydene.2020.09.138.
[28] C. Wang, N.T. Nguyen, T. Dang, J.A Bayer, Two-stage multiple criteria decision making for site selection of solar photovoltaic (PV) power plant: a case study in Taiwan, IEEE Access 9 (2021) 75509–75525, https://doi.org/10.1109/ACCESS.2021.3081995.
[29] C. Wang, T. Dang, H. Tibo, D. Duong, Assessing renewable energy production capabilities using DEA window and fuzzy TOPSIS model, Symmetry 13 (2) (2021) 334, https://doi.org/10.3390/sym13020334.
[30] A. Al-Refaie, A. Al-Hawadi, S. Fraij, Optimization models for clustering of solid waste collection process, Eng. Optim. (2020) 1–14, https://doi.org/10.1080/0305215X.2020.1843165.
[31] R.A. Ibikunle, I.F. Titiladunayo, B.O. Akinnuli, S.O. Dahunsi, T.M.A. Olayanju, Estimation of power generation from municipal solid wastes: a case study of ilorin metropolis, Nigeria, Energy Rep. 5 (2020) 126–135, https://doi.org/10.1016/j.egyr.2019.01.005.

An ensemble machine learning model for the prediction of danger zones: Towards a global counter-terrorism

Sangita Pal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sangitapal2@outlook.com*

Prakash Kumar Behera, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, prakash_behera21@gmail.com*

Sulochana Nanda, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, sulochanananda1@hotmail.com*

Subhalaxmi Nayak, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, subhalaxminayak715@outlook.com*

Keywords:
Terrorism
Ensemble machine learning
Danger zones
Support vector machine

A B S T R A C T

Terrorism can be described as the use of violence against persons or properties to intimidate or coerce a government or its citizens to some certain political or social objectives. It is a global problem which has led to loss of lives and properties and known to have negative impacts on tourism and global economy. Terrorism has also been associated with high level of insecurity and most nations of the world are interested in any research efforts that can reduce its menace. Most of the research efforts on terrorism have focused on measures to fight terrorism or how to reduce the activities of terrorists but there are limited efforts on terrorism prediction. The aim of this work is to develop an ensemble machine learning model which combines Support Vector Machine and K-Nearest Neighbor for prediction of continents susceptible to terrorism. Data was obtained from Global Terrorism Database and data preprocessing included data cleaning and dimensionality reduction. Two feature selection techniques, Chi-squared, Information Gain and a hybrid of both were applied to the dataset before modeling. Ensemble machine learning models were then constructed and applied on the selected features. Chi-squared, Information Gain and the hybrid-based features produced an accuracy of 94.17%, 97.34% and 97.81% respectively at predicting danger zones with respective sensitivity scores of 82.3%, 88.7% and 92.2% and specificity scores of 98%, 90.5% and 99.67% respectively. These imply that the hybrid-based selected features produced the best results among the feature selection techniques at predicting terrorism locations. Our results show that ensemble machine learning model can accurately predict terrorism locations.

1. Introduction

Terrorism is a global menace which has stayed with humanity since the ancient times. It is a global concern because it has led to loss of lives, properties and insecurity, both nationally and globally. Previous studies have shown that the level of insecurity and uncertainty caused by terrorism has influenced decision-making to the extent that many people now make more conservative and less risky decisions often as a way of compensating for the feelings of insecurity caused by the disasters associated with terrorism [1]. One of the most popular terrorist activities is the 9/11 which has not only been recorded as one of the deadliest single attacks in history but also has attracted world's attention to the dire need of investigating, predicting and cubing this social and economic enemy called terrorism. Although terrorism has often been defined in a way that is specific to the subject-matter of the particular convention, it is described by Title 22 of the U. S. code as politically motivated violence perpetrated in a clandestine manner against non-combatants [2]. It is often committed so as to create a fearful state of mind in an audience very often different from the victims. Although some have argued that terrorism has some positive implications, the fact is that it cannot be reasonable even if it was later discovered to be benign [3]. This was premised on the fact that any action exhibited against a representative public order is an illicit act and can be characterized as oppressive and illegitimate. This is because terrorism itself is associated with violence, extremism, intimidation and acts against public and social order [4]. Terrorism has also been found to be associated with anti-colonial movement, cruelty and rivalry among political opponents [5, 6].

The inclinations and impacts of terrorist activities are often quantified and assessed by the number of incidents and casualties [4]. The causes of terrorism can be categorized into three layers; the situational factor, the strategic factors and the individual factors [6]. The situational factors include conditions that allow the possibility of radicalization and motivate feelings against the enemy as well as specific triggers for actions. The factors may, in the short run, mean an act to advertise a course; but its long term (strategic) factors may point to a political change, nationalists, and revolution or separatist movements. It may also seek to disrupt and discredit the process of government, in-fluence public attitude and prevent good governance, instill fear and sympathy as well as provoke a counter-reaction to legitimize their grievances. The individual factors deal with the worldview, psychology and character traits of terrorists. It assumes that terroristic personality or predisposition exists in humans.

The danger of terrorism to lives and properties in a global sense and the need to cub it is a good justification for this work. The impact of the application of machine learning and artificial intelligence to curbing the spread of terrorism cannot be overemphasized, the techniques of which can help prevent and combat terrorism, help the government and other policy-makers make informed decisions, concertize citizens and pilgrims on the kind of terrorism activities a particular region is exposed to and, indeed provide a cost-effective means of protecting lives and properties of citizens [5-9]. Machine learning can be used to make predictions about terrorism with such information as financial transactions, travel patterns, activities, as well as publicly available information such as social media. The expected outcome of this study will highlight the importance of global terrorism data and the ability to obtain useful in-formation from them vis-à-vis counter-terrorism [10, 11].

Datasets for machine learning prediction may contain hundreds of attributes, many of which may be irrelevant to the mining task, hence, the need for feature selection. Feature selection is an important exercise for providing further insights and pre-insights into any given dataset. It can also form a crucial part of data preprocessing especially in the case of machine learning models [12, 13]. It helps in assigning a score to the predictive variables based on how the explain the target variable [14]. Feature importance can be manual, statistical or machine learning based. Chi-Square feature selection compares each feature against the target variable to measure their relative dependence which has been widely used in literature [15-17]. Information gain is another effective feature selection technique. It works by calculating the reduction in entropy by splitting the dataset according to a given value of a random variable. Information gain is defined in Eqs. (1), (2) and (3).

$$\mathrm{Gain}(S_j) = E(P_i) - E(S_j) \tag{1}$$

whereas

$$E(P_i) = \sum_{i=1}^{n} P_i \log_2 P_i \tag{2}$$

and

$$E(S_j) = \sum_{i=1}^{s_j} I_j * E(Y_j) \tag{3}$$

where P_i is the ratio of conditional attribute P in the given dataset, S_j is the index for each attribute. The information gain implementation yields scores for each attribute which ranks based on importance. The higher the information gain score, the more contributive the feature is to the target variable.

Some research works have been done in the application of statistical, machine learning and deep learning techniques to the Global Terrorism Database (GTD) towards countering global terrorisms. Some authors have applied machine learning techniques such as Naïve Bayes (NB), K-Nearest Neighbors (KNN), Decision Trees and Support Vector Machine (SVM) to predict the terrorist groups responsible for a given incident [18, 19]. Some authors have also developed a recommender system using deep learning to predict the rate at which terrorists spread online propaganda [20]. A hazard grading model has been developed for the quantification of terrorist attacks using K-Means clustering [21]. The success of terrorist activities has also been predicted using Decision Tree Algorithms [22]. Some researchers also predicted if a particular terrorist attack is targeted at a government official, civilians, military, business or others [23]. Similar work on the computational approaches to terrorism prediction have also been reported in previous studies [24, 25] and [26]. Whether a given terrorist attack will be claimed by a known group or not has also been established by recent researchers using different machine learning techniques [27]. In another dimension, it is noted that none of these works has predicted locations, that is, continents where a given kind of terrorism can occur. This is a novelty of this work. Such information can aid the War on Terrorism, improve security consciousness and serves as good advice for tourists. In this work, we proposed an ensemble computational model for the prediction of danger zones as it relates to terrorism attacks.

2. Materials and methods

2.1. Study workflow

We developed a workflow for the study to guide the project execution. This workflow includes preprocessing, feature selection, training, testing and prediction (Fig. 1). It shows the systematic approach taken by the researchers in actualizing the study. The Global Terrorism Database (GTD) was used for this study. The dataset was preprocessed and split into training and testing. The ensemble machine learning model consists of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). The system learned from the training set and its efficiency was evaluated with the testing set. This bottom up approach helped us to anticipate all that could be needed for the study and we made efforts towards getting them. These components are described in the following sub-sections.

2.2. Data description

In this study, we obtained terrorism dataset from the University of Maryland's online repository known as the Global Terrorism Database (GTD). The data is maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START). The data description shows that an incident is categorized and recorded as terrorism if it meets any two of these three criteria: (1) if it is intentional, (2) if it entails some level of violence or immediate threat of violence and (3) if the perpetrators of the violence are sub-national actors. This dataset contains incidents of terrorism and attacks that were collected from news sources all over the world [28]. The GTD contains 181,691 instances of recorded incidents of terrorism attacks recorded from July 1970 to December 2017 from all countries of the world (Fig. 2, Table 1). The dataset has 139 attributes, many of which are sparse due to missing data. Detailed description of this dataset is available elsewhere [28-30]. In order to give our model a very good statistical power and to reduce fitting error, we removed all the attacks that do not have complete information required for the modeling.

2.3. Data pre-processing

The GTD was preprocessed via data cleaning, discretization, duplicate removal and normalization. To clean the data, columns which describe the same features were combined and only one of such was retained and some extracted features replaced some feature subset. For instance, we removed the column for event ID and retained day, month and year of occurrence. Columns with 20% or less of missing data were also removed to reduce the effect of missing data on our model. This reduced the number of features from 139 to 46. Data discretization was

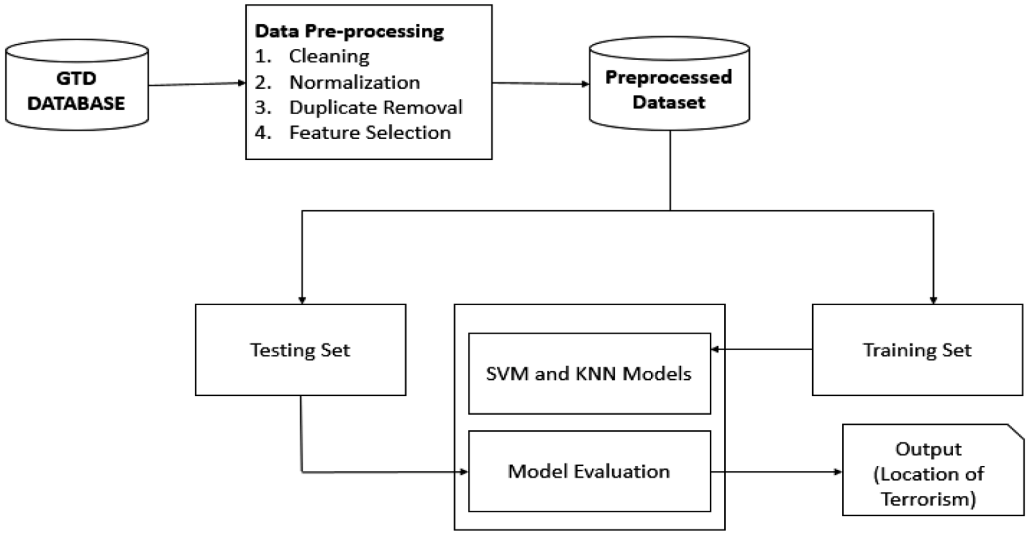
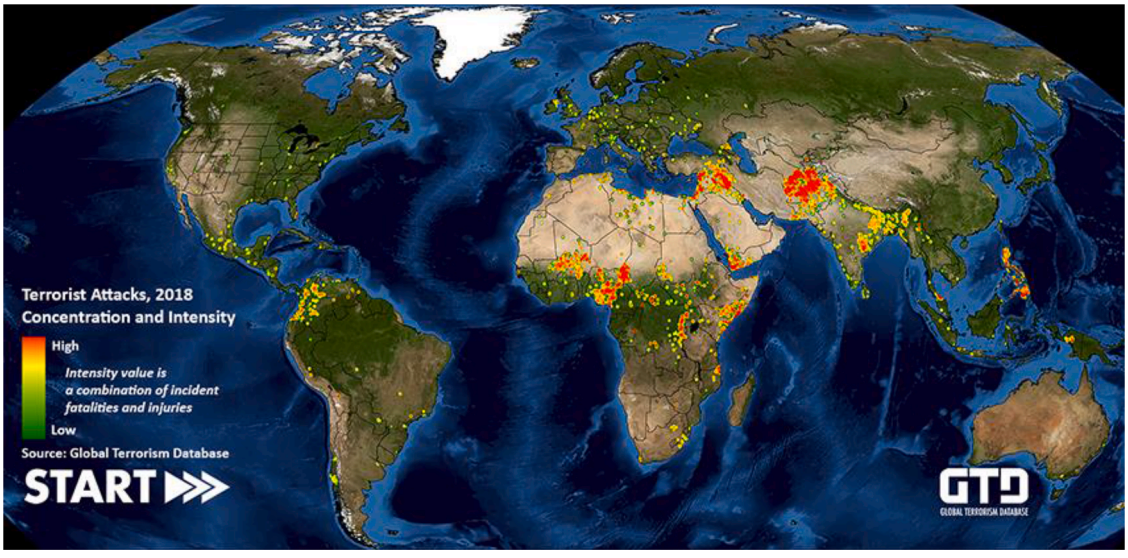


Fig. 1.. Workflow of the proposed model.



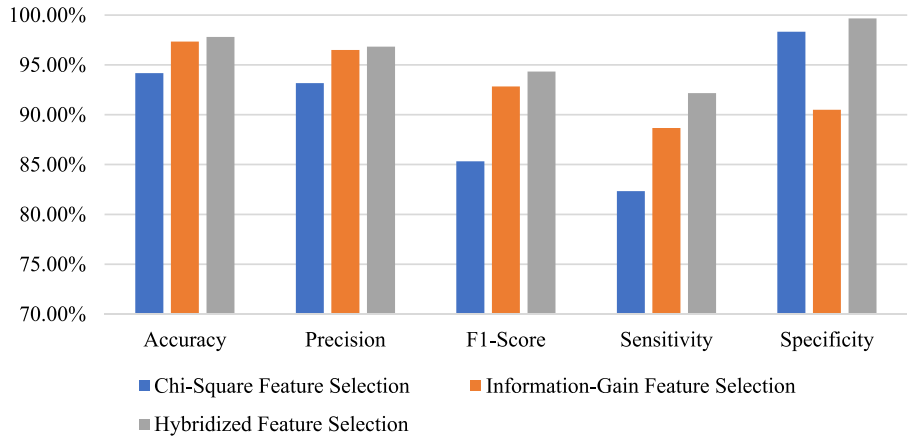


Fig. 2.. Map of the world showing the incidence of Terrorism
Source: Global Terrorism Data (https://start.umd.edu/news/global-terrorism-decreases-2018-recent-uptick-us-terrorist-attacks-was-sustained); Fig. 2: Performance metrics visualization.

Table 1.
Confusion matrix of the model predicting continents of terrorist attacks.

CLASS LABELS		Predicted						
		NA	AS	EU	SA	AF	OC	Total
	NA	4403	29	0	61	1	0	4494
Actual	AS	14	33,401	139	78	159	2	33,793
	EU	1	179	6747	0	73	0	7000
	SA	61	48	3	6109	0	0	6221
	AF	5	354	67	0	7920	3	8349
	OC	0	9	0	0	25	68	102
	Total	4484	34,020	6956	6248	8178	73	59,959

Note: NA → North America, AS → Asia, EU → Europe, SA → South America, AF → Africa, OC → Oceania.

done by converting the nominal fields in the GTD dataset into corresponding numerical values and data normalization was done using Eq. (4).

$$z = (x - \min(x))/(\max(x) - \min(x)) \quad (4)$$

where min and max are the respective minimum and maximum values for feature x and z is the normalized feature. Normalization is important because it reduces the variance of dataset thereby improving the fitness of our model and reducing bias. The countries were grouped into continents and each instance was assigned a continent code generated using longitude and latitude.

2.4. Feature selection

Two filter-based feature selection methods (Chi-Square and Mutual Information Gain) were used in this study to determine the "best fit" features in the pre-processed GTD datasets. This allowed us to:

I identify and focus on the most important features and
II reduce computation time because less features are relatively computationally less expensive.

Chi Squared and Mutual Information Gain helped in the individual ranking of the GTD features from which we selected features with higher scores. Hybrid selection was done by selecting the intersection of both selection techniques.

2.5. Modelling and implementation

Following data pre-processing and feature selection, we proceeded to the development of a predictive model using the selected features. Our target was to predict the continents of terrorist attack. Our prediction model is a mapping function which consists of the training dataset S of the original features. If n datasets are selected randomly for training the models using the known relevant attributes, the mapping $\xi : X_{j.k} \rightarrow Y_k$ defined as $\xi(X_{j.k}) = Y_k \ \forall$ terrorism incidents, k; where $X_{j.k}$ are the set of attributes, j is the incident counter Y is the output – predicted – class. Ensemble models are preferred to single models in order to reduce bias and increase the predictive power of the developed system. The motivation for using ensemble models is to reduce the generalization error of the prediction [31]. The base models used for the ensemble model were SVM and KNN. The dataset was divided to training and testing using ratio 70:30.

SVM and KNN are well known machine learning techniques used for classification and they are well described elsewhere [32-34]. SVM is a supervised machine learning technique used for labelled prediction. It uses the training set to learn the differences between groups to be classified. It is also called a maximum-margin classifier because it works by finding the optimal margin that best separates the groups to be classified SVM has a wide area of application because it can work with both linear and non-linear data. It also works well with high dimensional data and has high flexibility in modeling data from different sources [35]. In the SVM model, radial basis function (RBF) was used to handle the non-linearity in the data because it gave a very good result during experimentation. Details of KNN models can be found elsewhere [36-38]. KNN is also a supervised machine learning algorithm used for classification. Similar to SVM, KNN has application in many fields such as pattern recognition, data mining and intrusion detection. It is non-parametric which means that it does not make any assumption about the distribution of the data. It also uses the training set to learn the differences between groups to be classified. We evaluated the predictive power of the model using measures of sensitivity, specificity, accuracy and Area Under Curve (AUC).

3. Results

The data preprocessing reduced the number of features in the dataset to 21, these are Month of occurrence, Day of occurrence, Region of occurrence, Location, if it is intentional, if it entails some level of violence or immediate threat of violence and if the perpetrators of the violence must be sub-national actors, Is event clearly a terrorism, Connected to other Attacks, Is Suicide, Type of Attack, Target Type, Nationality of Target, Group Name Known, Not Affiliated to Group, Type of Weapon, Number of Fatalities, Property Damaged, Victims in Hostage, Ransom Given, Terrorism successful. The features selected from the hybridized feature selection process were collated and passed to the implemented ensemble classifier. The result of the prediction is presented in the confusion matrix in Table 1 while Table 2 shows the sensitivity, specificity and Area Under Curve (AUC).

The result shows that North America, Asia and South America have a high sensitivity of 0.98 while North America and Europe have a high sensitivity of 0.98. North America and South America produced the highest Area Under Curve of 0.99.

Table 2 shows that the AUC for the classes ranged from 0.83 to 0.99 with North America having the highest area and Oceania having the least. Also, the sensitivity ranged from 0.63 to 0.98 with the North America having the highest value and Oceania with the least value. The specificity ranged from 0.92 to 0.98 with North America having the highest value and South America having the least value.

Table 3 shows the summary result for the performance of the three models, i.e. the ensemble model using Chi Square feature selection, Information Gain feature selection and the hybridized method which combines the two. The result shows that the model built using the hybridized feature selection method gave the best performance in all the

Table 2.
Performance of the model per continent showing the sensitivity, specificity and AUC.

	Sensitivity	Specificity	AUC
NA	0.98	0.98	0.99
AS	0.98	0.96	0.98
EU	0.96	0.98	0.98
SA	0.98	0.92	0.99
AF	0.95	0.97	0.97
OC	0.67	0.93	0.83

Note: NA → North America, AS → Asia, EU → Europe, SA → South America, AF → Africa, OC → Oceania, AUC – Area Under the Curve.

Table 3.
Summary results of all the ensemble model using three feature selection techniques.

Performance Metrics	Chi-Square Feature Selection	Information-Gain Feature Selection	Hybridized Feature Selection
Accuracy	94.17%	97.34%	97.81%
Precision	93.17%	96.50%	96.83%
F1-Score	85.33%	92.83%	94.33%
Sensitivity	82.33%	88.67%	92.17%
Specificity	98.33%	90.50%	99.67%
Execution time	83.50s	81.47s	60.13s

performance metrics i.e. sensitivity, specificity, accuracy, precision and execution time. This is further visualized in Fig. 2.

4. Discussion

We developed 3 models for predicting the continent of terrorist attack using ensemble algorithm. The first one was based on Chi-Square future selection, the second was based on Information Gain feature selection while the third was based on the combination of Chi-Square and Information Gain called the hybridized method. Our results show that the hybridized model performed better than the other two. The hybridized model gave accuracy, sensitivity and specificity of 97.81%, 92.17% and 99.67% respectively which suggests that our model gave an excellent performance. These results depicts that the variables selected using the combination of Chi-Square and Information Gain predicts with a high accuracy the continent in which such a terrorism incident can occur

The main strength of this work is the use of machine learning in the prediction of location in which a given terrorism incident can occur. Previous works on the GTD have investigated features such as the extent of damage of attacks, likelihood of success of a terrorism attempt, likely targets of a terrorism incident and groups likely to be perpetrate an attack. To the best of our knowledge, none of the existing study worked on predicting the location of a terrorist attack. Another strength is the use of ensemble model that allowed us to combine two machine learning techniques namely SVM and KNN. This could explain the reason why the proposed model gave an excellent result. The choice of KNN and SVM was another strength. They are known to be fast and perform to be good with large dataset [39, 40]. GTD is a large dataset and the proposed model ran within 60.13 s which shows that it was time efficient.

One weakness identified in this study is that continent was predicted even though we believed that predicting countries will be better. However, this was avoided because this will amount to too many categories and machine learning models may not give a good result when the categories are many.

5. Conclusion

This study proposed an ensemble machine learning model which combines Support Vector Machine and K-Nearest Neighbor for prediction of continents susceptible to terrorism. Feature selection was used three feature selection techniques, Chi-squared, Information Gain and a hybrid of both methods. Our results show that ensemble machine learning model can accurately predict terrorism locations. Our results further showed that a combination of feature selection techniques offer an advantage over a single technique. This work established that terrorist location can be predicted using computational technique. This has a huge implication in the fight against terrorism and in protection of life and properties.

References

[1] K. Sacco, V. Galletto, E. Blanzieri, How has the 9/11 terrorist attack influenced decision making? Appl. Cogn. Psychol. 17 (2003) 1113–1127.
[2] T. Kapitan, Terrorism in the Arab-Israeli conflict, Terrorism (2004) 175–191.
[3] J.M. Sorel, Some questions about the definition of terrorism and the fight against its financing, Eur. J. Int. Law 14 (2003) 365–378.
[4] B.S. Frey, S. Luechinger, A. Stutzer, Calculating tragedy: assessing the costs of terrorism, J. Econ. Surv. 21 (2007) 1–24.
[5] M. Abdalsalam, C. Li, A. Dahou, S. Noor, A Study of the Effects of Textual Features on Prediction of Terrorism Attacks in GTD Dataset, Eng. Lett. 29 (2021).
[6] D.D. Atsa'am, R. Wario, F.E. Okpo, TerrorClassify: an algorithm for terror groups placement into hierarchical categories of casualties and consequences, J. Appl. Secur. Res. 15 (2020) 568–579.
[7] L. Hellmueller, V. Hase, P. Lindner, Terrorist Organizations in the News: A Computational Approach to Measure Media Attention towards Terrorism, Mass Communication and Society, 2021.
[8] S. Tickle, I. Eckley, P. Fearnhead, A Computationally efficient, High-Dimensional Multiple Changepoint Procedure with Application to Global Terrorism Incidence, arXiv preprint, 2020 arXiv:2011.03599.
[9] G.M. Campedelli, M. Bartulovic, K.M. Carley, Learning future terrorist targets through temporal meta-graphs, Sci Rep 11 (2021) 1–15.
[10] M.I. Uddin, N. Zada, F. Aziz, Y. Saeed, A. Zeb, S.A. Ali Shah, et al., Prediction of future terrorist activities using deep neural networks, Complexity 2020 (2020).
[11] A.S. Tawadros, Mapping terrorist groups using network analysis: egypt case study, J. Humanit. Appl. Soc. Sci. (2020).
[12] S. Hooker, D. Erhan, P.-.J. Kindermans, and B. Kim, "Evaluating feature importance estimates," 2018.
[13] G. König, C. Molnar, B. Bischl, M. Grosse-Wentrup, Relative feature importance, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 9318–9325.
[14] C. Liu, S. Gong, C.C. Loy, On-the-fly feature importance mining for person re-identification, Pattern Recognit 47 (2014) 1602–1615.
[15] X. Jin, A. Xu, R. Bie, P. Guo, Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles, in: International Workshop on Data Mining for Biomedical Applications, 2006, pp. 106–115.
[16] Y. Zhai, W. Song, X. Liu, L. Liu, X. Zhao, A chi-square statistics based feature selection method in text classification, in: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 160–163.
[17] I.S. Thaseen, C.A. Kumar, A. Ahmad, Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers, Arab. J. Sci. Eng. 44 (2019) 3357–3368.
[18] G.M. Tolan, O.S. Soliman, An experimental study of classification algorithms for terrorism prediction, Int. J. Knowl. Eng.-IACSIT 1 (2015) 107–112.
[19] M. Khorshid, T. Abou-El-Enien, G. Soliman, A comparison among support vector machine and other machine learning classification algorithms, IPASJ Int. J. Comp. Sci. 3 (2015) 26–35.
[20] R.S. Alhamdani, M.N. Abdullah, I.A. Sattar, Recommender System for Global Terrorist Database Based on Deep Learning, Int. J. Mach. Learn. Comput. 8 (2018).
[21] Y. Jun, X. Tong, H. Zhiyi, and L. Yutong, "Hazard Grading Model of Terrorist Attack Based on Machine Learning," 2019.
[22] T. Jani, "Predicting success of global terrorist," 2019.
[23] S.B. Salem, S. Naouali, Pattern recognition approach in multidimensional databases: application to the global terrorism database, Int. J. Adv. Comp. Sci. Appl. (IJACSA) 7 (2016).
[24] M.M. Khorshid, T.H. Abou-El-Enien, G.M. Soliman, Hybrid Classification Algorithms For Terrorism Prediction in Middle East and North Africa, Int. J. Emerg. Trend. Tech. Comp. Sci. 4 (2015) 23–29.
[25] S. Nieves, A. Cruz, Finding Patterns of Terrorist Groups in Iraq: a Knowledge Discovery Analysis, in: Ninth LACCEI Latin American and Caribbean Conference (LACCEI'2011), Engineering for a Smart Planet, Innovation, Information Technology and Computational Tools for Sustainable Development., Medellín, 2011.
[26] M. Adnan, M. Rafi, Extracting patterns from Global Terrorist Dataset (GTD) Using Co-Clustering approach, J. Independ. Stud. Res. 13 (2015) 7.
[27] V. Kumar, M. Mazzara, A. Messina, J. Lee, A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks, in: International Conference in Software Engineering for Defence Applications, 2018, pp. 146–158.
[28] G. LaFree, L. Dugan, Introducing the global terrorism database, Terror. Pol. Viol. 19 (2007) 181–204.
[29] L. Dugan, The Making of the Global Terrorism Database and Its Applicability to Studying the Life Cycles of Terrorist, The SAGE handbook of criminological research methods, 2011, p. 175.
[30] G. LaFree, The global terrorism database (GTD) accomplishments and challenges, Perspect. Terror. 4 (2010) 24–46.
[31] A. Diop, N. Emad, T. Winter, M. Hilia, Design of an Ensemble Learning Behavior Anomaly Detection Framework, Int. J. Comp. Inform. Eng. 13 (2019) 551–559.
[32] W.H. Land, J.D. Schaffer, The Support Vector Machine. The Art and Science of Machine Intelligence, ed, Springer, 2020, pp. 45–76.
[33] N. Shafiabady, L.H. Lee, R. Rajkumar, V. Kallimani, N.A. Akram, D. Isa, Using unsupervised clustering approach to train the Support Vector Machine for text classification, Neurocomputing 211 (2016) 4–10.
[34] Y. Zhang, G. Cao, B. Wang, X. Li, A novel ensemble method for k-nearest neighbor, Pattern Recognit 85 (2019) 13–25.
[35] B. Schölkopf, K. Tsuda, J. Vert, Kernel Methods in Computational Biology, ed, MIT Press, 2004, 2004.
[36] T. Thomas, A.P. Vijayaraghavan, S. Emmanuel, Nearest Neighbor and Fingerprint Classification. Machine Learning Approaches in Cyber Security Analytics, ed, Springer, 2020, pp. 107–128.

[37] R. Bandyopadhyay, Varying k-Nearest Neighbours: an Attempt to Improve a Widely Used Classification Model. Smart Intelligent Computing and Applications, ed, Springer, 2020, pp. 1–8.
[38] A.V. Joshi, Support vector machines. Machine Learning and Artificial Intelligence, ed, Springer, 2020, pp. 65–71.
[39] L. Demidova, I. Klyueva, A. Pylkin, Hybrid Approach to Improving the Results of the SVM Classification Using the Random Forest Algorithm, Procedia Comput. Sci. 150 (2019) 455–461.
[40] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, in: Advances in neural information processing systems, 2001, pp. 668–674.

Image steganography using genetic algorithm for cover image selection and embedding

Prasanta Kumar Sahoo, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, prasantakumarsahoo@outlook.com*

Anita Subudhi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, anitasubudhi89@gmail.com*

Binayini Pradhan, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, binayini.pradhan@gmail.com*

V. Raja, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, v_raja241@outlook.com*

Keywords:
Genetic Algorithm
Steganography
Authentication
Data integrity
Carrier image selection

A B S T R A C T

Data interchange through internet becomes an eminent technique and hence data security has become a big challenge in the field of communication with the increased use of internet. Demand for data authentication and effective means to control data integrity has been steadily increasing. Such a demand is due to the ease with which digital data can be tampered. Thus, cryptography and watermarking can be replaced with steganography for secure data communication and data privacy. In this paper, the carrier image is selected such that the payload/secret image and least significant bits of carrier image are matched with larger degree of compatibility and the hiding process introduces negligible changes in the resulted stego image based on genetic algorithm. In the proposed method we have achieved 30 to 40% improvements in the performance when compared to different existing methods. Selection of a suitable cover image and hiding the secret data to enhance the imperceptibility is a very challenging task. Genetic algorithm is used to ease the work of exploring an impossible task of selection from the trillions and millions of combinations.

1. Introduction

The aim of image steganography is to embed the secret data and prevent its presence for secret communication. The embedding of the secret data in the carrier image results in the stego image and the important objective of image steganography is to reduce the difference between the stego and carrier image [1]. The elementary parameters required to test steganography methods are imperceptibility, embedding capacity and security. Achieving all these parameters at once is a challenging task. For an instance, if we try to enhance the payload capacity, the visual quality or security factors may reduce similarly for other parameters [2]. LSB replacement steganography is the most conventional digital steganography technique where one can find some improvements in the evaluation parameters. In LSB steganography, the LSBs of the carrier image pixels are modified/ replaced by the secret image pixels. The payload image pixels can be either embedded in all the pixels of the carrier image or it can be embedded in a few selected pixels. The information about these selected cover pixels for embedding may be given through a secret stego-key [3]. In this paper, Image steganography using genetic algorithm is proposed to identify a suitable carrier image from a database of one hundred carrier images. The carrier image is identified based on the compatibility of the secret data and LSB's of carrier image such that the embedding process may leads to very little changes in stego image. In the proposed scheme along with identifying the carrier image genetic algorithm is used to find different ways to rearrange the secret data so that any changes in stego image can be minimized. Motivation and contribution of GA in image steganography is to identify a suitable carrier image and simultaneously embedding the payload data to enhance the visual quality of stego image which is a very challenging task. Hence to simplify the complexity incurred in the selection of carrier image from the huge database, Genetic algorithm is used. The proposed technique is implemented in two phases. The first phase discusses selection and comparison of eight cover images from the database of one hundred images. Statistical parameters of payload and carrier images are considered in this purpose. In the second phase, rearranging the payload data on the selected cover images and compares the conditions set, if the condition satisfies then it embeds the payload image. There are millions of ways to select the block size and their order for rearrangement. Exploring all these possibilities is the most difficult process. Hence GA is used to identify the nearest optimum possibility. The remaining paper is organized as follows: Section II brief out few related literature survey done on the topic selected. Section III explains

the basics of genetic algorithm and the proposed algorithm. The process of identifying the carrier images and GA implementation details are discussed in this section. In Section IV, the results of the proposed cover image selection and embedding method is discussed. Section V con-cludes the paper where the findings are summarized and scope for future works are highlighted.

2. Literature survey

Shah and Bichkar [4] described a Steganography technique based on Genetic Algorithm to identify best suitable place to embed the co-efficients selected and identifying quarter portion of an image to hide 2 bits pixel may leads to computation complexity. Shah and Bichkar [5] introduced evolutionary computation methods to select the most suitable locations and patterns for payload data hiding. Particle Swarm Optimization concept is used to select the possible sequence of data embedding and Genetic algorithm is used to find the best suitable patterns to alter payload data in order to generate least amount of modifications in carrier image and the PSNR obtained is around 42db. Goldberg [6] author of this book explained both normal and tutorial way - the computer concepts, mathematical tools, and research results which will provide both students and researchers to make use of genetic algorithms in order to solve problems in various fields. Pratik Shah and Rajankumar Bichkar [7] proposed payload data modification based image steganography scheme using genetic algorithm. The conditions and requirements used to alter and modify the payload data are controlled by genetic algorithm. Flexible chromosome concept is used in which genetic algorithm interpret the chromosome value in distinct ways and tries to identify the best suitable parameter that gives good visual quality stego images, since payload image is modified here extraction process at the receiver may becomes tedious. Jude Hemanth et al., [8] described the combined concepts of conventional and modified Genetic algorithm. Genetic algorithm associated with Fresnelet Transform and Discrete Ripplet Transform are used for embedding. Kanan and Bahram [9] proposed a tunable visually good quality image and spatial domain lossless data technique which depends on a genetic algorithm (GA) is proposed. PSNR and embedding capacity are used as evaluation parameters.

Bhattacharya et al.,[10] presented an approach to steganography where the payload image is first disturbed by Stego-Key and once again disturbed by a genetically developed, session based operator based on transposition technique. Followed by this step, the disturbed payload image is hidden within the cover image by variable bit exchange by making use of a hash function. At the receiver side, extraction of payload image from the disturbed image is retrieved by using bottom up concept. The session based TO and Stego Key are used to extract the secret image. Lin-Yu Tseng et al., [11] described Image embedding techniques used to hide a payload image into a carrier image. An improved genetic algorithm and a Pixel Adjustment optimization Process is employed to improve the visual quality of a Stego-image. Masoumeh Khodaei and Karim Faez [12] proposed an image embedding method by using LSB replacement method, where payload image is converted into useless picture by making use of a mapping function such that the error with respect hidden secret image pixels and LSB pixels of cover image is of minimal possible variables. Genetic algorithm is used for setting variables of mapping function to get the best criterion in the arrangement of the pixels. Shah and Bichkar [13] described a Steganography technique used to accomplish covert communication. The benefit of using steganography over other hiding techniques is its capacity to conceal the existence of secret communication. In steganography with image, the message is embedded in the carrier image, such that it generates very negligible modifications in the cover image. A genetic algorithm method is used to select a carrier image from a set of image database. The chosen carrier image should be more suitable with the given message. Further the transposition of the secret data is employed to improve the imperceptibility of the stego image. Pramanik et al., [14] Used image steganography to hide sensitive data inside a carrier image. To develop the technique, this method uses the combination of optimization based on Particle Swarm; wavelet transforms based on Bi-Orthogonal concept and Genetic Algorithm. Gu and Sun [15] described steganography technique based on the genetic algorithm, carry out image simulation on various types of mothers. For the images that need to be embedded, there may be change in sizes, the mother image can able perform good data embedding. There will not be much modification in the parent image. Deb et al., [16] discussed Multi-objective algorithms used for sorting of pixels based on non-domination and sharing. First fast approach on non-dominated sorting with less computational difficulty is presented. Second, an operator to select is presented which develops a crossing pool by the fusion of the parent and child populations and identify the best N possible solutions. Biswas and Bandyapadhay [17] introduced Steganography, the hiding technique, where the data is hidden in color image in transform domain by making use of Genetic Algorithm. Arbitrary multiple bits are considered to hide the message and for bits selection they used frequency domain and hash function. Yang and Honavar [18] used genetic algorithm to identify proper subsets, to get multi optional estimation in general terms like accuracy and costs in connection with the specific features. Lin et al., [19] developed a method used for data embedding is the LSB substitution. A genetic algorithm is used to find solution for the problem of embedding data in the rightmost m LSBs of the cover image, which may take a large computation time to solve for the optimal result even when m is large. Wu MN et al., [20] suggested a method to enhance the image quality of the stego-image, by making use of the LSB substitution and genetic algorithm. Two optimal strategies are suggested one for global optimization and the other for local estimation. Mahdi Ramezani and Shahrokh Ghaemmaghami [21] described an adaptive steganography method which depends on image contrast. Authors exploited the average difference between the pixels binary level values in 2×2 blocks of spatially non-overlapping blocks and their mean pixel binary level in order to find appropriate blocks for data hiding.

Ramezani, and Ghaemmaghami [22] make a study on feature-based steganalytic method using four different classification methods: Fisher linear discriminate, Gaussian naive Bayed, multilayer preceptor, and k nearest neighbor, are compared for steganalysis of doubted images. They exploited statistics of the histogram, wavelet transforms, center location characteristic operator of the histogram and co-occurrence blocks for feature extraction process. Genetic algorithm is used to minimize the proposed features sizes and to find the best subset. Shen Bian Yang and Xiamu Niu [23] recommended a heuristic genetic algorithm method for data embedding in a host image. After hiding the payload data in least significant bit of the host image, the values on the pixel of steg-image are altered by the genetic algorithm to maintain their statistic features. Mandal and Khamrui [24] presented an authentication/data hiding method through steganography approach and using GA. High dimension message/ image can be embedded in spatial domain of 3×3 masks from the carrier image in major row manner. Four bits of the authenticated secret image is hidden per byte of the host image in the least significant four bit of each pixel. Mutation is careered on the modified image. Nosrati and Karimi [25] authors reviewed at the usage of genetic algorithms in steganography. Image steganography along with genetic algorithm is one among them to provide good embedding capacity and visually less distorted stego image. Hanani et al., [26] the authors approached "before embedding hiding techniques" by trying to select suitable locations in the host image to hide the data with the less modifications of bits. Hence using genetic algorithm segmentation is done in order to transform the LSBs and data streams to the groups of blocks for embedding. After selecting the locations, payload data blocks are hidden and generated a key file to provide the data extraction at the given the data addresses. Sethi and Kapoor [27] proposed a Steganography technique adopting Elliptic Curve Cryptography and Discrete Cosine Transform steganography along with genetic algorithm.

Fig. 1. Representation of Gene, Chromosome and Population in a Genetic Algorithm

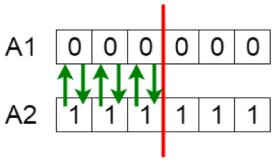


Fig. 2. Exchanging genes among parents

3. Genetic algorithm

Genetic Algorithm is evolved from Darwin's theory of evolution and adopts the concept of, "survival of the fittest". GA is a concept which depends on natural selection of fittest individuals in order to reproduce offspring. The reproduced off springs will compete among themselves for their existence and the one who fit for survival will seed for next generation. The number of individuals increases as the generation progress. Genetic algorithm undergoes following steps:

- Process of Natural Selection

The notion of natural selection begins by selecting the fittest individuals from the crowd. They produce offspring which retain some of the features of the parents and this will be added to the next generation. Survival of offspring's depends on the fitness of the parents and the process iterates to get fittest individuals in the crowd. This methodology can be used for a search problem.

In general five phases are there in a genetic algorithm.

- Initial population
- Fitness function
- Selection
- Crossover
- Mutation
- Initial Population

The process starts with a group of individuals called as Population. Each person can be used to dissolve a problem. A Person is characterized by a set of features known as Genes.

These variables are assigned into a string to develop a solution called as Chromosome. Genes are embedded in a chromosome. Binary ones and zeros are used to represent set of genes of an individual and is denoted by an alphabet as shown in Fig. 1.

• Function used for Fitness

The fitness function shows how an individual is able to fit in order to participate with other people. After participation algorithm assigns a fitness score to each person. The fitness score will be used to select an individual for next reproduction.

• Selection

The selection phase of an algorithm is used to identify the suitable individuals and allow them to circulate their genes to the next population. Based on fitness scores two pairs of characters are selected called as parents.

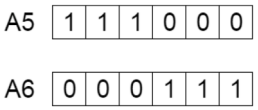


Fig. 3. New offspring generated.

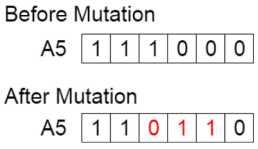


Fig. 4. Mutation concepts before and after

• Crossover

The important phase of a GA is the Crossover phase where a crossover point is selected randomly from the genes inside the population. For instance, let the crossover point to be 3 as shown in Fig. 2.

Fig. 3.

Then the new offspring generated denoted as A5 and A6 as shown in Fig. 1. 6.3 are included in to the population.

• Mutation

Few bits of the bit string may be transposed in a lower random probability which is subjected to a concept called mutation and the same is shown in Fig. 4.

In general there are seven different types of mutation functions as listed bellow;

- Bit string mutation (bits flipped at random places)
- Flip bit (selected bits are flipped)
- Boundary (genomes are replaced by upper and lower bounds randomly)
- Non –uniform (tunes solution at the later stages)
- Uniform (selected genome replaced by upper and lower bounds)
- Shrink (Replaced with the Gaussian variable)

• Termination

If the offspring produced are not significantly different from the previous population then the generation of the new off spring will be terminated by the algorithm. Thus the generated off springs itself will be the solution to the problem.

• Comments

The population will be of constant number. As population grows, lesser fitness characters will be vanished and provides place for newly generated offspring.

The above mentioned sequence of phases is repeated in order to generate individuals in each new population which should be better than the previous generation.

3.1. Cover image selection and embedding process

The selection of fittest from the group of individuals is based on natural selection. In the proposed method the crowd is considered as 100 cover images. For the selection of fittest individual in turn suitable cover image a method in genetic algorithm is considered called as **Fitness proportionate selection** or **roulette wheel selection**, it is a genetic operator used for the selection of suitable images foe embedding. In the RW selection process, the fitness function set a fitness value to the possible solutions or chromosomes. This fitness value is used as a probability function in the selection process of each single chromosome. If F_i is the fitness of an image i in the crowd then the probability of selecting suitable image is given by the probability function given by equation $P_{i,}$

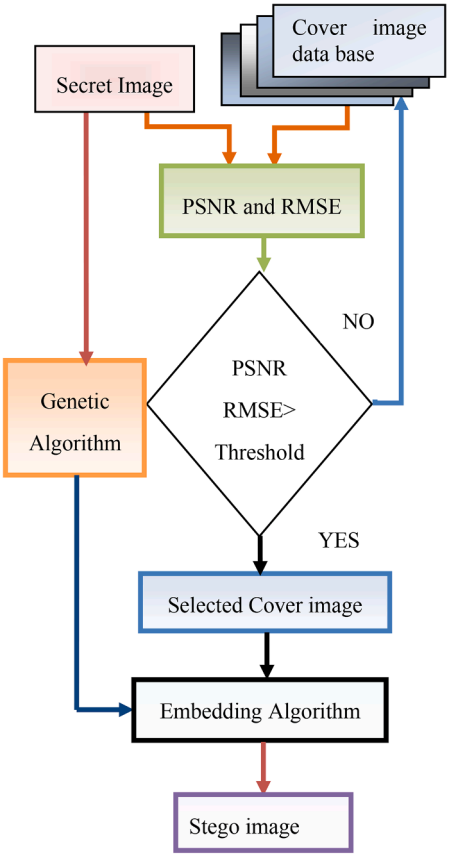


Fig. 5. Embedding Process

$$P_i = \frac{F_i}{\sum_{i=1}^{x} F_i}$$

where x is the number of images in the crowd.

This process of selection resembles to a casino Roulette wheel. In general based on the fitness value, a percentage of the wheel is allocated to each of the possible selections. The selection process can be done by dividing the fitness value of a selection by the total fitness of all the selections, hence normalizing them to 1. The initial population is created with 20 images and assigned fitness function in terms of mean square error and based on least error suitable cover image is selected and then crossover point 2 function is applied by exchanging the genes inside the population. From the cross over function child1 becomes parent 1 and child 2 becomes parent 2 and so on.

3.2. Mutation

Few bits of the bit string may be transposed in a lower random probability which is proportional to 1/L, where L is the length of binary vector. In the proposed method mutation rate can be varied from 0.1 onwards to measure the peak signal to noise ratios.

3.3. Termination

If the offspring produced are not significantly different from the previous population then the generation of the new off spring will be terminated by the algorithm. The target set in proposed algorithm is 80% PSNR and RMSE <1. If the obtained solution meets the target then algorithm terminates, if not the algorithm repeats infinite time to get the solution. Thus the generated off springs itself will be the solution to the problem.

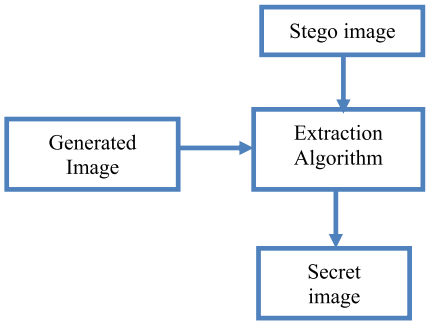


Fig. 6. Extraction Process of the proposed method

Table 1
PSNR comparisons between proposed method and existing method [7]

Payload image	Selected cover image	Existing method [15] PI size=512 × 256	Existing method [7]	Proposed method
Living Room	House	53.414	52.22	79.14
Lake	Pirate	53.714	52.17	78.65
Pepper	Baboon	53.74	52.25	79.38
Walkbridge	Cameraman	51.78	52.33	81.46

The flowchart of the proposed algorithm is shown in Fig. 5.
Algorithm to embed payload data:

- The crowd is created using set of cover image database 28,29
- The target or fitness value used to select suitable cover image for payload image embedding is in terms of PSNR ranges from 50 to 80% and RMSE<1. Genetic algorithm compares the payload image and the images from cover image database continuously to fit for the fitness value by generating chromosomes.
- The cover image selection is based on Roulette wheel algorithm. Once the images reach the fitness value, genetic algorithm generates its own image in which payload image is embedded using LSB substitution and the generated image fit the fitness value which is called as stego image.
- PSNR value is calculated between the generated image and the cover image selected.

3.4. Extraction process

The extraction process of the proposed method involves stego image as input to the extraction algorithm and the blockdiagram is shown in Fig. 6.

- The least significant bits are extracted and saved as array of bit strings
- Interpret with the chromosome generated by the genetic algorithm and modify accordingly the array of bit stream
- Compare with the stored database and rearrange the obtained LSBs in a proper order toget back the original payload image.

Table 2
PSNR and SSIM value comparisons between proposed method and existing method [13]

Selected cover image	Existing method [13]	Proposed method	SSIM [13]	SSIM
Flower	52.06	78.25	0.9989	0.9999
Beach	52.48	80.42	0.9986	0.9988
Car	52.40	79.76	0.9979	0.9981
Building	52.27	78.61	0.9986	0.9989

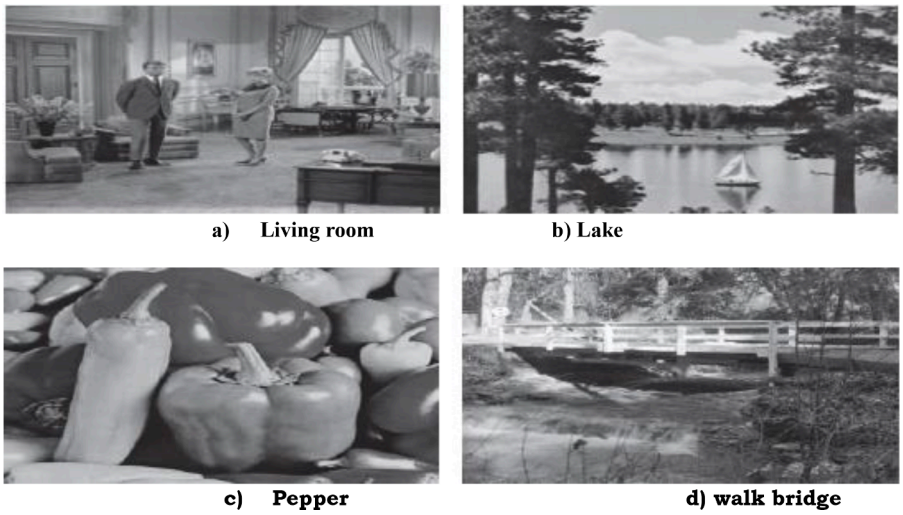


Fig. 7. show the different cover images and their stego images, which shows less distortion in the stego image and hence subjective analysis for the proposed method will results in no suspicious regarding steganography.

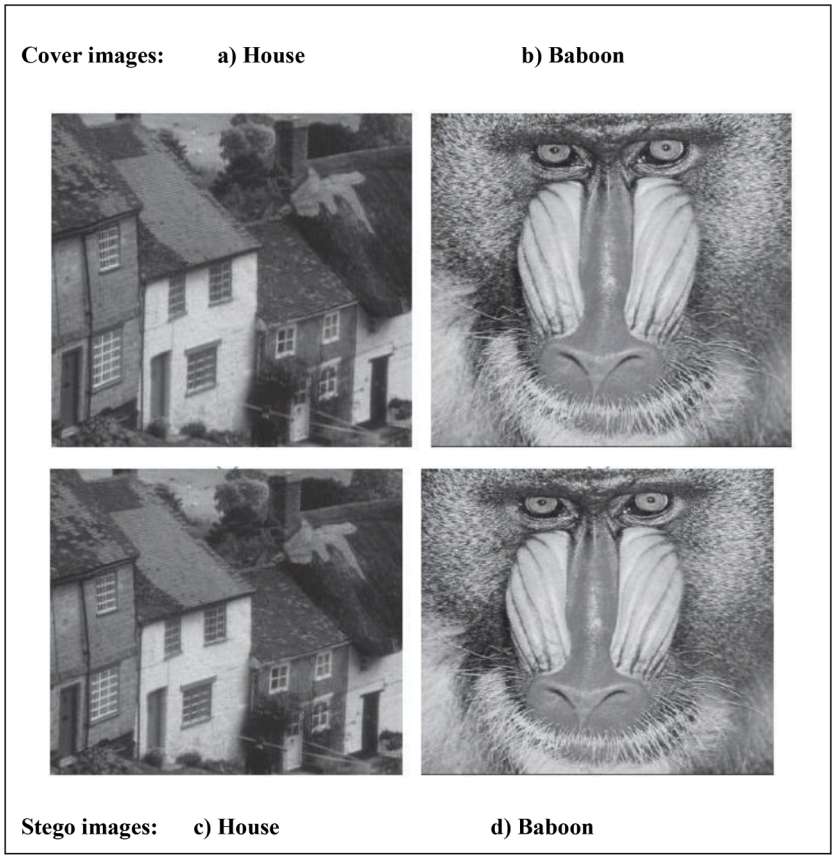


Fig. 8. Different cover and stego images

4. Experimental results and discussions

The proposed algorithm is executed using mat lab 2017a using USC-SIPI image data base set. The results are compared with the existing methods discussed in the literature. One bit per pixel is used for embedding and the results are tabulated based on parameters PSNR and SSIM. Table 1 list the PSNR values between proposed method and the existing method proposed by Pratik D. Shah and Rajankumar S. Bichkar [7] and [15] for different set of payload images and Payload image size of 512 × 256 table 2 describes PSNR vlues and SSIM between proposed method and existing method [13].

The root mean square value is a standard way of measuring the error of a model in predicting quantitative data. Formally it is defined as follows:

$$RMSE = \sqrt{\sum_{i=0}^{N} \frac{(y_{gi} - y_{oi})^2}{N}}$$

Where y_{gi} is the generated image, y_{oi} is the original image and i is the number of images.

SSIM is computed as in equation shown bellow, Where C1 and C2 are constants used to avoid instability when denominator terms become near to zero.

$$SSIM = \frac{(2\,\overline{x}\,\overline{y} + C_1)\,(2\sigma_{pq} + C_2)}{(\overline{x}^2 + \overline{y}^2 + C_1)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}$$

Results compared in Table 1 are obtained for the input cover images of size 256 × 256 and secret or payload image of size 64 × 128, results tabulated clearly explain PSNR value is significantly high and hence the difference between resulted stego image and cover image. Different payload images used are shown in Fig. 7.

Table 2 lists the PSNR and SSIM values of the existing method [13] and the proposed method considering input cover as 512 × 512 and payload as 256 × 256, which gives 1bpp embedding rate. Fig. 8 shows cover and payload images used for the testing purpose.

5. Conclusion

This chapter discussed image steganography approach using genetic algorithm for carrier image selection and embedding. The level of compatibility is measured between cover and payload image through several iterations among the cover image database called as crowd, by generating chromosomes. During the process of making genes strong, the algorithm generates new image which resembles the cover image selected and it is called as stego image, which consists of payload image embedded in it. The algorithm proposed is tested for 1bpp by making use of hundreds cover images and many secret images. The results obtained are tabulated and compared with the existing methods show the significant improvements in the results and hence proved with better performance. The main advantage of this method is in extraction process at the receiver. If the embedded payload data is damaged by the intruders during transmission, then there is all the possibility of retrieving the payload data using this algorithm. The algorithm compares the received stego image and if the extracted payload is damaged then the algorithm compares it with the predefined image in its database and reconstructs the complete payload image.

References

[1] Abbas Cheddad, Joan Condell, Kevin Curran, Paul Mc Kevitt, Digital image steganography: Survey and analysis of current methods, Signal Process. 90 (2010) 727–752.
[2] And G Bugár, V Bánoci, M Broda, D Levický, D Dupák, Data Hiding In Still Images Based On Blind Algorithm Of Steganography, in: 24th International Conference Radio elektronika (RADIOELEKTRONIKA), 2014, pp. 1–4.
[3] Kh Zaidoon, A.A. AL-Ani, B.B. Zaidan, Zaidan, Hamdan.O. Alanazi, Overview: main fundamentals for steganography, J. Comput. VOLUME 2 (ISSUE 3) (2010) 158–165, 0, ISSN 2151-9617.
[4] P. Shah, R. Bichkar, A secure spatial domain image steganography using genetic algorithm and linear congruential generator, in: International Conference on Intelligent Computing and Applications Advances in Intelligent Systems and Computing, 2018, pp. 119–129.
[5] P. Shah, R. Bichkar, Imperceptible steganography scheme with high payload capacity using genetic algorithm and particle swarm optimization, Int. J. Eng. Adv. Technol. 9 (1) (2019) 917–923. October.
[6] D. Goldberg, "Genetic Algorithms in Search Optimizations and Machine learning", Pearson Education India, 2006.
[7] Pratik D. Shah, Rajankumar S. Bichkar, Secret data modification based image steganography technique using genetic algorithm having a flexible chromosome structure", Eng. Sci. Technol. Int. J. (2021) 1–13.
[8] D Jude Hemanth, J Anitha, Daniela Elenab Popescu, Le Hoang Son, A modified genetic algorithm for performance improvement of transform based image steganography systems", Recent Advances in Machine Learning and Soft Computing, J. Intelligent Fuzzy Syst. 35 (1) (2018) 197–209.
[9] H. Kanan, N. Bahram, A novel image steganography scheme with high embedding capacity and tunable visual image quality based on a genetic algorithm, Expert Syst. Appl. 41 (14) (2014) 6123–6130.
[10] Tanmay Bhattacharya, Sandeep Bhowmik, S. Chaudhuri, A steganographic approach by using session based Stego-Key, genetic algorithm and variable bit replacement technique.", in: Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference on, IEEE, 2008.
[11] Lin-Yu Tseng, Yu-An-Ho Yung-kuan-chan, Yen-ping chu, Image hiding with an improved genetic algorithm and an optimal pixel adjustment process, in: Intelligent Systems Design and Applications, ISDA'08. Eighth International Conference on 3, IEEE, 2008.
[12] Masoumeh Khodaei, Karim Faez, Image hiding by using genetic algorithm and LSB substitution" ICISP'10, in: Proceedings of the 4th international conference on Image and signal processing/June Pages 44–411, 2010.
[13] P.D. Shah, R.S. Bichkar, Genetic algorithm based approach to select suitable cover image for image steganography, in: 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1–5. June.
[14] S. Pramanik, R.P. Singh, R. Ghosh, Application of bi-orthogonal wavelet transform and genetic algorithm in image steganography, Multimed. Tools Appl. 79 (2020) 17463–17482.
[15] X Gu, Y Sun, Image transformation and information hiding technology based on genetic algorithm, EURASIP J. Image Video Process. (2018) 1–10.
[16] K. Deb, S. Agrawal, A. Pratap, T Meyarivan, A fast elitist non-dominated sorting genetic algorithm for multi-objective optimisation": NSGA-II[C], in: International Conference on Parallel Problem Solving From Nature, Springer-Verlag, 2000, pp. 849–858.
[17] R. Biswas, S.K. Bandyapadhay, Random selection based GA optimization in 2D-DCT domain color image steganography, Multimed. Tools Appl. 79 (2020) 7101–7120.
[18] J. Yang, V. Honavar, Feature Subset Selection Using a Genetic Algoritltm, IEEE Intelligent Systems, 1998, pp. 44–49.
[19] C.F. Lin, R.Z. Wang, J.C. Lin, Image hiding by optimal lsb substitution and genetic algorithm, Pattern Recognit. 34 (2001) 671–683.
[20] Wu M N., Lin MH., Chang CC, "A LSB substitution oriented image hiding strategy using genetic algorithms". In: Chi CH., Lam KY. (eds) Content Computing. AWCC 2004. Lecture Notes in Computer Science, vol 3309. Springer, Berlin, Heidelberg.
[21] Mahdi Ramezani, Shahrokh Ghaemmaghami, Adaptive image steganography with Mod-4 embedding using image contrast, in: proceedings of the 7th IEEE Conference on Consumer Communications and Networking Conference, 2010, pp. 243–246.
[22] M. Ramezani, S. Ghaemmaghami, Towards genetic feature selection in image steganalysis, in: 6th IEEE International Workshop on Digital Rights Management, Las Vegas, USA, 2010.
[23] Yang Shen Bian, Xiamu Niu, A secure steganography method based on genetic algorithm, J. Inf. Hiding Multimed. Signal Process. c 2010 1 (1) (2010). January ISSN 2073-4212.
[24] J.K. Mandal, A. Khamrui, A Data Embedding Technique for Gray scale Image Using Genetic Algorithm (DEGGA), in: International Confrence on Electronic Systems (ICES-2011), 2021.
[25] M. Nosrati, R. Karimi, A survey on usage of genetic algorithms in recent steganography researches, World Appl. Program. 2 (3) (2012) 206–210.
[26] M. Nosrati, A. Hanani, R. Karimi, Steganography in image segments using genetic algorithm, in: 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, 2015, pp. 102–107.
[27] P. Sethi, V. Kapoor, A proposed novel architecture for information hiding in image steganography by using genetic algorithm and cryptography, in: Proc. Computer. Sci, 2016, pp. 61–66.
[28] xxx "University of Wisconsin–Madison public domain image database", 2021 [online] Available: https://homepages.cae.wisc.edu/~ece533/images/.
[29] xxx "USC-SIPI Image Database", 2021 [online] Available: http://sipi.usc.edu/database/.

Dynamic Pythagorean fuzzy probabilistic linguistic TOPSIS method with psychological preference and its application for COVID-19 vaccination

Soumya Mishra, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, soumyamishra96@gmail.com*

Malaya Tripathy, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, malaya.tripathy43@gmail.com*

Sushree Sangita Jena, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sushreesangita665.com*

Ipsita Samal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ipsitasamal55@gmail.com*

Keywords:
Probabilistic linguistic term set (PLTS)
Pythagorean fuzzy set (PFS)
Pythagorean fuzzy probabilistic linguistic term set (PFPLTS)
TOPSIS with psychological distance (Psy-TOPSIS)
Dynamic multicriteria group decision making
COVID-19 vaccination center

A B S T R A C T

The probabilistic linguistic term set (PLTS) has been widely used in multiple criteria group decision making (MCGDM) problems where the linguistic information is uncertain and hesitant. To reflect the different preferences and uncertainties, we propose a new PLTS with probability in the form of Pythagorean fuzzy set (PFS), called Pythagorean fuzzy probabilistic linguistic term set (PFPLTS). In addition, considering the information integrity, uncertainty and DMs' preferences, the operation and aggregation operators for PFPLTS are introduced. Then, the weight method based on minimum deviation and dual ideal point-vector projection is proposed, which considers the time-varying characteristics of the weights and combines multi-dimensional influencing factors. Next, the psychological distance measure is proposed by dividing the psychological space into multiple vectors. Based on the proposed dynamic weight method, three psychological distance measures and TOPSIS method, we develop a dynamic Pythagorean fuzzy probabilistic linguistic TOPSIS method with psychological distance (Psy-TOPSIS), the psychological index ranges from 1 to 40. Finally, a practical case, site selecting of COVID-19 vaccination center, is given and compared with three approaches to illustrate the effectiveness and practicality of PFPLTS and the proposed decision-making method.

1. Introduction

Multiple criteria group decision making (MCGDM) is an essential branch in the field of decision making. It refers to the way decision-makers (DMs) apply decision-making methods to select the best alternative with multiple criteria. With the increasing complexity of decision-making problems and their backgrounds, in many cases, DMs cannot accurately quantify the evaluation objects and can only use natural language to evaluate. For example, in the site selection of the COVID-19 vaccination center, DMs may say that "the traffic conditions in this location are not bad, but the transportation and transformation cost is too high." In this context, the words "not bad" and "too high" are described in linguistic terms. Although linguistic terms are intuitive, flexible, and close to people's cognition, they are difficult to calculate. To deal with this problem, Zadeh [1] proposed the fuzzy linguistic approach (FLA), which can aptly describe the fuzziness and uncertainty of information. To further reflect the uncertainty and the preference degree of linguistic terms, Pang [2] proposed the concept of PLTS. Since it was submitted, PLTS has been widely used in medical level assessment [3], supplier selection [4], venture capital [5], and so on. On the other hand, some contributions have been made to the conversion, computation, and aggregation method of PLTS [6–15]. In the follow-up research, many scholars have carried on the related expansion to PLTS. For example, multiple linguistic terms are utilized in the probabilistic uncertain linguistic term set (PULTS) to express the hesitation of evaluation information [16]. The Interval-valued probabilistic linguistic term set (IVPLTS) extends the corresponding probability of linguistic terms to interval values [17]. By adding the unknown probability to linguistic term, the uncertain probabilistic linguistic term set (UPLTS) can be established [18], etc. It is evident that only membership degree is used to describe the importance of linguistic terms in the existing probabilistic linguistic terms, which ignores the uncertainty and preference of DMs about a particular linguistic evaluation.

Dynamic decision-making is a process that changes with time. Regarding the time pressure and affairs unpredictability at different development stages, DMs must be more cautious and effective in dealing with the fuzziness and uncertainty of information. Additionally, it also means that decision information and criteria weights are affected by time. Therefore, it is necessary to determine the time weight and criteria weights of each stage. At present, there are many methods for weight

determination, mainly divided into three categories: subjective weighting method, objective weighting method, and combination weighting method. Subjective weighting methods mainly include BEM [19], AHP method [20], Delphi method [21], etc., which are obtained by the personal judgment of experts' experience and generally not affected by the attribute value. The advantage is that experts can determine the weights based on actual problems and their knowledge, and there will be no situation that contradicts the actual importance. However, the decision-making or evaluation results are subjective and vulnerable to the lack of decision-makers' knowledge. Because of the poor objectivity, it has great limitations in application. The objective weights mainly include the dispersion maximization method [22], entropy weight method [23], etc., calculated by attribute values. This is usually based on sound mathematical theories and techniques, so the weights are highly reasonable, and the method has a solid mathematical theoretical basis. However, this kind of empowerment method cannot reflect the preferences of DMs, and there may be situations that are contrary to the actual importance.

A combination weighting method was proposed to take into account the advantages of subjective and objective weighting methods. The most common is the linear weighted combination approach, and our personal perception often decides the combination coefficients. But in general, the mathematical theoretical basis of the subjective and objective integrated weighting method is relatively perfect, but the disadvantage lies in the high complexity of the algorithm. At the same time, many criteria weights methods are based on the needs of one or two sides of the DMs, which rarely contain the needs of multiple sides together. In addition, most of the existing weight methods are in a static environment. The criteria weights do not consider time-varying factors, so they cannot reflect the characteristics of dynamic decision-making.

The TOPSIS method is commonly used in MCGDM problems. It has been widely studied and applied, but it is rarely used in dynamic decision-making environments, and does not take into account the psychological changes of decision-makers. We propose a probabilistic linguistic term set that includes membership and non-membership of linguistic terms to overcome the above shortcomings. And on this basis, we have developed a new dynamic multi-criteria weight method that considers time-varying effects and a decision-making method that considers psychological changes. The main contributions of this paper can be summarized as follows:

1 Due to the advantages of PLTS in describing information, and its ignorance in linguistic preference and uncertainty, the corresponding probability is extended to a set that includes both the degree of membership and non-membership. Using the Pythagorean Fuzzy Set (PFS) to appropriately reduce the constraints on the set, we define the term set of Pythagorean probabilistic linguistic term set (PFPLTS). In addition, the basic operation and aggregation operators of PFPLTS are introduced.
2 In the existing criteria weights methods, DMs need to choose one or two angles for calculation, such as hesitation degree, identification degree, and so on. It is rare to consider the influence of multiple angles on the criteria weights, which leads to the inaccuracy of the fusion weight. Therefore, it is necessary to study the weight fusion method from multiple dimensions. In addition, the current weight methods are mostly single-stage and static [24,25], so they can not reflect the DMs' changing information and preferences. The subjective and objective time weight method of minimum deviation linear programming and a weight fusion method of dual ideal point-vector projection in a dynamic environment is proposed to solve these two problems.
3 The DMs' preferences for different alternatives and criteria will change accordingly with the decision-making background and environment. Therefore, we consider the multi-dimensional psychological space of DMs [26], and propose a dynamic psychological distance measurement that includes the psychological changes of DMs.
4 As a widely used multi-criteria decision-making method, the TOPSIS method has attracted extensive research from scholars since it was proposed. Based on the TOPSIS method and the aforementioned methods, a dynamic Pythagorean fuzzy probabilistic linguistic TOPSIS method with psychological preference is constructed and applied to the site selecting of the COVID-19 vaccination center.

The rest of this paper is organized as follows: In Section 2, some basic concepts about PLTS and PFS are briefly reviewed. Section 3 defines the concept of PFPLTS and proposes the related calculation method, operation and aggregation operators of PFPLTS. Section 4 constructs the time and criteria weights method with minimum deviation linear programming and dual ideal-point vector projection, respectively. Section 5 establishes the novel TOPSIS method with psychological distance measure and constructs an approach for MCGDM based on the Pythagorean fuzzy probabilistic linguistic psy-TOPSIS method. Section 6 gives a case study, site selection for the COVID-19 vaccination center, to illustrate the applicability and practicability of the proposed method. Finally, some conclusions are given in Section 7.

2. Preliminaries

In this section, we will briefly review some basic concepts related to PLTS and PFS.

2.1. Probabilistic linguistic term set

Definition 1. [2] Let $S = \{s_\alpha | \alpha = 0, 1, 2, \ldots, \tau\}$ be a linguistic term set, then the probabilistic linguistic term set (PLTS) can be defined as $L(p) = \{L^{(k)}(p^{(k)}) | L^{(k)} \in S, p^{(k)} \geq 0, k = 1, 2, \ldots, \#L(p), \sum_{k=1}^{\#L(p)} p^{(k)} \leq 1\}$. Where $L^{(k)}(p^{(k)})$ is the linguistic term $L^{(k)}$ associated with its corresponding probability $p^{(k)}$, and $\#L(p)$ is the cardinality of $L(p)$.

Definition 2. [27] Let $L(p) = \{L^{(k)}(p^{(k)}) | L^{(k)} \in S, p^{(k)} \geq 0, k = 1, 2, \ldots, \#L(p), \sum_{k=1}^{\#L(p)} p^{(k)} \leq 1\}$, $S = \{s_\alpha | \alpha = 0, 1, 2, \ldots, \tau\}$, and $\alpha^{(k)}$ is the subscript of linguistic term $L^{(k)}$. The score function and its inverse function are defined as:

$$g : [0, \tau] \to [0, 1], g(L(p)) = \left\{ \left[\frac{\alpha^{(k)}}{\tau} \right] (p^{(k)}) \right\} = L_\varpi(p), \varpi \in [0, 1],$$
$$g^{-1} : [0, 1] \to [0, \tau], g(L_\gamma(p)) = \{ s_{\tau\varpi}(p^{(k)}) \} = L(p)), \varpi \in [0, 1]. \quad (1)$$

Definition 3. [27] Let $L_1(p) = \{L_1{}^{(k)}(p_1{}^{(k)}) | k = 1, 2, \ldots, \#L_1(p)\}$, $L_2(p) = \{L_2{}^{(k)}(p_2{}^{(k)}) | k = 1, 2, \ldots, \#L_2(p)\}$ and $L_3(p) = \{L_3{}^{(k)}(p_3{}^{(k)}) | k = 1, 2, \ldots, \#L_3(p)\}$ be three finite and ordered PLTSs. λ is a positive real number, $\gamma_1^{(k)} \in g(L_1)$, $\gamma_2^{(l)} \in g(L_2)$, $\gamma_3^{(r)} \in g(L_3)$ and $k = 1, 2, \ldots, \#L_1(p)$, $l = 1, 2, \ldots, \#L_2(p)$, $r = 1, 2, \ldots, \#L_3(p)$.

1 $L_1(p) \oplus L_2(p) = g^{-1}(\cup_{\gamma_1^{(k)} \in g(L_1), \gamma_2^{(l)} \in g(L_2)} \{(\gamma_1^{(k)} + \gamma_2^{(l)})(p_1^{(k)} p_2^{(l)})\})$.

2 $L_1(p) \ominus L_2(p) = g^{-1}(\cup_{\gamma_1^{(k)} \in g(L_1), \gamma_2^{(l)} \in g(L_2)} \{\xi(p_1^{(k)} p_2^{(l)})\})$.

Where $\xi = \begin{cases} \dfrac{\gamma_1^{(k)} - \gamma_2^{(l)}}{1 - \gamma_2^{(l)}}, if \gamma_1^{(k)} \geq \gamma_2^{(l)} and \gamma_2^{(l)} \neq 1 \\ 0, otherwise \end{cases}$.

1 $L_1(p) \otimes L_2(p) = g^{-1}(\cup_{\gamma_1^{(k)} \in g(L_1), \gamma_2^{(l)} \in g(L_2)} \{(\gamma_1^{(k)} \gamma_2^{(l)})(p_1^{(k)} p_2^{(l)})\})$.

2 $L_1(p) \oslash L_2(p) = g^{-1}(\cup_{\gamma_1^{(k)} \in g(L_1), \gamma_2^{(l)} \in g(L_2)} \{\zeta(p_1^{(k)} p_2^{(l)})\})$.

$$\text{Where } \zeta = \begin{cases} \gamma_1^{(k)}, if \gamma_1^{(k)} \le \gamma_2^{(l)} and \gamma_2^{(l)} \ne 0 \\ \gamma_2^{(l)} \\ 0, otherwise \end{cases}.$$

1 $\lambda L_3(p) = g^{-1}(\cup_{\gamma_3^{(r)} \in g(L_3)} \{(1 - (1 - \gamma_3^{(r)})^{\lambda})(p_1^{(r)})\})$.

2 $L_3{}^{\lambda}(p) = g^{-1}(\cup_{\gamma_3^{(r)} \in g(L_3)} \{(\gamma_3^{(r)})^{\lambda}(p_1^{(r)})\})$.

3 $L_3{}^{-1}(p) = g^{-1}(\cup_{\gamma_3^{(r)} \in g(L_3)} \{(1 - \gamma_3^{(r)})(p_1^{(r)})\})$.

And the distance measure is defined as:

$$d(L_1(p), L_2(p)) = \frac{1}{2}\left(\sum_{k=1}^{\#L_1(p)} g\left(L_1{}^{(k)}\right)\left(p_1{}^{(k)}\right) - \sum_{k=1}^{\#PL_1(p)} g\left(L_2{}^{(k)}\right)\left(p_2{}^{(k)}\right) \right)$$

2.2. Pythagorean fuzzy set

Definition 4. [28] Let X be the universe of discourse, the Pythagorean fuzzy set is defined as

$$P = \left\{ < x, \mu_p(x), \nu_p(x) > \middle| x \in X, 0 \le \mu_p{}^2(x) + \nu_p{}^2(x) \le 1 \right\}$$

Where $\mu_p(x) : X \to [0,1]$ and $\nu_p(x) : X \to [0,1]$ are the degree of membership and non-membership of x belonging to P respectively. For any $x \in X$, the hesitation of x belonging to P is $\pi_p(x) = \sqrt{1 - \mu_p{}^2(x) - \nu_p{}^2(x)}$. For convenience, $< \mu_p, \nu_p >$ is called as Pythagorean fuzzy number (PFN), where $\mu_p, \nu_p \in [0,1]$, $\mu_p{}^2 + \nu_p{}^2 \le 1$, it is simply recorded as $P = < \mu_p, \nu_p >$.

3. Pythagoras-probabilistic linguistic term set

In order to describe the fuzziness and uncertainty of DMs, we introduce a new concept called PFPLTS. Then, the related comparison method, basic operation and aggregation operators are proposed.

3.1. The concept and comparation method of PFPLTS

By adding corresponding probability values to linguistic terms, PLTS has significantly progressed in describing uncertain information of hesitant fuzzy linguistic evaluation and comparative preference. But the fuzziness and uncertainty of linguistic terms have not been specifically expressed. With the increase of complexity and uncertainty of decision-making problems, the fuzziness of DMs' thinking, and the limitation of knowledge reserve, the probability value of hesitant linguistic evaluation is not completely certain. PFS introduced by Yager [28,29] on the basis of IFS can make the description of decision information more scientific and effective, because it contains membership and non-membership degrees whose square sum is not more than 1. For the fuzziness and uncertainty of PLTS and Pythagorean fuzzy sets in describing information, the Pythagorean fuzzy probabilistic linguistic term set (PFPLTS) is proposed.

Definition 5. Let X be a non-empty universe of discourse, $S = \{s_\alpha | \alpha = 0, 1, 2, \ldots, \tau\}$ is the linguistic term set, a PFPLTS can be defined as follows:

$$PL(p) = \left\{ \left[x, PL^{(k)}\left(\widetilde{p}^{(k)}\right) \right] \middle| x \in X, PL^{(k)} \in S, k = 1, 2, \ldots, \#PL(p) \right\}$$
$$= \left\{ \left[x, PL^{(k)} \langle \mu^{(k)}, \nu^{(k)} \rangle \right] \middle| x \in X, PL^{(k)} \in S, k = 1, 2, \ldots, \#PL(p) \right\}$$

where $PL^{(k)}(\widetilde{p}^{(k)})$ represents the linguistic term $PL^{(k)}$ associated with its uncertain PFS probability $\widetilde{p}^{(k)}$, and $\widetilde{p}^{(k)} = \langle \mu^{(k)}, \nu^{(k)} \rangle$, in which $\mu^{(k)}$ and $\nu^{(k)}$ represent the membership and non-membership degrees of linguistic term $PL^{(k)}$, $\alpha^{(k)}$ is the subscript of linguistic term $PL^{(k)}$, and $\#PL(p)$ is the cardinal number of $PL(p)$. The hesitation degree can be calculated by $\pi^{(k)} = \sqrt{1 - (\mu^{(k)})^2 - (\nu^{(k)})^2}$. We have noticed that the PFPLTS is an extension of the PLTS, when $PL^{(k)} = \langle \mu^{(k)} \rangle$, the PFPLTS degenerates into PLTS.

To compare two different PFPLTSs, the score and accuracy function are given in the following, and then the comparison method is developed.

Definition 6. Let X be a non-empty universe of discourse, and the score function of PFPLTSs on X can be defined as:

$$S(PL(p)) = \sum_{k=1}^{\#PL(p)} g\left(PL^{(k)}\right) \times \left[\left(\mu^{(k)}\right)^2 - \left(\nu^{(k)}\right)^2 \right] \tag{2}$$

The accuracy function of PFPLTSs on X can be defined as:

$$H(PL(p)) = \sum_{k=1}^{\#PL(p)} g\left(PL^{(k)}\right) \times \left[\left(\mu^{(k)}\right)^2 + \left(\nu^{(k)}\right)^2 \right] \tag{3}$$

Definition 7. Let X be a non-empty universe of discourse, for any two PFPLTSs $PL_1(p)$ and $PL_2(p)$.

(1) If $S(PL_1(p)) > S(PL_2(p))$, then $PL_1(p) \succ PL_2(p)$.
(2) If $S(PL_1(p)) < S(PL_2(p))$, then $PL_1(p) \prec PL_2(p)$.
(3) If $S(PL_1(p)) = S(PL_2(p))$, then:
   a If $H(PL_1(p)) > H(PL_2(p))$, then $PL_1(p) \succ PL_2(p)$.
   b If $H(PL_1(p)) < H(PL_2(p))$, then $PL_1(p) \prec PL_2(p)$.
   c If $H(PL_1(p)) = H(PL_2(p))$, then $PL_1(p) \sim PL_2(p)$.

Definition 8. Let X be a non-empty universe of discourse, give a PFPLTS $PL(p) = \{[x, PL^{(k)} \langle \mu^{(k)}, \nu^{(k)} \rangle] | x \in X, PL^{(k)} \in S, k = 1, 2, \ldots, \#PL(p)\}$, and $\alpha^{(k)}$ is the subscript of linguistic term $PL^{(k)}$, $PL(p)$ is called an ordered PFPLTS, if the elements (PFPLEs) $PL^{(k)}(\widetilde{p}^{(k)})$ in PFPLTS are sorted by the values of $S(PL^{(k)}(\widetilde{p}^{(k)}))(k = 1, 2, \ldots, \#PL(P))$ in ascending order.

3.2. Some basic operation for PFPLTS

With the introduction of PFPLTS, it is crucial to find the basic operation. Assume that all PFPLTSs are ordered and finite, then some basic operations are proposed as bellow:

Definition 9. Let $PL_1(p) = \{[x, PL_1{}^{(k)} \langle \mu_1{}^{(k)}, \nu_1{}^{(k)} \rangle] | x \in X, k = 1, 2, \ldots, \#PL_1(p)\}$ and $PL_2(p) = \{[x, PL_2{}^{(l)} \langle \mu_2{}^{(l)}, \nu_2{}^{(l)} \rangle] | x \in X, l = 1, 2, \ldots, \#PL_2(p)\}$ be any two PFPLTSs, then the distance measure is as follows:

$$d(PL_1(p), PL_2(p)) = \frac{1}{2}\left(\left| \sum_{k=1}^{\#PL_1(p)} g\left(PL_1{}^{(k)}\right)\left(\mu_1{}^{(k)}\right)^2 - \sum_{k=1}^{\#PL_1(p)} g\left(PL_2{}^{(k)}\right)\left(\mu_2{}^{(k)}\right)^2 \right| \right.$$
$$\left. + \left| \sum_{k=1}^{\#PL_1(p)} g\left(PL_1{}^{(k)}\right)\left(\nu_1{}^{(k)}\right)^2 - \sum_{k=1}^{\#PL_1(p)} g\left(PL_2{}^{(k)}\right)\left(\nu_2{}^{(k)}\right)^2 \right| \right) \tag{4}$$

Definition 10. Let $PL_1(p)=\{[x,PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle]|x\in X,k=1,2,\ldots,\#PL_1(p)\}$ and $PL_2(p)=\{[x,PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle]|x\in X,l=1,2,\ldots,\#PL_2(p)\}$ be any two PFPLTSs, then some basic operations are defined as follows:

(1) $PL_1(p)\oplus PL_2(p)=$
$$g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\\ \{(\gamma_1{}^{(k)}+\gamma_2{}^{(l)}-\gamma_1{}^{(k)}\gamma_2{}^{(l)})\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}\end{array}\right).$$

(2) $PL_1(p)\ominus PL_2(p)=g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\\ \{\xi\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}\end{array}\right)$,

where $\xi=\begin{cases}\dfrac{\gamma_1{}^{(k)}-\gamma_2{}^{(l)}}{1-\gamma_2{}^{(l)}}, if\gamma_1{}^{(k)}>\gamma_2{}^{(l)},\gamma_2{}^{(l)}\neq 0\\ 0, otherwise\end{cases}$.

PFPLTSs, $\lambda,\lambda_1,\lambda_2\geq 0$. *Then:*

(1) $PL_1(p)\oplus PL_2(p)=PL_2(p)\oplus PL_1(p)$.
(2) $(PL_1(p)\oplus PL_2(p))\oplus PL_3(p)=PL_1(p)\oplus(PL_2(p)\oplus PL_3(p))$.
(3) $\lambda(PL_1(p)\oplus PL_2(p))=\lambda PL_2(p)\oplus\lambda PL_1(p)$.
(4) $(\lambda_1+\lambda_2)PL_1(p)=\lambda_1PL_1(p)\oplus\lambda_2PL_1(p)$.
(5) $PL_1(p)\otimes PL_2(p)=PL_2(p)\otimes PL_1(p)$.
(6) $(PL_1(p)\otimes PL_2(p))\otimes PL_3(p)=PL_1(p)\otimes(PL_2(p)\otimes PL_3(p))$.
(7) $(PL_1(p)\otimes PL_2(p))^{\lambda}=(PL_2(p))^{\lambda}\otimes(PL_1(p))^{\lambda}$.
(8) $(PL_1(p))^{(\lambda_1+\lambda_2)}=(PL_1(p))^{\lambda_1}\otimes(PL_1(p))^{\lambda_2}$.
(9) $\lambda(PL_1(p)\ominus PL_2(p))=\lambda PL_1(p)\ominus\lambda PL_1(p)$.
(10) $\lambda_1PL_1(p)\ominus\lambda_2PL_1(p))=(\lambda_1-\lambda_2)PL_1(p)$.
(11) $(PL_1(p)\oslash PL_2(p))^{\lambda}=(PL_1(p))^{\lambda}\oslash(PL_2(p))^{\lambda}$.
(12) $(PL_1(p))^{\lambda_1}\oslash(PL_1(p))^{\lambda_2}=(PL_1(p))^{\lambda_1}\oslash(PL_1(p))^{\lambda_2}$

Proof

$(1)PL_1(p)\oplus PL_2(p)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\{(\gamma_1{}^{(k)}+\gamma_2{}^{(l)}-\gamma_1{}^{(k)}\gamma_2{}^{(l)})\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}\right)$
$=g^{-1}\left(\cup_{PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p),PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\{(\gamma_2{}^{(l)}+\gamma_1{}^{(k)}-\gamma_2{}^{(l)}\gamma_1{}^{(k)})\langle\mu_2{}^{(l)}\mu_1{}^{(k)},\nu_2{}^{(l)}\nu_1{}^{(k)}\rangle\}\right)$
$=PL_2(p)\oplus PL_1(p)$

$(2)(PL_1(p)\oplus PL_2(p))\oplus PL_3(p)$
$=g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\\ \{(\gamma_1{}^{(k)}+\gamma_2{}^{(l)}-\gamma_1{}^{(k)}\gamma_2{}^{(l)})\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}+\cup_{PL_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle\in PL_3(p)}\{\gamma_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle\}\end{array}\right)$
$=g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p),PL_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle\in PL_3(p)}\\ \{(\gamma_1{}^{(k)}+\gamma_2{}^{(l)}+\gamma_3{}^{(r)}-\gamma_1{}^{(k)}\gamma_2{}^{(l)}-\gamma_1{}^{(k)}\gamma_3{}^{(r)}-\gamma_2{}^{(l)}\gamma_3{}^{(r)}+\gamma_1{}^{(k)}\gamma_2{}^{(l)}\gamma_3{}^{(r)})\langle\mu_1{}^{(k)}\mu_2{}^{(l)}\mu_3{}^{(r)},\nu_1{}^{(k)}\nu_2{}^{(l)}\nu_3{}^{(r)}\rangle\}\end{array}\right)$
$=g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\{\gamma_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\}\\ +\cup_{PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p),PL_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle\in PL_3(p)}\{(\gamma_2{}^{(l)}+\gamma_3{}^{(r)}-\gamma_2{}^{(l)}\gamma_3{}^{(r)})\langle\mu_2{}^{(l)}\mu_3{}^{(r)},\nu_2{}^{(l)}\nu_3{}^{(r)}\rangle\}\end{array}\right)$
$=PL_1(p)\oplus(PL_2(p)\oplus PL_3(p))$.

$(4)(\lambda_1+\lambda_2)PL_1(p)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\left\{\left(1-(1-\gamma_1{}^{(k)})^{\lambda_1+\lambda_2}\right)\langle\mu^{(k)},\nu^{(k)}\rangle\right\}\right)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\left\{\left(1-(1-\gamma_1{}^{(k)})^{\lambda_1}\right)+\left(1-(1-\gamma_1{}^{(k)})^{\lambda_2}\right)-\left(1-(1-\gamma_1{}^{(k)})^{\lambda_1}\right)\left(1-(1-\gamma_1{}^{(k)})^{\lambda_2}\right)\langle\mu^{(k)},\nu^{(k)}\rangle\right\}\right)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\left\{\left(1-(1-\gamma_1{}^{(k)})^{\lambda_1}\right)\langle\mu^{(k)},\nu^{(k)}\rangle\right\}\right)\oplus g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\left\{\left(1-(1-\gamma_1{}^{(k)})^{\lambda_2}\right)\langle\mu^{(k)},\nu^{(k)}\rangle\right\}\right)$
$=\lambda_1PL_1(p)\oplus\lambda_2PL_1(p)$.

---

(3) $PL_1(p)\otimes PL_2(p)=g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\\ \{\gamma_1{}^{(k)}\gamma_2{}^{(l)}\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}\end{array}\right)$.

(4) $PL_1(p)\oslash PL_2(p)=g^{-1}\left(\begin{array}{l}\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(k)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\\ \{\zeta\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}\end{array}\right)$,

where $\zeta=\begin{cases}\dfrac{\gamma_1{}^{(k)}}{\gamma_2{}^{(l)}}, if\gamma_1{}^{(k)}\leq\gamma_2{}^{(l)},\gamma_2{}^{(l)}\neq 0\\ 0, otherwise\end{cases}$.

(5) $\lambda PL_1(p)=g^{-1}(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\{(1-(1-\gamma_1{}^{(k)})^{\lambda}\langle\mu^{(k)},\nu^{(k)}\rangle\})$.
(6) $(PL_1(p))^{\lambda}=g^{-1}(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\{(\gamma_1{}^{(k)})^{\lambda}\langle\mu^{(k)},\nu^{(k)}\rangle\}),\lambda\geq 0$.
(7) $(PL_1(p))^{-1}=g^{-1}(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p)}\{(1-\gamma_1{}^{(k)})\langle\mu^{(k)},\nu^{(k)}\rangle\})$.

Where $\gamma_1{}^{(k)}\in g(PL_1(p))$, $\gamma_2{}^{(l)}\in g(PL_2(p))$, $g(\cdot)$ is the score function. $\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle$ and $\langle\mu_2{}^{(k)},\nu_2{}^{(k)}\rangle$ are the uncertain probability values of hesitant linguistic evaluations, which are in the form of PFS.

Theorem 1. *Let* $PL_1(p)=\{[x,PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle]|x\in X,k=1,2,\ldots,\#PL_1(p)\}$, $PL_2(p)=\{[x,PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle]|x\in X,l=1,2,\ldots,\#PL_2(p)\}$ *and* $PL_3(p)=\{[x,PL_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle]|x\in X,r=1,2,\ldots,\#PL_3(p)\}$ *be any three*

$(5)PL_1(p)\otimes PL_2(p)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p)}\{\gamma_1{}^{(k)}\gamma_2{}^{(l)}\langle\mu_1{}^{(k)}\mu_2{}^{(l)},\nu_1{}^{(k)}\nu_2{}^{(l)}\rangle\}\right)$
$=PL_2(p)\otimes PL_1(p)$.

$(6)(PL_1(p)\otimes PL_2(p))\otimes PL_3(p)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p),PL_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle\in PL_3(p)}\right.$
$\left.\{\gamma_1{}^{(k)}\gamma_2{}^{(l)}\gamma_3{}^{(r)}\langle\mu_1{}^{(k)}\mu_2{}^{(l)}\mu_3{}^{(r)},\nu_1{}^{(k)}\nu_2{}^{(l)}\nu_3{}^{(r)}\rangle\}\right)=PL_1(p)\otimes(PL_2(p)\otimes PL_3(p))$.

$(6)(PL_1(p)\otimes PL_2(p))\otimes PL_3(p)$
$=g^{-1}\left(\cup_{PL_1{}^{(k)}\langle\mu_1{}^{(k)},\nu_1{}^{(k)}\rangle\in PL_1(p),PL_2{}^{(l)}\langle\mu_2{}^{(l)},\nu_2{}^{(l)}\rangle\in PL_2(p),PL_3{}^{(r)}\langle\mu_3{}^{(r)},\nu_3{}^{(r)}\rangle\in PL_3(p)}\right.$
$\left.\{\gamma_1{}^{(k)}\gamma_2{}^{(l)}\gamma_3{}^{(r)}\langle\mu_1{}^{(k)}\mu_2{}^{(l)}\mu_3{}^{(r)},\nu_1{}^{(k)}\nu_2{}^{(l)}\nu_3{}^{(r)}\rangle\}\right)=PL_1(p)\otimes(PL_2(p)\otimes PL_3(p))$.

$$(8)(PL_1(p))^{(\lambda_1+\lambda_2)}$$
$$= g^{-1}\Big(\cup PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p)\Big\{ \big(\gamma_1^{(k)}\big)^{(\lambda_1+\lambda_2)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \Big\}\Big) = (PL_1(p))^{\lambda_1} \otimes (PL_1(p))^{\lambda_2}$$

$$(8)(PL_1(p))^{(\lambda_1+\lambda_2)}$$
$$= g^{-1}\Big(\cup PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p)\Big\{ \big(\gamma_1^{(k)}\big)^{(\lambda_1+\lambda_2)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \Big\}\Big) = (PL_1(p))^{\lambda_1} \otimes (PL_1(p))^{\lambda_2}$$

$$(11)(PL_1(p) \oslash PL_2(p))^{\lambda}$$
$$= g^{-1}\left(\cup_{PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p), PL_2^{(l)}\langle \mu_2^{(k)}, \nu_2^{(l)}\rangle \in PL_2(p)} \left\{ \left(\frac{\gamma_1^{(k)}}{\gamma_2^{(l)}}\right)^{\lambda} \langle \mu_1^{(k)}\mu_2^{(l)}, \nu_1^{(k)}\nu_2^{(l)}\rangle \right\}\right)$$
$$= g^{-1}\left(\cup_{PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p), PL_2^{(l)}\langle \mu_2^{(k)}, \nu_2^{(l)}\rangle \in PL_2(p)} \left\{ \frac{\big(\gamma_1^{(k)}\big)^{\lambda}}{\big(\gamma_2^{(l)}\big)^{\lambda}} \langle \mu_1^{(k)}\mu_2^{(l)}, \nu_1^{(k)}\nu_2^{(l)}\rangle \right\}\right)$$
$$= (PL_1(p))^{\lambda} \oslash (PL_1(p))^{\lambda}, when \gamma_1^{(k)} \leq \gamma_2^{(l)}, \gamma_2^{(l)} \neq 0.$$

$$(12)(PL_1(p))^{\lambda_1} \oslash (PL_1(p))^{\lambda_2}$$
$$= g^{-1}\Big(\cup_{PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p)} \Big\{ \big(\gamma_1^{(k)}\big)^{\lambda_1} \langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \Big\}\Big) \oslash g^{-1}\Big(\cup_{PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p)} \Big\{ \big(\gamma_1^{(k)}\big)^{\lambda_2} \langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \Big\}\Big)$$
$$= g^{-1}\Big(\cup_{PL_1^{(k)}\langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \in PL_1(p)} \Big\{ \big(\gamma_1^{(k)}\big)^{\lambda_1-\lambda_2} \langle \mu_1^{(k)}, \nu_1^{(k)}\rangle \Big\}\Big)$$
$$= (PL_1(p))^{\lambda_1} \oslash (PL_1(p))^{\lambda_2}, when \gamma_1^{(k)} \leq \gamma_2^{(l)}, \gamma_2^{(l)} \neq 0.$$

3.3. The aggregation operators for PFPLTS

In order to make better use of PFPLTS in decision-making problems, some aggregation operators are proposed in this subsection.

Definition 11. Let $PL_i(p) = \{[x, PL_i^{(k)}\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle]|x \in X, PL_i^{(k)} \in S, k = 1, 2, \ldots, \#PL_i(p)\}$, $(i = 1, 2, \ldots, n)$, where $PL_i^{(k)}$ is the kth linguistic term, and $\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle$ is the corresponding uncertain probability. Then the Pythagorean fuzzy probabilistic linguistic average (PFPLA) operator is defined as:

$$PFPLA(PL_1(p), PL_2(p), \ldots, PL_n(p)) = \frac{1}{n}(PL_1(p) \oplus PL_2(p) \oplus \cdots \oplus PL_n(p)) \tag{5}$$

Definition 12. Let $PL_i(p) = \{[x, PL_i^{(k)}\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle]|x \in X, PL_i^{(k)} \in S, k = 1, 2, \ldots, \#PL_i(p)\}$, $(i = 1, 2, \ldots, n)$, where $PL_i^{(k)}$ is the kth linguistic term, and $\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle$ is the corresponding uncertain probability. Then the Pythagorean fuzzy probabilistic linguistic weighted average (PFPWLA) operator is defined as:

$$PFPLWA(PL_1(p), PL_2(p), \ldots, PL_n(p)) = \omega_1 PL_1(p) \oplus \omega_2 PL_2(p) \oplus \cdots \oplus \omega_n PL_n(p) \tag{6}$$

Where $\omega = (\omega_1, \omega_2, \ldots, \omega_n)^T$ is the weight vector of $PL_i(p)(i = 1, 2, \ldots, n)$, $\omega_i \geq 0$, $i = 1, 2, \ldots, n$, and $\sum_{i=1}^{n}\omega_i = 1$. Especially, if $\omega = (1/n, 1/n, \ldots, 1/n)^T$, then the PFPLWA operator degenerates to the PFPLA operator.

Definition 13. Let $PL_i(p) = \{[x, PL_i^{(k)}\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle]|x \in X, PL_i^{(k)} \in S, k = 1, 2, \ldots, \#PL_i(p)\}$, $(i = 1, 2, \ldots, n)$, where $PL_i^{(k)}$ is the kth linguistic term, and $\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle$ is the corresponding uncertain probability. Then the Pythagorean fuzzy probabilistic linguistic geometric (PFPLG) operator is defined as:

$$PFPLG(PL_1(p), PL_2(p), \ldots, PL_n(p)) = (PL_1(p) \otimes PL_2(p) \otimes \cdots \otimes PL_n(p))^{\frac{1}{n}} \tag{7}$$

Definition 14. Let $PL_i(p) = \{[x, PL_i^{(k)}\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle]|x \in X, PL_i^{(k)} \in S, k = 1, 2, \ldots, \#PL_i(p)\}$, $(i = 1, 2, \ldots, n)$, where $PL_i^{(k)}$ is the kth linguistic term, and $\langle \mu_i^{(k)}, \nu_i^{(k)}\rangle$ is the corresponding uncertain probability. Then the Pythagorean fuzzy probabilistic linguistic geometric (PFPLWG) operator is defined as:

$$PFPLWG(PL_1(p), PL_2(p), \ldots, PL_n(p)) = (PL_1(p))^{\omega_1} \otimes (PL_2(p))^{\omega_2} \otimes \cdots \otimes (PL_n(p))^{\omega_n} \tag{8}$$

Where $\omega = (\omega_1, \omega_2, \ldots, \omega_n)^T$ is the weight vector of $PL_i(p)(i = 1, 2, \ldots, n)$, $\omega_i \geq 0$, $i = 1, 2, \ldots, n$, and $\sum_{i=1}^{n}\omega_i = 1$. Especially, if $\omega = (1/n, 1/n, \ldots, 1/n)^T$, then the PFPLWG operator degenerates to the PFPLG operator.

4. The weight calculation method

The objectivity and accuracy of criteria weights are quite important in decision-making problems, which directly affect the validity of decision-making results. With the existing research methods and time-varying factors, we use the minimum deviation and the vector projection method to conduct in-depth research on the time weight method and the criteria's fusion weight method.

4.1. Time weight method based on minimum deviation of AHP and entropy method

In dynamic decision-making problems, the importance of different stages is quite different. To make the time weights at different stages more convincing and improve the accuracy of the decision-making

results, we combine the subjective and objective weights to calculate the time weights. The entropy method calculates the objective time weight, reflecting the advantages in illustrating data information. Then, the DMs compare the importance of different stages in pairs according to their experience and use the AHP method to determine their objective weights. Finally, a model is set to solve the distribution coefficient by minimizing the deviation of the subjective and objective weights to the combined weight.

Definition 15. Let X be a non-empty universe of discourse, $A = \{A_1, A_2, \cdots, A_m\}$ is the alternatives set, $C = \{c_1, c_2, \cdots, c_n\}$ is the criteria sets, $S = \{s_\alpha | \alpha = 0, 1, 2, \ldots, \tau\}$ is the linguistic term set. In the eth stage, the PFPLTS evaluation of the ith alternative under the jth criterion is recorded as $PL_{ij}(p) = \{[x, PL^{(ek)}\langle \mu^{(ek)}, \nu^{(ek)}\rangle] | x \in X, PL^{(ek)} \in S, k = 1, 2, \ldots, \#PL_{ij}(p)\}$. The entropy of eth stage is defined as:

$$H^{(e)} = -\frac{1}{2\ln mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left({P^{(e)}}_{ij}\ln{P^{(e)}}_{ij} + {Q^{(e)}}_{ij}\ln{Q^{(e)}}_{ij}\right) \tag{9}$$

Where, ${P^{(e)}}_{ij}, {Q^{(e)}}_{ij}$ represent the degree of membership and non-membership contribution at the eth stage, respectively. In order to obtain entropy $H^{(e)}$, the membership and non-membership contribution ${P^{(e)}}_{ij}$ and ${Q^{(e)}}_{ij}$need to be evaluated m*n times. And

$${P^{(e)}}_{ij} = \frac{\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\mu_{ij}^{(ek)}\right)^2}{\sum_{i=1}^{m}\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\mu_{ij}^{(ek)}\right)^2} \tag{10}$$

$${Q^{(e)}}_{ij} = \frac{\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\nu_{ij}^{(ek)}\right)^2}{\sum_{i=1}^{m}\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\nu_{ij}^{(ek)}\right)^2} \tag{11}$$

Where, the function $g(\cdot)$ is given as Definition 2. The entropy weight of the eth stage is defined as:

$$\theta^{(1)}(t_e) = \frac{1 - H^{(e)}}{\sum_{e=1}^{p}(1 - H^{(e)})} \tag{12}$$

Definition 16. The 1~9 scale method is used to construct the pairwise judgment matrix $A(a_{ij})_{q\times q}$ about q stages, it is necessary to have q(q-1)/2 pairwise comparisons. The normalized eth row vector is

$$\overline{\theta}^{(2)}(t_e) = p\sqrt{\prod_{i=1}^{p} a_{ei}} \tag{13}$$

And the subjective weight of the eth stage can be calculated as:

$$\theta^{(2)}(t_e) = \frac{\overline{\theta^2}(t_e)}{\sum_{e=1}^{p}\overline{\theta^2}(t_e)} \tag{14}$$

The consistency test is performed on the judgment matrix. When $CR < 0.1$, the judgment matrix passes the consistency test, where $CR = CI/RI$, CI is the maximum eigenvalue of the judgment matrix, and RI is directly obtained by looking up the table. Otherwise, the judgment matrix should be reconstructed.

Considering subjective and objective weights, we propose a combination weight method that minimizes the deviation between AHP and entropy weight method, which is defined as follows:

Definition 17. The objective weight is $\theta^{(1)}(t_e)$, the subjective weight is $\theta^{(2)}(t_e)$, and the combination weight is $\theta(t_e) = \alpha\theta^{(1)}(t_e) + \beta\theta^{(2)}(t_e)$, where $0 \le \alpha, \beta \le 1, \alpha + \beta = 1$. In order to obtain the coefficients, the nonlinear multi-objective programming model is constructed as follows:

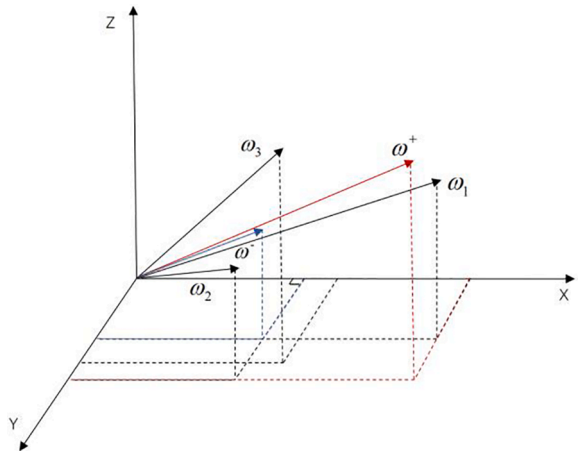


Fig. 1.. The cosine projected vector on the positive and negative ideal weights.

$$\begin{cases} \min\dfrac{1}{p-1}\displaystyle\sum_{e=1}^{p}\left(\left|\theta(t_e) - \theta^{(1)}(t_e)\right| + \left|\theta(t_e) - \theta^{(2)}(t_e)\right|\right) \\ \alpha + \beta = 1, 0 \le \alpha, \beta \le 1 \end{cases} \tag{15}$$

4.2. Fusion criteria weights based on time-varying

Since the criteria selection has been subjectively analyzed and selected by experts, this paper tends to use objective methods to determine the criteria weights. The objective weight method based on the idea of data analysis can avoid the error of subjective cognitive uncertainty and help reduce the pressure of DMs. This subsection selects three common objective weight analysis methods and proposes the dual ideal point-vector projection method to get the fusion criteria weights with time-varying factors.

4.2.1. Weight analysis method based on criterion recognition

Generally, a criterion has a higher recognition degree for the distinction and selection of alternatives, the more critical it is in the decision-making process. Then, it should be given greater weight. Conversely, if the criterion has a low degree of recognition in evaluating the alternatives, the criterion is not conducive to decision-making and should be given a smaller weight. The weight method based on criterion recognition of PFPLTS is proposed in this subsection.

Definition 18. Let ${PL_{ij}}^{(e)} = \{\langle c_j, {PL_{ij}}^{(ek)}(\langle {\mu_{ij}}^{(ek)}, {\nu_{ij}}^{(ek)}\rangle)\rangle | c_j \in C, {PL_{ij}}^{(ek)} \in S, k = 1, 2, \ldots, \#PL_{ij}(p)\}$, which means the PFPLTS evaluation information of alternative A_i about criterion c_j at the eth stage. The degree of recognition is defined as:

$$O_j^{(e)} = \sum_{i=1}^{m}\sum_{l=1, l\neq i}^{m} d\left({PL_{ij}}^{(e)}, {PL_{lj}}^{(e)}\right) \tag{16}$$

Where, $d({PL_{ij}}^{(e)}, {PL_{lj}}^{(e)})$ means the deviation between alternative A_i and A_l at the eth stage.

Definition 19. Let $\omega^{(1)} = (\omega_1^{(1)}, \omega_2^{(1)}, \ldots, \omega_n^{(1)})$ denote the fusion weight based on the recognition degree of the criteria, then the recognition weight $\omega_j^{(1)}$ with time varying of criterion c_j is

$$\omega_j^{(1)} = \sum_{e=1}^{q}\theta(t_e)\frac{O_j^{(e)}}{\sum_{p=1}^{n} O_p^{(e)}} = \sum_{e=1}^{q}\theta(t_e)\frac{\sum_{i=1}^{m}\sum_{l=1, l\neq i}^{m} d\left({PL_{ij}}^{(e)}, {PL_{lj}}^{(e)}\right)}{\sum_{p=1}^{n}\sum_{i=1}^{m}\sum_{l=1, l\neq i}^{m} d\left({PL_{ip}}^{(e)}, {PL_{lp}}^{(e)}\right)} \tag{17}$$

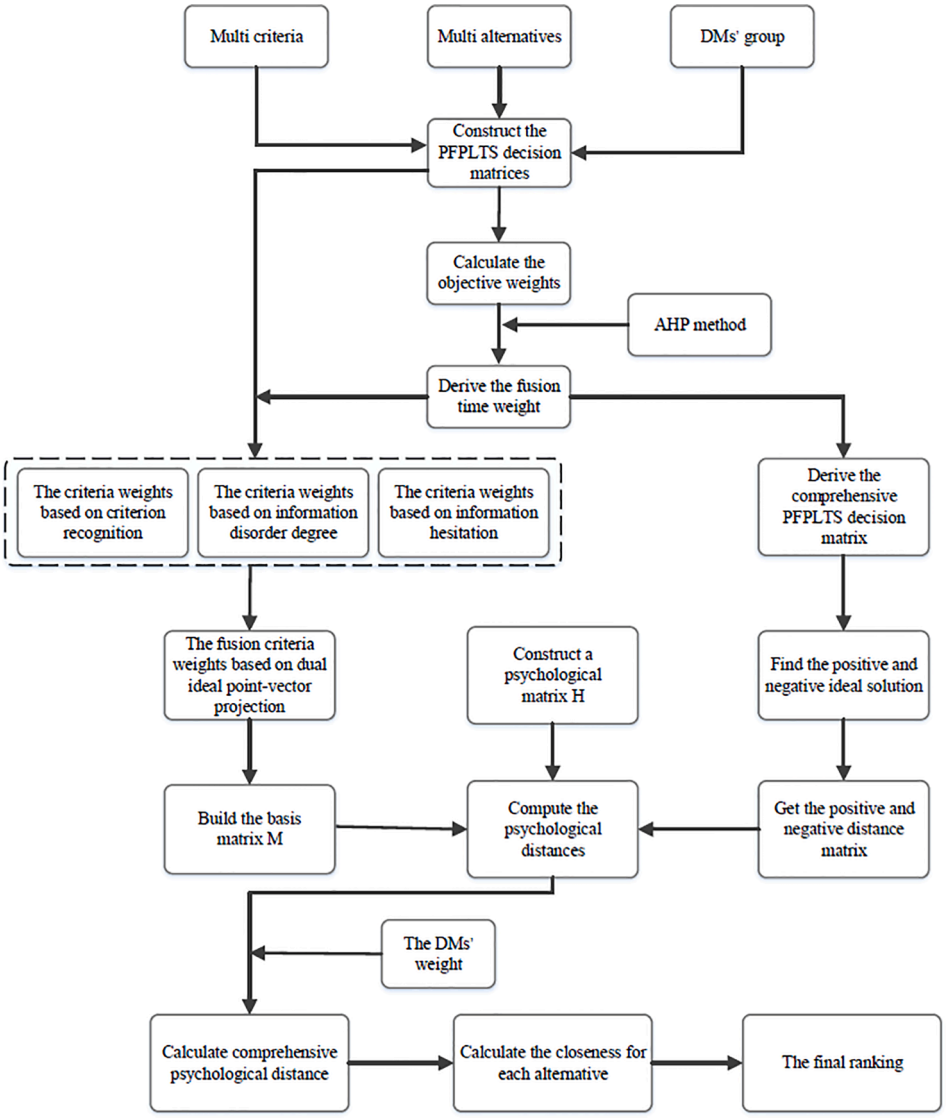


Fig. 2.. Process of PFPL-PT method for dynamic MCGDM.

In particular, if the identification degree of the criterion to the alternatives is 0, it is completely impossible to distinguish the pros and cons of the alternatives, which means that the criterion has no meaning in decision-making, and the weight should be set to 0.

4.2.2. Weight analysis method based on information disorder degree

The orderliness of decision-making information is also crucial to the impact of decision making, which can be described as the lower the disorder degree, the greater the utility value of the information. Information entropy is usually used to measure the disorder of information. The smaller the information entropy, the lower the information disorder degree, and the greater the effect value, then the greater the weight of the criterion. The weight method based on the information disorder degree of PFPLTS is proposed in this subsection.

Definition 20. Let $PL_{ij}^{(e)} = \{\langle c_j, PL_{ij}^{(ek)}(\langle \mu_{ij}^{(ek)}, \nu_{ij}^{(ek)}\rangle)\rangle | c_j \in C, PL_{ij}^{(ek)} \in S, k = 1,2,\ldots,\#PL_{ij}(p)\}$, which means the PFPLTS evaluation information of alternative A_i about criterion c_j at the eth stage. The information entropy of criterion c_j is defined as:

$$E_j^{(e)} = \frac{1}{2\ln m}\sum_{i=1}^{m}\left(P_{ij}^{(e)}\ln P_{ij}^{(e)} + Q_{ij}^{(e)}\ln Q_{ij}^{(e)}\right) \tag{18}$$

Where, $P_{ij}^{(e)}, Q_{ij}^{(e)}$ represent the degree of membership and non-membership contribution, respectively. And

$$P_{ij}^{(e)} = \frac{\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\mu_{ij}^{(ek)}\right)^2}{\sum_{i=1}^{m}\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\mu_{ij}^{(ek)}\right)^2} \tag{19}$$

$$Q_{ij}^{(e)} = \frac{\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\nu_{ij}^{(ek)}\right)^2}{\sum_{i=1}^{m}\sum_{k=1}^{\#PL_{ij}(P)} g\left(PL_{ij}^{(ek)}\right)\left(\nu_{ij}^{(ek)}\right)^2} \tag{20)}$$

Where the function $g(\cdot)$ is given as definition 2. Especially, $E_i^{(e)} = 1$, when $P_{ij}^{(e)} = Q_{ij}^{(e)} = \frac{1}{m}$.

Definition 21. Let $\omega^{(2)} = (\omega_1^{(2)}, \omega_2^{(2)}, \ldots, \omega_n^{(2)})$ denote the fusion weight based on information disorder degree, then the information disorder weight $\omega_j^{(2)}$ with time-varying of criterion c_j is

$$\omega_j^{(2)} = \sum_{e=1}^{q} \theta(t_e) \frac{1 - E_j^{(e)}}{\sum_{j=1}^{n} \left(1 - E_j^{(e)}\right)} \tag{21}$$

When the degree of membership and non-membership contributions tend to be the same, that is, $E_j^{(e)} = 1$, we can ignore this criterion in decision-making, and the corresponding weight is 0.

4.2.3. Weight analysis method based on information hesitation degree

The advantage of PFPLTS is that it can help DMs express the uncertainty of linguistic evaluation, reflect the probability and ambiguity of each linguistic evaluation. Obviously, the lower the hesitation of the DM's evaluation information, the more certain the decision-making is. It further shows that the evaluation information has higher accuracy, and the corresponding alternative should be given greater weight. The weight method based on the information hesitation degree of PFPLTS is proposed in this subsection.

Definition 22. Let $PL_{ij}^{(e)} = \{\langle c_j, PL_{ij}^{(ek)}(\langle \mu_{ij}^{(ek)}, \nu_{ij}^{(ek)} \rangle)\rangle | c_j \in C, PL_{ij}^{(ek)} \in S, k = 1,2,\ldots,\#PL_{ij}(p)\}$, which means the PFPLTS evaluation information of alternative A_i about criterion c_j at the eth stage. The information hesitation degree $H_j^{(e)}$ is defined as:

$$H_j^{(e)} = \sum_{i=1}^{m} \sum_{k=1}^{\#LP_{ij}(p)} \left[1 - \left(\mu_{ij}^{(ek)}\right)^2 - \left(\nu_{ij}^{(ek)}\right)^2\right] \tag{22}$$

Definition 23. Let $\omega^{(3)} = (\omega_1^{(3)}, \omega_2^{(3)}, \ldots, \omega_n^{(3)})$ denote the fusion weight based on information hesitation degree, then the information hesitation weight $\omega_j^{(3)}$ with time-varying of criterion c_j is

$$\omega_j^{(3)} = \sum_{e=1}^{q} \frac{1 - H_j^{(e)} \Big/ \sum_{j=1}^{n} H_j^{(e)}}{\sum_{j=1}^{n} \left(1 - H_j^{(e)} \Big/ \sum_{j=1}^{n} H_j^{(e)}\right)} \tag{23}$$

4.2.4. Fusion weight method based on dual ideal point-vector projection

The ideal point method has a wide range of applications in MCGDM problems, and many scholars have carried out in-depth research on it. The vector projection method has become a commonly used tool in studying multiple indexes due to its simple operation and easy understanding characteristics. The determination of the fusion weight is essentially a multi-index problem. In view of the advantages of the ideal point and vector projection method in multi-criteria problems, they will be integrated to get the fusion weight in this section.

In MCGDM problems, the effects of criteria are often divided into positive and negative effects. Generally, we expect that the positive criterion weight to be larger and the negative criterion weight to be smaller. Each weight vector is cosine projected on the positive and negative ideal weights, and the projection diagram is shown in Fig. 1.

The definitions of positive and negative ideal weight ω^+ and ω^- are presented below.

Definition 24. Let $\omega^{(i)} = (\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_n^{(i)}), i = 1,2,3$, which represent the criteria weight based on criterion recognition, information disorder degree and information hesitation degree, respectively. When the criterion is positive,

$$\omega^+ = \left(\max_j \omega_j^{(1)}, \max_j \omega_j^{(2)}, \max_j \omega_j^{(3)}\right) \tag{24}$$

$$\omega^- = \left(\min_j \omega_j^{(1)}, \min_j \omega_j^{(2)}, \min_j \omega_j^{(3)}\right) \tag{25}$$

When the criterion is negative,

$$\omega^+ = \left(\min_j \omega_j^{(1)}, \min_j \omega_j^{(2)}, \min_j \omega_j^{(3)}\right) \tag{26}$$

$$\omega^- = \left(\max_j \omega_j^{(1)}, \max_j \omega_j^{(2)}, \max_j \omega_j^{(3)}\right) \tag{27}$$

Definition 25. We have the criterion weight vector $\omega^{(i)} = (\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_n^{(i)}), i = 1,2,3$, positive and negative ideal weight ω^+ and ω^-. Let vector $\omega_j^{ide} = (\omega_j^{(1)}, \omega_j^{(2)}, \omega_j^{(3)})^T$. The positive projection intensity of the weight vector in the positive ideal weight is denoted as B_j, and the ratio of the positive projection intensity to the total sum is the positive fusion weight of the vector, denoted as ω_j^+.

$$B_j = \frac{< \omega^+, \omega_j^{ide} >}{\| \omega^+ \| \| \omega_j^{ide} \|} \| \omega_j^{ide} \| = \frac{< \omega^+, \omega_j^{ide} >}{\| \omega^+ \|} \tag{28}$$

$$\omega_j^+ = \frac{B_j}{\sum_{i=1}^{n} B_j} \tag{29}$$

The negative projection intensity of the weight vector in the negative ideal weight is denoted as D_j, and the ratio of the negative projection intensity to the total sum is the negative fusion weight of the vector, denoted as ω_j^-.

$$D_j = \frac{< \omega^-, \omega_j^{ide} >}{\| \omega^- \| \| \omega_j^{ide} \|} \| \omega_j^{ide} \| = \frac{< \omega^-, \omega_j^{ide} >}{\| \omega^+ \|} \tag{30}$$

$$\omega_j^- = \frac{D_j}{\sum_{j=1}^{n} D_j} \tag{31}$$

Then, the fusion weight is defined as:

$$\omega_j = \frac{\omega_j^- + \omega_j^+}{2} \tag{32}$$

5. Dynamic Pythagorean fuzzy probabilistic linguistic MCGDM with Psy-TOPSIS method

In this section, we mainly introduce the Pythagorean fuzzy probabilistic linguistic TOPSIS method with psychological distance measure (Psy-TOPSIS) method. Then an approach for MCGDM based on Pythagorean fuzzy probabilistic linguistic Psy-TOPSIS is proposed.

5.1. Novel TOPSIS method with psychological distance measure

Individuals do not consistently distribute their attention equally to each dimension when describing objects, and it differs their psychological space. A psychological distance measure, which can clarify the influence of different psychological factors and background information of the DMs and alternatives, is proposed. In psychological space, the preferential relationship between the alternatives is reflected by indifferent vectors $(v_{fj}(j = 1,2,\ldots,n-1))$ and dominant vector (v_d). The indifferent vectors v_{fj} quantitatively describe the relative gain due to criterion substitution and the dominant vector v_d can manifest the direction of the optimal alternative.

In MCGDM problems, the alternatives set is$A = (A_1, A_2, \ldots, A_m)$, $C = \{c_1, c_2, \cdots, c_n\}$ is the criteria sets, and the criteria weight is $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. To calculate the indifferent vectors, we compare each criterion weight with the position weighted average weight $(\widetilde{\omega})$ based on the normal distribution. Then, the indifferent vectors can be calculated as:

$$v_{fj} = \left(-\frac{\omega_{j+1}}{\widetilde{\omega}}, 0, \ldots, 0, \frac{\omega_1}{\widetilde{\omega}}\right)^T \tag{34}$$

Table 1
Individual PFPLTS decision matrix D^{11}.

	c_1	c_2	c_3	c_4	c_5
A_1	$\{s_5\langle 0.8,0.1\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_2\langle 0.7,0.3\rangle, s_3\langle 0.2,0.4\rangle\}$	$\{s_3\langle 0.2,0.5\rangle, s_4\langle 0.7,0.2\rangle\}$	$\{s_5\langle 0.8,0.1\rangle\}$
A_2	$\{s_2\langle 0.5,0.2\rangle, s_3\langle 0.3,0.1\rangle\}$	$\{s_3\langle 0.6,0.2\rangle, s_4\langle 0.3,0.2\rangle\}$	$\{s_5\langle 0.9,0.2\rangle\}$	$\{s_3\langle 0.8,0.1\rangle\}$	$\{s_1\langle 0.9,0.1\rangle\}$
A_3	$\{s_4\langle 0.6,0.1\rangle, s_5\langle 0.3,0.2\rangle\}$	$\{s_4\langle 0.6,0.3\rangle, s_5\langle 0.3,0.1\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_4\langle 0.9,0.1\rangle\}$	$\{s_1\langle 0.9,0.1\rangle\}$
A_4	$\{s_1\langle 0.9,0.1\rangle\}$	$\{s_1\langle 0.6,0.2\rangle, s_2\langle 0.4,0.3\rangle\}$	$\{s_2\langle 0.5,0.2\rangle, s_3\langle 0.4,0.3\rangle\}$	$\{s_3\langle 0.7,0.2\rangle, s_4\langle 0.3,0.2\rangle\}$	$\{s_1\langle 0.7,0.3\rangle, s_2\langle 0.2,0.2\rangle\}$

where $\frac{\omega_1}{\omega}$ is at the $(j+1)$th position. According to Berkowitsch [26], the dominance vector v_d is orthogonal to all indifference vectors $(v_{fj})_{n-1}$, which means $v_d \cdot v_{fj} = 0, j = 1,2,\ldots,n-1$. Then, the dominant vector is

$$v_d = \left(\frac{\omega_1}{\widetilde{\omega}}, \frac{\omega_2}{\widetilde{\omega}}, \ldots, \frac{\omega_n}{\widetilde{\omega}}\right)^T \tag{35}$$

To calculate the projection in each direction, the basis matrix(M) can be built as

$$M = \left(v_{f1}, v_{f2}, \ldots, v_{fn-1}, v_d\right)^T \tag{36}$$

And the basis matrix (M^*) after length normalization is

$$M^* = \left(\frac{v_{f1}}{\| v_{f1} \|}, \frac{v_{f2}}{\| v_{f2} \|}, \ldots, \frac{v_{fn-1}}{\| v_{fn-1} \|}, \frac{v_d}{\| v_d \|}\right)^T \tag{37}$$

It is significant to weigh the dominance vector more strongly in psychological distance measure by adjusting the parameter w_{dom}. Hence, we construct a psychological matrix H:

$$H = diag(1,1,\ldots,1,w_{dom}) \tag{38}$$

Then, the psychological distance between alternatives A_k and A_l can be computed as follows:

$$\| D \|_i = \| HM^{*-1}d' \|_i, i = 1,2,\infty. \tag{39}$$

Where, d is the standard distance matrix between two alternatives A_k and A_l, $d(A_k,A_l) = (PL_{k1} - PL_{l1}, PL_{k2} - PL_{l2}, \ldots, PL_{kn} - PL_{\ln})$, PL_{ij} is the PFPLTS evaluation information of alternative A_i about criterion c_j. Especially, $\| D \|$is 1-norm, when $i = 1$. $\| D \|$is 2-norm, when$i = 2$. $\| D \|$is infinity-norm, when $i = \infty$.

5.2. An approach for MCGDM problems based on PFPLF-PT method

We mainly introduce the dynamic Pythagorean fuzzy probabilistic linguistic MCGDM with the Psy-TOPSIS method. The procedure is visualized in Fig. 2 and summarized as follows:

Step1: Collect and preprocess the evaluation information of the PFPLTS decision matrices $D^{(ge)} = (PL_{ij}^{(ge)})_{m\times n}$ given by the gth DM under the same criteria at eth stage.

Step2: Calculate the objective weights vector $\theta_1^{(g)}(t_e)$ of the gth DM according to the method in Definition 15.

Step3: To subjectively evaluate the importance of the three stages involved in the problem, use the 1~9 scale method in Definition 16 to construct pairwise judgment matrix $A(a_{ij})_{q\times q}$, and calculate the subjective weight vector $\theta_2^{(g)}(t_k)$ through the AHP method.

Step4: Combine the subjective and objective time weights, use a nonlinear optimization model to calculate the coefficients of the time combination weights, and then get the time combination weights vector $\theta^{(g)} = (\theta^{(g)}(t_1,t_2,\ldots,t_q))$, which is assigned by the method in Section 4.1.

Step5: Derive the comprehensive PFPLTS decision matrix $D^{(g)} = (V_{ij}^{(com-g)})_{m\times n}$ of each expert. And

Table 2.
Individual PFPLTS decision matrix D^{12}.

	c_1	c_2	c_3	c_4	c_5
A_1	$\{s_2\langle 0.7,0.4\rangle, s_3\langle 0.3,0.4\rangle\}$	$\{s_4\langle 0.7,0.3\rangle, s_5\langle 0.3,0.1\rangle\}$	$\{s_2\langle 0.7,0.3\rangle, s_3\langle 0.2,0.4\rangle\}$	$\{s_4\langle 0.9,0.2\rangle, s_5\langle 0.1,0.3\rangle\}$	$\{s_4\langle 0.7,0.2\rangle, s_5\langle 0.2,0.5\rangle\}$
A_2	$\{s_2\langle 0.5,0.3\rangle, s_3\langle 0.4,0.1\rangle\}$	$\{s_4\langle 0.7,0.2\rangle\}$	$\{s_5\langle 0.8,0.1\rangle\}$	$\{s_3\langle 0.9,0.1\rangle\}$	$\{s_2\langle 0.7,0.4\rangle, s_3\langle 0.1,0.3\rangle\}$
A_3	$\{s_4\langle 0.7,0.2\rangle, s_5\langle 0.2,0.3\rangle\}$	$\{s_3\langle 0.6,0.3\rangle, s_4\langle 0.4,0.2\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_4\langle 0.8,0.2\rangle\}$	$\{s_1\langle 0.8,0.3\rangle, s_2\langle 0.2,0.2\rangle\}$
A_4	$\{s_1\langle 0.8,0.2\rangle, s_2\langle 0.2,0.4\rangle\}$	$\{s_1\langle 0.7,0.3\rangle, s_2\langle 0.3,0.3\rangle\}$	$\{s_3\langle 0.6,0.2\rangle\}$	$\{s_3\langle 0.9,0.1\rangle\}$	$\{s_1\langle 0.8,0.2\rangle\}$

Table 3.
Individual PFPLTS decision matrix D^{13}.

	c_1	c_2	c_3	c_4	c_5
A_1	$\{s_3\langle 0.7,0.3\rangle, s_4\langle 0.3,0.4\rangle\}$	$\{s_4\langle 0.9,0.1\rangle\}$	$\{s_2\langle 0.8, 0.2\rangle\}$	$\{s_4\langle 0.9,0.2\rangle, s_5\langle 0.1,0.3\rangle\}$	$\{s_4\langle 0.7,0.3\rangle, s_5\langle 0.2,0.5\rangle\}$
A_2	$\{s_2\langle 0.5,0.3\rangle, s_3\langle 0.4,0.1\rangle\}$	$\{s_4\langle 0.8,0.2\rangle\}$	$\{s_5\langle 0.8, 0.1\rangle\}$	$\{s_3\langle 0.9,0.1\rangle\}$	$\{s_1\langle 0.9,0.1\rangle\}$
A_3	$\{s_4\langle 0.7,0.3\rangle, s_5\langle 0.2,0.4\rangle\}$	$\{s_3\langle 0.6,0.3\rangle, s_4\langle 0.4,0.3\rangle\}$	$\{s_5\langle 0.9, 0.1\rangle\}$	$\{s_4\langle 0.8,0.1\rangle\}$	$\{s_4\langle 0.7,0.3\rangle, s_5\langle 0.3,0.2\rangle\}$
A_4	$\{s_1\langle 0.7,0.3\rangle, s_2\langle 0.3,0.1\rangle\}$	$\{s_1\langle 0.8,0.2\rangle, s_2\langle 0.2,0.1\rangle\}$	$\{s_3\langle 0.7, 0.1\rangle\}$	$\{s_3\langle 0.9,0.1\rangle\}$	$\{s_1\langle 0.9,0.1\rangle\}$

Table 4.
Individual PFPLTS decision matrix D^{21}.

	c_1	c_2	c_3	c_4	c_5
A_1	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_3\langle 0.6,0.4\rangle, s_4\langle 0.3,0.1\rangle\}$	$\{s_3\langle 0.8,0.2\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$
A_2	$\{s_2\langle 0.5,0.2\rangle, s_3\langle 0.3,0.1\rangle\}$	$\{s_3\langle 0.5,0.3\rangle, s_4\langle 0.5,0.4\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_2\langle 0.6,0.4\rangle, s_3\langle 0.4,0.1\rangle\}$	$\{s_1\langle 0.9,0.2\rangle\}$
A_3	$\{s_4\langle 0.8,0.1\rangle\}$	$\{s_5\langle 0.8,0.2\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_4\langle 0.8,0.3\rangle, s_5\langle 0.1,0.3\rangle\}$	$\{s_1\langle 0.9,0.2\rangle\}$
A_4	$\{s_3\langle 0.8,0.2\rangle\}$	$\{s_2\langle 0.7,0.3\rangle, s_3\langle 0.2,0.1\rangle\}$	$\{s_3\langle 0.6,0.3\rangle, s_4\langle 0.4,0.3\rangle\}$	$\{s_3\langle 0.8,0.1\rangle\}$	$\{s_1\langle 0.8,0.3\rangle, s_2\langle 0.2,0.1\rangle\}$

Table 5.
Individual PFPLTS decision matrix D^{22}.

	c_1	c_2	c_3	c_4	c_5
A_1	$\{s_2\langle 0.9,0.2\rangle, s_3\langle 0.1,0.1\rangle\}$	$\{s_4\langle 0.9,0.2\rangle\}$	$\{s_3\langle 0.7,0.3\rangle\}$	$\{s_4\langle 0.8,0.3\rangle, s_5\langle 0.2,0.1\rangle\}$	$\{s_4\langle 0.8, 0.2\rangle\}$
A_2	$\{s_3\langle 0.6,0.2\rangle\}$	$\{s_4\langle 0.8,0.2\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_2\langle 0.3,0.3\rangle, s_3\langle 0.7,0.2\rangle\}$	$\{s_3\langle 0.9, 0.2\rangle\}$
A_3	$\{s_4\langle 0.8,0.1\rangle\}$	$\{s_3\langle 0.7,0.3\rangle, s_4\langle 0.3,0.2\rangle\}$	$\{s_5\langle 0.9,0.1\rangle\}$	$\{s_4\langle 0.9,0.2\rangle\}$	$\{s_1\langle 0.8, 0.2\rangle\}$
A_4	$\{s_2\langle 0.9,0.1\rangle\}$	$\{s_2\langle 0.7,0.3\rangle\}$	$\{s_3\langle 0.7,0.2\rangle, s_4\langle 0.3,0.2\rangle\}$	$\{s_4\langle 0.7,0.3\rangle, s_5\langle 0.3,0.4\rangle\}$	$\{s_1\langle 0.8, 0.2\rangle\}$

Table 6.
Individual PFPLTS decision matrix D^{23}.

	c_1	c_2	c_3	c_4	c_5
A_1	$\{s_3\langle 0.8, 0.2\rangle\}$	$\{s_4\langle 0.9,0.2\rangle\}$	$\{s_3\langle 0.8, 0.2\rangle\}$	$\{s_4\langle 0.9,0.2\rangle, s_5\langle 0.1, 0.3\rangle\}$	$\{s_4\langle 0.8, 0.2\rangle\}$
A_2	$\{s_3\langle 0.7, 0.2\rangle\}$	$\{s_4\langle 0.8,0.2\rangle\}$	$\{s_4\langle 1\rangle\}$	$\{s_2\langle 0.5,0.3\rangle, s_3\langle 0.5, 0.2\rangle\}$	$\{s_1\langle 0.9, 0.1\rangle\}$
A_3	$\{s_4\langle 0.8, 0.2\rangle\}$	$\{s_3\langle 0.7,0.3\rangle, s_4\langle 0.3, 0.2\rangle\}$	$\{s_5\langle 0.9, 0.1\rangle\}$	$\{s_4\langle 0.9,0.2\rangle\}$	$\{s_4\langle 0.8, 0.2\rangle\}$
A_4	$\{s_2\langle 0.9, 0.1\rangle\}$	$\{s_2\langle 0.8,0.2\rangle\}$	$\{s_3\langle 0.8, 0.2\rangle\}$	$\{s_4\langle 0.8,0.2\rangle\}$	$\{s_1\langle 0.8, 0.2\rangle\}$

Table 7.
Judgment matrix $A(a_{ij})_{q\times q}$.

	t1	t2	t3
t1	1	1/3	5
t2	3	1	7
t3	1/5	1/7	1

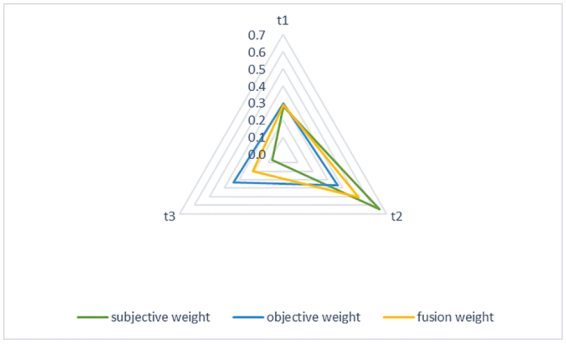


Fig. 3. . The time weight radar map of E1.

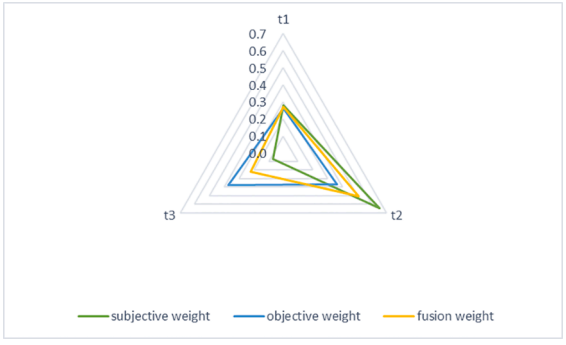


Fig. 4. The time weight radar map of E2.

$$V_{ij}^{(com-g)} = \sum_{e=1}^{p} \theta(t_e) PL_{ij}^{(ge)} \tag{40}$$

Where, $PL_{ij}^{(ge)} = [PL_{ij}^{(gk)}(t_e)\langle \mu_{ij}(t_e)^{(gk)}, \nu_{ij}(t_e)^{(gk)}\rangle]$, $PL_{ij}^{(gk)}(t_e) \in S, k = 1, 2, \ldots, \#PL_{ij}^{(ge)}$. According to formula in Definition 6,

$$V_{ij}^{(com-g)} = \sum_{e=1}^{p} \theta(t_e)\left[PL_{ij}^{(g)}(t_e)\left\langle \mu_{ij}(t_e)^{(g)}, \nu_{ij}(t_e)^{(g)}\right\rangle\right]$$
$$= \sum_{e=1}^{p} \cup_{PL_{ij}^{(gk)}(t_e)\left\langle \mu_{ij}(t_e)^{(gk)}, \nu_e(t_e)^{(gk)}\right\rangle \in PL_{ij}^{(ge)}}$$
$$\left\{\left(1-\left(1-PL_{ij}^{(gk)}(t_e)\right)^{\theta(t_e)}\right)\left\langle \mu_{ij}(t_e)^{(gk)}, \nu_{ij}(t_e)^{(gk)}\right\rangle\right\} \tag{41}$$

Step6: According to the score function in Definition 2 and Definition 7, find the positive and negative ideal solution of each DM, respectively.

Step7: Calculate the distance from each alternative to the positive and negative ideal solution, and get the positive and negative distance matrix of every DM, respectively.

Step8: Calculate the criteria weights with time-varying based on criterion recognition by Eq. (17).

Step9: Calculate the criteria weights with time-varying based on information disorder degree by Eq. (21).

Step10: Calculate the criteria weights with time-varying based on information hesitation degree by Eq. (23).

Step11: Calculate the fusion criteria weights based on the dual ideal point-vector projection method as Eq. (32). The time-varying fusion weight vector is recorded as $\omega^{[g]} = (\omega_1{}^{[g]}, \omega_2{}^{[g]}, \ldots, \omega_n{}^{[g]})^T$.

Step12: Calculate the dominance vector v_d and the indifference vectors $(v_{fj})_{n-1}$ by the fusion criteria weight $\omega^{[g]}$. Then, build the basis matrix M.

Step13: Construct a psychological matrix $H = diag(1, 1, \ldots, 1, w_{dom})$

Step14: Compute the psychological distances $d_{psy}^{(g)}(V_{ij}{}^{(com-g)}, V_j{}^{+})$ and $d_{psy}^{(g)}(V_{ij}{}^{(com-g)}, V_j{}^{-})$, $V_j{}^{+}$ and $V_j{}^{-}$are the positive ideal solution and negative ideal solution of $V_{ij}{}^{(com-g)}$respectively.

Step15: Through the weighted average method, the comprehensive psychological distance $d_{psy}^{(g)}(A_i, A^{+})$ and $d_{psy}^{(g)}(A_i, A^{-})$ are aggregated based on each DM's psychological distance in Step 14, and the DMs' weight vector is $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_g)^T$.

Step16: Calculate the closeness η_i for each alternative A_i.

Step17: Sort the alternatives according to η_i and choose the best one.

6. Case study and analysis

This section will give an example about the site selecting of the COVID-19 vaccination center to illustrate the proposed method.

6.1. *Case study: site selecting of the COVID-19 vaccination center*

With the control of the COVID-19 in the country, production and life are slowly recovering. China successfully eliminated COVID-19 and

Table 8
Time comprehensive decision matrix D^1.

	c_1	c_2
A_1	$\{s_5\langle 0.8, 0.056\rangle\}$	$\{s_5\langle 0.81, 0.004\rangle\}$
A_2	$\{s_2\langle 0.125, 0.018\rangle, s_{2.2335}\langle 0.1, 0.006\rangle, s_{2.3329}\langle 0.075, 0.009\rangle, s_{2.5405}\langle 0.06, 0.003\rangle, s_{2.5604}\langle 0.1, 0.006\rangle, s_{2.7503}\langle 0.08, 0.002\rangle, s_{2.8311}\langle 0.06, 0.003\rangle, s_3\langle 0.048, 0.001\rangle\}$	$\{s_{3.7773}\langle 0.336, 0.008\rangle, s_{3.9999}\langle 0.168, 0.008\rangle\}$
A_3	$\{s_{3.9999}\langle 0.294, 0.006\rangle, s_5\langle 0.435, 0.009\rangle\}$	$\{s_{3.3641}\langle 0.216, 0.027\rangle, s_{3.5758}\langle 0.144, 0.027\rangle, s_{3.8512}\langle 0.144, 0.018\rangle, s_{3.9999}\langle 0.096, 0.018\rangle, s_5\langle 0.3, 0.03\rangle\}$
A_4	$\{s_{1.546}\langle 0.441, 0.012\rangle, s_{1.7391}\langle 0.189, 0.004\rangle, s_{2.1912}\langle 0.189, 0.012\rangle, s_{2.3482}\langle 0.081, 0.004\rangle\}$	$\{s_1\langle 0.336, 0.012\rangle, s_{1.2236}\langle 0.084, 0.006\rangle, s_{1.32}\langle 0.224, 0.018\rangle, s_{1.5257}\langle 0.056, 0.009\rangle, s_{1.546}\langle 0.144, 0.012\rangle, s_{1.7391}\langle 0.036, 0.006\rangle, s_{1.8223}\langle 0.096, 0.018\rangle, s_2\langle 0.024, 0.009\rangle\}$

	c_3	c_4	c_5
A_1	$\{s_2\langle 0.3920, 0.018\rangle, s_{2.3329}\langle 0.112, 0.024\rangle, s_{2.5604}\langle 0.112, 0.024\rangle, s_{2.8311}\langle 0.032, 0.032\rangle\}$	$\{s_{3.7773}\langle 0.162, 0.02\rangle, s_{3.9999}\langle 0.567, 0.008\rangle, s_5\langle 0.171, 0.147\rangle\}$	$\{s_5\langle 0.648, 0.056\rangle\}$
A_2	$\{s_5\langle 0.576, 0.002\rangle\}$	$\{s_3\langle 0.648, 0.001\rangle\}$	$\{s_{1.546}\langle 0.567, 0.004\rangle, s_{2.1912}\langle 0.081, 0.003\rangle\}$
A_3	$\{s_5\langle 0.729, 0.001\rangle\}$	$\{s_{3.9999}\langle 0.576, 0.002\rangle\}$	$\{s_{1.9684}\langle 0.504, 0.009\rangle, s_{2.1912}\langle 0.216, 0.006\rangle\}$
A_4	$\{s_{2.7503}\langle 0.21, 0.004\rangle, s_3\langle 0.168, 0.006\rangle\}$	$\{s_3\langle 0.567, 0.02\rangle, s_{3.3641}\langle 0.243, 0.002\rangle\}$	$\{s_1\langle 0.504, 0.006\rangle, s_{1.32}\langle 0.144, 0.004\rangle\}$

Table 9
Time comprehensive decision matrix D^2.

	c_1	c_2	c_3
A_1	$\{s_5\langle 0.864, 0.006\rangle\}$	$\{s_5\langle 0.81, 0.004\rangle\}$	$\{s_3\langle 0.336, 0.024\rangle, s_{3.3413}\langle 0.168, 0.006\rangle\}$
A_2	$\{s_{2.7685}\langle 0.21, 0.008\rangle, s_3\langle 0.126, 0.004\rangle\}$	$\{s_{3.7943}\langle 0.32, 0.012\rangle, s_4\langle 0.448, 0.008\rangle\}$	$\{s_5\langle 0.81, 0\rangle\}$
A_3	$\{s_4\langle 0.512, 0.002\rangle\}$	$\{s_5\langle 0.968, 0.03\rangle\}$	$\{s_5\langle 0.729, 0.001\rangle\}$
A_4	$\{s_{2.3112}\langle 0.648, 0.002\rangle\}$	$\{s_{1.9999}\langle 0.392, 0.018\rangle, s_{2.3112}\langle 0.336, 0.012\rangle\}$	$\{s_3\langle 0.336, 0.012\rangle, s_{3.3413}\langle 0.28, 0.008\rangle, s_{3.5956}\langle 0.24, 0.006\rangle, s_{3.8352}\langle 0.2, 0.004\rangle\}$

	c_4	c_5
A_1	$\{s_{3.7943}\langle 0.576, 0.012\rangle, s_5\langle 0.568, 0.018\rangle\}$	$\{s_5\langle 0.576, 0.004\rangle\}$
A_2	$\{s_{1.9999}\langle 0.216, 0.036\rangle, s_{2.2559}\langle 0.288, 0.012\rangle, s_{2.3112}\langle 0.18, 0.018\rangle, s_{2.5406}\langle 0.24, 0.006\rangle, s_{2.5603}\langle 0.288, 0.012\rangle, s_{2.7685}\langle 0.384, 0.004\rangle, s_{2.8135}\langle 0.24, 0.006\rangle, s_3\langle 0.32, 0.002\rangle\}$	$\{s_{3.0276}\langle 0.729, 0.004\rangle\}$
A_3	$\{s_4\langle 0.648, 0.012\rangle, s_5\langle 0.405, 0.008\rangle\}$	$\{s_1\langle 0.576, 0.008\rangle\}$
A_4	$\{s_{3.283}\langle 0.448, 0.006\rangle, s_{3.7943}\langle 0.32, 0.002\rangle\}$	$\{s_1\langle 0.512, 0.006\rangle, s_{1.2988}\langle 0.32, 0.004\rangle\}$

Table 10
The positive and negative ideal solution of each DM.

	c_1	c_2	c_3	c_4	c_5
$A^{(1)+}$	$\{s_5\langle 0.8, 0.056\rangle\}$	$\{s_4\langle 0.9, 0.2\rangle\}$	$\{s_2\langle 0.3920, 0.018\rangle, s_{2.3329}\langle 0.112, 0.024\rangle, s_{2.5604}\langle 0.112, 0.024\rangle, s_{2.8311}\langle 0.032, 0.032\rangle\}$	$\{s_{3.7773}\langle 0.162, 0.02\rangle, s_{3.9999}\langle 0.567, 0.008\rangle, s_5\langle 0.171, 0.147\rangle\}$	$\{s_5\langle 0.648, 0.056\rangle\}$
$A^{(1)-}$	$\{s_2\langle 0.125, 0.018\rangle, s_{2.2335}\langle 0.1, 0.006\rangle, s_{2.3329}\langle 0.075, 0.009\rangle, s_{2.5405}\langle 0.06, 0.003\rangle, s_{2.5604}\langle 0.1, 0.006\rangle, s_{2.7503}\langle 0.08, 0.002\rangle, s_{2.8311}\langle 0.06, 0.003\rangle, s_3\langle 0.048, 0.001\rangle\}$	$\{s_1\langle 0.336, 0.012\rangle, s_{1.2236}\langle 0.084, 0.006\rangle, s_{1.32}\langle 0.224, 0.018\rangle, s_{1.5257}\langle 0.056, 0.009\rangle, s_{1.546}\langle 0.144, 0.012\rangle, s_{1.7391}\langle 0.036, 0.006\rangle, s_{1.8223}\langle 0.096, 0.018\rangle, s_2\langle 0.024, 0.009\rangle\}$	$\{s_{2.7503}\langle 0.21, 0.004\rangle, s_3\langle 0.168, 0.006\rangle\}$	$\{s_3\langle 0.567, 0.02\rangle, s_{3.3641}\langle 0.243, 0.002\rangle\}$	$\{s_1\langle 0.504, 0.006\rangle, s_{1.32}\langle 0.144, 0.004\rangle\}$
$A^{(2)+}$	$\{s_5\langle 0.864, 0.006\rangle\}$	$\{s_5\langle 0.968, 0.03\rangle\}$	$\{s_5\langle 0.81, 0\rangle\}$	$\{s_{3.7943}\langle 0.576, 0.012\rangle, s_5\langle 0.568, 0.018\rangle\}$	$\{s_5\langle 0.576, 0.004\rangle\}$
$A^{(2)-}$	$\{s_{2.7685}\langle 0.21, 0.008\rangle, s_3\langle 0.126, 0.004\rangle\}$	$\{s_{1.9999}\langle 0.392, 0.018\rangle, s_{2.3112}\langle 0.336, 0.012\rangle\}$	$\{s_3\langle 0.336, 0.024\rangle, s_{3.3413}\langle 0.168, 0.006\rangle\}$	$\{s_{3.283}\langle 0.448, 0.006\rangle, s_{3.7943}\langle 0.32, 0.002\rangle\}$	$\{s_1\langle 0.576, 0.008\rangle\}$

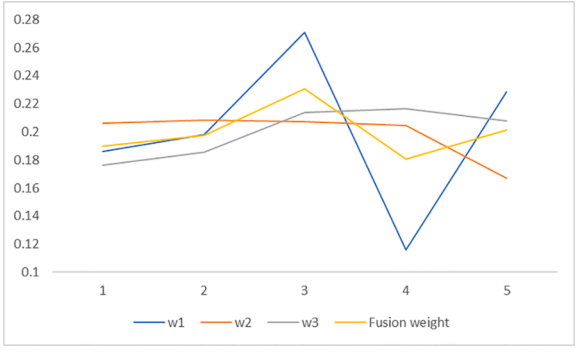


Fig. 5. The criteria weights relation of E1.

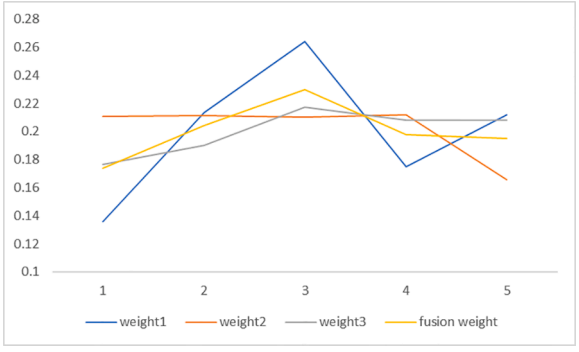


Fig. 6. The criteria weights relation chart of E2.

became a non-epidemic area. However, no country can afford the cost of isolation from the world. At most, it can only temporarily block personnel exchanges with a few countries. When many countries outside of China treat the COVID-19 as the flu, we cannot stand alone. How to adapt to this possible prospect is a major challenge and a subject worth considering.

The good news is that China has sufficient industrial and health strength to provide vaccination for a large-scale population. At the same time, the State Council promised that individuals would not bear any vaccination costs. The principle of vaccination is to make the body produce antibodies naturally by accessing the deactivated COVID-19. Vaccinating can significantly enhance the immunity of this virus. As of June 18, 2021, 31 provinces (autonomous regions and municipalities directly under the central government) and the Xinjiang Production and Construction Corps reported a total of 99.257 million doses of COVID-19 vaccination. Choosing the site of the COVID-19 vaccination center will be related to the public recognition, vaccination efficiency and vaccination success rate, and ultimately affect the essential role of the vaccine. Temporary vaccination centers are generally transformed from some public places, such as gymnasiums, conference halls. The choice of vaccination sites is an MCGDM problem. Suppose that the health administration of District A needs to establish a vaccination site. After preliminary screening, four alternatives $\{A_1, A_2, A_3, A_4\}$are formed, and two DMs $\{E_1, E_2\}$ will evaluate the alternatives from the following five criteria:

(1) Traffic conditions(c_1): Reasonable and efficient traffic condition is conducive to vaccinators to save traffic costs. In addition,

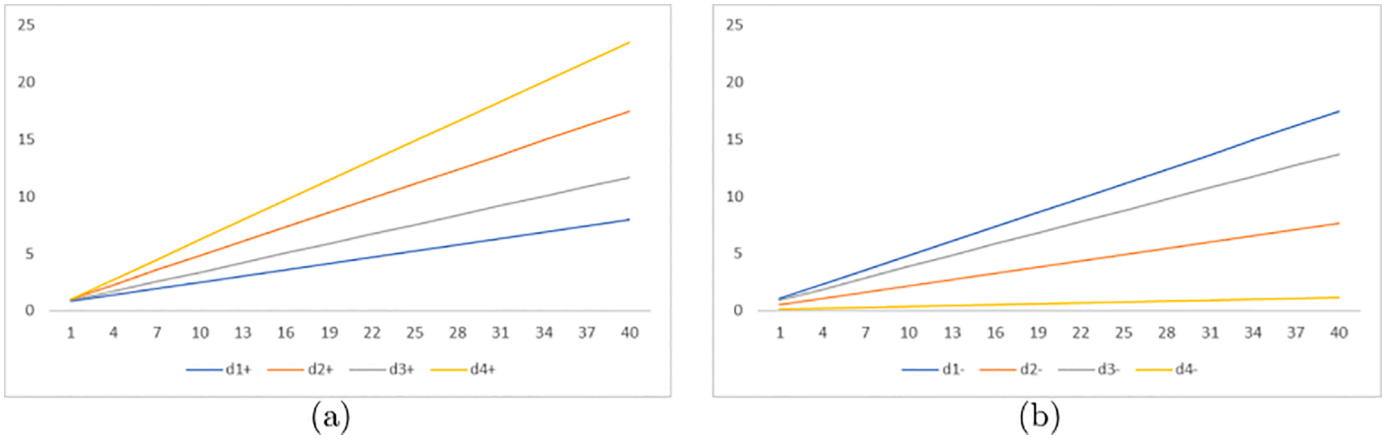


Fig. 7. The influence of w_{dom} on $d_{psy}(A_i, A^+)$ and $d_{psy}(A_i, A^-)$.

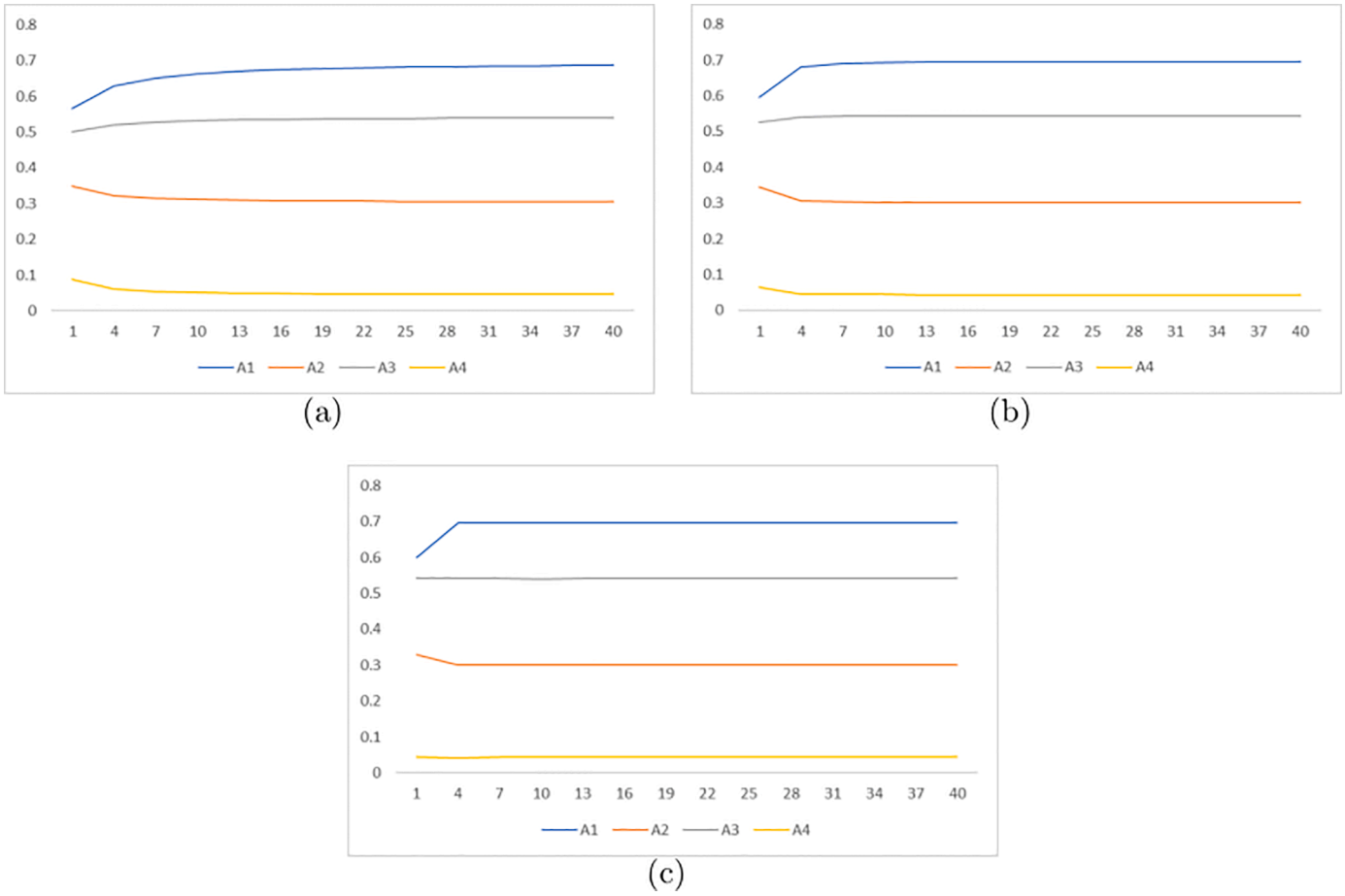


Fig. 8. The ranking result with different norms.

convenient transportation also helps increase vaccine acceptance.

(2) Flow density (c_2): There are many people waiting to be vaccinated, and most of them are vulnerable to infection. To prevent the possibility of the spread of COVID-19, the flow density of the site should not be too high.

(3) Internal facilities(c_3): There should be sufficient material storage and turnover space and good emergency equipment to prevent particular circumstances such as trampling and chaos.

(4) Surrounding supporting construction(c_4): Complete security facilities and drainage systems are needed. Moreover, it should also consider surrounding eating and resting places so that the vaccinators can eat and rest nearby.

(5) Transportation and modification costs(c_5): During the transportation of vaccines, due to factors such as the length of transportation time and temperature, the vaccine's potency is reduced or invalid. The transportation cost should be considered on the premise that the vaccine will not be damaged. At the same time, the cost and complexity of site reconstruction should not be too high. Then, the steps to solve the problem are as follows:

Step1: Collect and preprocess the evaluation information of PFPLTS

Table 11
Final ranking by different ranking aggregation methods.

Ranking aggregation method	Final ranking	Optimal location
Method in [17]	$A_1 \succ A_3 \succ A_2 \succ A_4$	A_1
Method in [2]	$A_1 \succ A_2 \succ A_3 \succ A_4$	A_1
Method in [30]	$A_1 \succ A_3 \succ A_2 \succ A_4$	A_1
Proposed method(1-norm)	$A_1 \succ A_3 \succ A_2 \succ A_4$	A_1
Proposed method(2-norm)	$A_1 \succ A_3 \succ A_2 \succ A_4$	A_1
Proposed method(infinity-norm)	$A_1 \succ A_3 \succ A_2 \succ A_4$	A_1

decision matrices $D^{(ge)} = (PL_{ij}^{(ge)})_{m\times n}$ given by each DM under the same criteria at three different stages. The decision matrices are shown in Tables 1–6, which already meet the standard and do not need to be processed.

Step2: Calculate the objective weight $\theta_1^{(g)}$ of the gth DM at different stages according to the method in Definition 15. The objective time weights of the two DMs are $\theta_1^{(1)} = (0.30, 0.37, 0.33)^T$and $\theta_1^{(2)} = (0.26, 0.37, 0.37)^T$, respectively.

Step3: Use the 1∼9 scale method to construct pairwise judgment matrix $A(a_{ij})_{q\times q}$, which is shown in Table 7. Calculate the subjective weight θ_2 through the AHP method.

The objective time weight is $\theta_2 = (0.28, 0.65, 0.07)^T$.

Step4: To integrate the subjective and objective time weights, we use the nonlinear optimization model to calculate coefficients of the time combination weights by the method in Section 4.1, and then get the time combination weights $\theta^{(1)} = (0.29, 0.51, 0.20)$, $\theta^{(2)} = (0.27, 0.51, 0.22)$. The radar map of the fusion time weight, subjective and objective time weight is shown in Figs. 3 and 4.

Step5: Derive the comprehensive PFPLTS decision matrix $D^{(g)} = (V_{ij}^{(com-g)})_{m\times n}$ of each expert with time weight. The comprehensive time decision matrices are shown in Tables 8 and 9.

Step6: According to the score function in Definition 2,7, the positive and negative ideal solutions of each DM are shown in Table 10.

Step7: Calculate the distance from each alternative to the positive and negative ideal solution, get the positive and negative distance matrix of every DM, respectively.

Step8: According to Eq. (17), the criteria weights vectors with time-varying based on criterion recognition are $\omega_1{}^{(1)} = (0.19, 0.20, 0.23, 0.18, 0.20)^T$ and $\omega_1{}^{(2)} = (0.14, 0.21, 0.26, 0.17, 0.21)^T$.

Step9: According to Eq. (21), the criteria weights vectors with time-varying based on information disorder degree are $\omega_2{}^{(1)} = (0.21, 0.21, 0.21, 0.20, 0.17)^T$ and $\omega_2{}^{(2)} = (0.21, 0.21, 0.21, 0.21, 0.17)^T$.

Step10: According to Eq. (23), the criteria weights vectors with time-varying based on information hesitation degree are $\omega_3{}^{(1)} = (0.18, 0.19, 0.21, 0.22, 0.21)^T$ and $\omega_1{}^{(2)} = (0.18, 0.19, 0.22, 0.21, 0.21)^T$.

Step11: Based on the dual ideal point-vector projection method as Eq. (32). The fusion weight vectors are recorded as $\omega^{(1)} = (0.19, 0.20, 0.23, 0.18, 0.20)^T$ and $\omega^{(2)} = (0.17, 0.20, 0.23, 0.20, 0.19)^T$. And the relation between these weights can be shown in Figs. 5 and 6.

Step12: Calculate the dominance vector v_d and the indifference vectors $(v_{fj})_{n-1}$ by the fusion criteria weight $\omega^{(1)}$ and $\omega^{(2)}$. Then, the basis matrix M is obtained.

Step13: Construct a psychological matrix $H = diag(1, 1, \ldots, 1, w_{dom})$, and let $w_{dom} = 10$.

Step14: Compute the psychological distance $d_{psy}^{(g)}(V_{ij}{}^{(com-g)}, V^+)$ and $d_{psy}^{(g)}(V_{ij}{}^{(com-g)}, V^-)$ with the comprehensive value in Step 5.

Step15: The DMs' weight vector is $\sigma = (0.5, 0.5)^T$. Then, with the weighted average method, the comprehensive psychological distances $d_{psy}^{(g)}(A_i, A^+)$ and $d_{psy}^{(g)}(A_i, A^-)$ are aggregated based on each DM's psychological distance in step 14.

Step16: Calculate the closeness η_i for each alternative A_i. The comprehensive closeness $\eta_i = (0.6631, 0.3120, 0.5324, 0.0513)$.

Step17: In accordance with η_i, the final alternative ranking is $A_1 \succ A_3 \succ A_2 \succ A_4$.

In the case calculation, keep the dominance vector at an appropriate importance and set the psychological index to 10. According to the evaluation information given by the two experts and the criteria weights calculated according to the time weights of different stages, the recommended ranking of the COVID-19 vaccination center selection in the four alternatives is $A_1 \succ A_3 \succ A_2 \succ A_4$. Therefore, it is suggested that the health administration of District A establish a vaccination center at A_1.

6.2. Comparison and discussion

In the former section, we mention that the variable w_{dom} reflects the preference of DMs. Different distance measures and decision-making methods make the Psy-TOPSIS method more robust and flexible. Hence, to further illustrate the effectiveness of the proposed method, the analysis is conducted with different distance measures, decision-making methods and the value of w_{dom} ranges from 1 to 40.

6.2.1. Sensitivity analysis

We draw figures to show the influence of parameter w_{dom} and different distance measures on the alternative rankings. Fig. 7(a)-(b) describes the influence of varying w_{dom} on the distance from each alternative to the positive and negative ideal solutions when 1-norm is used. It is easy to find that both $d_{psy}(A_i, A^+)$ and $d_{psy}(A_i, A^-)$ increase with $w_{dom}$from Fig. 7(a)-(b), and $d_{psy}(A_i, A^+)$ increases faster than $d_{psy}(A_i, A^-)$.

DMs can express their unique personal preferences by providing diverse w_{dom}, which determines the weight between dominant and indifferent directions. Especially, with 1-norm, when $w_{dom} = 1$, which means the preferential relationship has no difference between dominant vector and indifferent vectors. From Fig. 8(a)-(c), we can conclude that the rankings of the alternatives are identical when w_{dom} or the norms are different, yet the closeness gap increases with the increase of w_{dom}, no matter which norm is used. And when w_{dom} is greater than 4, the closeness gradually stabilizes, and the ranking of the alternatives becomes stable, which shows that our method is adequately effective and robust. Hence, DMs can select the value of w_{dom} that will affect the alternative closeness and final rankings to describe their psychological preference of the dominant vector and indifferent vectors.

6.2.2. Comparison with different ranking aggregation methods

To illustrate the effectiveness of the proposed aggregation method, we compare the ranking results of several ranking aggregation methods with ours. As we can see from Table 11, the optimal location obtained by all ranking aggregation methods is A_1, but the final rankings are slightly different by these ranking aggregation methods. The results produced by all these methods are not very different or even basically the same. The most important reason is that we use the same weights and weight method in the calculation, that is, the weight determination method proposed in Section 4 above. The traditional TOPSIS method [17] can effectively avoid data subjectivity and well depict the comprehensive influence of multiple indicators. The advantage of the OWA operator [2] is that it can reflect the importance of the information itself, as well as the importance of the location of the information. The GBWM method has a broad application prospect because of its low time complexity in computation. However, these methods can only determine the only decision alternatives ranking according to the known weight and decision information, that is, time varying factors and decision maker's psychology are not considered.

In the Psy-TOPSIS method, we can fully use the changing and uncertain information to derive the time and criteria weights, which can directly influence the final ranking. At the same time, the DMs can fully

reflect their psychological preference for different alternatives between dominant vector and indifferent vectors, which indicates the effective-ness and superiority of the proposed method to other aggregation methods.

7. Conclusion

PLTS is a valuable technique in linguistic evaluation. However, DMs may be uncertain and self-denying about the given linguistic terms. To reflect the uncertainty and hesitation of DMs, we have extended the traditional PLTS to a new fuzzy linguistic set named PFPLTS. Then, some corresponding basic operations and aggregation operators have been proposed. A linear programming method with minimum deviation and the vector projection method to determine the time and criteria weights are submitted, respectively, which can determine the importance of different stages in dynamic Pythagoras fuzzy probabilistic linguistic MCGDM problems and make full use of the hesitation and uncertainty of the evaluation information. Furthermore, DMs' psychological preference information has been considered. With the new time and criteria weights method, the TOPSIS method with psychological distance has been developed. Finally, the validity and feasibility are verified with a numerical example, site selecting of COVID-19 vaccination center.

In the future study, the proposed method can be applied in other MCGDM problems, such as medical diagnosis and investment decisions combined with forecasting model. In addition, the weight methods and the Psy-TOPSIS method can be further used in other fuzzy sets, such as intuitionistic fuzzy set and so on.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service or company that could be construed as influencing.

Acknowledgement

This work was supported by Natural Science Foundation Project of Chongqing (Grant No. cstc2019jcyj-msxmX0075).

References

[1] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning-part I, Inf. Sci. (Ny) 8 (3) (1975) 199–249.
[2] Q. Pang, H. Wang, Z. Xu, Probabilistic linguistic term sets in multi-attribute group decision making, Inf. Sci. (Ny) 369 (2016) 128–143.
[3] C. Wei, N. Zhao, X. Tang, Operators and comparisons of hesitant fuzzy linguistic term sets, IEEE Trans. Fuzzy Syst. 22 (3) (2014) 575–585.
[4] C. Bai, R. Zhang, L. Qian, et al., Comparisons of probabilistic linguistic term sets for multi-criteria decision making, Knowl. Based Syst. 119 (2017) 284–291.
[5] Alexander, Lutz Brusche, Gaps in academic literature on venture capitalists' decision-making on funding for early-stage, high-tech ventures, Technol. Transf. Entrep. 3 (2) (2016) 82–89.
[6] H. Liu, J. Le, Z Xu, Entropy measures of probabilistic linguistic term sets, Int. J. Comput. Intell. Syst. 11 (1) (2017) 45–87.
[7] D. Liang, A. Kobina, Q. Wei, Grey relational analysis method for probabilistic linguistic multi-criteria group decision-making based on geometric bonferroni mean, Int. J. Fuzzy Syst. 20 (7) (2017) 2234–2244.
[8] X.L. Zhang, A novel probabilistic linguistic approach for large-scale group decision making with incomplete weight information, Int. J. Fuzzy Syst. 20 (7) (2018) 2245–2256.
[9] P.D. Liu, X.L You, Probabilistic linguistic TODIM approach for multiple attribute decision-making, Granular Comput. 2 (4) (2017) 333–342.
[10] H. Liao, L. Jiang, Z. Xu, et al., A linear programming method for multiple criteria decision-making with probabilistic linguistic information, Inf. Sci. (Ny) 415 (2017) 341–355.
[11] X. Zhang, X. Gou, Z. Xu, et al., A projection method for multiple attribute group decision making with probabilistic linguistic term sets, Int. J. Mach. Learn. Cybern. 10 (1) (2018).
[12] X. Wu, H. Liao, An approach to quality function deployment based on probabilistic linguistic term sets and ORESTE method for multi-expert multi-criteria decision making, Inf. Fusion 43 (2018) 13–26.
[13] P. Liu, F. Teng, Probabilistic linguistic TODIM method for selecting products through online product reviews, Inf. Sci. (Ny) 485 (2019) 441–455.
[14] X. Wang, J. Wang, H. Zhang, Distance-based multicriteria group decision-making approach with probabilistic linguistic term sets[J], Expert Systems 36 (2) (2019) 1–18.
[15] X. Wu, Liao, A consensus-based probabilistic linguistic gained and lost dominance score method, Eur. J. Oper. Res. 272 (3) (2018).
[16] M. Lin, Z. Xu, et al., Multi-attribute group decision-making under probabilistic uncertain linguistic environment, J. Oper. Res. Soc. (2018).
[17] C. Bai, Z. Ren, S. Shuang, et al., Interval-valued probabilistic linguistic term sets in multi-criteria group decision making, Int. J. Intell. Syst. 33 (6) (2018) 1301–1321.
[18] J. Chen, W. Hai, Z Xu, Uncertain probabilistic linguistic term sets in group decision making, Int. J. Fuzzy Syst. 21 (3) (2019).
[19] J. Rezaei, Best-worst multi-criteria decision-making method, Omega (Westport) 53 (2015) 49–57, jun.
[20] Zhou L., Huang P., Chi S., et al. Structural health monitoring of offshore wind power structures based on genetic algorithm optimization and uncertain analytic hierarchy process. 2020, 218.
[21] N. Caglayan, A. Yesil, O. Kabak, et al., A decision-making approach for assignment of ecosystem services to forest management units: a Case Study in Northwest Turkey, Ecol. Indic. (2020).
[22] J. Wang, G. Wei, C. Wei, et al., Maximizing deviation method for multiple attribute decision making under q-rung orthopair fuzzy environment, Def. Technol. (2020).
[23] Q. Cheng, The stucture entopy weight method to confrm of evaluating index, Syst. Eng. Theory Pract. 30 (7) (2010) 1225–1228.
[24] P.D. Liu, Method for multi-attribute decision-making under risk with the uncertain linguistic variables based on prospect theory, Control Decis. 26 (2011) 893–897.
[25] J. Hu, Y. Liu, Dynamic stochastic multi-criteria decision-making method based on cumulative prospect theory and set pair analysis, Syst. Eng. Procedia 1 (2011) 432–439.
[26] N. Berkowitsch, B. Scheibehenne, J. Rieskamp, et al., A generalized distance function for preferential choices, Br. J. Math. Stat. Psychol. 68 (2) (2015).
[27] X. Gou, Z. Xu, Novel basic operational laws for linguistic terms, hesitant fuzzy linguistic term sets and probabilistic linguistic term sets, Inform. Sci. 372 (2016) 407–427.
[28] R.R. Yager, A.M Abbasov, Pythagorean membership grades, complex numbers, and decision making, Int. J. Intell. Syst. 28 (5) (2013) 436–452.
[29] R.R Yager, Pythagorean membership grades in multicriteria decision making, IEEE Trans. Fuzzy Syst. 22 (4) (2014) 958–965.
[30] G. Haseli, R. Sheikh, J. Wang, H. Tomaskova, E.B. Tirkolaee, A novel approach for Group Decision Making Based on the Best–Worst Method (G-BWM): application to supply chain management, Mathematics 9 (16) (2021) 1881.

Epileptic seizure identification in EEG signals using DWT, ANN and sequential window algorithm

Pravat Kumar Subudhi, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, pk.subudhi11@outlook.com*

S. Sivasakthiselvan, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, s.sivasakthiselavan@live.com*

Abhishek Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, abhishekdas2256@gmail.com*

Anita Subudhi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, anitasubudhi89@gmail.com*

Keywords:
Artificial neural network
Discrete wavelet transform
EEG signal classification
False positive rate
Seizure detection
Sequential window algorithm

A B S T R A C T

A patient-specific novel systematic methodology is described in this study for automatic seizure detection from raw electroencephalogram (EEG) signals. Filtering process by means of band-pass finite impulse response (FIR) filter with the frequency range of 0.5–40 Hz is implemented at the outset to eliminate different artifacts and noises mixed with raw EEG signals. As EEGs are highly non-linear and non-stationary signals in nature, discrete wavelet transform (DWT) is then used to analyze the signals in time-frequency domain. DWT with four level decomposition is performed using db6 mother wavelet for feature extraction. A new feature set, composed of eleven non-linear statistical features extracted from each sub-bands resulting from due to wavelet decomposition, is then fed to the input of artificial neural network (ANN) to classify the signal accurately. Finally, a novel algorithm named sequential window algorithm is carried out to improve the classification performance. 99.44% mean classification accuracy, 80.66% average sensitivity, 4.12 s mean latency and 0.2% average false positive rate (FPR) are achieved in this study. This study successfully reduces the latency time with more accuracy and significantly low FPR.

1. Introduction

Epilepsy is a very much concerning issue as it may lead the patient to serious injuries and death [1–3]. Around 0.6–0.8% of world's population, close to 50 million people in the world, are suffering from epilepsy, nearly 80% of them live in low- and middle-income countries, as stated by World Health Organization (WHO) [4]. If good diagnosis and treatment can be ensured, then it is anticipated that maximum of 70% epilepsy persons can live without seizure. But there is a lack of sufficient treatment for the rest of the epilepsy persons [2,5]. Identification of epileptic seizure is really critical.

There are some drawbacks for detection of epileptic seizure in traditional clinical operation. Firstly, the neurophysiologist is to diagnosis a extensive size of electroencephalogram (EEG) signals recorded visually by monitoring continuously for long term which kills time, is sometimes boring and might lead to escalation the probability of error. Secondly, bio-signals are exceedingly subjective, so there is a high chance of disagreement among the physicians during the analysis of the seizure signals. Hence, a method which can diagnosis epileptic seizure automatically with trustworthy accuracy has great importance [6,7].

The EEG signal is a good diagnostic tool for automatic identification of epilepsy. The characteristics of this type of signal are complex, non-linear and non-stationary. So, it is convenient to analyze this signal in time-frequency domain, and, wavelet transform (WT) is often used for this purpose [3,6,8–10] to extract the features. After feature extraction an expert classifier is needed. Different types of neural networks, for instance, artificial neural network (ANN), probabilistic neural network, wavelet neural network, recurrent neural network and convolutional neural network have been used extensively to detect epileptic seizure due to its competency of finding the relationship between rapid variations of EEG recordings, characteristics of fault tolerance, enormous parallel processing ability and adaptive learning competency [10–13].

Recent works on automatic detection of seizure from EEG signals have concentrated on patient-specific predictors, where a classifier is trained and tested on the records of the same person [14–16]. Some recent studies on seizure detection using the Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) scalp EEG database [17,18] are seen in literatures. In 2014, Kiranyaz et al. [14] recommended a patient-specific seizure detection model. They carried out the experiment on 21 patients excluding the data for patients # 6, 12 and 16 of the database with common 18 channels and collected less than 2 min of seizure data and 24 min of non-seizure data on average from

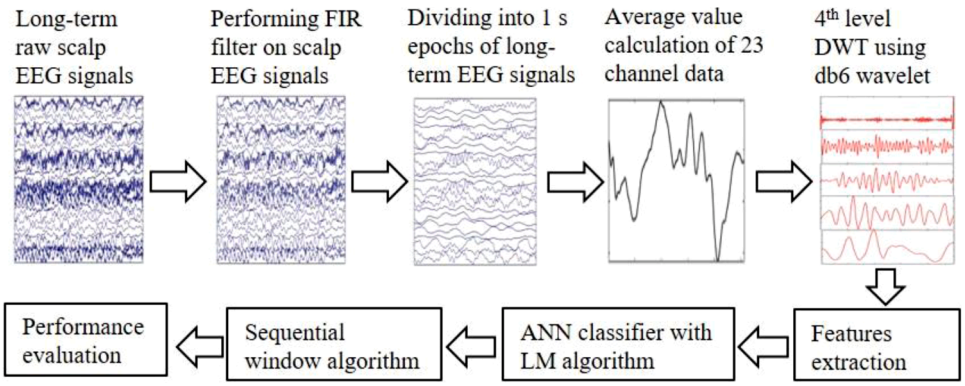


Fig. 1. Work flow diagram of whole study.

each patient. They divided the signal into non-overlapping 1 s segments. To remove physiological and electrical noise and artifacts, they used linear phase band pass FIR filter between the frequency range of 0.5–30 Hz with Parks-McClellan algorithm. They extracted morphological, time domain, frequency domain, time-frequency domain and non-linear mel frequency cepstral coefficient (MFCC) features. Multi-dimensional particle swarm optimization (PSO), collective network of binary classifiers (CNBC) ensemble and support vector machine (SVM) classifiers were used to measure performance.

In 2015, Fergus et al. [15] also proposed a patient-specific seizure detection model in which they used common 23 channels record for 24 patients, but only 171 seizure files out of 198 were used. They formed 60 s data blocks for each ictal files and 171 non-seizure data blocks extracted randomly from non-seizure signals. In pre-processing stage, second order Butterworth band pass filters were used in between delta, theta, alpha and beta frequency ranges. They extracted, in total 20 features, in both time domain and frequency domain; and Fourier transform was performed to extract frequency domain features. Different classifiers, such as, linear discriminant, quadratic discriminant, uncorrelated normal density-based classifier, polynomial, logistic, *k*-class nearest neighbor classifier (KNNC), decision tree, Parzen and SVM classifiers were used to classify the signal.

In 2018, Harpale and Bairagi [19] proposed a non-patient specific seizure detection model. They used 22 patients' data, recorded using common 23 channels, for study among them first 6 patients' data were used for testing, 4 patients' data were used for validation, and 12 patients' data were used for testing. Independent component analysis (ICA) technique was used to remove artifacts from the raw EEG signal. Using continuous wavelet transform (CWT), they extracted time-frequency features: mean, variance, coefficient of variation, root mean square (RMS), kurtosis, power spectral density (PSD) from wavelet coefficients and averaged the features value to form final feature value. This final feature was divided into three category: normal signal, pre-seizure signal (30 s before the seizure signal) and seizure signal. The above-mentioned time-frequency features were as the f1 feature vector of fuzzy classifier and f2 feature set (mean, RMS, standard deviation and PSD) extracted by pattern adaptable WT constructed from single channel seizure pattern from training samples.

In 2019, Jiang et al. [16] suggested a patient-specific onset seizure detection model using redundancy removed dual-tree discrete wavelet transform (DWT) and SVM classifier. They used 181 out of 198 seizures which has common 23 channels and divided the recorded data into 30 s segments. For each seizure ictal case, they used only one 30 s segment and they randomly down sampled the inter-ictal data to 1:3 ratio between ictal and inter-ictal data. Energy and modified multi-scale entropy features have been extracted. They extracted total 2254 dimensional feature vectors for each segment. Therefore, these features contain huge amount of redundant information, so, auto-weighted feature selection via global redundancy minimization (AGRM) algorithm has been used to select the feature as well as to remove redundant information. Finally, leading 50 features were selected for training and testing. Then the SVM classifier was used to classify the data.

In 2019, Mansouri et al. [20] carried out a non-patient specific on the same database but they only concentrated on the age range between 4

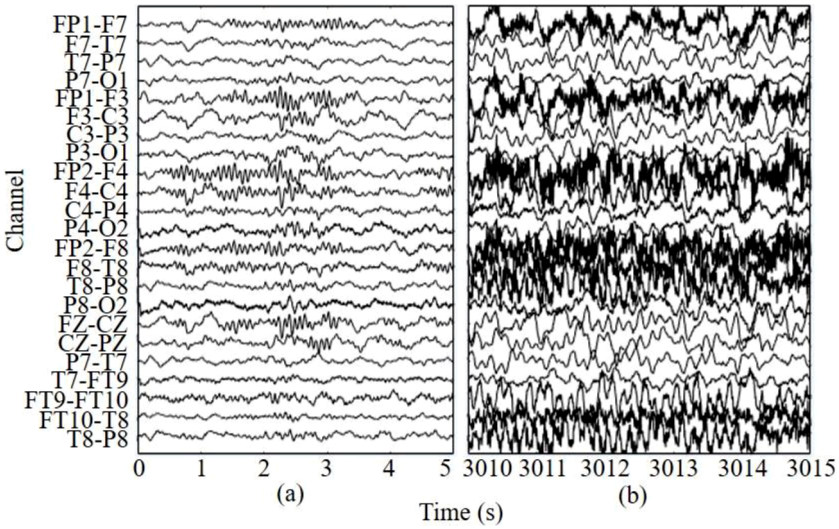


Fig. 2. Graphical representation of EEG signals (a) seizure free period and (b) seizure period.

Table 1
Description of datasets including training and testing datasets.

PN[1]/ G[1]	Age	Training			Testing		
		Record name	SN[1]	SD[1]	Record name	SN[1]	SD[1]
1/F	11	chb01_03, chb01_04, chb01_15, chb01_16	4	158	chb01_18, chb01_21, chb01_26	3	284
2/M	11	chb02_16, chb02_19	2	91	chb02_16+	1	81
3/F	14	chb03_01, chb03_02, chb03_03, chb03_04, chb03_34	5	285	chb03_35, chb03_36	2	117
4/M	22	chb04_05, chb04_08	2	160	chb04_28	2	218
5/F	7	chb05_06, chb05_13, chb05_16	3	321	chb05_17, chb05_22	2	237
6/F	15	chb06_01, chb06_04, chb06_09	6	100	chb06_10, chb06_13, chb06_18, chb06_24	4	53
7/F	14.5	chb07_12, chb07_19	2	229	chb07_13	1	96
8/M	3.5	chb08_02, chb08_05, chb08_21	3	625	chb08_11, chb08_13	2	294
9/F	10	chb09_06, chb09_19	2	126	chb09_08	2	150
10/M	3	chb10_12, chb10_20, chb10_27, chb10_30	4	228	chb10_31, chb10_38, chb10_89	3	219
11/F	12	chb11_99	1	752	chb11_82, chb11_92	2	54
12/F	2	chb12_06, chb12_08, chb12_09, chb12_33, chb12_38	15	475	chb12_10, chb12_11, chb12_23, chb12_36, chb12_42	12	514
13/F	3	chb13_19, chb13_21, chb13_40, chb13_55, chb13_59	7	309	chb13_58, chb13_60, chb13_62	5	226
14/F	9	chb14_03, chb14_04, chb14_06, chb14_11	5	111	chb14_17, chb14_18, chb14_27	3	58
15/M	16	chb15_06, chb15_10, chb15_15, chb15_17, chb15_20, chb15_22, chb15_28, chb15_31, chb15_40, chb15_46	11	1267	chb15_49, chb15_52, chb15_54, chb15_62	9	725
16/F	7	chb16_10, chb16_11, chb16_14, chb16_16	4	38	chb16_17	4	31
17/F	12	chb17a_03, chb17a_04	2	205	chb17b_63	1	88
18/F	18	chb18_29, chb18_30, chb18_31	3	148	chb18_32, chb18_35, chb18_36	3	169
19/F	19	chb19_28, chb19_29	2	155	chb19_30	1	81
20/F	6	chb20_12, chb20_13, chb20_14	4	136	chb20_15, chb20_16, chb20_68	4	168
21/F	13	chb21_19, chb21_20	2	106	chb21_21, chb21_22	2	93
22/F	9	chb22_20, chb22_25	2	132	chb22_38	1	72
23/F	6	chb23_09	4	244	chb23_06, chb23_08	3	180
24/-	-	chb24_01, chb24_03, chb24_04, chb24_06, chb24_11, chb24_14	10	303	chb24_07, chb24_09, chb24_13, chb24_15, chb24_17, chb24_21	6	208

[1]patient number (PN), gender (G), seizure number (SN) and seizure duration (SD).

and 21 years old, thus, they excluded patients # 4, 6, 10, 12 and 13 except patient # 8 (3.5 years old). They divided signal into 10 s epochs with 5 s overlapping and pre-processing was performed to remove DC offset and 60 Hz power line noise for each epochs. Using fast Fourier transform (FFT), they decomposed each epoch into five frequency bands and calculated power in band of interest (PBI) from the FFT coefficients for each epoch. They computed an adaptive threshold value based on PBI values in three blocks of epochs, to determine whether the current epoch is ictal or not. If PBI value is greater than the adaptive threshold value, current epoch is ictal. After that they designed a distance network (DN) to determine spreading technique of seizure in brain, but to locate focus of a seizure in brain they designed correlation network (CN).

In this research, we have developed a patient-specific computer-aided model which will detect epileptic seizure automatically and more accurately. FIR filter, DWT, ANN with suitable backpropagation algorithm and newly proposed sequential window algorithm (SWA) have chronologically been employed to develop the model. Performance is duly measured using statistical parameters.

2. Materials and methods

The work flow of this study is shown in Fig. 1, and, then the steps are described sequentially as follows.

2.1. Description of EEG data

The experimental database was collected from the Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) scalp EEG database [17,18]. The database consists of EEG recordings with intractable seizures recorded from 22 pediatric patients who are 5 males and 17 females. Sampling rate of all signals is 256 samples per second with a resolution of 16 bits. For recording the signals international 10–20 system of EEG electrode positions and nomenclature were used. Different channel configurations were used to record the signals, but in most cases 23 channels configuration was used. Most records are 1 h long but some records are also 2, 3 and 4 h long. Records that contain seizure are called seizure records and that do not contain seizure are called non-seizure records. There are total 676 records in the database among them 141 seizure records. In 141 seizure records, there are 198 seizures and total seizure duration is 11621 s. A typical non-seizure signals from 1 s to 5 s and seizure signals from 3010 s to 3015 s are illustrated in Fig. 2 of patient # 1 from record # chb01_03 for 23 channels.

In this study, we have used only the seizure records that were recorded using these common 23 channels: FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ, CZ-PZ, P7-T7, T7-FT9, FT9-FT10, FT10-T8

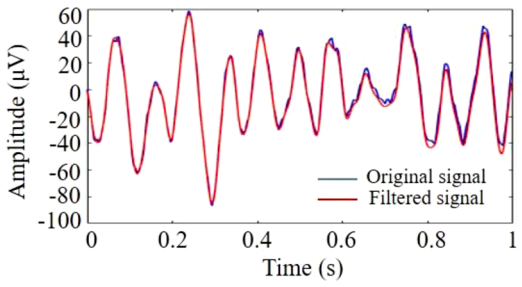


Fig. 3. Original and filtered EEG signals.

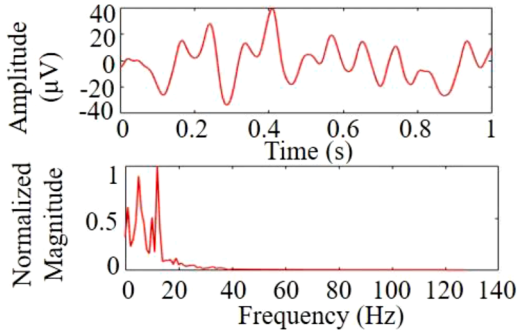


Fig. 4. Average value of the 23 channels data after dividing into 1 s epoch: (a) 1 s epoch and (b) its frequency spectrum.

and T8-P8. Among the 141 seizure records, 4 records (chb12_27, chb12_28, chb12_29 and chb16_18) were recorded using different channel configurations. Thus, we excluded these 4 records. In total, 137 seizure records were used where 80 seizure records (duration 5565 s) and 57 seizure records (duration 5555 s) were used for testing and training, respectively. Table 1 gives a brief overview of database and training and testing data which were used in the study.

2.2. Preprocessing the EEG signals

Different types of artifacts, such as: eye blinks, muscle movements, movement of EEG sensors, power line interference, environment, etc., may be incorporated with EEG signals while recording. Hence, it is necessary to remove these artifacts from the raw EEG signals for extracting the actual features from the signals. High pass, low pass, band pass and notch filters are often used to remove these artifacts based on the artifacts' frequency range. The frequency ranges of subjective and electrical artifacts hardly exceed 50 Hz [21], instrument artifacts also rarely exceed 30–40 Hz [21], power line interferences remain in 50/60 Hz [22,23]. Interestingly, most brain activity occurs in between 3 and 29 Hz [23,24]. Additionally, below 0.5 Hz, no cerebral activity occurs [21]. Actually, signals below 0.5 Hz indicates motion or other electrical activity [15]. Consequently, in this study, firstly, filtering was performed using band-pass finite impulse response (FIR) filter where the lower cut-off frequency is 0.5 Hz and higher cut-off frequency is 40 Hz. For example: Fig. 3 is depicted for both raw and filtered EEG signals for 1 s data and this 1 s data is collected from 2 s to 3 s data of FP1-F7 channel for patient # 1 and record # chb01_03. Secondly, the filtered EEG signals was then divided into 1 s epoch for all 23 channels. Finally, average value was calculated for all 23 channels data, as shown in Fig. 4.

2.3. Discrete wavelet transform (DWT)

WT decomposes a signal into a set of coefficients which are known as wavelet coefficients. It provides precise frequency and time information at low and high frequency due to its suitability of using variable size windows [6]. It is well-known that DWT, compared to CWT, offers a more flexible time-frequency window function, which narrows when observing high frequency information and widens when analyzing low frequency resolution. It is implemented by decomposing the signal into coarse approximation and detail information by using successive low and high pass filtering [7]. The low pass filter produces coarse approximation coefficients, whereas the high pass filter outputs the detail coefficients. The size of the approximation coefficients and detail coefficients decreases by a factor of 2 at each successive decomposition. At each step frequency resolution is doubled and time resolution is halved by down sampling. Selecting the appropriate number of decomposition level is important for DWT. For the EEG signal analysis, the number of decomposition levels can be determined directly, based on their dominant frequency components and the number of levels is chosen in such a way that those parts of the signals which correlate well with the frequencies required for the classification of EEG signals are retained in the wavelet coefficients [1,13,25].

Table 2
Decomposition levels coefficients and their frequency range.

Level #	Coefficient Vector	Frequency Range (Hz)
1	D1	64–128
2	D2	32–64
3	D3	16–32
4	D4	8–16
4	A4	0–8

In this work, four level wavelet decomposition has been performed by using db6 mother wavelet function. The results of four level decomposition are four detail coefficients D1, D2, D3 and D4, and one approximate coefficient A4, shown in Table 2. The graphical demonstrations are also shown in Fig. 5 for understanding how the wavelet decomposition works. So, D1, D2, D3, D4 and A4 coefficients were used for extracting the features.

2.4. Feature extraction

The following eleven significant features, such as, minimum value, maximum value, mean, variance, energy, log entropy energy, fractal dimension, kurtosis, skewness, median value, inter quartile range, have been adopted for more accurately classifying the epileptic seizures.

Minimum value (min):	$\min = X_{min}$	(1)
Maximum value (max):	$\max = X_{max}$	(2)
Mean (μ):	$\mu = \frac{\sum_{i=0}^{N-1} X_i}{N-1}$	(3)
Variance (σ^2):	$\sigma^2 = \frac{\sum_{i=0}^{N-1} \lvert X_i - \mu \rvert^2}{N-1}$	(4)
Energy (E):	$E = \sum_{i=0}^{N-1} \lvert X_i \rvert^2$	(5)
Log entropy energy (E_{log}):	$E_{log} = \sum_{i=0}^{N-1} \log(X_i^2)$	(6)
Fractal dimension (FD):	$FD = \frac{\sum_{i=0}^{N-1} \lvert X_{i+1} - X_i \rvert}{N-1}$	(7)
Kurtosis (K):	$K = \frac{Fourth\ moment}{Second\ mooment^2}$	(8)
Second Moment (SM):	$SM = \frac{\sum_{i=0}^{N-1} (X_i - \mu)^2}{N-1}$	(9)
Third Moment (TM):	$TM = \frac{\sum_{i=0}^{N-1} (X_i - \mu)^3}{N-1}$	(10)
Fourth moment (FM):	$FM = \frac{\sum_{i=0}^{N-1} (X_i - \mu)^4}{N-1}$	(11)
Skewness (S):	$S = \frac{Third\ moment}{Second\ mooment^{3/2}}$	(12)
Median value (M):	$M = X_{\frac{N+1}{2}}$ when N is odd $M = \frac{1}{2}\left[X_{\frac{N}{2}} + X_{\left(\frac{N}{2}+1\right)}\right]$ when N is even	(13)
Inter quartile range (IQR):	$IQR = Q_3 - Q_1$	(14)

Here, X_{min}, X_{max}, X_i and N are minimum value, maximum value, the

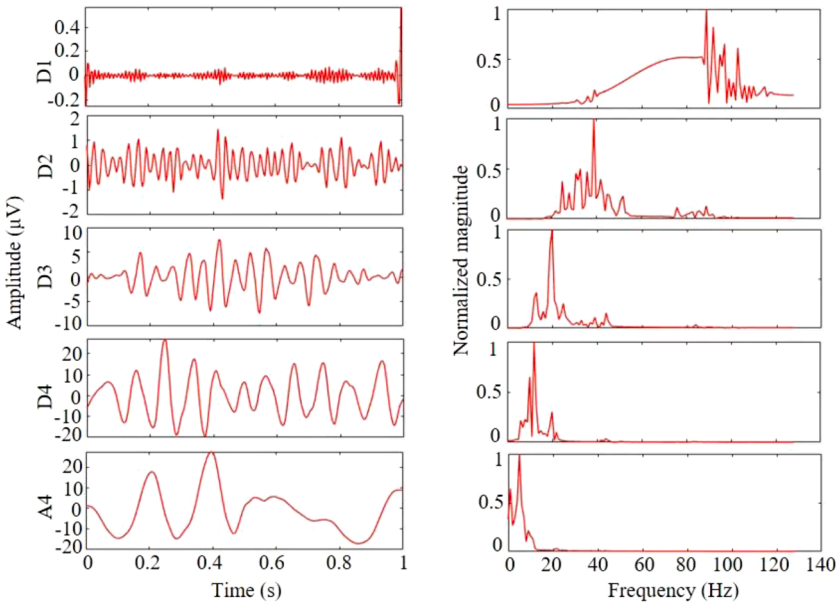


Fig. 5. Graphical demonstration of 4th level wavelet decomposition using db6 wavelet.

*i*th sample value and total samples' number of a dataset X, respectively. Again, Q_1 and Q_3 are defined as the first and third quartile, respectively.

These eleven statistical features are extracted from each decomposition coefficient. Thus, 55 features, from D1, D2, D3, D4 and A4 sub-bands, are extracted in total.

2.5. Identification

Now, we are in need of an expert classifier to classify the signal based on the extracted features from the decomposed coefficients. There exist several classifiers, such as SVM, KNN, ANN and so on. The ANN classifier is considered in this work because it's fast operations, easy implementation and ability to learn and generalize. Its working process is similar to human cognition process. ANN is an expert machine learning information processing system made up of many computational neural units, which are called nodes, and these are inter-connected. The network is trained by adjusting the weights, which is a links between the connecting nodes, and biases based on cost function to produce the desired output. To train the network, training algorithms is an integral part for a model development. An appropriate topology may still fail to give a better model, unless trained by a suitable training algorithm. A good training algorithm will shorten the training time, while achieving a better accuracy. In fact, training process is an important characteristic of the ANNs, whereby representative examples of the knowledge are iteratively presented to the network, so that it can integrate this knowledge within its structure [1,3,5,7,9,12,13]. It is reported that Levenberg-Marquardt (LM) algorithm is the fastest method for training moderate-sized feed forward neural networks (up to several hundred weights). LM algorithm combines the advantages of gradient-descent (GD) and Gauss-Newton method [26,27].

In this research, we designed ANN with 55 input nodes because we have extracted 55 feature vectors from five sub-bands and a single hidden layer with 55 hidden nodes. To select the hidden layer nodes we have optimized our networks with different number of hidden nodes. The best classification accuracy is found using 55 hidden nodes. Tangent sigmoid activation function is used at the hidden layer output. For classification problems, target output 0 is set for seizure free datasets and 1 is set for seizure dataset. Our network have been trained using the training dataset until the best performance is achieved. After that by using the trained network, testing dataset have been classified, and the performance is measured by the statistical parameters.

2.6. Sequential window algorithm (SWA)

After classification using ANN of the epochs, due to different artifacts and noise, some epochs classified incorrectly. For instance: consider the output from ANN for record # chb01_18 of patient # 1, as shown in Table 3, where epochs from 1724 to 1799s exhibits irregular patterns, such as, four consecutive epochs are seizure (1724–1727s) after that two consecutive epochs (1728–1729s) are non-seizure then one seizure epoch 1730s and so on, though according to the database description the seizure starts from 1720s and lasts 1810s.

Actually this result is not desirable and this non-linear characteristics increases false detection rate (FDR). Therefore, analyzing all of the 24

Table 3
Output from the ANN classifier for record chb01_18 of patient # 1.

Epochs	1	2	3…	1723	1724	1725	1726	1727	1728
Output	0	0	0	0	1	1	1	1	0
Epochs	1729	1730	1731	1732…	1749	1750	1751	1752	1753…
Output	0	1	0	1	1	0	1	0	1
Epochs	1764	1765	1766	1767	1768	1769	1770	1771	1772
Output	1	0	1	1	1	0	1	0	1
Epochs	1773	1774	1775	1776	1777	1778	1779	1780	1781
Output	1	1	1	0	1	1	1	1	0
Epochs	1782	1783	1784…	1795	1796	1797	1798	1799…	2204
Output	1	0	1	1	1	0	1	0	0
Epochs	2205	2206	2233	2234	2235…	3493	3495	3496…	3600
Output	1	0	0	1	0	0	1	0	0

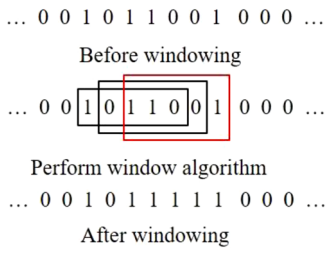


Fig. 6. Basics of performing sliding window algorithm.

patients tested records, we developed an algorithm which can improve FDR as well as sensitivity and specificity. In addition to, making false positive rate (FPR) rate zero or close to zero was main target. This algorithm is accomplished in eleven steps and there are total nine sliding windows which function one after another. Each step window perform during the entire duration of the record sliding after 1 s and update detection rate, shown in Fig. 6, after that the next step window will perform and so on.

Windows length was chosen based on the patterns. Firstly, at step 3, to improve the following patterns '1 0 1 0 1′ or '1 1 0 0 1′ or '1 0 0 1 1′, we made 5 s sliding window and updated the patterns to '1 1 1 1 1′. At step 4, a 6 s sliding window was made to update the following patterns '1 0 1 1 0 1′ or '1 1 0 1 0 1′ or '1 0 1 0 1 1′ or '1 0 0 1 1 1′ or '1 1 0 0 1 1′ or '1 1 1 0 0 1′, and converted the patterns to '1 1 1 1 1 1′. At step 5, to remove the single epoch which has been detected as seizure but there is no seizure epoch either before or after the detected seizure epoch i.e. the patterns '1 0′ or '0 1′, we made 2 s sliding window for selecting the pattern '1 1′ and kept the pattern unchanged otherwise for patterns '1 0′ or '0 1′, we converted the patterns to '0 0′. At step 6, again we made 5 s sliding window for selecting the patterns '1 0 1 1 1′ or '1 1 0 1 1′ or '1 1 1 0 1′ and updated the patterns to '1 1 1 1 1′. At step 7, for selecting the following patterns '1 0 0 0 1 1′ or '1 1 0 0 0 1 1′ or '1 1 1 0 0 0 1′ or '1 0 1 0 0 1 1′ or '1 1 0 1 0 0 1′ or '1 1 0 0 1 0 1′ or '1 0 1 0 1 0 1′ or '1 0 0 1 1 0 1′ or '1 0 1 1 0 0 1′ or '1 0 0 1 0 1 1′, we made 7 s sliding window and converted the patterns to '1 1 1 1 1 1 1′. At step 8, we developed 8 s sliding window and select those patterns whose first and last epoch value must be '1′ and contains total 4 epochs whose value is '1′, for instance: window '1 1 0 0 0 0 1 1′. Then converted all epochs value of selected windows to '1 1 1 1 1 1 1 1′. At step 9, to remove the consecutive four epochs which have been detected as seizure, we made 5 s sliding window for selecting the pattern '1 1 1 1 1′ and kept the pattern unchanged otherwise for patterns '1 1 1 1 0′ or '0 1 1 1 1′, we converted the patterns to '0 0 0 0 0′. At step 10, a 11 s sliding window was made to select those patterns whose first and last epoch value must be '1′ and contains total 3 epochs whose value is '1′, for instance: window '1 0 0 0 0 0 0 0 0 0 1′ and then converted the selected patterns to '1 1 1 1 1 1 1 1 1 1 1′. Finally, at step 11, to remove the consecutive nine epochs which have been detected as seizure, we developed a 10 s sliding window for selecting the pattern '1 1 1 1 1 1 1 1 1 1′ and kept the pattern unchanged otherwise for patterns '1 1 1 1 1 1 1 1 1 0′ or '0 1 1 1 1 1 1 1 1 1′ we converted the patterns to '0 0 0 0 0 0 0 0 0 0′.

Algorithm:

(1) Calculate the length of the testing record's epochs and put the value in a variable called *r*.
(2) Create an array namely "outAnn" whose size is ($r \times 1$) and put the detected value of all the epochs using ANN classifier.
(3) Make 5 s sliding window using following steps:
  a Initialize a variable, $c = 0$ to count the number of 1 s.
  b Make a loop from $i = 1$ to (r-4) and increase it every time by 1.
  c Check if outAnn(i,1) = 1 and outAnn(i + 4,1) = 1, if yes then
    i Make a loop j = i to (i + 4) and increase it every time by 1.
    ii Check if outAnn(j,1) = 1, if yes then $c = c + 1$.
    iii End the loop.
    iv Check if $c = 3$, if yes then
      1 Make a loop k = i to (i + 3) and increase every time by 1.
      2 Set outAnn(k,1) = 1.
      3 End the loop.
  d Set $c = 0$.
  e End the loop.
(4) Make 6 s sliding window using following steps:
  a Set $c = 0$.
  b Make a loop from $i = 1$ to (r-5) and increase it every time by 1.
  c Check if outAnn(i,1) = 1 and outAnn(i + 5,1) = 1, if yes then
    i Make a loop j = i to (i + 5) and increase it every time by 1.
    ii Check if outAnn(j,1) = 1, if yes then $c = c + 1$.
    iii End the loop.
    iv Check if $c = 4$, if yes then
      1 Make a loop k = i to (i + 4) and increase every time by 1.
      2 Set outAnn(k,1) = 1.
      3 End the loop.
  d Set $c = 0$.
  e End the loop.
(5) Make 2 s sliding window using following steps:
  a Initialize an another array namely "outAnn2″ whose size is ($r \times 1$) and set outAnn2(1:r,1) = 0 and also set $c = 0$.
  b Make a loop from $i = 1$ to (r-1) and increase it every time by 1.
  c Check if outAnn(i,1) = 1 and outAnn(i + 1,1) = 1, if yes then
    i Make a loop j = i to (i + 1) and increase it every time by 1.
    ii Check if outAnn(j,1) = 1, if yes then $c = c + 1$.
    iii End the loop.
    iv Check if $c = 2$, if yes then
      1 Make a loop k = i to (i + 1) and increase every time by 1.
      2 Set outAnn2(k,1) = 1.
      3 End the loop.
  d Set $c = 0$.
  e End the loop.
(6) Again make 5 s sliding window using following steps:
  a Set $c = 0$.
  b Make a loop from $i = 1$ to (r-4) and increase it every time by 1.
  c Check if outAnn2(i,1) = 1 and outAnn2(i + 4,1) = 1, if yes then
    i Make a loop j = i to (i + 4) and increase it every time by 1.
    ii Check if outAnn2(j,1) = 1, if yes then $c = c + 1$.
    iii End the loop.
    iv Check if $c = 4$, if yes then
      1 Make a loop k = i to (i + 3) and increase every time by 1.
      2 Set outAnn2(k,1) = 1.
      3 End the loop.
  d Set $c = 0$.
  e End the loop.
(7) Make 7 s sliding window using following steps:
  a Set $c = 0$.
  b Make a loop from $i = 1$ to (r-6) and increase it every time by 1.
  c Check if outAnn2(i,1) = 1 and outAnn2(i + 4,1) = 1, if yes then
    i Make a loop j = i to (i + 6) and increase it every time by 1.
    ii Check if outAnn2(j,1) = 1, if yes then $c = c + 1$.
    iii End the loop.
    iv Check if $c = 4$, if yes then
      1 Make a loop k = i to (i + 5) and increase every time by 1.
      2 Set outAnn2(k,1) = 1.
      3 End the loop.
  d Set $c = 0$.
  e End the loop.
(8) Make 8 s sliding window using following steps:
  a Set $c = 0$.
  b Make a loop from $i = 1$ to (r-7) and increase it every time by 1.
  c Check if outAnn2(i,1) = 1 and outAnn2(i + 4,1) = 1, if yes then
    i Make a loop j = i to (i + 7) and increase it every time by 1.
    ii Check if outAnn2(j,1) = 1, if yes then $c = c + 1$.
    iii End the loop.

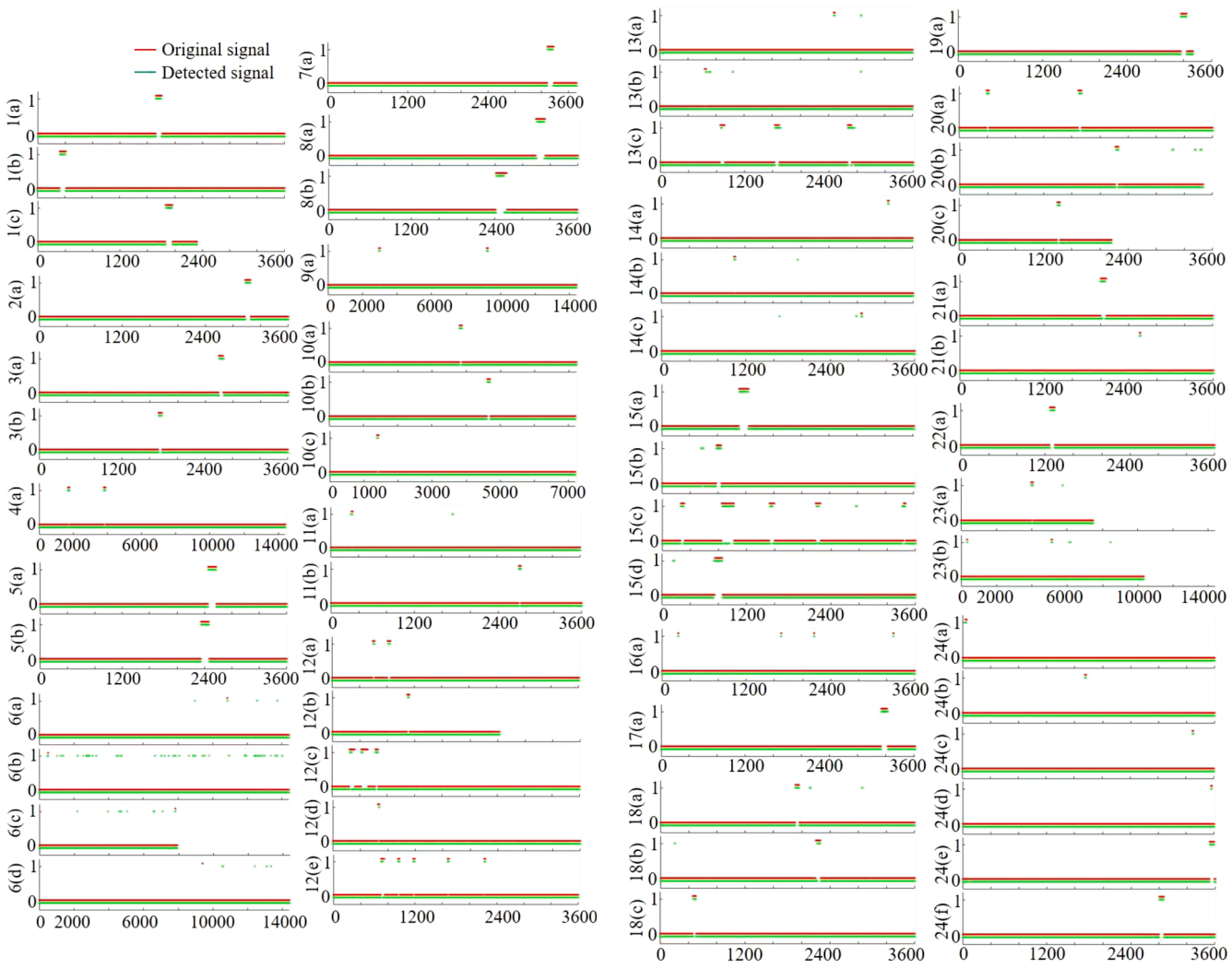


Fig. 7. Patient-specific results for CHB-MIT database where 1(a) chb01_18; 1(b) chb01_21; 1(c) chb01_26; 2(a) chb02_16+; 3(a) chb03_35; 3(b) chb03_36; 4(a) chb04_28; 5(a) chb05_17; 5(b) chb05_22; 6(a) chb06_10; 6(b) chb06_13; 6(c) chb06_18; 6(d) chb06_24; 7(a) chb07_13; 8(a) chb08_11; 8(b) chb08_13; 9(a) chb09_08; 10(a) chb10_31; 10(b) chb10_38; 10(c) chb10_89; 11(a) chb11_82; 11(b) chb11_92; 12(a) chb12_10; 12(b) chb12_11; 12(c) chb12_23; 12(d) chb12_36; 12(e) chb12_42; 13(a) chb13_58; 13(b) chb13_60; 13(c) chb13_62; 14(a) chb14_17; 14(b) chb14_18; 14(c) chb14_27; 15(a) chb15_49; 15(b) chb15_52; 15(c) chb15_54; 15 (d) chb15_62; 16(a) chb16_17; 17(a) chb17b_63; 18(a) chb18_32; 18(b) chb18_35; 18(c) chb18_36; 19(a) chb19_30; 20(a) chb20_15; 20(b) chb20_16; 20(c) chb20_68; 21(a) chb21_21; 21(b) chb21_22; 22(a) chb22_38; 23(a) chb23_06; 23(b) chb23_08; 24(a) chb24_07; 24(b) chb24_09; 24(c) chb24_13; 24(d) chb24_15; 24(e) chb24_17 and 24(f) chb24_21.

iv Check if $c = 4$, if yes then
1 Make a loop $k =$ i to $(i + 6)$ and increase every time by 1.
2 Set outAnn2(k,1) = 1.
3 End the loop.
d Set $c = 0$.
e End the loop.

(9) Again make 5 s sliding window using following steps:
a Initialize a third array namely "outAnn3" whose size is $(r \times 1)$ and set outAnn3(1:r,1) = 0 and also set $c = 0$.
b Make a loop from $i = 1$ to (r-4) and increase it every time by 1.
c Check if outAnn2(i,1) = 1 and outAnn2$(i + 4,1) = 1$, if yes then
i Make a loop $j =$ i to $(i + 4)$ and increase it every time by 1.
ii Check if outAnn2(j,1) = 1, if yes then $c = c + 1$.
iii End the loop.
iv Check if $c = 5$, if yes then
1 Make a loop $k =$ i to $(i + 1)$ and increase every time by 1.
2 Set outAnn3(k,1) = 1.
3 End the loop.
d Set $c = 0$.
e End the loop.

(10) Make 11 s sliding window using following steps:
a Set $c = 0$.
b Make a loop from $i = 1$ to (r-10) and increase it every time by 1.
c Check if outAnn3(i,1) = 1 and outAnn3$(i + 10,1) = 1$, if yes then
i Make a loop $j =$ i to $(i + 10)$ and increase it every time by 1.
ii Check if outAnn3(j,1) = 1, if yes then $c = c + 1$.
iii End the loop.
iv Check if $c >= 2$, if yes then
1 Make a loop $k =$ i to $(i + 9)$ and increase every time by 1.
2 Set outAnn3(k,1) = 1.
3 End the loop.
d Set $c = 0$.
e End the loop.

(11) Make 10 s sliding window using following steps:
a Initialize final array namely "outFinal" whose size is $(r \times 1)$ and set outFinal(1:r,1) = 0 and also set $c = 0$.
b Make a loop from $i = 1$ to (r-9) and increase it every time by 1.
c Check if outAnn3(i,1) == 1 and outAnn3$(i + 9,1) = 1$, if yes then

i Make a loop $j = i$ to $(i + 9)$ and increase it every time by 1.
ii Check if outAnn3(j,1) = 1, if yes then $c = c + 1$.
iii End the loop.
iv Check if $c = 10$, if yes then
1 Make a loop $k = i$ to $(i + 9)$ and increase every time by 1.
2 Set outFinal(k,1) = 1.
3 End the loop.
d Set $c = 0$.
e End the loop.

For patient # 6 and 16, due to very short seizure duration, we applied steps 1–9. Moreover, for patient # 16, in step 9, we applied 4 s sliding window.

2.7. Performance evaluation

The performance of our proposed method is evaluated by computing the statistical parameters. These are:

Sensitivity (Sens.): $$Sens. = \frac{TP}{TP+FN} \times 100 \quad (15)$$

Specificity (Spec.): $$Spec. = \frac{TN}{TN+FP} \times 100 \quad (16)$$

Classification accuracy (CA): $$CA = \frac{TP+TN}{TP+FN+TN+FP} \times 100 \quad (17)$$

Latency: The amount of delay to detect seizure epochs by the algorithm as compare to the ground truth marked onset by the clinicians.

False positive rate (FPR): $$FPR = \frac{FP}{TN+FP} \times 100 \quad (18)$$

3. Results and discussion

Basic FIR filter within the frequency range of 0.5 Hz–40 Hz was used to remove artifacts and noise fused together with EEG signals while recoding. The database that we used in this study contains records with different channel configurations, and, first, when we trained classifier using data from different channel configurations, it's performance was very low. Thus, we selected the records which had same channel configuration that's why we used these common 23 channels configuration. In seizure record, both seizure signals and non-seizure signals exist. To train the ANN classifier, it is required unique seizure signal and non-seizure signal. So, each record was divided into 1 s segments. Some existing studies have made 5 s or 10 s segments, but making segments length greater than 1 s creates some trouble, for instance, when we have tried to divide the record # chb06_18 of patient # 6 whose duration is 7928 s into 5 s segments, we could not make any segments for the last 3 s data. At next stage, average value was calculated for all the 23 channels data to reduce computational time and size of the feature vectors. For time-frequency feature extraction DWT was performed for each 1 s epoch. By averaging the channels data, feature vector size decreases from 1265 ($= 23 \times 5 \times 11$) to 55 ($= 5 \times 11$) feature vectors for four level wavelet decomposition. This technique has made our model more time efficient and less complexity. One study, Jiang et al. [16] extracted total 2254 dimensional feature vectors for each segment and they used AGRM algorithm to select the best features. This procedure requires additional computation time. After extracting non-linear time-frequency statistical features, these were fed to the input of ANN with smart training algorithm. LM algorithm is more appropriate than GD algorithm because it requires significantly very low computational time. In training stage to train the classifier, for selecting seizure and non-seizure epochs from Training records, shown in Table 1, we used all seizure epochs and for non-seizure epochs selection, at first, we checked the validity of the epochs by measuring their classification performance then we selected the epochs which showed greater performance. For most of the patients all seizure and non-seizure epochs were used, whereas, for some patients majority of the non-seizure epochs were used. In addition to, from the Training records we trained the classifier and then using this classifier Tested records were tested. Moreover, Training and Testing records were selected based on seizure duration. In this study, we tried to divide the records into Training and Testing record so that each side contains almost equal amount of seizure duration. After the classification, the output of the classifier contains countable amounts of false detection and these false detection occurs due to trace of artifacts and noises, therefore, a novel 'sequential window algorithm' was developed to improve FDR. Conveniently, by using this algorithm, FDR is reduced significantly and performance is enhanced. We graphically illustrated the performance of our model, as shown in Fig. 7, and the measurement results are demonstrated in Table 4.

Table 4
Calculated values of sensitivity, specificity, accuracy, latency and FPR after application of SWA.

Patient #	Sensitivity	Specificity	Accuracy	Mean Latency (s)	FPR (%)
1	87.68	100	99.63	4	0
2	90.12	100	99.78	8	0
3	85.47	99.82	99.58	7.5	0.18
4	84.40	100	99.76	4.5	0
5	98.31	100	99.94	1.5	0
6	52.83	99.26	99.21	1	0.74
7	88.54	100	99.70	0	0
8	75.85	100	99.01	9.5	0
9	89.33	100	99.88	0.5	0
10	86.30	99.98	99.84	3	0.02
11	81.48	99.72	99.58	0	0.28
12	53.11	100	98.57	5.92	0
13	57.52	98.97	98.10	8.2	1.01
14	81.03	99.83	99.73	3.67	0.17
15	76.97	99.28	98.16	5.12	0.68
16	70.97	100	99.75	1	0
17	97.73	99.75	99.70	2	0.25
18	82.25	99.65	99.38	8.67	0.35
19	96.30	100	99.91	0	0
20	80.36	99.27	98.93	6.75	0.71
21	68.82	99.90	99.5	0	0.09
22	86.11	100	99.72	8	0
23	81.11	99.58	99.39	7.33	0.42
24	83.17	100	99.84	2.83	0

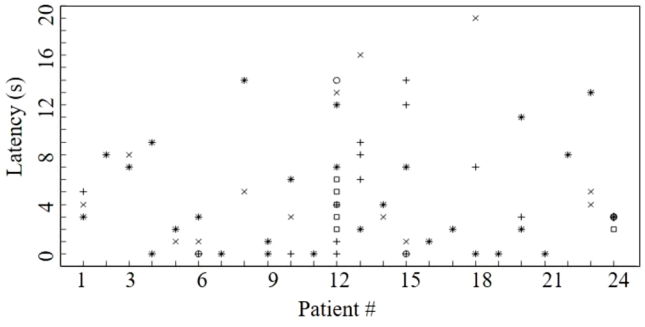


Fig. 8. Latency time for each tested seizure record where the symbols '*', '×', '+', '○', '□' and '◇' represents for all (a), (b), (c), (d), (e) and (f), respectively from previous Fig. 7.

From the performance analyses as seen in Fig. 7 and Table 4, it is noticed that patient # 6, 12 and 13 gives very low sensitivity rate which is supposed to be happened due to residual artifacts in the signals and also because of either age or gender of the patients. Patient # 12 and 13 are female, they are 2 and 3 years old, respectively, thus, their data contains a lots of artifacts, though, patient # 8 and 10 are also 3.5 and 3 years old, respectively, but they are male.

We achieved 99.44% mean classification accuracy, 80.66% average sensitivity, 99.79% mean specificity and 0.2% average FPR. Additionally, 0% FPR is achieved for patient # 1, 2, 4, 5, 7, 8, 9, 12, 16, 19, 22 and 24. However, FPR is relatively high for patients # 6, 13, 15, 20 and 23. Moreover, latency time for different records is depicted in Fig. 8. We achieved average latency time 0 s for patients # 7, 11, 19 and 21. For

Table 5
Comparisons with other existing studies for CHB-MIT database.

Refs. No.	Researchers (Year)	Methods1	Patient #	Sens. (%)	Spec. (%)	CA (%)	Latency (s)	FPR
[29]	Khan et al. (2012)	DWT, LDA	5	83.60	100	91.8	-	-
[28]	Ahammad et al. (2014)	DWT (db2), LC	24	98.5	-	84.2	1.76	-
[14]	Kiranyaz et al. (2014)	CNBC, PSO	21	89.01	94.71	-	-	-
[15]	Fergus et al. (2015)	PSD, KNNC	24	93	94	-	-	-
[30]	Xiang et al. (2015)	FE, SVM	18	98.27	98.36	98.31	-	-
[31]	Zabihi et al. (2016)	PCA, LDA, NBC	23	88.27	93.21	93.11	5.03	4.86/h
				89.10	94.80	94.69	4.65	3.04/h
[32]	Truong et al. (2018)	STFT, CNN	13	81.2	-	-	-	0.16/h
[19]	Harpale et al. (2018)	ICA, WT, FC	22	96.52	-	96.48	-	0.352 (%)
[33]	Selvakumari et al. (2019)	PCA, SVM, NBC	24	97.5	94.5	96.28	-	-
			23	97.92	95.01	96.77		
[20]	Mansouri et al. (2019)	PBI, DN, CN	19	72	-	-	8	2.82/h
-	This Work (2020)	DWT, ANN, SWA	24	80.66	99.79	99.44	4.12	0.2%

[1]linear discriminant analysis (LDA), linear classifier (LC), fuzzy classifier (FC), fuzzy entropy (FE), naive Bayesian classifier (NBC).

some records we also found latency time 0 s, such as, records # chb04_28, chb06_18, chb06_24, chb09_08, chb10_89, chb12_23, chb15_54, chb15_62, chb18_32 but there are few records, such as, records # chb08_11, chb12_10, chb12_11, chb12_36, chb13_60, chb15_54, chb18_35, chb20_15, chb20_16, chb23_06, we acieved 14, 12, 13, 14, 16, 14, 19, 11, 13 s latency time, respectively, and due to these later records our mean latency time increased to 4.12 s. A comparison is done with other existing studies, and is shown in Table 5. Some existing studies [14,15,19,29,31,32] have showed greater sensitivity rate but poor specificity rate and accuracy. In case of long-term EEG data high specificity rate close to 100% is very important because if specificity rate is a small amount lower than 100%, a lot of non-seizure signals will be classified as seizure signals. For instance: from Fig. 7 it is seen that for patient # 6, a noticeable amount of green color where there are no corresponding red color, as a result these green color signals are detected as seizure signals but actually they are non-seizure signals though specificity rate is 99.26% and accuracy rate is 99.21%. Apart from this patient, patients # 11, 13, 14, 15, 18, 20 and 23 have also very few green color whereas there are no seizure signals actually.

4. Conclusion

This research is mainly motivated on to design a methodology by which epileptic seizure can be detected more accurately from raw scalp EEG signals. Firstly, basic FIR filter with the frequency range of 0.5–40 Hz is performed. Secondly, DWT with fourth level wavelet decomposition by using db6 mother wavelet function is executed of the filtered signal to characterize it in time-frequency domain. Eleven time-frequency features extracted from each coefficient are then fed to the ANN's input. The network is trained by using LM back propagation algorithm. The network is optimized in terms of hidden layer nodes to attain greater classification accuracy. After classification, the proposed SWA is used to improve FDR. Finally, we achieved 99.44% mean classification accuracy, 80.66% average sensitivity, 99.79% mean specificity, 4.12 s mean latency and 0.2% average FPR, which are relatively better compare to other existing studies. Significantly, FPR value is really low even there are some patients which have zero FPR rate, and, also latency rate is very reasonable.

References

[1] A. Subasi, E. Ercelebi, Classification of EEG signals using neural network and logistic regression, Comput. Methods Programs Biomed. 78 (2) (2005) 87–99, https://doi.org/10.1016/j.cmpb.2004.10.009.
[2] L. Guo, D. Rivero, J. Dorado, J.R. Rabunal, A. Pazos, Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks, J. Neurosci. Methods. 191 (1) (2010) 101–109, https://doi.org/10.1016/j.jneumeth.2010.05.020.
[3] E. Juarez-Guerra, V. Alarcon-Aquino, P. Gomez-Gil, Epilepsy seizure detection in EEG signals using wavelet transforms and neural networks, in: K. Elleithy, T. Sobh (Eds.), New Trends in Networking, Computing, E-learning, Systems Sciences and Engineering, Springer, 2015. Lecture notes in Electrical EngineeringCham. vol. 312pp. 261-269http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-319-06764-3_33.
[4] World Health Organization, Geneva, Switzerland, Epilepsy. https://www.who.int/newsroom/fact-sheets/detail/epilepsy, 2019 (accessed 05 August 2019).
[5] A.T. Tzallas, M.G. Tsipouras, D.I. Fotiadis, Automatic seizure detection based on time-frequency analysis and artificial neural networks, Comput. Intell. Neurosci. (2007), https://doi.org/10.1155/2007/80510, 80510.
[6] H. Ocak, Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy, Expert Syst. Appl. 36 (2) (2009) 2027–2036, https://doi.org/10.1016/j.eswa.2007.12.065.
[7] Z. Zainuddin, L.K. Huong, O. Pauline, Reliable epileptic seizure detection using an improved wavelet neural network, Australas. Med. J. 6 (5) (2013) 308–314, https://doi.org/10.4066/AMJ.2013.1640.
[8] F. Shahlaei, S. Banakar, H. Salempoor, M. Aflaki, S.M.S.B. Keyvan, Feature classification of EEG signal using signal energy in multi-resolution analysis (MRA) and radial basis function (RBF) for detecting seizure and epilepsy, Int. J. Electromagn. Appl. 7 (1) (2017) 1–8, https://doi.org/10.5923/j.ijea.20170701.01.
[9] U. Orhan, M. Hekim, M. Ozer, EEG signals classification using the K-means clustering and a multilayer perceptron neural network model, Expert Syst. Appl. 38 (10) (2011) 13475–13481, https://doi.org/10.1016/j.eswa.2011.04.149.
[10] I. Guler, E.D. Ubeyli, Application of adaptive neuro-fuzzy inference system for detection of electrocardiographic changes in patients with partial epilepsy using feature extraction, Expert Syst. Appl. 27 (2004) 323–330, https://doi.org/10.1016/j.eswa.2004.05.001.
[11] I. Guler, E.D. Ubeyli, Detection of ophthalmic artery stenosis by least-mean squares backpropagation neural network, Comput. Biol. Med. 33 (4) (2003) 333–343, https://doi.org/10.1016/S0010-4825(03)00011-8.
[12] W.G. Baxt, Use of an artificial neural network for data analysis in clinical decision making: the diagnosis of acute coronary occlusion, Neural Comput. 2 (4) (1990) 480–489, https://doi.org/10.1162/neco.1990.2.4.480.
[13] A. Subasi, EEG signal classification using wavelet feature extraction and a mixture of expert model, Expert Syst. Appl. 32 (4) (2007) 1084–1093, https://doi.org/10.1016/j.eswa.2006.02.005.
[14] S. Kiranyaz, T. Ince, M. Zabihi, D. Ince, Automated patient-specific classification for long term electroencephalography, J. Biomed. Inform. 49 (2014) 16–31, https://doi.org/10.1016/j.jbi.2014.02.005.
[15] P. Fergus, D. Hignett, A. Hussain, D. Al-Jumeily, K. Abdel-Aziz, Automatic epileptic seizure detection using scalp EEG and advanced artificial intelligence techniques, Biomed Res. Int. (2015), https://doi.org/10.1155/2015/986736, 2015 (986736).
[16] X. Jiang, K. Xu, R. Zhang, H. Ren, W. Chen, Redundancy removed dual-tree discrete wavelet transform to construct compact representations for automated seizure detection, Appl. Sci. 9 (2019), https://doi.org/10.3390/app9235215 (23) (5215).
[17] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J. E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220, https://doi.org/10.1161/01.CIR.101.23.e215 [Online].

[18] A.H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment [Ph.D. thesis], MIT Division of Health Sciences and Technology, Harvard University, 2009. http://hdl.handle.net/1721.1/54669.
[19] V. Harpale, V. Bairagi, An adaptive method for feature selection and extraction for classification of epileptic EEG signal in significant states, J. King Saud Univ. Comput. Inf. Sci. (2018), https://doi.org/10.1016/j.jksuci.2018.04.014.
[20] A. Mansouri, S.P. Singh, K. Sayood, Online EEG signal detection and localization, Algorithms 12 (2019), https://doi.org/10.3390/a12090176 (9) (176).
[21] M.H. Libenson, Practical approach to electroencephalography, First ed., Saunders, United States, 2009.
[22] P. LeVan, E. Urrestarazu, J. Gotman, A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification, Clin. Neurophysiol. 117 (4) (2006) 912–927, https://doi.org/10.1016/j.clinph.2005.12.013.
[23] M.E. Saab, J. Gotman, A system to detect the onset of epileptic seizures in scalp EEG, Clin. Neurophysiol. 116 (2) (2005) 427–442, https://doi.org/10.1016/j.clinph.2004.08.004.
[24] A. Aarabi, F. Wallois, R. Grebe, Automated neonatal seizure detection: a multistage classification system through feature selection based on relevance and redundancy analysis, Clin. Neurophysiol. 117 (2) (2006) 328–340, https://doi.org/10.1016/j.clinph.2005.10.006.
[25] O. Faust, U.R. Acharya, H. Adeli, A. Adeli, Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis, Seizure 26 (2015) 56–64, https://doi.org/10.1016/j.seizure.2015.01.012.
[26] J.J. More, The Levenberg-Marquardt algorithm: Implementation and Theory, in: G. A. Watson (eds.), Numerical Analyses, Lecture Notes in Mathematics, vol. 630. Springer, Berlin, Heidelberg. 10.1007/BFb0067700.
[27] L. Zinn-Bjorkman, D.R. Harp, V. Vesselinov, Numerical optimization using the levenberg-marquardt algorithm, (2011). 10.13140/RG.2.2.11253.01760.
[28] N. Ahammad, T. Fathima, P. Joseph, Detection of epileptic seizure event and onset using EEG, Biomed. Res. Int. (2014), https://doi.org/10.1155/2014/450573, 2014 (450573).
[29] Y.U. Khan, N. Rafiuddin, O. Farooq, Automated seizure detection in scalp EEG using multiple wavelet scales, in: Proceedings of the IEEE International Conference on Signal Processing, Computing and Control, ISPCC, Waknaghat Solan, India, 2012, https://doi.org/10.1109/ISPCC.2012.6224361, 1-5.
[30] J. Xiang, C. Li, H. Li, R. Cao, B. Wang, X. Han, J. Chen, The detection of epileptic seizure signals based on fuzzy entropy, J. Neurosci. Methods 243 (2015) 18–25, https://doi.org/10.1016/j.jneumeth.2015.01.015.
[31] M. Zabihi, S. Kiranyaz, A.B. Rad, A.K. Katsaggelos, M. Gabbouj, T. Ince, Analysis of high-dimensional phase space via poincaré section for patient-specific seizure detection, IEEE Trans. Neural Syst. Rehabil. Eng. 24 (3) (2016) 386–398, https://doi.org/10.1109/TNSRE.2015.2505238.
[32] N.D. Truong, A.D. Nguyen, L. Kuhlmann, M.R. Bonyadi, J. Yang, S. Ippolito, O. Kavehei, Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram, Neural Netw. 105 (2018) 104–111, https://doi.org/10.1016/j.neunet.2018.04.018.
[33] R.S. Selvakumari, M. Mahalakshmi, P. Prashalee, Patient-specific seizure detection method using hybrid classifier with optimized electrodes, J. Med. Syst. 43 (5) (2019), https://doi.org/10.1007/s10916-019-1234-4.

A fuzzy proximity relation approach for outlier detection in the mixed dataset by using rough entropy-based weighted density method

Bhagaban Sri Ramakrishna, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, maheswarinath1@outlook.com*

Purnya Prava Nayak, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, purnyaprava.nayak26@gmail.com*

Manisha Pradhan, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, manishapradhan456@gmail.com*

Swaha Pattnaik, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, swaha.pattanayak@gmail.com*

Keywords:
Data Mining
Entropy
Fuzzy Proximity
Mixed Data
Rough sets
Weighted Density

A B S T R A C T

Data mining is an emerging technology where researchers explore innovative ideas in different domains, particularly detecting anomalies. Instances in the dataset which considerably deviate from others by their common patterns are known as anomalies. The state of being ambiguous and not affording certainty of data exists in this world of nature. Rough Set Theory is a proven methodology which deals with ambiguity and uncertainty of data. Research works that have been done until this point were focused on numeric or categorical type, which fails when the attributes are mixed type. By using fuzzy proximity and ordering relations, the numerical data has been converted to categorical data. This article presented an idea for detecting outliers in mixed data where the weighted density values of attributes and objects are calculated. The proposed approach has been compared with existing outlier detection methods by taking the hiring dataset as an example and benchmarked with Harvard dataverse datasets to prove its efficiency and performance.

1. Introduction

Data can be defined as any matter, numerals, or content easily handled by a system. Nowadays, companies have a huge volume of data in various styles and aspects. It comprises operational information such as stock and finance, non-operational information like weather forecasting and monetary information, and meta information (the information about the information itself), like the design of different databases or definitions for a word given in a dictionary [3]. Modeling of data or providing the link between these objects will provide some information. The point of sale system provides information regarding when the products are sold. The information can be translated into knowledge based on previous facts and by future predictions. The point of sale system can be improved by knowing the buying behavior of the customers. In recent years, massive data acquisition are amassed at the supermarkets, images produced by the satellites, and data present in the networking system [29]

A dataset may contain instances that have not adhered to normal behavior or deviate from the rest of the objects are termed as outliers [11]. A dataset may be comprised of numerical, categorical, or mixed types of data. It also alludes to discovering designs in the information system that does not adjust to expected behavior. Exceptions have likewise been alluded to as abnormalities, dissonant perceptions, exemptions, issues, abandons, distortions, commotion, or contaminants in various application domains. In earlier days, outliers were discarded as noise or exceptions.

An anomaly may demonstrate wrong information. For instance, the information may have been coded mistakenly, or the analysis might not have been run accurately[16]. If the outlying point is erroneous, then it can be corrected or removed from the dataset. It may not be conceivable to decide whether an outlying point has invalid information. If the information contains critical anomalies, we may need to think about the utilization of powerful,measurable systems[6]. But nowadays, much importance will be given to identify outliers. Because sometimes it may hold some valuable information. It is vital to identify outliers in major domains such as criminal activities like misuse of mobile phones and credit card activities, pattern recognition of malignant tumors, secured communication in the presence of third parties, malfunction of an airplane engine, and artificial intelligence [2]

The intra region anomalies are determined by density-based and inter-region anomalies are determined by distance-based methods[10]. Also, outliers can be identified in exceptional cases and the generation of novel patterns. Mostly clustering techniques provide efficient outlier detection rather than classification method. The statistical approach, probability model, will also be used to determine outliers[17].

Outliers are reported in two categories: the labeled objects are

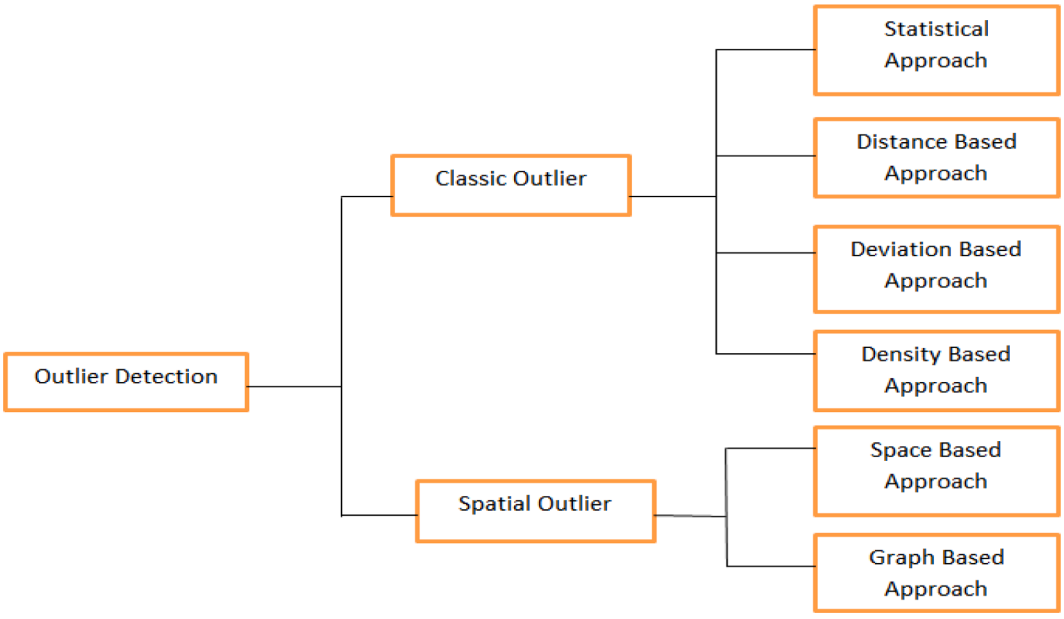


Fig. 1. Different Methodologies of outlier detection.

treated as normal objects, and the remaining objects that are not labeled are identified as outliers. Each pattern will be assigned an outlier score by fixing the threshold value to determine the degree of outliers [18]. The similitude of data cannot be correspondingly measured if there exists much noise. But the similarity measure and density measure do not suit high dimensional data [8]. Nowadays, researchers are focussing on detecting outliers in high-dimensional data. Because so many works were carried out to detect outliers for qualitative and quantitative data [23]. The proposed approach probably suits mixed data with a high level of significance. Different methodologies for outlier detection techniques are shown in Fig. 1.

2. Outlier detection method

2.1. Supervised method

This technique displays data uniformity and anomaly. The specialists label similar objects and objects that do not coordinate the model of ordinary objects as exceptions or outliers [1]. The normal data objects appear much than the outlier objects. This method has two classes (normal and outliers) which are imbalanced. The small amount of sample data taken for training will not suitably be considered for outlier distribution. But labeling the true object as an outlier should not be allowed. It is more important than outlier detection.

2.2. Unsupervised methods

In a few applications, labeling objects as "usual" or "exception" are not made. Consequently, an unsupervised learning technique must be utilized. Clustering can be done between normal objects and outlier objects[9]. Objects which deviate from normal behavior form one cluster, and the remaining objects fall into a normal category. The issues in unsupervised strategies are sometimes data that does not belong to any group might be considered noise but not an outlier[35]. Also, it is regularly expensive to design clusters first and to find anomalies. It is typically expected that outlier objects are distant than objects which are considered to be normal.

2.3. Semi-supervised methods

It can be viewed as the utilization of semi-supervised learning strategies. In particular, while accessing labeling objects, it can be utilized, or with the closer unlabelled objects that are close by preparing a layout for normal objects. The layout of the ordinary object at that point can be utilized to identify outliers - the items which do not fit into the layout of normal objects are anomalies [4]. To enhance the nature of exception location, one can get assistance from models of unsupervised strategies.

3. Rough set theory and fuzzy approximation space

During the 1980s, Zdzislaw Pawlak[27], a Polish mathematician, developed a mathematical tool called rough sets with lower and upper approximation concepts, which have crisp sets. However, it does not need any prior or extra information about concerned data. There exists a strict association between vague and uncertain data. The rough set approach demonstrates a clear association between these two ideas. Vagueness is associated with sets, while uncertainty is associated with components of sets. The data analysis with rough sets uses decision tables with structured rows and columns[12]. The columns of a table are attributes classified into two groups: condition and decision attributes. Each row specifies an object which induces some decision or result. If some conditions are satisfied, then the decision rule is certain; otherwise, it is uncertain.

It also implicates the thought of similarity. Let us consider the information table $IT=(W, X, Y, Z)$ where W is the universe which should be nonempty, X is the set of attributes, Y and Z are the conditional and decisional attributes[13]. The components of W are objects, entities, items, or investigations. Attributes are also represented as features, aspects, or characteristics.

Assume $S=(V, RT)$ then the subset $Y \subseteq V$ and an equivalence relation $RT \in IND(S)$. The subsets of X, such as lower and upper approximation, are defined as follows:

$$\underline{RT}Y = \cup^{\{X \in V/RT:\ X \subseteq Y\}}$$
$$\overline{RT}Y = \cup^{\{X \in V/RT:\ X \cap Y = \varnothing\}}$$
$$or$$
$$x \in \underline{RT}X \ \ if\ and\ only\ if\ \ [x]_{RT} \subseteq X$$
$$x \in \overline{RT}X if and only if [x]_{RT} \cap^{X} \neq 0$$

From this, $Boundary(X) = \overline{RT}\,X - \underline{RT}\,X$ will be called the RT boundary of X. The boundary sets are included in the upper approximation but not in the lower approximation. Rough sets are defined through the lower

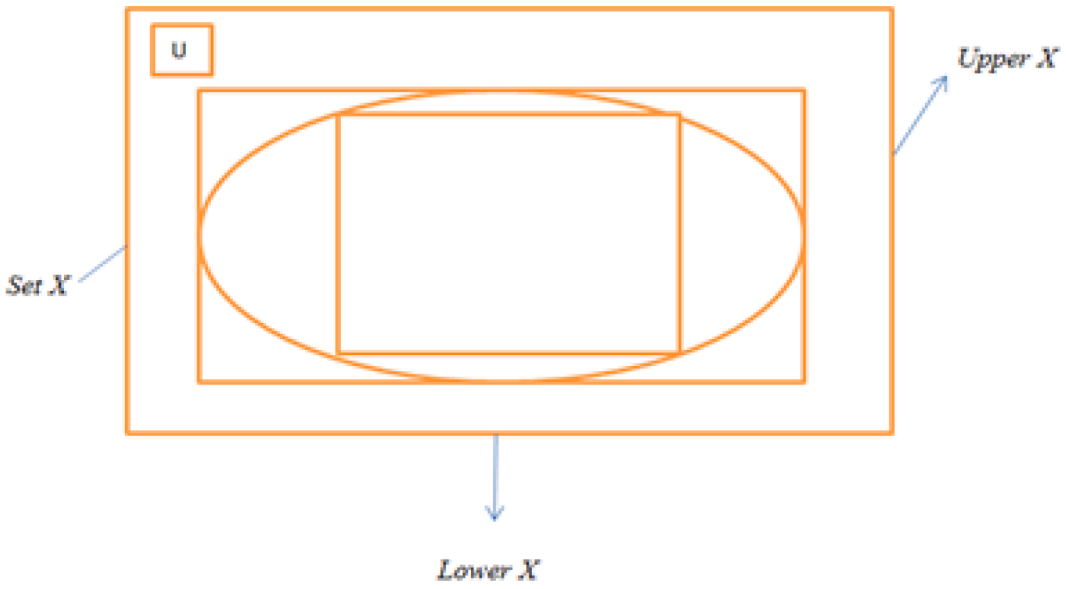


Fig. 2. Set Approximation

and upper approximation. Also, a boundary region is a devoid set ($\overline{RT}\,X \neq \underline{RT}\,X$).

3.1. Membership relation and approximation

Membership relation is derived from approximation spaces. Both membership and set approximation are related to knowledge only[28]. The representation is shown below:

$l \underline{\in}_T Lthenl \in \underline{T}L$
$l_T \overline{\in} Lthenl \in \overline{T}L$

In which,$\underline{\in}_T$reads "l surely belongs to L for T" and $\overline{\in}T$, "l possibly belongs to L concerning T", is the lower and upper membership relation, respectively. Fig. 2 depicts the set approximation.

3.2. Fuzzy approximation space with Rough Sets

In general, fuzzy sets are used to handle the issues in understandability of the patterns, incomplete and noisy data, multimedia information, and intercommunication between persons resolves quickly within the determined time [14]. The minimal and maximal approximation for a fuzzy set B in Z as the fuzzy sets $T\downarrow B$, $T\uparrow B$ in Z *as*

$(T\downarrow B)(r) = inf_{s\in S}\,(R(s,\ r),\ A(s))$
$(T\uparrow B)(r) = sup_{s\in S} T(R(s,\ r),\ A(s))$

$T\downarrow B$ and $T\uparrow B$ can also be determined as how much inclusion T_r in B and overlap of T_r and B respectively [10], which is related to $r \in T\downarrow A$*only* $[r]_T \underline{\subseteq} B$ and $r \in T \uparrow B$ *only* $[r]_T \cap B \neq 0$.

4. Related work

Datasets make different clusters based on different labeling techniques. A data item to be compared with these formed clusters that don't belong to any cluster will be identified as an outlier[7]. For a single class classification, a support vector data description (SVDD) method was used. It determines a hypersphere that includes all normal data within its space. The objects which out lies from the hypersphere are termed outliers. In k-means clustering, objects that are found to be similar under a feature vector are formed into clusters, and any object that does not group under any cluster is outliers. In the local outlier factor method (LOF), the relative distance of an object with its neighborhood points is to be calculated. If the value has a high deviation, then it is an outlier [34].

Multivariate Outlier Detection (MOD) is a traditional strategy for the detection of outliers. It regularly demonstrates those perceptions that are found generally a long way from the focal point of the information distributed. A few distance measures are executed for such detection [19]. The Mahalanobis distance is an outstanding rule which relies upon evaluated parameters of the multivariate distribution.

The rough membership function also is used to detect outliers from the real-world dataset. One of the most popular distance-based approaches is the Manhattan distance. When the threshold value increases, this technique outdoes the performance of statistical approaches and distance-based methods. The efficiency can be improved by fixing the proper threshold value. The clustering technique provides more accuracy than the distance-based approach[37,38]. Small clusters can be constructed by using Partitioning Around Medoids (PAM) to detect outliers from the dataset.

In neural networks, the data will be trained and tested. It is used to clear the ambiguity in patterns and is also an effective tool to retrieve knowledge from large databases. The rough set method with the neural network is defined well to handle data mining problems. A backpropagation algorithm has been employed using rough sets to avoid inconsistencies between data. The neural system learning model uses backpropagation. Neurobiologists and therapists initially ignited this field to create and test neuron's computational analog. The neural system is arranged so that input/yield units are associated with weights related to it [25].

Backpropagation learns by preparing an informational index of tuples iteratively, which contrasts the system's expectation of an individual tuple with the known target. The objective target might be the class name known for the preparation tuple (characterization issues) or consistent instance (forecast). Each preparation tuple has weights altered to limit the error of mean squared value between the system's expectation and the real target instance [20]

Rough Entropy has been used to measure the uncertainty of data. Each object and attribute is calculated with a weighted density value to detect outliers. But clustering of data had not been done[21]. The clustering approach can be improved by using RKM (rough K-means) with a preliminary centroid selection method [22]. Cluster validity index will be achieved by improved entropy-based rough K-means (ERKM) method. In multi granulation rough sets, the decision was made by "OR" instead of "AND" logic. When the two attributes have

Table 1
Study on different outlier detection methods

S. No	Outlier Detection Method	Advantages	Disadvantages
1	Support Vector Data description (SVDD)	It detects outliers well in smaller sample sizes and produces effective results for more intricate and scanty datasets.	If the sample sizes become larger, outlier detection is difficult.
2	k means clustering	Even if the dataset is huge,outlier detection is possible	Generally, outliers are to be discarded, but in this method, outliers form a separate group.
3	Local Outlier Factor(LOF)	The point at the smallest distance is considered an outlier to the cluster, which is at a denser level. But in general outlier detection approaches, the point at the smallest distance will not be considered an outlier.	The threshold value will be fixed to detect outliers. The fixation of the threshold value will be based on the problem and the user.
4	Multivariate Outlier Detection (MOD)	It detects outliers (of n features)in n-dimensional space.	Finding distributions of n-dimensional space is difficult, so training of the dataset would be required.
5	Partitioning Around Medoids (PAM)	When compared to other available partitioning algorithms, outliers are less noticeable in the PAM method	Choosing k medoids is random; it gives a different result for the same dataset.
6	Backpropagation Method	A deeper understanding of the data is not required.	Particularly sensitive to noisy data.
7	Rough k Means(RKM)	The weighted density method uses the Gaussian function to detect outliers in a vague dataset.	When separating objects which are overlapped between clusters, the approach is susceptible.
8	Entropy Rough k Means (ERKM)	Effectively outliers are removed, which results in the formation of quality clusters.	Centroid selection is random based on the Rough k means Method(RKM)

contradictions and inconsistencies, multi granulation with a rough set framework has been used [40]. So that, it needs effective computation.

The traditional approach of outlier detection was the statistical method where it applies to single-dimensional datasets alone. The model suitably fits for perceptible real-world datasets where the categorical data has been converted into numerical data for the processing of statistical methods [5]. So it increases the processing time for tangled datasets. The simple outlier detection method with no prior information needed for the processing of data is the proximity-based technique. But, the calculation of distances between all objects results in high exponential growth. The number of objects n and its dimensionality m is directly proportional to its time complexity. So it will not be suitable for high dimensional data.

The parametric method is suitable for larger datasets because it has a built-in distribution model. If any model fits the prescribed dataset, then the outcome will be accurate. The data model grows with paradigmatic complexity, not with the size of data. The only condition is the predefined model should be fit for the available dataset. The nonparametric methods need prior information to process.

In some cases, the prior knowledge will not be available, or the computation cost will be high[32]. Many datasets use not only a determined data model but also follow a random distribution model. It may be applicable for regression and principal component analysis methods. In the pre-processing stage, parameter settings are to be made, and later they should be processed.

An outer perception, or anomaly, seems to diverge extraordinarily from other individuals where it occurs. A perception (or a subset of perceptions) gives off an impression of conflicting with the rest of the data [24]. Exceptions are defined as the focuses lie outwards from the cluster but at the same time are isolated from the noise[30]. Patterns with the well-defined notion of normal behavior, which are not confirmed, are outliers, and the regions of network structure differs from expected under the normal behavior [26].

Social network anomaly detection focuses on outlier detection techniques developed in machine learning and statistical domains [31]. Intrusion detection with anomaly detection was proposed through system calls[33]. First, evaluate decision-makers preferences for each choice and introduce the concept of pre-decisions, resulting in an incomplete fuzzy decision system[43]. Then, using the defined similarity relation, the weighted conditional probabilities are determined. The concept of relative utility functions is next introduced, followed by a method for determining relative utility function values. Then, in incomplete fuzzy decision systems, we build a three-way decision model and apply it to the modeling of incomplete multi-attribute decision--making issues[44].

On IFVIS(intuitionistic fuzzy-valued information systems), three alternative sorting decision-making procedures include subtracting intuitionistic fuzzy numbers, sorting functions, and intimacy coefficients [45].We create the outranked set for each alternative and present a hybrid information table that includes a Multi-Attribute Decision-Making matrix and a loss function table.Multi-attribute decision-making (MADM) is a crucial component of modern decision sciences[46]. It refers to a decision problem of selecting the best alternative or ranking alternatives based on numerous attributes.A three-way decision has been included in a multi-scale decision information system, which offers a novel approach to addressing multi-attribute decision-making concerns in a multi-scale decision information system[47]. In addition, a review has been made for outlier detection using data mining methods. The pros and cons of different outlier detection methods are shown in Table 1.

5. Proposed model

Detecting outliers is a major data mining technique that has significant consideration inside different research groups and application domains. Numerous methods have been created to identify outliers but only on numerical data. Those methods cannot be applied directly to categorical data. So the fuzzy proximity relation is introduced to convert numerical data to categorical[36]. Then the Density and uncertainty of every object and attribute are calculated. For a stable dataset, the fixation of the threshold value is high, and for the unstable dataset, the lower threshold value is fixed. In this way, outliers are removed incredibly to improve the execution of data mining algorithms. In Fig. 3, at the pre-processing stage, the mixed data is converted to categorical by using fuzzy proximity relation in post-processing. Finally, a rough set entropy-based weighted density outlier detection approach is applied to determine outliers.

5.1. Roughset entropy-based weighted density outlier detection algorithm

A dataset may include missing data and some negative and null values, which are outliers. So the dataset is defined to be vague and incomplete. To handle this scenario, a rough set with a weighted density-based outlier detection method is proposed. In the pre-processing stage, numerical data is converted to categorical data by using fuzzy proximity relation, and then it is ordered. In the post-processing stage, similar objects are identified concerning attributes using indiscernibility relation, and complement entropy measure is used to calculate uncertainty values; the weighted density values are calculated by identifying indiscernible objects divided by the total number of objects to each attribute. Finally, the user fixes the threshold value. If the calculated value is lesser than the threshold, then they are treated as outlier objects. The following definitions will be used to detect outliers when the table has been converted from mixed to categorical type,

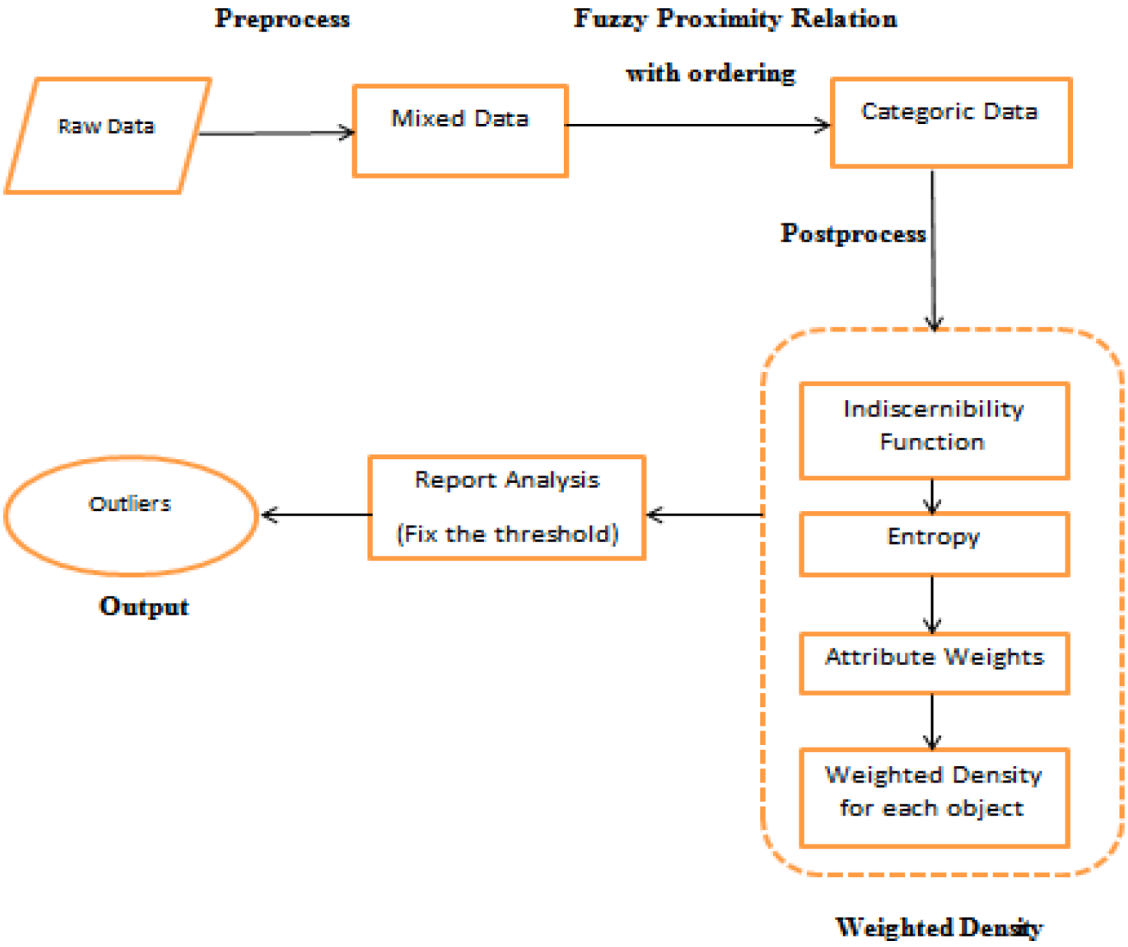


Fig. 3. Proposed Model for Outlier Detection using Rough Sets.

which is discussed below:

Definition 1:A. dataset *DS* is defined by the triplet $DS=(Z, R, C)$ where *Z* represents the universe, *R* represents the objects, and *C* represents the attributes in a dataset.

Definition 2. Let $DS=(Z, R, C)$ and $RT \subseteq C$. The indiscernibility relation *RT* for *r* in *R* or *s* in *C* is represented as

$$\{Z|IND(RT)\} = \{[r]_{RT}|\ r \epsilon Z\ \}$$

Definition 3:Let. $DS=(Z, R, C)$, and $RT \subseteq C$ and $\frac{Z}{IND(RT)} = \{C_1, C_2, \ldots C_m\}$.The complement entropy *(CPE)* with respect to *RT* is defined as

$$CPE(RT) = \sum_{j=1}^{n} \frac{|C|}{|R|} \frac{|C|_j^c}{|R|}$$

where C_j^c denotes complement set of C_j, which is $C_j^c = R - C$;

Definition 4. Let $DS=(Z, R, C)$, the weight of every attribute for *C* is defined as

$$Weight(C) = \frac{1 - CPE(RT)}{\sum_{j=1}^{n}(C_j)}$$

Definition 5:The. average Density of each attribute will be determined as

$$AverageDensity(R_j) = \frac{|[R_j]_C|}{|Z|}$$

From that, the weighted Density of each object will be determined as follows:

$$WeightedDensity(C) = \sum_{r_i \epsilon R} \left(Average\ Density(R_j), Z(C)\right)$$

Definition 6. Let us consider the dataset $DS=(Z, R, C)$, and θ is a fixed threshold value from the weighted density objects. If the value of *Weighted Density(R)* $< \theta$ then *r* is termed an outlier.

6. An empirical study on hiring dataset

A fabricated mixed dataset "Hiring" is designed with four conditional attributes Degree, Experience, French, and Reference for the effective proposed approach. The attribute experience has numerical values, and the remaining attributes such as degree, french, and reference have categorical values. So many algorithms are available for numerical data to detect outliers. But, our proposed method uses fuzzy proximity relation to convert numerical data to categorical data. The $FPR(o_i, o_j)$, which derives binary relation for the numerical attribute experience by using the formula,finds the almost similarity among the objects o_i &o_j.

$$FPR(o_i, o_j) = 1 - \frac{|o_i - o_j|}{(o_i + o_j)}$$

Based on calculated values, the attribute experience is ordered. The proposed algorithm has been applied to this dataset to detect outliers and graphs have also been plotted using the nominal values. The author has been conducted evaluations by comparing existing methods with the proposed method for hiring dataset. The hiring dataset with 10 objects and mixedattributes are shown in Table 2 and Table 3 shows fuzzy proximity relation for the attribute experience.

Table 2
Hiring Dataset - Mixed Type

Objects	*Degree*	*Experience*	*French*	*Reference*
E_1	*MBA*	*5.2*	*Yes*	*Excellent*
E_2	*MSc*	*4.3*	*Yes*	*Good*
E_3	*MSc*	*3.4*	*No*	*Neutral*
E_4	*MBA*	*2.5*	*No*	*Good*
E_5	*MCA*	*6.2*	*Yes*	*Good*
E_6	*MCA*	*3.1*	*Yes*	*Neutral*
E_7	*MBA*	*2.2*	*Yes*	*Excellent*
E_8	*MSc*	*3.2*	*No*	*Excellent*
E_9	*MCA*	*2.7*	*No*	*Good*
E_{10}	*MBA*	*2.4*	*Yes*	*Neutral*

Let the almost indiscernibility be $\omega \geq 90\%$, From Table 2, thus, the objects E_1, E_2, E_5 are ω- identical. Similarly, $E_3, E_4, E_6, E_7, E_8, E_9, E_{10}$ are ω – identical.

$$U/R_1^{\omega} = \{\{E_1, E_2, E_5\}, \{E_3, E_4, E_6, E_7, E_8, E_9, E_{10}\}\}$$

Based on the similarity value of ω, the attribute experience is ordered into two groups. The numerical values of the attribute experience for objects $\{E_1, E_2, E_5\}$ are having greater values. So it is classified as High and the remaining objects $\{E_3, E_4, E_6, E_7, E_8, E_9, E_{10}\}$ are classified as Low. Now the numeric type of experience attribute is converted to categorical, which is shown in Table 4.

Obtain indiscernible relation for each attribute. Objects that possess indiscernible values for attributes are:

$$U/IND\ (Degree) = \{E_1, E_4, E_7, E_{10}\}, \{E_2, E_3, E_8\}, \{E_5, E_6, E_9\}$$

$$U/IND(Experience) = \{E_1, E_2, E_5\}, \{E_3, E_4, E_6, E_7, E_8, E_9, E_{10}\}$$

$$U/\ IND(French) = \{E_1, E_2, E_5, E_6, E_7, E_{10}\}, \{E_3, E_4, E_8, E_9\}$$

$$U/IND(Reference) = \{E_1, E_7, E_8\}, \{E_2, E_4, E_5, E_9\}, \{E_3, E_6, E_{10}\}$$

The complement entropy function is to be calculated for each attribute with the obtained indiscernible relation.

$$CE(Degree) = \frac{4}{10}\left(1-\frac{4}{10}\right) + \frac{3}{10}\left(1-\frac{3}{10}\right) + \frac{3}{10}\left(1-\frac{3}{10}\right) = \frac{33}{50}$$

$$CE(Experience) = \frac{3}{10}\left(1-\frac{3}{10}\right) + \frac{7}{10}\left(1-\frac{7}{10}\right) = \frac{21}{50}$$

$$CE(French) = \frac{24}{50}\ ; CE(Reference) = \frac{33}{50}$$

Calculate each attribute weight by adding the total number of attributes with the complement entropy function.

$$Weight\ of\ Attribute(Degree) = \frac{17}{54};\ Weight\ of\ Attribute(Experience) = \frac{29}{54}$$

$$Weight\ of\ Attribute(French) = \frac{26}{54};\ Weight\ of\ Attribute(Reference) = \frac{17}{54}$$

The weight of each object should be calculated by the summation of the product of the weight of attributes with indiscernible objects.

$$W\ (E_1) = \frac{4}{10}\times\frac{17}{54} + \frac{3}{10}\times\frac{29}{54} + \frac{6}{10}\times\frac{26}{54} + \frac{3}{10}\times\frac{17}{54} = 0.67;$$

$$W(E_2) = 0.67; W(E_3) = 0.75; W(E_4) = 0.82; W(E_5) = 0.67\ ;$$

$$W(E_6) = 0.85; W(E_7) = 0.88; W(E_8) = 0.75;\ W(E_9) = 0.78;$$

$$W(E_{10}) = 0.88.$$

If $\theta < 0.7$, then the objects $E_1, E_2$and E_5 are outliers. The normal and outlier objects are shown in Fig. 4.

7. Experimental results

The working model of outlier detection algorithm in mixed datasets will be understood by, conducted experiments on a hiring dataset that has 120 objects with four conditional attributes of numerical and categorical values. It has been implemented with Processor-Intel Pentium,1GigaByte RAM, and the Windows10 operating system. Existing methods like Distance-based, Density-based, Local Outlier Factor and Class outlier factor were analyzed using Rapid Miner 7.0. The concept of Rough sets was implemented using C. It is a flexible language that is used to implement mathematical models. The proposed algorithm has been run on a hiring dataset that is of mixed type. The fuzzy proximity relation method was used to convert numerical value to categorical value, and then it was ordered.

A rough set entropy-based weighted density outlier detection

Table 4
Converted Table – Mixed to Categoric Type

Objects	*Degree*	*Experience*	*French*	*Reference*
E_1	*MBA*	*High*	*Yes*	*Excellent*
E_2	*MSc*	*High*	*Yes*	*Good*
E_3	*MSc*	*Low*	*No*	*Neutral*
E_4	*MBA*	*Low*	*No*	*Good*
E_5	*MCA*	*High*	*Yes*	*Good*
E_6	*MCA*	*Low*	*Yes*	*Neutral*
E_7	*MBA*	*Low*	*Yes*	*Excellent*
E_8	*MSc*	*Low*	*No*	*Excellent*
E_9	*MCA*	*Low*	*No*	*Good*
E_{10}	*MBA*	*Low*	*Yes*	*Neutral*

Table 3
Fuzzy Proximity Relation -Experience Attribute

R_1	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
E_1	1.0000	0.9053	0.7907	0.6494	0.9123	0.747	0.5946	0.7620	0.6836	0.6316
E_2	0.9053	1.0000	0.8832	0.7353	0.8191	0.8379	0.677	0.8534	0.7715	0.7165
E_3	0.7907	0.8832	1.0000	0.8475	0.7084	0.9539	0.7858	0.9697	0.8853	0.8276
E_4	0.6494	0.7353	0.8475	1.0000	0.5748	0.8929	0.9362	0.8772	0.9616	0.9796
E_5	0.9123	0.8191	0.7084	0.5748	1.0000	0.6667	0.5239	0.6809	0.6068	0.5582
E_6	0.747	0.8379	0.9539	0.8929	0.6667	1.0000	0.8302	0.9842	0.9311	0.8728
E_7	0.5946	0.677	0.7858	0.9362	0.5239	0.8302	1.0000	0.8149	0.898	0.9566
E_8	0.762	0.8534	0.9697	0.8772	0.6809	0.9842	0.8149	1.0000	0.9153	08572
E_9	0.6836	0.7715	0.8853	0.9616	0.6068	0.9311	0.8980	0.9153	1.0000	0.9412
E_{10}	0.6316	0.7165	0.8276	0.9796	0.5582	0.8728	0.9566	0.8572	0.9412	1.0000

Fig. 4. Showing Normal and Outlier objects.

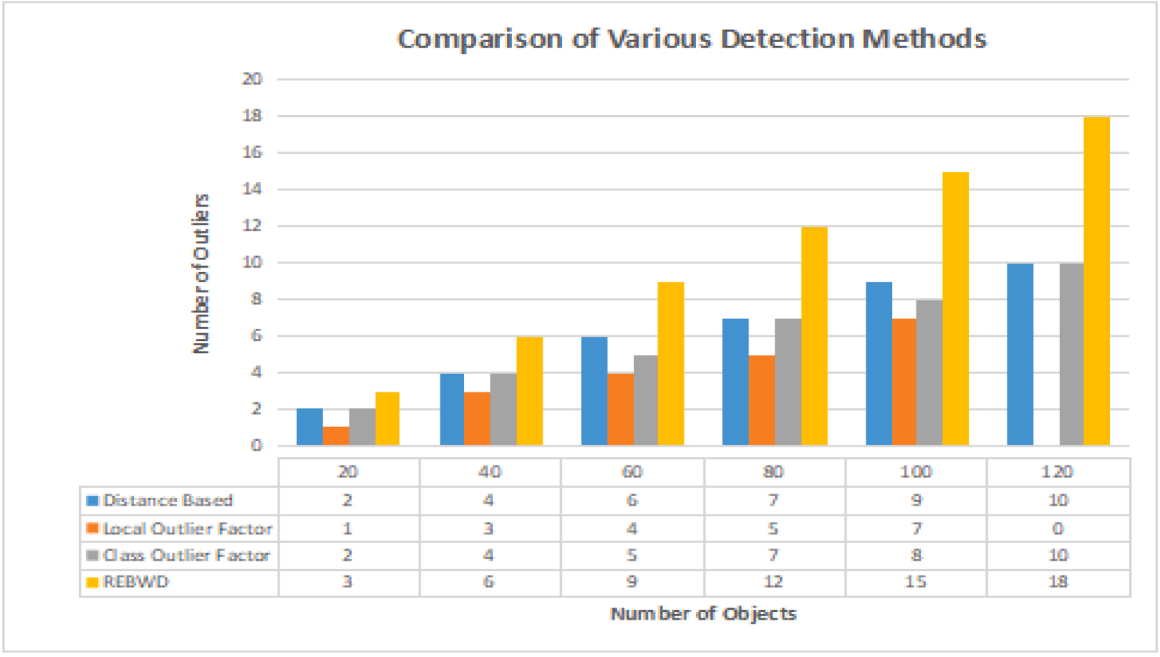
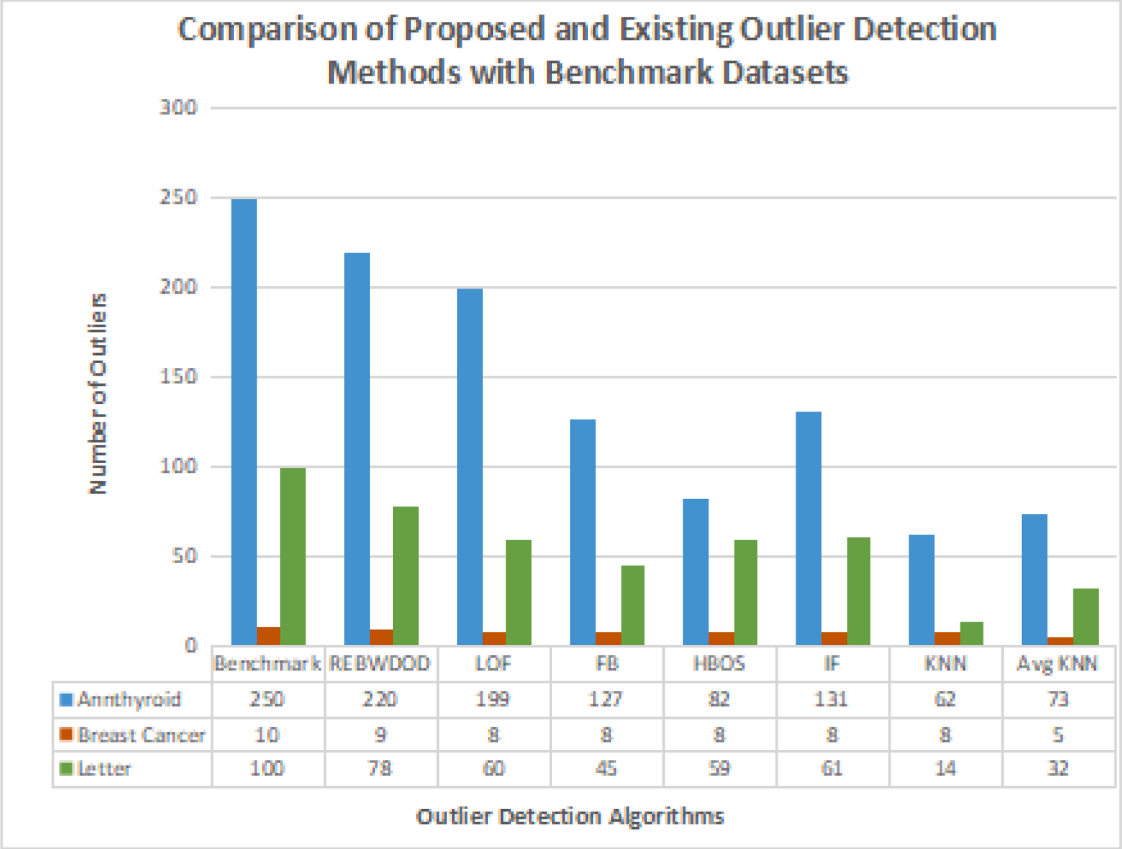


Fig. 5. Comparison Chart for Existing Methods with Proposed Method

method has been applied for effective outlier detection. Fig. 5 shows that the comparison chart for an existing and proposed method for outlier detection. In the distance-based outlier detection method, each data point has been ranked based on the distance to its *k-th* nearest neighbor [39] so that the top *n* data points are declared as outliers. It detects ten outlier objects. In density-based outlier detection method *DensityBased (p, P)*, an object that deviates at least P distance from the p, the proportion of all data objects is considered outliers. This method does not detect any outlier objects. In the local outlier factor method, each object should be calculated with a local outlier factor based upon the local density measure. Then it is compared with their *l* nearest neighbors [41].

The objects which are having lower density values when compared with their neighbors are termed to be outliers. It detects seven outlier objects. In the class outlier factor method, each data point in the sample will be ranked based on ClassOutlierFactor=(S, N) where S represents top-class outlier and N represents the number of nearest neighbors. This algorithm detects ten outlier objects. Further, our proposed method rough set entropy-based weighted density outlier detection method detects outliers by computing the weighted density value of all objects and attributes. It detects 18 outlier objects[42]. Our proposed algorithm's performance and efficiency are high compared to existing methods because it calculates weighted density values for every object and attribute so that a true object will never be detected as an outlier. The comparison chart showing various outlier detection methods is shown in Fig. 5.

Also, benchmark datasets such as the annthyroid dataset, breast cancer dataset, and letter dataset have been taken from Harvard dataverse to show the proposed algorithm efficiency, which has been compared with other existing outlier detection methodologies such as local outlier factor (LOF), feature-based(FB), isolation forest(IF), K-



	Benchmark	REBWDOD	LOF	FB	HBOS	IF	KNN	Avg KNN
Annthyroid	250	220	199	127	82	131	62	73
Breast Cancer	10	9	8	8	8	8	8	5
Letter	100	78	60	45	59	61	14	32

Fig. 6. Comparison of Proposed and Existing Outlier Detection Methods with Benchmark Datasets.

nearest neighbor (KNN), average KNN and histogram-based outlier score (HBOS). The local outlier factor determines the Density of an object with the distance of its neighbors. Feature bagging selects features of the subsamples randomly and finally combines the values of all base detectors, using the local outlier factor. Isolation forest observes data by constructing a tree[15]. The isolated value score is determined as outliers that are well suitable for high dimensional data. By constructing histograms, outliers are detected in the histogram-based outlier score approach. It is an unsupervised learning method that generates scores by considering the independent features.KNN identifies the nearest neighbor to an object. Based on the distance, it calculates scores, and outliers are identified. In the Average KNN method, super samples are constructed for individual classes. The test data is given as an input, and Average KNN searches samples available in super samples or closer. Others are identified as outliers. The comparison chart of the proposed method with existing outlier detection algorithms for the benchmark datasets is shown in Fig. 6.

Other approaches like fuzzy bipolar soft set and Pythagorean fuzzy bipolar soft set are compared with the proposed method to prove its efficiency. The fuzzy-based bipolar soft set is used to analyze the patients with the help of membership degrees and decide whether the patient is hypomania, depression, or bipolar. Mostly it is used in decision-making systems.On the other hand, the pythagorean fuzzy bipolar soft setis mostly used in group decision-making situations. Personalization of the findings acquired is avoided because a common idea is derived from the opinions of all doctors. But the proposed method identifies indiscernible values, computesEntropy, and then calculates each object's weighted density value and attribute to detect outliers.

7.1. Measures for performance evaluation

The performance evaluation of benchmark datasets is measured by calculating their accuracy, specificity, sensitivity, precision, and F1 score. The accuracy of a classifier is calculated as the total number of objects which are correctly classified to the total number of objects available. The formula to calculate accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

where TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative. Thus, sensitivity or recall measures the true positive values proportion, which is correctly identified, whereas specificity measures the true negative values proportion, which is correctly detected. The values are obtained from the formulas shown below:

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

Precision or positive predictive value is the one that measures relevant objects from the retrieved objects. The formula to calculate precision is as follows:

$$Precision = \frac{TP}{TP + FP}$$

F1 score measure provides the balance between precision and recalls when the distribution of classes is not even. It becomes worse when its value is 0 and best when it is 1. The formula to calculate the F1 score is as follows:

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 5
Performance Evaluation - Annthyroid Dataset

Sl.No	Measures	LOF	REBWDOD
1	Accuracy	98.16%	99.57%
2	Specificity	1.0	1.0
3	Sensitivity	0.9813	0.9955
4	Precision	1.0	1.0
5	F1 Score	0.9906	0.9978

Table 6
Performance Evaluation - Breast Cancer Dataset

Sl.No	Measures	LOF	REBWDOD
1	Accuracy	99.18%	99.46%
2	Specificity	1.0	1.0
3	Sensitivity	0.99	0.99
4	Precision	1.0	1.0
5	F1 Score	0.9958	0.9972

Table 7
Performance Evaluation - Letter Dataset

Sl.No	Measures	LOF	REBWDOD
1	Accuracy	97.56%	98.69%
2	Specificity	1.0	1.0
3	Sensitivity	0.97	0.98
4	Precision	1.0	1.0
5	F1 Score	0.9872	0.9930

The performance evaluation of benchmark datasets such as annthyroid, breast cancer, and letter dataset are shown in Table 5, Table 6, and Table 7.

7.2. Analysis of efficacy

The following three sorts of tests have been conducted to see how each algorithm's performance changes as factors change, such as the size of the data set, the dimensionality of the data set, and the number of outliers[48].In comparison to the local outlier factor method, the WDOD approach takes less time in terms of data size, data dimensionality, and mark the number of outliers.

The WDOD technique appears to be particularly suitable for big data sets with high dimensionality and data sets with a high number of outliers, based on the results of these experiments. The WDOD algorithm's growing rate of execution time is substantially slower than the local outlier factor algorithm. As a result, when the data size is huge, attributes are more the suggested WDOD algorithm can ensure efficient execution in detecting outliers which are shown in Fig. 7, Fig. 8, and Fig. 9.

8. Conclusion

In this paper, outlier detection for a mixed dataset has been proposed. In the pre-processing stage, fuzzy proximity relations with order information rules convert numeric to categorical attributes. The rough set-based Entropy weighted density outlier detection method has been carried out to detect outliers in the post-processing stage. Research works carried out so far detect outliers only for numeric or categorical data, where mixed data was not considered. The proposed model detects outliers in the hiring dataset, which has mixed data, by calculating their weighted density value so that the normal object will not be detected as an outlier. However, the proposed algorithm is benchmarked with Harvard dataverse datasets such as the annthyroid dataset,breast cancer dataset, and letter dataset compared with the existing local outlier factor outlier method to prove its efficiency and performance level. As the number of increasing objects and attributes, the proposed method ensures efficient execution in detecting outliers.Future work will be focused on detecting outliers where input is dynamic and in multi-granulation sets. The proposed work has some limitations, such that the fixation of threshold value sometimes results in regular objects become outlier and outliers become regular objects.This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

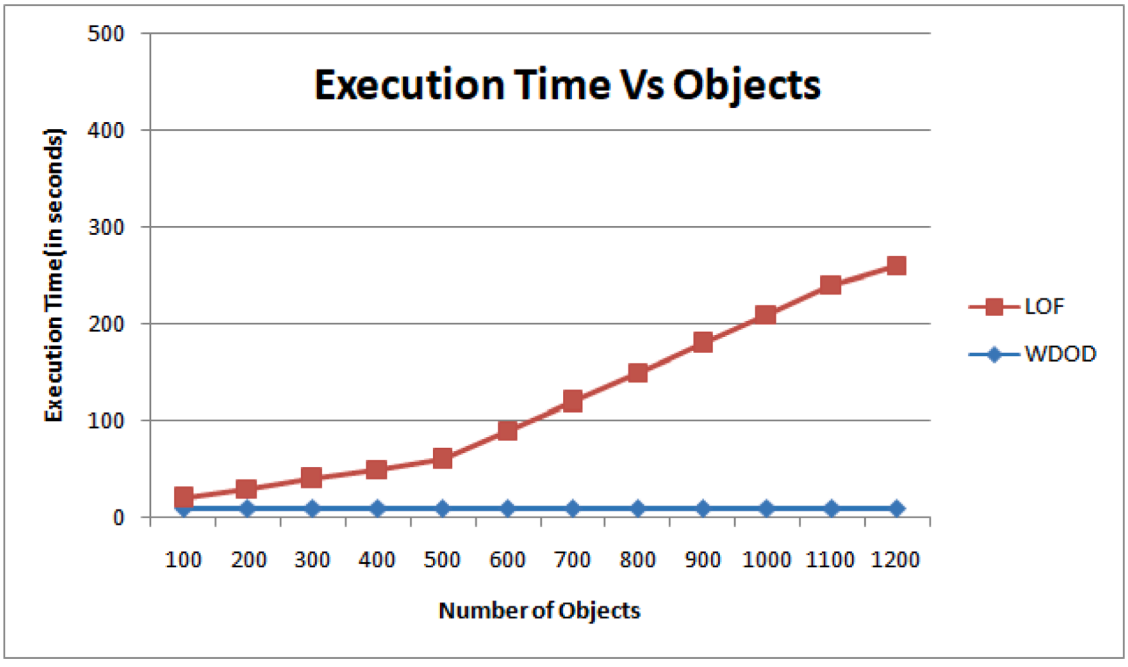


Fig. 7. Comparing execution time as the number of objects grows

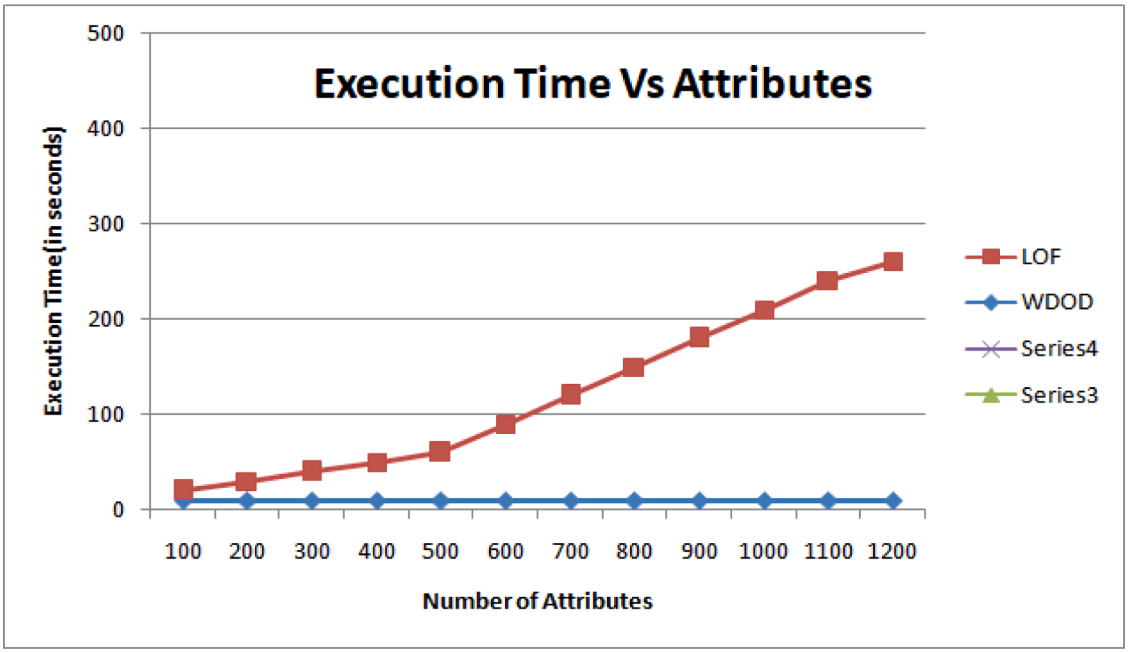


Fig. 8. Comparing execution time as the number of attributes grows

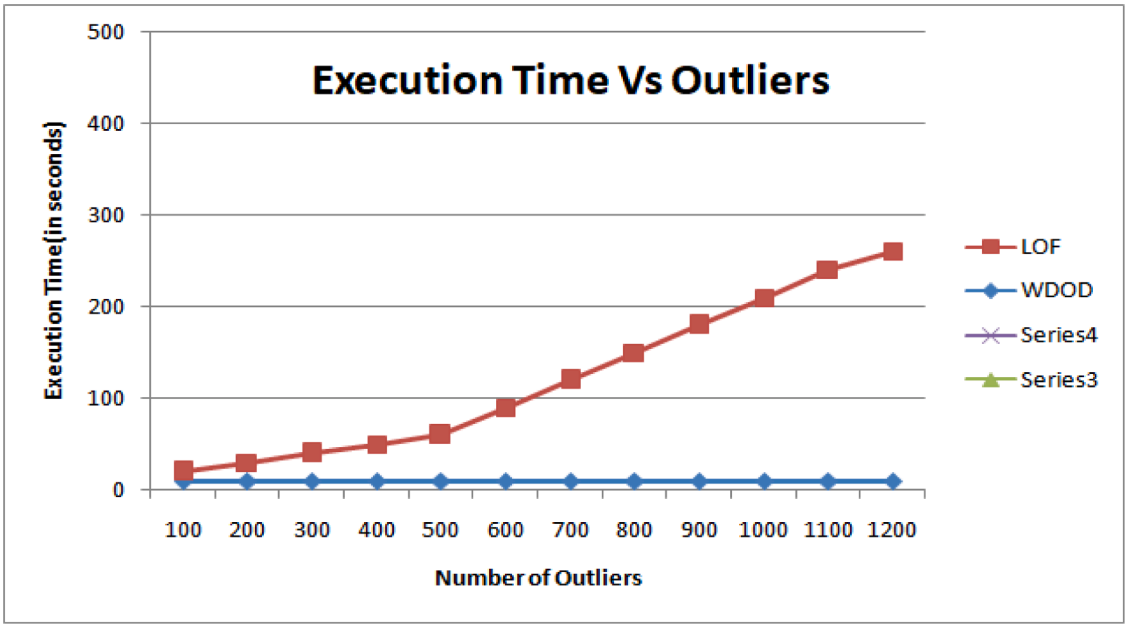


Fig. 9. Comparing execution time with an increased number of outliers

Algorithm 1
The algorithm for the proposed model has been shown below:

Input: Dataset $DS(W,\alpha,\beta)$ and θ be threshold values.
Output: Set S holds outlier objects.

Step 1:Start
Step 2: Input the dataset of mixed type.
Step 3: Use fuzzy proximity relation and ordering to convert numeric into categorical data.
Step 4: Let $S = \varnothing$
Step 5: For each attribute $\beta_i \in \beta$
Step 6: Calculate the indiscernibility function $U/IND(\alpha_i)$ according to definition2;
Step 7: Calculate the complement entropy according to definition3;
Step 8:For each attribute $\beta_i \in \beta$, compute weighted Density according to definition4;
Step 9:For each object $\alpha_i \in W$, calculate the weighted Density according to definition5;
Step 10: If (*Weighted Density*$(\alpha_i) < \theta$)
Step 11: $S = S \cup \{\alpha_i\}$.
Step 12: Return S.
Step 13: Stop.

References

[1] H.P. Achtert, L. Kriegel, E Reichert, R. Schubert, Zimek Wojdanowski, A Visual Evaluation of Outlier Detection Models, in: *Proc.* International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan, 2010. https://link.springer.com/chapter/10.1007/978-3-642-12098-5_34.
[2] C.C. Aggarwal, P.S. Yu, Outlier detection for high dimensional data, in: Proc.ACM-SIGMOD, Int. Conf. Management of Data (SIGMOD'01), 2001, pp. 37–46. Santa Barbara,CA, https://dl.acm.org/doi/abs/10.1145/375663.375668.
[3] A. Arning, R. Agrawal, &P. Raghavan, A linear method for deviation detection in large databases, in: Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, 1996. OR, https://www.aaai.org/Papers/KDD/1996/KDD96-027.pdf.
[4] P. Ashok, &G.M.K. Nawaz, Outlier Detection Method on UCI Repository Dataset by Entropy-Based Rough K-means, J. Defence Science 11 (2016) 113–121.
[5] V. Barnett, &T. Lewis, Outliers in statistical data, John Wiley and sons, 1994.
[6] S.D. Bay, &M. Schwabacher, Mining distance-based outliers in near-linear time with randomization and a simple pruning rule, in: Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Washington, DC., 2003. https://dl.acm.org/doi/abs/10.1145/956750.956758.
[7] R.J. Beckman, R.D. Cook, Outliers Technometrics 25 (2) (1983) 119–149, https://doi.org/10.1080/00401706.1983.10487840.
[8] M.M. Breunig, H.P. Kriegel, J. Sander, Identifying density-based local outliers, in: Proc Acm Sigmod Conference, 2021, pp. 93–104. https://dl.acm.org/doi/abs/10.1145/342009.335388.
[9] V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection A Survey, ACM ComputingSurveys 41 (1) (2011) 58–66, 2011, https://dl.acm.org/doi/abs/10.1145/1541880.1541882.
[10] A.G. Christy, M. Gandhi, S.V. Subramaniyan, Cluster based outlier detection for cluster data 5 (5) (2012) 363–387.
[11] D. Dasgupta, F.A. Nino, Comparison of negative and positive selection algorithms in novel pattern detection, in: Proceedings of the IEEE International Conferenceon Systems, Man, and Cybernetics. Nashville, TN 1, 2000, pp. 125–130. https://ieeexplore.ieee.org/abstract/document/884976.
[12] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf?source=post_page.
[13] F. Jiang, Y. Sui, C. Cao, Outlier Detection Based on RoughMembership Function, Rough Sets and Current Trends in Computing 4259 (2006) 388–397. https://link.springer.com/chapter/10.1007/11908029_41.
[14] S. Forrest, C. Warrender, B. Pearlmutter, Detecting intrusions using system calls: Alternate data models, in: Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society, Washington, DC, USA, 1999, pp. 133–145. https://ieeexplore.ieee.org/abstract/document/766910.
[15] A. Ghoting, S. Parthasarathy, M. Otey, Fast mining of distance-based outliers in high dimensional spaces, in: Proc SIAM Int Conf on Data Mining(SDM) dimensional spaces, Bethesda, ML, 2006. https://link.springer.com/article/10.1007/s10618-008-0093-2.
[16] F.E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics 11 (1) (1969) 19–21. https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657.
[17] D. Hawkins, Identification of outliers, Monographs on Applied Probability and Statistics (1980).
[18] V. Hodge, J.A. Austin, Survey of outlier detection methodologies, Artificial Intelligence Review 22 (2) (2004) 85–126. https://link.springer.com/article/10.1023/B:AIRE.0000045502.10941.a9.
[19] J. Han, M. Kamber & J. Pei, Data Mining concepts and techniques, 2012.
[20] E.M. Knorr, R.T. Ng, A unified approach for mining outliers, in: Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON), Toronto, Canada, 1997. https://www.aaai.org/Papers/KDD/1997/KDD97-044.pdf.
[21] E.M. Knorr, N.G. RT, Algorithms for mining distance-based outliers in large datasets, in: Proc. Int. Conf. on Very Large Data Bases (VLDB), New York, NY, 1998. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.8026&rep=rep1&type=pdf.
[22] E.M. Knorr, R.T. Ng, Finding intensional knowledge of distance-based outliers, in: Proc. Int. Conf. onVeryLargeDataBases(VLDB)Edinburgh Scotland on Very Large Data Bases (VLDB), Edinburgh, Scotland, 1999. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9005&rep=rep1&type=pdf.
[23] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J.A. Srivastava, Comparative study of anomaly detection schemes in network intrusion detection, in: Proceedings of theThirdInternational Conference on Data Mining. *SIAM*, 2021. https://epubs.siam.org/doi/abs/10.1137/1.9781611972733.3.
[24] A. McCallum, K. Nigam, L.H. Ungar, Efficient clustering of high-dimensional data sets with application to reference matching, in: Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Boston, MA, 2000. https://dl.acm.org/doi/abs/10.1145/347090.347123.
[25] M. Markou, S. Singh, Novelty detection: A Review - Part1:Statistical Approaches, Signal Processing 83 (12) (2003) 2481–2497. https://www.sciencedirect.com/science/article/abs/pii/S0165168403002020.
[26] M. Markou, S. Singh, Novelty detection: A Review-Part 2: Neural Network based approaches, Signal Processing 83 (12) (2003) 2499–2521. https://www.sciencedirect.com/science/article/abs/pii/S0165168403002032.
[27] Z. Pawlak, Rough Sets, J. Computer and Information Sciences 11 (1982) 341–356. https://link.springer.com/article/10.1007/BF01001956.
[28] M.I. Petrovskiy, Outlier detection algorithms in data mining systems, Programming and Computer Software 29 (4) (1988) 228–237. https://link.springer.com/article/10.1023/A:1024974810270.
[29] F. Preparata & M. Shamos, Computational Geometry: an Introduction. Springer Verlag.
[30] R. Banshal, N. Gaur, S.N. Singh, Outlier detection-Applications and Techniques in Data Mining, IEEE Conference 44 (12) (2016) 2862–2870. https://ieeexplore.ieee.org/abstract/document/7508146.
[31] R. Kannan, H. Woo, C.C. Aggarwal, Outlier Detection for text data-An extended version 23 (2000) 61–69. https://arxiv.org/abs/1701.01325.
[32] R. Kaur, S. Singh, A review of social network-centric anomaly detection techniques 23 (1) (2016) 61–69. https://www.inderscienceonline.com/doi/abs/10.1504/IJCNDS.2016.080582.
[33] P. J. Rousseeuw, & A.M. Leroy Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.
[34] P.I.J. Ruts, Rousseeuw, Computing depth contours of bivariate point clouds, Computational Statistics and Data Analysis 23 (2021) 153–168. https://www.sciencedirect.com/science/article/abs/pii/S0167947396000278.
[35] D. Synder, MS thesis, Department of Computer Science, Florida State University, 2001.
[36] S. Mitra, Sankar, P. Mitra, Data Mining in soft computing framework: A Survey, J. IEEE transactions on neural networks, 13 (2002) 132. https://ieeexplore.ieee.org/abstract/document/977258.
[37] E. Savage, G. Becker, R. Zamudio, A survey of link mining and anomaly detection 34 (3) (2015) 645–654.
[38] J. Tang, Z. Chen, A.W. Fu, D.W. Cheung, Capabilities of outlier detection schemes in large datasets, framework, and methodologies, Knowledge and Information Systems 11 (1) (2006) 45–84. https://link.springer.com/article/10.1007/s10115-005-0233-6.
[39] V. Kumar, S Kumar, A.K. Singh, Outlier detection – A clustering based approach, IJISME I (7) (2021) 383–387. ISSN:2319-6386.
[40] X. Zhao, J. Liang, F. Cao, A simple and effective outlier detection algorithm for categorical data, J.Machine Learning and Cybernetics 5 (2014) 469–477. https://link.springer.com/article/10.1007%2Fs13042-013-0202-4.
[41] Z.A. Bakar, R. Mohamed, A. Ahmad, M.M. Deris, A comparative study for outlier detection techniques in data mining 10 (4) (2006) 371–395. https://ieeexplore.ieee.org/abstract/document/4017846.
[42] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, in: Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Montreal, Canada, 2021. https://dl.acm.org/doi/abs/10.1145/235968.233324.
[43] J. Zhan, J. Ye, W. Ding, P. Liu, A novel three-way decision model based on utility theory in incomplete fuzzy decision systems, IEEE Transactions on Fuzzy Systems (2021).

[44] K. Zhang, J. Zhan, W.Z. Wu, On multi-criteria decision-making method based on a fuzzy rough set model with fuzzy α-neighborhoods, IEEE Transactions on Fuzzy Systems (2020).
[45] J. Zhan, H. Jiang, Y. Yao, Three-way multi-attribute decision-making based on outranking relations, IEEE Transactions on Fuzzy Systems (2020).
[46] J. Ye, J. Zhan, W. Ding, H. Fujita, A novel fuzzy rough set model with fuzzy neighborhood operators, Information Sciences 544 (2021) 266–297.
[47] W. Wang, J. Zhan, C. Zhang, Three -way decisions based multi-attribute decision making with probabilistic dominance relations, Information Sciences 559 (2021) 75–96.
[48] J. Deng, J. Zhan, W.Z. Wu, A three-way decision methodology to multi-attribute decision-making in multi-scale decision information systems, Information Sciences 568 (2021) 175–198.

Bus journey simulation to develop public transport predictive algorithms

Namrata Khamari, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, namrayakhamari@outlook.com*

Tapas Ranjan Baitharu, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, tkbaitharu11@gmail.com*

Biraja Nayak, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, birajanayak21@gmail.com*

Susmita Mohapatra, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, susmitamohapatra963@gmail.com*

Keywords:
Public transport
Arrival time prediction
Data simulation
Data quality
Machine learning
Deep learning

A B S T R A C T

Encouraging the use of public transport is essential to combat congestion and pollution in an urban environment. To achieve this, the reliability of arrival time prediction should be improved as this is one area of improvement frequently requested by passengers. The development of accurate predictive algorithms requires good quality data, which is often not available. Here we demonstrate a method to synthesise data using a reference curve approach derived from very limited real world data without reliable ground truth. This approach allows the controlled introduction of artefacts and noise to simulate their impact on prediction accuracy. To illustrate these impacts, a recurrent neural network next-step prediction is used to compare different scenarios in two different UK cities. The results show that a realistic data synthesis is possible, allowing for controlled testing of predictive algorithms. It also highlights the importance of reliable data transmission to gain such data from real world sources. Our main contribution is the demonstration of a synthetic data generator for public transport data, which can be used to compensate for low data quality. We further show that this data generator can be used to develop and enhance predictive algorithms in the context of urban bus networks if high-quality data is limited, by mixing synthetic and real data.

1. Introduction

Cities around the world are trying to shift personal traffic to public transport to reduce congestion and environmental impact. A crucial part of such a strategy is to make public transport as convenient as possible. Bus passengers often rely on Real-Time Passenger Information (RTPI) systems at bus stops, online and in mobile apps. These RTPI systems can be unreliable [1] which is inconvenient for passengers. In general, passengers assign different priorities to certain aspects of public transport. Reliability and safety are considered the two most important [2].

The importance of making especially buses as attractive as possible in comparison to private vehicles is highlighted in the historical statistical records. In the UK, 4.8 billion bus trips were made in 2018/19, accounting for 58% of all public transport journeys [3]. These journeys amounted to 27.4 billion km travelled and saved approximately 96 million tonnes of CO_2 [4]. However, since 1985, bus travel has been steadily decreasing by a total of 0.7 billion journeys. As other public transport modes such as trains in most areas cannot be a replacement for local bus services, this suggests that a larger share of passengers opt for private vehicles. This is mirrored in the continuous upward trend of car traffic on British roads [3]. To encourage potential passengers to use public transport, it is crucial to make it as attractive as possible to reverse the above trends, ultimately having a positive impact on the environment as well as congestion levels in urban settings. However, the mentioned data are pre-pandemic, thus the long-term impact of the pandemic on public transport cannot currently be anticipated.

Other studies also highlighted the importance of accurate Estimated Time of Arrival (ETA) predictions to improve customer experience [5]. Many public transport providers have developed mobile apps, which give ‘live’ positions of vehicles. Passengers can use such technology to decide when to leave the house to catch a bus without having long wait times at a bus stop. However, we previously noted the latency of this information caused by delays of wireless network infrastructure and the fact that the data in our operational area passes through a number of 3rd party systems [6]. Therefore, the RTPI system might suggest a vehicle is further away than it is in reality. This could cause a passenger to miss a bus and thus unnecessarily inconvenience them. In Bournemouth, one of the two cities used as an example in this study, the latency of the internet-based ‘live position’ is approximately 30–40 s. To alleviate this issue, we have proposed a short-horizon prediction which will be useful in the further development of ETA and long-term predictions, and in bringing the ‘live’ locations closer to reality. The commonly deployed Automatic Vehicle Location (AVL) systems [7], could supply data for such approaches.

To compare any potential model, the assessment of their performance is of crucial importance, this has to be reported in a way

that allows to replicate and compare the results. However, this is not possible in all cases as some authors report relative errors [8–10] and no consistency in the reported parameters can be distinguished. The precondition for all machine learning algorithms should be verifiable, and the RoyalSociety's report highlights this as a central feature [11]. This has also been recognised in the healthcare sector where guidelines for the development and reporting of predictive models exist [12]. The difference in standards might be explained because ETA predictions do not affect the health or safety of a passenger and a spurious algorithm might at most cause inconvenience rather than physical harm. However, for an operating company, this might cause a loss of revenue through a decline in patronage, and the society as a whole might be subjected to more congestion that could simply be reduced by providing accurate ETA predictions. Furthermore, the doctrine of science is replicability. The reproducibility crisis is most prominently known from psychological research [13] however due to its notoriety, it has been actively addressed [14]. It has also been identified as a problem in 'harder' sciences such as biomedicine [15] and also artificial intelligence [16]. Although results gained from machine learning techniques might be considered hard evidence, because the final model is based on mathematical concepts, they often suffer from similar problems as seen in psychology where the research is often subjective to the researcher. The similarities between the two fields are that the findings cannot usually be explained due to the 'black box' effect. The field of psychology has now started to apply lessons from problems seen in machine learning research [14]. A suggested way of addressing such problems is meta-science that could shed light on the true accuracy of findings [17]. However, this relies on comparable measurements of accuracy, which is not found in a large proportion of the public transport literature. Therefore, comprehensive standards of reporting are urgently needed in the field of predictive bus transportation research. This as a consequence poses the issue that high-quality data is required to develop good predictive models. We and other researchers have highlighted that data quality issues need to be considered in the context of public transport research [6,18–20]. Therefore, in this study we demonstrate a method to synthesis bus journeys based on limited and low quality data. This allows on the one hand to generate a hybrid dataset to develop models from. On the other hand it has the potential to be used to generate synthetic datasets that can be used for benchmarking in an attempt to combat the highlighted replicability issues faced by public transport research.

In our data, a notable lack of quality hampers the development of predictive algorithms. The quality issues include the lack of clear journey identification, linkable to a timetable, artefacts such as gaps in recordings, falsely reported line numbers, and direction of travel (inbound vs. outbound). These quality issues make it impossible to develop accurate predictive algorithms. Unfortunately, the simplest solution of recording high-quality historical data is not feasible due to closed source data collection by 3rd party companies. To address this issue, this study describes a reference curve-based synthetic data generator, which bases its assumptions on limited real-world data. This allows to test algorithms in a controlled environment and enables the injection of user-defined artefacts into the dataset to test their effect on prediction quality. We also show that mixing real and synthetic data improves the prediction accuracy.

2. Background

Methods for ETA prediction can include simple historical averages or be based on statistical models. However, due to the complexity of the ETA prediction, machine learning methods have become increasingly popular [21]. In recent years, artificial Neural Networks (NN) have revolutionised a number of other domains. Therefore, NNs should be expected to have similar potential when applied to bus ETA prediction problems. A comprehensive review specifically investigating NN applications in public transport [22] found that only 16% (12) addressed ETA of buses, whereas the rest of the studies applied the technique to other modes of transport. This suggests that the area of bus ETA prediction using NNs might be underrepresented in the context of public transport research. This relative absence of NNs to predict bus ETA is striking as NNs have revolutionised other areas of data science such as image and speech recognition [23,24].

The challenge of all machine learning approaches is to fine tune the model parameters, one solution is to use genetic algorithms [25] to optimise machine learning algorithms inspired by nature. Several innovative variations have been demonstrated in the recent literature, such as an algorithm inspired by the mating of red deer populations [26], or the simplification of parameter search with a simplified metaheuristic [27]. The same authors also demonstrate methods applicable to supply chain management using the Taguchi method to outperform conventional genetic algorithms [28] as well as the potential use of blockchain algorithms in the management of supply chains [29], additionally they show applications to predict photovoltaic electricity generation [30] as well as bioremediation [31].

Nowadays, the majority of buses have onboard AVL systems, which are equipped with GPS sensors and transmit the location of the bus at frequent intervals, typically ranging between 20 and 60 s. The availability of vehicle locations are the basis for any ETA prediction and are accessible through the AVL system and do not necessarily need any additional investment in static sensors.

The biggest hurdle in developing machine learning solutions generally is the difficulty to acquire enough good-quality data to develop a useful algorithm. In some fields, this has led to the use of simulated data ranging from medicine [32] to geophysics [33]. Regarding public transport journey simulation, the literature is scarce. Some examples related to bus data simulation include bus platooning [34] as well as traffic simulation [35]. However, to the best of our knowledge, no study has investigated the use of simulated data to train a next step prediction model for urban bus networks. In many areas of machine learning research, benchmark datasets are common [36]. These allow researchers to objectively compare algorithms against each other. This is missing in the field of urban bus networks. Therefore, the presented data generator could allow the generation of a standardised benchmark dataset that could lay the foundation for further research in public transport.

3. Real-world data processing

3.1. Data collection

Data is accessible via the infrastructure of our collaborators, and two British cities have been selected with the largest number of vehicles and access to recorded travel data. AVL data was collected from two different bus operators from Reading (UK) line 17 and Bournemouth (UK) line 1 (Fig. 1). Each vehicle transmits its position approximately every 40 s, which is recorded by the company providing the Electronic Ticketing Machines (ETMs) with the integrated AVL-system. Due to data handling by several independent entities, only a limited amount of information is transmitted. The available data are:

- Timestamp
- Position (latitude and longitude)
- Line number
- Direction (outbound or inbound)

For the Bournemouth operator, it became apparent that the transmitted directions are often incorrect and so are the line numbers when a vehicle changes its line during an operational run. The data collected in Reading had a better integrity with reliably transmitted direction, thus simplifying the data processing steps. Based on this limited information, it is not possible to match a vehicle to a timetable corresponding to the journey it is currently serving. A journey is a specific trip found in the

Fig. 1. Location of both example cities and the journey shape used for all experiments. The line 1 in Bournemouth is shown yellow and the line 17 in Reading in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

timetable of a bus line, e.g., the outbound 9 AM service 1. In contrast, a route pattern (also referred to as 'shape') is the route as travelled on the road, which can vary slightly for each journey for the same bus service. In the example of line 1 in Bournemouth, there are several patterns which can include different starting points along the route, resulting in shorter overall journeys or slightly different routes. In both cities, reliably matching a vehicle directly to a specific route pattern is not possible as the unique route pattern identifiers were not accessible to us. Therefore, one route pattern for each city was arbitrarily selected and used to generate synthetic data, which is an acceptable approach as in the selected cities the differences between patterns are negligible.

3.2. Identifying route sections for filtering

The bus route used in Bournemouth is line 1, starting in the town centre towards Christchurch (Fig. 1). The complete route shape includes longer journeys and therefore needs to be truncated. In the second example of Reading line 17 was used, which can have up to 90 different route patterns per direction with different runtimes and minor variations in route shapes (Fig. 1). Additionally, a complicating factor is that the route follows a one-way system in the city centre, meaning that the routes are different depending on the served direction. Therefore, a two-pronged approach was used. To initially filter journeys that were too far away from the shape, all available shapes for both directions were combined to a template shape. Any journey outside a radius of 3 x the mean distance to the template shape was excluded. The final filtering with the ability to enforce the direction was done using an arbitrarily selected route pattern from the many different patterns available for each line covering the entire length of the route. In the case of Reading these route patterns are mostly identical, however, in Bournemouth the patterns can be very different. We have described these issues previously [6].

3.3. Identification of individual journeys

Due to the lack of explicit journey identification, a heuristic approach was used to separate individual journeys that will then be used as a basis to generate synthetic data.

Bournemouth operator does not reliably transmit the direction a vehicle currently serves. However, an observation made was that at the end of a journey vehicles stopped transmitting data for a short period of time. Thus, once it reappears in the data stream, a gap in the timestamps can be detected. A new journey was defined as a time gap of more than 15 min. If such a gap is detected, it is assumed a new journey has started.

Reading operator reliably reports the direction of travel, making the identification of an individual journey easier. Furthermore, vehicles tend to serve the same line and do not change lines between runs, by selecting a single direction, large gaps in transmission timestamps can be observed, making the separation of journeys accurate.

3.4. Trajectory generation

It is assumed that the vehicles follow the identified outbound journey shape. This allows us to represent a journey as a trajectory which is the distance travelled along the route shape. Using such a trajectory, a journey can be represented in two dimensions based on the distance travelled and the run time from the start of the journey.

3.5. Additional processing steps

To ensure a clean dataset, repetitions at the start where the vehicle did not move further than 10 m were removed and a journey is assumed to start once the vehicle has moved further than this threshold. The journey was presumed to have ended as soon as it had reached its maximum trajectory.

4. Synthetic data generation

The data generation process uses a heuristic data-based approach to generate synthetic journeys. This process is broken down into several sub steps:

- The interpolation of the route shape as the reported points are not evenly distributed along the route.
- The identification of the normal run time for a journey is based on historical data, which also allows the identification of delays.
- The probability-based simulation of the delays.

The above steps are described in detail in the following subsections.

4.1. Interpolating the journey based on the route shape

A synthetic journey is generated based on future timetables. To avoid all vehicles starting at the same point, a time offset is added to the start time of the timetable, which is a random number between 0

and 40 s (the transmission interval). This is added to the scheduled start time. The distance that should be offset is then calculated by multiplying the offset by the average speed observed in the real world data 8 m/s (30 km/h). The timestamps are then interpolated to a user-defined interval — 40 s in the presented example. Calculating the time difference between two subsequent stops on the route segment gives the overall runtime. This can be divided by the transmission frequency of 40 s to give the number of transmissions expected on this route section. By assuming the vehicle travels at a constant speed, the progress along the shape can be estimated and the coordinates of the shape at the transmission points can be extracted. However, the coordinates of the reference journey pattern are not equidistant; the distances between consecutive reported locations vary between 6 m and 100 m. Therefore, interpolation solely based on the shape would give very different speeds depending on the road shape. This is avoided by generating an interpolation based on the distance along the route. The closest calculated distance of the shape coordinates is used to calculate the difference between the interpolation coordinate and the shape coordinate. If this distance is greater than 5 m, the two neighbouring points on the shape are used to interpolate the positions between these two coordinates to make the data more realistic. This does not account for variations in the speed or the curvature of the earth, but as the distance is at most 100 m, it is a reasonable omittance. Additionally, it appears that wider gaps are found on straight road sections and the frequency increases in meandering sections, making the proposed approach a good compromise.

4.2. The problem of determining delays

As arrival times at bus stops are not recorded, it cannot be determined whether a vehicle was running on time or was delayed. An additional difficulty is that the journey times vary and depend on the time of the day and weekdays. This variation in timetabled runtime compensates for the expected traffic status. TomTom, a location technology company, records congestion characteristics for different cities based on consumer GPS data. The data for Bournemouth indicates the percentage of delay that needs to be added to a journey at a certain time of day. The maximum in Bournemouth is on a Wednesday afternoon with an expected 71% increase in travel time (pre-pandemic) [37].

Most times of the day, the timetable overestimates the travel time compared to the expected time based on TomTom's data. However, it needs to be kept in mind that the vehicles travel between Bournemouth and Christchurch and the data only accounts for Bournemouth. Furthermore, stops to let passengers board or debark are not considered in the TomTom dataset. This means the timetable accounts for expected variations in traffic conditions and thus cannot be used to simulate vehicle delays.

Another avenue explored was the use of Google services to predict delays based on consumer data, which was not possible as buses travel in bus lanes, making the route very different from a prediction based on Google Maps.

4.2.1. Probability based simulation of delays

By assessing all journeys within the real-world dataset by weekday and hours of day, a reference trajectory can be derived. This reference trajectory is simply the mean trajectory of all observed journeys (Fig. 2(a)). As a result, the outliers are removed and the reference curve represents the baseline of a 'normal' journey (Figs. 2(b) and 2(c)). This allows to calculate the probability that a journey will be delayed or early for every time of each week day. Reference curves were generated using a centred moving 3 h window except for the first and last hour where a truncated window was used. This gives the advantage that the time dependency of delays is simulated, meaning that a vehicle following a delayed bus will most likely also be delayed, thus approximating the delay propagation along a single line.

4.2.2. Journey generation

To generate a journey, the timetables of one week are queried and used as a template. The reason for this approach is that although the timetables for Bournemouth are available until the end of the current calendar year, this is not the case in Reading where only one week is available. As the timetable normally does not drastically change within the same year, this is a justifiable approach. Subsequently, the reference curve queried and the following relevant data points are extracted:

- The mean reference trajectory.
- The standard deviation as well as 95% confidence intervals.
- The probabilities of delayed or early arrival with respect to the reference curve (Fig. 2).

4.2.3. Delays

Based on the reference curve, the probability of a journey being delayed or early can be calculated. Whether a journey is delayed is decided by sampling from a normal distribution for each entry of the reference table, a random number r is generated and stored in a probability list $\{r_o \ldots r_n\}$. These parameters double as a modification parameter to generate the delay or time gain. To remove variations of the list of probabilities, a Savitzky–Golay filter is applied with a window of 7 and a polynomial order of 3. A decision whether a vehicle will be on time, early or delayed is made based on the smoothed probability list. A vehicle will arrive early if $r < p_{early}$. If $p_{early} < r < p_{early} + p_{delayed}$ vehicle is delayed. If neither of the conditions is true, the vehicle is assumed to be on time. To simulate the variations in time gained, the initially expected runtime t of the reference curve is calculated as well as the difference of the last position of the reference curve γ. The ratio of expected variation is calculated based on the confidence interval of the reference curve ν. Thus, the progress along the trajectory under the influence of a time gain can be calculated as follows:

$$\nu = (\sigma_i/\gamma_i) * (R = \{\ )$$
$$P = P_{i-1} + t - (t \times ((0.9 \times \nu) \times 1.25))$$

Where: ν=volatility, γ=reference, P=position, t =expected time at position

If the next position will be delayed, a random modification factor m is generated by sampling from a beta continuous random distribution (α=1, β=2). This tailed distribution was chosen as it makes large reductions in delay less likely and a vehicle will in most circumstances make up no or very little time. The delay *volatility* is defined as the ratio of the reference curve standard deviation to the reference curve itself multiplied by m. Additionally, the delay of the previous step d_{i-1} is calculated and subtracted from the current delay to prevent an exponential increase in delay. To account for random major changes outside the 'norm' of delay or time gains observed in the real data, GPS *noise* is generated using a uniformly sampled random number R which also acts as a weight of the additional delay. Thus, a position with simulated noise can be described as:

$$\eta = \nu \times (R = \{\ \ +1)$$
$$P = P_{i-1} + (t + [(\nu \times m) - d_{i-1} \pm \eta])$$

Where: η= noise to be added, ν=volatility, P=position, t =expected time at next position

If the bus is most likely on time, the probability p of it being on time is used to generate an adjustment towards the reference curve as follows:

$$P = [P_{i-1} + t] - [p \times t]$$

Where: P=position, p = probability a vehicle is on time t =expected time at next position

The generated trajectory is then interpolated to give positions in time intervals of 40 s consistent with the transmission rates of the recorded data.

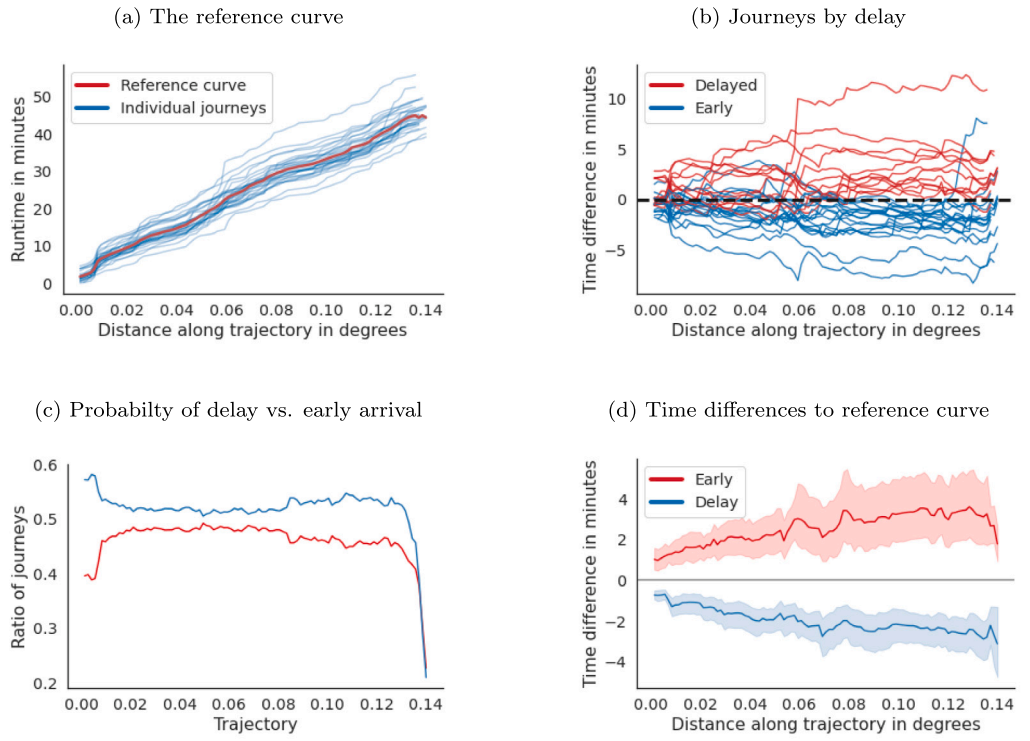


Fig. 2. (a) The historical trajectories of a one day block in Bournemouth (Tuesday 9–12 am). (b) The relative difference from the reference curve along the trajectory. Journeys delayed at more than 60% of the positions are highlighted in red. (c) Probability of travelling early or late on the trajectory. The discrepancy in the sum of the two conditions represents the fraction of vehicles that arrive on time. (c) The average time difference to the reference curve with the uncertainty highlighted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3. Injection of artefacts

The original data is affected by artefacts caused by the behaviour of vehicles as well as data collection issues. Three noteworthy artefacts have been incorporated into the simulation of the synthetic data and are described below.

4.3.1. Injection of GPS noise

GPS recordings are affected by noise which can depend on the surrounding environment, such as high-rise buildings. In the cities used in this study, buildings tend to be low and thus effects due to reflection of the GPS signal are unlikely and have not been observed. To simulate the inaccuracies of the GPS recording, random noise sampled from a normal distribution (mean=0, σ=7) is added to latitude and longitude.

4.3.2. Injection of repeated locations

Due to operational reasons, journeys have scheduled buffers to allow vehicles to catch up with the timetable. This means that the vehicle often repeatedly transmits the same location at the start or end of a journey. At the journey start, 83% of the journeys have repeated locations, whereas end-repetitions are seen in 67% of journeys. The number of repeats varies depending on how long a vehicle is stationary. A skew-normal distribution [38] was fitted to both the start and end repetitions and this reference distribution is used to sample the number of repeats at either end of the journey. This artefact is optional and datasets with as well as without have been generated as in theory it is possible to gather journey data only for the journey itself without buffer times at either end.

4.3.3. Geofencing artefacts

The original data collected contained characteristic circular patterns. We empirically demonstrated previously [6] that the origin of such characteristic artefacts are the geofencing methods used by some AVL-systems to determine if a vehicle has arrived at a bus stop [6]. Unless the bus has been very close to the stop, the AVL-system 'snaps' the real position of the vehicle to a circular geofencing boundary with a radius of 10 m. As this is an unusual artefact, it is generated optionally.

4.4. Data generation

For both cities, datasets were generated for 145 days and for three different conditions:

- a journey only with GPS noise,
- a journey with GPS noise and circular artefacts,
- a journey with GPS noise, and start and end repeats.

Additionally, a hybrid dataset was generated for the city of Reading containing 5000 journeys, of which 50% were synthetically generated and the remaining half were taken from the original dataset.

5. Prediction methods

5.1. Benchmarks

Two naïve benchmark algorithms were used to compare all models against.

Average speed: This method uses the average speed of a vehicle since the start of its current journey. Thus, it does not reflect any short-term speed variation. The calculated speed is used to interpolate the position of the vehicle from the trajectory of its journey pattern for the next 40 s.

Current speed: This method uses the last three transmitted positions of a vehicle to calculate its current average speed, hence accounting for temporary speed variations. The prediction is made by interpolating the position for the next 40 s from the journey trajectory.

5.2. Target representation

The target was represented as a trajectory, by projecting the coordinates onto the route pattern of a journey. This ensures that inaccuracies locating a vehicle off-route are removed. In practice, this method predicts a number representing the progress along the trajectory with a max of 1, which is the final destination. To illustrate the performance of the model, the trajectory can be decoded into coordinates to allow the calculation of a Haversine distance between the predicted and actual location, which is more intuitive than a loss based on the trajectory. Two variations of this target representation were used: **a.** the unconstrained progress along the trajectory, which could lead to a vehicle appearing to move backwards, **b.** the distance travelled in the next time interval added to the last known position, which enforces a forward prediction.

5.3. Input features

The features included were: coordinates normalised to a bounding box representing the operational area of the bus company, the time delta between consecutive recordings, the elapsed time from the start of the journey, and time embeddings as described below. The input features were min–max normalised.

5.4. Handling of time

The time information was split into its components to make it possible for the algorithms to learn periodic patterns. To achieve this, the timestamp was translated into the minute of the day, the hour of the day, and day of the week. These were embedded in a multidimensional space as detailed in the architecture description 5.6.

5.5. Input windows

A moving window was applied to each journey. The window size was a minimum of 10 data points growing by one time step at a time until the end of the journey. This ensures a realistic simulation of the progress of a journey as would be observed in a real world application.

5.6. Architecture

Two neural networks were used with identical architecture except for the Recurrent Neural Network (RNN) module [39], which was either a Gated Recurrent Unit (GRU) [40] or a Long Short Term Memory (LSTM) network [41]. The time embeddings were learned by the network in a multidimensional space. The dimensions were chosen as half of the possible number of values for each embedded variable. As an example, the hour of the day was embedded in 12 dimensions as the maximum number of hours is 24. These embeddings with a total of 52 dimensions were fed into a linear layer to reduce their dimensions back to the original number of time-based features. The output of the linear layer was concatenated with the remaining input features and fed into either a GRU or LSTM layer followed sequentially by a 1D batchnorm, a linear layer, a leaky ReLU, a second batchnorm and a final linear layer. To ensure the outputs were bounded, a sigmoid function was applied.

5.7. Hyper-parameters

To allow for direct comparison between the models, all training hyper-parameters were kept constant between the two cities. It is appreciated that this might not always yield the best performance but will illustrate the influence of the modifications made on the performance. The variables used were chosen through empirical exploration following the recommendations described by [42]. Each model was trained for 50 epochs using the one-cycle policy [42] with a maximum learning rate of 10^{-1} (Bournemouth) and 10^{-2} (Reading). As a loss function, the mean average error (MAE) was used.

6. Results and discussion

It is crucial to compare predictive algorithms using several different metrics to ensure a balanced interpretation of the results. Furthermore, it has to be kept in mind that in the presented example the two cities are considerably different. The most striking difference is the practice regarding journey shapes. The idea behind a journey shape is that it gives the exact route along the road of a certain journey. This, however, is handled differently by the bus operators. In the example of Reading each journey has an individual shape amounting to 90 shapes a day. These are mostly very similar or identical. In the example of Bournemouth fewer shapes are used, however, the shapes are significantly different in length as well as route, highlighting the need for standardisation of public transport data. As a result, only a subset of the journeys in Bournemouth are similar enough to be simulated in one approach, thus this dataset contains fewer journeys than the dataset generated for Reading (17,115 vs 7839 journeys). These differences have to be kept in mind and are crucial for the interpretation of the results. The median accuracies for mean speed benchmarks in Reading are lower in all datasets compared to the current speed benchmark and are shown in Fig. 3. The current speed benchmark for Bournemouth is comparable to the average speed benchmark. In the example of Reading this is not the case and the current speed benchmark suffers from higher prediction errors compared to the average speed benchmark (Fig. 3). An explanation could be that vehicles in Reading are more likely to stop for brief periods, which is reflected in a 13% increase of standard deviation of the travelling speed compared to Bournemouth. Interestingly, the histogram for the Reading benchmarks shows a peak around 80 m for the dataset with repeated start and ends (Fig. 4). This is explained by the benchmarking method, which uses the last three positions to estimate the average speed. Thus, a vehicle's speed can change from stationary to moving within 120 s or vice versa. Considering this time frame, 80 m/120 s corresponds to an average speed of 24 km/h, which is a realistic prediction for an urban bus network and in accordance with the estimated speed from the mean speed benchmark (Figs. 3 & 4)).

6.1. Perfect journeys

The first set of experiments shows the 'perfect' synthetic journey. These are generated without any of the discussed artefacts and therefore, should represent the simplest prediction problem. Poor performance of both architectures can be observed in the Bournemouth dataset. Both architectures perform virtually identical with a mean error of 63.8 m (σ=55 m) (Fig. 5(a)). This is an accuracy comparable to the benchmarks (current speed: 64.2 m, mean speed: 62.1 m). This underwhelming performance could be explained by the smaller dataset compared to the Reading data, however, a more likely explanation is the variability of the journey shape and routes in Bournemouth, which naturally results in less realistic synthetic data. As a consequence, it is difficult to identify individual journeys from the original data. Furthermore, the data generation suffers from the fact that the vehicles do not follow a consistent route, which would be expected to cause unrealistic synthetic journeys. In contrast, the prediction for Reading performs well with a mean error of 41.5 m (σ=46.5) and 47.5 m (σ=47.2) for the GRU and LSTM respectively (Fig. 5(a)). Both models significantly improve on the error compared to the benchmark (current speed: 68 m, mean speed 50.7 m). As mentioned previously, this dataset contains more journeys per day, however, the most likely explanation of this performance improvement can be attributed to the uniform journey shape, which will reduce errors in the data generation.

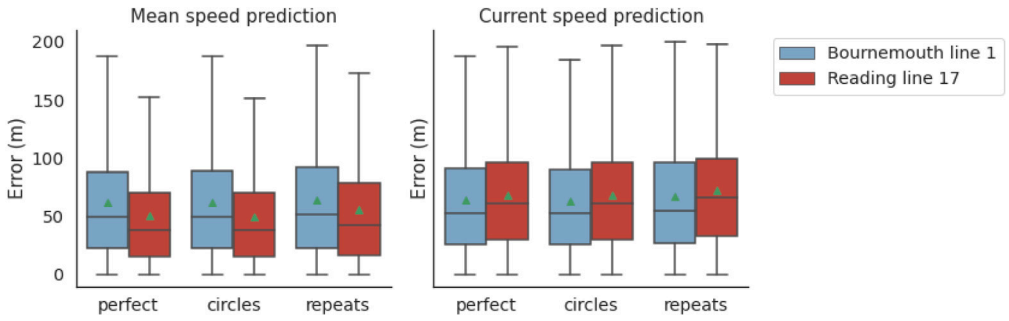


Fig. 3. Boxplot illustrating the prediction errors of the two nïve benchmark algorithms for both cities.

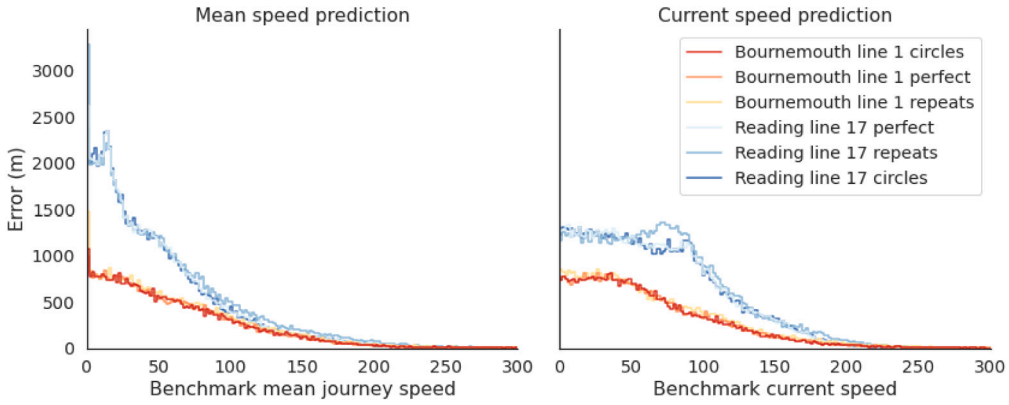


Fig. 4. Boxplot illustrating the prediction errors of the two nïve benchmark algorithms for both cities.

6.2. Ticketing machine artefacts

The introduction of the characteristic circular artefacts into the dataset would be expected to make any prediction more difficult. This is indeed observed in the predictions for Bournemouth. The average GRU performance was reduced by 2.5 m compared to the artefact free journeys. Notably, the performance of the LSTM did not significantly decrease and remained at 63.9 m (Fig. 5(a)). Similar findings were observed in Reading where the mean error of the GRU increased by 5 m. Interestingly, the mean error of the LSTM decreased by 2 m.

6.3. Repeats at start and end

The introduction of repeats at the start and end of the journey did have a strong impact on the prediction performance. The mean prediction error in Bournemouth increased by 5 m and 2 m for the GRU and LSTM, respectively. In Reading, the GRU prediction worsened drastically by 24 m, whereas the LSTM was not affected and remained at 47.8 m (Fig. 5). This is an intuitive response of the LSTM which, due to its ability to forget irrelevant information, is able to focus on the data relevant for the next step prediction.

6.4. Using hybrid data to improve predictions

The described hybrid dataset was used to demonstrate a possible application. As an intuition, it was assumed that the addition of synthetic data, which are cleaner and not affected by uncontrollable artefacts, should improve the overall prediction. When using an unconstrained prediction along the trajectory, this however is not observed and a model trained on purely synthetic or hybrid data performs worse on inference on real data (Fig. 5). This, however, is not the case if the prediction is forced forward as described in Section 4.4. If the prediction space is limited, an improvement in the inference accuracy of networks trained on both the real world dataset can be observed both in the purely synthetic and the hybrid dataset. The largest improvement can be observed if hybrid data were used for training (Fig. 5(b)).

6.5. Discussion of results

The results of this study show that the addition of synthetic data can improve predictive algorithms, which suffer from data quality issues. The use of synthetic data is used in many settings [43], such as healthcare settings to preserve privacy [44] but is also used in the assessment of algorithms such as feature selection methods where the control of features is important [45]. Some authors have also used synthetic data to estimate the upper theoretical limits of predictive algorithms [46]. The generation of hybrid datasets consisting of both real and synthetic data is less common, but examples such as from computer vision exist [47] or for classification problems with heavily unbalanced data [48]. Furthermore, some studies used synthetic data to augment small datasets, for example to improve pandemic datasets and the associated machine learning models [49]. Examples from the field of public transport are rare and mostly focus on optimisation of transport networks and specifically bus routes to minimise delays [50–52]. However, in general, a knowledge gap appears to prevent the combination of simulated data with machine learning algorithms [53], which could be beneficial to improve many areas especially in public transport research. This study demonstrates the use of such hybrid datasets to improve prediction quality. Furthermore, it highlights the lack of framework previously noted by us [54]. A prediction accuracy comparison with the wider literature for this study is not possible as similar research aims to solve different problems. The reason for this is that the research focus regarding short horizon predictions are focused on time frames of >5 min [55,56] or are defined as a distance rather than a time horizon [57]. Shorter prediction horizons are found in the literature but are aimed at predicting different metrics such as speed [58] or the elimination of bus-bunching [59]. As there are, to the author's knowledge, no examples in the literature predicting the position of urban buses in an ultrashort prediction horizon, a comparison with other studies cannot be drawn. Additionally, this study does not claim predictive superiority but demonstrates that the use of hybrid

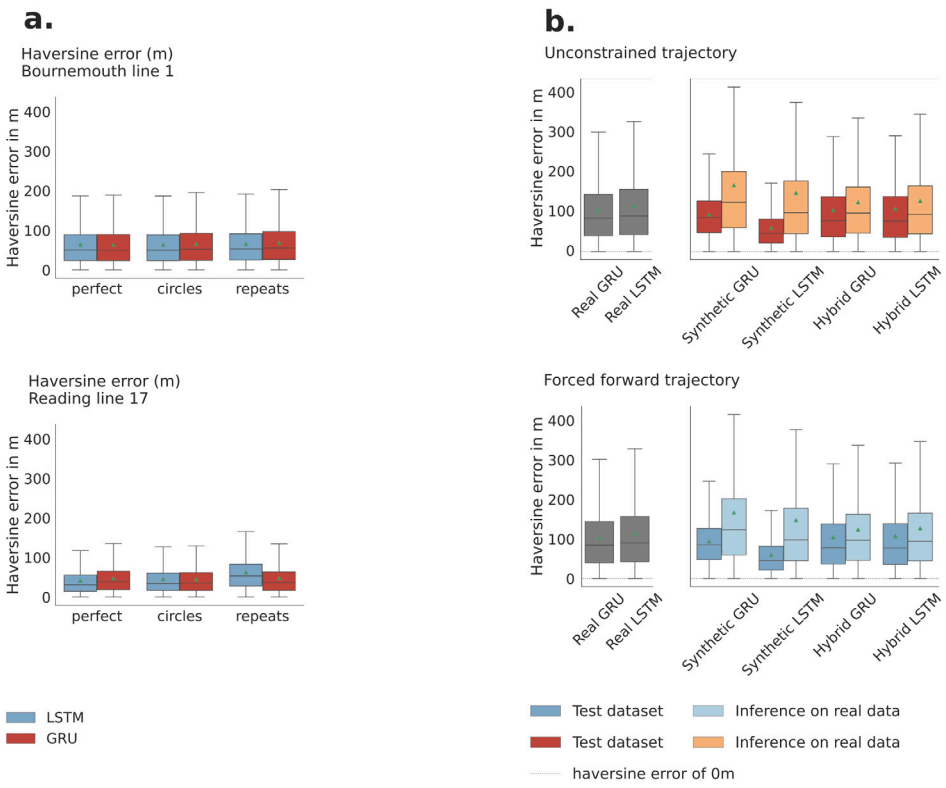


Fig. 5. **(a)** Boxplots for both cities and for each of the dataset and network architecture combinations. It is apparent that the performance in Reading is considerably better and the expected deterioration with the introduction of artefact can be observed. **(b) top:** Boxplots showing the error ranges in meters for the unconstrained networks the grey boxes show a network trained on real data as reference. The red boxes show the error of the holdout portion of the synthetic or hybrid dataset the orange boxes show the inference errors on the real dataset. **(b) bottom:** Boxplots showing the error ranges in meters for the forced forward networks the grey boxes show a network trained on real data as reference. The dark blue boxes show the error of the holdout portion of the synthetic or hybrid dataset the light blue boxes show the inference errors on the real dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data can improve prediction accuracy. This knowledge will be of value to public transport researchers and can be applied to any prediction problem as well as to any model architecture to push the limits of the available data.

7. Conclusion

The importance of making public transport as convenient as possible is self-evident and could help increase passenger numbers and reduce urban congestion and pollution. Reliable predictions of current vehicle position and arrival times play a crucial part in this endeavour. However, this is being inhibited by the lack of reliable data, making any such algorithm development difficult.

Therefore, the described method of generating realistic journeys builds a bridge between the low quality recordable data and the real world. As a result, it is a platform to develop algorithms in a simulated and controlled environment, which can later be deployed in a real world scenario. Additionally, this platform allows simulation of user-specified artefacts as demonstrated by the repetition of positions or geofencing based disturbances. This study has highlighted several areas of improvement for urban bus network data to allow the development of reliable predictive solutions. The most striking observation was that any RNN based predictions in Bournemouth barely outperformed the naïve benchmark. This is due to the varied route shapes and lengths of the same bus line, making generalisation unfeasible. Thus, it can be recommended from a managerial as well as software development point of view that either route shapes should be standardised between the lines or that the lines are subdivided based on their route shapes. This will greatly improve the potential of the data collected and the development of data-based software solutions.

The second observation was that the prediction performance can be improved if the data is as clean as possible. This means that technology providers need to collaborate to ensure the best possible outcome for public transport as a whole. Although geofencing methods to determine the arrival at a stop are useful, the produced artefacts of some systems do have a negative impact on the tested predictive algorithms. Furthermore, an indication whether a vehicle has started or ended a journey will help in the overall prediction accuracy. The differences between the two example cities highlight the need for a national standard if accurate predictions are desired, universally preventing the need to develop a predictive system from the ground up for each city and operational line. This would be a big step forward to an implementation of mobility as a service and would benefit all public transport operators.

The limitations of this study are that the ground truth can only be approximated due to the lack of high-quality data. This, however, is also the driving force behind the demonstrated approach to further advance this research and any other research relying on public transport data, the following key points should be considered for future research:

- Develop a standardised framework to transmit and record public transport data.
- Standardise the use of route patterns to ensure they can be used for data driven applications.
- Develop a benchmarking framework specifically for predictive algorithms in urban bus networks.

In the meantime, until such standardisations become reality, our data generation method described here is a good approximation of reality and a useful tool in simulating effects on urban bus networks.

References

[1] M.M. Salvador, M. Budka, T. Quay, Automatic transport network matching using deep learning, Transp. Res. Proc. 31 (2016) (2018) 67–73, http://dx.doi.org/10.1016/j.trpro.2018.09.053, URL https://linkinghub.elsevier.com/retrieve/pii/S2352146518301273.

[2] G.-J. Peek, M. van Hagen, Creating synergy in and around stations: Three strategies for adding value, Transp. Res. Rec. J. Transp. Res. Board 1793 (1) (2002) 1–6, http://dx.doi.org/10.3141/1793-01, URL http://trrjournalonline.trb.org/doi/10.3141/1793-01.

[3] Department for Transport, Transport Statistics Great Britain 2019 Moving Britain Ahead, Tech. rep., 2019, p. 10.

[4] Department for Business Energy & Industrial Strategy, Greenhouse gas reporting: conversion factors 2019, Research and Analysis (2019) URL https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2019.

[5] R.G. Mishalani, M.M. Mccord, J. Wirtz, Passenger wait time perceptions at bus stops : Empirical results and impact on evaluating real- time bus arrival information, J. Publ. Transp. 9 (2) (2006) 89–106, http://dx.doi.org/10.5038/2375-0901.9.2.5.

[6] T. Reich, M. Budka, D. Hulbert, Impact of data quality and target representation on predictions for urban bus networks, in: 2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, IEEE, 2020, pp. 2843–2852, http://dx.doi.org/10.1109/SSCI47803.2020.9308166, URL https://ieeexplore.ieee.org/document/9308166/.

[7] M. Hickman, Bus automatic vehicle location (AVL) systems, in: Assessing the Benefits and Costs of ITS, Kluwer Academic Publishers, Boston, 2006, pp. 59–88, http://dx.doi.org/10.1007/1-4020-7874-9{_}5, URL http://link.springer.com/10.1007/1-4020-7874-9_5.

[8] Z. Junyou, W. Fanyu, W. Shufeng, Application of support vector machine in bus travel time prediction, Int. J. Syst. Eng. 2 (1) (2018) 21–25, http://dx.doi.org/10.11648/j.ijse.20180201.15, URL https://www.tandfonline.com/doi/full/10.1080/21680566.2017.1353449.

[9] J. Li, J. Gao, Y. Yang, H. Wei, Bus arrival time prediction based on mixed model, China Commun. 14 (5) (2017) 38–47, http://dx.doi.org/10.1109/CC.2017.7942193, URL http://ieeexplore.ieee.org/document/7942193/.

[10] L. Meng, P. Li, J. Wang, Z. Zhou, Research on the prediction algorithm of the arrival time of campus bus, in: Advances in Intelligent Systems Research, AISR, Vol. 142, 2017, pp. 31–33.

[11] The Royal Society, Machine learning: the power and promise of computers that learn by example, in: Report By the Royal Society, Vol. 66, The Royal Society, 2017, p. 125, http://dx.doi.org/10.1126/scitranslmed.3002564, URL https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

[12] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, T.B. Ho, S. Venkatesh, M. Berk, Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view., J. Med. Int. Res. 18 (12) (2016) e323, http://dx.doi.org/10.2196/jmir.5870, http://www.ncbi.nlm.nih.gov/pubmed/27986644.

[13] M. Baker, Over half of psychology studies fail reproducibility test, Nature (2015) http://dx.doi.org/10.1038/nature.2015.18248, http://www.nature.com/doifinder/10.1038/nature.2015.18248 http://www.nature.com/articles/nature.2015.18248.

[14] T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: Lessons from machine learning, Perspect. Psychol. Sci. 12 (6) (2017) 1100–1122, http://dx.doi.org/10.1177/1745691617693393.

[15] M. Baker, 1,500 Scientists lift the lid on reproducibility, Nature 533 (7604) (2016) 452–454, http://dx.doi.org/10.1038/533452a, URL http://www.nature.com/doifinder/10.1038/533452a.

[16] M. Hutson, Artificial intelligence faces reproducibility crisis unpublished code and sensitivity to training conditions make many claims hard to verify, Science 359 (6377) (2018) 725–726, http://dx.doi.org/10.1126/science.359.6377.725, URL http://www.ncbi.nlm.nih.gov/pubmed/29449469.

[17] J.W. Schooler, Metascience could rescue the 'replication crisis', Nature 515 (7525) (2014) 9, http://dx.doi.org/10.1038/515009a, URL http://www.nature.com/doifinder/10.1038/515009a.

[18] W. Rasdorf, H. Cai, C. Tilley, S. Brun, H. Karimi, F. Robson, Transportation distance measurement data quality, J. Comput. Civ. Eng. 17 (2) (2003) http://dx.doi.org/10.1061/(ASCE)0887-3801(2003)17:2(75).

[19] S. Robinson, B. Narayanan, N. Toh, F. Pereira, Methods for pre-processing smartcard data to improve data quality, Transp. Res. C 49 (2014) http://dx.doi.org/10.1016/j.trc.2014.10.006.

[20] R. Arbex, C.B. Cunha, Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data, J. Transp. Geogr. 85 (2020) http://dx.doi.org/10.1016/j.jtrangeo.2020.102671.

[21] R. Choudhary, A. Khamparia, A.K. Gahier, Real time prediction of bus arrival time: A review, in: 2016 2nd International Conference on Next Generation Computing Technologies, NGCT, IEEE, 2016, pp. 25–29, http://dx.doi.org/10.1109/NGCT.2016.7877384, URL http://ieeexplore.ieee.org/document/7877384/.

[22] E. Pekel, S.S. Kara, A comprehensive review for artificial neural network application to public transportation, Sigma J. Eng. Nat. Sci. 35 (1) (2017) 157–179.

[23] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, 2014, pp. 1–9, http://dx.doi.org/10.1007/s10107-014-0839-0, URL http://arxiv.org/abs/1409.3215.

[24] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging 9 (4) (2018) 611–629, http://dx.doi.org/10.1007/s13244-018-0639-9.

[25] O. Theophilus, M.A. Dulebenets, J. Pasha, Y.y. Lau, A.M. Fathollahi-Fard, A. Mazaheri, Truck scheduling optimization at a cold-chain cross-docking terminal with product perishability considerations, Comput. Ind. Eng. 156 (March) (2021) http://dx.doi.org/10.1016/j.cie.2021.107240.

[26] A.M. Fathollahi-Fard, M. Hajiaghaei-Keshteli, R. Tavakkoli-Moghaddam, Red deer algorithm (RDA): a new nature-inspired meta-heuristic, Soft Comput. 24 (19) (2020) 14637–14665, http://dx.doi.org/10.1007/s00500-020-04812-z.

[27] A.M. Fathollahi-Fard, M. Hajiaghaei-Keshteli, R. Tavakkoli-Moghaddam, The social engineering optimizer (SEO), Eng. Appl. Artif. Intell. 72 (April) (2018) 267–293, http://dx.doi.org/10.1016/j.engappai.2018.04.009.

[28] M.R. Islam, S.M. Ali, A.M. Fathollahi-Fard, G. Kabir, A novel particle swarm optimization-based grey model for the prediction of warehouse performance, J. Comput. Design Eng. 8 (2) (2021) 705–727, http://dx.doi.org/10.1093/jcde/qwab009.

[29] J. Moosavi, L.M. Naeni, A.M. Fathollahi-Fard, U. Fiore, Blockchain in supply chain management: a review, bibliometric, and network analysis, Environ. Sci. Pollut. Res. (July) (2021) http://dx.doi.org/10.1007/s11356-021-13094-3.

[30] N. Ghadami, M. Gheibi, Z. Kian, M.G. Faramarz, R. Naghedi, M. Eftekhari, A.M. Fathollahi-Fard, M.A. Dulebenets, G. Tian, Implementation of solar energy in smart cities using an integration of artificial neural network, photovoltaic system and classical delphi methods, Sustainable Cities Soc. 74 (2021) 103149, http://dx.doi.org/10.1016/j.scs.2021.103149, https://www.sciencedirect.com/science/article/abs/pii/S2210670721004315 https://linkinghub.elsevier.com/retrieve/pii/S2210670721004315.

[31] M. Mohammadi, M. Gheibi, A.M. Fathollahi-Fard, M. Eftekhari, Z. Kian, G. Tian, A hybrid computational intelligence approach for bioremediation of amoxicillin based on fungus activities from soil resources and aflatoxin B1 controls, J. Environ. Manag. 299 (2021) 113594, http://dx.doi.org/10.1016/j.jenvman.2021.113594, URL https://linkinghub.elsevier.com/retrieve/pii/S030147972101656X.

[32] J.M. Huttunen, L. Kärkkäinen, H. Lindholm, Pulse transit time estimation of aortic pulse wave velocity and blood pressure using machine learning and simulated training data, PLoS Comput. Biol. 15 (8) (2019) e1007259, http://dx.doi.org/10.1371/journal.pcbi.1007259.

[33] K.B. Withers, M.P. Moschetti, E.M. Thompson, A machine learning approach to developing ground motion models from simulated ground motions, Geophys. Res. Lett. 47 (6) (2020) http://dx.doi.org/10.1029/2019GL086690, https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2019GL086690 https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL086690.

[34] G. Sethuraman, X. Liu, F.R. Bachmann, M. Xie, A. Ongel, F. Busch, Effects of bus platooning in an urban environment, in: 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 974–980, http://dx.doi.org/10.1109/ITSC.2019.8917041.

[35] Y. Ding, S.I.-J. Chien, N.A. Zayas, Simulating bus operations with enhanced corridor simulator: Case study of new jersey transit bus route 39, Transp. Res. Rec. (1731) (2000) 104–111, http://dx.doi.org/10.3141/1731-13.

[36] P. Ristoski, G.K.D. De Vries, H. Paulheim, A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9982 LNCS, Springer Verlag, 2016, pp. 186–194, http://dx.doi.org/10.1007/978-3-319-46547-0{_}20, URL https://link.springer.com/chapter/10.1007/978-3-319-46547-0_20.

[37] TomTom, Bournemouth traffic report | TomTom traffic index, 2020, Bournemouth Traffic URL https://www.tomtom.com/en_gb/traffic-index/bournemouth-traffic/.

[38] A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew normal distribution, J. R. Stat. Soc. Ser. B Stat. Methodol. 61 (3) (1999) 579–602, http://dx.doi.org/10.1111/1467-9868.00194.

[39] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Read. Cogn. Sci. A Perspect. Psychol. Artif. Intell. (V) (2013) 399–421, http://dx.doi.org/10.1016/B978-1-4832-1446-7.50035-2.

[40] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Vol. 4, (January) Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, pp. 1724–1734, http://dx.doi.org/10.3115/v1/D14-1179, URL http://arxiv.org/abs/1409.3215 http://aclweb.org/anthology/D14-1179.

[41] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[42] L.N. Smith, A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018, pp. 1–21, arXiv CoRR abs/1803.09820 URL http://arxiv.org/abs/1803.09820.

[43] M. Hittmeir, A. Ekelhart, R. Mayer, On the utility of synthetic data: An empirical evaluation on machine learning tasks, PervasiveHealth Pervasive Comput. Technol. Healthc. (2019) http://dx.doi.org/10.1145/3339252.3339281.

[44] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, G. Epelde, Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing, JMIR Med. Inform. 8 (7) (2020) http://dx.doi.org/10.2196/18910.

[45] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowl. Inf. Syst. 34 (3) (2013) 483–519, http://dx.doi.org/10.1007/s10115-012-0487-8.

[46] Y.I. Kuchin, R.I. Mukhamediev, K.O. Yakunin, One method of generating synthetic data to assess the upper limit of machine learning algorithms performance, Cogent Eng. 7 (1) (2020) http://dx.doi.org/10.1080/23311916.2020.1718821.

[47] D. Dai, C. Sakaridis, S. Hecker, L. Van Gool, Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding, Int. J. Comput. Vis. 128 (5) (2020) 1182–1204, http://dx.doi.org/10.1007/s11263-019-01182-4.

[48] H. Kaur, H.S. Pannu, A.K. Malhi, A systematic review on imbalanced data challenges in machine learning: Applications and solutions, ACM Comput. Surv. 52 (4) (2019) http://dx.doi.org/10.1145/3343440.

[49] H.P. Das, R. Tran, J. Singh, X. Yue, G. Tison, A. Sangiovanni-Vincentelli, C.J. Spanos, Conditional synthetic data generation for robust machine learning applications with limited pandemic data, 19, 2021, URL http://arxiv.org/abs/2109.06486.

[50] R.K. Kalle, P. Kumar, S. Mohan, M. Sakata, Simulation-driven optimization of urban bus transport, WIT Trans. Built Environ. 186 (2019) 97–108, http://dx.doi.org/10.2495/UT190091.

[51] S.M.H. Moosavi, A. Ismail, C.W. Yuen, Using simulation model as a tool for analyzing bus service reliability and implementing improvement strategies, PLoS One 15 (5) (2020) 1–26, http://dx.doi.org/10.1371/journal.pone.0232799.

[52] S.R. Pells, An approach to the simulation of bus passenger journey times for the journey to work, Transp. Plan. Technol. 14 (1) (1989) 19–35, http://dx.doi.org/10.1080/03081068908717411, https://www.witpress.com/Secure/elibrary/papers/UT19/UT19009FU1.pdf http://www.tandfonline.com/doi/abs/10.1080/03081068908717411.

[53] L. von Rueden, S. Mayer, R. Sifa, C. Bauckhage, J. Garcke, Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12080 LNCS, Springer International Publishing, 2020, pp. 548–560, http://dx.doi.org/10.1007/978-3-030-44584-3{_}43.

[54] T. Reich, M. Budka, D. Robbins, D. Hulbert, Survey of ETA prediction methods in public transport networks, 2019, arXiv.

[55] T. Zhou, W. Wu, L. Peng, M. Zhang, Z. Li, Y. Xiong, Y. Bai, Evaluation of urban bus service reliability on variable time horizons using a hybrid deep learning method, Reliab. Eng. Syst. Saf. 217 (May 2021) (2021) 108090, http://dx.doi.org/10.1016/j.ress.2021.108090.

[56] E. Hans, N. Chiabaut, L. Leclercq, R.L. Bertini, Real-time bus route state forecasting using particle filter and mesoscopic modeling, Transp. Res. C 61 (2015) 121–140, http://dx.doi.org/10.1016/j.trc.2015.10.017.

[57] C. Coffey, A. Pozdnoukhov, F. Calabrese, Time of arrival predictability horizons for public bus routes, in: 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science 2011, CTS'11, in Conjunction with ACM SIGSPATIAL GIS 2011, 2011, pp. 1–5, http://dx.doi.org/10.1145/2068984.2068985.

[58] Q. Ye, W.Y. Szeto, S.C. Wong, Short-term traffic speed forecasting based on data recorded at irregular intervals, IEEE Trans. Intell. Transp. Syst. 13 (4) (2012) 1727–1737, http://dx.doi.org/10.1109/TITS.2012.2203122.

[59] B. Varga, T. Tettamanti, B. Kulcsár, Energy-aware predictive control for electrified bus networks, Appl. Energy 252 (August) (2019) 113477, http://dx.doi.org/10.1016/j.apenergy.2019.113477.

The energy of a photon, on the geometrical perspective

Soumya Mishra, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, soumyamishra96@gmail.com*

Malaya Tripathy, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, malaya.tripathy43@gmail.com*

Bhagaban Sri Ramakrishna, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, maheswarinath1@outlook.com*

Namrata Khamari, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, namrayakhamari@outlook.com*

ARTICLE INFO

Keywords:
Geometrical energy-form
Illusive mass
Morphing
Photon's energy
Rest mass
Wave-particle duality

ABSTRACT

The complex number in quantum mechanics and Riemannian geometry in relativity theory is the basic framework for understanding the physical reality on small and large scales. Special theory of relativity has been produced the famous mass-energy equivalence relation. This article investigates photon's energy theoretically based on photon's rest mass. And represents the energy equation in terms of a complex number and constitutes by the surface's geometry. Photon possesses both wave nature and particle nature; due to this reason, photon reveals its various mass forms as zero rest mass, as real and imaginary rest mass. Indeed, the rest mass and energy of the photon is a complex number on the surface of the matter. The photon's energy depends upon the scalar curvature of the surface of the matter as well as on the wavelength of the photon. In reality, the photon itself reveals the illusion posing with its rest mass, and alongside, it hides itself energy. And this energy equation followed all energy forms of the photon with certain norms.

1. Introduction

In order to understand the nature of reality, we generally use real number and we consider the imaginary number to elucidate some mysterious phenomena like black magic which have no relation with reality. In Physics, a complex number has been used to elaborate the real phenomena. In quantum mechanics, the complex number plays a pivotal role to describe the dynamics of sub-atomic world. In such phenomena, the state vector and wave function are designated by the complex number [1]. The role of complex number has been well established in the digital systems as well [2]. According to Roger Penrose, complex numbers are more fundamental than the real numbers [3]. Another mathematical framework, geometry, became an integral part of modern physics. Riemannian geometry is the building block of Einstein's general theory of relativity. Further, quantum field theory, quantum gravity, string theory has been formulated based on the geometrical aspect [4,5]. This article unfolds the unrevealed reality of the photon's energy in terms of the complex numbers, manifesting in the geometrical flavor. Under specific conditions, the energy equation of photon, leads to the usual energy equation of photon. It is well established that mass and energy are the same things with different manifestation [6,7]. Here, the energy equation of photon is based on the illusive mass of the photon [8,9].

Although the rest mass of a photon is zero in free space [10], there is an experimental evidence that it has non-zero real rest mass [11–13] and imaginary rest mass in a transparent medium [14]. According to recent experimental data, the value of photon rest mass is 1.5×10^{-54} Kg [12]. The value of photon's imaginary rest mass is significantly less and is comparable with the rest mass of the neutrino [14]. Photon's rest mass in free space also depends upon the wavelength [13]. These various mass forms of the photon are explained by one rest mass equation of photon that is called as the illusive mass equation [8,9].

According to special theory of relativity, the rest mass of the particle is considered to be $m_0 = m\sqrt{1 - v^2/c^2}$ where m is the relativistic mass, v is the velocity of the particle and c is the speed of light in free space. The energy of the particle would be

$E = mc^2$, the linear combination of kinetic energy and rest-mass energy m_0c^2 [15]. Quantum mechanics treats light not only as the particle or as the wave, consider both forms as a particle and as a wave and energy in the form of wave of light is $E = h\nu$, where h is the plank's constant and ν is the frequency of the wave [16]. But there is no single equation which holds or explain both energy forms of a photon. However, here we will derive the energy equation of photons, which depends on the scalar curvature of the surface of the matter and explain both energy forms.

2. Methods

In general relativity of theory, tensor is the main structural part and scalar curvature has taken the crucial role. The Ricci scalar curvature $\mathbf{R} = g^{\mu\nu}R_{\mu\nu}$, where $g^{\mu\nu}$ is the inverse metric tensor and $R_{\mu\nu}$ is the Ricci curvature tensor. The Ricci tensor is obtained by contacting the indices of Riemannian-Christoffel tensor $R^{\rho}_{\mu\nu\rho}$, here μ, ν, ρ indices represents four-dimensional space-time coordinates. We consider only the three-dimensional space part [17,18] in deriving the complex nature of energy and rest mass of the photon. So, we use $I^n = g^{\mu\nu}R_{\mu\nu}$ instead of $\mathbf{R}$, which physically represents every point of the surface of matter [8,9]. Consider light with frequency ν_{ph} and wavelength λ morphing into a particle form and fall onto the surface of matter with velocity c_p and rest-mass $m_0(\lambda)$. The energy of such particle is given by $E_p = m_0(\lambda)c^2$, here c is the velocity of light in a free space. Its energy corresponding to matter wave velocity c_w, wavelength λ_p and frequency ν_p can be expressed as $E = h\nu_p$. Using $c^2 = c_pc_w$ [8,9], the energy of photon can be written as $E_p = m_0(\lambda)c_pc_w$. The energy of the photon when it is in contact with the surface of the matter is different from these two energies. Here, we show how to calculate this new form of energy of photon under this special circumstance using rest mass equation of the photon. We can also retrieve the two fundamental energy equations from this new form of energy expression.

When a photon comes in contact from free space onto the surface of matter, the velocity of photon in free space and velocity on the surface becomes approximately equal i.e. $\upsilon_{ph} \approx c_p$. Hence the rest mass of the photon is a linear combination of these two masses and can be expressed as

$$\mathbf{m} = m_1 + m_2 \tag{1}$$

In which, m_1 is the wavelength dependent rest mass of the photon and m_2 is its curvature dependent rest mass [8,9].

$$\mathbf{m} = m(\lambda)\sqrt{1 - \frac{\upsilon_{ph}^2}{c^2}} + i\ m'(\lambda_p)\ I^n\ \sqrt{E_i^2 - 1} \tag{2}$$

$$= m_x + i\ m_y \tag{3}$$

So, considering $m_x = m(\lambda)\sqrt{1 - \frac{\upsilon_{ph}^2}{c^2}}$ and $m_y = m'(\lambda_p)\ I^n\ \sqrt{E_i^2 - 1}$ Eq. (3) represents the complex nature of the rest mass of photon. Hence, this complex form of mass of the photon is called as the illusive mass. In the absence of matter $I^n = 0$, and hence $\mathbf{m} = m_x$, which is the Einstein's rest mass equation.

3. Results and discussion

In order to obtain the energy of photon on the surface of matter in terms of its complex mass, consider the Einstein's mass-energy equation $\varepsilon = mc^2$.

When the photon touches the surface of matter, its mass can be considered in the complex form and the energy expression takes the form

$$\varepsilon = \frac{\mathbf{m}\ c^2}{\sqrt{1 - \frac{c_p^2}{c^2}}}$$

Now, substituting the value of $\mathbf{m}$

$$\varepsilon = \frac{c^2}{\sqrt{1 - \frac{c_p^2}{c^2}}}\left[m(\lambda)\sqrt{1 - \frac{\upsilon_{ph}^2}{c^2}} + i\ m'(\lambda_p)\ I^n\sqrt{E_i^2 - 1}\right]$$

$$= \frac{m(\lambda)\ c^2\sqrt{1 - \frac{\upsilon_{ph}^2}{c^2}}}{\sqrt{1 - \frac{c_p^2}{c^2}}} + i\ \frac{m'(\lambda_p)\ I^n c^2\sqrt{E_i^2 - 1}}{\sqrt{1 - \frac{c_p^2}{c^2}}}$$

$$= m(\lambda)\ c^2 + i\ m'(\lambda_p)\ I^n\ \sqrt{E_i^2 - 1}\ c^2\left(1 - \frac{c_p^2}{c^2}\right)^{-1/2} \quad [\text{as } \upsilon_{ph} \approx c_p]$$

$$= m(\lambda)\ c^2 + i\ m'(\lambda_p)\ I^n\ \sqrt{E_i^2 - 1}\ c^2\left(1 + \frac{1}{2}\frac{c_p^2}{c^2}\right) \quad [\text{using Binomial approximation}]$$

$$= m(\lambda)\ c^2 + i\ m'(\lambda_p)\ I^n \sqrt{E_i^2 - 1}\left(c^2 + \frac{c_p^2}{2}\right)$$

$$= m(\lambda)\ c^2 + i\ \left[m'(\lambda_p)\ I^n \sqrt{E_i^2 - 1}\ c_y^2\right] \quad \left[\text{where, } c_y^2 = c^2 + \frac{c_p^2}{2}\right]$$

$$\varepsilon = E_x + i\ E_y, \tag{4}$$

Here, the energy of the photon in its complex form, where $E_x = m(\lambda)\ c^2$ and $E_y = m'(\lambda_p)\ I^n \sqrt{E_i^2 - 1}\ c_y^2$ So, Eq. (4) represents the complex form of photon's energy, where depends upon the scalar curvature I^n of every point of the surface whenever the photon comes in contact with the matter. In the absence of matter, we have $I^n = 0$ and hence energy becomes $\varepsilon = m(\lambda)c^2$, again this is the usual Einstein's mass-energy relation. When photon transforms from particle to wave, then it will possess the energy $\varepsilon = h\nu$.

Quantum mechanics showed that the photon possesses dual nature, viz., particle and wave [19]. The photon can continuously transform from particle to wave and vice-versa, called as morphing [20]. Due to the dual nature of light, the photon has various forms of rest mass (mass posing), zero; nonzero real and imaginary value. Eventually, the photon's rest mass is a complex number and advent in nature itself with several forms of the rest mass [9] as massless, non-zero real mass and non-zero imaginary and each form depends upon some special circumstances. In free space, the rest mass of photons is zero when it behaves like a wave. When it behaves like a particle, the rest mass is a real number and within the dispersive medium, its value becomes imaginary. And on the surface of matter, it becomes complex quantity. The reason for this mass posing of photon is wave-particle duality of light [9].

Since mass; energy are the same entity and in this context of illusive mass, photon's energy is complex number in the particle aspect. The formulation of the photon's energy has embodied the geometry. That already has had corporeal in large scale (general relativity) as well as in small scale (string theory). This is the first time in which, this energy equation has shown the complex form by considering the wave-particle duality and Rieannian geometry. That explain both energy forms, particle and wave with certain conditions. On the surface of matter photon hides its energy due to surface geometry of matter, mathematically represents complex number. This unreveal energy form is intrinsic nature of photon, arises for the reason of wave-particle duality and illusive nature. From this unreveal energy form, photon could show the reveal forms of energy $m(\lambda)c^2$ when surface of matter will be absence i.e. without any contact of matter. And when photon transform into wave nature, energy would be $h\nu$.

4. Conclusions

The energy of a photon depends upon the scalar curvature of the surface of the matter when it comes in contact with the matter. Using the fundamental principles of scalar curvature of surface of matter, we show that the energy of the photon can have a complex behavior because of the complex form of its rest mass. In the absence of matter, the energy of the photon equation looks like the real number which is a known energy form of the particle. So, when a photon transforms from particle to wave in the absence of matter, then the energy of photon would be frequency dependable. We obtained the real form of energy of photon from its complex form. In reality, due to the complex form of energy, the photon might hides its energy itself. One needs more experimental evidence to prove this complex nature of the photon's energy. Our theoretical computations could pave the way towards unraveling the dark

References

[1] Philip Carter, Imaginary Physics, 2014.
[2] Richard G. Shoup, A complex logic for computation with simple interpretations for physics workshop on physics and computation, IEEE Comput. Soc. (1992).
[3] Roger Penrose, The Road to Reality: A Complete Guide to the Laws of the Universe, Random House, 2006.
[4] Michael Atiyah, Robbert Dijkgraaf, Nigel Hitchin, Geometry and physics, Phil. Trans. R. Soc. A 368 (1914) (2010) 913–926.
[5] Edward Witten, Physics and geometry. PRE-30537, 1987.
[6] Eugene Hecht, Did einstein really discover $E = mc^2$?, Amer. J. Phys. 56 (2) (2009) 799–806.
[7] W.L. Fadner, Einstein on mass and energy, Amer. J. Phys. 77 (9) (1988) 114–122.

[8] Mahendra Goray, Ramesh Naidu Annavarapu, A novel way of understanding the linear momentum of photon, Optik 223 (2020) 165488.
[9] Mahendra Goray, Ramesh Naidu Annavarapu, Rest mass of photon on the surface of matter, Results Phys. 16 (2020) 102866.
[10] Liang-Cheng Tu, Jun Luo, George T. Gillies, The mass of the photon, Rep. Progr. Phys. 68 (1) (2004) 77.
[11] G.V. Chibisov, Astrophysical upper limits on the photon rest mass, Sov. Phys. Uspekhi 19 (7) (1976) 624.
[12] D.D. Ryutov, Using plasma physics to weigh the photon, Plasma Phys. Control. Fusion 49 (12B) (2007) B429.
[13] R. Lakes, Experimental limits on the photon mass and cosmic magnetic vector potential, Phys. Rev. Lett. 80 (9) (1998) 1826.
[14] C. Tan, Imaginary rest mass of a photon in a dispersive medium, Optik 126 (24) (2015) 5304–5306.
[15] Louis de Broglie, Research on the theory of quanta, Ann. Phys. 10 (3) (1925).
[16] R. Ionicioiu, D.R. Terno, Proposal for a quantum delayed-choice experiment, Phys. Rev. Lett. 107 (23) (2011) 230406.
[17] Satya Pal Puri, General Theory of Relativity, Pearson Education India, 2013.
[18] Lee C. Loveridge, Physical and geometric interpretations of the Riemann tensor, Ricci tensor, and scalar curvature, 2004, arXiv preprint arXiv:gr-qc/0401099.
[19] Roman Kolesov, et al., Wave–particle duality of single surface plasmon polaritons, Nat. Phys. 5 (7) (2009) 470–474.
[20] A.S. Rab, E. Polino, Z.-X. Man, N.B. An, Y.-J. Xia, N. Spagnolo, R.L. Franco, F. Sciarrino, Entanglement of photons in their dual wave-particle nature, Nature Commun. 8 (1) (2017) 1–7.

Designing and manufacturing of interference notch filter with a single reflection band

Sangita Pal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sangitapal2@outlook.com*

Biraja Nayak, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, birajanayak21@gmail.com*

S. Sivasakthiselvan, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, s.sivasakthiselavan@live.com*

Abhishek Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, abhishekdas2256@gmail.com*

ARTICLE INFO

Keywords:
Notch filters
Reactive magnetron sputtering
Discrete layer structure
Thin films

ABSTRACT

The main idea of this work is to design a notch filter structure with a narrow notch width and maximum reflection while reducing fabrication challenges. In addition, using anti-reflection layers in the outermost part of the designed structure, the pass-band ripples are reduced. In this study, we considered $[(nHnL)^s\ (mHmL)^p\ (nHnL)^z]$ structure with $n=5$ and $m=3$. Using this form of design and combining 3 and 5 quarter-wave coefficients instead of 1 and 3, we could reach a narrower NW in fewer periods of HL layers. The stability of the deposition conditions and the density of the layers affect their quality and consequently the result of environmental tests. Hence, to construct the designed structure, we employed the sputtering method with RF and DC sources. In our experiments, we showed that the use of a simple shield prevents the oxidation of targets' surfaces as well as reduces the deposition rate and increases the stability of deposition processes. Fabricated Samples have been subjected to a variety of environmental tests, including humidity, hard and soft abrasion, temperature, and adhesion tests with satisfactory results.

1. Introduction

During the last decades, notch filters (NFs) have been used in different signal processing [1] applications, including speech and image processing, data transmission, seismology [2], biomedical engineering [3] and other applications like Raman and fluorescence spectroscopy, laser systems, laser protective coatings [4,5], protection of optical equipment against high beams [6] in order to control the wavelength of light and eliminate some frequencies [7–9]. NFs are optical instruments based on the interference phenomenon, which rejects frequencies inside a narrow band and allows other frequencies to pass without change [10–14].

Single-band and multi-band are of two main kinds of NFs that have been mostly used in industrial engineering. The main methods that have been used to produce such filters are rugate [15], discrete layer (HL) designs [4,16], and holographic technologies [5]. The rugate designs with a small difference between refractive index of layers, which have less ripple in the transmission regions [17], are well known in the literature. The latter design requires the deposition of layers by a refractive index gradient or so-called flip-flop structures, which is very complex and difficult to produce and economically not useful.

In the discrete layer (HL) method, materials with different refractive indices are used to manufacture NFs. H and L indicate the quarter-wave layer with the higher and lower refractive index, respectively. The optical thickness of these layers is a quarter of the central wavelength [4]. The challenging parts of this design are the deposition of the layers in a repeating order with high

accuracy, reducing transmission band ripples, making high layer quality. In addition, low level of particle defects, reasonable stress in the layers, and less manufacturing cost per sample are among important factors. In this method, the layers are placed on top of each other periodically that leads to ease in the design and manufacturing, but we should note that the transmission band ripples increase due to the sudden change of the refractive index between the layers with high and low refractive index. on the other hand, with fewer layers can be reached to a certain level of reflection in the center of the reflection band [10,18].

The deposition method and selecting the appropriate materials are very effective in the quality of the filter made. The reflectance bandwidth decreases by reducing the difference between the refractive index of the two materials in the structure. In this study, we use appropriate materials in the HL structure to provide a functional design with high reproducibility for the NFs at a center wavelength of 532 nm. Then, using anti-reflect layers, we reduce the transmission band ripples. We use the sputtering method with two RF and DC sources to manufacture the designed structure. Finally, we present the result of quality assessments conducted on the fabricated samples.

2. Design methodology

2.1. Initial design

The first step to create an optical instrument is to choose a suitable design method. An important issue that arises before selecting a design method is how to perform the final design practically. The method of quarter-wave stack structure, in the deposition stage, is more straightforward than other methods, and fewer layers are used in the structure, which make it more economical and less challenging. Furthermore, fewer layers also reduce the stress between layers and lead to better resistance to damage from environmental conditions.

In this method, the layers are placed alternately to form structures such as $(\mathrm{nHmL})^s$ or $(\mathrm{nLmH})^s$. In this notation, m and n determine the number of layers (odd integer) of each quarter-wave, and s expresses the number of pairs of layers used in the design. The reflection bandwidth or notch width (NW), as one of the characteristics of NF, depends on the difference between the refractive indices of the two materials with high and low refractive indices or the ratio between them. Therefore, the selection of materials should be such that the difference between the refractive indices of two materials optimize NW and the maximum reflection to desired values. The NW is given by [19]

$$\mathrm{NW} = \frac{4}{\pi}\arcsin\left[\frac{(\rho^2 + 2\rho\cos\gamma + 1)^{\frac{1}{2}}}{\rho + 1}\right], \tag{1}$$

where

$$\gamma = \left(\frac{s-1}{s}\right)\pi \quad \text{and} \quad \rho = \frac{n_{\mathrm{H}}}{n_{\mathrm{L}}}. \tag{2}$$

Eq. (1) then, in a simpler form, becomes [19]

$$\mathrm{NW} = 2\Delta g = \frac{4}{\pi}\arcsin\left(\frac{n_{\mathrm{H}} - n_{\mathrm{L}}}{n_{\mathrm{H}} + n_{\mathrm{L}}}\right). \tag{3}$$

The maximum reflection depends on the number of repetitions of the main period, s, in the structure [19,20] and can be calculated by [19]

$$R_{\max} = \frac{n_{\mathrm{sub}} - \rho^{2s} n_{\mathrm{H}}^2}{n_{\mathrm{sub}} + \rho^{2s} n_{\mathrm{H}}^2}, \tag{4}$$

where n_{sub} is the refractive index of the substrate. In this study, we used Y_2O_3 and Al_2O_3 as the materials with high refractive index (H) and SiO_2 as the material with low refractive index (L). We used BK7 with the refractive index of 1.52 as the substrate, and we considered air with that of 1 as the incidence medium in the design. The central wavelength of design is $\lambda = 532$ nm, which is located in the center of the reflection band. The physical thickness (PT) of the layers is such that the layers have optical thicknesses equal to odd integer coefficients of one-quarter of the central wavelength. Details of the designed structures with two pairs of materials, Y_2O_3-SiO_2 and Al_2O_3-SiO_2, are given in Table 1. Although using combination of 3 and 5 coefficients instead of single quarter-wave layers increases the thickness of structures, but significantly reduces the transmission band ripples and the number of total layers required in the design. The transmission diagrams of both structures are given in Fig. 1. From Fig. 1 and the information of Table 1, it can be concluded that reducing the difference between the refractive indices of the two materials reduces the NW. Meanwhile, we need more layers to achieve the maximum reflection.

2.2. Final design optimization

We use anti-reflect layers, which are made of the same materials used in the main structure, in the outermost layers of the design to reduce ripples more. Regarding this method, the strong scattering of the equivalent refractive index of the multilayer is compensated by the scattering of the anti-reflection structure and the transmission band ripples are improved [21]. In Section 2.1, the initial Notch filter designs were based on $(\mathrm{nHmL})^s$ stack. This structure can provide a narrow notch width and over 95% reflection in all of these designs, but they cannot suppress side ripples to an acceptable level. Therefore, we added antireflection pairs next to the outmost layer to increase the average pass-band transmittance.

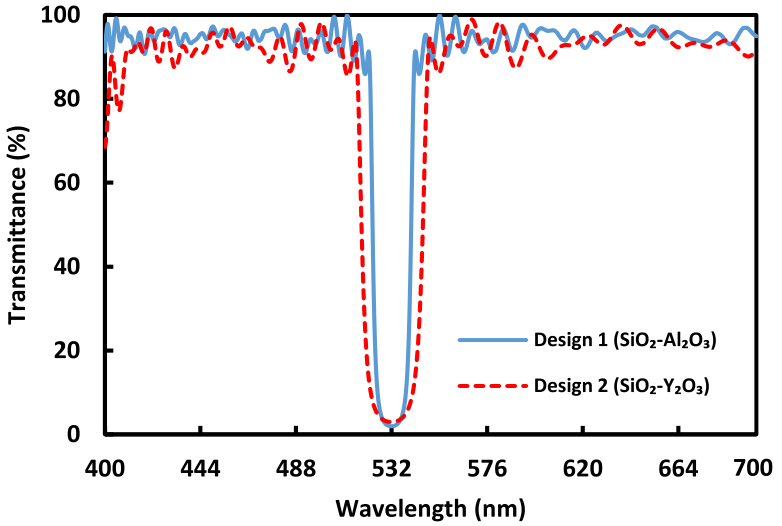


Fig. 1. Transmittance diagrams of designed structures with SiO_2-Y_2O_3 and SiO_2-Al_2O_3. The information about the structures is listed in Table 1.

Table 1
Details of designed structures with SiO_2-Y_2O_3 and SiO_2-Al_2O_3.

Design number	1	2
λ (nm)	532	532
Materials	SiO_2-Al_2O_3	SiO_2-Y_2O_3
Number of layers	44	26
R_{max} (%)	98.84	98.53
NW (nm)	17	26
Total PT (μm)	14	8

Table 2
Comparative performance of NF with various stack. The table presents the stack formula, maximum reflection R_{max}, average ripple reflectivity ARR, and the notch width NW.

Design number	Stack formula	R_{max} (%)	ARR (%)	NW (nm)
1	$[(5H5L)^4(3H3L)^{15}(5H5L)^3]$	98.05	6.11	17
1	$[(5H5L)^4(3H3L)^{15}(5H5L)^3]2HLHL$	98.84	3.02	17
2	$[(5H5L)^2(3H3L)^9(5H5L)^2]$	97.04	9.57	26
2	$[(5H5L)^2(3H3L)^9(5H5L)^2]2HL$	98.60	3.77	26

First, in Design 1, which was made by SiO_2–Al_2O_3 materials with the structure of $[(5H5L)^4\ (3H3L)^{15}\ (5H5L)^3]$ by adding two pairs of antireflection layers, the initial stack had been modified as $[(5H5L)^4\ (3H3L)^{15}\ (5H5L)^3]2HLHL$. Table 2 and Fig. 2 demonstrate that this design significantly improves the results concerning side ripples and increases peak reflections.

Second, for Design 2, in which SiO_2–Y_2O_3 materials were used, we added an antireflection pair to the end of the stack, and it was changed from $[(5H5L)^2(3H3L)^9(5H5L)^2]$ to $[(5H5L)^2\ (3H3L)^9\ (5H5L)^2]2HL$. As a result, the average ripple reflectivity decreased, and there is a minor increase in peak reflection. Fig. 3 and the second two lines of Table 2 can show us these results.

In the next section, we discuss the feasibility of manufacturing the structure containing Al_2O_3. This structure has a smaller NW in regards to the smaller difference between its refractive index and SiO_2. Moreover, it is more resistant and cheaper than Y_2O_3.

3. Experiment

3.1. Manufacturing method

The layers were deposited employing Reactive magnetron sputtering method, which is a suitable method for dielectric materials deposition, and it can produce dense layers with good resistance to environmental stresses. Increasing the deposition rate by increasing the applied power, reducing sputtering of the substrate and vacuum chamber, which prevents contamination, are the advantages of this method. Reducing the substrate's heating during deposition enables the sputtering process to be used for different substrates. Also, this method needs less gas pressure during the deposition and is economically affordable due to its material consumption compared to other methods.

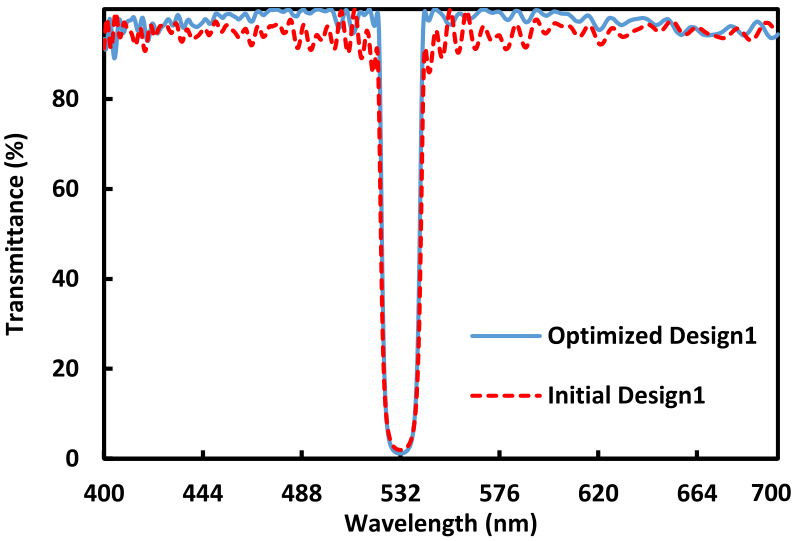


Fig. 2. Initial and optimized designs of SiO_2-Al_2O_3. The information about structures is listed in Table 2.

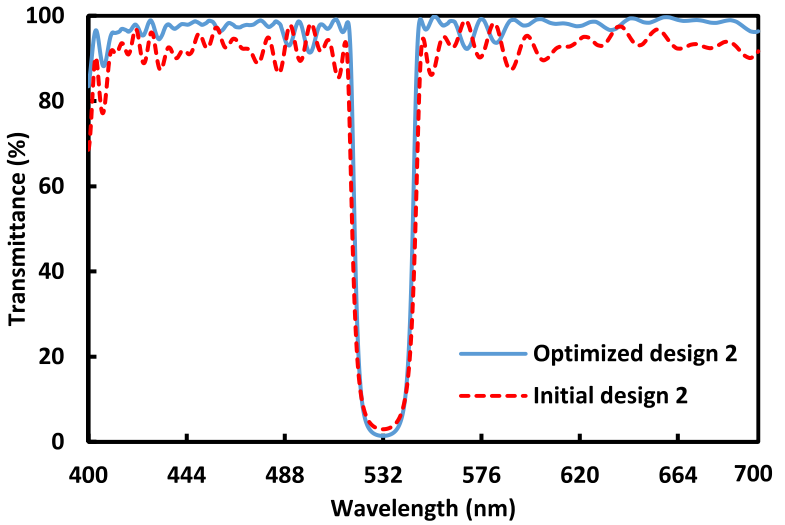


Fig. 3. Initial and optimized designs of SiO_2-Y_2O_3. The information about structures is listed in Table 2.

3.2. Deposition rate optimization

One of the most important challenges related to metal targets used in the reactive magnetron sputtering process is the oxidation of the metal targets' surfaces because the lower sputtering rate of the oxide than that of the metal has been observed. After oxygen enters the vacuum chamber, the target surface oxidizes, and consequently, the deposition rate suddenly decreases. Target surface oxidation also disturbs the stability of the deposition process. Due to the fact that achieving stoichiometric oxide thin films on the substrate surface requires a certain amount of oxygen, it is not possible to reduce the oxygen flow too much. So, oxidation of the target surface and reducing the deposition rate are inevitable. Accordingly, without reducing the oxygen flow, we must prevent the oxidation of the target surface. This can be done by using a simple shield with holes in the vacuum chamber, which is situated above the sputtering sources, and changing the position of inlet oxygen and argon gases. To prevent oxidation of the targets' surfaces, we place the gas inlet system for oxygen near the substrate and inlet argon gas near the DC and RF sources. Fig. 4 shows a schematic representation of the reactive magnetron sputtering vacuum chamber with a DC and an RF source. In this setup, we used a shield containing two holes above the sources to prevent excessive oxygen concentration around the sputtering sources and to achieve a stable deposition process. The location of the shield and the diameter of its holes are two important parameters that can affect the deposition rate and stability of the process. In order to achieve the highest shield performance, these two parameters must be optimized for a specific coating device.

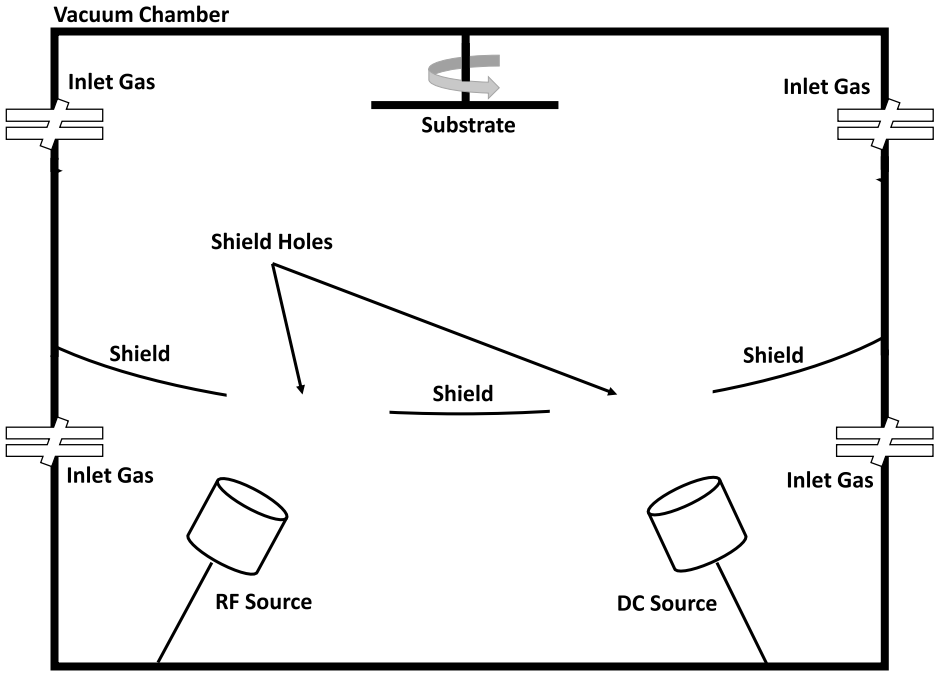


Fig. 4. Schematic representation of the reactive magnetron sputtering vacuum chamber with two DC and RF sources.

Table 3
Comparison of deposition rates of SiO_2 and Al_2O_3 with and without shield.

Target material	Si	Al
Oxide	SiO_2	Al_2O_3
Argon flow (sccm)	25	25
Oxygen flow (sccm)	11	3
Power (w)	480	240
Initial rate (Å/s)	0.9	0.4
Optimized rate (Å/s)	6	1.5

Table 3 shows the applied power and deposition rate values before and after the presence of the shield.

The oxygen and argon flows, sources' powers, and all the deposition conditions, which are mentioned in detail in the next section, are fixed in both stages. The only difference is the presence and the absence of the shield. Eventually, we can prevent the deposition rate reduction by this method and provide a stable deposition process for both alumina and silica thin films.

3.3. SiO_2 and Al_2O_3 single layers

In the design software, we design a single layer of SiO_2 and a single layer of Al_2O_3 at the center wavelength of 532 nm. In this step, we use SF6 and BK7 as substrates for SiO_2 and Al_2O_3 single layer, respectively. The deposition was done using DC source for SiO_2 and RF source for Al_2O_3 with silicon and aluminum targets with 99.99% purity and after complete cleaning of the substrate. Optimal conditions for deposition of SiO_2 and Al_2O_3 were obtained in 11 sccm and 3 sccm oxygen flow, respectively. Deposition rates of SiO_2 and Al_2O_3 are 6 Å/s and 1.5 Å/s respectively. The deposition was performed for both single layers at 100 °C, chamber pressure of 3 mTorr during deposition, the argon gas flow of 25 sccm, and the start pressure of 8×10^{-6} Torr.

Fig. 5 shows the fabricated and designed spectra of Al_2O_3 and SiO_2 single layers. There are good agreements between the designed and manufactured curves of both single layers. It can prove that we obtained correct stoichiometry for deposited materials. But we might need more optimization when we want to coat more layers. In the following steps, we will study different stack optimizations on the movement towards NF fabrication.

3.4. $(3H3L)^2$ stack

Optimal conditions for deposition of stacks are the same as ones used for single layers. However, before deposition the final design structure and in order to achieve better results, it is necessary to deposit a simpler structure with fewer layers. Fig. 6 shows the transmittance spectra of the deposited and the designed structures are in good agreement. It indicates that the obtained deposition

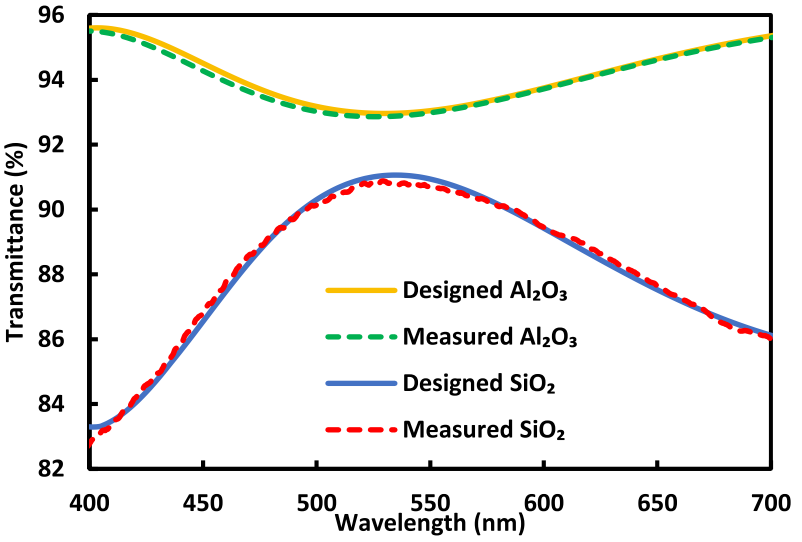


Fig. 5. The spectral transmission of the SiO_2 and Al_2O_3 single layers.

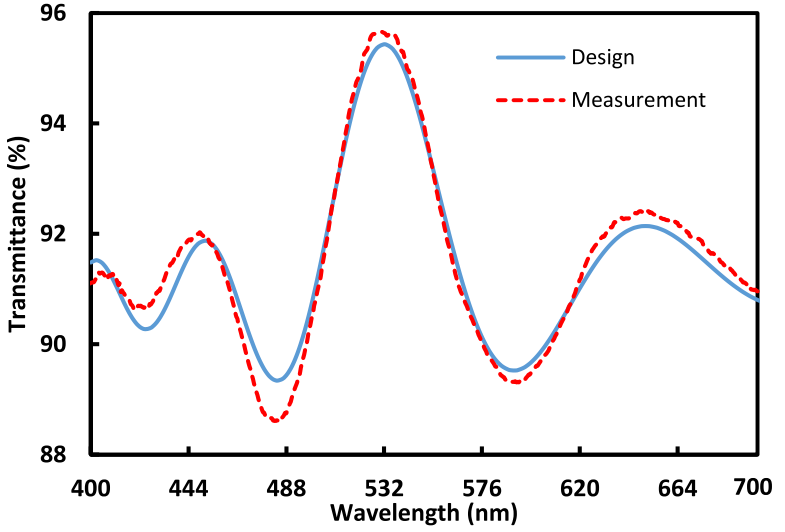


Fig. 6. Measured and designed transmittance spectra for $(3H3L)^2$ stack.

conditions in the previous sections for single layers are approximately the optimal ones. Section 3.4 presents the results of the optimization for samples located on the rotating segment, and then the necessary corrections to Packing Density and to Tooling Factor have been made.

However, there are some minor mismatches between designed and fabricated curves. The reason for this is that two important parameters should be considered in the manufacturing step, namely the Packing Density of deposited layers and the Tooling Factor. The density of thin layers is less than the density of bulk form of specific material, and the ratio between them is known as Packing Density (P). Since the P influences the refractive index and thickness of thin layers, initial tests should be conducted by comparing the spectra of designed and manufactured stacks to determine the practical P. Tooling factor should also be taken into account during the deposition process. On the substrate, the thicknesses appear to differ from the thicknesses the device shows us during the deposition process. The mismatches can be eliminated by crossing the practical Tooling Factor with the designed layer thicknesses. These two parameters can affect the obtained manufactured spectrum, so after deposition of each stack by comparing the designed and manufactured transmittance curves, we obtained new optimal Tooling Factor and Packing Density to modify the mismatches. In the next step, we will discuss other stacks.

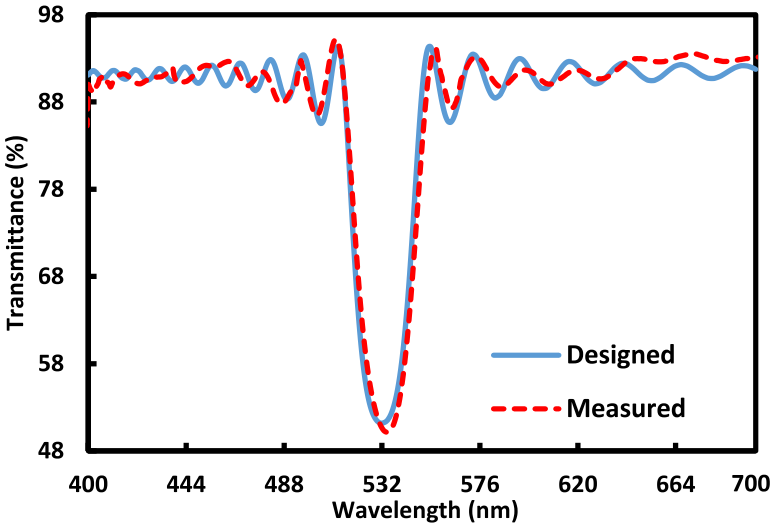


Fig. 7. Measured and designed transmittance spectra for $(3H3L)^{10}$ stack.

3.5. $(3H3L)^{10}$ stack

To achieve better optimization in this step, we designed and manufactured $(3H3L)^{10}$ stacks with more layers than the previous stacks. This stack has eight more periods than $(3H3L)^{2}$, while maintaining the same 3H3L structure. Fig. 7 shows the transmittance spectra of the designed and fabricated stack in which there is a good adaptation between the general form of the two curves. Comparing the manufactured curve to the designed curve, it is evident that it is approximately 2 nm ahead of the designed curve. It is due to mismatches between the deposited and designed thickness, which can be controlled by Tooling Factor. Based on a comparison of two curves, we modified our Tooling Factor to adapt to the designed and fabricated curves of future stacks.

3.6. $[(5H5L)^{4}(3H3L)^{14}]$ stack

For the purpose of deposition of our NF, we considered a stack in which both 5H5L and 3H3L periods are present. The current stack is approximately twice as thick as the previous stack. Fig. 8 shows precise agreement between designed and manufactured curves, especially in reflection band position, as we expected. As mentioned in the last sections, Tooling Factor and Packing Density have a vital role in the manufacturing process. On the other hand, optimizing deposition parameters such as inlet gas flows and the deposition rate is significant. According to Fig. 8, we have gained the correct deposition parameters and can begin fabricating the designed NF using SiO_2–Al_2O_3.

3.7. NF deposition

According to design 1, we use silica and alumina materials as H and L refractive indices, respectively. Regarding the optimized structure, see Fig. 5, the final design at the central wavelength of 532 nm is given by the design formula number 1 (see Table 1).

In Section 3.4, by performing several experiments, design and fabrication of single layers of materials in the structure, optimizing deposition conditions, and finally designing and deposition of 4-layer stacks were done.

The conditions for deposition are similar to those described in Section 3.3. It is essential to control the deposition rate as well as the thickness of the layers. 15.8 h were required for NF deposition with a 14 μm micrometer thickness.

In Section 3.4, the optimization was performed for samples, which are located on the rotating segment, and the necessary corrections were made. Therefore, we expect the samples to show good compliance with the designed structure. In the transmittance spectra, second side of the sample has a reflection, so first in the design software, we applied the effect of the reflection of the second side of the samples, and then we compare design and measured transmittance spectra. Fig. 9 shows the measured and designed spectra of the central sample.

As shown in Fig. 9, the final deposited structure and the final designed structure at the location of the central wavelength and the reflectance bandwidth are well matched. By extracting the data obtained from the transmission spectrum, the center of the reflection band is located at 530 nm, which is a perfect match with the design. We can see the ripples only in some areas of the transmission band. In the transmission band, there is a mismatch at some point between the transmittance spectrum of the deposited structure and the designed structure. Generally, the mismatches between the transmittance spectra of designed and deposited structures in the coating technology are due to the mismatches between the thickness of the same layers, the refractive indices, and the packing density in the deposited and the designed structures. Absorption in some coating areas, contamination, and error of spectrometer

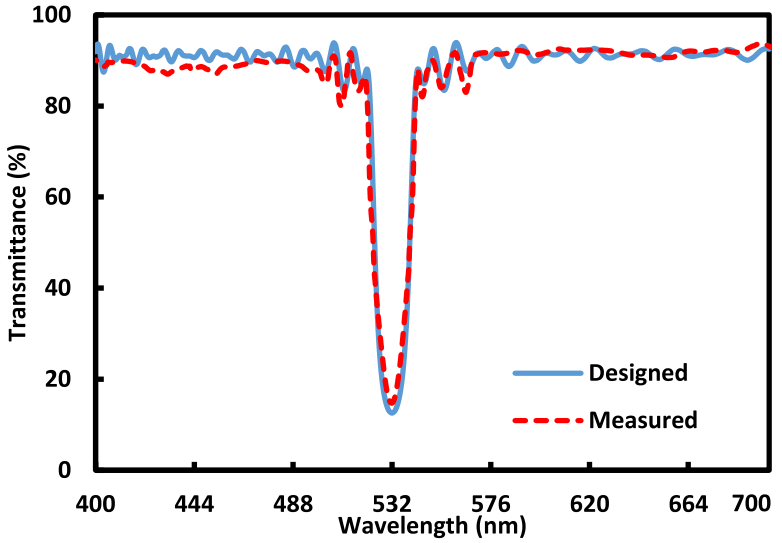


Fig. 8. Measured and designed transmittance spectra for $[(5H5L)^4(3H3L)^{14}]$ stack.

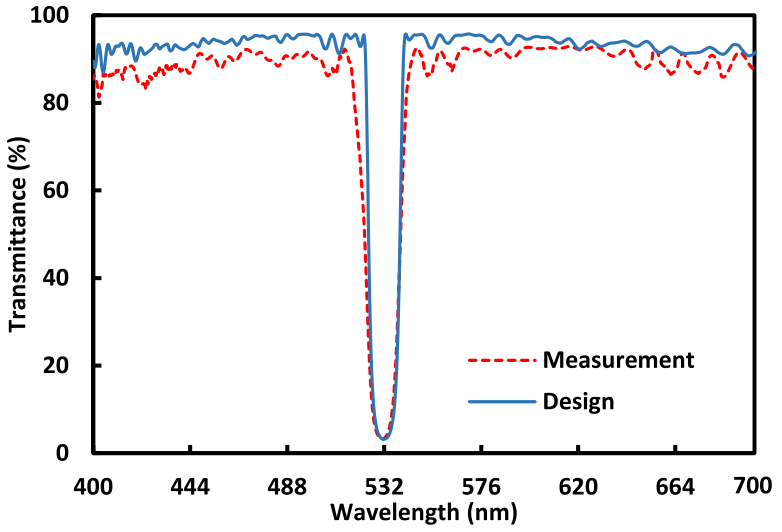


Fig. 9. Measured and designed transmittance spectra for NF (SiO_2-Al_2O_3).

and operator are other factors. In Fig. 9, although the spectra of the designed and the deposited structures are very well matched in many wavelengths, the control and stability of the deposition process must be very high to achieve spectra closer to the designed structure. Some of these mismatches are inevitable because the final designed structure of NF consists of 52 layers with different thicknesses. In the manufactured NF, the NW was 17 nm, the maximum reflectance at central wavelength was more than 95%, and the average transmittance was approximately 90%.

3.8. Environmental stability tests

Based on the environment and conditions in which each type of optical filter is placed, the type and characteristics of the stability tests that are required will vary. As part of the current study, manufactured samples are exposed to different environmental tests, including humidity, hard and soft abrasions, temperature, and adhesion tests.

3.8.1. Adhesion test

A layer's adhesion to the substrate and to each other is necessary. Cleaning of the substrate surface, the application of glow discharge, proper deposition conditions, the compatibility of the materials of the deposited layers, as well as the inherent stresses of the layers, are the factors that affect adhesion. This test uses standard adhesive tape. In order to carry out the test, adhesive tape

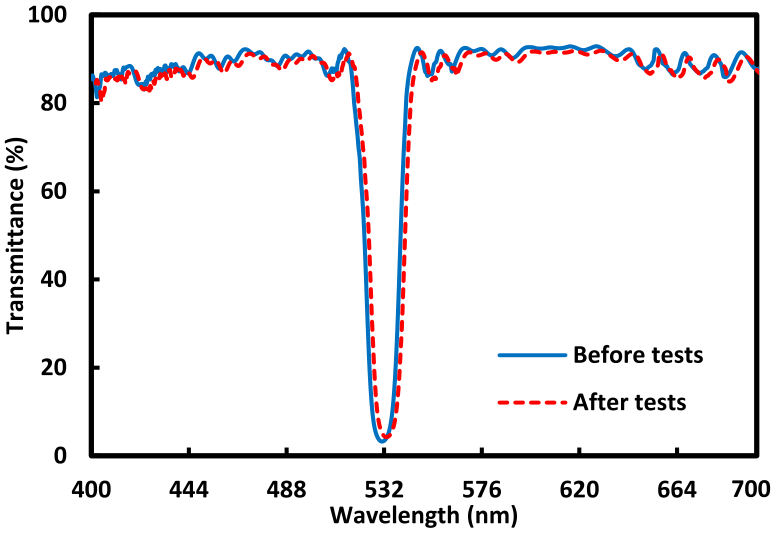


Fig. 10. The transmission spectral of a fabricated NF before and after environmental tests.

is pasted to the NF's surface and, after removing bubbles, it is peeled off vertically and suddenly. The inspection is done with the naked eye in standard lighting conditions. The result of this test on our NF was successful, and the sample surface before and after all environmental stability tests did not change, and no damage was seen after cleaning the surface.

3.8.2. Humidity test

Following the adhesion test, the NF was subjected to a humidity test. The sample was placed in a test chamber for a period of at least 24 h and exposed to a temperature of 49 ± 1 °C and a humidity of 95% to 100%. The sample was removed from the container after this period of time and then cleaned and dried according to the cleaning procedure for optical elements. Then its physical quality was examined. The sample surface did not exhibit any signs of damage, cracks, blebs, or spots. As the second comparison we made, we compared the transmittance spectra of the exposed samples before and after all of the environmental tests. The comparison is shown in Fig. 10. According to Fig. 10, there is no noticeable difference between the spectra before and after the tests. It should be noted that raising the substrate temperature results in better adhesion between the first layer and the substrate surface.

3.8.3. Temperature test

Interference filters are typically used at room temperature. Due to this, temperature changes may affect the physical structure of the stack, the shape of the spectrum, and the location of the extremum. For this reason, the sample was kept at each temperature of 62 ± 1 and 70 ± 1 °C for 2 h in this step. The rate of temperature change should not exceed two degrees Celsius per minute. After each step, the sample was placed at a temperature between 16 and 32 degrees Celsius and evaluated for cracks, peeling, etc. using an adhesion test. No signs of scaling, cracking, blurring, or blistering was observed after the NF surface had been tested. The transmittance spectra before and after this test are shown in Fig. 10. Our product passed this test well.

3.8.4. Abrasion test

NFs may need to be cleaned frequently before being installed on any device. Sometimes cleaning is not done in good condition. For example, cotton or fabric may be pulled over the surface while dust particles are still on the surface. These particles sometimes act as abrasives, and if the NF's surface is not tough enough, surface defects may appear on it. Selection of suitable material for layers and good deposition conditions, especially deposition temperature, are influential factors in achieving abrasion resistance. Our NF was made with metal oxide layers, which are among the hardest and durable materials. Its deposition conditions have provided a reasonable basis for achieving excellent abrasion resistance. At first, a soft abrasion test was done with a gross linen fabric. The thickness of the fabric is 6.4 mm, and its width is 9.5 mm. This fabric is fixed on the tester, and its surface is completely covered. The tester moved the width of the sample from one point to another in one direction for 25 full turns (50 rounds) with a force of at least 1 pound applied continuously. The path length should be at least 2 times bigger than the width of the fabric, and the force should always be applied vertically. Second, for the hard abrasion test, The tester was being moved for a complete 20 rounds with a force of 2 to 2.5 pounds. The length of the path should be approximately 3 times the diameter of the tester surface. After soft and hard abrasion tests, no traces of removing and scratching were observed on the NF's surface, so manufactured NF could pass all environmental tests successfully.

4. Conclusions

There are three main methods to design and manufacture NFs. In this study, the discrete layer (HL) method has been used in which there are fewer challenges than both rugate and holographic methods in the manufacturing step. In this method, the differences between the refractive indices of the materials used in structures are the determining factor in the NW and the maximum reflection. Furthermore, the $(mHnL)^s$ structure is usually used to design the HL layers, in which the same HL layers periods are repeated s times all over the structure where n and m are generally equal to 1 or 3. Still, in this study, we considered $[(nHnL)^s (mHmL)^p (nHnL)^z]$ structure with $n=5$ and $m=3$. Using this design form and combining 3 and 5 quarter-wave coefficients instead of 1 and 3, we could reach a narrower NW in fewer periods of HL layers. Consequently, less total thickness and our structure are beneficial because these combined coefficients significantly reduce the transmittance band ripples compared to the previous one. On the other hand, In this paper, by antireflection layers and a half-wave layer used as the interface layer in the outermost layers of the final structure, the transmission band ripples were appropriately reduced while we kept all the H and L coefficients as integers, so it is beneficial in the deposition process.

We used reactive magnetron sputtering to achieve high-quality layers and reduce material wastage with aluminum and silicon targets in the manufacturing step. Moreover, the materials used in the manufactured NF are very tough and inexpensive. We have proved that a simple shield above the metal targets has a substantial role in preventing the oxidation of targets' surfaces and reducing deposition rate and the stability of the deposition process.

Finally, fabricated samples were subjected to different environmental tests such as humidity, hard and soft abrasion, temperature, and adhesion tests which yielded positive results. It should be noted that the stability of the coating rate and precise control of the inlet gases to the chamber during the coating process has a significant impact on the fabricated thin films' quality. In addition, increasing the substrate temperature and perfect pre-deposition cleaning process causes better adhesion of the first layer to the surface of the substrate.

References

[1] J. Minguillon, M.A. Lopez-Gordo, D.A. Renedo-Criado, M.J. Sanchez-Carrion, F. Pelayo, Blue lighting accelerates post-stress relaxation: Results of a preliminary study, PLoS One 12 (10) (2017) e0186399.
[2] C.-C. Lee, Optical interference coatings for optics and photonics, Appl. Opt. 52 (1) (2013) 73–81.
[3] Z. Juan, G. Sen, L. Xue, Design of optical notch filters based on equivalent relations, in: Advanced Materials Research, vol. 679, Trans Tech Publ, 2013, pp. 47–52.
[4] V. Janicki, M. Lappschies, B. Görtz, D. Ristau, U. Schallenberg, O. Stenzel, N. Kaiser, Comparison of gradient index and classical designs of a narrow band notch filter, in: Advances in Optical Thin Films II, vol. 5963, International Society for Optics and Photonics, 2005, p. 59631O.
[5] R.W. Sprague, B. Shnapir, G.L. Minott, Rugate notch filters find use in laser-based applications, Laser Focus World 40 (9) (2004) 107–111.
[6] P.G. Verly, Design of complex rugate filters, in: Optical Interference Coatings, Optical Society of America, 2007, p. WA5.
[7] S.V. Nikolić, G.Z. Stančić, S. Cvetković, Design of nearly linear-phase double notch digital filters with close notch frequencies, IET Signal Process. 12 (9) (2018) 1107–1114.
[8] A. Nehorai, A minimal parameter adaptive notch filter with constrained poles and zeros, IEEE Trans. Acoust. Speech Signal Process. 33 (4) (1985) 983–996.
[9] M.-Y. Jeong, J.Y. Mang, Continuously tunable optical notch filter and band-pass filter systems that cover the visible to near-infrared spectral ranges, Appl. Opt. 57 (8) (2018) 1962–1966.
[10] M. Scherer, U. Schallenberg, H. Hagedorn, W. Lehnert, B. Romanov, A. Zoeller, High performance notch filter coatings produced with PIAD and magnetron sputtering, in: Advances in Optical Thin Films III, vol. 7101, International Society for Optics and Photonics, 2008, p. 71010I.
[11] J. Zhang, M. Fang, Y. Jin, H. He, Narrow line-width filters based on rugate structure and antireflection coating, Thin Solid Films 520 (16) (2012) 5447–5450.
[12] A. Thelen, Design of optical minus filters, J. Opt. Soc. Am. 61 (3) (1971) 365–369.
[13] A.J. Thelen, Design of optical interference coatings 1992, in: Lens and Optical Systems Design, vol. 1780, International Society for Optics and Photonics, 1993, p. 17802G.
[14] H. Khazraj, F.F. da Silva, C.L. Bak, S. Golestan, Analysis and design of notch filter-based PLLs for grid-connected applications, Electr. Power Syst. Res. 147 (2017) 62–69.
[15] O. Lyngnes, J. Kraus, Design of optical notch filters using apodized thickness modulation, Appl. Opt. 53 (4) (2014) A21–A26.
[16] T.D. Rahmlow Jr., J.E. Lazo-Wasem, E.J. Gratrix, Narrow band infrared filters with broad field of view, in: Infrared Technology and Applications XXXII, vol. 6206, International Society for Optics and Photonics, 2006, p. 62062S.
[17] W.H. Southwell, Using apodization functions to reduce sidelobes in rugate filters, Appl. Opt. 28 (23) (1989) 5091–5094.
[18] U. Schallenberg, B. Ploss, M. Lappschies, S. Jakobs, Design and manufacturing of high-performance notch filters, in: Modern Technologies in Space-and Ground-Based Telescopes and Instrumentation, vol. 7739, International Society for Optics and Photonics, 2010, p. 77391X.
[19] H.A. Macleod, Thin-Film Optical Filters, CRC Press, 2017.
[20] D.W. Diehl, N. George, Holographic interference filters for infrared communications, Appl. Opt. 42 (7) (2003) 1203–1210.
[21] A. Shankar, S. Kumar, Using a pair of antireflection layer in a minus filter to suppress ripples in the HUD-beam combiner, J. Opt. 43 (3) (2014) 257–259.

Photoinduced charge transfer in two-photon absorption

Rudra Prasad Nanda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, rudraprasad858@gmail.com*

Manoranjan Sahoo, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, manoranjansahoo14@gmail.com*

Swaha Pattnaik, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, swaha.pattanayak@gmail.com*

Sambhunath Biswas, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, s.biswas09@outlook.com*

ARTICLE INFO

Keywords:
External electric field
Charge differential density
Sequential charge transfer
Super-exchange charge transfer
Two-photon absorption

ABSTRACT

The visual method is used to theoretically discuss the influence of different external electric fields on the photo-induced charge transfer of the donor-bridge-acceptor (D-B-A) system in the two-photon absorption (TPA). We dynamically observe the transition process in different spatial dimensions. The final results show that the external electric field can manipulate the orientation and type of intermolecular charge transfer in the TPA process of the D-B-A system. The changes in the proportions of sequential charge transfer and super-exchange charge transfer in TPA provide a strong proof for the conclusion. This non-monotonic change has theoretically deepened our understanding of the charge transfer characteristics in TPA, and we have a further understanding of the optical properties of the D-B-A system.

1. Introduction

With the advent of lasers in the 1960s, the research of two-photon absorption (TPA) began to develop. It has great potential in three-dimensional (3D) information storage, medical diagnosis, and fluorescence microscopy imaging (Goppert-Mayer, 1931; Denk et al., 1990; Dorfman et al., 2016; Kaiser and Garrett, 1961; Honig and Jortner, 1967). Essentially, TPA is a nonlinear optical effect (Albota et al., 1998; Mongin et al., 2003; Kato et al., 2004). The two photons generated by light of different frequencies interact with atoms or molecules, so that energy is converted between light radiation and matter. The photons absorbed in this process promote the transition of matter from the ground state to the excited state through the intermediate virtual state. Here, the transition probability in the transition process can be expressed as (Guo et al., 2003; Sun et al., 2008)

$$\delta_{tp} = 8 \sum_{\substack{j\neq g \\ j\neq f}} \frac{|\langle f|\mu|j\rangle|^2 |\langle j|\mu|g\rangle|^2}{\left(\omega_i - \omega_f/2\right)^2 + \Gamma_f^2}\left(1 + 2\cos^2\theta_j\right) + 8\frac{\left|\Delta\mu_{fg}\right|^2 |\langle f|\mu|g\rangle|^2}{\left(\omega_f/2\right)^2 + \Gamma_f^2}\left(1 + 2\cos^2\varphi\right) \quad (1)$$

where $|g\rangle$ and $|f\rangle$ represent the ground state and the final state, respectively. $|j\rangle$ can represent the intermediate virtual state in the TPA process, which can be any excited state. μ is the electric dipole moment operator, so $\Delta\mu_{fg}$ represents the dipole moment difference between states ($\Delta\mu_{fg} = \langle f|\mu|f\rangle - \langle g|\mu|g\rangle$). The former part of the formula containing $|j\rangle$ is called "three states", and the corresponding latter part is called "two states". Among them, the "three-state" part contains the denominator of frequency and life, and its value corresponds to the difference between the energy of the intermediate state and the final state. The smaller the denominator, the closer the energy is, and the probability of two-photon absorption at this time increases. The resonance enhancement phenomenon can be considered in the energy level angle. In the "two-state" part, the dipole moment difference in the numerator often has a positive correlation effect on the probability of two-photon absorption. This shows that the difference in symmetry of the wave function cannot be ignored. The process of visualizing the "three states" mainly observes the electron-hole coherence and charge transfer.

In this paper, the D-B-A system is one of the important structural types of third-order nonlinear optical materials (Paulson et al., 2005; Giacalone et al., 2004; Davis et al., 2001; Mirebeau et al., 2000). Here, D represents an electron donor, A represents an electron acceptor, and B represents a conjugate bridge composed of large π electrons. In 2002, Junya et al. experimentally reported the intramolecular charge transfer behavior of the porphyrin-oligothiophene-fullerene system (Junya and Kazuo, 2002). In 2005, Sun et al. made theoretical calculations for this ternary compound system based on quantum chemistry methods, and did research on the transition energy, oscillator strength, and electron-hole coherence in the frontier molecular orbital (HOMO, LUMO) of the polymer system (Sun et al., 2005). Later, there were also related studies on the recombination energy, electronic coupling, and free energy of electron transfer reactions of the polymer D-B-A system. In this paper, we are exploring the charge transfer characteristics of the D-B-A system in two-photon absorption under the control of an

external electric field. Porphyrin-oligothiophene-fullerene has unique advantages in application. In organic photovoltaic materials, people's attention to the efficiency of converting light energy is mainly manifested in the charge transfer efficiency and energy transfer inside the device. With the development of organic solar cells, porphyrins and fullerenes have gradually become ideal donor and acceptor materials due to their electrical properties and unique chemical structures. Thiophene acts as an intermediate "bridging group" connecting the acceptor and the donor. Because it is a five-membered ring structure containing sulfur atoms, it has stable chemical properties, low oxidation sites, strong structural tunability, and low cost. The length of the bridge base (the distance between the donor and the acceptor) in the D-B-A structure composed of the three will affect the transfer of energy and electrons. Therefore, we want to study the influence of the external electric field on the charge transfer of the system with-out changing the length of the intermediate conjugate bridge in the two-photon absorption of the typical nonlinear optical effect. We the-oretically studied the different charge transfer characteristics of the current system's TPA process, especially the difference in external elec-tric fields of different orientations. In addition, for the two-step transi-tion process represented by the first term in the formula, the classical quadratic response theory cannot accurately analyze the overall charge transfer. Using our developed TPA calculation program based on the sum-of-states (SOS) model (Mu et al., 2019), we can finally achieve the purpose of visually analyzing the TPA transition character-istics (Kang et al., 2020). In particular, reciprocal transfer may occur. Here, the visualization method mainly refers to the "three-state" model based on the TPA process. We use the calculated matrix and dif-ferential density to draw graphics in two-dimensional and three-dimensional real spaces. The charge transfer that occurs in the transi-tion process is clearly located on a local single atom, which promotes better revealing of different charge transfer characteristics.

2. Results and discussion

Fig. 1 shows the composition of the D-B-A system (Otsubo et al., 2002; Ikemoto et al., 2002). Fullerenes in hyperconjugated systems have strong electron delocalization capabilities. The long chain structure composed of four thiophene units in the middle acts as a "bridge" for electron transfer between the two hyperconjugated systems (Song et al., 2013, 2015; Mu et al., 2020). The X axis in Cartesian coordinates in the figure is parallel to the direction of the "bridge". All external electric fields also act on the X-axis orientation.

The absorption spectra of porphyrin-tetra-thiophene-fullerene in different external electric fields are shown in Fig. 2. First, Fig. 2(a) and 2(b) are the OPA spectra of two opposite electric fields on the X axis. By observing the spectrum contrast chart in the negative direction, it can be seen that with the increase of the electric field intensity, the strong absorption peak in the visible light region (410 nm–480 nm) produces an obvious red shift and a gradual weakening of the intensity. On the contrary, the absorption peak in the same region of the spectrum in the forward direction did not show a significant decrease. Only when the intensity of electric field is $F = 1.5 \times 10^{-3} au$, the spectrum shows a distinguishable red shift. Secondly, Fig. 2(c) and 2(d) are the TPA spectra of two opposite electric fields on the X axis. Compared with the spectrum without electric field, new peaks will be generated in different regions. In Fig. 2(d), there are super strong absorption peaks in the two largest external electric fields of $F = 1.2 \times 10^{-3} au$ and $F = 1.5 \times 10^{-3} au$. The difference in the spectra caused by the external electric field is considered to be due to the deviation of the front molecular orbital distribution in the molecule. The decrease in the number of transition electrons leads to a decrease in the required excitation energy, and finally the increase in wavelength results in a peak red shift. Depending on the regulation of the external electric field, the optical properties of electronic transitions need to be further revealed by visualization methods.

Compared with simple orbital analysis, charge differential density (CDD) can fully consider the orbital transition and configuration coefficients. It is a commonly used method to analyze the change in charge distribution caused by the difference in electron density during the excitation process. However, CDD also has certain limitations in the study of charge transfer characteristics. In previous studies, the charge transfer characteristics within the molecule are not single. When there are both electrons and holes in the same area in the CDD diagram, it is necessary to specifically analyze the electron-hole source from local excitation or charge transfer. At this point, you can refer to the density matrix to locate a single atom to assist in the analysis.

Below we list and analyze the TPA cross section and transfer characteristics under several external electric fields. Fig. 3 shows the TPA and OPA process of $F = -9 \times 10^{-4} au$. In particular, there is a strong absorption peak mainly contributed by the "three-state" term at 780 nm in Fig. 3(a). (The cyan dotted line in the figure dominates the peak). After calculation, we found that the strong absorption states S_{12}, S_{13}, and S_{15} in Fig. 3(b) may all be intermediate transition states in the TPA process. Among them, S_{20} can use S_{13} as a intermadiate transition channel ($S_0 \rightarrow S_{13}\text{-}>S_{20}$). S_{19} has four transition channels ($S_0 \rightarrow S_4\text{-}>S_{19}$, $S_0 \rightarrow S_{12}\text{-}>S_{19}$, $S_0 \rightarrow S_{13}\text{-}>S_{19}$, $S_0 \rightarrow S_{15}\text{-}>S_{19}$.). Therefore, this article focuses on the analysis of the transition process of S_{19}. Fig. 3(c) shows the two transition channels in S_{19}. The first channel is S_{12} as the intermediate state. At this time, the first step of excitation is the strong local excitation of the fullerene structure, accompanied by charge transfer between the thiophene chain and the fullerene. The second step is completely different. The main thing is that fullerenes provide electrons for super-exchange charge transfer to the porphyrin structure. The second channel selects S_{15} as the intermediate state. Unlike the previous channel, in the first transition, there is a charge transfer behavior of the thiophene chain from the middle to the donor and acceptor on both sides. At this time, weak local excitation occurs on the fullerene. The transition characteristics of the second step are the same as those of the previous channel. The bright area appears on the off-diagonal line in the transition density matrix (TDM) diagram, which is the characteristic of super exchange charge transfer. Therefore, the "bridge" of the thiophene chain is more like providing electrons for the second transition and promoting the occurrence of super-exchange transfer at the macro level.

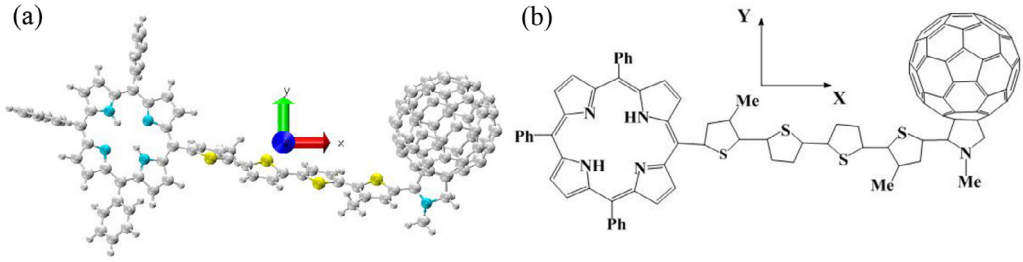


Fig. 1. The molecular structure of porphyrin-tetra-thiophene-fullerene (PTF).

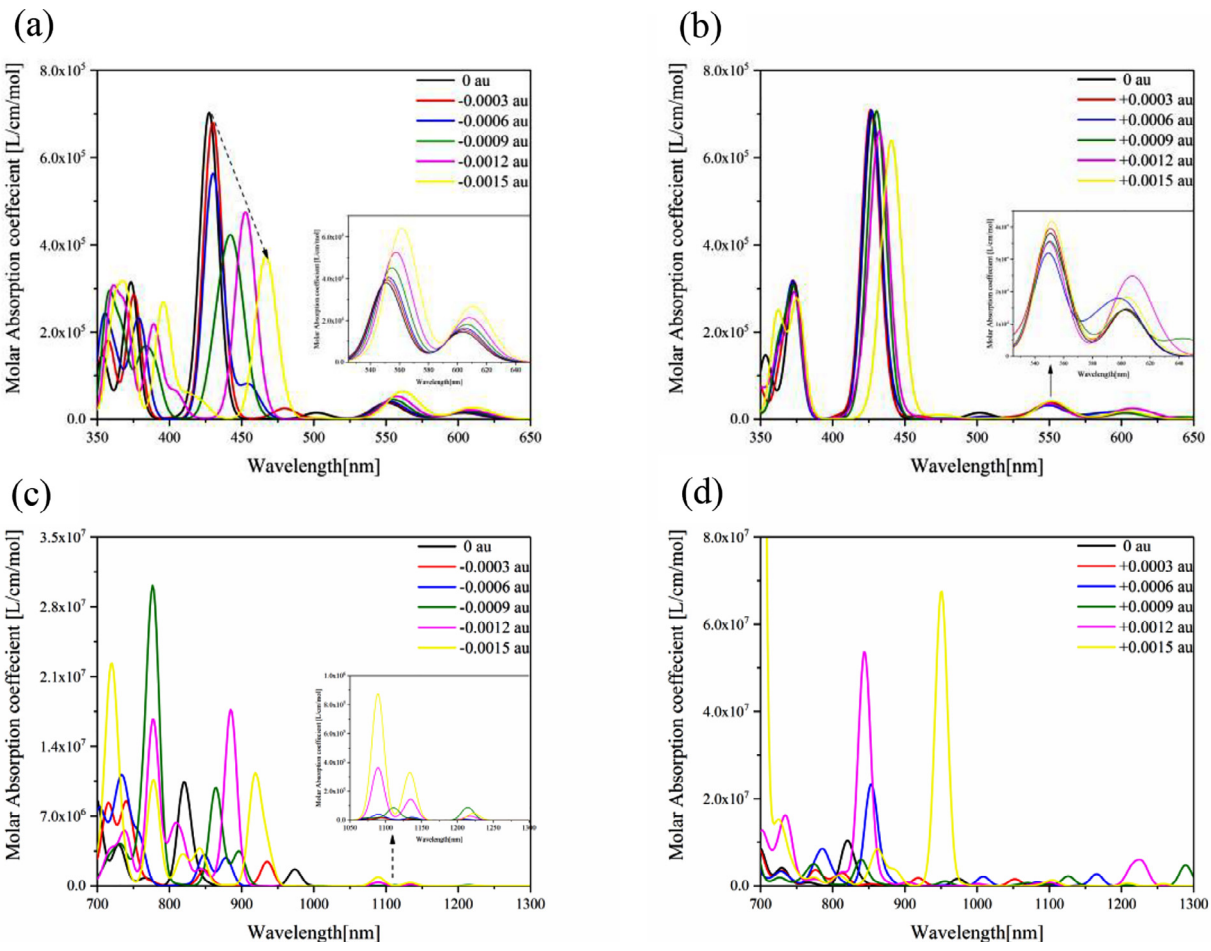


Fig. 2. (a), (b) are the one-photon absorption (OPA) spectra of porphyrin-tetra-thiophene-fullerene (PTF). (c), (d) are the two-photon absorption (TPA) spectra of porphyrin-tetra-thiophene-fullerene. "+" and "−" represent the positive and negative directions of the external electric field along the X axis, respectively.

The TPA and OPA spectra of the D-B-A system in the external electric field $F = -1.5 \times 10^{-3} au$ are shown in Fig. 4(a) and (b). We analyzed the S_{22} at the same peak position as S_{19} in Fig. 4(a), and S_{11} is one of the channels in the OPA spectrum. The 2D and 3D images on the left in Fig. 4(c) can visualize the two-step transition process. First of all, during the transition from S_0 to S_{11}, local excitation occurs on fullerenes and tetra-thiophene, and some electrons on tetra-thiophene are transferred to porphyrin. During the transition from S_{11} to S_{22}, there is an electron-hole coherence phenomenon on fullerenes. But observe that there is no bright area in the lower left corner of the TDM graph. At this time, the electron-holes come from tetra-thiophene and porphyrin. The first transition process of the current channel mainly consists of two parts: one part is the super-exchange charge transfer from porphyrin to fullerene structure; the other part is the sequential charge transfer between thiophene and fullerene. From the 3D picture, the drawbacks of CDD have been revealed. The electron-holes accumulated in fullerenes do not clearly specify the current charge transfer characteristics. The combination with 2D graphics is necessary. In addition, in the channel of another intermediate state S_8, initially the thiophene oligomer transfers electrons to fullerene, which is the $p-\pi*$ transition. Then, two transfer characteristics appeared: one is the sequential charge transfer from fullerene to thiophene chain, and then from thiophene chain to porphyrin. This transfer is similar to the feeling of "relay". And this type of transition has a strong transition density. The other is super exchange charge transfer directly from fullerene to porphyrin. Compared with the previous one, the density is weaker. Although there are bright areas in the 2D images, it can be distinguished that the local excitation presents the characteristics of flocculent aggregation, and the sequential charge transfer presents a grid-like shape with gaps in the middle. And in the two-step process, a charge recombination process occurred between the thiophene chain and the fullerene.

Next, we analyze the electric field in the positive direction. The TPA spectrum and OPA spectrum in the external electric field $F = 1.2 \times 10^{-3} au$ are shown in Fig. 5(a), (b). In Fig. 5(a), the contribution of the "three-state" term and the "two-state" term under the S_{25} absorption peak is almost equal. Therefore, we analyzed $S_0 \rightarrow S_{23}$->S_{25} and $S_0 \rightarrow S_{25}$ for the main visual analysis of S_{25}. First, the two-step transition process of $S_0 \rightarrow S_{23}$->S_{25}, $S_0 \rightarrow S_{23}$ is mainly a local excitation on the thiophene chain. During the $S_{23} \rightarrow S_{25}$ transition, we can see a clear hole isosurface in the CDD diagram. Therefore, it is speculated that there is a process in which electrons are sequentially transferred from porphyrin to thiophene chain, and then from thiophene chain to fullerene. This is mainly due to the lack of super-exchange charge transfer characteristics in the TDM diagram. Secondly, due to the large value of $\left|\Delta\mu_{25.0}\right|^2$ in the formula, the two-state term of the direct transition from $S_0 \rightarrow S_{25}$ cannot be ignored. After visual analysis, there is an optical characteristic that is completely different from that of TPA. The whole transfer process is mainly the super exchange charge transfer between the porphyrin structure and the two thiophene unit structures connected to it and the fullerene.

In order to verify that the difference between the "three-state" item and the "two-state" item really exists. We have added TPA spectrum and OPA spectrum in $F = 1.5 \times 10^{-3} au$, as shown in Fig. 6(a) and (b). In Fig. 6(a), the TPA section of the "three-state" term of S_{19} is

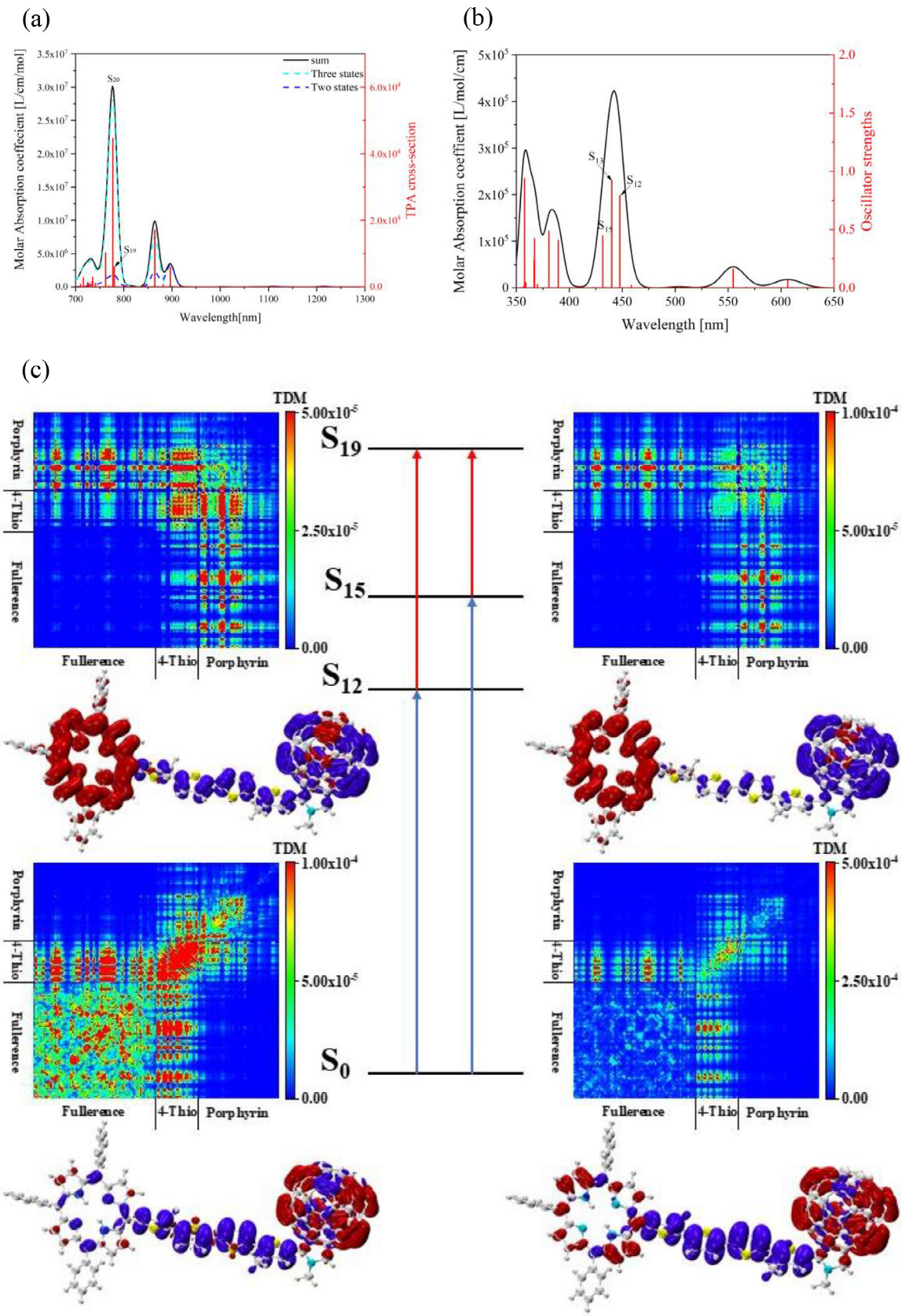


Fig. 3. The external electric field is the TPA spectrum (a) and OPA spectrum (b) of $F = -9 \times 10^{-4} au$. (c) Two-dimensional (2D) graphs representing transition density and three-dimensional (3D) graphs representing electron-hole coherence from different channels. Among them, red and blue represent electrons and holes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

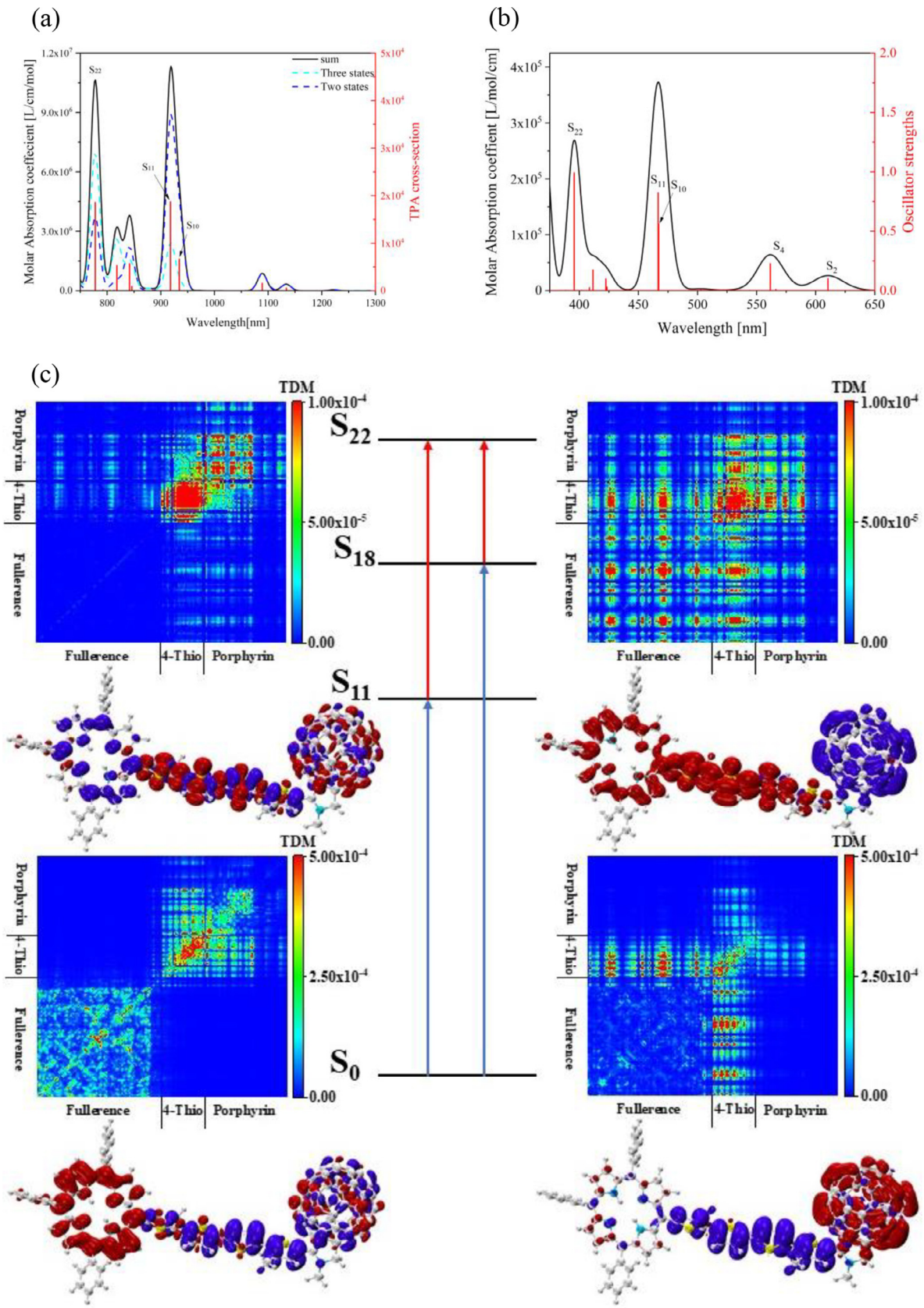


Fig. 4. The external electric field is the TPA spectrum (a) and OPA spectrum (b) of $F = -1.5 \times 10^{-3} au$. (c) Two-dimensional (2D) graphs representing transition density and three-dimensional (3D) graphs representing electron-hole coherence from different channels. Among them, red and blue represent electrons and holes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

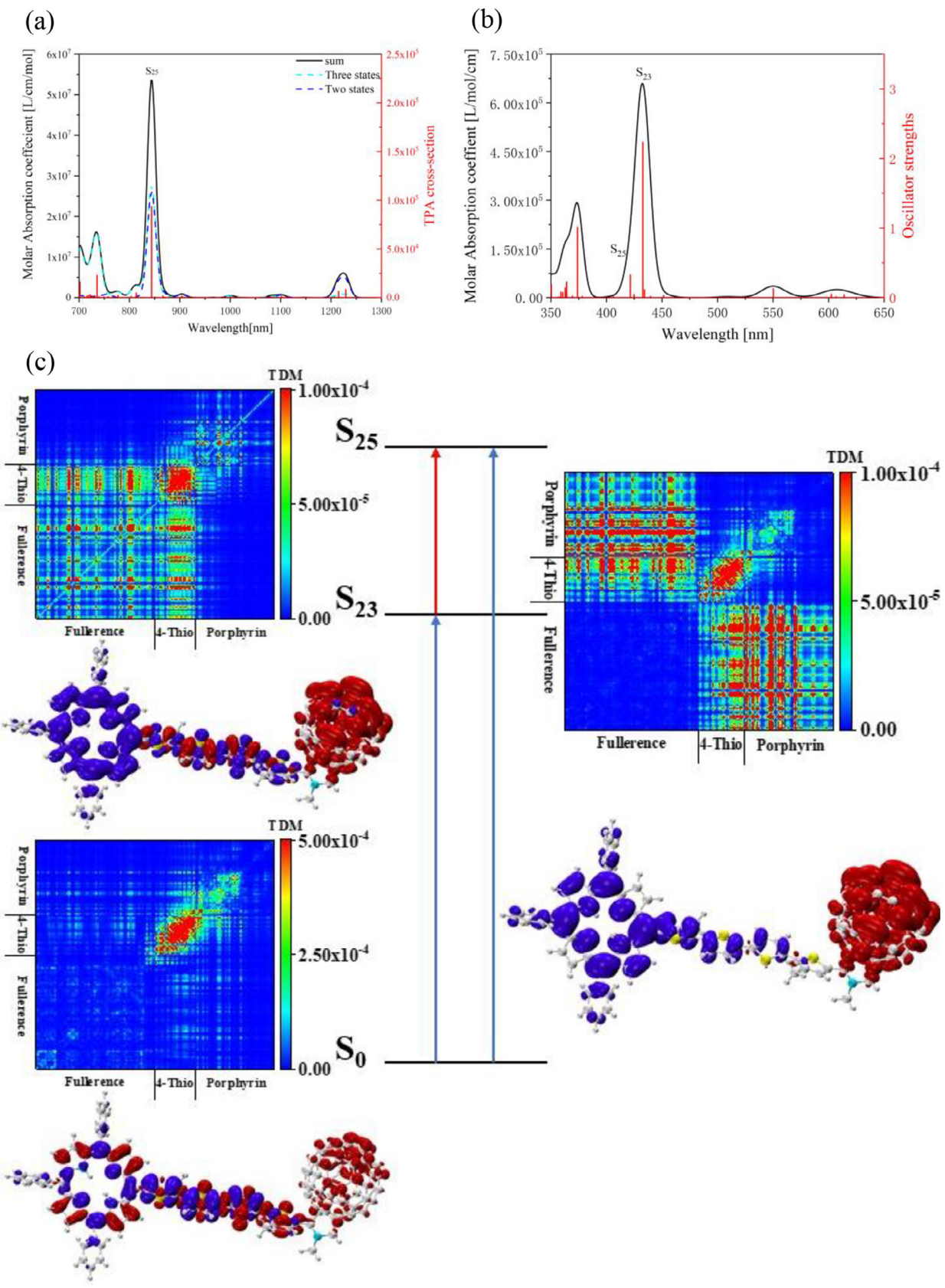


Fig. 5. The external electric field is the TPA spectrum (a) and OPA spectrum (b) of $F = 1.2 \times 10^{-3} au$. (c) Two-dimensional (2D) graphs representing transition density and three-dimensional (3D) graphs representing electron-hole coherence from different channels. Among them, red and blue represent electrons and holes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

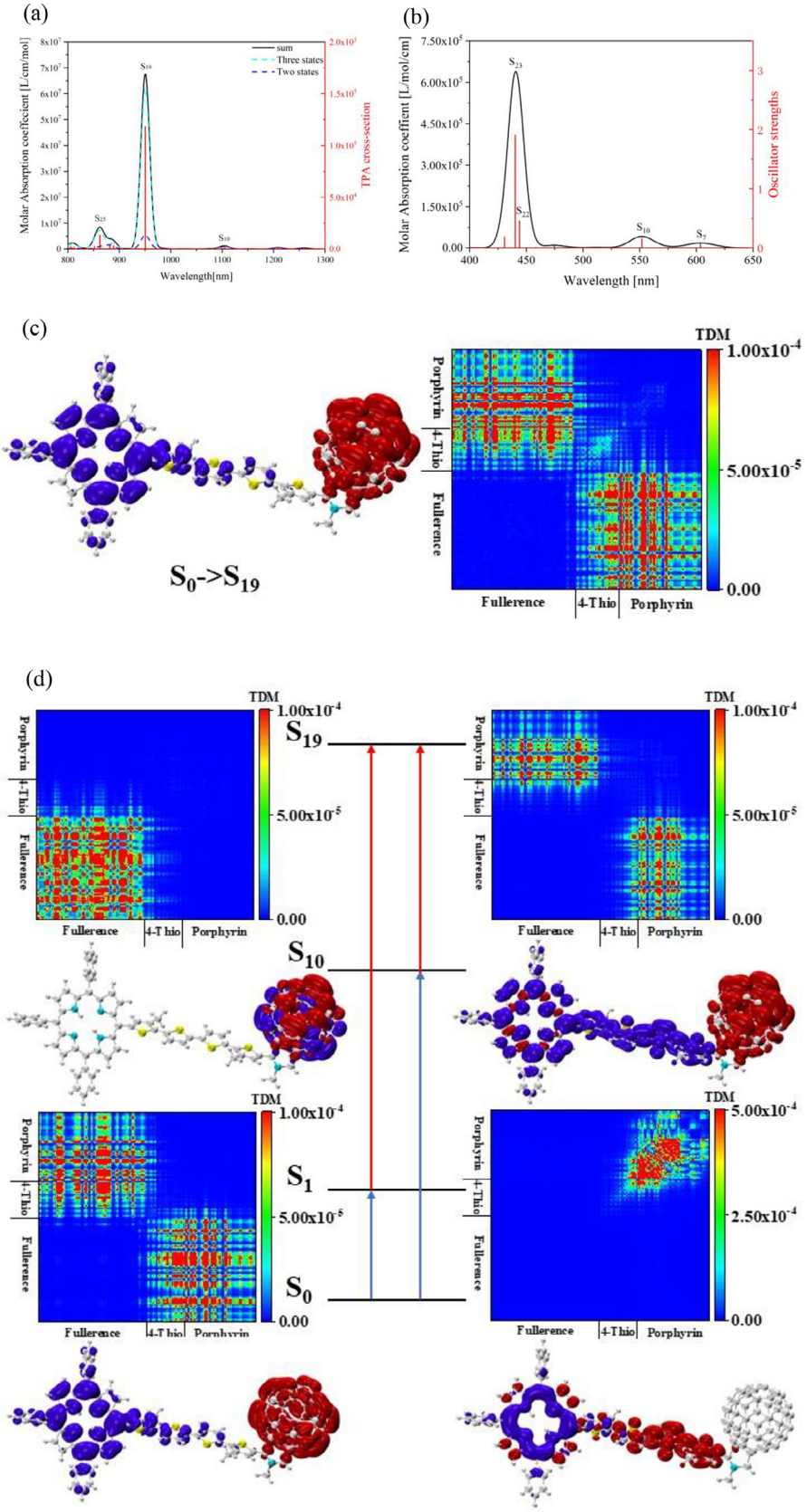


Fig. 6. The external electric field is the TPA spectrum (a) and OPA spectrum (b) of $F = 1.5 \times 10^{-3} au$. (c), (d) Two-dimensional (2D) graphs representing transition density and three-dimensional (3D) graphs representing electron-hole coherence from different channels. Among them, red and blue represent electrons and holes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

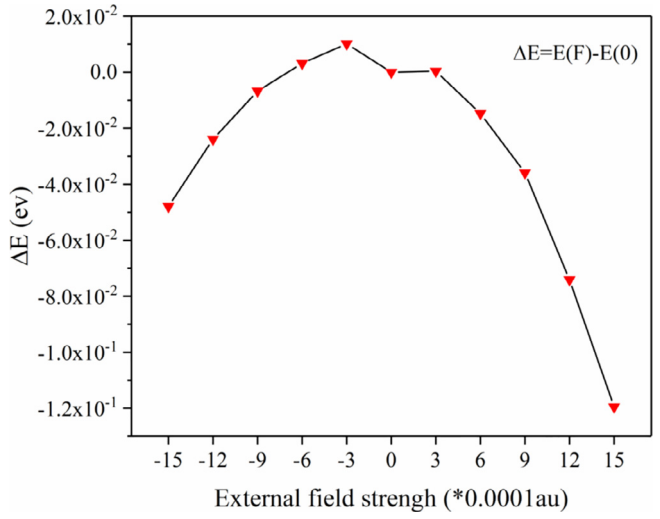


Fig. 7. The correlation between the energy difference of the ground state and different external electric field strengths.

Table 1
The net charge transfer between the three fragments of porphyrin, tetra-thiophene and fullerene.

Transferred electrons between fragments (Net)			
	Porphyrin -> Tetra-thiophenes	Porphyrin -> Fullerene	Tetra-thiophenes -> Fullerene
−0.0003 au	−0.12529	−0.00107	0.02891
−0.0006au	−0.22283	0.01448	0.10277
−0.0009 au	−0.05508	−0.07733	0.07750
−0.0012au	−0.36605	−0.00903	0.03453
−0.0015au	−0.16873	−0.11745	0.11740
+0.0003au	−0.01918	0.00280	0.04483
+0.0006au	0.05616	0.07083	0.15079
+0.0009au	0.02875	0.01560	0.03462
+0.0012au	0.09419	0.04977	0.08620
+0.0015au	0.07575	0.03064	−0.01555

about 7 times that of the "two-state" term. However, the S_{19} excited state is very strong, so the "two states" of the current excited state is worth analyzing compared to other excited states (see Fig. 6(c)). Similarly, similar to the previous external electric field situation, the $S_0 \rightarrow S_{19}$ channel is completely super exchange charge transfer. It is worth discussing the two-step transition process of S_{19}, as shown in Fig. 6(d). When S_1 is used as an intermediate state, the first step is the superexchange transfer phenomenon of porphyrin to fullerene. In the second step, all isosurfaces are gathered on the fullerene structure, which is the local excitation of the fullerene. When S_{10} is the intermediate state, the first step is the sequential charge transfer from porphyrin to thiophene chain. After adding a certain amount of electrons to the thiophene oligomer, the second step is the superexchange transfer of the porphyrin and the thiophene oligomer to the fullerene, respectively.

Macroscopically, when the direction of the external electric field is the negative direction of the X-axis, the strong absorption state in the TPA is used as the final state in the CDD diagram, and the electronic isosurface is more likely to appear on the porphyrin structure. When the direction of the external electric field is the positive direction of the X axis, the electron isosurface is more likely to appear on the fullerene structure. Based on this phenomenon, we divide the D-B-A structure into three segments, which is intended to further confirm the promotion of the same direction charge transfer by an external electric field. (see Table 1).

It can be seen from the table that with the increase of the electric field intensity in the negative direction, the net charge transfer amount of the porphyrin fragment to the fullerene fragment gradually becomes negative. The data shows that the charge is transferred from fullerene and thiophene to porphyrin as a whole, accompanied by the transfer of thiophene to fullerene. We have to consider that the latter may provide part of the electrons for fullerenes. With the increase of the electric field strength in the positive direction, the three net charge values all show an overall trend of increasing gradually. It shows that electrons are also transferred along the positive direction of the X axis. In addition, the number of electrons transferred by sequential charge transfer is larger, and the proportion is larger.

In addition, we observed the lowest potential energy of the ground state in different external electric fields, see Fig. 8. Taking the energy without electric field as the benchmark, the absolute value of the ground state energy difference gradually increases as the intensity of the external electric field in two opposite directions increases. It is found that $|\Delta E = E(F) - E(0)| \leqslant 0.12ev$. The difference can indicate that the ground state energy will be affected by the external electric field and cannot be ignored.

In the previous analysis, we observed the charge transfer characteristics of excited states with strong absorption peaks in the TPA spectra. Light-induced electron transfer has a certain sequence in the current system. The energy of the charge separated state is lower than that of the light excited state. Therefore, we compare and analyze the CDD diagrams of the lowest charge transfer state in different external electric fields in Fig. 8. The results show that the strength and direction of the electric field can significantly manipulate the charge transfer in the lowest charge transfer excited state. As shown in Fig. 8, when the direction of the electric field is along the positive X-axis, the porphyrin structure and the tetrathiophene unit structure transfer electrons to the fullerene structure. When $F = 1.5 \times 10^{-3} au$, the hole density of the long thiophene chain close to the fullerene side is small. It shows that the charge transfer at this time is mainly the super exchange charge transfer from the porphyrin structure to the fullerene structure. As the electric field decreases, the sequential charge transfer from the thiophene chain to the fullerene structure increases. Local excitation on the porphyrin and sequential charge transfer from the thiophene chain to the porphyrin can be observed in the case of the negative electric field and no electric field. When the electric field is $F \leqslant -9 \times 10^{-4} au$, the hole density appears on the fullerene, which may be accompanied by the appearance of super exchange charge transfer.

Therefore, in order to further explore the characteristics of these lowest charge transfer excited states, we recorded their excitation energy and absorption intensity (see Fig. 9). The excitation energy changes relatively smoothly in the five negative electric fields, while the change trend in the positive direction is greater. The changes in the excitation energy of the electric field in the five negative directions are relatively stable, while the change trend in the electric field in the positive direction is greater. It shows that in the charge transfer controlled by the external electric field, the electric field direction is the main influencing factor, followed by the electric field size. The addition of an external electric field may bring about a shift in the distribution of frontier molecular orbitals, resulting in a decrease in the number of transition electrons. So, the energy of the excited state in Fig. 9 is reduced, and at the same time, the red shift of the spectrum in Fig. 2 is caused.

3. Conclusion

In summary, through the analysis of the strong absorption peak in TPA, we found that the applied electric field has an effect on the

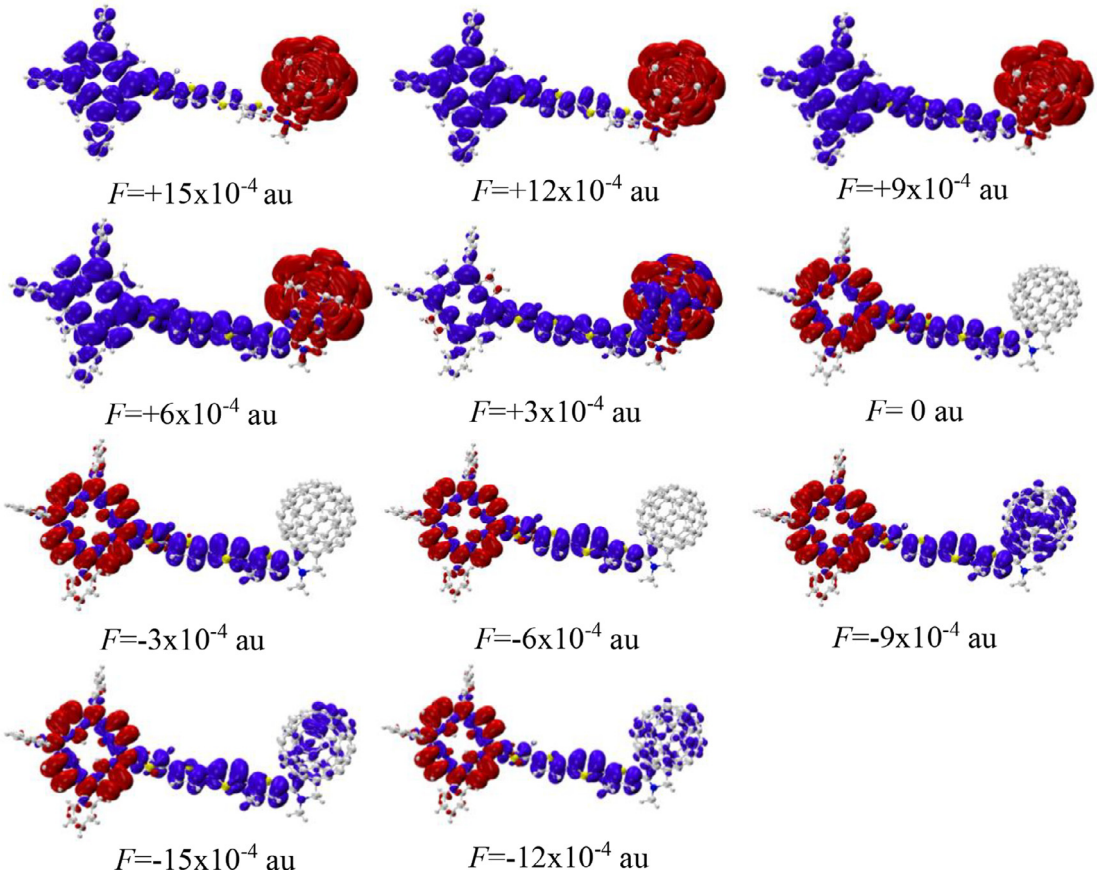


Fig. 8. The charge density difference plot of the lowest charge transfer excited state. Red and blue represent the distribution of electrons and holes, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

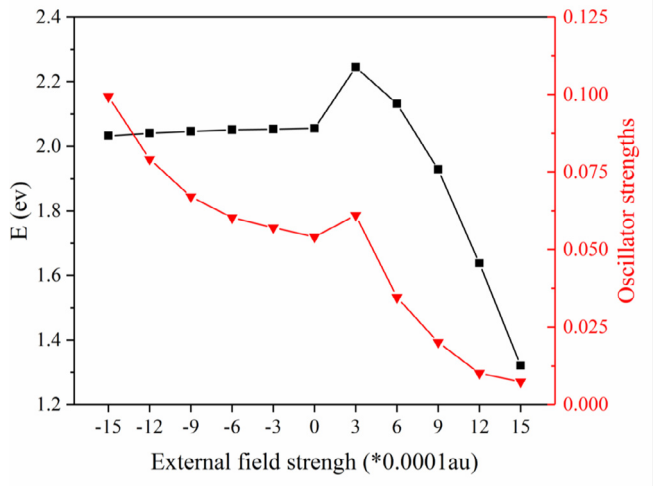


Fig. 9. Excitation energy and absorption intensity of the lowest charge transfer excited state in different electric fields.

charge transfer characteristics of the D-B-A system. In a nutshell, the direction and strength of the applied electric field promotes the occurrence of intra-molecular charge transfer. These include sequential charge transfer and super exchange charge transfer in TPA. During the two-step transition process of the current system TPA, we observe the density matrix and the charge differential density map, and find that when the electric field direction is the same as the direction of the donor pointing to the acceptor, the strength of the charge separation at this time is strong, and the phenomenon of super exchange charge transfer is prone to occur; Conversely, the phenomenon of charge recombination may also occur. This macroscopic adjustment is related to the optical properties of the same direction on the atomic scale. More importantly, this discovery can provide theoretical guidance for solar cells, dyes whose colors can be adjusted by electric fields and photocatalysis.

4. Methods

All the quantum chemical calculations are performed with Gaussian 16 software. S1 The molecular structure of porphyrin-tetra-thio phene-fullerene is optimized by the B3LYP functional (Frisch et al., 2016) theory in density functional theory (DFT) (Becke, 1988) combined with 6-31G(d) basis set (Kohn and Sham, 1965). In addition, the density functional theory D3 correction (Grimme et al., 2011) method is added to optimize the calculation accuracy. On the basis of the optimized structure, the properties of the excited state are calculated using time-dependent density function theory (TD-DFT) (Gross and Kohn, 1985); Cam-B3LYP (Yanai et al., 2004) functional and 6-31G(d) basis set. In our calculations, 200 excited states are calculated, which provides reasonable results for the SOS method. (Lu and Chen, 2012).

Multiwfn 3.6 program (Humphrey et al., 1996) is used for charge differential density (CDD) and transition density matrix (TDM). The VMD program [32] can draw a schematic diagram of the electron-hole pair isosurface. In this article, the red isosurface represents electrons in all CDD images (the electron density in this area increases), and

the blue isosurface represents holes (the electron density in this area decreases). All isosurfaces have a size of 0.0002.

The TPA cross section can be written as

$$\sigma_{tp} = \frac{4\pi^2 a_0^5 \alpha}{15 c_0} \frac{\omega^2 g(\omega)}{\Gamma_f} \delta_{tp} \tag{2}$$

where the c_0 is the speed of light, Γ_f is the lifetime of final state, a_0 is Bohr radius, α is the fine structure constant, ω is the energy of the incident light, and $g(\omega)$ expresses the spectral line profile, which is assumed to be a δ function. The specific manifestation of the transition probability δ_{tp} in Eq. (2) has been given in Eq. (1).

The transition matrix elements in TDM are defined as

$$P_{\mu\nu}^{tran} = \begin{bmatrix} C_{\nu i} C_{\mu i} & C_{\nu i} C_{\nu j} \\ C_{\mu i} C_{\mu j} & C_{\mu i} C_{\nu j} \end{bmatrix} \tag{3}$$

where the $C_{\nu i}$ is the i th orbital expansion coefficient of the μth atom. This formula expresses the transition density matrix between arbitrary excited states.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

National Natural Science Foundation of China (91436102, 113734353); Fundamental Research Funds for the Central Universities (grant 06500067).

References:

Albota, M.A., Xu, C., Webb, W.W., 1998. Two-photon fluorescence excitation cross sections of biomolecular probes from 690 to 960 nm. Appl. Opt. 37, 7352–7356.
Becke, A.D., 1988. Density-functional exchange-energy approximation with correct asymp-totic behavior. Phys. Rev. A 38 (6), 3098–3100.
Davis, W.B., Ratner, M.A., Wasielewski, M.R., 2001. Conformational Gating of Long Distance Electron Transfer through Wire-like Bridges in Donor-Bridge-Acceptor Molecules. J. Am. Chem. Soc. 123, 7877–7886.
Denk, W., Strickler, J., Webb, W., 1990. Two-photon laser scanning fluorescence microscopy. Science 248 (4951), 73–76.
Dorfman, K.E., Schlawin, F., Mukamel, S., 2016. Nonlinear optical signals and spectroscopy with quantum light. Rev. Mod. Phys. 88, 045008.
Frisch, M., Trucks, G., Schlegel, H., Scuseria, G., Robb, M., Cheeseman, J., Scalmani, G., Barone, V., Petersson, G., Nakatsuji, H., et al, 2016. Gaussian 16, revision A03. Gaussian, Inc..
Giacalone, F., Segura, J.L., Martin, N., Guldi, D.M., 2004. Exceptionally Small Attenuation Factors in Molecular Wires. J. Am. Chem. Soc. 126, 5340–5341.
Goppert-Mayer, M., 1931. Elementarakte mitzwei Quanten-sprüngen. Über. Ann. Phys. 401, 273–294.
Grimme, S., Ehrlich, S., Goerigk, L., 2011. Effect of the damping function in dispersion corrected density functional theory. J. Comput. Chem. 32, 1456–1465.
Gross, E.K.U., Kohn, W., 1985. Local Density-Functional Theory of Frequency-Dependent Linear Response. Phys. Rev. Lett. 55, 2850–2852.
Guo, D., Wang, C., Luo, Y., Agren, H., 2003. Influence of electron-acceptor strength on the resonant two-photon absorption cross sections of diphenylaminofluorene-based chromophores. PCCP 5, 3869–3873.
Honig, J., Jortner, J., 1967. Theoretical Studies of Two-Photon Absorption Processes. II. Model Calculations. J. Chem. Phys. 47, 3698–3703.
Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: visual molecular dynamics. J. Mol. Graph. 14 (1), 33–38.
Ikemoto, J., Takimiya, K., Aso, Y., Otsubo, T., Fujitsuka, M., Ito, O., 2002. Porphyrin-oligothiophene-fullerene triads as an efficient intramolecular electron-transfer system. Org. Lett. 4, 309–311.
Junya, I., Kazuo, et al, 2002. Porphyrin-Oligothiophene-Fullerene Triads as an Efficient Intramolecular Electron-Transfer System. Org. Lett. 4 (3), 309–311.
Kaiser, W., Garrett, C.G.B., 1961. Two-Photon Excitation in CaF2: Eu2+. Phys. Rev. Lett. 7, 229–231.
Kang, D., Zhu, S., Liu, D., Cao, S., 2020. One- and Two-Photon Absorption: Physical Principle and Applications, M. Sun. Chem. Rec. 20, 894–911.
Kato, S., Matsumoto, T., Ishi-i, T., Thiemann, T., Shigeiwa, M., Gorohmaru, H., Maeda, S., Yamashita, Y., Mataka, S., 2004. Strongly red-fluorescent novel donor-π-bridge-acceptor-π-bridge-donor (D-π-A-π-D) type 2,1,3-benzothiadiazoles with enhancedtwo-photon absorption cross-sections. Chem. Commun. 20, 2342–2343.
Kohn, W., Sham, L.J., 1965. Self-consistent equations including exchange and correlation effects. Phys. Rev. 140, A1133–A1138.
Lu, T., Chen, F., 2012. Multiwfn: a multifunctional wavefunction analyzer. J. Comput. Chem. 33 (5), 580–592.
Mirebeau, I., Goncharenko, I.N., Cadavezperes, P., Bramwell, S.T., Gingras, M.J., Gardner, J., 2000. Pressure-induced crystallization of a spin liquid. Nature 420 (6911), 54–57.
Mongin, O., Porres, L., Katan, C., Pons, T., Mertz, J., Blanchard-Desce, M., 2003. Synthesis and two-photon absorption of highly soluble three-branched fluorenylene-vinylene derivatives. Tetrahedron Lett. 44, 8121–8125.
Mu, Xijiao, Wang, Jingang, Sun, Mengtao, 2019. Visualization of Photoinduced Charge Transfer and Electron-Hole Coherence in Two-Photon Absorption. J. Phys. Chem. C 123 (23), 14132–14143.
Mu, Xijiao, Wang, Xinxin, Quan, Jun, Sun, Mengtao, 2020. Photoinduced Charge Transfer in Donor-Bridge-Acceptor in One- and Two-photon Absorption: Sequential and Superexchange Mechanisms. J. Phys. Chem. C 124 (9), 4968–4981.
Otsubo, T., Aso, Y., Takimiya, K., 2002. Functional Oligothiophenes as Advanced Molecular Electronic Materials. J. Mater. Chem. 12, 2565–2575.
Paulson, B.P., Miller, J.R., Gan, W.X., Closs, G., 2005. Superexchange and Sequential Mechanisms in Charge Transfer with a Mediating State between the Donor and Acceptor. J. Am. Chem. Soc. 127, 4860–4868.
Song, Peng, Li, Yuanzuo, Ma, Fengcai, Pullerits, Tõnu, Sun, Mengtao, 2013. External Electric Field-Dependent Photoinduced Charge Transfer in a Donor-Acceptor System for an Organic Solar Cell. J. Phys. Chem. C 117 (31), 15879–15889.
Song, Peng, Li, Yuanzuo, Ma, Fengcai, Sun, Mengtao, 2015. Insight into external electric field dependent photoinduced intermolecular charge transport in BHJ solar cell materials. J. Mater. Chem. C 3 (18), 4810–4819.
Sun, M.T., Chen, J.N., Xu, H.X., 2008. Visualizations of transition dipoles, charge transfer, and electron-hole coherence on electronic state transitions between excited states for two-photon absorption. J. Chem. Phys. 128, 064106.
Sun, M., Peng, S., Chen, Y., et al, 2005. Intramolecular charge transfer in the porphyrin–oligothiophene–fullerene triad. Chem. Phys. Lett. 416 (1–3), 94–99.
Yanai, T., Tew, D.P., Handy, N.C., 2004. A new hybrid exchange – correlation functional using the Coulomb-attenuating method (CAM-B3LYP). Chem. Phys. Lett. 39, 351–357.

Dynamic control over group speed of light in plasma cladded optical fiber: An analytical approach

Sushree Sangita Jena, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sushreesangita665.com*

Susmita Mohapatra, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, susmitamohapatra963@gmail.com*

Smruti Samantray, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, smrutisamantray23@hotmail.com*

Supriya Nayak, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, supriyanayak443@gmail.com*

ARTICLE INFO

Keywords:
Slow wave structure
Group index
Dispersion management
Unconventional optical waveguides

ABSTRACT

The ability of plasma frequency to tune the guiding properties of plasma cladded optical fiber is explored and numerically investigated. The identified parameter to manipulate the propagation characteristics of the investigated structure is electron–ion density which can be controlled by varying electric potential. It is shown that the curve of group index vs optical signal frequency can be manipulated significantly by manipulating plasma frequency. Judicious consideration of fiber's parameters and the values of plasma frequency allow for obtaining minimum or desired group velocity dispersion at a desired optical frequency. Further, ability of manipulation of the slope of group index vs signal frequency curve allows for dispersion management. Possibility of significant variation in group-index is the highlight of the present work and the most impressive feature is that it could be achieved in online condition (by varying plasma frequency with the help of tuning electric field). Since, ratio of plasma to signal frequency is the only important factor in manipulating the propagation properties, the idea presented here can be extended to frequency domains in the range of Terahertz, Microwaves etc. An optical modulator is proposed in the last section on the basis of present investigation.

1. Introduction

The analyses of varied propagation properties of optical signal in bounded structures has always remained of prime interests amongst scientists and engineers to explore its vivid applications in integrated optical devices and processing systems. Different kinds of optical waveguides with various schemes of cross-sectional geometries (Xiao et al., 2019; Kim et al., 1987; Singh and Kumar, 2009) and core-cladding materials (Anicin, 2000; Hu and Wei, 2001; Shen and Pao, 1991; Chatterton and Shohet, 2007; Mishra et al., 2013) were proposed and investigated by many researchers to address the recent applications related issues. Emphasis continues to design an optical waveguide structure that may efficiently transport image data, enables strong light-matter interaction, all optical computing and more importantly tunable guiding features.

Cylindrical optical waveguides loaded with unconventional materials such as chiral materials (Janeiro et al., 2002), metamaterials (Yamunadevi et al., 2016), semiconductors (Ballato et al., 2010), plasma etc. have shown to be very useful to design novel, chip-scale, ultrafast devices for applications in terahertz wireless communications and in all-optical computing. However, efforts continue to identify some dynamic parameters (particularly with electrical tuning) that may manipulate the guiding properties online in order to make the fiber versatile with tunable guiding properties. Some authors have also explored the guiding properties of a bi-waveguide (Foteinopoulou and Vigneron, 2013) and reported some interesting results. This bi-waveguide consists of slabs of positive index and negative index materials.

Amongst various optical waveguide configurations proposed so far, there has been growing interest in optical fibers considering plasma either in its core (Shen, 1991) or in cladding (Singh et al., 2010; Hairong et al., 2007; Mishra and Singh, 2015) due to its electrically controlled frequency dependent refractive index. Considering plasma in the cladding – the so called plasma cladded optical fiber (PCOF), has potential to address the issues of two-fold coupling of propagating hybrid modes in image transferring fiber optic system (KianiMajd et al., 2018). Moreover, the unique feature of PCOF which distinguishes it from other optical waveguide structures is its ability to

manipulate the propagation properties dynamically (in online condition). This dynamic manipulation of propagation properties of modes is possible by changing the electrical potential which in turn alters the values of electron ion densities and hence plasma frequency.

Recent works on PCOF suggest for its numerous applications in optical storage devices and modulating systems (Mishra, 2021).

Present work on PCOF encompasses the studies of group index and effective index and is extended to explore the possibilities of its device applications. The effects of various fiber parameters on group index is also explored. The proposed optical fiber (Fig. 1) consists of a circular dielectric core surrounded with concentric, isotropic, homogeneous, low density and cold plasma. Further, to avoid the radiation loss, it is assumed that the plasma region extends infinitely in transverse direction. The characteristic dispersion relation of PCOF is derived using Maxwell's equations and appropriate boundary conditions. The identified dynamic parameter of PCOF is plasma frequency which can be tuned electrically. The novel feature of PCOF is its electrically controlled refractive index of cladding region that is exploited to explore some of manipulable propagation characteristics of guided modes.

2. Physical structure and formulation of the problem

The physical model of the waveguide structure (Fig. 1) under investigation is considered to be a dielectric core with coaxial cold plasma cladding. The wave is supposed to be essentially varying harmonically in space and time coordinates and be propagating in the z –direction. The refractive index of plasma cladding region (n_2) is defined below by the relation (Krall et al., 1973)

$$n_2 = \sqrt{1 - \frac{\omega_p^2}{\omega\,(\omega - i\gamma)}} \tag{1}$$

where ω is the signal frequency, γ is loss constant and ω_p is the plasma frequency. For a collision-less plasma which is judiciously tenable at low pressure, the loss constant γ becomes zero, hence, Eq. (1) reduces in the form

$$n_2 = \sqrt{1 - \frac{\omega_p^2}{\omega^2}} \tag{2}$$

The behavior of Electromagnetic wave propagating through a plasma media is quite dissimilar under low and high frequency conditions. Under low frequency conditions ($\omega < \omega_p$), the refractive index of plasma turns to be complex and hence waves get attenuated while in high frequency conditions ($\omega > \omega_p$), plasma behaves like a loss-less dielectric media. Furthermore, to restore the assumptions specific to cold plasma, the thermal motion of electrons and ions are neglected (Krall et al., 1973). Thus, the plasma frequency for a collision-less and cold plasma can be expressed by the relation

$$\omega_p = \sqrt{\frac{e^2 n}{m_e\,\varepsilon_0}} \tag{3}$$

The constants e, n, ε_0 and m_e are the charge, electron density, permittivity of free space and mass of the electron respectively. Some of the existing plasma generation techniques being used for practical purposes, with their corresponding parameters is enlisted in Table 1.

The derivation of characteristic dispersion equation primarily involves the derivation of wave equation which essentially requires to be solved to get the longitudinal and transverse field components in core and cladding regions (Pollock, 1995) separately. These field components are then matched at core-cladding boundary ($r = a$) using appropriate boundary conditions. Following the usual mathematical steps as applicable for a standard circular step-index dielectric waveguide, the dispersion relation of the PCOF comes out to be

$$\left\{\frac{J_\nu'(u\,a)}{u J_\nu(u\,a)} + \frac{K_\nu'(w\,a)}{w K_\nu(w\,a)}\right\} \times \left\{\frac{k^2 n_1^2}{u}\frac{J_\nu'(u\,a)}{J_\nu(u\,a)} + \frac{k^2 n_2^2}{w}\frac{K_\nu'(w\,a)}{K_\nu(w\,a)}\right\} - \frac{\beta^2 \nu^2}{a^2}\left\{\frac{1}{u^2} + \frac{1}{w^2}\right\}^2 = 0 \tag{4}$$

where the representative of fields J_ν and K_ν are Bessel and modified Bessel functions of first and second kind respectively, β is longitudinal propagation constant, ν is modal index and core and cladding parameters are respectively defined below as

$$u^2 = k^2 n_1^2 - \beta^2 \text{ and}$$

$$w^2 = \beta^2 - k^2 n_2^2$$

The prime over J_ν and K_ν represent the first order differentiation with respect to argument. A detailed discussion about cylindrical Bessel functions and their properties can be obtained in (Arfken and Weber, 2005). Eq. (4) is the standard dispersion equation of a step-index optical fiber in the sense that one can derive separate dispersion relation for different modes (transverse symmetric and hybrid modes) from above relation. Using Bessel function identities and making use of Eq. (2), the dispersion relation of fundamental mode of the PCOF follows

$$\left\{-\frac{J_{\nu+1}(ua)}{u J_\nu(ua)} - \frac{K_{\nu-1}(wa)}{w K_\nu(wa)} + \frac{\nu}{a}\left(\frac{1}{u^2} + \frac{1}{w^2}\right)\right\} \times k^2\left\{-n_1^2\left(\frac{J_{\nu+1}(ua)}{u J_\nu(ua)} - \frac{\nu}{a\,u^2}\right) - \left(\frac{K_{\nu-1}(wa)}{w K_\nu(wa)} + \frac{\nu}{a\,w^2}\right)\right\} - \frac{\beta^2 \nu^2}{a^2}\left(\frac{1}{u^2} + \frac{1}{w^2}\right)^2 = 0 \tag{5}$$

In our forthcoming analyses, we have considered only fundamental mode (HE_{11}) in our discussions since it is one of the most investigated fiber mode due to its outstanding propagation properties carried by it.

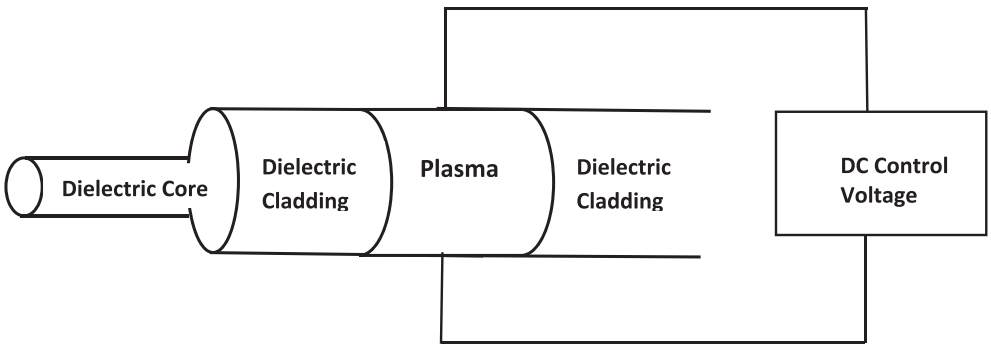


Fig. 1. Cross-sectional view of circular optical fiber with constant value of dielectric core's refractive index and a plasma cladding with varying refractive index (Mishra, 2021).

Table 1
Plasma generation techniques and their corresponding properties (Mehmood et al., 2018).

Source of Plasma	Nominal Power/W	Electron number density (m^{-3})	Plasma frequency (Hz) (Calculated using Eq. (3))
RF glow discharge	200–500	10^{17}	~10^{11}
DC glow discharge	100–300	10^{16}	~10^{10}
Inductively coupled	500–2000	10^{18}	~10^{11}
Electron cyclotron resonance	300–1000	10^{18}	~10^{11}
DC welding arc	500–2000	10^{19}–10^{23}	~10^{12}–10^{15}
Helicon	500–2000	10^{18}–10^{19}	~10^{11}–10^{12}
DC plasma jet	1000–20000	10^{20}–10^{23}	~10^{12}–10^{15}

However, one can derive dispersion relation for transverse symmetric modes considering $\nu = 0$ in Eq. (4) and making use of appropriate Bessel function identities.

3. Numerical results

The study of group index is an important aspect in determining the group velocity dispersion and slow or fast nature of propagating modes. It is conveniently defined by the relation;

$$N_g = n - \lambda \frac{dn}{d\lambda} \tag{6}$$

The regions where an increase in wavelength causes a decrease in refractive index indicates for slow light. In order to explore explicitly the effect of core radius and core refractive index, we present here two more equations;

$$v_g = \frac{d\omega}{d\beta} \tag{7}$$

$$\beta^2 = \frac{b^2 V^2}{a^2} + \frac{1}{c^2}\left(\omega^2 - \omega_p^2\right) \tag{8}$$

where v_g is group velocity. b and V are two other important dimensionless fiber parameters defined below

$$b = \sqrt{\frac{\left(\frac{\beta}{k}\right)^2 - n_2^2}{n_1^2 - n_2^2}} \tag{9}$$

$$V = a\left(\sqrt{u^2 + w^2}\right) = k\,a\sqrt{n_1^2 - n_2^2} \tag{10}$$

Eq. (6) together with Eqs. (7) and (8) gives the values of group index. The allowed values of β can be obtained by solving Eq. (5) which is pre-required for calculating normalized propagation constant (b).

(i) effect of variation of core radius on group index

A curve between group index and signal frequency is plotted for different values of plasma frequency and core radius. It is fairly evident from Fig. 2 that group index is dominantly affected with the variations in both plasma frequency and core radius. Compared to other considered values of plasma frequency, the slope of group index is obviously very large for $\omega_p \approx \omega$. It may further be noted from the curve that for $a = 400$ nm and $\omega_p = 6 \times 10^{14}$ Hz, the slop of group index is almost flat in C- Band of communication spectrum which indicates minimum group velocity dispersion. It is important to mention here that higher values of group index reflect slowing of light which is very crucial in controllable optical delay devices, optical buffers and modulation systems (Boyd et al., 2006). It may also be noted that, within C-band, the slope of the group index vs signal frequency curve also changes significantly with the plasma frequency. Possibility of slope manipulation in online condition should have potential applications in dispersion management. Further, the sign of the slope could be changed within C-band using plasma frequency which should be crucial in dispersion compensation.

(ii) effect of variation of core refractive index on group index

Knowledge of the effect of variation of core refractive index and core radius on group index is crucial from fabrication point of view. The curve between group index and signal frequency (Fig. 3) at constant core radius and different values of plasma frequency displays the varied effect of core refractive index. It may be noted that the curves are highly grouped and the slope of curves varies very slightly by varying core refractive index (keeping ω_p fixed). Point of intersection of the intersecting curves indicate same group speed but different group dispersion at the optical signal frequency corresponding to the point of intersection. Points of intersections in the C-Band and beyond may be noted in the figure.

(iii) effective index

The study of effective index (n_{eff}) is very important in order to quantify the optical response of media. It is conveniently defined by the relation

$$n_{eff} = \sqrt{\frac{b^2 V^2}{k^2 a^2} + n_2^2} \tag{11}$$

To draw substantial information from effective index curve (Fig. 4), we define here new characteristic plasma parameter;

$$\delta = \frac{\omega_p}{\omega} \tag{12}$$

As can be seen in effective index curves that unlike with standard dielectric optical fibers, the magnitude and slope of n_{eff} of PCOF is quite dissimilar for ω_p equal to 10^{15} Hz and 10^{14} Hz. This dissimilarity in magnitude and slope of n_{eff} primarily appears due to variations in electrically controllable n_2. This feature of PCOF clearly suggests its utility in optical storage devices and modulating systems wherein the slow- speed of optical signal plays a crucial role.

(iv) the optical modulator:

A modulator could easily be proposed on the basis of present investigation which is shown in the Fig. 5. The proposed modulator is consisting of a Mach-Zehnder Interferometer (MZI) with one arm of a simple single mode fiber (SMF) and the other of a PCOF.

When an optical beam is launched at the input port, it splits into two equal parts. One part propagates through the SMF and its counterpart through the PCOF. The speed of the part propagating through the PCOF could be changed (modulated) by modulating the electric potential of PCOF. When modulated part obtained through PCOF combines with the reference part through SMF, intensity modulated beam is obtained at the output.

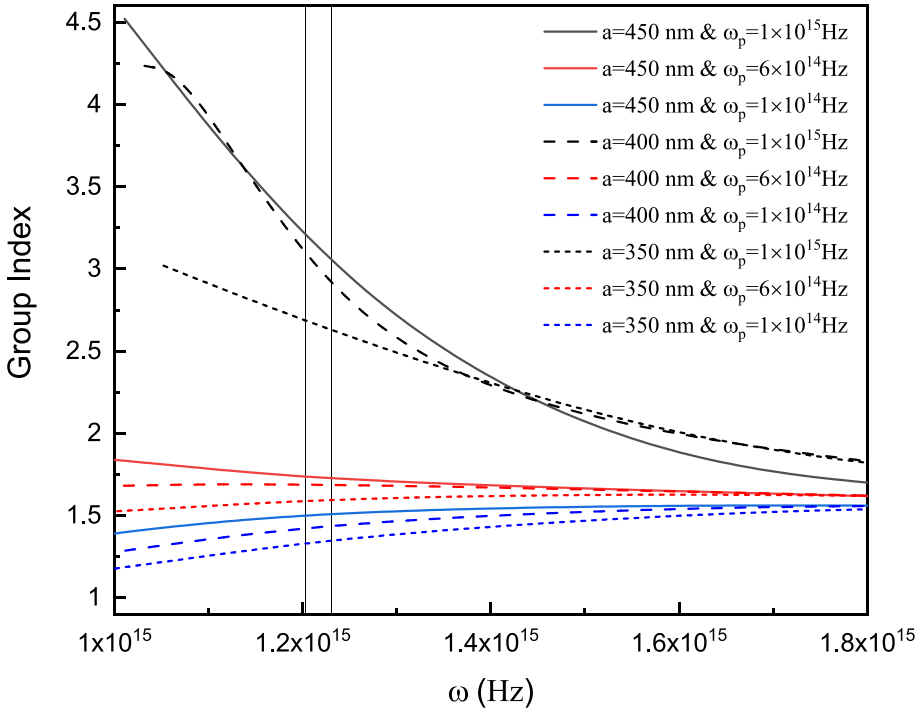


Fig. 2. Variation of group index vs signal frequency of fundamental mode as a function of core radius and plasma frequency. The part of the curve under two vertical lines correspond to the C-Band of communication spectrum.

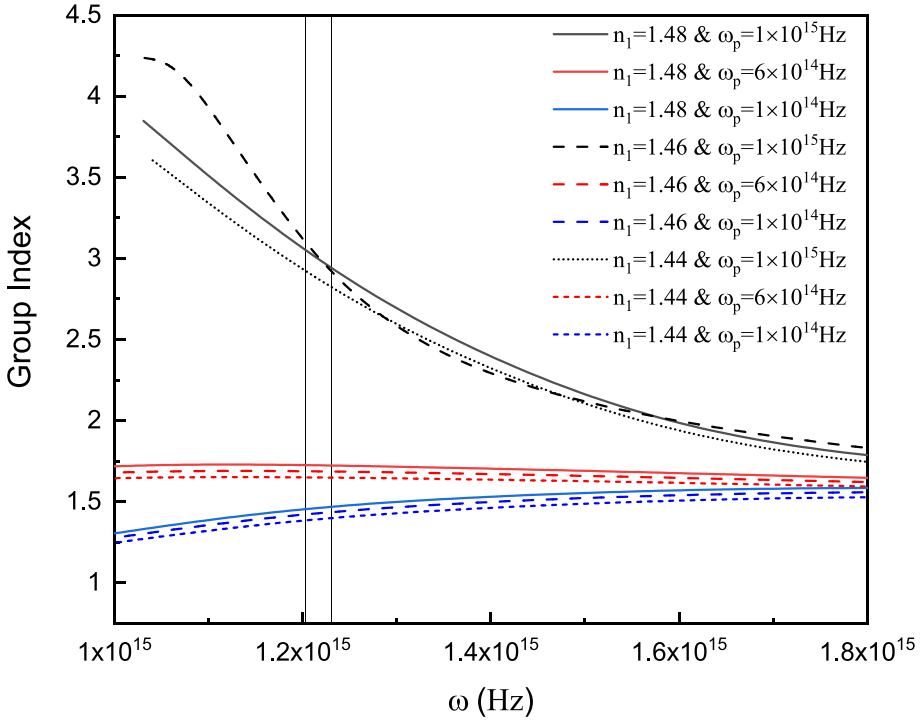


Fig. 3. Variation of group index vs signal frequency of fundamental mode as a function of core refractive index and plasma frequency. The part of the curve under two vertical lines correspond to the C-Band of communication spectrum.

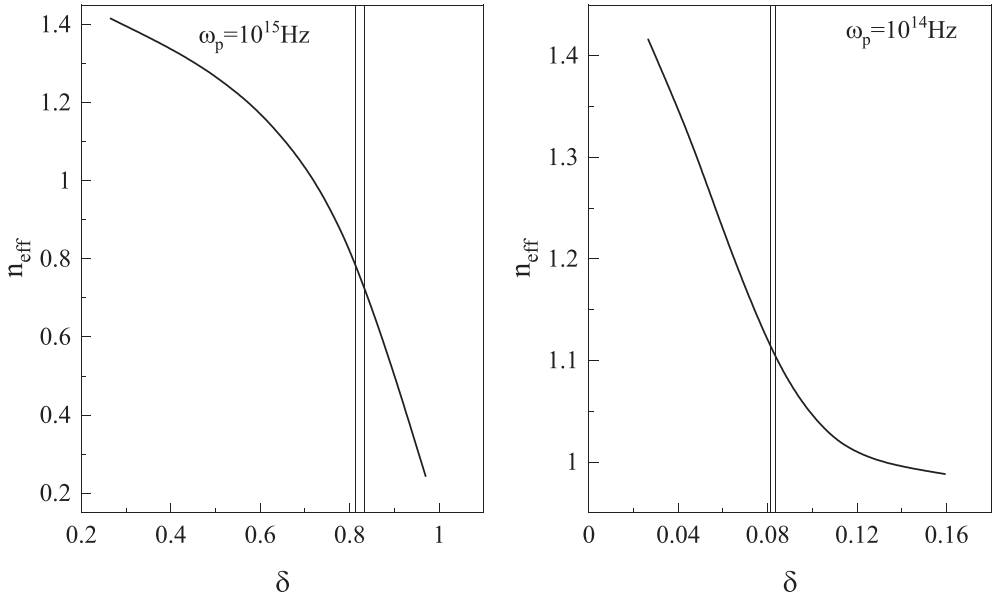


Fig. 4. Variation of effective index with optical frequency. The part of curves under two parallel vertical lines corresponds to the C-band of communication spectrum. A variation in δ at a fixed ω_p entails a corresponding variation in ω according to Eq. (12).

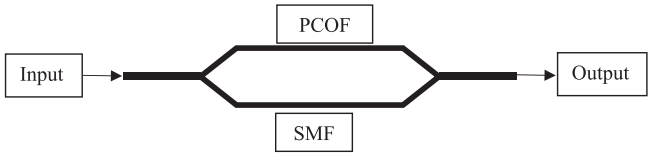


Fig. 5. Proposed optical modulator consisting of a Mach-Zehnder Interferometer (MZI) with one arm of simple single mode fiber (SMF) and the other of a PCOF.

4. Conclusions

Some of the new propagation features of electrically tunable PCOF is explored and based on the derived results, an optical modulator is proposed. Within C-band, the slope of the group index vs signal frequency curve changes significantly with the plasma frequency. Possibility of slope manipulation in online condition should have potential applications in dispersion management. For $a = 400nm$ and $\omega_p = 6 \times 10^{14} Hz$, a variation of group index with signal frequency is almost flat in C-Band which clearly suggest for minimum group velocity dispersion (GVD). It is further impressive to mark here that the slope of these GVD curves for $a = 450nm$, $\omega_p = 6 \times 10^{14} Hz$ and $a = 450nm$, $\omega_p = 1 \times 10^{14} Hz$ is approximately opposite in nature thereby leading to be instrumental in dispersion management/compensation. It may also be noted further that the sign of the slope could be changed within C-band using plasma frequency. Intersecting GVD curves are observed in our investigation. Point of intersection of the intersecting GVD curves indicate same group speed but different group dispersion at an optical signal frequency corresponding to the point of intersection. The points of intersections can however, be shifted in C-Band by varying the refractive index of core. An optical modulator based on Mach-Zehnder interferometer is also proposed in the last section. The obtained results of the investigated optical fiber structure are appearing to be very promising and may attract other peer investigators for more advanced studies in addition to its fair possibilities of applicability in integrated optical devices and optical processing systems.

References

Anicin, B.A., 2000. Plasma loaded helical waveguide. J. Phys. D: Applied Phys. 33, 1276–1281.
Arfken, G.B., Weber, H.J., 2005. Mathematical Methods for Physicists. Academic Press, California.
Ballato, J., Hawkins, T., Foy, P., Yazgan-Kokuoz, B., McMillen, C., Burka, L., Morris, S., Stolen, R., Rice, R., 2010. Advancements in semiconductor core optical fiber. Opt. Fiber Technol. 16 (6), 399–408.
Boyd, R.W., Gauthier, D.J., Gaeta, A.L., 2006. Applications of slow light in telecommunications. Opt. Photonics News 17 (4), 18. https://doi.org/10.1364/OPN.17.4.000018.
Chatterton, J.D., Shohet, J.L., 2007. Guided modes and loss in a plasma filled Bragg waveguide. J. Appl. Phys. 102 (6), 063304. https://doi.org/10.1063/1.2776374.
Foteinopoulou, S., Vigneron, J.P., 2013. Extended, slow-light field enhancement in positive-index/negative-index heterostructures. PRB 88, 195144-1–8.
Hairong, L., Changjian, T., Pukun, L., 2007. Mode theory of plasma cladding waveguide. J. Phys. D: Appl. Phys. 40, 2002–2009.
Hu, B.J., Wei, G., 2001. Numerical analysis of a plasma waveguide with finite thickness of cladding in external magnetic field. IEE Proc. Microwave Antennas Propagation 148 (2), 115. https://doi.org/10.1049/ip-map:20010218.
Janeiro, F.M., Paiva, C.R., Topa, A.L., 2002. Guidance and leakage properties of chiral optical fibers. J. Opt. Soc. Am. B 19 (11), 2558. https://doi.org/10.1364/JOSAB.19.002558.
KianiMajd, M., Hasanbeigi, A., Mehdian, H., Hajisharifi, K., 2018. Dispersion properties of plasma cladded annular optical fiber. Phys. Plasmas 25 (5), 053505. https://doi.org/10.1063/1.5019669.
Kim, B., Blake, J.N., Huang, S.Y., Shaw, H.J., 1987. Use of highlly elliptical core fibers for two-mode fiber devices. Opt. Lett. 12, 729–731.

N.A. Krall, A.W. Trivelpiece, Principles of Plasma Physics, McGraw-Hill, New Delhi, 1973.

F Mehmood, T Kamal, U Ashraf, Generation and applications of plasma (an academic review), Preprints (2018) 10.20944/preprints201810.0061.v1.

Mishra, A.K., Medhekar, Sarang, Parashar, J., Kumar, Mukesh, 2021. Investigation of plasma cladded optical fiber for dynamic manipulations of its propagation properties. Optik. https://doi.org/10.1016/j.ijleo.2021.166537.

Mishra, A.K., Kumar, M., Kumar, D., Singh, O.N., 2013. Modal study of plasma cladded cylindrical optical fiber. J. Electromagnet. Waves Appl. 27 (7), 868–876.

Mishra, A.K., Singh, O.N., 2015. Simplified study of guided modes in plasma cladded step-index optical fiber. Opt. Commun. 345, 120–124.

Pollock, C.R., 1995. Fundamentals of Optoelectronics. Tom Casson, Chicago.

Shen, H.M., 1991. Plasma waveguide: A concept to transfer electromagnetic energy in space. J. of Appl. Phys. 69, 6827–6835.

Shen, H.-M., Pao, H.-Y., 1991. The plasma waveguide with a finite thickness of cladding. J. Appl. Phys. 70 (11), 6653–6662.

Singh, S.P., Janam, R., Jatan, R., Singh, V., Singh, B.D., 2010. Modal analysis and cutoff condition of a circular doubly clad optical waveguide with plasma in inner cladding region. J. Infrared Millimeter Terahertz Waves 31, 1381–1389.

Singh, V., Kumar, D., 2009. Modal dispersion characteristics of a Bragg fiber having plasma in the inner cladding regions. Prog. Electromagnet. Res. 89, 167–181.

Xiao, H., Li, H., Wu, B., Dong, Y., Xiao, S., Jian, S., 2019. Elliptical hollow-core optical fibers for polarization-maintaining few-mode guidance. Opt. Fiber Technol. 48, 7–11.

R Yamunadevi, D Shanmuga Sundar, A Sivanatha Raja, Characteristics analysis of metamaterial based optical fiber, Optik 127 (2016) 9377-9385.

Realization and optimization of optical logic gates using bias assisted carrier-injected triple parallel microring resonators

Asheerbad Pradhan, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, asheerbadpradhan46@gmail.com*

Smruti Samantray, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, smrutisamantray23@hotmail.com*

Priya Chandan Satpathy, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, priyachandan.satpathy57@outlook.com*

Supriya Nayak, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, supriyanayak443@gmail.com*

ARTICLE INFO

Keywords:
Microring resonator
Optical logic gates
Carrier injection
Bandfilling
Bandgap shrinkage
Free carrier absorption

ABSTRACT

We propose a p-i-n diode embedded parallel triple microring resonator (MRR) configuration to simultaneously realize optical OR and AND, or NAND and NOR logic gates using a bias-assisted carrier injection mechanism. The applied bias on the rings induces refractive index change in the intrinsic region through bandfilling, bandgap shrinkage and free carrier absorption effects, leading to intensity variation at the output ports of the MRR due to respective resonant wavelength shift. The optical logic gate operational outputs are represented as the light intensities at the output ports of the MRR with the wavelength of the input optical signal launched into the input port being at the resonant wavelength of the microring resonator, while the operands are represented as the bias applied onto the rings. The proposed microring resonator configuration is theoretically optimized for achieving a high contrast ratio and an optical confinement factor by optimizing the intrinsic region width, applied bias, coupling coefficient between ring-bus waveguide and lifetime of carriers in the intrinsic region.

1. Introduction

Optical logic gates are the key elements needed to realize high speed optical signal processing systems and optical computing. Hardy and Shamir (2007) introduced a new platform in signal processing, which simultaneously enables high speed operation and control convenience. The high-speed and control benefits of optics and electronics can be collectively used to develop electro-optics-based logic circuits whereby electrons and photons are used for signal control/switching and signal operation, respectively (Soref, 2011). Optical directed logic gates open up computational parallelism, thus lead to higher packaging density in integrated circuits (ICs). Performing optical operations independently on each switching element in a network enables boolean logic functions to be realized with reduced latency and over-all processing time. Optical logic gates based on III-V semiconductors are highly recommended for monolithic integration with other III-V devices (e.g., lasers, photodetectors, amplifiers, etc.) in chip platform. To enable control/switching in logic gates, several switching mechanisms such as electro-optic (EO) effect (Kumar et al., 2014), electro-absorption (EA) effect (Fayza and Sooraj, 2020) and thermo-optic (TO) effect (Tian et al., 2011) can be used. However, each of these mechanisms has its own drawbacks. The EO effect is relatively weak in III-V semiconductors, and hence, the length of an EO device must be long enough in order to attain a large change in optical output with a practical applied bias voltage, which makes it incompatible for chip integration. EA based logic gates are polarization sensitive and the wavelength operation of these devices should be chosen near the bandgap of the material as EA effect is strongest only for those wavelengths that are near the bandedge wavelength. Group III-V semiconductor materials with bandgap close to lowest attenuation and dispersion wavelength window (around 1300 nm and 1550 nm, respectively) are Indium-based ternary and quaternary materials, such as InAlAs/InGaAs and InGaAsP/InP. The major drawbacks of these materials are their scarcity and immature processing technologies. The TO based logic gates faces thermal mismatch issues along with poor stability and reliability. An alternative switching mechanism that works well with III-V semiconductor materials is carrier injection (CI) (Ishida et al., 1987). The major attractive features of CI mechanism are polarization independent nature, operational simplicity, high contrast ratio and freedom to operate over a wavelength band far away from

the bandgap. Through a bias-assisted CI mechanism, varying the bias voltage applied to the III-V semiconductor material changes its refractive index through bandfilling (BF), bandgap shrinkage (BGS) and free carrier absorption (FCA) effects. Unlike silicon-based logic gate devices, GaAs-based devices enable monolithic integration with other semiconductor optoelectronic devices. The limitation faced by CI modulation-based devices implemented using Mach Zehnder Interferometer and directional coupler structures in ICs is their bulky nature (Ito and Tanifuji, 1988; Abdalla, 2004). However, a solution that overcome this limitation is to use microring resonator (MRR) structures (Dominik, 2007).

MRR configurations have been widely used in photonic-electronic ICs to realize different combinational and sequential logic gates. The basic single MRR structure, shown in Fig. 1 comprises a ring waveguide closely coupled to bus waveguides. The unique properties provided by optical MRRs are small foot print (in few $\mu m \times \mu m$ range), narrow band filtering and high Q factor. The optical signal launched at the input port of a bus waveguide will eventually gets coupled to the ring waveguide, resulting in constructive interference for resonant wavelengths. The resonance condition in a MRR having a resonant wavelength λ_R is represented as $m\lambda_R = 2\pi R n_{eff}$, where R is the radius of ring, m is an integer and n_{eff} is effective refractive index of ring. In Fig. 1, α, τ, κ and φ represents the absorption coefficient, transmission coefficient, coupling coefficient between ring and bus waveguide, and signal phase as propagating through the ring, respectively (Dominik, 2007). The coupling between the waveguides and ring is assumed to be lossless for the proposed logic gate configuration (i.e., $\tau^2 + \kappa^2 = 1$). At resonance, the launched input signal which is resonant ($\lambda = \lambda_R$) appears at the drop port (D port) while any non-resonant wavelength ($\lambda \neq \lambda_R$) appears at throughput port (T port). Due to resonance nature, MRRs are effective in realizing a large contrast ratio (CR) at the output ports for a very small refractive index change.

In this paper, we introduce a novel bias-assisted CI-based p-i-n diode embedded parallel triple MRR configuration that simultaneously realizes OR and AND gates at the output ports (T port and D ports) when a resonant wavelength (resonant at zero bias) is launched at the input ports. The bias applied to the rings acts as the operands while the operational results are collected at the respective output ports. The performance of the proposed logic gate configuration is optimized by optimizing the applied bias, intrinsic region width, relaxation time of carriers and coupling coefficient between ring and bus waveguides. Launching an optical signal whose wavelength satisfies the resonance condition at non-zero bias at the input port leads to the simultaneous realization of NOR and NAND gates at the respective output ports. Section 2 discusses the bias-assisted CI mechanism by considering the BF, BGS and FCA effects, along with the modelling of CI-based triple parallel MRR configured logic gates. The output intensities and respective optimization results obtained for the proposed logic gate configuration are presented in Section 3 and the paper is concluded in Section 4.

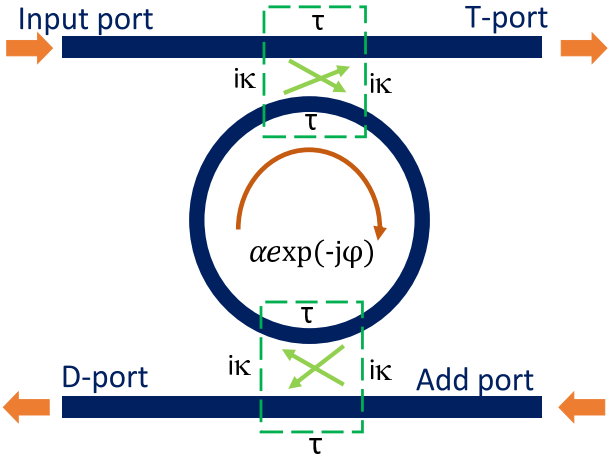


Fig. 1. Schematic diagram of a single MRR. Signals launched at the Input port that satisfies the resonant condition ($m\lambda_R = 2\pi R n_{eff}$) couples to the ring and gets collected at D port while the non-resonant signal bypasses the ring and appears at T port. Additional signals can be launched into MRR using Add port.

2. Modeling of carrier injected parallel triple ring resonators based optical gates

The perspective view of the proposed logic gate configuration is shown in Fig. 2. The structure consists of three identical parallel rings of identical radii, R, seperated by a sufficient minimum gap that the cross coupling between the individual rings is negligible. The entire logic gate structure is grown on a GaAs substrate. The device is comprised of p-i-n diodes having GaAs intrinsic (i) regions, which act as light carrying media (core regions), and surrounding p-AlGaAs and n-AlGaAs regions, which have lower refractive indices than the intrinsic region and act as claddings that confine the optical signal to the intrinsic regions (cores). Injection of carriers into i-GaAs is enabled by depositing p and n electrodes on the top of the rings and beneath the GaAs substrate. Since the bandgap wavelength of the GaAs material is typically far less than the low attenuation and dispersion wavelengths, efficient operation can be attained at these long wavelengths. The material exhibits appreciable refractive index change with a small applied bias voltage. The injection of carriers in a forward-biased ring resonator structure induces a refractive index change that shifts the resonant wavelength, thus routing a portion of optical signal launched into the input port to the T and D ports of the subsequent MRRs. The resulting intensity variation at the T port and D port of MRR3 (refer to Fig. 2) leads to the realization of OR and AND or NAND and NOR gates. The principle of operation and realization of OR and AND or NAND and NOR logic gates using carrier injective p-i-n embedded parallel triple ring resonators are discussed in the next section.

2.1. Principle of operation of carrier injected p-i-n diode embedded MRRs

Applying through the electrodes, forward biases to the p-i-n structured MRRs results in electron and hole injection into the intrinsic regions. The change in refractive index in the i-GaAs is attributed to the change in absorption associated with BF (Burstein-Moss phenomenon, i.e. bandgap of a semiconductor material increases as the absorption edge is shifted to higher energies due to conduction band state population.), BGS and FCA (plasma effect) due to the injected carriers (Bennet et al., 1990). Note that, BF and BGS are interband absorption effects while FCA is intraband absorption effect. In the following calculations of switching characteristics, the band shapes are assumed to be parabolic. By injecting carriers, photons of energies slightly greater than the nominal bandgap energy experience lower absorption because the electrons and holes fill the conduction and valance band, thus leading to the BF effect. The BGS effect happens when the injected carrier concentration density exceeds the critical carrier concentration density (material parameter dependent). This effect lowers the conduction band edge and increase the valance band edge, thus shrinking of the bandgap energy, which increases the absorption for photon of energies less than the nominal bandgap energy. Note that, while calculating the BF effect, the change in the band gap due to BGS must also be included and typically the contributions of the BF and BGS effects can be combined together in the calculations (Ravindran et al., 2012). The critical carrier density estimation (χ_{cr}) and bandgap energy reduction ($\Delta E_g(\chi)$) is modeled according to Bennet et al. (1990) and Bennet and Soref (1987). The resulting change in absorption ($\Delta\alpha(\chi, E)_{BF+BGS}$) due to the combined BF and

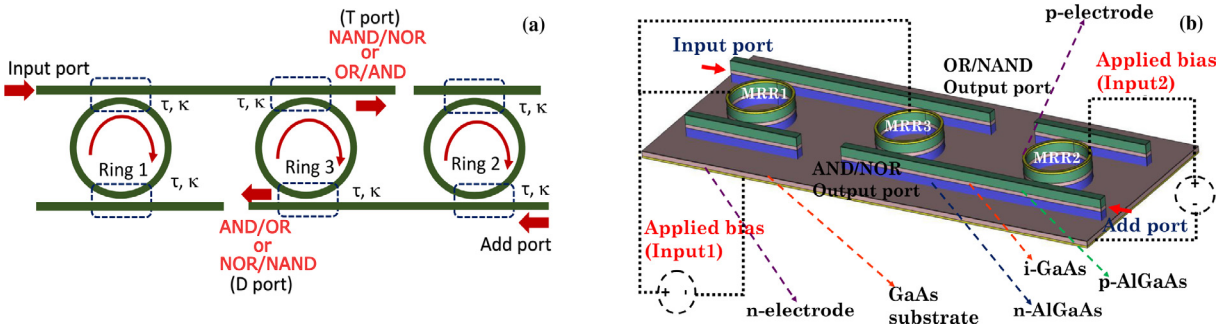


Fig. 2. (a) Parallel triple ring MRR configuration for OR and AND or NOR and NAND gate realization. (b) Perspective view of MRRS for realizing OR and AND or NOR and NAND gate based on CI. The T port of MRR3 acts as an OR gate and the D port as an AND gate at zero bias for the resonant signal. The T port of MRR3 acts as a NAND gate and D port as a NOR gate at a non-zero bias (the bias applied on MRR3 is same as MRR1).

BGS effects is expressed as a function of carrier density (χ) in i-GaAs and photon energy (E) as:

$$\Delta\alpha(\chi,E)_{BF+BGS} = \frac{C_{hh}}{E}((f_v(E_{ah})) - (f_c(E_{bh})) \times \sqrt{E-(E_g-\Delta E_g(\chi))} + \frac{C_{lh}}{E}((f_v(E_{al})) - (f_c(E_{bl})) \times \sqrt{E-(E_g-\Delta E_g(\chi))} - \left(\frac{C_{hh}}{E}\sqrt{E-E_g} + \frac{C_{lh}}{E}\sqrt{E-E_g}\right) \quad (1)$$

Here, E_{ah} and E_{bh} denotes the state of energy levels in valance band and conduction band by considering heavy holes, while E_{al} and E_{bl} denotes the state of energy levels in valance band and conduction band by considering light holes respectively. The Fermi–Dirac probability distribution function in these respective states are represented by f_v and f_c. The constants (C_{hh} and C_{lh}) are obtained from Bennet et al. (1990) by correcting with a multplicative term $\sqrt{\hbar}$ and units as cm^{-1}-eV. Here E_g denotes the bandgap energy of the core region. The change in absorption is accompanied by a change in refractive index (Δn_{BF+BGS}) and it can be calculated by applying Kramers–Kronig integration to Eq. (1) and substituting the appropriate Cauchy principal value. The maximum change in refractive index due to BF and BGS occurs near the bandgap region and become negligible for photon energies much less than the bandgap energy.

FCA is due to the absorption of photons by free carriers (electrons and holes) and the intraband transition of free carriers occurring from one energy state to a higher energy state. FCA-based refractive index change (Δn_{FCA}) dominates at lower photon energies or higher operating wavelengths, since Δn_{FCA} is proportional to the square of the operating wavelength. Δn_{FCA}, which depends on the injected electron and hole densities ($\chi_{e,p}$), refractive index (n) and propagating wavelength through the core i-GaAs region, is given as

$$\Delta n_{FCA} = -\frac{e^2\lambda^2}{8\pi^2c^2\varepsilon_0 n}\left(\frac{\chi_e}{m_e}+\frac{\chi_p}{m_h}\right) \quad (2)$$

where e, c, ε_0, m_e and m_h are the electronic charge, velocity of light in free space, free space permittivity and effective masses of electron and holes, respectively. The total change in refractive index (Δn_{total}) is the sum of the Δn_{BF+BGS} and Δn_{FCA}.

2.2. Realization of triple parallel MRRs based optical logic gates

Incorporating the CI modulation mechanism into the proposed triple parallel MRRs configuration leads to realization of different logic gates. The launched resonant photon energy of the signal at the input port of MRR1 and add port of MRR2 in the proposed configuration is selected to be far less than bandgap energy of the intrinsic region in the rings. Three identical rings are used in the proposed configuration shown in Fig. 2, where L is the circumference of each of the rings. The input power of the launched light signal is denoted as P_I and the normalized output power levels at the T port of MRR1 (ring1) and D port of MRR2 (ring2) are represented as P_{T1} and P_{D2} respectively. MRR3 acts as a Fredkin gate with the T port output of Ring1 fed as an input into the input port of Ring3 and the add port of Ring3 fed with the D port output of Ring2 (Shamir et al., 1986). The normalized power levels at T1 and D2 ports are given by

$$P_{T1,D2} = \left|\tau + \frac{(i\kappa)^2\tau\exp\left(\frac{i2\pi Ln_{1,2}}{\lambda}-\alpha L\right)}{1-\tau^2\exp\left(\frac{i2\pi Ln_{1,2}}{\lambda}-\alpha L\right)}\right|^2 \quad (3)$$

while P_{T3} and P_{D3} are the normalized powers at T port and D port of MRR3 (ring3) and are given by

$$P_{T3} = \left|\tau + \frac{(i\kappa)^2\tau\exp\left(\frac{i2\pi Ln_3}{\lambda}-\alpha L\right)}{1-\tau^2\exp\left(\frac{i2\pi Ln_3}{\lambda}-\alpha L\right)}\right|^2 \times P_{T1} + \left|\frac{(ik)^2\exp\left(\frac{i\pi Ln_3}{\lambda}-\alpha L/2\right)}{1-\tau^2\exp\left(\frac{i2\pi Ln_3}{\lambda}-\alpha L\right)}\right|^2 \times P_{D2} \quad (4a)$$

$$P_{D3} = \left|\tau + \frac{(i\kappa)^2\tau\exp\left(\frac{i2\pi Ln_3}{\lambda}-\alpha L\right)}{1-\tau^2\exp\left(\frac{i2\pi Ln_3}{\lambda}-\alpha L\right)}\right|^2 \times P_{D2} + \left|\frac{(ik)^2\exp\left(\frac{i\pi Ln_3}{\lambda}-\alpha L/2\right)}{1-\tau^2\exp\left(\frac{i2\pi Ln_3}{\lambda}-\alpha L\right)}\right|^2 \times P_{T1} \quad (4b)$$

The refractive indices of the rings (n_1 for Ring1, n_2 for Ring2, n_3 for Ring3) is n_{eff} at zero applied, and changes to n_{eff}-Δn_{total} at an applied bias. Note that the add port of Ring1 must be isolated from the input port of Ring2 as shown in Fig. 2 to restrict the entry of any optical signal through the add port of Ring1, which can result in nonfunctioning of the gates. The bias voltage applied to Ring1 and Ring2 is considered as the operands, and the logic output is collected as light intensities at the T port and the D port of Ring3. An optical signal which is resonant either at zero bias or at an applied bias is continuously launched through the input port of Ring1 and the add port of Ring2, depending whether an OR/AND or a NAND/NOR function is needed. The bias applied to Ring3 controls the ports through which OR/AND and NAND/NOR functions are obtained. If the bias applied to Ring3 is same as that applied to Ring1, then, depending on whether the launched optical signal is resonant at zero bias or non-zero bias, an OR or a NAND gate is realized at the T port of Ring3, while the D port of Ring3 realizes the AND or NOR gate, respectively. If the bias applied on Ring3 is same as that applied to Ring2, then an OR or a NAND gate is realized at the D port of Ring3, while the T port of Ring3 realizes an AND or a NOR gate, respectively.

The operation of the device as a logic gate is subsequently described. Here, we considered that the applied bias on Ring3 is the same as that on Ring1. Also consider that $\lambda = \lambda_{0R}$ is the wavelength

which satisfies resonance condition with zero bias being applied to the rings, while $\lambda = \lambda_{1R}$ satisfies the resonance condition when the rings are biased with a non-zero voltage.

Case 1: (Input1 = 0, Input2 = 0, $\lambda = \lambda_{0R}$): A zero bias applied to Ring1 and Ring2 results in resonance conditions at Ring1, Ring2 and Ring3 for optical signals of wavelength λ_{0R} launched into the input port of MRR1 and the add port of MRR2. This scenario results in low optical intensities (logic 0) at the T port and the D port of Ring3, as no input optical signals enter into the ports of Ring3.

(Input1 = 0, Input2 = 0, $\lambda = \lambda_{1R}$): A zero bias applied to Ring1 and Ring2 results in non-resonance conditions at Ring1, Ring2 and Ring3 for optical signal of wavelength λ_{1R} launched into the input port of MRR1 and the add port of MRR2. This scenario results in high optical intensities (logic 1) at the T port and the D port of Ring3.

Case 2: (Input1 = 0, Input2 = 1, $\lambda = \lambda_{0R}$): When the optical signals of wavelength λ_{0R} are launched at the input port of Ring1 and the add port of Ring2, a non-zero bias applied to Ring2 (i.e., MRR2 becomes out of resonance, thus bypassing Ring2) yields a high-intensity optical signal at the drop port of Ring2. The drop port of Ring2 is connected to the add port of Ring3, and thus, the optical signal launched into the add port of Ring2 couples into Ring3 (Ring3 is at resonance as no bias is applied onto it) and appears at the T port of Ring3, resulting in a high-intensity optical signal (logic 1) at the T port and a low-intensity optical signal (logic 0) at the D port of Ring3 (as the T port intensity of MRR1 is low).

(Input1 = 0, Input2 = 1, $\lambda = \lambda_{1R}$): When optical signals of wavelength λ_{1R} launched at the input port of Ring1 and the add port of Ring2, a non-zero bias applied on Ring2 (i.e., MRR2 becomes resonant, thus couples to Ring2) yields a high-intensity optical signal at the T port of Ring2. A zero bias applied to Ring1 makes MRR1 out of resonance, yielding a high-intensity optical signal at the T port of MRR1. Since the T port of MRR1 is connected to the input port of Ring3, as Ring3 is driven with the same bias as Ring1, MRR3 becomes out of resonance, and this results in a high-intensity optical signal (logic 1) at the T port of MRR3 and a low-intensity optical signal (logic 0) at the D port of MRR3.

Case 3: (Input1 = 1, Input2 = 0, $\lambda = \lambda_{0R}$): A non-zero bias applied to Ring1 (i.e., MRR1 becomes out of resonant, thus bypasses Ring1) yields a high-intensity optical signal at the T port of Ring1. Since the T port of Ring1 is connected to the input port of Ring3, the optical signal launched into the input port of Ring1 bypasses Ring3 (Ring3 is out of resonance, as a non zero bias is applied onto it) and appears at the T port of Ring3, resulting in a high-intensity optical signal (logic 1) at the T port and a low-intensity optical signal (logic 0) at the D port of Ring3 (since the optical signal intensity at the D port of MRR2 is low and at the T port is high).

(Input1 = 1, Input2 = 0, $\lambda = \lambda_{1R}$): A non-zero bias applied to Ring1, at which MRR1 becomes resonant, yields a high-intensity optical signal at the D port of Ring1. A zero bias applied to Ring2 drives MRR2 out of resonance and thus, a high-intensity optical signal appears at the D port of MRR2 and the intensity of optical signal at T port of MRR2 becomes low. As the D port of MRR2 is connected to the add port of Ring3, when both Ring3 and Ring1 share the same bias, MRR3 is driven into resonance, yielding a high-intensity optical signal (logic 1) at the T port and a low-intensity signal (logic 0) at the D port of MRR3.

Case 4: (Input1 = 1, Input2 = 1, $\lambda = \lambda_{0R}$): A non-zero bias applied to both Ring1 and Ring2 results in non-resonance conditions for MRR1 and MRR2, and hence, zero bias resonant-wavelength signals launched at the input port of Ring1 and the add port of Ring2 bypass Rings 1, 2 and 3, yielding high-intensity optical signal (logic 1) at both the T port and the D port of Ring3.

(Input1 = 1, Input2 = 1, $\lambda = \lambda_{1R}$): A non-zero bias applied to both Ring1 and Ring2 drive them into resonance, and hence, non-zero bias resonant-wavelength signals launched to the input port of MRR1 and the add port of MRR2 couple into Rings 1 and 2, and appear at the D port of MRR1 and the T port of MRR2, resulting in low-intensity optical signals (logic 0) at both the T port and the D port of MRR3.

The above-described operations of the proposed triple parallel MRR based logic gate configuration demonstrate its ability to realize OR and AND gate at T port and D port for a zero bias resonant wavelength λ_{0R} or NAND and NOR gate at T port and D port for a non-zero bias resonant wavelength λ_{1R}, respectively.

3. Results and discussions

Based on the principle of the proposed triple parallel MRR configuration for operation as OR/AND and NAND/NOR logic gates, shown in Fig. 2 and as discussed in the previous section, the logical outputs are collected at the T port and D port of MRR3, while the input continuous-wave resonant (at zero or non-zero bias) wavelength signals are launched at input port of MRR1 and add port of MRR2. The average carrier density injected into the i-GaAs region during forward biasing is calculated by solving the electron and hole transport equations (Piprek, 2003). The induced carrier density leads to resonant wavelength shift (due to a change in the refractive index), which results in changes in the intensities of the optical signals at the output ports. The light intensity change at the D and T port is decided by the change in the refractive index which in turn depends on the applied bias and therefore on injected carrier density in the intrinsic region of the p-i-n diode. The available carrier density is also decided by the carrier life time and the active volume availabe for the carriers to recombine. A smaller life time leads to smaller carrier concentration as the rate at which carriers recombine would be fast. Similarly, a larger active volume would also result in smaller carrier concentration as the carrier recombination would be large (Piprek, 2003). The coupling coefficient variation affects the output port intensities (Yariv, 2000) in the proposed MRR logic gate configurations and the optimum coupling coefficient to attain maximum CRs at output ports depends upon the refractive index change in the rings resulting from the applied bias. The optimization of the proposed logic gate configuration is therefore achieved by optimizing the width (w) of the i-GaAs region, the applied bias, the lifetime of carriers in the i-GaAs region and the coupling coefficients of the ring-bus waveguides for maximizing contrast ratio and optical field confinement in the light carrying core region of the proposed logic-gate configuration. As confinement increases, the cladding losses decreases and leads to appearance of higher fraction of intensities at the output ports. Thus helps to achieve better contrast ratio between the output port intensities representing the logic 0 and logic 1 conditions in MRR logic gate.

The absorption and refractive index change are estimated by considering the contribution of the BF, BGS and FCA effects (Eqs. (1) and (2)) as described in Section 2.1. Increasing the width of the i-GaAs region enables better optical signal confinement, however, the average carrier concentration injected into the core region during forward biasing gets reduced, due to the increase in electron–hole recombination rate. Note that, increasing the applied bias increases the average concentration in the core region, however, also increases the current and therefore the power dissipation. A larger bias also results in reduced optical signal confinement in the intrinsic GaAs region, since the refractive index difference between i-GaAs core and AlGaAs cladding reduces. All these effects constrains the range of the applied bias to 1.4 V-3.0 V and intrinsic region width of 0.4–0.5 μm.

The microrings in the proposed configuration are assumed to be identical of radius 5 μm. The launched resonant wavelength at zero bias is selected to be 1559.7 nm, a wavelength that falls in the window where the fiber attenuation is minimal. The dependency of n_a on the width of the intrinsic region, the carrier lifetime and the applied bias is depicted in Fig. 3, and the corresponding Δn and $\Delta \alpha$ is calculated by referring Eqs. (1) and (2). It is obvious from Fig. 3 that the average carrier concentration increases with increasing both the bias voltage and

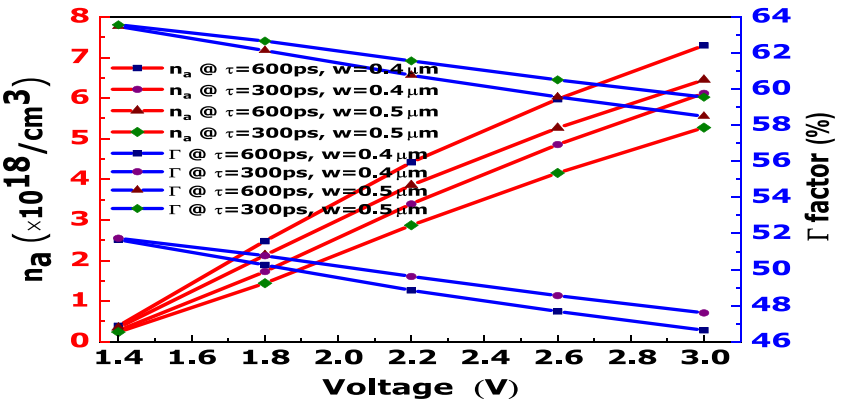


Fig. 3. Average carrier concentration and optical confinement factor versus the applied bias voltage for different intrinsic-region width (w) and carrier lifetimes (τ) in the i-GaAs region.

the carrier lifetime, whereas increasing the width of intrinsic region reduces the carrier concentration. An increase in lifetime results in higher n_a (leading to higher refractive index change) with respect to increase in voltage, however, this may not be always favorable for attaining higher CR at the output ports, because a higher refractive index change could now cause the signal to be resonant at both zero and non-zero biased condition. In addition, the optical confinement increases when the intrinsic region width increases, however, it decreases with increasing the applied bias. This is because, at higher bias, a larger average carrier density is present in the core region, which greatly reduces the refractive index of the core, thereby reducing the refractive index difference between the core and the cladding, and thereby reducing the optical mode confinement. Fig. 4 shows $\Delta\alpha_{BF+BGS}$ and Δn_{total} versus wavelength for an intrinsic region width of 0.4 μm and τ = 300 ps and different bias voltages applied to the ring configuration. The change in absorption and refractive index increases with increasing the applied bias voltage. FCA typically contributes to the change in refractive index at higher wavelengths. As shown in Fig. 4(b), at 1559.7 nm, a refractive index change of 0.0266 is attained when the bias changes from 1.4 V to 3.0 V.

The modified output intensities at the MRR output ports can be obtained by substituting Δn in the transfer functions of proposed MRR logic gate configuration (Eq. (4)). The bias voltage applied to Ring3 is assumed to be the same bias applied on Ring1. Fig. 5(a) and (b) show the normalized output optical intensity at the T port and D port of MRR3, versus wavelength at different bias voltages applied to MRRs 1 and 2, when the proposed configuration is operated as an OR/NAND gate and AND/NOR gate, respectively. As shown in Fig. 5(a), for λ = 1559.7 nm and κ = 0.55, a high output optical signal intensity (logic 1) is obtained at the T port of MRR3, when either Ring1 or Ring2 are biased, leading to OR gate realization. For λ = 1549.6 nm and κ = 0.55, a low output optical signal intensity (logic 0) is obtained at the T port of MRR3, only when both Ring1 and Ring2 are biased at 2.6 V, leading to NAND gate realization. Referring to Fig. 5(b), for λ = 1559.7 nm and κ = 0.55, the D port of MRR3 collects a high-intensity optical signal only when both rings are biased, thus realizing an AND gate. For λ = 1549.6 nm and κ = 0.55, the D port of MRR3 collects a high-intensity optical signal only when both the rings are not biased at 2.6 V, thus realizing a NOR gate. The truth table for the proposed carrier injection based OR/AND triple MRR logic gate configuration is depicted in Table 1 which demonstrates the OR and AND logic operations. The threshold value is chosen as 0.3 times the input light intensity. Hence, if output intensity is above 0.3 times the input light intensity it will be considered as logic 1, otherwise it is considered as logic 0.

CR is defined as the intensity difference between the high and low states in the gate. Therefore, CR1, CR2 and CR3 for the OR gate are the intensity difference between the outputs for (Input1, Input2) = (1, 1) and (Input1, Input2) = (0, 0), (Input1, Input2) = (1, 1) and (Input1, Input2) = (0, 1), (Input1, Input2) = (1, 1) and (Input1, Input2) = (1, 0), respectively. Note that the coupling coefficient between the ring and bus waveguides needs to be optimized in order to obtain a high CR for given intrinsic region width, applied bias voltage and relaxation time. Fig. 6(a)–(d) show the normalized output optical intensities at the T port (OR gate) of MRR3 versus the coupling coefficient for different CR values and for a launched wavelength of 1559.7 nm, τ = 300 ps, w = 0.4μm for different logic 1 bias conditions of V = 1.4 V, 1.8 V, 2.2 V and 2.6 V, respectively. see Table 2.

As shown in Fig. 6, the optimum coupling coefficient required to maximize the CRs for the OR gate at the T port depends on the applied bias voltage (representing logic 1 condition). The maximum CR attained for the OR gate is at an applied bias voltage of 2.6 V when κ = 0.586. Fig. 7(a)–(d) show the normalized output optical signal intensities at the D port (AND gate) of MRR3 versus coupling coeffi-

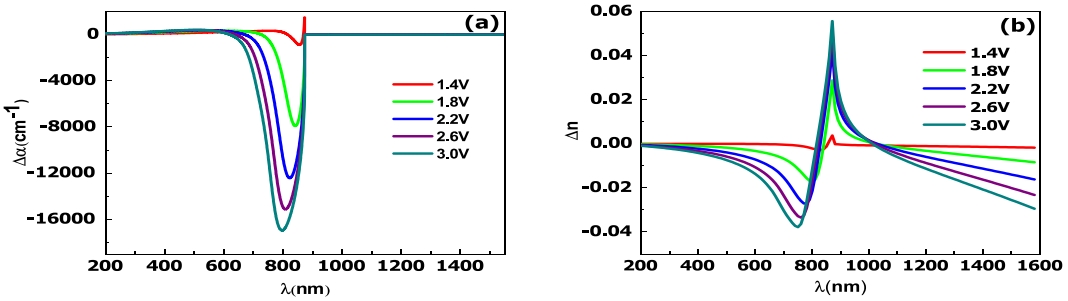


Fig. 4. (a) BF and BGS effect based absorption change spectrum for a w = 0.4 μm and τ = 300 ps at different appied bias. (b) Total refractive index change ($\Delta n_{BF+BGS+FCA}$) as a function of wavelength for a w = 0.4 μm and τ = 300 ps at different appied bias.

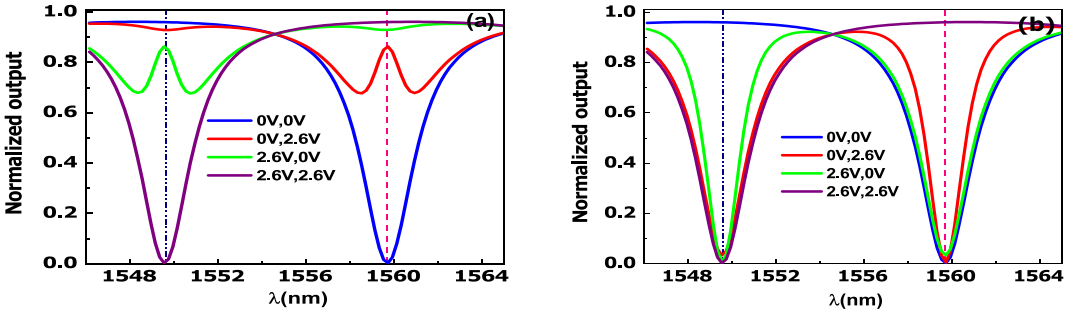


Fig. 5. (a) Normalized T port intensity spectrum with OR gate realized at 1559.7 nm (vertical pink dashed line) and NAND gate realized at 1549.6 nm (vertical blue dot dashed line). (b) Normalized D port intensity spectrum with AND gate realized at 1559.7 nm (vertical pink dashed line) and NOR gate realized at 1549.6 nm (vertical blue dot dashed line). Logic 1 operand is represented as 2.6 V on the p-i-n structured MRR while 0 V represents logic 0 operand condition.

Table 1
Truth table for OR and AND gate at applied bias as 0 V and 2.6 V and $\kappa = 0.55$, $\lambda = 1559.7$ nm, w = 0.4 μm and $\tau = 300$ ps. The value in the square brackets represent the optical power at the respective ports.

Input1 (Ring1)	Input2 (Ring2)	OR (T port)	AND (D port)
0 (0 V)	0 (0 V)	0 [0.0025]	0 [0.0025]
0 (0 V)	1 (2.6 V)	1 [0.8631]	0 [0.0052]
1 (2.6 V)	0 (0 V)	1 [0.9292]	0 [0.0339]
1 (2.6 V)	1 (2.6 V)	1 [0.9603]	1 [0.9603]

cient for different CR values and for a launched wavelength of 1559.7 nm, $\tau = 300$ ps, w = 0.4μm, for different logic 1 bias conditions of V = 1.4 V, 1.8 V, 2.2 V and 2.6 V, respectively. As shown in Fig. 7, the optimum coupling coefficient required to maximize the CRs for the AND gate at the D port depends on the applied bias voltage (representing logic 1 condition). For an applied bias voltage of 2.6 V when $\kappa = 0.431$, a maximum CR is attained for AND gate for an intrinsic width of 0.4 μm and operating wavelength as $\lambda_R = 1559.7$ nm. The maximum CRs (CR1, CR2 and CR3) attained at an applied bias of 1.4 V is less compared to the maximum CRs obtained at an applied bias of 2.6 V with the optimized coupling coefficient values. Typically, the optimum κ for maximizing the CR depends on applied bias voltage, intrinsic region width and carrier recombination lifetime. It is noteworthy that the CR attained at the optimum coupling coefficient increases with the applied bias voltage. From Figs. 6 and 7, it is interesting to note that for the designed logic OR and AND gate configurations, the tolerance of the output intensity to the variation in κ increases as the applied bias increases. The intensity-versus-κ curves at higher bias voltages has almost a constant value, enabling the selection of a single optimum coupling coefficient (instead of two different optimum coupling coefficients) for both the OR and AND gate. The "flat-top" nature of intensity-versus-κ during higher applied bias results from the large change in the refractive index due to which the wavelength that would have been resonant with the ring under applied bias shifts far away from the wavelength that was resonant to the ring with no applied bias.

Imperfections that occur during the fabrication of MRR can cause deviations in the gap between ring and bus waveguides, resulting in significant difference in κ value obtained and κ value designed. This, however, will not be a major issue in the proposed MRR OR/AND logic gate configuration when operated at higher bias due to an almost con-

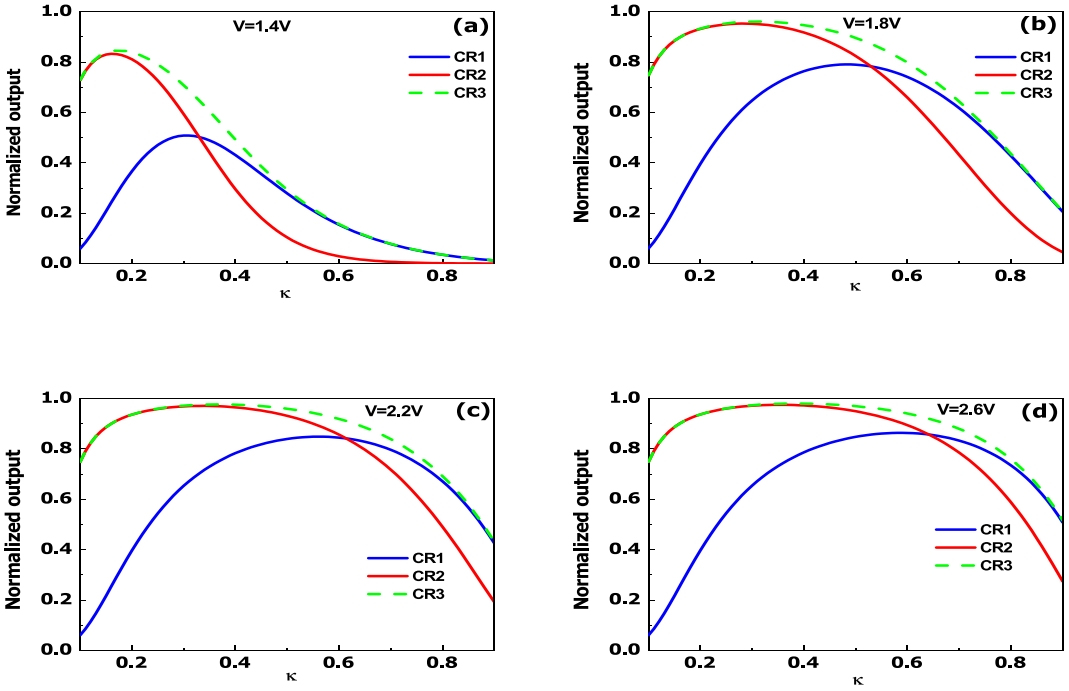


Fig. 6. T port CR dependency on coupling coefficient of triple CI MRR based OR gate with Ring3 fed with Input1 with an applied bias of (a) 1.4 V (b) 1.8 V (c) 2.2 V (d) 2.6 V as input logic 1 condition and operating wavelength $\lambda_R = 1559.7$ nm.

Table 2
Truth table for NAND and NOR gate at applied bias as 0 V and 2.6 V, $\kappa = 0.55$ for NAND gate and $\kappa = 0.44$ for NOR gate, $\lambda = 1549.6$ nm, w = 0.4 μm and $\tau = 300$ ps. The value in the square brackets represent the optical power at the respective ports.

Input1 (Ring1)	Input2 (Ring2)	NAND (T port)	NOR (D port)
0 (0 V)	0 (0 V)	1 [0.9573]	1 [0.9825]
0 (0 V)	1 (2.6 V)	1 [0.9237]	0 [0.0217]
1 (2.6 V)	0 (0 V)	1 [0.8631]	0 [0.0168]
1 (2.6 V)	1 (2.6 V)	0 [0.0032]	0 [0.0078]

stant appreciable CRs attained over a wider range of κ values. Also, the flat-topped intensity-versus-κ curves at higher bias voltages facilitate to design the MRRs at relatively smaller κ values instead of higher κ values, as higher κ values demand for smaller gap between ring and bus waveguides with advanced device fabrication process. This considerably relaxes the fabrication complexity. However, the fabrication tolerance admissible at higher bias will be at the expense of higher power dissipation, which should be minimized for logic gates.

The same triple parallel MRR logic gate configuration with CI modulation mechanism can also work as NAND and NOR gates using the T port and D port of MRR3, when the launched wavelength satisfies resonance condition at an applied bias voltage for the MRRs. Each MRR satisfies the resonance condition at an operating wavelength, $\lambda = 1549.6$ nm, with an applied bias of 2.6 V, leading to the realization of NAND and NOR gates using the T port and D port (refer Fig. 5(a) and (b)). Note that, a change in absorption of the rings through electroabsorption mechanism can also result in a change in the refractive index of the rings and can be used to realize logic gates using MRR configurations. However, in realization of logic gates using electroabsorption (Fayza and Sooraj, 2020), it was observed that the electroabsorption-based modulation mechanism fails to realize NAND/NOR gates using triple parallel MRR configuration. This is because, at an applied bias, absorption of the rings increase due to the reduction of the Q factor during the propagation of resonant wavelength through the rings when a non-zero bias voltage is applied to the rings. On the other hand, usage of CI mechanism in triple parallel MRR configuration successfully overcomes this drawback and helps in realizing both OR and AND or NAND and NOR logic gates at the output ports, thus making the proposed device more versatile. The coupling dependency to yield maximum CR in NAND/NOR gate is similar to the calculated optimum κ for the OR/AND gate configuration. The truth table for the NAND/NOR gate for the launched wavelength $\lambda = 1549.6$ nm and at their optimized coupling condition is shown in Table 2 for an applied bias voltage of 2.6 V (logic 1 operand condition), and demonstrates the NAND and NOR logic operations.

We have proposed a novel triple parallel MRR configuration for realizing OR and AND gate or NAND and NOR gate at the output ports. In Kumar et al. (2014), the authors have reported XOR/XNOR and AND gate which are realized using Mach-Zehnder interferometer, which is bulky in nature and hence not suitable for large scale integration. In Tian et al. (2011), the authors have reported AND/NAND and OR/NOR logic gates realized using MRR, however the switching is enabled using thermo-optic effect. This switching mechanism faces poor stability and reliability. Our proposed device on the other hand relies on carrier injection based switching mechanism which requires only a low operating voltage and making it more reliable, thermally stable and consumes only lesser footprint. Furthermore, the proposed triple parallel configuration provides better functionality as non-conjugate gates (OR and AND or NAND and NOR gates) are obtained simultaneously in the MRR output ports instead of obtaining conjugate gates (like AND and NAND gates, OR and NOR gates) simultaneously at the output ports as in conventional configurations introduced in the literature. CI based triple parallel MRR logic gate realizes NAND and NOR gate at the output port, which is not possible to realize in electroabsorption based triple parallel MRR (Fayza and Sooraj, 2020) due to higher absorption effect as the operating wavelength is needed to be selected near bandgap edge of the core material.

4. Conclusion

A novel logic gate configuration based on the use of triple parallel MRRs in conjunction with carrier injection is proposed. Results shows

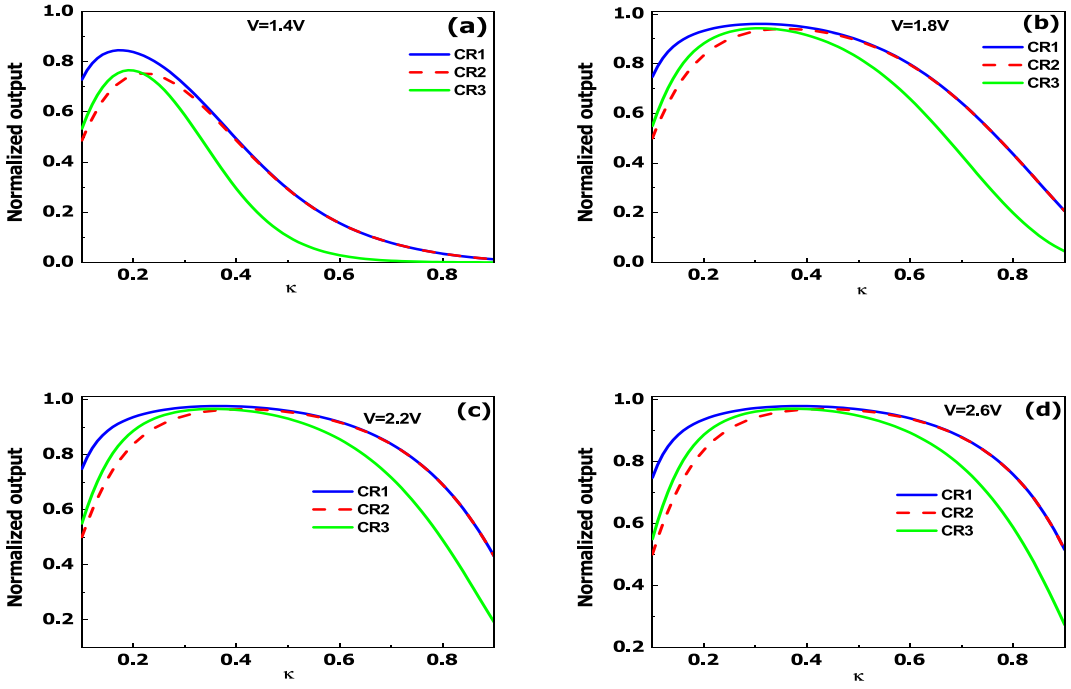


Fig. 7. D port CR dependency on coupling coefficient for triple CI MRR based AND gate with Ring3 fed with Input1 with an applied bias of (a) 1.4 V (b) 1.8 V (c) 2.2 V (d) 2.6 V as logic 1 condition and operating wavelength $\lambda_R = 1559.7$ nm.

that by using the bias voltage applied to MRRs as the operands and the output intensities at the T port and D port as the logic gate outputs, the proposed logic gate configuration simultaneously realize OR and AND gate or NAND and NOR gate operations depending on the wavelength of the optical signal launched into the input port and add port of the logic gate configuration. The dependency of the contrast ratio of the logic gate outputs on the applied bias voltage, relaxation time, intrinsic region thickness and coupling coefficient between the ring and bus waveguide have been simulated, and the results have shown that, by increasing the bias voltage, a higher fabrication tolerance (high contrast ratio obtained over a wider value range of coupling coefficient) is achieved. The fabrication tolerance attained at higher applied bias enables the selection of a single optimum coupling coefficient (instead of two different optimum coupling coefficients) to simultaneously realize both the OR and AND gates or NAND and NOR gates.

References

Abdalla, S. et al, 2004. Carrier injection-based digital optical switch with reconfigurable output waveguide arms. IEEE Photon. Technol. Lett. 16, 1038 1040.

Bennet, B., Soref, R., 1987. Electrorefraction and electroabsorption in InP, GaAs, GaSb, InAs, and InSb. IEEE J. Quant. Electr. 23, 2159–2166.

Bennet, B.R., Soref, R.A., Del Alamo, J.A., 1990. Carrier-induced change in refractive index of InP, GaAs and InGaAsP. IEEE J. Quant. Electr. 26, 113–122.

Dominik, G.R., 2007. Ring resonators: theory and modeling. Integrated Ring Resonators, Springer 127, 3–40.

Fayza, K.A., Sooraj, R., et al, 2020. Advanced realization and characterization of directed optical logic gates using electroabsorptive quantum-well-based micro ring resonator. Optik 221, 164426.

Hardy, J., Shamir, J., 2007. Optics inspired logic architecture. Opt. Exp. 15, 150–165.

Ishida, K., Nakamura, H., Matsumura, H., Kadoi, T., Inoue, H., 1987. InGaAsP/InP optical switches using carrier induced refractive index change. Appl. Phys. Lett. 50, 141–142.

Ito, Fumihiko, Tanifuji, Tadatoshi, 1988. Carrier-injection-type optical switch in GaAs with a 1.06-1.55 μm wavelength range. Appl. Phys. Lett. 54, 134.

Kumar, Ajay, Kumar, Santosh, Raghuwanshi, Sanjeev Kumar, 2014. Implementation of XOR/XNOR and AND logic gates by using Mach-Zehnder interferometers. Optik 125, 5764–5767.

Piprek, Joachim, 2003. Semiconductor Optoelectronic Devices, Academic Press, pp. 49–82..

Ravindran, Sooraj, Datta, Arnab, Alameh, Kamal, Lee, Yong Tak, 2012. GaAs based long-wavelength microring resonator optical switches utilising bias assisted carrier-injection induced refractive index change. Opt. Exp. 26, 15610–15627.

Shamir, J., Caulfield, H., Micelli, W., Seymour, R., 1986. Optical computing and the Fredkin gates. Appl. Opt. 25, 1604–1607..

Soref, R., 2011. Reconfigurable integrated optoelectronics. Adv. OptoElectr. 2011, 15.

Tian, Yonghui, Zhang, Lei, Ji, Ruiqiang, et al, 2011. Proof of concept of directed OR/NOR and AND/NAND logic circuit consisting of two parallel microring resonators. Opt. Lett. 36, 1650–1652.

Yariv, A., 2000. Universal relations for coupling of optical power between microresonators and dielectric waveguides. Electr. Lett. 36, 321–322.

Compact and efficient polarization splitter based on dual core microstructured fiber for THz photonics

Madhusudan Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, madhusudandas55@hotmail.com*

Ipsita Samal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ipsitasamal55@gmail.com*

Prangya Paramita Padhi, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, prangya.p.padhi@gmail.com*

Prasanna Kumar Chhotaray, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pkchhotaray85@gmail.com*

ARTICLE INFO

Keywords:
THz Photonics
Microstructured fiber
Polarization splitter
Device length
Extinction ratio

ABSTRACT

We consider a dual core polyethyelene microstructured fiber with teflon and air inclusions in the core region for achieving compact and efficient polarization splitting at THz frequencies. The teflon inclusions form slit-like structure that helps in shifting the degeneracy point to lower frequencies. The y-polarized light couples from one core to the other within a very small device length of a few cm. Modal fields, effective indices, device length, extinction ratio and different losses have been calculated, which provide a handle to reduce the polarization splitting length at frequencies below and above ~ 0.7THz. Extinction ratios are found to be significantly high in a broad frequency range, e.g. a value of ~ 70 dB for x-polarization and ~ 45 dB for the y-polarization with about 2.1 cm device length at 1 THz. This device length is more than 30 times shorter than previously reported values. However, it attains a maximum device length of 21.6 cm at ~ 0.7 THz, which is also shorter than the corresponding earlier reported value.

1. Introduction

Polarization splitting with short device length is desirable at terahertz (THz) frequencies. There are various areas in THz science and technology, e.g., communication, sensing, security, imaging, spectroscopy, medical diagnosis, etc. (Koch et al., 2014; Hidayat et al., 2008; Burrow et al., 2017; Morris et al., 2012), which demand efficient polarization splitting devices of compact size. Traditionally, photonic crystals have been the choice of the researchers in the field of design and development of polarization splitters due to the flexibility in the design aspect of theirs (Sesay et al., 2013; Park et al., 2010). However, it must also be noted that all those devices have a limitation in terms of the operational bandwidth (Zhong and Sheng, 2015; Mo and Li, 2016) for reliable polarization splitting and extensive research work has been done for improving upon that. As such, efficient polarization splitters have been demonstrated out of photonic crystal fibers (PCFs) which offer a flexibility to achieve desired mode propagation characteristics by design engineering and optimization of dispersion, effective beam area, etc. However, most of them have been studied at the visible and infrared frequencies. By choosing appropriate materials, the above concepts can also be passed over to devices suitable at THz frequencies.

Dual and triple core PCFs have been suggested in the literature for achieving efficient polarization splitting by spatial separation at specific frequencies from visible to mid infrared (MIR), where, during propagation within the splitting length, there is not much change in the beam shapes. In the dual core configuration, either one polarization couples between them or coupling for both the polarizations can take place but to have different coupling lengths for the two orthogonal polarizations (Jiang et al., 2014). On the other hand, in case of triple core PCFs, two cores can be kept nearly same for their configuration and material while the third core has different coupling lengths for the two orthogonal polarizations (Saitoh et al., 2004). Gold filled dual core PCFs were suggested for compact and ultra-broadband polarization splitting at optical frequencies (Khaleque and Hattori, 2015). Again, in the optical range, pure silica based PCFs were found suitable. These contain dual cores made of elliptical air holes and a glycerol filled central air hole between them (Wang et al., 2018). Triple core PCFs combining a bimetal coating and liquid filling also perform well in the optical range (Liu et al., 2019). However, the complicated structure and large number of materials involved in the later make the whole system quite complicated. BK7 glass based dual core PCF containing hexagonal pattern of liquid crystal filled regions around one core, was shown as a polarization splitter at 1.3 μm wavelength (Younis et al., 2018). However, the limited flexibility of the liquid crystals can allow only an extremely narrow bandwidth of just a few nm. At the MIR, polarization splitters were designed out of silicon

based dual core PCFs (Qu et al., 2020) providing a large bandwidth (~1μm) for operation.

For THz waves, polymer based dual core microstructured fibers have been studied in the literature for achieving polarization splitting. For example, in an elliptically shaped dual core fiber, the TE and TM modes get separated out due to a difference in the corresponding frequency-dependent coupling lengths in the two cores (Chen et al., 2016). The main limitations in this proposal were due to the long device length and low-value of extinction ratios at all the THz frequencies considered. Also, ellipticity in the cores cannot be maintained along the full length of the fiber required for efficient polarization splitting. In another study, using air holes, which form nearly orthogonal patterns in the two cores of an index guided photonic crystal fiber (IGPCF), efficient polarization splitting was achieved in ~ 0.4 0.65THz_

window (Li et al., 2013). However, in this case the device length increases exponentially from < 2 cm at 0.4 THz to greater than 20 cm at 0.65THz, thereby, limiting its potential use at the com-monly used frequencies near 1THz. Recently, we have proposed a polyethylene based dual core refractive index gradient IGPCF by which efficient polarization splitting is possible with device length of < 60 cm at 1THz (Kumar et al., 2020). Still, polarization splitting length of just a few cm is desirable in a broad range of frequencies around 1THz.

In the current study, we have considered polyethylene based IGPCF containing specifically arranged air columns in the clad region and a dual core of air and teflon inclusions at the center. With proper gradient arrangement of the air and teflon inclusions and their dimensions in the dual core region of such a IGPCF, we show that a short device length of just 2.1 cm can be obtained for efficient polarization splitting operation at 1THz. We have checked the viability of operation at wide range of frequencies. It is found that in the 0.4-1THz range, the polarization splitting length varies from ~ 2 to ~ 21 cm. The device length is maximum at ~ 0.65THz but quickly reduces on either side of this frequency. Consequently, splitting is found to be better with a short device length at both the high and low frequencies, thereby making it highly suitable for commonly used THz sources at ~ 1THz. Also, the polarization splitter has a high extinction ratio in the whole fre-quency range of ~ 0.4-1THz band.

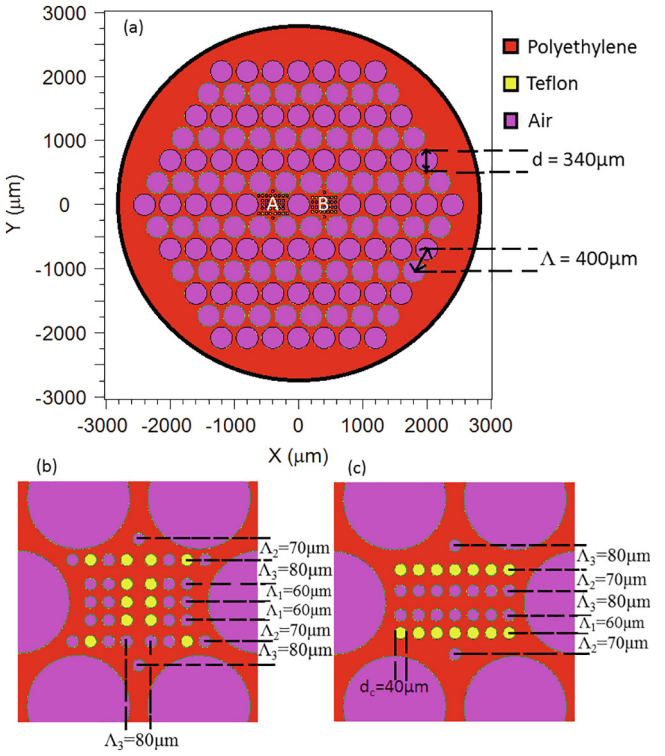


Fig. 1. (a) Transverse cross-sectional view of the polyethylene based dual core microstructured fiber (Fiber-AB). The cores have been identified as A and B. Magnified views of (b) the core A, and (c) the core B regions of the fiber containing air and teflon inclusions. Different pitch values (Λ) and diameters (d) of the teflon and air columns are indicated. All remaining pitches in x-direction are same, i.e., $\Lambda_1 = 60$ μm.

2. Fiber structure and numerical analysis

Polyethylene is highly suitable for device fabrication for THz frequencies (Piesiewicz et al., 2007). The actual fabrication of the polymer based microstructured fibers can be carried out using the popular methods such as stack and draw method, sol–gel method, etc. (Pysz et al., 2016; Chow et al., 2012; Islam et al., 2020). Fig. 1 (a) shows cross-sectional view (xy-plane) in the clad and the dual core regions of the polyethylene based IGPCF considered in our current theoretical study. The design consists of air columns arranged in a hexagonal ring structure in the clad region, and rectangular arrays of air and teflon inclusions in the two cores. In the clad, pitch $\Lambda = 400$ μm and the air hole diameter $d = 340$ μm have been used. In the dual core region, a rather interesting arrangement of the air and teflon inclusions has been used and that differs in the two cores also (see Fig. 1 (b) and 1(c)). The dimensions in μm along the reference x- and y-directions have been indicated in the figure. In the two cores, the pitches are $\Lambda_1 = 60$ μm, $\Lambda_2 = 70$ μm, $\Lambda_3 = 80$ μm and diameter of air holes is $d_c = 40$ μm. These values of structural parameters in the two cores and their configurations with air and teflon inclusions are such that effective indices (n_{eff}) of y-polarized modes in the two cores become nearly same in an extended frequency range. Arrangement of teflon inclusions in the two cores form slit-like pattern. Below, we have discussed results for three cases: Fiber-A, i.e., the fiber containing only core A at its usual position while an air hole at the position of core B, Fiber-B, i.e., the fiber containing only core B at its usual position while an air hole at the position of core A, and third one (Fiber-AB) where both the cores A and B are present.

The two core structures have been configured in such a way that a large difference between the n_{eff} values of x-polarized modes in the two single core fibers, i.e., Fiber-A and Fiber-B, is achieved. The corresponding difference for the y-polarized modes in the two individual fibers should be minimum to ensure good coupling along the length of the fibers (Piesiewicz et al., 2007). Proper arrangement of air holes and high-index teflon rods in the two cores provides high gradient in the effective index. The slit-like patterns formed by the teflon rods are orthogonal in the two cores, where, different spacing between any two arrays of teflon rods is dictated by different pitch values as indicated in Fig. 1. The slit-like patterns of teflon rods in the two cores and the arrangement of air holes (Fig. 1) provide a handle for shifting the frequency at which n_{eff} curves of the y-polarized modes in Fiber-A and Fiber-B intersect with each other. With proper positioning of the teflon rods, the crossing frequency is pushed below 0.65THz, while difference between n_{eff} values for the y-polarized modes in the Fiber-AB is increased at higher frequencies leading to a shorter polarization splitting length. It is to be noticed that two innermost teflon rods in the core A are missing in the last row, the impact of which, will be a slight decrease in splitting length at higher frequencies than what can be achieved when they are included. Actually the splitting length vs. frequency curve in the high frequency region gets bifurcated from it being little higher when these two rods are present.

Full vector beam propagation method (Pedrola, 2016) was used to calculate the frequency-dependent n_{eff} values. Briefly, the method involves generating equations via tridiagonal matrices from the vector Helmholtz equations, which are then used to find modal fields at each step (x, y, z) during the propagation along the length (z-direction) in a

continuous manner. For each propagation step at (x, y, z), the block size was taken to be 4 μm × 4 μm × 5 μm. Transparent boundary con-dition at the interface was applied to suppress reflections at the inter-faces. While calculating the modal profiles and the refractive indices, a tolerance of 10^{-7} in n_{eff} was considered. The propagation loss for the fundamental mode was ~ 3.5 dB/m at 1THz. We may point out that the propagation loss for next four higher order modes was ~ 20 dB/m, which is significantly high. Hence, they are not included in our analysis. BEAMPROP and FEMSIM modules in the commercially available software, RSOFT, were used for determining the modal fields and the effective indices for different polarizations in the entire frequency spectrum of ~ 0.4-1THz. These results were further used in MATLAB to perform calculations for achieving the spectral splitting length, extinction ratios, and losses due to propagation and bending. For validation of the results, the finite element method was also applied for the design of the structure. Here, we decompose the entire structure into a number of meshes, on the nodes of which, we consider values of the electric field. Using these values of the electric field and order of the mesh, we get the values of the electric field for one element by interpolation. Subsequently, the vector Maxwell equation is applied in this interpolation function to form elemental equation. Combining all the equations from each element and applying boundary condition, the global system of equations is formed. By solving this system of equations, we were able to find the modal field profiles, from which the values of the effective indices are obtained with the help of stan-dard algorithms (Rahman, 2013). A number of structure iterations were applied so as to achieve the desired output characteristics. The values of the effective indices were very close to each other in the two methods mentioned above.

3. Results and discussion

The frequency dependence of n_{eff} for the x-polarized and y-polarized modes of Fiber-A and Fiber-B, termed as x(A), y(A) and x (B), y(B), respectively, have been presented in Fig. 2(a) for a large frequency range (0.4-1THz). Due to large difference between n_{eff} values of x(A) and x(B), the coupling between them is completely avoided in the entire frequency window. However, the difference in n_{eff} values for the y(A) and y(B) modes is comparatively much smaller, hence, coupling between them is possible. In fact, the n_{eff} curves for these two modes cross at a frequency of ~ 0.65 THz, the degenerate point. The crossing frequency has shifted to a lower value that what was reported earlier for a similar microstructured polyethylene fiber but without the teflon inclusions and the slit-like pattern in their arrangement in the cores (Kumar et al., 2020). The dual core IGPCF (Fig. 1(a)), i.e., Fiber-AB, will have two fundamental modes for the x- and y-polarizations. Correspondingly, they have been named as x-First, y-First and x-Second, y-Second in Fig. 2(b), where we have shown their n_{eff} vs frequency behavior. Clearly, the x-First and x-Second modes are quite apart from each other in the entire frequency window, while, the y-First and y-Second modes have their n_{eff} very close to each other around the frequency of ~ 0.68 THz. This has been highlighted in the inset of Fig. 2(b). This allows effective power coupling between y-First and y-Second.

The x- and y-polarized modal electric field distributions in Fiber-AB, as calculated using the full vector beam propagation method, are shown in Fig. 3. These distributions have been drawn at three representative frequencies around the degenerate point, i.e., at 0.6THz in Fig. 3(a), 0.7THz in Fig. 3(b) and 0.8THz in Fig. 3(c). The corresponding values of the effective indices at those frequencies have also been indicated in the figure. The figure clearly indicates that the modal field confinement is more for higher frequencies. Each x-polarized mode is confined in one or the other core as expected from their effective indices being very different from each other. On the other hand, due to proximity of the effective indices, the y-polarized modes are confined in both the cores of Fiber-AB. As we come close to minimum difference frequency (~0.68THz), modal profiles for y-polarization look same (Fig. 3(b)).

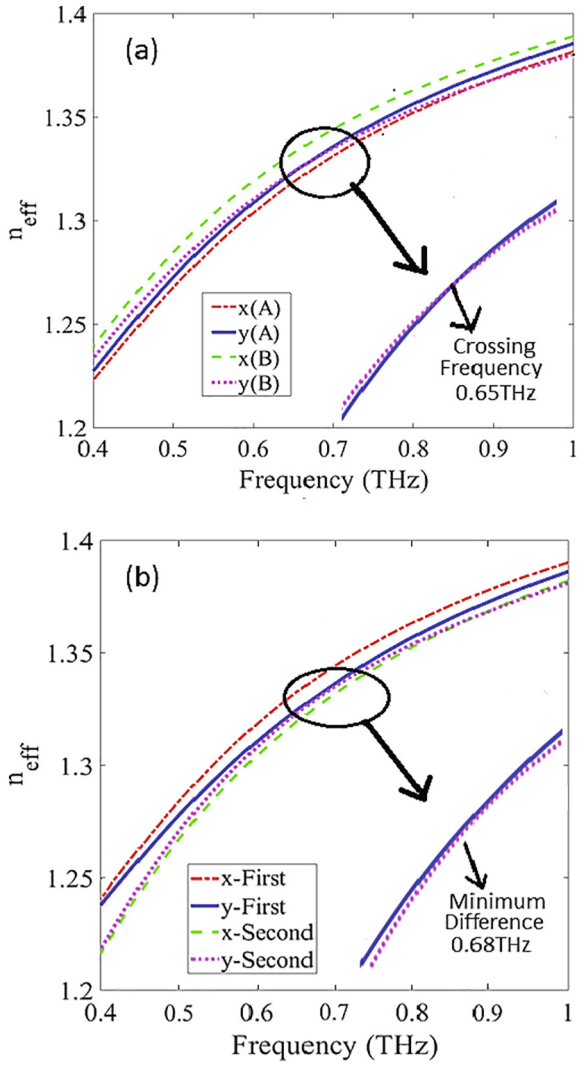


Fig. 2. Effective indices, n_{eff} for the x- and y-polarizations of (a) Fiber-A and Fiber-B, and (b) the dual core fiber, Fiber-AB. Inset in (a) is the magnified view around the crossing frequency at ~ 0.65 THz. Inset in (b) is the magnified view to clearly see the difference among the curves in the frequency.

It is important to mention that the teflon inclusions in the two cores have not perturbed the orthogonality of x- and y-polarized modes. For determining the coupling and hence overlapping of the modes, input power at both the x- and y-polarizations is injected at core A of the Fiber-AB. The phase matching between the two individual core modes, i.e., Fiber-A and Fiber-B, and the overlap between the input powers with these modes in Fiber-AB (Fig. 3) describe the distribution of power for two polarizations in two cores. For input power in x-polarization at core A, power propagates to the output via x-Second mode without much coupling due to higher overlap in x-Second mode. This weak coupling for x-polarization is due to the phase mismatch between the two individual core modes. On the other hand, for input power in y-polarization at core A, power flows through both y-First and y-Second modes due to phase matching between y-polarized modes of the two cores in Fiber-AB.

Due to the special configuration in the cores of the polyethylene based dual core IGPCF in our study here (Fiber-AB), it acts like a y-polarization coupler and simultaneously as an x- and y-polarization

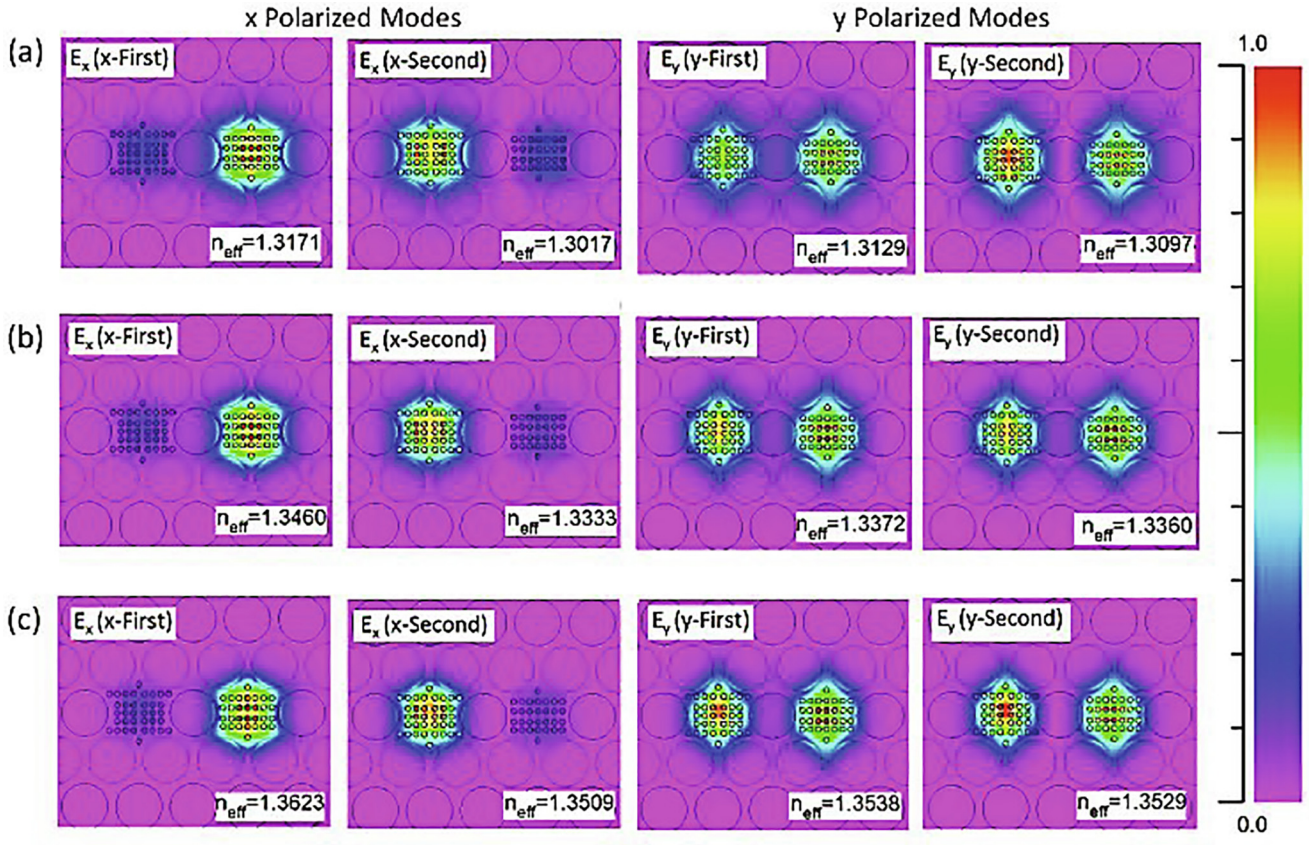


Fig. 3. First and second modal electric field distributions in the dual core microstructured fiber (Fiber-AB) for the x- and y-polarizations at a frequency of (a) 0.6 THz, (b) 0.7 THz, and (c) 0.8 THz. The color bar represents the variation in the electric field strength.

splitter. Coupling of power between core A and core B for the y-polarized modes of Fiber-AB takes place at the splitting length L_s (also known as coupling length, L_c in the literature) that is given by the relation,

$$L_s = \frac{\lambda}{2(n_1 - n_2)} \quad (1)$$

Here, λ is the wavelength; n_1 and n_2 represent the n_{eff} values of the y-First and y-Second modes, respectively of the Fiber-AB. The behavior of n_{eff} in Fig. 2(b) ensured by the slit-like pattern of the teflon inclusions in the dual core configuration, the polarization splitting length for Fiber-AB gets reduced at frequencies above 0.68 THz as shown in Fig. 4. This is contrary to the observation of an increasing trend in the splitting length vs. frequency (see inset of Fig. 4) that was made earlier for a simple dual core configuration of IGPCF without having rod inclusions of high refractive index material (Kumar et al., 2020). Below 0.68 THz, the increasing trend in L_s with the frequency is due to tighter mode confinement at increasing frequencies in the individual cores. Hence, the poor coupling demands higher value of L_s for complete power transfer. Furthermore, above the minimum difference frequency of 0.68THz, the difference in n_{eff} for the y-polarized modes of Fiber-AB keeps increasing with the frequency. Therefore, at frequencies above 0.68THz, the splitting length in Fiber-AB decreases with the increasing frequency per the Eq. (1). For a comparison, the results for L_s for similar dual core IGPCFs considered in Refs. (Li et al., 2013; Kumar et al., 2020), have been drawn in the inset of Fig. 4. Clearly, in comparison to the previous reports, the design of Fiber-AB provides a much smaller device length for efficient polarization splitting in the entire frequency range considered. For example, at the frequency of 0.85THz, which is the maximum frequency of operation in Ref. (Kumar et al., 2020), our device length is more than 6 times smaller than the prior. Similarly, at 1 THz, we achieve a device length of just

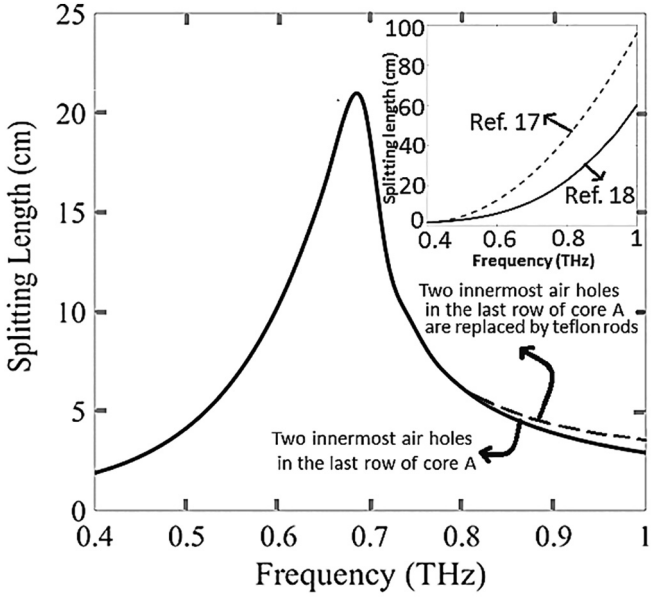


Fig. 4. Spectral variation of the polarization splitting device length (L_s) (solid curve). The superimposed dashed curve (bifurcated at higher frequencies) is for the structure when two innermost air holes in the last row of core A (see Fig. 1) are replaced with teflon rods. Inset: A comparison between the current results and those from Refs. (Li et al., 2013; Kumar et al., 2020).

2.1 cm, which is about 30 times smaller than previously reported value (Li et al., 2013). This was possible due a control achieved by the inclusion of teflon rods in the two cores of our design. The inclusion of two

innermost air holes instead of teflon rods in the last row in core A (Fig. 1(a)) provides a further reduction in the device length at the high frequencies. With and without those teflon rods, there is a small difference in the corresponding n_{eff} values of y-polarized modes of Fiber-AB. This difference in the L_s in the two cases is shown by the bifurcated curve in Fig. 4 at frequencies between 0.8THz and 1THz.

The other important characteristic of the polarization splitters is extinction ratio (ER) between the leaked power of the unwanted polarized mode and the power in the desired polarized mode. For Fiber-AB, we define the extinction ratios for x- and y-polarized modes as ER_x and ER_y, respectively. They are given by the following relations,

$$ER\_x = 10 \times log_{10} \frac{P_{out}(y(A))}{P_{out}(x(A))} \quad (2)$$

$$ER\_y = 10 \times log_{10} \frac{P_{out}(x(B))}{P_{out}(y(B))} \quad (3)$$

Here, $P_{out}(i(J))$ is the output power at the J^{th} core of Fiber-AB (either A or B in Fig. 1) for the i^{th} polarized mode (i is either x- or y-polarization). The extinction ratios of a polarization splitting device are sensitive to and should be checked with the deviation in the splitting length. The spectral variation of the extinction ratios for x-polarization is illustrated in Fig. 5(a) at different percentage deviation (ΔL) in the device length. We find that for 1% deviation in the splitting length, ER_x is very high of ~ 70 dB, but gets reduced to just 20 dB at 5% deviation in the splitting length. Therefore, ER_x is highly sensitive to the device length in our case. This can be understood because of the fact that due to coupling there will be oscillatory nature of y polarized power in core A, which is sensitive to length. The spectral variation of the extinction ratios for y-polarization is illustrated in Fig. 5(b) at different percentage deviation (ΔL) in the device length. On the other hand, ER_y is almost insensitive to the deviation in the splitting length. This also can be understood from the fact that $P_{out}(x(B))$ remains very small, which is also more dependent on frequency. The extinction ratios for both splitted polarizations are very high at higher frequencies around 1THz, compared to the previously reported structure in Ref. (Kumar et al., 2020).

As usual, the confinement loss in Fiber-AB decreases with the increasing frequency as shown in Fig. 6(a). We have kept sufficient number of air holes in the clad to control the confinement loss to the minimum value. The spectral response of the total propagation loss (including effective material loss and the confinement loss) is also shown in Fig. 6(a). As desired, the total propagation loss in Fiber-AB is much lower than those in the previously reported dual core IGPCFs. (Li et al., 2013). We may mention that for determining the total propagation loss, imaginary part of the n_{eff} was used, which itself was calculated using full vector beam propagation method after incorporating

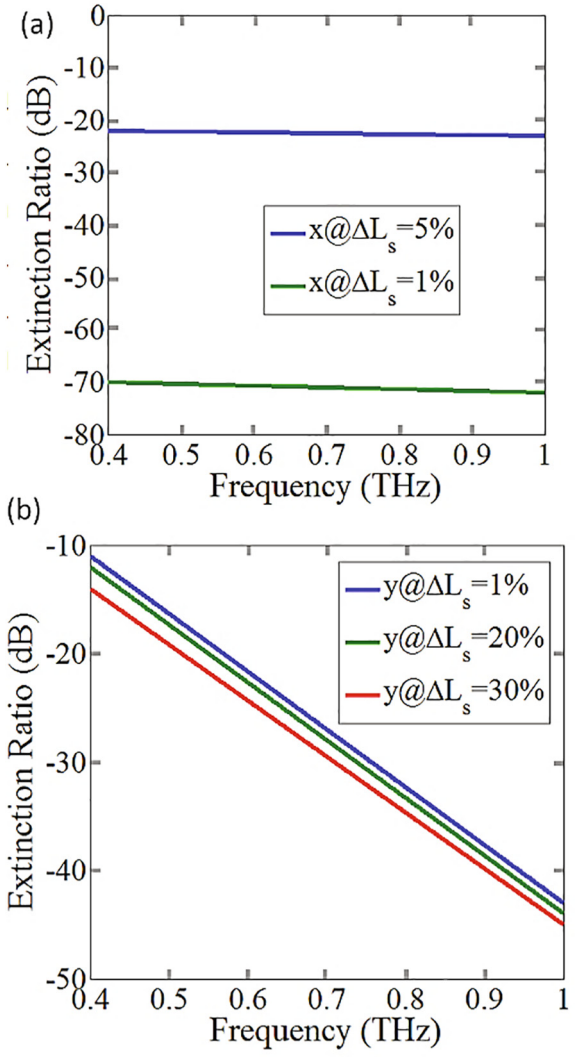


Fig. 5. Spectral variation of the extinction ratios for (a) x-polarization and (b) y-polarization at different percentage deviation (ΔL) in the device length.

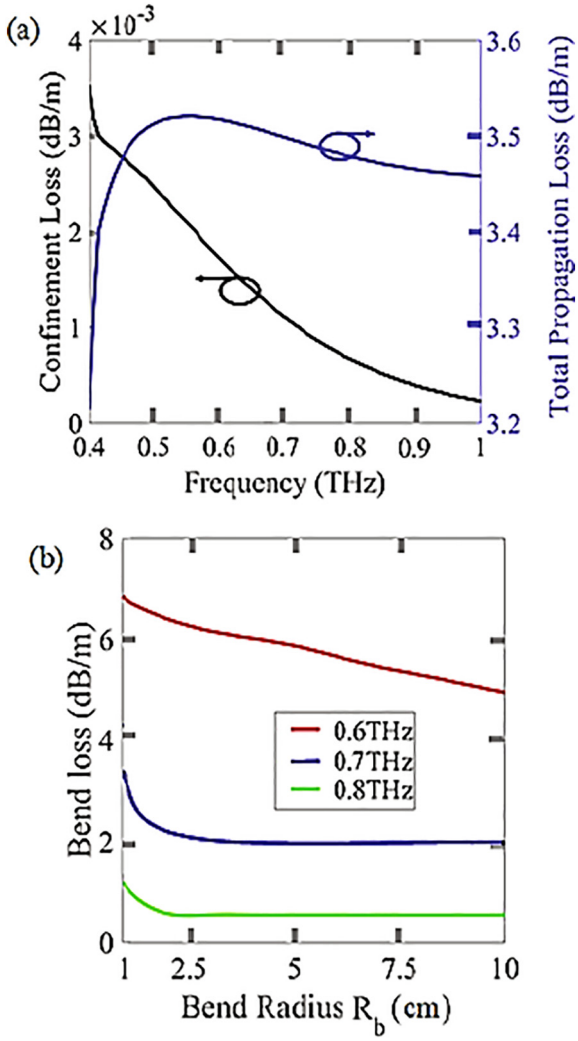


Fig. 6. (a) Frequency dependence of the confinement loss and the total propagation loss in the dual core polyethylene microstructured fiber (Fiber-AB). (b) Bending loss in dB/m calculated with variation in bend radius at various fixed THz frequencies.

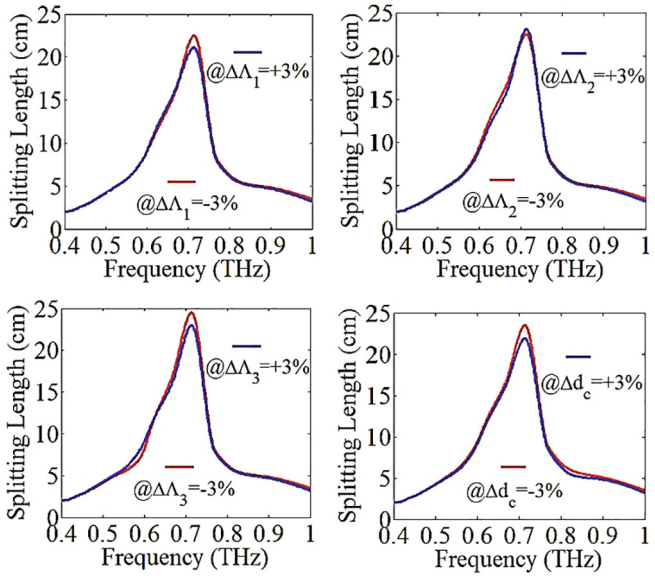


Fig. 7. Spectral variations of splitting length with ± 3% deviation in various design parameters. Here, Δ represents deviation in parameters.

the actual values of the extinction coefficients for all the materials (Piesiewicz et al., 2007) in the design of the Fiber-AB. Fig. 6(b) presents the bend loss characteristics of the Fiber-AB determined at various frequencies using the overlap vector integral method (Kumar et al., 2020). It can be seen that the bend loss is significantly higher at lower frequencies. However, this is not an issue as the splitting length is sufficiently small at the low frequencies that bending of the fiber is not required in practical applications.

The dimensional parameters of the proposed structure are in the range of 0.1 mm, which can be easily achieved in recent fabrication technologies (e.g., stacking method (Pysz et al., 2016; Chow et al., 2012). However, in order to see fabrication tolerances of the proposed structure, we have also studied changes in the performance characteristics of the device by considering a typical change (±3%) in the various designed parameters (e.g., various pitches, core air hole diameter (Λ_1, Λ_2, Λ_3 and d_c)). Some of these results are shown in Fig. 7. This figure clearly indicates that the performance characteristics of the device remain almost same due to low splitting length.

4. Conclusion

We have proposed an efficient, very compact and low loss broadband THz polarization splitter based on dual core photonic crystal index guided fiber. The fiber consists of specially configured porous dual core structure with inclusion of teflon rods to form slit-like pattern. The structure is designed to achieve low device length at higher frequencies. From the proper engineering in the configuration of the dual core, it allows us to achieve a device length of ~ 2.1 cm at 1THz, which, to our best knowledge, is the shortest one achieved in such structures. Moreover, we have achieved large extinction ratio, e.g. ~ 70 dB for x-polarization and ~ 45 dB for y-polarization at 1 THz, which remains very high at high frequencies. The bend loss and the propagation loss are significantly small suggesting efficient performance of the proposed dual core PCF as a compact tool for polarization splitting in THz photonics.

References

Koch, B., Noé, R., Sandel, D., Mirvoda, V., 2014. Versatile endless optical polarization controller/tracker/demultiplexer. Opt. Express 22, 8259–8276.

Hidayat, A., Koch, B., Zhang, H., Mirvoda, V., Lichtinger, M., Sandel, D., Noé, R., 2008. High-speed endless optical polarization stabilization using calibrated waveplates and field-programmable gate array-based digital controller. Opt. Express 16, 18984–18991.

Burrow, J.A., Yahiaoui, R., Sarangan, A., Agha, I., Mathews, J., Searles, T.A., 2017. Polarization-dependent electromagnetic responses of ultrathin and highly flexible asymmetric terahertz metasurfaces. Opt. Express 25, 32540–32549.

C. M. Morris, R. V. Aguilar, A. V. Stier, and N. P. Armitage, "Polarization modulation time-domain terahertz polarimetry," in Imaging and Applied Optics Technical Papers, OSA Technical Digest (online) (Optical Society of America, 2012), paper SW4C.2

Sesay, M., Jin, X., Ouyang, Z., 2013. Design of polarization beam splitter based on coupled rods in a square-lattice photonic crystal. J. Opt. Soc. Am. B 30, 2043–2047.

Park, J.M., Lee, S.G., Park, H.R., Lee, M.H., 2010. High-efficiency polarization beam splitter based on a self-collimating photonic crystal. J. Opt. Soc. Am. B 27, 2247–2254.

Zhong, S., Sheng, L., 2015. Terahertz wave polarization splitter using full band-gap photonic cystal. Journal of infrared, millimeter and terahertz waves 36, 255–261.

Mo, G.Q., Li, J.S., 2016. Compact terahertz wave polarization beam splitter using photonic crystal. Appl. Opt. 55, 7093–7097.

Jiang, H., Wang, E., Jhiang, J., Hu, L., Mao, Q., Li, Q., Xie, K., 2014. Polarization splitter based on dual core photonic crystal fibers. Optics Express 22, 30461–30466.

Saitoh, K., Sato, Y., Koshiba, M., 2004. Polarization splitter in three core photonic crystal fibers. Optics Express 22, 3940–13046.

Khaleque, A., Hattori, H.T., 2015. Ultra-broadband and compact polarization splitter based on gold filled dual core photonic crystal fiber. Journal of applied physics 118, 143101.

Wang, J., Pei, L., Weng, S., Wu, L., Li, J., Ning, T., 2018. Ultrashort polarization beam splitter based on liquid filled dual-core photonic crystal fiber. Applied Optics 57, 3847–3852.

Liu, Chao, Wang, Liying, Wang, Famei, Chunhong, Xu., Liu, Qiang, Liu, Wei, Yang, Lin, Li, Xianli, Sun, Tao, Chu, Paul K., 2019. Tunable single-polarization bimetal-coated and liquid-filled photonic crystal fiber filter based on surface plasmon resonance. Appl. Opt. 58, 6308–6314.

Younis, B.M., Heikal, A.M., Hameed, Mohamed Farhat O., Obayya, S.S.A., 2018. Highly wavelength-selective asymmetric dual-core liquid photonic crystal fiber polarization splitter. J. Opt. Soc. Am. B 35, 1020–1029.

Qu, Y., Yuan, J., Zhou, X., Li, F., Yan, B., Wu, Q., Wang, K., Sang, X., Long, K., Yu, C., 2020. Surface plasmon resonance-based silicon dual-core photonic crystal fiber polarization beam splitter at the mid-infrared spectral region. J. Opt. Soc. Am. B 37, 2221–2230.

Chen, H., Yan, G., Forsberg, E., He, S., 2016. Terahertz polarization splitter based on a dual-elliptical-core polymer fiber. Applied Optics 55, 6236–6242.

Li, S., Zhang, H., Hou, Y., Bai, J., Liu, W., Chang, S., 2013. Terahertz polarization splitter based on orthogonal microstructure dual-core photonic crystal fiber. Appl. Opt 52, 3305–3310.

Kumar, V., Varshney, R.K., Kumar, S., 2020. Design of a compact and broadband terahertz polarization splitter based on gradient dual-core photonic crystal fiber. Applied Optics 59, 1974–1979.

Piesiewicz, R., Jansen, C., Wietzke, S., Mittleman, D., Koch, M., Kürner, T., 2007. Properties of Building and Plastic Materials in the THz Range. International Journal of infrared and millimeter waves 28 (5), 363–371.

Pysz, D., Kujawa, I., Stępień, R., Klimczak, M., Filipkowski, A., Franczyk, M., Kociszewski, L., Buźniak, J., Haraśny, K., Buczyński, R., 2016. "Stack and draw fabrication of soft glass microstructured fiber optics," Boulletin of the Polish Academy of. Science 62 (2), 667–682.

D. M. Chow, S. R. Sandoghchi and F. R. M. Adikan, "Fabrication of photonic crystal fibers," 2012 IEEE 3rd International Conference on Photonics, Pulau Pinang, Malaysia, 2012, pp. 227-230, doi: 10.1109/ICP.2012.6379830.

Islam, M.S., Cordeiro, C.D.M., Franco, M.A.R., Sultana, J., Cruz, A.L.S., Abbott, D., 2020. Terahertz optical fibers [Invited]. Opt. Express 28, 16089.

Pedrola, L., 2016. Beam Propagation Methods for Design of Optical Waveguide Devices. Wiley.

Rahman, B.M.A., 2013. Finite Element Modeling Methods for Photonics. Artech house.

Kumar, V., Varshney, R.K., Kumar, S., 2020. "Terahertz generation by four-wave mixing and guidance in diatomic teflon photonic crystal fibers,"Optics. Communications 454, 124460.

Fiber-coupling optical system for high-power and multi-wavelength diode laser bars oriented to integrated biomedical imaging systems

Madhulita Mohapatra, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, madhulitamohapatra@gmail.com*

Ashis Singh, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, asish.singh2156@gmail.com*

Premananda Sahu, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, premanandasahu8@live.com*

Srimanta Mohapatra, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srimantamohaparta66@gmail.com*

ARTICLE INFO

Keywords:
Biomedical imaging
Photoacoustic endoscopy
Diode laser bars
Optical beam shaping
Fiber-coupled system
Near infrared
High power diode lasers

ABSTRACT

The use of pulsed high-power diode lasers (HPDLs) in the near infrared (NIR) range as multispectral sources has attracted much interest in biomedical applications due to their relatively low cost and small size compared to nanosecond Nd:YAG optical parametric oscillator (OPO) lasers. One of the main limitations of these sources is the availability to combine different wavelengths with high power density in the same beam. Various works have shown that the use of linear arrays of emitters as diode laser bars (DLBs) or stacked arrays of emitters as diode laser stacks (DLSs) allows the combination of multiple wavelengths while maintaining high power densities. Nevertheless, the highly asymmetric beam profile emitted by such laser sources between fast and slow axes implies the need for suitable beam shaping for efficient fiber coupling. In this work, we investigate a novel beam shaping technique to homogenize the beam quality of six DLBs in the wavelength range between 790 nm and 980 nm.

We consider fast-axis collimating lenses (FAC) and beam twisters to reduce the beam asymmetry of the individual bars. The beams from the DLBs are then combined into a single multispectral beam using reflective mirrors, dichroic mirrors, and a polarizing beam splitter cube (PBC), and effectively coupled into a 400 μm core-diameter/N.A. = 0.22 optical fiber using a pair of cylindrical lenses. Simulation shows high coupled power densities with ~ 1.8 MW/cm^2 at the output of the fiber. The coupling efficiency reaches 89.4%. The use of sub-millimeter fiber optic probes is particularly promising for photoacoustic endoscopy (PAE) applications requiring minimally invasive examination of internal organs.

1. Introduction

Non-ionizing imaging techniques based on NIR radiation offer several potential advantages over existing radiological techniques: they provide radiation that is harmless to the patient; they assist the therapist in discriminating between soft tissues because of their differential absorption and scattering at NIR wavelengths that cannot be distinguished with other modalities; and they allow extraction of functional information through absorption of natural chromophores. NIR imaging research has focused on a variety of potential clinical applications aimed at screening breast cancer and other early growth stage diseases so that tumors can be distinguished from surrounding healthy tissue before metastases form and treatment becomes more complicated. Among the various biomedical applications, optical coherence tomography (OCT) (Israelsen et al., 2019), diffuse optical imaging (DOI) (Hoshi et al., 2016), photoacoustic imaging (PAI) and spectroscopy (PAS) (Sun et al., 2020; Dumitras et al., 2020), and fluorescence lifetime imaging (Lian et al., 2019) stand out. In particular, photoacoustic techniques are novel and promising biomedical imaging modalities that have received considerable attention in recent decades due to their ability to combine the high imaging contrast of optical imaging and the high penetration depth of ultrasound imaging (Gao et al., 2017; Attia et al., 2019; Yu et al., 2019; Zhou and Jokerst, 2020; Hosseinaee et al., 2020). These technologies combine the advantages of optical and ultrasonic techniques and find fertile ground in the NIR range. In this regard, multi-wavelength laser sources are necessary for a functional PAS by acquiring differentiated image data over the NIR spectrum. Numerous studies have experimentally investigated laser-excited photoacoustic tomography (PAT) to detect chromophores in turbid tissues with different optical properties (Park et al., 2021; Xia et al., 2014; Wang and Yao, 2016). Both Nd:YAG-OPO nanosecond lasers and near-infrared pulsed HPDLs have been used for PAI applica-

tions (Kolkman et al., 2006; Shu et al., 2015; Upputuri and Pramanik, 2015), depending on the scope of each experiment. Although Nd:YAG-OPO nanosecond lasers provide much higher peak power compared to HPDLs, such lasers are expensive, bulky, and provide limited imaging speed due to their low pulse repetition rate (tens of Hz). Moreover, their fixed wavelength (532 nm or 1064 nm) limits their application in biomedical imaging.

On the other hand, power combining using linear arrays of pulsed near-infrared HPDLs are considered as scalable and cost-effective alternatives for PAI (Erfanzadeh et al., 2017; Yao et al., 2017; Upputuri and Pramanik, 2018), offering kHz repetition rates and sufficient power to penetrate deep into the tissue. Moreover, these arrays can be scaled down by using linear arrays of diode laser cavities (i.e., DLBs or DLSs) (Leggio et al., 2017; Yu et al., 2017; Lin et al., 2020), which emit high peak powers desired for PAI applications in the 700 1100 nm wavelength range (NIR region), where soft tissues exhibit high absorption. To couple the power of such sources into an optical fiber, the beam must be focused to a small spot (typically 1000 μm) (Sullins, 2002; Polese et al., 2020; Gawali et al., 2016), which in turn requires relatively good beam quality (i.e., Gaussian-like beam profile intensity). However, the beam profile of these diode laser sources becomes increasingly elliptical in the far field, since the width of each emitter (50 200 μm) is much larger than its respective height (typically 1 μm). This elliptical beam shape of diode laser emitters is colloquially referred to as astigmatism. In general, the beam divergence angle of a single diode laser emitter at 95% power content is about 36-66° along the fast axis and 7-13° along the slow axis (Diehl, 2000).

Therefore, their highly divergent and asymmetric beam emission must be homogenized by an optimized beam shaping system to achieve efficient coupling into an optical fiber. The high asymmetry of the beam emitted by the DLBs causes their beam parameter product (*BPP*) in both axes to be inconsistent with that of the optical fiber, thus affecting the quality of the fiber coupling. To overcome this limitation, it is necessary to design an effective and compact beam shaping scheme to homogenize the *BPP* ratio between fast axis and slow axis before fiber coupling.

In a previous work (Yu et al., 2016), the design of a beam shaping system based on the use of beam twisters to improve the beam quality of DLBs was proposed for the first time. That study, together with experimental results, specifically demonstrated an improvement in the beam quality of a mini-bar of 9 emitters by using a FAC microlens to collimate the bar beam, a beam twister to rearrange the beam shape, a SAC lens to collimate the beam on the slow axis, and a FAC lens array to perform re-collimation and further reduce beam divergence along the fast axis. Our initial concern was to find a way to reduce the optical elements in the system, and we eventually considered that the SAC lens and the FAC lens array of Ref. (Yu et al., 2016) could be replaced by a single element. We also felt that the proposed system could be further improved and extended to a wider range of wavelengths for biomedical applications.

In this regard, this work describes a novel optical beam shaping scheme modeled in Zemax to homogenize the beam quality of six multi-wavelength DLBs coupled to a 400 μm core-diameter optical fiber. Specifically, we model cataloged DLBs (model QD-Q1901-A1 with 19 emitters, *Quantel*) emitting ns pulses in the wavelength range between 790 nm and 980 nm.

Previous works (Kalva et al., 2019a, 2019b; Upputuri and Pramanik, 2015; Shabairou et al., 2020; Ji et al., 2015) reported the use of pulsed lasers to achieve PAI, PAT, and PAE at specific wavelengths. However, the systems presented there can be extended by using a multispectral compact system based on DLBs.

In this paper, a beam shaping technique is presented to homogenize the beam quality of six DLBs and achieve multispectral fiber coupling. The proposed technique is based on beam transformation systems (BTSs, i.e., two FAC lenses with beam twister), reflective mirrors, and dichroic mirrors as beam homogenization elements. Specifically, we use the BTSs to collimate and homogenize the light beam of the individual bars, the dichroic mirrors with selective transmission bands and a PBC to combine the beams in a multispectral source, and a pair of cylindrical lenses with a 90° offset to enable beam coupling into a 400 μm core-diameter/ N.A. = 0.22 optical fiber.

The obtained results make the proposed system suitable for integrated PAE systems that require a variable amount of power in a multispectral configuration (Zhou and Jokerst, 2020). This method achieves beam symmetrization of the individual beams with a significant improvement in *BPP* ratios, resulting in a very high coupling efficiency into the fiber of 89.4% of the multispectral source and an output power density of ~ 1.8 MW/cm^2.

2. The origin of astigmatism in diode laser emitters

At a few micrometers distance from the active layer facet of DLBs, the beam size in the fast axis direction is equal to the beam size in the slow axis and the beam cross section is still circular. Beyond this distance, the beam cross section becomes elliptical but with the principal diameter in the fast axis direction. As shown in Fig. 1, the imaginary beam source points P_x and P_y are located at different positions, because $w_{0y} > w_{0x}$ and $\theta_{//half} < \theta_{\perp half}$, where w_{0x} and w_{0y} are the beam waist radii along the fast axis and the slow axis at the output facet of the diode laser, respectively, whereas $\theta_{\perp half}$ and $\theta_{//half}$ are the corresponding half-angle divergences of the beam along the perpendicular axis (fast axis) and the parallel axis (slow axis), respectively. In the same figure, it can be seen that the beam source point P_x is located on the output facet of the active layer due to the tiny size of w_{0x} (typically 1 μm). Consequently, by tracing the far-field beam backwards, one can locate the beam source point P_y inside the active layer. The distance between P_x and P_y is the magnitude of the astigmatism.

Considering a Gaussian beam, at a position z after the diode laser facet, the spot size parameter can be described by the combination of beam waist radii $w_x(z)$ and $w_y(z)$ in the fast axis and slow axis, respectively. These two parameters are given by the following formulas (Svelto and Hanna, 2013):

$$w_x(z) = w_{0x} \cdot \left(1 + \left(\frac{\lambda z}{\pi w_{0x}^2}\right)^2\right)^{1/2}, \quad (1)$$

And

$$w_y(z) = w_{0y} \cdot \left(1 + \left(\frac{\lambda z}{\pi w_{0y}^2}\right)^2\right)^{1/2}, \quad (2)$$

where λ is the source wavelength. In this context, the calculation of the beam size at a certain position z after the output facet of the emitter is necessary to build the proper optical system for symmetrizing the opti-

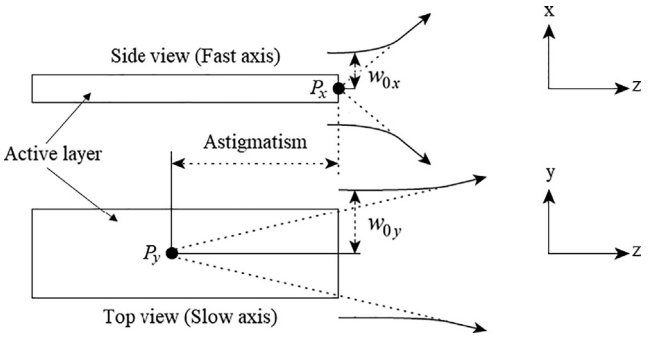


Fig. 1. Sketch of the side view and top view of a diode laser cavity. The phenomenon of astigmatism results from the distance of the source points on the fast axis and the slow axis (www.newport.com/resourceListing/technical-notes).

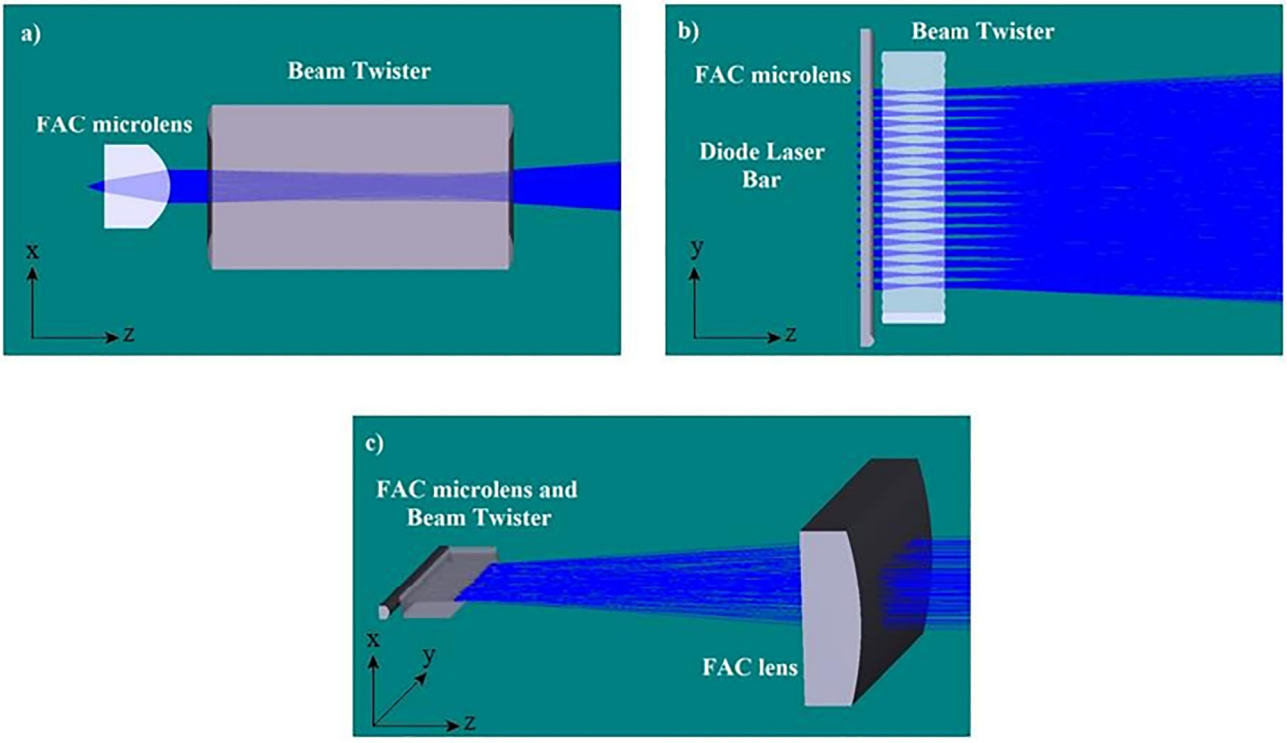


Fig. 2. Beam collimation and homogenization using a FAC microlens and a beam twister: a) side view, b) top view, c) perspective view. First, the FAC microlens collimates the large beam divergence in the fast axis, then the beam twister rotates the beam by 90° after an internal focusing (a, b). A second FAC lens collimates the residual divergence coming from the beam twister (c).

cal beam. In the case of DLBs, it is necessary to consider the number of emitters along the y dimension, so Eq. (2) must be reformulated as follows:

$$w_y(z) = w_{0y} \cdot \left(1 + \frac{\lambda z}{\pi w_{0y}^2}\right)^{2}\Bigg)^{1/2} \cdot N_y/FF_{bar}, \tag{3}$$

where N_y is the number of the emitters of each bar placed along the slow axis and FF_{bar} is the bar fill factor corresponding to the ratio between the emitter width and the emitter pitch.

3. Beam quality and symmetrization of the diode laser bars

The *BPP* is an essential parameter for estimating beam quality in both the fast and slow axes and is closely related to the measure of how efficiently a light beam can be focused. The higher the beam quality, the easier it will be to focus the beam on a smaller spot size at the input of the optical fiber, also depending on the optical system used. Due to the highly asymmetric rectangular shape of DLBs, the beam profile of these laser sources is highly asymmetric in the far field between the fast axis and the slow axis. Therefore, it is necessary to estimate the *BPP* in both fast and slow axes. In the fast axis, it is defined as the product of the beam waist radius w_{0x} (i.e., half the vertical size in the fast axis; in our case $w_{0x} = 0.5\mu m$) with the half-angle divergence $\theta_{\perp half}$ of the beam along the perpendicular axis (fast axis). Conversely, in the slow axis, it corresponds to the product of the beam waist radius w_{0y} (i.e., half the emitter width in the slow axis) with the half-angle divergence $\theta_{//half}$ of the beam along the parallel axis (slow axis). Hence, the *BPPs* along fast axis and slow axis (*BPPx* and *BPPy*) are respectively expressed by:

$$BPP_x = w_{0x} \cdot \theta_{\perp half}, \tag{4}$$

$$BPP_y = w_{0y} \cdot \theta_{//half} \tag{5}$$

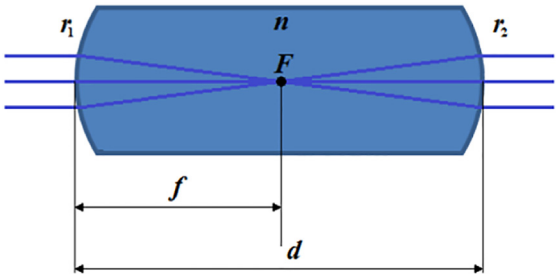


Fig. 3. Schematic diagram of a single lenslet of a beam twister.

When considering a DLB, Eq. (5) must be multiplied by the repetition factor k. Since the emitters in this case are spaced apart, their repetition factor must be normalized to a constant value expressing the percentage of the longitudinal space they occupy. Therefore, Eq. (5) is multiplied to $k = N_y/FF_{bar}$, as follows:

$$BPP_y = \frac{w_{0y} \cdot \theta_{//half} \cdot N_y}{FF_{bar}}, \tag{6}$$

In order to achieve the necessary beam quality for efficient fiber coupling, the *BPP* ratio (*BPPy/BPPx or BPPx/BPPy*) must be reduced using beam shaping optical systems. Ideally, this corresponds to a perfectly rectangular shape of the beam, independent of the distance to the source, and can be achieved by beam symmetrization. According to previous theoretical studies (Loosen et al., 2007), the *BPP* ratio is usually around 500:1 or even higher for a single bar. Consequently, beam symmetrization is necessary to homogenize the *BPP* ratio. Specifically, the line shape of the beam is cut into individual segments that are optically rearranged so that *BPPx* is increased by a multiple while *BPPy* is decreased, or vice versa. In this context, the beam qual-

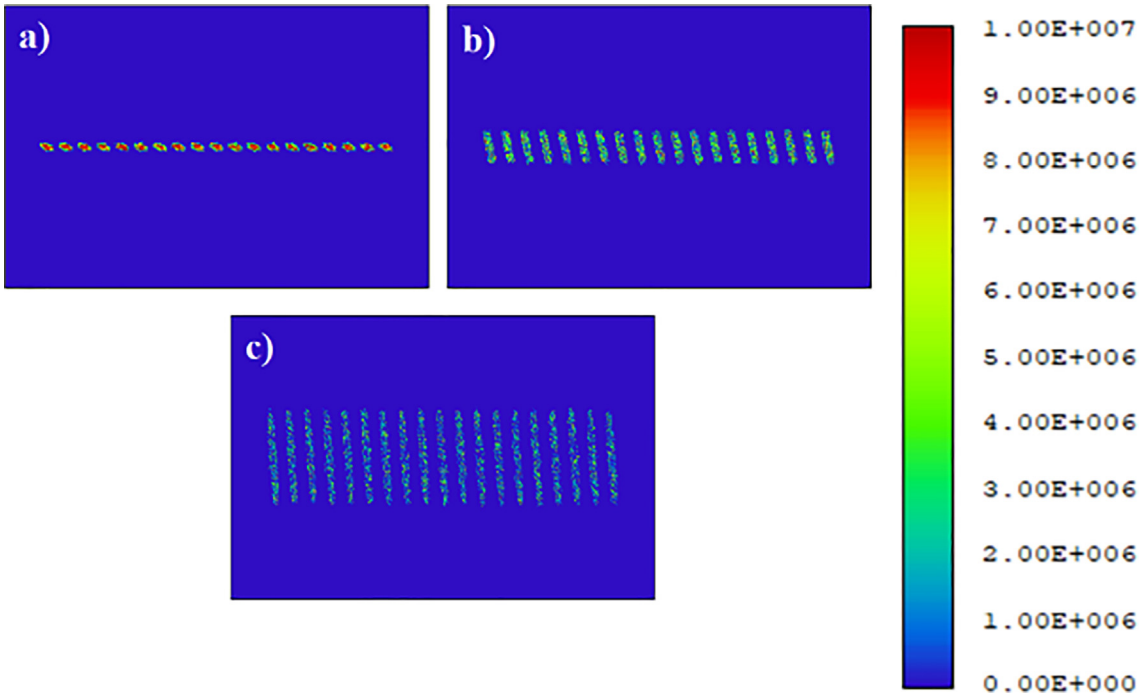


Fig. 4. Beam profiles projected by a beam twister: a) immediately after the beam twister at z = 2.4 mm, b) after 6 mm, c) after 14.4 mm (just before the second FAC lens).

ity of the optical system can be described by an overall *BPP*, namely *BPPtot,rms*, which is defined by the diagonal between *BPPx* and *BPPy* according to the following equation (Loosen et al., 2007):

$$BPP_{tot,rms} = \sqrt{BPP_x^2 + BPP_y^2}. \tag{7}$$

In ideal condition of perfect beam symmetrization $BPP_y/BPP_x = 1$, so that Eq. (7) is reduced to:

$$BPP_{tot,rms,ideal} = \sqrt{2}BPP_x = \sqrt{2}BPP_y. \tag{8}$$

This suggests a beam symmetrization that homogenizes the beam between the two axes. After beam symmetrization, the *BPPs* are swapped between the fast and slow axes: the new *BPP* in the vertical axis is obtained by splitting *BPPy* into *N* subsections, while the new *BPP* in the horizontal axis is obtained by multiplying *BPPx* by *N*, as expressed by the following equations (Loosen et al., 2007):

$$BPP_v = BPP_y/N, \tag{9}$$

and,

$$BPP_h = BPP_x \cdot N, \tag{10}$$

where BPP_v and BPP_h are the rearranged *BPPs*. In this way, the *BPPs* are adjusted by shifting the beam quality from one axis to the other. In this line, the beam quality of the optical system after the beam rearrangement $BPP'_{tot,rms}$ can be defined by the following equation (Loosen et al., 2007):

$$BPP'_{tot,rms} = \sqrt{BPP_v^2 + BPP_h^2}. \tag{11}$$

In order to achieve an efficient fiber coupling of the beam, the value of $BPP'_{tot,rms}$ must not exceed the value of BPP_{fiber}, as expressed by the following equation:

$$BPP'_{tot,rms} \leqslant BPP_{fiber} = \frac{d}{2} \cdot \theta_{half\,\max}, \tag{12}$$

where BPP_{fiber} and d are the *BPP* and the core diameter of the optical fiber, respectively, and $\theta_{half\ max}$ is the maximum half beam divergence admitted by the input, being expressed as follows:

$$\theta_{half\max} = arcsin(\text{N.A.}), \tag{13}$$

where N.A. is the numerical aperture of the fiber. Eq. (12) shows that the upper limit of the *BPP* of the optical system can be defined by the physical parameters of the fiber. In an ideal, though rare, situation, the *BPPs* would be equal in the two axes. The beam twisters rotate the beam of each laser emitter in the bar by 90°, shifting the beam quality from one axis to the other (Fig. 2). After the beam twister, the *BPPs* of each bar can be expressed as follows, respectively:

$$BPP_v = w_{0y} \cdot \theta_{//half}, \tag{14}$$

and

$$BPP_h = \frac{w_{0x} \cdot \theta_{\perp half} \cdot N_y}{FF_{bar}}. \tag{15}$$

The above equations apply to the general case of a beam twister positioned after the source, but in a real design (Fig. 2) it is necessary to consider a FAC microlens that changes the beam divergence along the fast axis immediately after the source. The new fast axis divergence is denoted as $\theta_{\perp \mathrm{FAC}\ half}$, so that Eq. (15) is rewritten as:

$$BPP_{FAC\ h} = \frac{w_{0x} \cdot \theta_{\perp FAC\ half} \cdot N_y}{FF_{bar}}. \tag{16}$$

After successive beam collimation by a second FAC lens, the beam is nearly homogenized between fast axis and slow axis with a residual divergence (nearly < 10 mrad). Therefore, taking into account the rearranged beam sizes and divergences, the *BPP* equations can be restated as follows:

$$BPP_v^{rearr} = s_v \cdot \theta_{v\ half}, \tag{17}$$

$$BPP_h^{rearr} = \frac{s_h \cdot \theta_{h\ half} \cdot N_y}{FF_{bar}^{rearr}}, \tag{18}$$

where s_v, θ_{vhalf}, s_h, and θ_{hhalf} are the rearranged beam waist and residual half-divergence along vertical axis and horizontal axis, respectively, and FF_{bar}^{rearr} is the rearranged fill factor along the slow axis. Since it is difficult to estimate the residual divergences after BTS in this case, we can roughly assume their values to be < 10 mrad.

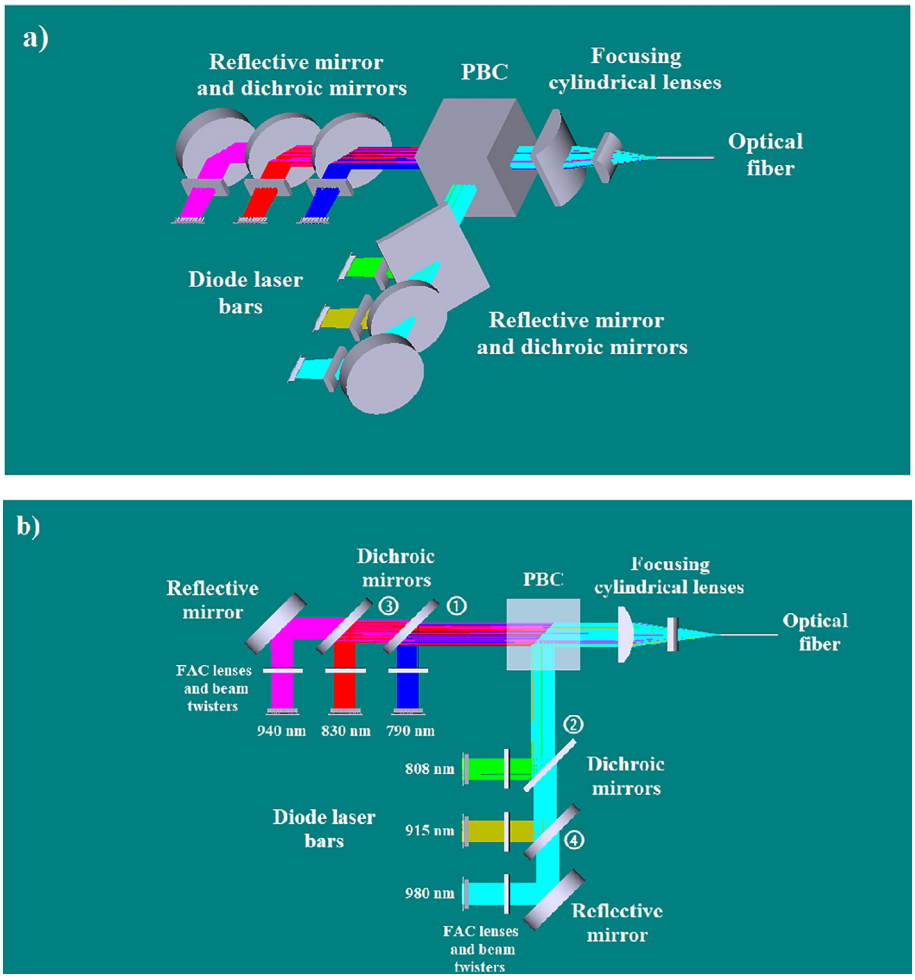


Fig. 5. Beam shaping design of a fiber-coupled multi-wavelength system based on DLBs by using realistic elements: a) Perspective view, b) Top view with annotation for the dichroic mirrors (referred to with circled numbers in the text).

Table 1
Parameters of the DLBs (model QD-Q1901-A1) (www.quantel-laser.com).

Parameter	Symbol	Value	Unit
Wavelengths	λ	790, 808, 830, 915, 940, 980	nm
Fast axis beam divergence[1] ($1/e^2$)	$\theta_{\perp}$	36	degrees
Slow axis beam divergence[1] ($1/e^2$)	$\theta_{//}$	8	degrees
Emitter width	$2 * w_{0y}$	100	μm
Emitter pitch[2]	p_e	400	μm
Number of emitters per bar	N_y	19	Units
Output power per bar	P_{bar}	400	W
Bar width	w	7.3	mm

[1] The divergences are referred to each bar (FWHM).
[2] The emitter pitch refers to the spacing between emitters.

3.1. Key elements for beam symmetrization

The beam emitted from a DLB has a line shape with poor beam quality (Loosen et al., 2007), so the *BPP* ratio of each DLB must be reshaped using optical or micro-optical elements. A crucial optical element required for such beam reshaping is the BTS (i.e., the FAC lenses with the beam twister, Fig. 2(a)). The FAC microlens is designed to reduce the large beam divergence in the fast axis, while the beam twister, which consists of an array of biconvex 45°-tilted cylindrical lenses, rotates the beam by 90° around the optical axis. More precisely, after collimation of the fast axis, the 45°-tilted input facet of the beam twister lenslet internally focuses the beam collimated by the FAC microlens at its center (Fig. 2(b)), with a 45° rotation. After internal focusing, the beam from the output facet is again rotated by 45°, having the same radius of curvature as the input facet. In this way, the horizontal and vertical components (beam waists and divergence angles along the fast and slow axes) are swapped and the beams are stacked on top of each other (Loosen et al., 2007).

At the output of the beam twister, the emerging beam is still slightly divergent in both axes, but now the divergence can be corrected by an additional FAC lens (Fig. 2(c)). After beam shaping by the BTS, the resulting beam can be collimated and then coupled into an optical fiber with additional cylindrical lenses, approaching condition of Eq. (8). The schematic diagram of a single lenslet of a beam twister with refractive index n is shown in Fig. 3. The radii of curvature and the lens thickness satisfy the relations $r_1 = -r_2$ and $d = 2f$, respectively, where d is the thickness and f is the focal length of the lenslet. The inner focal point F is located at the center of the lenslet. The beam profiles emerging from a beam twister after three different distances are shown in Fig. 4.

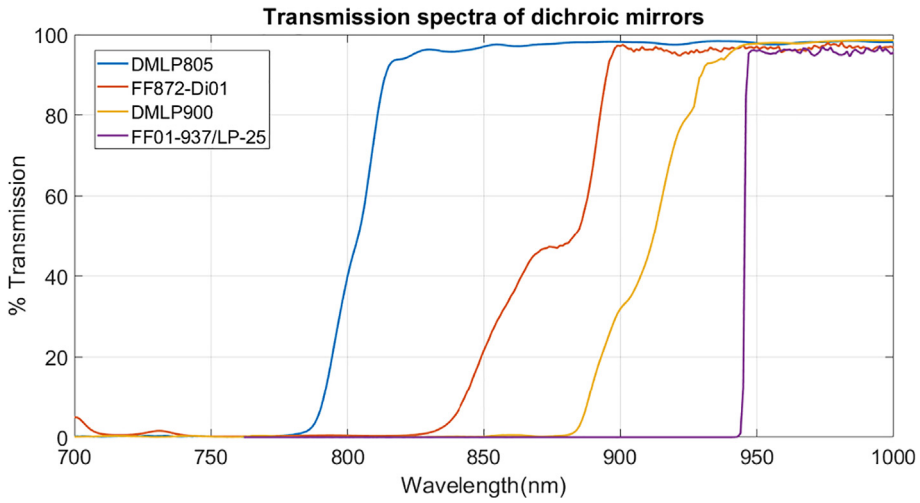


Fig. 6. Transmission spectra of the dichroic mirrors used in the design of the multi-wavelength system (www.thorlabs.com; www.semrock.com).

Table 2
List and characteristics of the optical components of the setup (www.thorlabs.com; www.semrock.com; www.limo.de).

Components	Manufacturer	Model	Characteristics
Beam transformation systems	Limo GmbH	BTS-HOC 200-400 (only BTS)	[1]L = 12 mm, [2]CT = 0.354 mm, [3]H = 0.45 mm (FAC microlens), L = 10.5 mm, CT = 1.659 mm, H = 0.98 mm (beam twister), L = 14 mm, CT = 2 mm, H = 6 mm (FAC lens)
Reflective mirrors	Thorlabs Inc.	PF10-03-P01	[4]D = 25.4 mm, [5]TH = 6 mm [6]R > 97.5% for 450 nm - 2 μm
① Dichroic mirror	Thorlabs Inc.	DMLP805	D = 25.4 mm, TH = 3.2 mm R = 92.34 % (790 nm) [7]T = 96.30 % (830 nm) T = 98.34 % (940 nm)
② Dichroic mirror	Semrock Inc.	FF872-Di01	L = 25 mm, H = 36 mm, TH = 2 mm R = 99.69 % (808 nm) T = 96.51 % (915 nm) T = 97.09 % (980 nm)
③ Dichroic mirror	Thorlabs Inc.	DMLP900	D = 25.4 mm, TH = 3.2 mm R = 99.82 % (830 nm) T = 95.28 % (940 nm)
④ Dichroic mirror	Semrock Inc.	FF01-937/LP-25	D = 25 mm, TH = 3.5 mm R = 99.99 % (915 nm) T = 95.03 % (980 nm)
Polarizing beamsplitter cube	Thorlabs Inc.	PBS252	25.4 mm x 25.4 mm x 25.4 mm (composed by two triangular prisms)
Focusing cylindrical lens (slow axis)	Thorlabs Inc.	LJ1212L1-B	[8]EFL = 29.99 mm, L = 22 mm, H = 20 mm, [9]r = 15.5 mm, CT = 5.7 mm, [10]ET = 2 mm, [11]BFL = 26.3 mm
Focusing cylindrical lens (fast axis)	Thorlabs Inc.	LJ1636L1-B	EFL = 15 mm, L = 12 mm, H = 10 mm, r = 7.8 mm, CT = 3.8 mm, ET = 2 mm, BFL = 12.5 mm
Optical fiber	Thorlabs Inc.	AFM400H	N.A. = 0.22, Ø400 μm, multimode

[1] L = Length, [2]CT = Center Thickness, [3]H = Height, [4]D = Diameter, [5]TH = Thickness, [6]R = Reflectance, [7]T = Transmittance, [8] EFL = Effective Focal Length, [9]r = Radius of Curvature, [10]ET = Edge Thickness, [11]BFL = Back Focal Length

4. Design of the optical system

In a previous work (Leggio et al., 2017), we presented a method for beam shaping five DLBs (*Jenoptik Optical Systems LCC*) emitting between 808 nm and 980 nm using microoptical FAC lenses and beam twisters. Subsequently, beam combining was performed using dichroic mirrors and a PBC, and the resulting multispectral beam was focused into a 400-μm optical fiber. Similarly, in this work, six DLBs, each consisting of 19 emitters, are used as multispectral sources in the range between 790 nm and 980 nm, as shown in Fig. 5. The output beams of the bars are individually collimated along the fast axis using FAC microlenses nearby the sources and homogenized by the beam twisters (Fig. 2). The beams are then collimated along the vertical axis before being combined using the dichroic mirrors and a PBC (Fig. 5). Six commercial DLBs (model QD-Q1901-A1 with 19 emitters, *Quantel*) are considered for the simulation (see Table 1). In Fig. 5, the six DLBs are shown as follows: 790 nm (blue), 808 nm (green), 830 nm (red), 915 nm (yellow), 940 nm (magenta), and 980 nm (cyan). In addition, optical elements such as the BTSs (first part of BTS-HOC 200–400*[1], *Limo GmbH*), the two reflective mirrors (PF10-03-P01, *Thorlabs Inc.*), the dichroic mirrors (① DMLP805 and ③ DMLP900 from *Thorlabs Inc.*, ② FF872-Di01 and ④ FF01-937/ LP-25 from *Semrock Inc.*), and a

[1] * Only the BTS part.

Table 3
Beam waist radii at the input face of the FAC microlenses.

Wavelength (nm)	Beam waist $w_x(z)$ (mm)	Beam waist $w_y(z)$(mm)
790	0.04728	~3.8
808	0.04836	~3.8
830	0.04967	~3.8
915	0.05476	~3.8
940	0.05625	~3.8
980	0.05865	~3.8

Table 4
BPP ratios of the beams at the output of a single bar and at the fiber input.

Beam	Position	Value
Single bar beam	After single bar	~1690
Multispectral beam	Before fiber	~4.09

Table 5
BPP parameters evaluated for fiber coupling.

Parameter	Element	Value
$BPP_{tot,rms}$	Focused beam	24.64 mm·mrad
BPP_{fiber}	200-μm optical fiber	22.18 mm·mrad
BPP_{fiber}	400-μm optical fiber	44.36 mm·mrad

PBC (PBS252, *Thorlabs Inc.*) were also taken from commercial catalogs. The transmission spectra of the dichroic mirrors are shown in Fig. 6. The cylindrical lenses (LJ1212L1-B and LJ1636L1-B, *Thorlabs Inc.*) and the optical fiber were also taken from catalog (AFM400H, N.A. = 0.22, Ø400 μm, multimode, *Thorlabs Inc.*). All cataloged components are listed in Table 2.

5. Results and discussion

In this section, two types of beam analysis are performed: the first considers the size of the emitted beam spot as a function of wavelength and how it affects beam shaping; the second involves the analysis of beam quality in terms of calculations of *BPPs* along the beam path until it reaches the optical fiber.

5.1. Analysis of beam spot size vs. Wavelength

As shown in Fig. 1, the position of the beam waist differs from the fast axis and the slow axis. It is important to note that the beam divergences and the beam waists are the same for all wavelengths (Table 1) at the output facet of the emitters, so that the positions of the source points P_x and P_y do not change. After beam emission, a FAC microlens (Fig. 2) is positioned at 94 μm in front of each DLB. Since the beam waist radii $w_x(z)$ and $w_y(z)$ depend on the wavelength (Eqs. (1) and (3)), they will be different for each case (Table 3).

Considering that the radius of curvature of the FAC microlenses is r = 0.235 mm, the beam waist $w_x(z)$ is between ~ 4.97 (at 790 nm) and ~ 4.0 (at 980 nm) times smaller, representing a small variation in the wavelength range. Conversely, the beam waist $w_y(z)$ is nearly unchanged. As a result, the beam spot size at the output facet of the FAC microlens varies between ~ 0.176 mm · 7.30 mm (at 790 nm) and ~ 0.178 mm · 7.30 mm (at 980 nm). This small difference has no effect on beam shaping; however, a beam twister is placed immediately after the FAC microlens to rearrange the beam, and a second FAC lens is responsible for removing residual divergence (Fig. 2).

5.2. Analysis of beam quality

Considering Eqs. (4) and (6), the *BPP* ratio is $\frac{BPP_y}{BPP_x} = 265.29\ mm \cdot mrad / 0.157\ mm \cdot mrad \cong 1690$ at the output of each bar. After the beam collimation performed by the FAC microlens, fast axis half beam divergence $\theta_{\perp \mathrm{FAC}\ half}$ of DLB reduces to ~ 10 mrad, thus enabling the focusing inside the beam twister. Hence, according to Eqs. (14) and (16), after beam twister the *BPP* ratio of each bar reduces to ~ 9.19, and is ~ 1.05 (Eqs. (17) and (18)) after a collimation on the vertical axis, considering a new beam size of $s_v \cdot s_h = 2\ \mathrm{mm} \cdot 0.32\ \mathrm{mm}$ and a residual half beam divergence of $\theta_{\mathrm{v\ half}} \cdot \theta_{\mathrm{h\ half}} = 4\ mrad \cdot 1\ mrad$ for each single bar (in this case the bar fill factor is 0.8). The pair of focusing cylindrical lenses (Fig. 5) equalizes the focal point between the two axes and focuses the beam into the optical fiber with a beam size of $140\ \mu m \cdot 380\ \mu m$ and half divergence of $83.6\ mrad \cdot 126\ mrad$. Hence, the *BPP* ratio at the input of the optical fiber is ~ 4.09 (Eqs. (4) and (5)) and the $BPP_{tot,rms}$ is 24.64 $mm \cdot mrad$ (Eq. (7)). This value is compatible with a multimode fiber with 400 μm-core diameter and N.A. = 0.22 that means $BPP_{tot,rms} < BPP_{fiber} = \frac{400\ \mu m}{2} \cdot 221.814\ mrad = 44.36\ mm \cdot mrad$ (Eq. (12)), considering a the N.A. of the fiber of 0.22 and $\theta_{half\,\max}$ of 221.814 $mrad$ (Eq. (13)). This

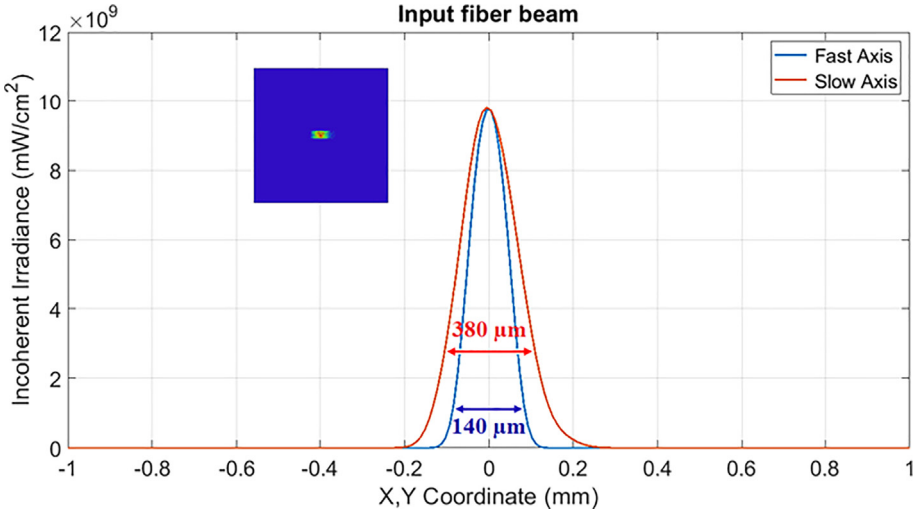


Fig. 7. Beam profile expressed in terms of incoherent irradiance and intensity pattern. The multispectral beam is focused in a 140 μm × 380 μm spot at the entrance of the optical fiber with a core-diameter of 400 μm.

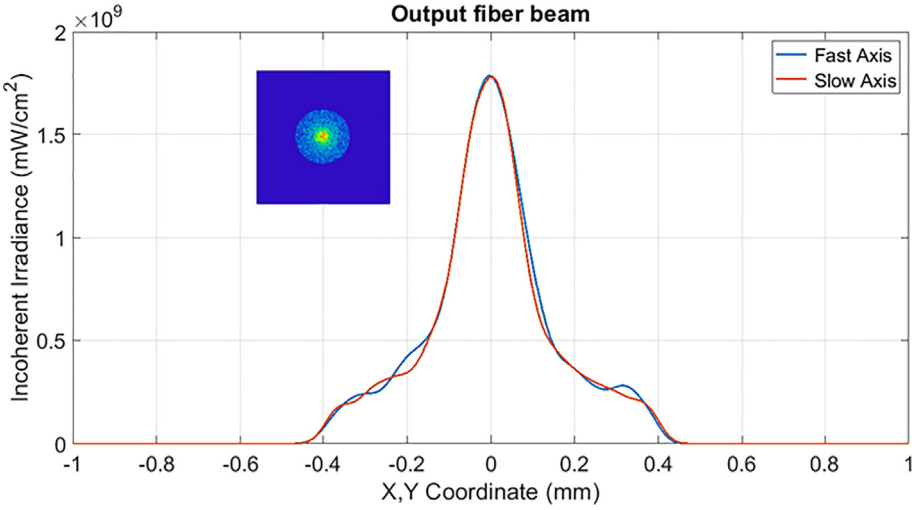


Fig. 8. Beam profile emerging from the fiber output (L = 20 cm) expressed in incoherent irradiance and intensity pattern.

means that the beam is effectively coupled into the fiber. Conversely, a fiber with 200 μm core-diameter and N.A. = 0.22 would have a BPP_{fiber} of 22.18 $mm \cdot mrad$, being incompatible with the condition of efficient fiber coupling (Eq. (12)).

Overall values are summarized in Tables 4 and 5, assuming a light detector with a resolution of 1024 *x* 1024 pixels first at the input and then at the output of the fiber. The beam profile at the input of the fiber (140 μm × 380 μm) is shown in Fig. 7, while the output profile after propagation in 20 cm fiber is shown in Fig. 8. The power coupling efficiency reaches 89.4% at the fiber output with a power density of ~1.8 MW/cm^2.

6. Conclusions

We modeled a compact beam shaping design in Zemax to homogenize the beam quality of six DLBs (790 nm − 980 nm) using realistic FAC lenses and beam twisters to significantly reduce the *BPP* ratio between fast axis and slow axis. The addition of reflective and dichroic mirrors allowed the six wavelengths to be combined in a single beam, which was coupled through a pair of cylindrical lenses into a 400-μm / N.A. = 0.22 optical fiber. A power density of ~ 1.8 MW/cm^2 is calculated at the output of the fiber with a coupling efficiency of 89.4%. The successful coupling of a multispectral source, such as the one described in this paper, into a 400-μm optical fiber enables new opportunities for applications in PAE requiring non-invasive inspection of the body, as well as for other emerging bio-imaging techniques where high-power multi-wavelength emission is required. The prototype numerical model presented in this paper has been proposed for possible future implementation and verification. However, there are no important technical limitations that would hinder the experimental corroboration of the results of this study by a laboratory equipped with the appropriate infrastructure.

References

Israelsen, N.M., Petersen, C.R., Barh, A., Jain, D., Jensen, M., Hannesschläger, G., Tidemand-Lichtenberg, P., Pedersen, C., Podoleanu, A., Bang, O., 2019. Real-time high-resolution mid-infrared optical coherence tomography. Light Sci. Appl. 8 (11). https://doi.org/10.1038/s41377-019-0122-5.

Hoshi, M.D., Yamada, Y., 2016. Overview of diffuse optical tomography and its clinical applications. Journal of Biomedical Optics 21 (9), 091312. https://doi.org/10.1117/1.JBO.21.9.091312.

Sun, A., Guo, H., Gan, Q., Yang, L., Liu, Q., Xi, L., 2020. Evaluation of visible NIR-I and NIR-II light penetration for photoacoustic imaging in rat organs. Opt. Express 28 (6), 9002–9013. https://doi.org/10.1364/OE.389714.

Dumitras, D.C., Petrus, M., Bratu, A.M., Popa, C., 2020. Applications of near infrared photoacoustic spectroscopy for analysis of human respiration: a review. Molecules 25 (7), 1728. https://doi.org/10.3390/molecules25071728.

Lian, X., Wei, M.-Y., Ma, Q., 2019. Nanomedicines for near-infrared fluorescent lifetime-based bioimaging. Front. Bioeng. Biotechnol. 7, 386. https://doi.org/10.3389/fbioe.2019.00386.

Gao, F., Zhang, R., Feng, X., Liu, S., Ding, R., Kishor, R., Qiu, L., Zheng, Y., 2017. Phase-domain photoacoustic sensing. Appl. Phys. Lett. 110 (3), 033701. https://doi.org/10.1063/1.4974326.

Attia, A.B.E., Balasundaram, G., Moothanchery, M., Dinish, U.S., Bi, R., Ntziachristos, V., Olivo, M., 2019. A review of clinical photoacoustic imaging: current and future trends. Photoacoustics 16, 100144. https://doi.org/10.1016/j.pacs.2019.100144.

Yu, L., Sun, J., Lv, X., Feng, Q., He, H., Zhang, B., Ding, Y., Liu, Q., 2019. High-contrast photoacoustic imaging through scattering media using correlation detection of adaptive time window. Sci. Rep. 9 (1), 17262. https://doi.org/10.1038/s41598-019-53990-w.

Zhou, J., Jokerst, J.V., 2020. Photoacoustic imaging with fiber optic technology: a review. Photoacoustics 20, 10021. https://doi.org/10.1016/j.pacs.2020.100211.

Hosseinaee, Z., Le, M., Bell, K., Reza, P.H., 2020. Towards non-contact photoacoustic imaging [review]. Photoacoustics 20, 100207. https://doi.org/10.1016/j.pacs.2020.100207.

Park, S., Villa, U., Brooks, F.J., Su, R., Oraevsky, A.A., Anastasio, M., 2021. Three-dimensional quantitative functional optoacoustic tomography to estimate vascular blood oxygenation of the breast. Photons Plus Ultrasound: Imaging and Sensing 2021. Proceedings of SPIE 11642, SPIE Bios 2021, 116421N. https://doi.org/10.1117/12.2585341.

Xia, J., Yao, J., Wang, L.H.V., 2014. Photoacoustic tomography: principles and advances. Prog. Electromagn. Res. 147, 1–22. https://doi.org/10.2528/PIER14032303.

Wang, L.V., Yao, J., 2016. A practical guide to photoacoustic tomography in the life sciences. Nat. Methods 13 (8), 627–638. https://doi.org/10.1038/nmeth.3925.

Kolkman, R.G.M., Steenbergen, W., van Leeuwen, T.G., 2006. In vivo photoacoustic imaging of blood vessels with a pulsed laser diode. Lasers Med. Sci. 21 (3), 134–139. https://doi.org/10.1007/s10103-006-0384-z.

Shu, W., Ai, M., Salcudean, T., Rohling, R., Abolmaesumi, P., Tang, S., 2015. Image registration for limited-view photoacoustic imaging using two linear array transducers. Photons Plus Ultrasound: Imaging and Sensing, Proceedings of SPIE 9323, Photonics West 2015, 932348. doi: 10.1117/12.2077740.

Upputuri, P.K., Pramanik, M., 2015. Performance characterization of low-cost, high-speed, portable pulsed laser diode photoacoustic tomography (PLD-PAT) system. Biomed. Opt. Express 6 (10), 4118–4129. https://doi.org/10.1364/BOE.6.004118.

Erfanzadeh, M., Kumavor, P.D., Zhub, Q., 2017. Laser scanning laser diode photoacoustic microscopy system. Photoacoustics 9, 1–9. https://doi.org/10.1016/j.pacs.2017.10.001.

Yao, Q., Ding, Y., Liu, G., Zeng, L., 2017. Low-cost photoacoustic imaging systems based on laser diode and light-emitting diode excitation. J. Innovat. Opt. Health Sci. 10 (4), 1730003. https://doi.org/10.1142/S1793545817300038.

Upputuri, P.K., Pramanik, M., 2018. Fast photoacoustic imaging systems using pulsed laser diodes: a review. Biomed. Eng. Lett. 8 (2), 167–181. https://doi.org/10.1007/s13534-018-0060-9.

Leggio, L., Wiśniowski, B., Gawali, S.B., Rodríguez, S., Sánchez, M., Gallego, D., Carpintero, G., Lamela, H., 2017. "Multi-wavelength photoacoustic system based on

high-power diode laser bars," Photons Plus Ultrasound: Imaging and Sensing, Proceedings of the SPIE 10064, Photonics West 2017, 1006441. doi: 10.1117/12.2254871.

Yu, H., Zhao, X., Wu, X., Zou, Y., Ma, X., Jin, L., Xu, Y., Zhang, H., 2017. Integrated beam shaping and polarization beam combining design for fiber-coupled semiconductor laser stacks system. Appl. Opt. 56 (34), 9510–9514. https://doi.org/10.1364/AO.56.009510.

Lin, G., Zhao, P., Dong, Z., Lin, X., 2020. Beam-shaping technique for fiber-coupled diode laser system by homogenizing the beam quality of two laser diode stacks. Opt. Laser Technol. 123, 105919. https://doi.org/10.1016/j.optlastec.2019.105919.

Sullins, K., 2002. Diode laser and endoscopic laser surgery. Vet. Clin. N. Am. Small Anim. Pract. 32 (3), 639–648. https://doi.org/10.1016/s0195-5616(02)00013-x.

Polese, L., La Raja, C., Fasolato, S., Frigo, A.C., Angeli, P., Merigliano, S., 2020. Endoscopic diode laser therapy for gastric hyperplastic polyps in cirrhotic patients. Lasers Med. Sci.. https://doi.org/10.1007/s10103-020-03127-7.

Gawali, S.B., Leggio, L., Sánchez, M., Rodríguez, S., Dadrasnia, E., Gallego, D.C., Lamela, H., 2016. Combining high power diode lasers using fiber bundles for beam delivery in optoacoustic endoscopy applications. Semiconductor Lasers and Laser Dynamics VII, Proceedings of the SPIE 9892, Photonics Europe 2016, 98921W. doi: 10.1117/12.2227697.

Diehl, R., 2000. High-power Diode Lasers: Fundamentals, Technology, Applications. Springer. https://doi.org/10.1007/3-540-47852-3.

Yu, J., Guo, L., Wu, H., Wang, Z., Gao, S., Wu, D., 2016. Optimization of beam transformation system for laser-diode bars. Opt. Express 24 (17), 19728–19735. https://doi.org/10.1364/OE.24.019728.

Kalva, S.K., Upputuri, P.K., Austria Dienzo, R., Pramanik, M., 2019. Pulsed laser diode based photoacoustic tomography system using multiple acoustic reflector based single element ultrasound transducers. Photons Plus Ultrasound: Imaging and Sensing 2019, Proceedings of the SPIE 10878, Photonics West 2019, 1087831. doi: 10.1117/12.2508281.

Kalva, S.K., Upputuri, P.K., Pramanik, M., 2019. High-speed, low-cost, pulsed-laser-diode-based second-generation desktop photoacoustic tomography system. Opt. Lett. 44 (1), 81–84. https://doi.org/10.1364/OL.44.000081.

Upputuri, P.K., Pramanik, M., 2015. Pulsed laser diode based optoacoustic imaging of biological tissues. Biomedical Physics & Engineering Express 1 (4). https://doi.org/10.1088/2057-1976/1/4/045010.

Shabairou, N., Lengenfelder, B., Hohmann, M., Klämpfl, F., Schmidt, M., Zalevsky, Z., 2020. All-optical, an ultra-thin endoscopic photoacoustic sensor using multi-mode fiber. Sci. Rep. 10, 9142. https://doi.org/10.1038/s41598-020-66076-9.

Ji, X., Xiong, K., Yang, S., Xing, D., 2015. Intravascular confocal photoacoustic endoscope with dual-element ultrasonic transducer. Opt. Express 23 (7), 9130–9136. https://doi.org/10.1364/OE.23.009130.

www.newport.com/resourceListing/technical-notes.

Svelto, O., Hanna, D.C., 2013. Principles of Lasers. Springer.

Loosen, Peter, Knitsch, Alexander, 2007. In: Springer Series in Optical Sciences High Power Diode Lasers. Springer New York, New York, NY, pp. 121–179.

Identifying and Locating Connection Fault of Layer Winding Turn in Distribution Transformer

Ajaya Kumar swain, *Department of Electrical and Electronics Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ajayaswain9@gmail.com*

Prativa Barik, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, p.barik213@gmail.com*

Romeo Jena, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, romeo_jena2@hotmail.com*

Pranay Rout, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pranayrout93@gmail.com*

Abstract *– Transformers are considered as the most important and expensive distribution and transfer networks of electrical energy. Among different fault identifying in transformers, identifying winding fault is not easily recognizable because of lower domain effect in terminal voltages and currents. In this article, frequency response analysis method is used as an efficient method to recognize turn's connection fault. By comparing frequency response in fault and intact conditions, fault recognizing in winding becomes possible. In order to determine frequency response, the described winding model is used. Analyzing the model is done by MATLAB software. The accuracy of this model is very dependent on determining its parameters. In order to exact calculation of parameters of described winding model, Finite Element Method based on winding design information is used and in order to increase accuracy, parameter dependency to frequency is considered. Finally, the effect of turn's connection fault and its location on layer winding of a distribution transformer is evaluated.*

Keywords*: Frequency Response, Finite Element Method, Turn's Connection Fault, Locating*

INTRODUCTION

Power transformers are considered as the most important and expensive elements of distribution and transfer networks of electrical energy and hence by the increase of demand for intact and safe electrical energy, avoiding fault occurrence in power transformers especially faults which result in transformer fails, became more important for network beneficiaries. Statistical studies show that 70-80 % of power transformers' fails come from inside faults [1-2].

An estimation of transformer faults is shown in figure (1), which shows 10 % occurrence of total winding faults. Among these faults, winding turn is challenging for monitoring and identifying, especially in lower domains of fault current. Usually, this fault begins with a turn connection in winding and gradually result in phase to phase short-circuit fault to earth.

Since various methods are presented and investigated in order to identifying and locating short-circuit faults of wining in power transformers. Some of these methods are based on laboratory work and some other are based on modeling. In each of these methods, an index is used in order to fault identifying.

Differential relays are one of identifying methods for inside faults. In this method the difference between initial and secondary phase currents are monitored continuously as a parameter to identify fault [3].

DGA is one of general fault detection methods which is based on analyzing solution gases. In this method solution gases are analyzed in oil so fault or normal performance of transformer is recognized [3].

Wavelet and neural network transform are among methods that can be used in order to fault detection of transformer [2] and [4-7].

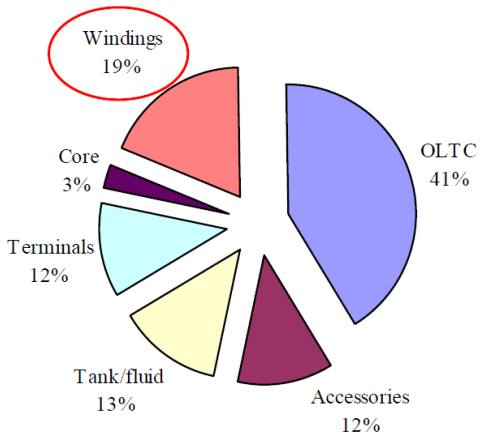


Figure 1- Different fault percentage in transformer [3]

In another method the base of symmetric components' theory or in more exact words sequence currents, is negative. The existence of significant amount of negative sequence currents in transformer's terminal quantities, by itself is a sign of emergence of a disturbance or an index that show non-symmetry that is coupled with this under investigation transformer [4].

In [8] differential currents are calculated from three phases of turns and measurements' ratios. The set of differential currents are transformed by transforming the park into d-q components. The curves among d and q prove fault of turn connection.

Online analysis of transformer's leakage flux can be used as an efficient method to evaluate accuracy of the machine and recognize the existence of insulating failures during its primary steps. The base of this method is based on recognizing changes in leakage flux because of insulating failure [9].

According to high sensitivity of frequency response to short-circuitfault of wining, among appropriate methods in order to testing winding turn connection, there is wining's frequency response analyze (FRA) [3]. This method is based on this principle that each winding of transformer has its unique transform function and frequency response which is sensitive to changes that are performed in the structure of winding such as resistance, inductance and capacitive changes that finally result in internal faults in transformer. So, identification of winding transformation, winding connection and winding movement and core is possible by this method. By comparing frequency response in fault and intact condition there is the possibility of identification of fault in winding.

Dick and Erven initially demonstrated this method in 1978 [13] for detecting winding deformation under the influence of short circuits. The response in FRA can be categorized broadly as low-frequency, medium-frequency, and high-frequency response. Under the condition of direct short-circuited turns, magnetizing characteristics of the core are changed which change low-frequency response in FRA. Medium-frequency response in FRA gets affected mainly due to the mechanical movement of the winding, e.g., winding deformation, buckling, etc. Localized winding damage causes seemingly random changes in the high-frequency response in FRA [14]. Thus, during inter-turn fault in transformer, low- and high-frequency responses are significant in FRA [15]. In the conventional FRA method, impedance spectrum over the range of frequencies is analyzed using Discrete Fourier Transform (DFT). However, it is mostly observed that the low- and medium-frequency components are insufficient while analyzing using DFT. Thus, the advanced techniques such as synthetic spectral analysis (SSA) based on cut-and-concatenation (CCM) method are used. After spectral analysis of the current set and the reference set, various diagnostic criterions such as sum-squared-error, correlation coefficient, sum squared ratio error, sum squared max-min ratio error, and absolute sum of logarithmic error (ASLE) can be used to determine the fault in the transformer. The combination of SSA (based on CCM) and ASLE has been proved to be most pertinent criterion [16]. Different winding structures give different frequency response in FRA. Winding-to- winding interaction between two different phases and delta connected winding arrangement also affect the frequency response [17]. Although the FRA is generally used to detect the electrical and mechanical faults in windings, its applicability for detection of core fault can also be observed in the literature [18]. The core parameters, such as magnetic permeability, conductivity, and magnetizing impedance, can also be obtained at high frequencies with the help of FRA [19]. However, results obtained from FRA applied to detect the winding faults are not independent of core magnetization. The governing factor in this phenomenon is magnetic viscosity, which is defined as the time dependence of magnetization under a constant magnetic field. The impedance measurement, mainly below 10 kHz, is observed to be significantly dependant on DC magnetization, instances when power supply switches off, and demagnetization [20].

Although, FRA is well-known and popular method in fault diagnostics, this method requires additional sophisticated instruments for the detection. Also, the prediction about the operating condition from the complex admittance-signature is not straightforward and always needs an expert's opinion or evidential reasoning (ER) approach [21].

In this article an efficient method is used to identification and locating turn connection fault. For this reason layer winding of a transformer of a sample distribution is evaluated. To obtain winding frequency response of transformer, first winding is being modeled by described model and parameters are obtained from finite element method, then by the help of MATLAB software winding frequency response is obtained and by comparing fault and intact conditions, turn connection fault is identified and finally fault location is investigated with an appropriate index.

Modeling transformer's winding

The model is being used to describe transformer in frequency finite higher than 10 KHz of described model that is among physical methods of modeling winding. Figure 2 shows the descriptive equivalent turn of transformer. Each unite of this circuit is a turn of winding that includes self and against inductances with other turns. For each turn, capacitor is considered against other turns and with the earth and also ohmic resistance, series with self-inductance. Insulating losses between each two turns are also considered [10]. The smallest component of a winding is its turns. Discs result from being placed on top of each other in radius direction and their connection in vertical direction results in layers of winding.

In order to obtain frequency response, the analysis of equivalent circuit is done in node method and in frequency field. The base of this method is to form admittance matrix from circuit elements. Node equations and admittance matrix are expressed in Equation 1.

$$Y.V = I \quad (1)$$

Y, Admittance matrix, is a $2n \times 2n$ matrix that n is the number of winding turns. The vector of nodes' voltage and the current of inductor branches are considered as below.

$$V = [V_1, V_2, ..., V_n, I_{L1}, I_{L2}, ..., I_{Ln}]^T_{2n\times 1} \quad (2)$$

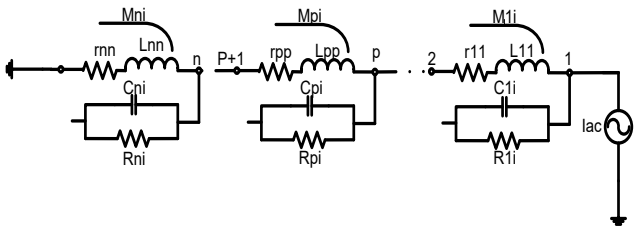


Figure 2- equal circuit of n turns by using described model [10]

Stimulation current vector is expressed as below that I_t is circuit stimulation current.

$$I = [I_t, 0, ..., 0, 0]^T \quad (3)$$

Admittance matrix elements of the node are determined as follow that can be different from frequency.

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \quad (4)$$

Where:

$$Y_{11} = [Cs] + [G(s)], Y_{22} = [Ls] + [R(s)] \quad (5)$$

$$Y_{21} = \begin{bmatrix} -1 & 1 & 0 & . & 0 \\ 0 & -1 & 1 & . & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & . & 0 & -1 \end{bmatrix}, Y_{12} = \begin{bmatrix} 1 & 0 & . & 0 & 0 \\ -1 & 1 & . & 0 & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & -1 & 1 \end{bmatrix} \quad (6)$$

By using equation (7) node's voltage is obtained:

$$V = [Y]^{-1}.[I] \quad (7)$$

To calculate input impedance, winding has allocated value 1 to the first line of stimulation vector in various frequencies and the values of other lines are equal to zero. In this situation by solving equation (7), the value of node 1 is equal to winding's input impedance. By placing s=jω, the first line of vector V is a mixed number which its size is impedance value and its phase shows inductor or capacitive property in the frequency. Among various windings the order of nodes to each other is different and naturally parameters of the circuit are different in various windings.

Characteristics of transformer and winding

This transformer is a real transformer that is designed and built by Iran University of science and technology, Academic center of education, culture and research. The modeling of transformer and its geometric dimensions are presented in Figures (3) and (4). High-voltage winding belongs to a dry distribution transformer $760^v / 380^v, 10kVA, Dy, U_k = 4\%$.

This winding is composed of 594 turns that are wrapped 6 layers around the core. In each turn one copper conductor with diameter of 1 mm is used. In each turn a green paper is used. According to the existing information we consider the dielectric coefficient of green paper equal to 3, the amount of dielectric coefficient of the insulator equal to 2.8 and the dielectric coefficient for the fiber equal to 4.6.

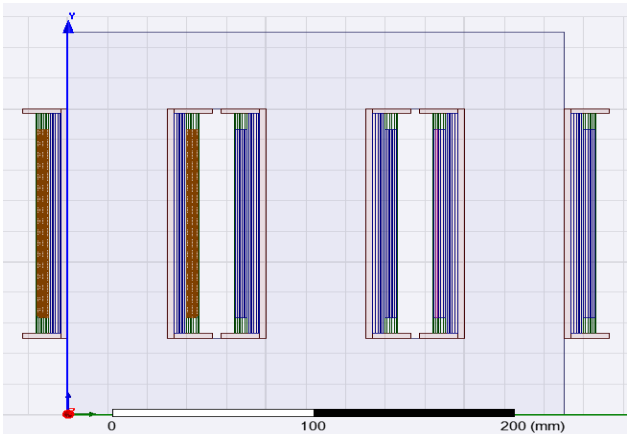


Figure 3- The under investigation transformer

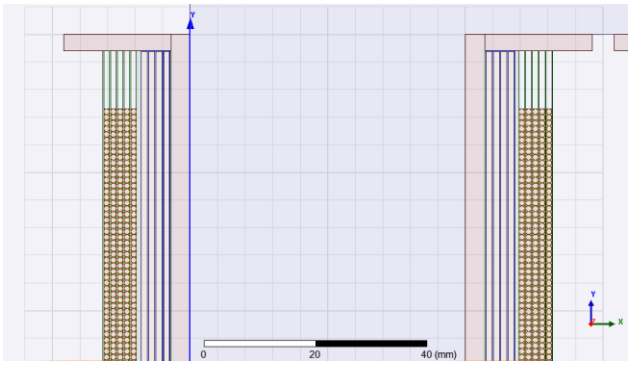


Figure 4 - A view of the under investigation winding

Calculation of descriptive model's parameters

The ability of described model in reconstruction of winding's fluctuation behavior is dependent on the accuracy of calculation of model's parameters. In this article, in order to increase accuracy, finite element method is used to determine winding parameters. For this reason finite element software, MAXWELL 2D, is used.

Inductance calculation: Self-induction related to magnetic flux involves winding which the amount of this flux depends on the density of magnetic flux (B), on the other hand the density of magnetic flux depends on the magnetic permeability of the material used in winding space. The aim of this article is to investigate high

frequency behavior ($10^4 - 10^7 Hz$) of transformer winding. In this frequency range, eddy currents are so much that discharge field lines completely out of the core. The center of the core is practically without flux. Filed lines do not enter to the core, but they block their path of air. Iron core behaves like an empty cylinder that there is current in its body. Therefore, we can model the set of iron core and low-voltage winding by an empty cylinder [11]. So the effect of core is not taken into consideration. By dimension determination of turns and identification of materials, finite elements software calculates self and against inductance based on the following equation:

$$L_{ij} = \frac{4W_{ij}}{I^2_{peak}} \quad (8)$$

Where W_{ij} is medium energy which is calculated via field calculations. This software considers the peak current amount for each turn as 1 Ampere, thus the inductance is simplified as $4W_{ij}$. Table 1 shows the amounts of a 6 layer inductance.

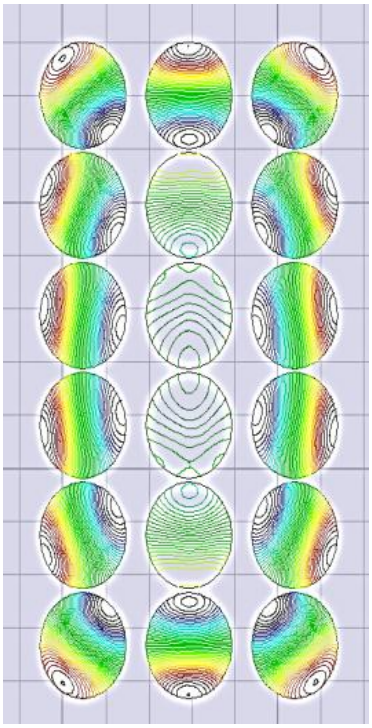
TABLE 1
SELF AND AGAINST INDUCTANCE MATRIX (MH) OF 6 TURNS OF TRANSFORMER WINDING

Turns number	1	2	3	4	5	6
1	L11=2.047	L21=1.854	L31=1.724	L41=1.647	L51=1.591	L61=1.548
2	L12=1.854	L22=2.072	L32=1.859	L42=1.726	L52=1.647	L62=1.591
3	L13=1.724	L23=1.859	L33=2.067	L43=1.864	L53=1.729	L63=1.649
4	L14=1.647	L24=1.726	L34=1.864	L44=2.062	L54=1.865	L64=1.73
5	L15=1.591	L25=1.647	L35=1.729	L45=1.865	L55=2.072	L65=1.862
6	L16=1.548	L26=1.591	L36=1.649	L46=1.73	L56=1.862	L66=2.076
7	L17=1.513	L271=1.548	L37=1.592	L47=1.65	L57=1.728	L67=1.863

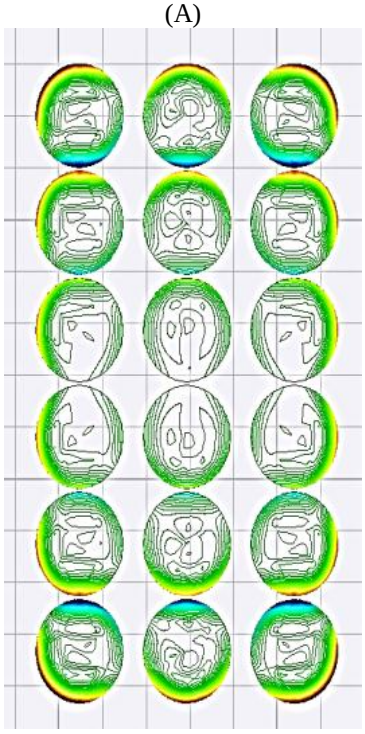
Eddy currents in conductors are because of changing fields with time and we should involve their effect in calculation of inductance matrix, of course, considering this effect in analytic relations is very difficult. The resulted eddy current in conductors cause that the conductor center becomes current free and this results, in the increase of current density in conductor surface, so inductance value decreases. But this effect is more significant on self-inductance value [12].

Figure 5 shows the effect of dermal and proximity influence on current density in several turns in one and thousand KHz frequencies. This element influences the amount of self and against inductances of winding. The amount of inductance of a turn is appropriate to its radius. The decreasing rate of inductance of various turns is approximately the same. In frequency area that inductance amount reaches to its final limit, has relationship with conducting coefficient and conducting transient coefficient.

If there is an investigation in time field, analyzing changing elements with time is very difficult. Therefore, if we can have a good fitting from changing curve of self and against inductance according to frequency, circuit analysis is very useful in frequency field.



(A)



(B)

Figure 5 - Current distribution in some turns of the winding. A: 1KHz B: 1000KHz

The most proper curve which is in harmony with self-inductance changing can be expressed by the following formula.

$$L(f) = ae^{(bf)} + ce^{(df)} \quad (9)$$

So the amounts of inductance are calculated by finite elements in various frequencies and then we can express inductance changes with curve fitting method mathematically. Against inductance changes are few, for this reason we omit against inductance changes than frequency in the performed modeling [10]. By fitting the curve by the help of MATLAB software, fitting curve coefficients are calculated as follow.

$$a = 0.0003106, b = -2.935e-8,$$
$$c = 0.0007683, d = -3.571e-13$$

Capacity calculation: Capacities between various elements of windings and core in high frequencies play an important role. The method of distribution of primary voltage on winding is determined by internal capacitive network of the transformer. As the previous part, we determine the capacity amount by modeling in finite elements software. Changes of dielectric coefficient do not make problems in our analysis and usually frequency behavior of dielectric coefficient is not identified. Table 2 shows the amounts of 5 turn capacity.

Resistance calculation

Resistances damp the fluctuations done in inductor and capacitive complex network of winding. It is obvious that without modeling system's resistants we cannot distinguish unstable fluctuations from stable fluctuations. In order to have an accurate resistant modeling, its changes according to frequency are also considered. Resistant increase occur because of the increase of frequency stimulation source due to the movement of current from conducting center to its surface Figure 5, which obtaining its approximate relations analytically is very complicated. By using measurement or finite element method, we can find ohmic resistant amount for each required frequencies.

The method of finding resistant change function according to frequency in order to analyze frequency is very useful. In order to find ohmic resistant fitting according to frequency, the best curve is the exponential curve which is shown below.

$$(10)\ R(f) = ae^{(bf)} + ce^{(df)}$$

In table (3) resistant change of one turn of transformer winding is shown in various frequencies. Fitting coefficients by using MATLAB software are as follow.

$$a = 1.426, b = 4.605e-12,$$
$$c = -1.287, d = -1.84e-8$$

TABLE 2

CAPACITY MATRIX (PF)OF FIVE TURNS OF TRANSFORMER WINDING USING MAXWELL2D SOFTWARE

Turns number	1	2	3	4	5
1	C11=403.7	C21=-262.27	C31=-15.98	C41=-9.6689	C51=-6.6206
2	C12=-262.27	C22=611.1	C32=-254.04	C42=-11.516	C52=-6.875
3	C13=-15.98	C23=-254.04	C33=612.39	C43=-252.43	C53=-10.82
4	C14=-9.6689	C24=-11.516	C34=-252.43	C44=614.22	C54=-253.62
5	C15=-6.6206	C25=-6.875	C35=-10.82	C45=-253.62	C55=615.59

TABLE 3

THE AMOUNTS OF THE RESISTANT (OHM) OF ONE TURN OF TRANSFORMER WINDING IN VARIOUS FREQUENCIES

f	1	2	3	4	5	6	7	8	9
R	0.029008	0.029011	0.029014	0.029015	0.029016	0.029017	0.029017	0.029018	0.029018
f	10	20	30	40	50	60	70	80	90
R	0.029018	0.029019	0.029019	0.029019	0.029019	0.02902	0.02902	0.029021	0.029021
f	100	200	300	400	500	600	700	800	900
R	0.029021	0.029025	0.02903	0.029034	0.029039	0.029044	0.02905	0.029056	0.029063
f	1000	2000	3000	4000	5000	6000	7000	8000	9000
R	0.02907	0.029169	0.029321	0.029527	0.029787	0.030101	0.030467	0.030885	0.031353

Calculation of insulating conduction: Because of ideal insulators, their electrical conduction coefficient is against zero. This causes losses in transformer's insulator. By considering a resistant, parallel to capacitor, we can enter insulator losses in the equations. Insulator losses coefficient for capacitor – resistant parallel to capacitor- is as follow.

$$\tan\delta = \frac{1}{R_p C_p \omega} \quad (11)$$

So by having known insulator losses efficient of one winding unite, we can calculate unknown parallel resistant in the described model. This amount depends on frequency and the properties of insulator loss coefficient impregnated with oil. The amount of insulating loss coefficients for compounds of insulators that exist in one winding unite is estimated through measurement.

$$\tan\delta = 1.082\times10^{-8}\omega + 5.0\times10^{-3} \quad (12)$$

For frequencies of multiple thousand KHz, fixed amount of 0.01 is used as insulating loss coefficient of a winding unite [10].

Model implementation

To solve the described model of transformer, we can get help from circuit solving softwares. To analyze circuit, MATLAB software is used. In order to obtain frequency response a program is written in the context of this software based on node method in frequency field. By using obtained parameters from finite element method and program in MATLAB, input impedance of transformer winding is obtained in various situations. Figure 6 shows frequency response of transformer winding in normal situations.

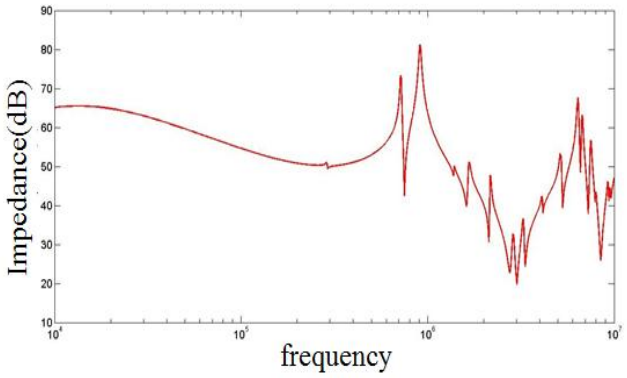


Figure 6- Input impedance of transformer winding (dB) to frequency

Turn-connection modeling

Turn connection occur because of losing insulator between conductors. Qualitatively, we can identify the turn connection equal to increase of insulating conductor between two turns which we can use this method for modeling turn connection.

To form connection matrix, there has been occurred a connection between nodes I and j, with resistant of r_f, we can make elements (j,j), (I,i) of connection matrix equal to $1/r_f$, and make (j,i), (I,j) equal to $-1/r_f$ and the rest of elements equal to zero. Connection matrix is from conduction and it should be coupled with conduction matrix in node method.

Figure 7 shows input impedance of intact winding with stretch line and turn connection with dashed lines. As we can see, turn connection results in domain decrease of input impedance peak and its shift to the right of frequency axis.

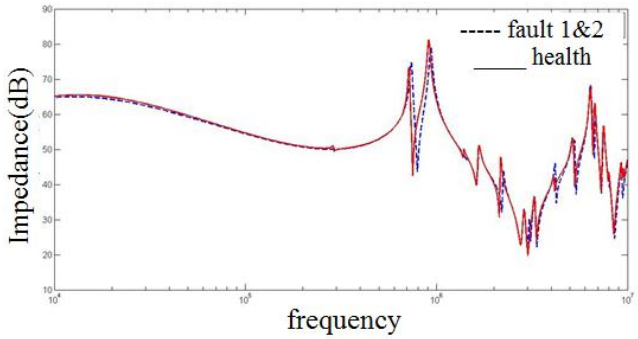


Figure 7- Input impedance of winding (dB) in intact condition and turn connection between the first and second rounds to the frequency

Location of turn connection

Figure 8 shows frequency responses in the existence of turn connection in various places in the top half of the first layer of winding. You can see that the location of turn connection is effective on frequency response.

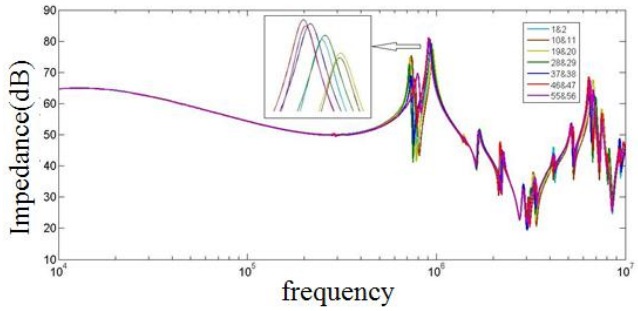


Figure 8- Frequency response of winding with the existence of various faults in the first layer

As it is obvious in Figure 8 the amount of deviations and domain changes depend on fault location. However these changes are very low and there will be obtained no information from visual comparison, so we get help from indexes in order to better comparison. In this step resonance points are studied. Absolute amounts of relative difference in the domain (and frequency) between resonance points have the maximum domain that are chosen as index. This reasoning is obvious in Figure (9).

Following equations express the way of calculation of index.

$$\frac{Df_i}{f_i} = [\frac{f_{o,i} - f_{k,i}}{f_{o,i}}] \qquad (13)$$

$$\frac{DA_i}{A_i} = [\frac{A_{o,i} - A_{k,i}}{A_{o,i}}] \qquad (14)$$

To investigate the under investigation transformer winding in the first layer, 14 parts turn connection winding were done between two turns. The results are in Figures 10-12.

As it is obvious in these figures, obtained indexes are symmetric to winding center. This issue reveals the existence of symmetry between bottom-half of transformer and its top half.

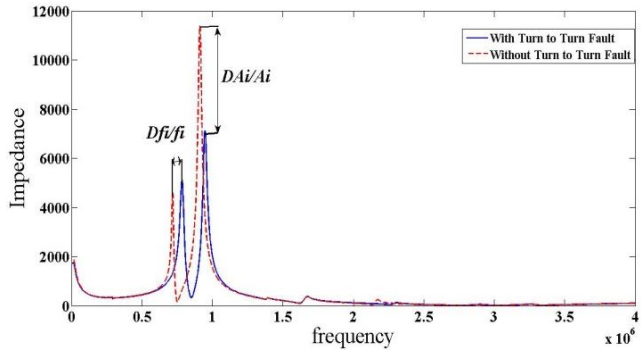


Figure 9- The effect of short circuit on transform function

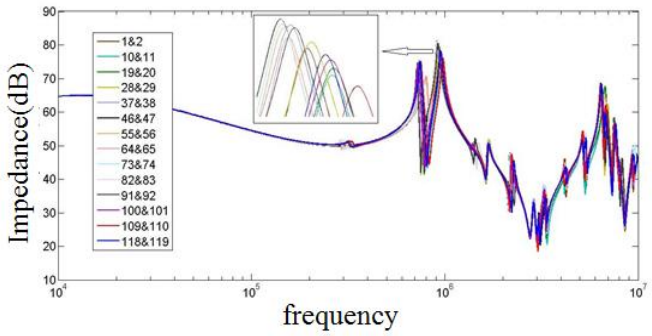


Figure 10- Input impedance of winding (dB) in the existence of fault in various locations of the first layer to frequency

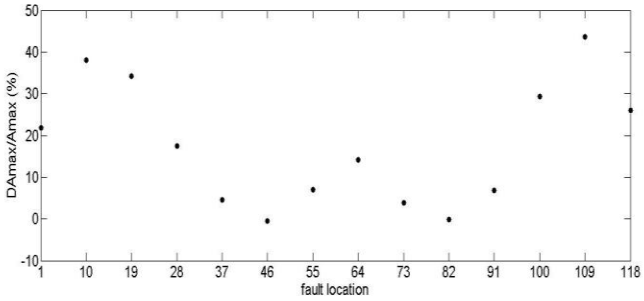


Figure 11- Points dependency to input impedance domain to fault location in the first layer

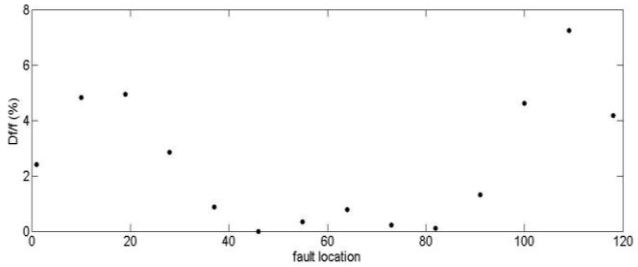


Figure 12- Frequency dependence of maximum points of input impedance to fault location in the first layer

We can also see this symmetry for other layers respectively in Figures 13-20.

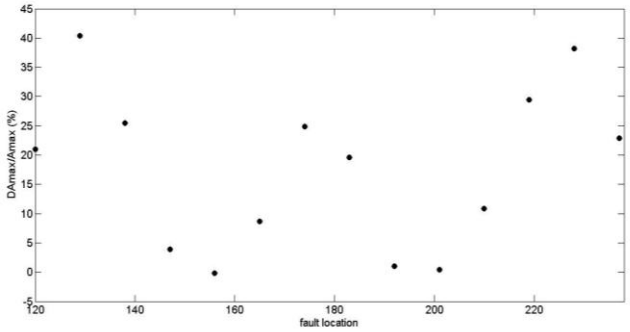


Figure 13- Dependency of maximum points of input impedance to fault location in the second layer

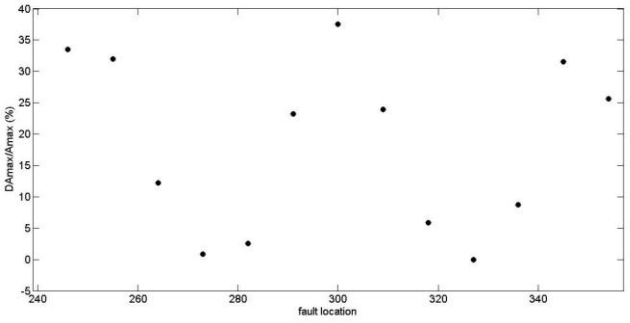


Figure 14- Dependency of maximum points of input impedance to fault location in the third layer

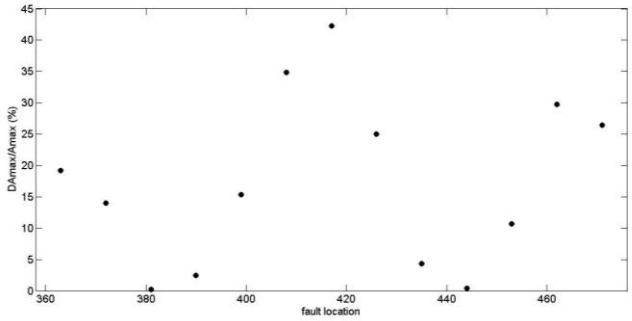


Figure 15- Dependency of maximum points of input impedance to fault location in the fourth layer

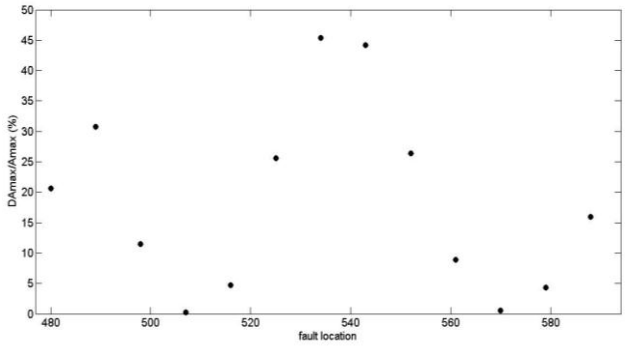


Figure 16- Dependency of maximum points of input impedance to fault location in the fifth layer

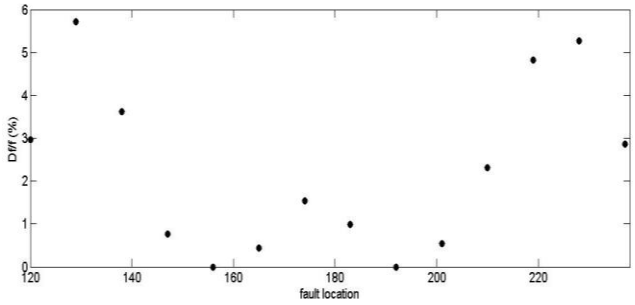


Figure 17- Frequency dependence of maximum points of input impedance to fault location in the second layer.

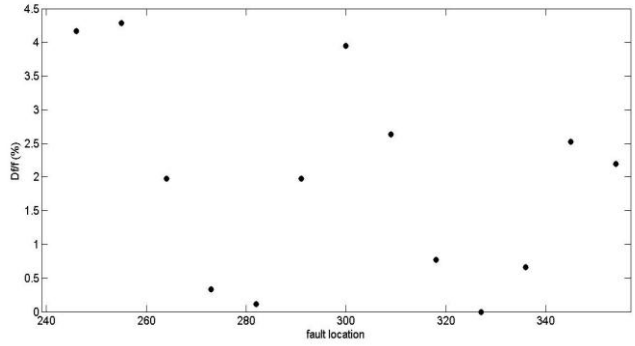


Figure 18- Frequency dependence of maximum points of input impedance to fault location in the third layer

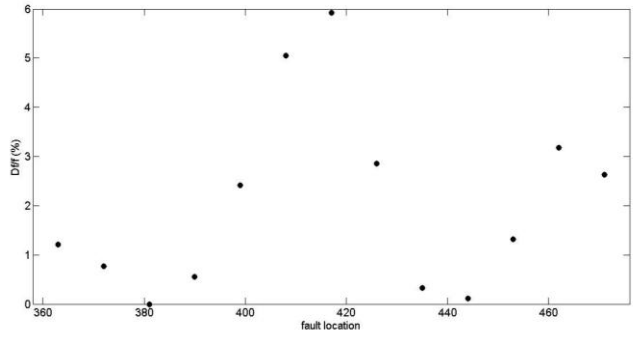


Figure 19- Frequency dependence of maximum points of input impedance to fault location in the fourth layer

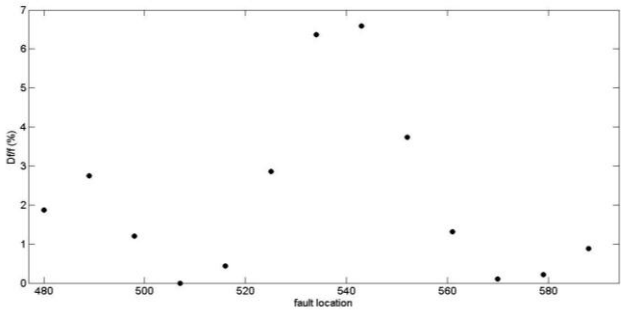


Figure 20- Frequency dependence of maximum points of input impedance to fault location in the fifth layer

Now we focus on fault location change on various layers. Figures (21) and (22) show the indexes for connecting a turn above a winding according to layer number.

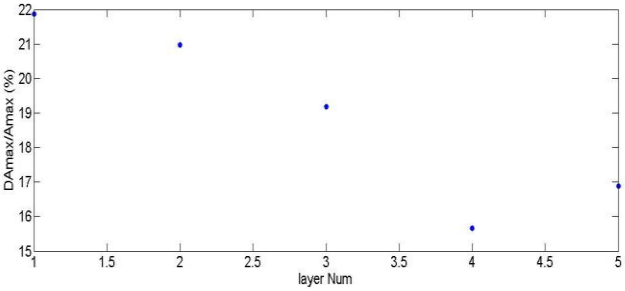


Figure 21- Dependency of maximum points of input impedance according to layer number

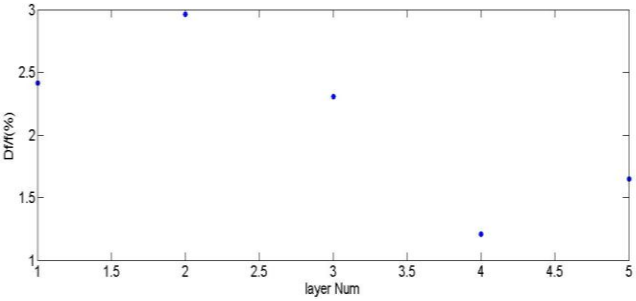


Figure 22- Frequency dependence of maximum points of input impedance according to layer number.

It is seen that fault location on various layers effects on the form of frequency response.

CONCLUSION

In this article frequency response method is used in order to identification and determination of fault location of turn connection. In order to modeling winding, the described model is used. The accuracy of this model is very dependent on its parameter determination. In order to calculate parameters of the described model of winding transformer, finite element method is used which has a good accuracy. By comparing intact and fault mode of

frequency response, in resonance points of frequency response, it is obtained that fault mode causes domain change and frequency shift of severe points. The obtained responses in this article have proved the sensitivity of frequency response to fault location in winding. The results of creating connection in various locations of winding show that we have symmetry between top half transformer and bottom half transformer.

REFERENCES

[1] N. Y. Abed and O. A. Mohammed, "Modeling and characterization of transformers internal faults using finite element and discrete wavelet transforms," IEEE Transactions On Magnetics, Vol. 43, No. 4, April 2007.

[2] O .Ozgonenel, E .Kilic, D .Thomas and A .E . Ozdemir, "Identification of transformer internal faults by using an RBF network based on dynamical principle component analysis," Powereng 2007, April 12-14, Setubal, Portugal, 2007.

[3] R. S. Bhide, M. S. S. Srinivas, A. Banerjee and R. Somakumar, "Analysis of Winding Inter-turn Fault in Transformer: A Review and Transformer Models," IEEE ICSET 6-9, Kandy, Sri Lanka, Dec 2010.

[4] Vahid behjat, "online monitoring of power transformers for detection of winding short circuit faults using finite element analysis", PHD thesis, 2010.

[5] Ch. P. Babu1, M. Surya kalavathi, and B.P.Singh, "Use of wavelet and neural network (BPFN) for transformer fault diagnosis," IEEE Ann. Rep. Conf. Elect. Ins. Diel. Pheno., pp. 93-96, 2006.

[6] H. Firoozi, M. Kharezi, and H. Bakhshi, "Turn-to-Turn Fault Localization of Power Transformers Using Neural Network Techniques," Int. Conf. Pro. Appl. Diel. Mat. B-11, pp. 249- 252, jul 2009.

[7] A .Ngaopitakkul and A .Kunakorn, "Internal fault classification in transformer windings using combination of discrete wavelet transforms and back-propagation neural networks," International Journal of Control, Automation, and Systems, vol. 4, no. 3, pp. 365-371, June 2006.

[8] Luís. M. R. Oliveira and A. J. Marques Cardoso, "A permeance-based transformer model and its application to winding interturn arcing fault studies," IEEE Transactions on Power Delivery, Vol. 25, No. 3, July 2010.

[9] F. Cabanas, G. Melero, F .Pedrayes, H. Rojas, A. Orcajo, M. Cano, G. Iglesias, and F .Nuño, "A new online method based on leakage flux analysis for the early detection and location of insulating failures in power transformers: application to remote condition monitoring," IEEE Transactions On Power Delivery, Vol. 22, No. 3, July 2007.

[10] H. Mobaraki, A. Vahedi, "Turn on Turn Fault's Effect on Inverted Winding's FRA of Power Transformer", The International Conference for Inductive & Electromagnetic Component (CWIEME) Inductica, Mumbai, India, November 2009.

[11] E. Mombello, K. Moller, "Impedances for the calculation of electromagnetic transient phenomena and resonance in transformer windings," Electric Power Systems Research, no. 54 pp. 131-138, 2000.

[12] E.E. Mombello, "A novel linear equivalent circuit of a transformer winding considering the frequency-dependence of the impedances," Electric Power Systems Research, pp. 885–895 2007.

[13] E. P. Dick and C. C. Erven, "Transformer diagnostic testing by frequency respponse analysis," IEEE Trans. Pwr. App. Syst, vol. PAS- 97, no. 6, nov/dec. 1978, pp. 2144-2153.

[14] S. A. Ryder, "Diagnosing transformer faults using frequency response analysis," IEEE Elec. Ins.Magz. vol. 19, no. 2, mar/apr 2003, pp. 16-22.

[15] M. Florkowski and J. Fargul, "A high-frequency method for determining winding faults in transformers and electrical machines," Rev. Sci. Instrum., vol. 76, nov., 2005, pp. 114701-1-114701-6.

[16] J. W Kim, B. Park, S. C. Jeong, S. W. Kim, and PG Park,, "Fault diagnosis of a power transformer using an improved frequency-response analysis," IEEE Trans. Pwr. Deliv. vol. 20, no. 1, jan 2005.

[17] Z. Wang, Jie Li and D. M. Sofian, "Interpretation of Transformer FRA Responses—Part I: Influence of Winding Structure," IEEE Trans. Pwr. Deliv. vol. 24, no. 2, apr. 2009, pp. 703-710.

[18] J. Pleite, C. González, J. Vázquez, and A. Lázaro, "Power Transfomer Core Fault Diagnosis Using Frequency Response Analysis," IEEE MELECON, Benalmádena (Málaga), Spain, may. 16-19, pp. 1126-1129.

[19] A. Shintemirov, W. H. Tang, and Q. H. Wu, "Transformer Core Parameter Identification Using Frequency Response Analysis," IEEE Trans. Magn. vol. 46, no.1. jan 2010, pp. 141-149.

[20] N. Abeywickrama,Y. V. Serdyuk, and Stanislaw M. Gubanski, "Effect of Core Magnetization on Frequency Response Analysis (FRA) of Power Transformers," IEEE Trans. Pwr. Deliv. vol. 23, no. 3, jul. 2008, pp. 1432-1438.

[21] A. Shintemirov, W.H. Tang, and Q.H. Wu, "Transformer winding condition assessment using frequency response analysis and evidential reasoning ," IET Elec. Pwr. Appl. vol. 4, no. 3, 2010, pp. 198-212.

Optimal Design of Bearingless Permanent Magnet-Type Synchronous Motors for Generating Maximum Levitation Force

Srinivas, *Department of Electrical and Electronics Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, srinivas225@gmail.com*

Alekha Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, alekha.sahoo241@gmail.com*

Subhrajit Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, subhrajitsahoo23@yahoo.co.in*

Subhendu Sahoo, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, s.sahoo95@outlook.com*

Abstract *– One maintenance task that still exist with conventional motors, are bearing lubrication and renewal. Bearingless motors are replaced with conventional motor that uses a magnetic levitation force to suspend a rotor without any mechanical contact. In bearingless motors, additional windings are wound together with motor windings in stator slots. In this paper, a bearingless permanent magnet-type synchronous motor (BPMSM) Has been studied. First, the generation of radial levitation forces is discussed and then the optimum permanent magnet thickness is determined to produce maximum levitation force. After that the effect of additional winding pole-pair in the amount of levitation force is investigated. The simulation is done in Maxwell software.*

Keywords*: Bearingless Permanent Magnet Synchronous Motor, Maximum Levitation Force, Optimization, Thickness of PM.*

INTRODUCTION

In some applications, bearing maintenance is still a significant problem. For example, the bearings can present a major problem in motor drive applications in outer space, and also in harsh environments with radiation and poisonous substances. In addition, lubrication oil cannot be used in high vacuum, ultra high and low temperature atmospheres and food and pharmacy processes. In these cases magnetic bearing is used. Magnetic bearing is a machine element uses a magnetic levitation force to suspend a rotor. The characteristics of magnetic bearing are no friction, no wear, no lubrication, high speed, high precision, and long operating life. The applications of magnetic bearing are in canned pumps and drives, compact pumps, high-speed flywheel storage system, artificial hearts, Spindle drives and Semi-conductor processing. But magnetic bearing cause long axial length of the rotor shaft and complicated structure and large size of the motor [1-5].

For overcoming these problems, bearingless motors are proposed. Bearingless motor is an electric motor that combines the functions of both torque generation and magnetic suspension together in a single motor. A bearingless motor have two kinds of windings. The conventional windings and suspension winding are wound together in the stator of motor to produce torque and radial suspension force simultaneously. In comparison with magnetic bearing, bearingless motor have Simple structure, higher speed, compactness and lower cost [6-8].

So far, various bearingless motors have been proposed but bearingless permanent magnet synchronous machine (BPMSM) due to its advantages such as simple structure, high efficiency, high torque density, are actively researched and developed around the world. The aim of this paper is to determine the thickness of permanent magnet and to consider the number of suspension windings in order to produce maximum levitation force in BPMSM [9].

MATERIAL AND METHODS

Principles of radial force generation

Figure 1 shows the cross section of a primitive bearingless motor under different conditions. In Figure 1(a), there is a symmetrical 4-pole flux distribution. The solid curves illustrate the flux paths circulating around the four conductors 4a, these conductors are located in the stator slots. The 4-pole flux wave produces airgap poles in the order N, S, N and S in the airgap sections 1, 2, 3 and 4 respectively.

Since the flux distribution is symmetrical, the flux density magnitudes in airgap sections 1, 2, 3 and 4 are of the same value at the same point in the pole section. There are attractive magnetic forces between the rotor poles and stator iron. The amplitudes of these attractive radial forces are the same, but the directions are equally distributed so that the sum of radial force acting on the rotor is zero.

Figure 1(b) shows the principle of radial force generation. Two conductors 2a are located in the stator slots. With the current direction as shown in the figure, a 2-pole flux wave is generated. In airgap section 1, the flux density is increased because the direction of the 4-pole and 2-pole fluxes is the same. However, in airgap section 3, the flux density is decreased because the direction of these fluxes is opposite. The magnetic forces in the airgap sections 1 and 3 are no longer equal, i.e., the force in airgap 1 is larger than in airgap 3. Hence a radial force F

results in the x-axis direction. It follows that the amplitude of the radial force increases as the current value in conductors 2a increases. Figure 1(c) shows how a negative radial force in the x-axis direction is generated. The current in conductors 2a is reversed so that the flux density in airgap section 1 now decreases while that in airgap section 3 increases. Hence the magnetic force in airgap section 3 is larger than that in airgap section 1, producing a radial force in the negative x-axis direction.

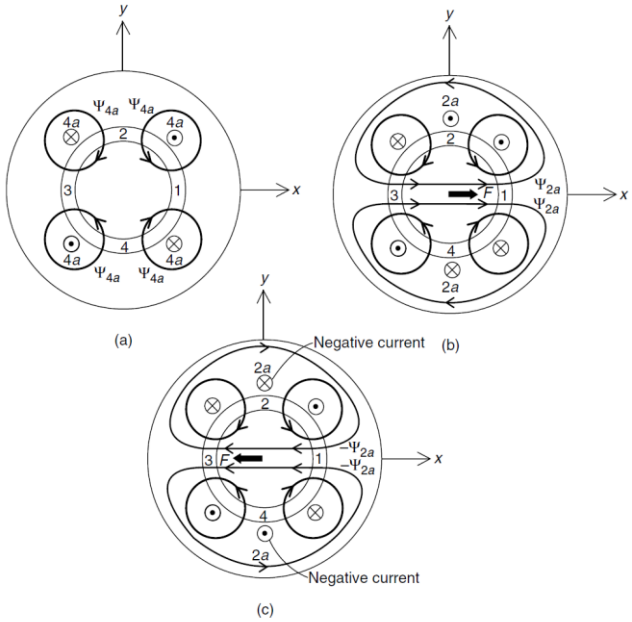


Figure 1. Principles of radial force generation: (a) 4-pole symmetrical flux; (b) x-direction radial force; (c) negative x-direction force [11].

Figure 2 shows radial force generation in the y-axis direction. Two conductors 2b, which have an MMF centred on the y-axis, are added to the stator. A similar flux density imbalance occurs but this time between airgap sections 4 and 2, hence producing a force on the y-axis. The polarity of the current will dictate the direction of the force. These are the principles of radial force generation in x- and y-axis directions. The force values are almost proportional to the current in conductors 2a and 2b (assuming constant 4-pole current). The vector sum of these two perpendicular radial forces can produce a radial force in any desired direction and with any amplitude [11-13].

In order to generate the torque and radial suspension force simultaneously in a bearingless machine the following conditions are

$$\begin{cases} P_M = P_B \pm 1 \\ \omega_M = \omega_B \end{cases} \quad (1)$$

Where and are the pole-pair number and current frequency of the torque winding, and are the pole-pair number and current frequency of the suspension force winding.

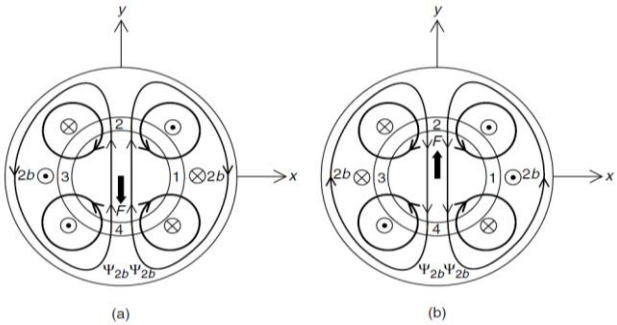


Figure 2. Y-Direction radial force: (a) negative y-direction radial force; (b) y-direction radial force [12].

Figure 3 shows cross-sectional view of a 3-phase bearingless permanent magnet-type synchronous motor. The torque and suspension force windings are wound together in the same stator slots of the BPMSM. Coil sides 4u, 4v and 4w are for the torque windings of phase u, v and w and Coil-sides 2u, 2v and 2w are for suspension force windings of phase u, v and w, respectively. In the figure the torque winding and suspension force winding are 4-pole and 2-pole respectively [14, 15].

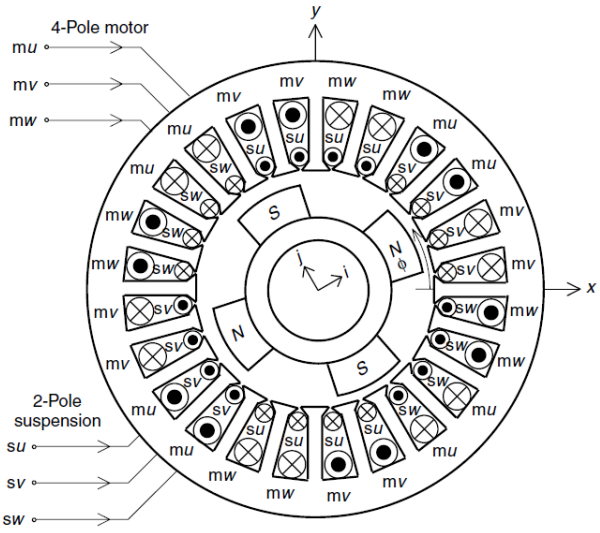


Figure 3. Bearingless permanent magnet bearingless motor [13].

Radial levitation force equation

In this paper, two-phase model of BPMSM is used for simplicity. All of the variables are in the synchronous rotating reference frame.

The relationship between the radial levitation forces and the currents of suspension force winding can be expressed as

$$\begin{cases} F_{ix} = (K_M \pm K_L).(i_{2d}\psi_{1d} + i_{2q}\psi_{1q}) \\ F_{iy} = (K_M \pm K_L).(i_{2q}\psi_{1d} - i_{2d}\psi_{1q}) \end{cases} \quad (2)$$

F_{ix}, F_{iy} are the radial levitation forces. K_M, K_L is Maxwell and Lorentz forces constant. i_{2d}, i_{2q} are current components of radial levitation windings. ψ_{1d}, ψ_{1q} are the airgap flux linkages components of motor windings. The generated Maxwell force F_{sx}, F_{sy} are proportional to the displacement and can be written as

$$\begin{cases} F_{sx} = K_s x \\ F_{sy} = K_s y \end{cases} \quad \text{where} \quad K_s = k\frac{\pi r l B^2}{\mu_0 \delta} \quad (3)$$

Where is the force-displacement Coefficient, is the free space permeability and is the airgap length. So the radial levitation force and can be expressed as

$$\begin{cases} F_x = F_{ix} + F_{sx} \\ F_y = F_{iy} + F_{sy} \end{cases} \quad (4)$$

Substituting (1) , (2) into (3), equation (3) can be written as

$$\begin{cases} F_x = (K_M \pm K_L).(i_{2d}\psi_{1d} + i_{2q}\psi_{1q}) + K_s x \\ F_y = (K_L \pm K_M).(i_{2q}\psi_{1d} - i_{2d}\psi_{1q}) + K_s y \end{cases} \quad (5)$$

When $P_B = P_M + 1$, (4) can be written as

$$\begin{cases} F_x = (K_M + K_L).(i_{2d}\psi_{1d} + i_{2q}\psi_{1q}) + K_s x \\ F_y = (K_L + K_M).(i_{2q}\psi_{1d} - i_{2d}\psi_{1q}) + K_s y \end{cases} \quad (6)$$

When $P_B = P_M - 1$, (4) can be written as

$$\begin{cases} F_x = (K_M - K_L).(i_{2d}\psi_{1d} + i_{2q}\psi_{1q}) + K_s x \\ F_y = (K_L - K_M).(i_{2q}\psi_{1d} - i_{2d}\psi_{1q}) + K_s y \end{cases} \quad (7)$$

The stator flux linkage equation is as follows

$$\begin{cases} \psi_{1d} = L_d i_{1d} + \psi_r \\ \psi_{1q} = L_q i_{1q} \end{cases} \quad (8)$$

Where ψ_r is rotor flux linkages L_d and L_q are the self-inductance of motor Windings [7].

Demagnetization of permanent magnet in BPMSM

In the surface-mounted permanent magnet machines, thin permanent magnets with small airgap can generate the radial levitated forces more effectively. However, thin permanent magnets can simply demagnetize. Thus, it is very important to consider demagnetization of permanent magnets. Moreover irreversible demagnetization of the permanent magnets is a more serious problem in these motors than in conventional electric motors.

Figure 4 shows an example with q-axis motor current and suspension current flux paths. The rotor angular position is 0 deg. When the q-axis motor flux is generated as shown, torque is generated in the counter-clockwise direction. In this case, goes through the permanent magnets denoted as A, B, C and D in the opposite direction to their magnetization. In addition, the suspension flux is shown. Note that both and go through permanent magnet D against the magnetization so that D is the most critical permanent magnet. The q-axis motor flux is synchronously rotating with the rotor but the suspension flux is rotating with a frequency twice that of the rotor, i.e., is not synchronized to the revolving rotor magnetic field. Therefore, not only D but also A, B and C will experience the same degree of demagnetization at some point with the possibility of irreversible demagnetization.

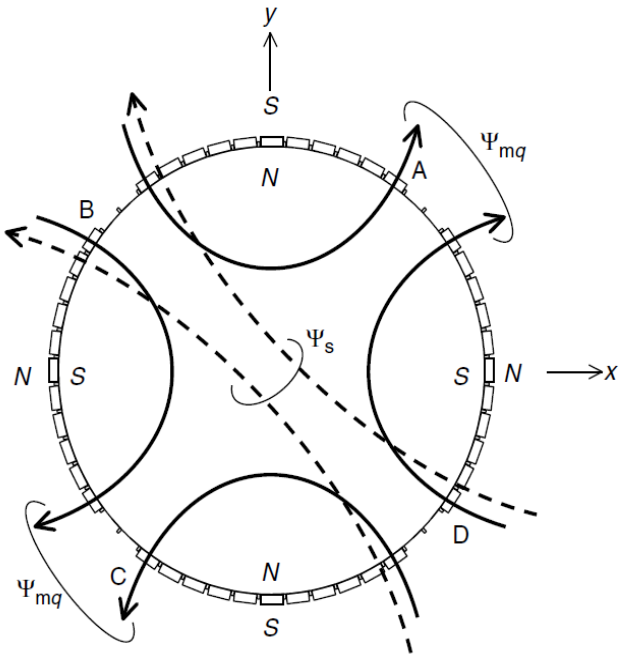


Figure 4. Position of permanent magnet with possibility of irreversible demagnetization [10].

RESULTS

Results and Simulink of model in Maxwell

In this section, FEA is employed to determine the optimum thickness of permanent magnet and number of pole-pair of suspension windings to produce maximum levitation force. The specification of surface-mounted permanent magnet type bearingless motor is shown in Table 1.

Table 1. Specification of BPMSM

Pole pair of torque winding	4
Pole pair of suspension force winding	6
The outer diameter of stator (mm)	155
The inner diameter of stator (mm)	98
The outer diameter of rotor (mm)	88
Axial length of machine (mm)	105
Air-gap length (mm)	4
Residual flux density of PM (NdFeB) (T)	1.28

This motor has 4-pole torque winding and 6-pole suspension force windings. Figure 5 shows the surface mounted permanent magnet bearingless motor in Maxwell software. Since the machine has symmetric flux distribution just one fourth of the machine is sketched in Maxwell software. In the figure, torque and suspension force windings are showed. Mesh diagram produced by finite element analysis is also shown in Figure 6.

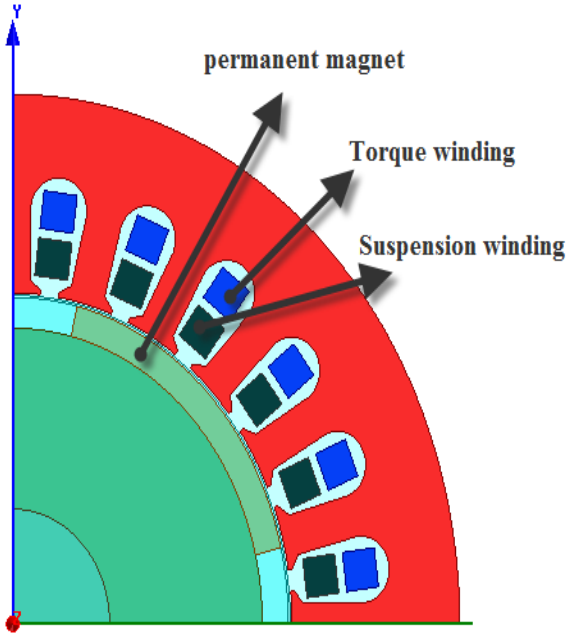


Figure 5. Surface mounted permanent magnet bearingless motor.

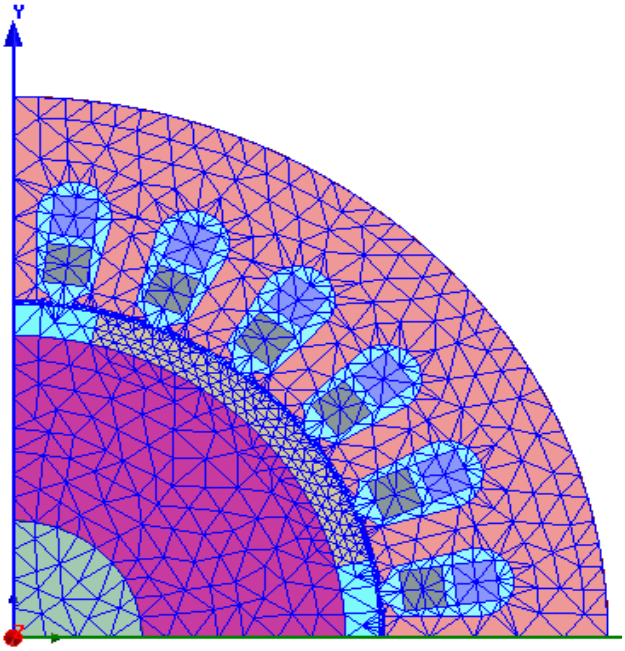


Figure 6. Mesh produced by FEA.

As mentioned earlier, in a permanent magnet motor with fixed airgap length, there is an optimal thickness to generate maximum radial levitation force. The variations of radial levitation force corresponding to the thickness of permanent magnets is analysed in Maxwell software. The results are showen in table (II) and figure 7. In table (II) for different value of PM, the radial levitation force is obtained. It can be seen form figure 7 that levitation force changes corresponding to the thickness of permanent magnets when the length of airgap is fixed. In this paper, the airgap length of motor is 4 mm. As permanent magnet thickness is 1.8 mm, the maximum levitation force is generated (45 N).

Table 2. Thickness of permanent magnet and corresponding levitation force

Thickness of PM (mm)	Radial levitation force (N)
0	0
0.5	9.89
1	28.91
1.5	42.78
1.8	45
2	44.55
2.5	43.25
3	38.23
3.5	35.21
3.8	34.25

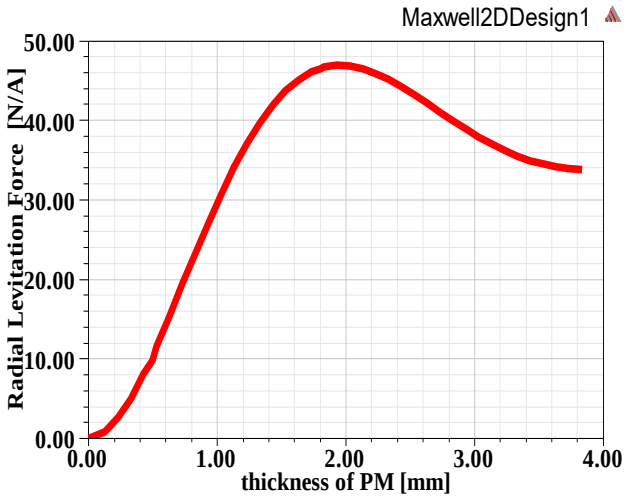


Figure 7. Relationships of radial force and thickness of Permanent magnet.

Now, with this thickness of PM, we must check whether the PM demagnetized or not. Therefore flux density on the surfaces of permanent magnets is derived in Maxwell software and shown in figure 8. It can be seen that the minimum flux density is more than zero, so there is no demagnetization.

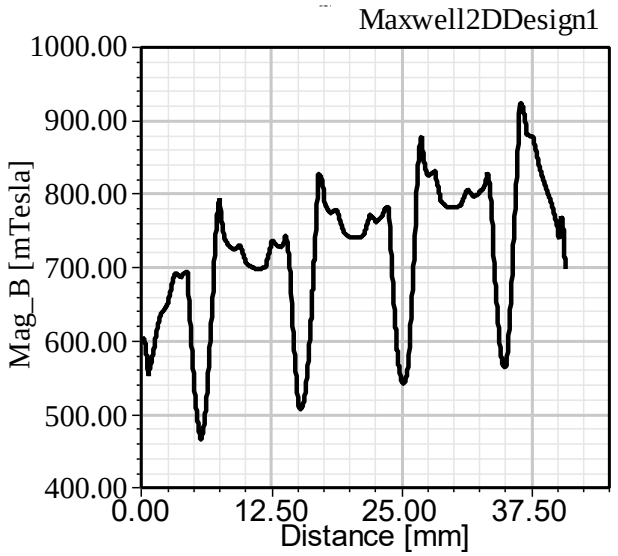


Figure 8. Flux densities on the surfaces of permanent magnet.

In next step, the effect of different pole-pair of suspension winding in the amount of radial levitation force are analyzed. As discussed earlier, for generating levitation force, the number of pole-pair of suspension winding is one more or less than the number of pole-pair of torque windings. So if the pole-pair of torque winding is two, the number of pole-pair of suspension winding is one or three. The effect of different pole-pair of suspension winding in radial levitated force is done in Maxwell software. Results are shown in Figure 9.

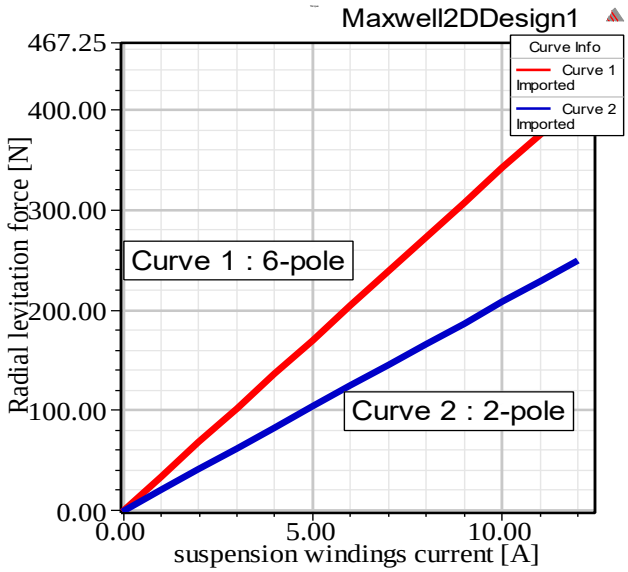


Figure 9. Relationships of the radial levitation force and suspension winding current.

Figure 7 show relation of suspension winding current and radial levitation force with different pole pairs of suspension winding current. Curve 1 of figure 7 shows the relationship between radial levitation forces and 6-pole suspension windings current. Curve 2 of figure 7 shows the relationship between radial levitation force and 2-pole suspension windings current. As shown in Figure 7, it can be seen that radial levitation forces generated by 6-pole suspension windings are larger than that of 2-pole suspension windings under the same pole-pair of torque winding.

CONCLUSION

The aim of this paper is to design a surface mounted permanent magnet-type bearingless motor to produce maximum radial levitation force. First the effect of permanent magnet thickness was evaluated. The optimal permanent magnet thickness is derived 1.8 mm. Second the effect of pole-pair of suspension winding on radial levitation force was evaluated. The result shows that with 6-pole suspension winding the levitation force is greater than that motor with 2-pole suspension winding. So by considering these two notes in design, in addition to increase the radial levitation force, the size of motor is also reduced.

REFERENCES

[1] T. Schuhmann, W. Hofmann, and R. Werner, “Improving operational performance of active magnetic bearings using Kalman filter and state feedback control,” IEEE Trans. Ind. Electron., vol. 59, no. 2, pp. 821–829, Feb. 2013.

[2] Y. Ren and J. Fang, “Current-sensing resistor design to include current derivative in PWM H-bridge unipolar switching power amplifiers for magnetic bearings,” IEEE Trans. Ind. Electron., vol. 59, no. 12, pp. 4590–4600, Dec. 2012.

[3] J. Fang and Y. Ren, “Self-adaptive phase-lead compensation based on unsymmetrical current sampling resistance network for magnetic bearing switching power amplifiers,” IEEE Trans. Ind. Electron., vol. 59, no. 2, pp. 1218–1227, Feb. 2012.

[4] T. Chiba, Fukao, O. Ichikawa, M. Oshima, M. Takemoto, and D. G. Dorrell, Magnetic Bearings and Bearingless Drives. Amsterdam, the Netherlands: Elsevier, Mar. 2005.

[5] T. Reichert, T. Nussbaumer, and J. W. Kolar, “Bearingless 300 W PMSM for bioreactor mixing,” IEEE Trans. Ind. Electron., vol. 59, no. 3, pp. 1376–1388, Mar. 2012.

[6] J. Asama, Y. Hamasaki, T. Oiwa, and A. Chiba, “Proposal and analysis of a novel single-drive bearingless motor,” IEEE Trans. Ind. Electron., vol. 60, no. 1, pp. 129–138, Jan. 2013.

[7] X. Wang, Q. Zhong, Z. Deng, and S. Yue, “Current-controlled multiphase slice permanent magnetic bearingless motors with open-circuited phases: Fault-tolerant controllability and its verification,” IEEE Trans. Ind. Electron., vol. 59, no. 5, pp. 2059–2072, May 2012.

[8] R. Warberger, Kaelin, T. Nussbaumer, and J. W. Kolar, "50 N • m/ 2500 W bearingless motor for high-purity pharmaceutical mixing," IEEE Trans. Ind. Electron., vol. 59, no. 5, pp. 2236–2247, May 2012.

[9] Li and W. Hofmann, "Speed regulation technique of one bearingless 8/6 switched reluctance motor with simpler single winding structure," IEEE Trans. Ind. Electron., vol. 59, no. 6, pp. 2592–2600, Jun. 2012.

[10] H. Grabner, W. Amrhein, S. Silber, and W. Gruber, "Nonlinear feedback control of a bearingless brushless DC motor," IEEE/ASME Trans. Mechatronics, vol. 15, no. 1, pp. 40–47, Feb. 2010.

[11] T. Schneeberger, T. Nussbaumer, and J. W. Kolar, "Magnetically levitated homopolar hollow-shaft motor," IEEE/ASME Trans. Mechatronics, vol. 15, no. 1, pp. 97–107, Feb. 2010.

[12] L. S. Stephens and K. Dae-Gon, "Force and torque characteristics for a slotless Lorentz self-bearing servomotor," IEEE Trans. Magn., vol. 38 ,no. 4, pp. 1764–1773, Jul. 2002.

[13] M. T. Bartholet, T. Nussbaumer, and J.W. Kolar, "Comparison of voltagesource inverter topologies for two-phase bearingless slice motors," IEEE Trans. Ind. Electron., vol. 58, no. 5, pp. 1921–1925, May 2011.

[14] N. Quang Dich and S. Ueno, "Analysis and control of nonsalient permanent magnet axial gap self-bearing motor," IEEE Trans. Ind. Electron., vol. 58, no. 7, pp. 2644–2652, Jul. 2011.

[15] F. Rodriguez and J. A. Santisteban, "An improved control system for a split winding bearingless induction motor," IEEE Trans. Ind. Electron., vol. 58, no. 8, pp. 3401–3408, Aug. 2011.

Current Measurement with Optical Current Transformer

Smruti Ranjan Panda, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, sr_panda@outlook.com*

Prakash Chandra Sahu, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, pksahoo88@gmail.com*

Manoj Mohanta, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, manoj.mohanta62@outlook.com*

Pranay Rout, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pranayrout93@gmail.com*

Abstract *–The development of power electrical systems causes attention to accuracy of protection elements and measurement values. So we need some new measuring devices with high accuracy for power transmission lines. One of these devices is Optical Current Transformer (OCT). Optical current transformers are suitable for power system protection and can replace the magnetic current transformers. In this paper, we described a comparison between optical current transformers and conventional current transformers. Using an optical current transformer has several advantages, e.g. high accuracy, low weight, easy installation and no-saturation. But it has some disadvantages also, like if the magnetic fields induced by the currents through other conductors is sufficiently high, the fault current measured will have some errors. In general, the optical current transformer device is more reliable and suitable for new power systems.*

Keywords*: Optical Current Transformer, Fiber Optic, Current Sensor, Protection, Current Transducer*

INTRODUCTION

Two of the most important components of a high voltage substation are CT and PT. Their task is the sampling of voltage and current for use in measurement and protection [15]. We use this equipment since the voltage and current at high voltage substations cannot be directly used for measurement and protection and voltage and current transformers must be used to bring down large amounts. In conventional CT and PT are used the current and voltage transformers that have the primary winding core, secondary winding that convert voltage and current. Resin, oil and gas insulating are used to insulate primary and secondary voltage. For various uses such as the protection and the measurement are used of separate cores. It causes measurement device to become bulky. The information of voltage and current are conducted by a wire to control room and there they are used specially. Conventional CT and PT were and are the best in the market for high voltage substations but their Grandeur decrease too much by introducing optical equipment lately. When with the increasing of Core or Burden for the substations, we have to replace CT due to the limitations of rated current or we want add a relay. We will not be worried due to this new method for this problem, because new CT does not have these problems. Its reason is that they can convert a current from the lowest amount to nearly 4000 ampere and core support by main sensor [12].

The information transition has by optical fiber from the main equipment and it caused a connection at CVT or a disruption at the Beginning CT to terminate that it eliminates the worry of equipment explosion. In many of the new optical CT and PT, Faraday principle was used that it is explained in this paper. In some creative factories, Faraday principle was not used and they use optical fiber for the information transition from equip to control room digitally that the methodology of it is explained in this paper. In years, recently with the improvement in big transition networks, the recognition of connections has to do with the measurement of current and voltage rapidly that it's possible with new methods.

Theory

Faraday found that when a piece of special glass is affected by a strong magnetic field, it becomes active. And optic surface spin when a flat polarized optic forwards through a glass in parallelism with magnetic lines. Since the Faraday's discovery this phenomenon was seen in many solids, liquids and gases. The amount of whirl in each material is proportional to the amount of magnetic field and the distance that an optic go in a material impractically [3,5,6,9,10,11,13,14].

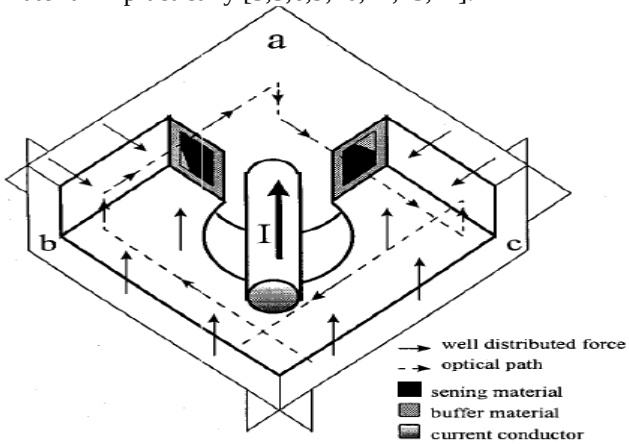


Figure 1. Construction of MOCT

Magnetic - optical or Faraday Effect was reported by Michael Faraday first in 1845 AD [6,10,12,14]. Serious research and development for the application of Faraday Effect on highly accurate flow measurement applications began in the late 1970s. These efforts led to a series of successes in the development of measurement systems using optical sensors [12].

Faraday equations are explained briefly [3,5,8,11]:

$$\theta=\mu VHL \quad (1)$$

θ = Rotation angle of the plane of polarization
μ = Sample absorption Magnetic factor
V = Verdet Constant
H = Field intensity
L = Length of light beam

The distance travelled by light in a glass. Verdet Constant for a particular material represents the intensity of the effect Faraday that it expresses based on the amount of turning on the field intensity unit multiply at the unit of distance. The exact relationship between the magnetic field (H) and electric current (I) depends on the relative geometric position to another. If this relationship expresses to form the coefficient of K so [11,12,14]:

$$\theta=\mu V(KI)L \quad (2)$$

In the design OCT that shown in this figure[3,6,9], the concentrator field concentrator senses the uniform field relatively that Faraday create it so Eq.2 is honest about it, on the other hand, because the light flow a full round around the conductor carrying, Eq.1 become the following form:

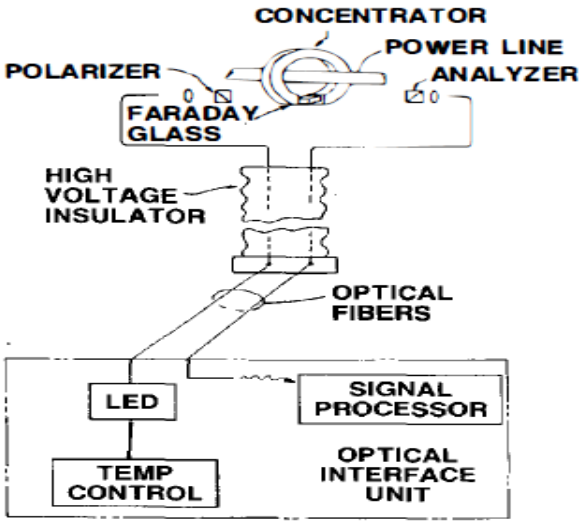


Figure 2. Schematic configuration of optical current measurement [11].

$$\theta=\mu V\O Hdl \quad (3)$$

Eq.3 is written based on amp spin:

$$\theta=\mu VIN \quad (4)$$

The number N is the conductor spins.

Fig.3 shows an example of a Faraday sensor that the axis transfer, subscription and the parser have than to another a 45-degree rotation [3,5,9,10,11,12]. This rotation causes the release light intensity to do modulation in all systems. Meanwhile, light intensity or TD (optical power) on the Manifest is as follows [3,5,16]:

$$TD = (\tau 0/2)(1+\sin 2\theta) \quad (5)$$

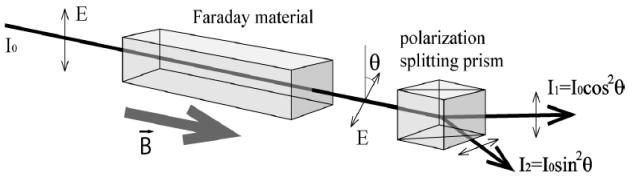


Figure 3. The three key optical elements in the sensor head of an optical current transducer

The Electrical Signal processor admeasurement AC component of waveform (τ0 sin2θ/2) on DC component (τ0/2). So the resulting waveform is independent of the range of light intensity so an output optical signal become as the following form [2,3,11]:

$$m = \sin 2\theta \quad (6)$$

Then the processor combines the two equations (2 and 5) and produces an AC output signal that is proportional directly to the primary flow in a high pressure conductor, for example in an OCT of concentrator plant [11]:

$$I = \sin\text{-}1(m)/2\mu kl \quad (7)$$

In the OCT of circular plan:

$$I = \sin\text{-}1(m)/2\mu VN \quad (8)$$

In all of these relationships m is the measuring quantity and another is constant. The analog processor circuit produces an output that is directly proportionate its domain with pipeline flow and a type of optical CT also use in phase with that. It is expressed theoretically in this paper.

MATERIAL AND METHODS

OCT Optical-CT was proposed by the creators of several methods by using common theory:

1. Conventional CT with optical readout [1, 5]:

In this type, one channel of optical and completely insulated information connects to the output of CT so instead of the typical copper wire is used the optic fiber in

the output data transmission. The methodology of converting the CT output to optical signal form is out of the discussion of this article, but we can say that do not open the output heads of CT is the benefit of this CT that it is the most important factor in the explosion of the CT

Magnetic concentrator (core) with optical measurement [5,11,14]: A magnetic circuit arises around the conductor by ferrous core. The difference with the traditional CT is it that an air gap is generated in the core and magnetic field in the core measure in this air gap with optical instruments. The advantage of this design is that the path of light is short and simple, and smaller optical elements are required.

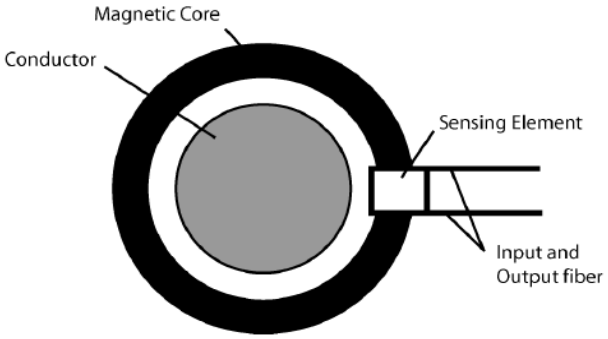


Figure 4A. Schematic of Faraday current sensor using a magnetic concentrator

2. Optical path around the conductor [2,4,5,6]:

If the path is put around the conductor carrying of current that through it the magnetic field effects on the light ray. This optical package path around the conductor measures current similar to normal CT core. In our Division, this is the first plan that is not including the Ferromagnetic component. This type includes two alternatives.

A piece of light-sensitive [5,14]: In this alternative of light path, actually, a piece of optical active materials that one round surrounds around the conductor according to Fig.4.

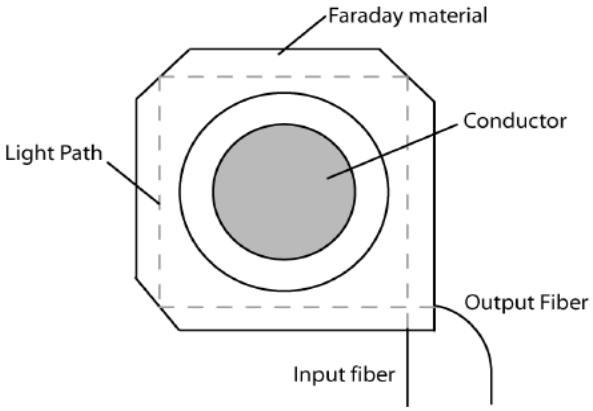


Figure 4B. Schematic of Faraday current sensor using a Bulk Optic

3. Fiber Optics [11,13,14]:

Here the light path around the conductor consists of an optical fiber that it is wrapped to the number of rounds that it required to achieve the desired sensitivity.

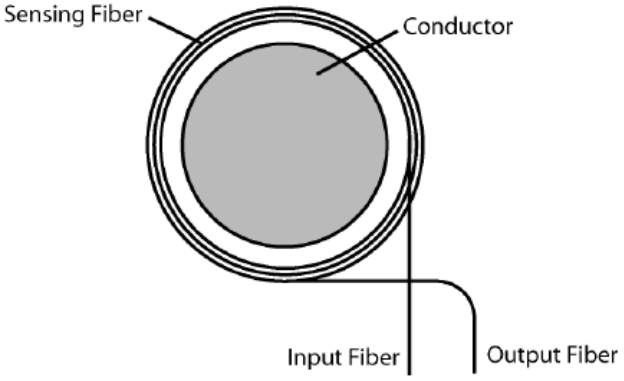


Figure 4C. Fiber Optics Based Current Measurement

4. Witness Sensor [5,14]:

This converter is the latest type in our assortment and it is only type that measurement of it does not include the surrounding of conductor completely. Instead, as shown in Figure 6, the magnetic field at a point closer to the conductor affects on the light distribution.

And therefore, it is not a real current transformer. Although it can be said a field constant distribution around the conductor is a function of its current. It can be said that light with arbitrary polarization is composed of two independent components. In the case of linearly polarized components can be simply said that two components are perpendicular to each other.

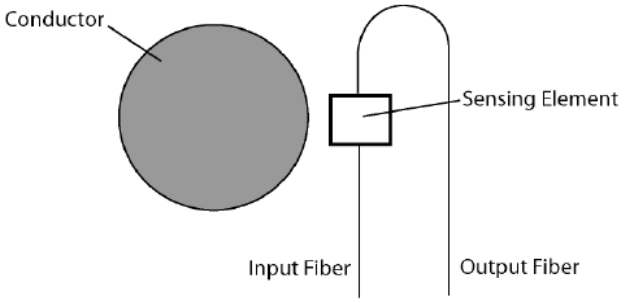


Figure 4D. Schematic of a Faraday-effect sensor, unlinked (Witness) type geometry

In the OCT a light becomes polarized linearly initially. Any type of non-polarized light can be easily polarized light by passing through a polarizer. The next issue is to measure the spin of polarization surface. In fact, it is not directly measurable. Light detectors in any way is not sensitive to the polarization of light, rather they measure the power of incoming waves. In fact, it is proportional to field square. The methods described are too short in this paper:

Method A:

In this method after the exiting of light from the optical sensor, it is entered to the other polarizer that it is called a parser. By placing the light parser at the side and right angle, we can be extracted the range of (value) component linearly polarized in any direction so the amount of light power is important [2,5].

If the angle between the transmission axis of a polarizer and a parser call α and the power of light wave when entering a detector called a pin. Then we have [11]:

PDET is the light power of the detector.

In most designs, angle A is considered the amount of π and can be proved that the measured current is independent of the input light power [3,5].

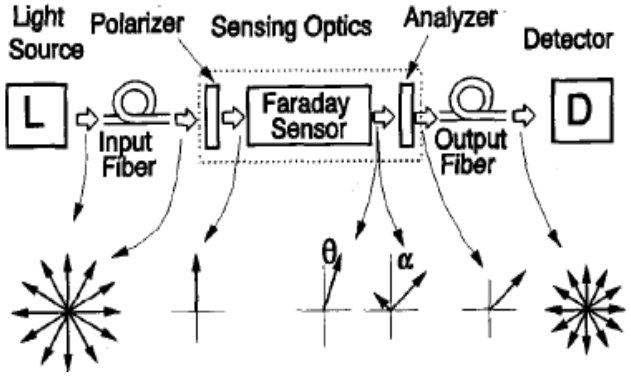


Fig5. Arrangement of Optical Components in Faraday sensor [5]

Method B:

In this method, the output of the parser that they are in the amount of relative angle equivalent π to each other is used. This set are within a single optical element called a polarization splitter. Then outputs subtract from each other and it is divided by the sum. A result is similar to method A; it is not different from method A In terms of sensitivity and failure. But in terms of hardware is a complex plan so it's less used [5,9].

Method C:

Since Faraday materials are not ideal double refractive index so some distortion and noise are created in the route of Sensor due to internal stress or temperature and it causes elliptical polarization finally. And as a result, a complete analysis of output light requires using different angles of polarization and check frequently of output [3,5,7,11].

OPD

OPD optical voltage transformer: Faraday's law can be applied to PTs, like OCTs. Voltage measurement has a fundamental problem and it's that MOCT has been just sensitive to the magnetic field. To create such a flow, connecting a resistor between the line high voltage and ground is the easiest way. Despite the simplicity of this method will cause a great loss, in addition, the resistance depends on the temperature cause accurately measure to lose. The best way is to use a capacitor that has no resistance problems [3,6].

The capacitor I = Cdv/dt so I = -j2πfCv, 90 degree phase difference between current and voltage value is easily compensated by electronic circuits [6]. Fig.2 shows the MOVT.

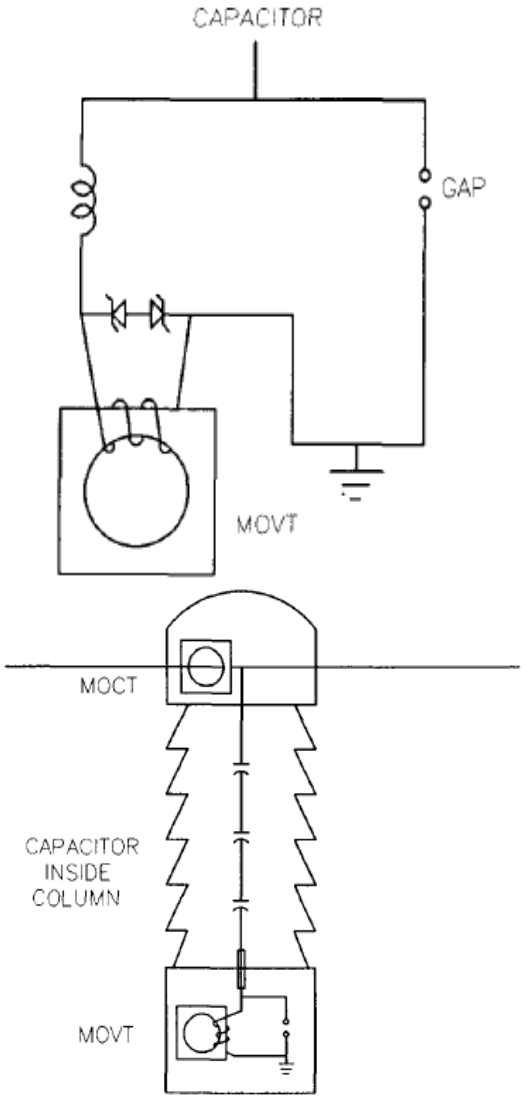


Figure 6. MOVT Schematic & Insulating Column

Calculations show that current through the capacitor is not sufficient that a MOCT be measurable with a spin of current. We know that MOCT like CT with iron Core is sensitive to the number of current carrying wire Spin so with the increased of these numbers, we can create the sufficient magnetic field for glass sensor. Fig.6 shows it.

MOCT can be used together and a MOPT and Base and bushing are put. The digital method can also be used for CVT that it has advantages of DOCT. Fig.7 shows it:

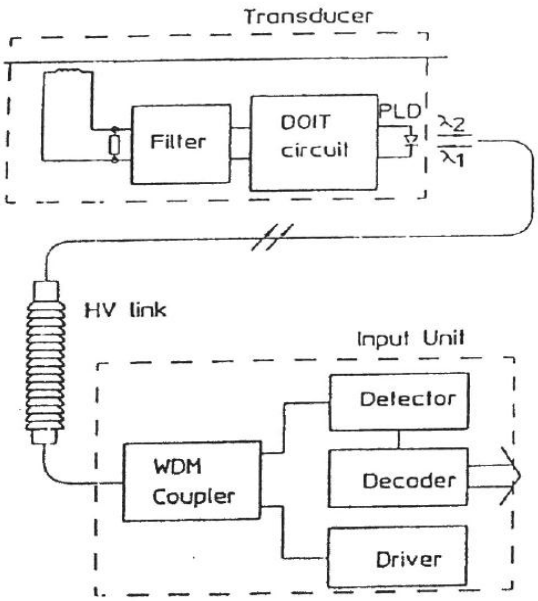


Figure 7A. Digital method for using CVT

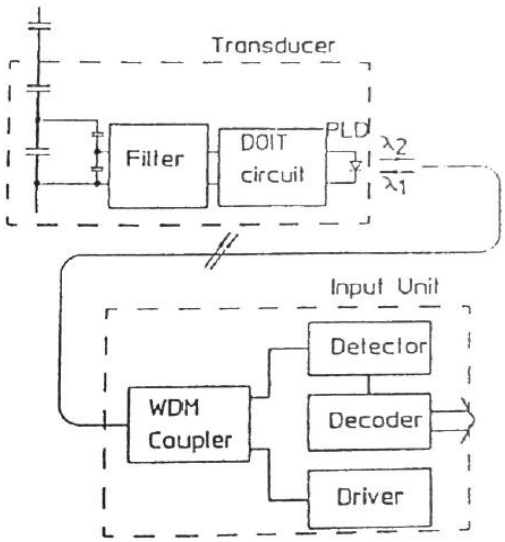


Figure 7B. Digital method for using CVT

Advantages and disadvantages

The use of MOIT and DOIT (Digital Optic Instrument Transformer) has many advantages that they are reasons of Expand its use. The equipment is used in measurement and protection units of substations from the distribution voltage to high voltage.

The use of electronic–Monitoring systems creates measurement, protection with high reliability and wide dynamic range.

A. Advantages:

1. The subject of creating noise is ruled out due to the use of optical fiber [3,5,13,15].

2. Resistance to Acoustic and Electromagnetic parasites is excellent that it plays an important role in protecting [2,14].

3.The internal problems of MOITs and DOITs cause connecting with lines and substations so it causes the traditional equipment to explode very dangerous. In addition, it can damage surrounding equipment [10,14].

4.In terms of size and weight, they have small size and low weight. The size of equipment can be a great help in substations that have land problems and they cause the foundations and structures to remove and they have high costs. The probable installation and relocation does not require heavy machinery that it is huge economy [5,12,13,14].

5.During installation, do not require cutting the insulated conductors. It causes re-insulating to prevent [2,13].

6.Having sufficient electrical insulation resistance [5,14].

7.Lack of magnetic saturated. Due to Lack of core, there is no saturation that it solves many protection and Measurement problems [10,14].

8.Lack of Ferro resonance phenomena and hysteresis [6,10,14].

9.The measurement error is less, than 0.3% [1,2,3,16].

10.The power supply is not used in the HV section in DOITs [8].

11.Self-supervision is possible in DOITs [8].

12.In DOIT and MOIT, the lack of transmission signal attenuation is problematic in conventional CTs.

13.The ability of high mechanical withstands [6].

14.The ability to install common with other equipment [5].

15.Complete isolation in high voltage section and control room [11].

16.Full compatibility the output with computers that will have the most responsibility in the future and new substations [11].

17.The output of conventional CT and PT can support new optical systems (Relays and measuring devices) and conversely.

18.Protection against the opening of output in CTs and shorts PTs [3,15].

19.PCB that is severe environmental pollutants are not used [6,10].

20.The measurement and analysis of current with higher harmonics [5,7,9,12].

21.There are not any Burden and Class [5,12].

B. Disadvantages:

1. If the magnetic fields induced by the currents through other conductors is sufficiently high, the fault current measured will have some errors. Since many conductors in the substation adopted air insulation, it is

possible for the Faraday sensor to detect the fields resulting from the fault current by other conductors [5].

2. The effect of temperature on glass sensitive sensor causes duality of the refractive index and it causes polarized light to distort and Light with linear polarization has become into elliptical and it causes unwanted disturbances to create (refractive index becomes various amounts at different temperatures) [3]. Of course, with modern methods, this problem has been solved. One of these methods is the use of diamagnetic glass. It is independent of the effects of temperature and it can be used without danger from 50 to 110° C [2,11].

3. Faraday Effect is related to the wavelength of light in the system. To remain constant wavelength of led, a temperature controller is used. Thermal expansion and vibration effect on the magnetic field adversely and this problem is solved by loop system.

4. In the optical fiber sensors being used on materials such as adhesives, Resin, thin strips of electrical insulation and these materials evaluated carefully. Because it is difficult to assess each material separately, longevity or aging experiments are done for optical sensors overall.

5. Bending of Optical fiber amply causes the refractive index of the fiber of changes and response sensitivity to Reduce and the influence of temperature to increase, especially in hoops that have several rounds of wire.

6. Because output quantities of MOITs are negligible, despite the simplicity of the building they need to have very high accuracy when they design and build.

7. In MCVTs, optical fiber must pass through the oil and the inside of the tank and it causes its Coating to be destroyed so must forecast that it reinforce.

CONCLUSION

Rapid advances in the quality of performance and costs of the optical fiber and electronic equipment to encourage development of measuring trances based on new technology.

We have provided an experimental comparison of the performance of the optical current transformers with conventional magnetic current transformers. The results have confirmed that the OCTs are suitable for power system protection and can replace the magnetic CTs. Similar comparison can be performed with other technologies of optical CTs, like magneto-optic and fiber optic current sensors.

Future goals are systems for advanced tariff measurement and protective relays for this technology. Waveform of Measurement and Protection by MOCT was set both in terms of time zones and the frequency very excellent. Values of RMS of Output current in conventional CTs and Moct are measured by less than 0.1% difference from each other that it showed the high quality of the Moct output signal, it can use with a great inductor to connect with applications and devices and observed that MOCT does cloning and replication expected harmonics in the power system.

The results of parallel comparing experiments (Two types CT) showed that during a 4-month period of operation only a 0.4 percent difference has occurred in the counter numeration.

Continuous activities and research investments in the development that has been done by different companies and manufacturers decrease costs naturally and make advanced technology in future.

REFERENCE

[1] J. H. Harlow, Electric Power Transformer Engineering, The Electric Power Engineering Series; 9. CRC Press LLC, 2004.

[2] M. Xianyun, L.Chengmu, "A Method to Eliminate Birefringence of A Magneto-optic AC Current Transducer With Glass Ring Sensor Head",IEEE Transactions on Power Delivery, Vol. 13, No. 4, pp. 1015-1019, 1998

[3] T. Sawa, K. Kurosawa, T. Kaminishi, T. Yokota,"Development of Optical Instrument Transformers", IEEE Trans. on Power Delivery, 1990, PWRD-5(2):884-891.

[4] Y. Nie, X. Yin and Z. Zhang, "Optical Current Transducer Used in High Voltage Power System", IEEE

[5] Optical Current Transducers for Power Systems: A Review, IEEE Trans. on Power Delivery, 1994, PWRD-9(4): 1778-1 788.

[6] T.W. Cease, J.G. Driggans and S.J. Weikel, "Optical voltage and current sensors used in a revenue metering system". IEEE Trans. Pwr. Delivery, Vol 6, No 4, pp. 1374-1379, 1991.

[7] M. Brojboie, V. Ivanov, S.M. Diga, "Implementation of The Optical Current and Voltage Transducers in The Power Systems", Int. Conf. Electromechanical and Power Systems, Romania, October 2009, pp. 27-33.

[8] C.H. Einvall, M. Adolfsson, P.Lindberg, J.Samuelsson, "A new optoelectronic measuring system for EHV substations", ELECTRA, 1988.

[9] Z. Araujo, M. Dávila, E. Mora, L. Maldonado, "Analysis of the Behavior of an Optical Current Transformer using an equivalent circuit", IEEE Computer Society, DOI 10.1109/Andescon.2012.28, pp. 81-84, 2012.

[10] T. W. Cease, P. Johnston, "A Magneto-Optic Current Transducer", IEEE Transactions on Power Delivery, Vol. 5, No. 2, pp. 548-555, 1990.

[11] E.A. Ulmer, Jr., "A High-Accuracy Optical Current Transducer For Electric Power Systems", IEEE Transactions on Power Delivery, Vol. 5, No.2, pp. 892-898, 1990.

[12] T.D. Maffetone, T.M. McClelland, "345 kV Substation Optical Current Measurment System for Revenue Metering and Protective Relaying", IEEE Transactions on Power Delivery, Vol. 6, No. 4, pp. 1430-1437, 1991.

[13] E. Aikawa, A. Ueda, M. Watanabe, H. Takahashi, "Development of New Concept Optical Zero-Sequence Current/Voltage Transducer for Distribution Network", IEEE Transactions on Power Delivery, Vol. 6, No. 1, pp. 414-420, 1991.

[14] S. Liehr, "Optical Measurement of Currents in Power Converters", Master's Degree Project in Electrical Measurement Technology report no. XR-EE-MST 2006:001.

[15] S. Kucuksari, G.G. Karady, "Experimental Comparison of Conventional and Optical Current Transformers", IEEE TRANSACTIONS ON POWER DELIVERY, VOL. 25, NO. 4, pp. 2455-2463, 2010.

[16] Y. Yamagata, T. Oshi, H. Katsukawa, S. Kato, "Development of Optical Current Transformers and Application to Fault Location Systems for Substations", IEEE Transactions on Power Delivery, Vol. 8, No. 3, pp. 866-873, 1993.

Reliability Constrained Energy and Reserve Scheduling of Microgrids Including High Penetration of Renewable Resources

Smruti Ranjan Nayak, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, smrutinayak@live.com*

Pratik Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, pratikmohanty92@hotmail.com*

Nabnit Panigrahi, *Department of Electrical and Electronics Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, nabnitpanigrahi@gmail.com*

Alekha Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, alekha.sahoo241@gmail.com*

Abstract *– Due to environmentally and economically advantages, high deployment of renewable energy sources (RES) such as wind or photovoltaic (PV) units in Microgrids (MG) has been increasing in recent decades. On the other hand, random and uncertain nature of the RES poses a challenge to Microgrid operators (MGO) for energy and reserve scheduling considering reliability constraints. To address this problem, a novel probabilistic energy and reserve scheduling method is proposed in this paper. The proposed method maximizes the Microgrid net benefit and reliability so that the optimal requirement reserve is determined by a tradeoff between reliability and economics.*

Keywords*: Microgrids, renewable energy sources (RES), energy and reserve scheduling, expected energy not supplied (EENS).*

INTRODUCTION

Increasing deployment of renewable energy sources in Microgrids implies that MGOs will need to handle the random and uncertain nature of RESs like wind and PV in order to continuously preserve the supply-demand balance [1]. Microgrid, as a low voltage small distribution network illustrated in Fig. 1, integrating renewable and conventional energy sources, energy storage and loads in order to local production of electricity as well selling power back to the upstream network [2]. Energy and reserve scheduling of a MG has substantial differences from the large power system due to the flexibility and usability of MGs depend on their composition [4]. The MGs energy scheduling, including renewable generation in a MG, have been studied in many works [5-7].

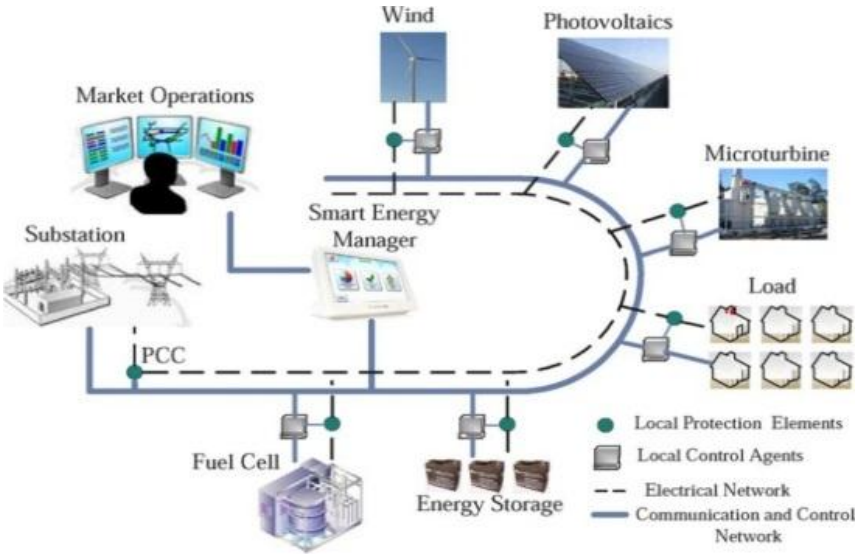


Figure 1. Microgrids [3]

In study of Foo Eddy et al. [5], a multi-agent based system for coordination of MG is proposed. Although the developed multi-agent system ensures proper operation scheduling of MG but uncertainties of renewable energy sources is ignored. In study of Xiong Wu et al. [6], a hierarchical framework for MGs energy scheduling is

proposed. The upper level the MG operation cost is minimize while the lower level combines scheduling of renewable energy sources and energy storage unit in order to minimize uncertainties of power generation of RESs. However, reliability constraints are not investigated by authors. In study of Duong Nguyen et al. [7], an optimal stochastic day-ahead energy scheduling framework of MGs has been proposed incorporating the uncertainties of renewable energy sources. However, procurement reserve for satisfying wind and PV power uncertainties has not been taken into account within the proposed stochastic day-ahead MG scheduling.

The study of Khodaei [8], explores resiliency-oriented MGs optimal scheduling problem. The proposed model minimizes the load shedding by optimally MG scheduling when supply of power from the upstream network is disconnected for a specific period of time. However, uncertainties of renewable energy sources are not included in the proposed model. In study of Cecati at al. [9], a combined operation of renewable energy sources and responsive loads is optimized in the MGs. By combining supply and demand scheduling, it permits a better use of renewable energy sources and a decrease in the payment cost of responsive loads. Islanded MGs scheduling model including renewable energy sources and battery storage unit is proposed in [10]. In this regard, the MG operation cost including battery life loss cost, operation and maintenance cost, fuel cost, and environmental cost is minimized. In study of Duong Nguyen [11], an optimal bidding strategy for MGs is presented. The proposed method enables MGO to determine optimal day-ahead hourly bids that maximize the MG profit by a risk constrained stochastic programming approach.

To the best of our knowledge, no probabilistic energy and reserve scheduling method in Microgrids with high deployment of renewable sources considering reliability constraints has been reported in the papers. Accordingly, this paper address this issue by proposing a probabilistic approach for energy and reserve scheduling of Microgrids in which reliability constraints are taken into account.

The remaining paper is organized as follows. The uncertainties of wind, PV and Microgrid demand are described in section 'Uncertainties modeling'. The proposed method is introduced in section 'energy and reserve scheduling '. A case study is examined in section 'simulation results'. Conclusion section is explained in section 'Conclusion'.

MATERIAL AND METHODS

UNCETAINTIES MODELING

The random output power of renewable energy sources such as wind and PV units is caused a significant uncertainty into MG scheduling. As well, the load forecasting uncertainty at the MG level is high [12]. In this section, the uncertainty of wind power, PV power and load is modeled.

A. Load Demand

The MG load demand forecasting uncertainty can be obtained from historical data set. According to power system, a normal distribution probability as shown in Fig. 2 with a large standard deviation is used [7].

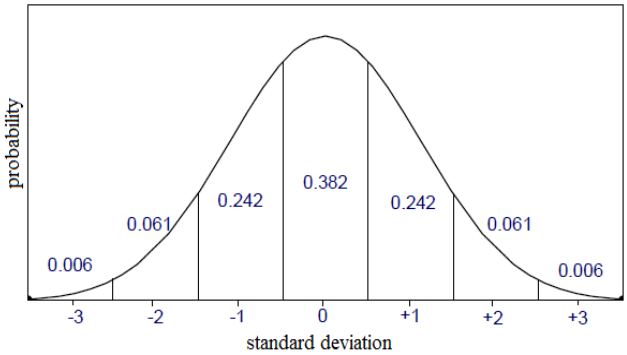


Fig. 2 – Probabilistic Load Model

B. Wind Power

The probability density function of wind speed which is modeled by the Weibull distribution [2] is shown in Fig. 3 and given by:

$$f(v) = (k/c)(v/c)^{(k-1)} e^{-(v/c)^k}, \quad 0 < v < \infty \qquad (1)$$

Where, v, k and c represent wind speed, shape and scale factor respectively.

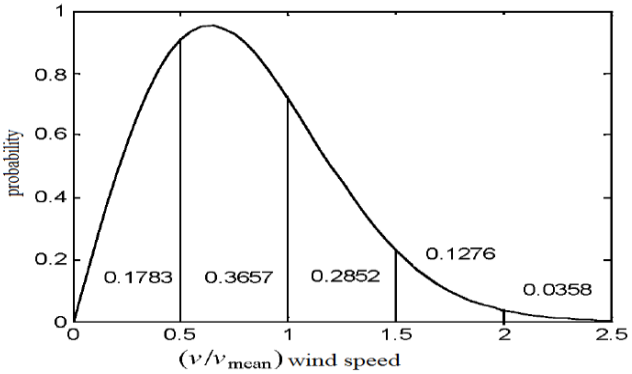


Fig. 3 – Wind speed probability Model

The power generation of the wind unit can be calculated by:

$$w = \begin{cases} 0 & for\ v < v_i\ and\ v > v_o \\ w_r (v - v_i)/(v_r - v_i) & for\ v_i \le v \le v_r \\ w_r & for\ v_r \le v \le v_o \end{cases} \qquad (2)$$

Where, v_i, v_o and v_r are cut-in, cut-out and rated speed of wind turbine, respectively. Also, w_r is rated power of wind turbine.

C. PV Power

The probability density function of solar irradiance which is modeled by the bimodal distribution [11] is shown in Fig. 4 and given by:

$$f(g) = \omega(k_1/c_1)\left(g/c_1\right)^{(k_1-1)} e^{-(g/c_1)^{k_1}} + (1-\omega)(k_2/c_2)\left(g/c_2\right)^{(k_2-1)} e^{-(g/c_2)^{k_2}}, \quad 0 < g < \infty \tag{3}$$

Where, g and ω are solar irradiation and weight factor, respectively, k_1 and k_2 are shape factors, c_1 and c_2 are scale factors.

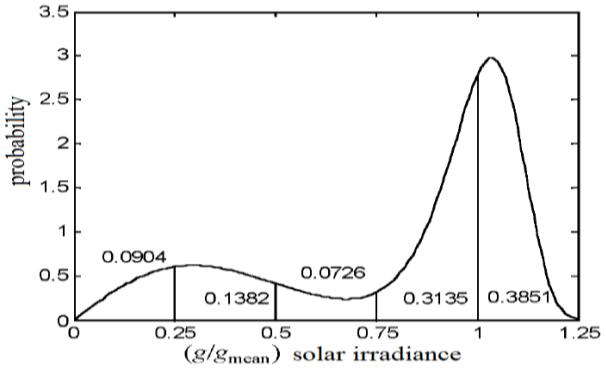


Fig. 4 – Solar irradiance probability Model

The power generation of the PV unit can be calculated by:

$$p = \eta^{PV} S^{PV} g \tag{4}$$

Where, p is PV power generation, S^{PV} and η are PV total area and efficacy, respectively.

D. Scenario Generation

Monte Carlo sampling method [11] is used to sample day-ahead load demand, wind speed and solar irradiance according to their aforementioned probability distributions. Then wind and PV power generation can be calculated using equations (2) and (4), respectively.

ENERGY AND RESERVE SCHEDULING

In this section, the mathematical formulation of the proposed method is described.

A. Objective function and constraints

The objective of proposed method is to maximize the total benefit of the MG:

$$\max\left\{\sum_{t=1}^{T}\sum_{i=1}^{I} MP(t)\times P_{i,t} - \sum_{t=1}^{T}\sum_{i=1}^{I}\left[a_i.U_{i,t} + b_i.P_{i,t} + SC_i \times K_{i,t}\right] - \sum_{t=1}^{T}\sum_{i=1}^{I} q_{i,t}\times R_{i,t} - EENS\times VOLL\right\} \tag{5}$$

Where, $MP(t)$ is wholesale market price, $P_{i,t}$ and $U_{i,t}$ are power generation and status of unit i during period t, respectively. a_i, b_i and SC_i are cost coefficients and start-up cost of unit i, respectively. $q_{i,t}$ and $R_{i,t}$ are reserve cost and its value related to unit i, respectively. Reliability cost is equal to expected energy not supplied (EENS) caused by load demand, wind and PV power uncertainties and value of loss of load (VOLL) is taken into account.

The constraints of proposed model are depicted as follow:

- Load balance constraint

$$\sum_{i=1}^{I} P_{i,t} = P_t^D \quad \forall\ t = 1,...,T \tag{6}$$

Where, P_t^D is forecasted MG load demand during period t.

- Reliability constraint

$$LOLP_t \le LOLP_t^{\max} \tag{7}$$

Where, LOLP is probability of loss of load caused by load demand, wind and PV power uncertainties.

- Reserve constraint

$$\begin{cases} R_{i,t} \le P_i^{\max} U_{i,t} - P_{i,t} \\ R_{i,t} \le U_{i,t}.R_{i,t}^{up} \end{cases} \tag{8}$$

Where, $R_{i,t}^{up}$ is ramp up rate of unit i during period t. In addition, each unit is subject to its own operating limits, which consist of the maximum and minimum unit limits, minimum up and down time limits and ramp up and ramp down limits [9].

B. Reliability formulations

The mathematical formulation of LOLP and EENS are presented as bellow:

$$EENS = \sum_{t=1}^{T}\sum_{n=1}^{N} p_t^0 p_{n,t} b_{n,t}^0 (\Delta P_{n,t} - R_t) + \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{n=1}^{N} p_{i,t}^1 p_{n,t} b_{i,n,t}^1 (P_{i,t} + R_{i,t} + \Delta P_{n,t} - R_t) + \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j>i}^{I}\sum_{n=1}^{N} p_{i,j,t}^2 p_{n,t} b_{i,j,n,t}^2 (P_{i,t} + R_{i,t} + P_{j,t} + R_{j,t} + \Delta P_{n,t} - R_t) \tag{9}$$

$$LOLP_t = \sum_{n=1}^{N} p_t^0 p_{n,t} b_{n,t}^0 + \sum_{i=1}^{I}\sum_{n=1}^{N} p_{i,t}^1 p_{n,t} b_{i,n,t}^1 + \sum_{i=1}^{I}\sum_{j>i}^{I}\sum_{n=1}^{N} p_{i,j,t}^2 p_{n,t} b_{i,j,n,t}^2 \tag{10}$$

Where, k, m and l are numbers of load demand, wind speed and solar irradiance generated scenarios in Monte Carlo method, respectively.

$$\Delta P_{k,t}^{WT} = \sum_{i=1}^{nWT}\left[f_i^{WT}\left(v_{i,t}^{fcst}\right) - f_i^{WT}\left(v_{i,t}^{scen}\right)\right] \tag{11}$$

$$\Delta P_{m,t}^{PV} = \sum_{i=1}^{nPV}\left[f_i^{PV}\left(g_{i,t}^{fcst}\right) - f_i^{PV}\left(g_{i,t}^{scen}\right)\right] \tag{12}$$

$$\Delta P_{l,t}^{D} = P_t^D - P_{l,t}^D \tag{13}$$

Where, load demand, wind speed and solar irradiance scenarios are $P_{l,t}^D$, $v_{i,t}^{scen}$ and $g_{i,t}^{scen}$, respectively.

$$b_{k,l,m,t}^0 = \begin{cases} 1 & if\ \ \Delta P_{k,t}^{WT} + \Delta P_{m,t}^{PV} + \Delta P_{l,t}^{D} - R_t > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The p_t^0, $p_{i,t}^1$ and $p_{i,j,t}^2$ are forced outage probabilities [13] of zeros, one and two units, respectively.

$$p_t^0 = \prod_{i=1}^{I}\left(1-u_i U_{i,t}\right) \quad (15)$$

$$p_{i,t}^1 = u_i U_{i,t} \prod_{j=1, j\neq i}^{I}\left(1-u_j U_{j,t}\right) \quad (16)$$

$$p_{i,j,t}^2 = u_i u_j U_{i,t} U_{j,t} \prod_{k=1,k\neq i,j}^{I}\left(1-u_k U_{k,t}\right) \quad (17)$$

Where, u_i is forced outage rate (FOR) of unit i.

RESULTS

In this section, the proposed model is implemented on the MG as illustrated in Fig. 5. Hourly load, wind speed as well wind turbine, solar irradiation as well PV array and price data are drive from [2].

Technical and economic data of units was taken from works of Chen at al. [3]. Reserve cost and FOR let fix and equal to 0.04 $/kWh and 0.006, respectively, while VOLL is set to be 1000 $/kWh.

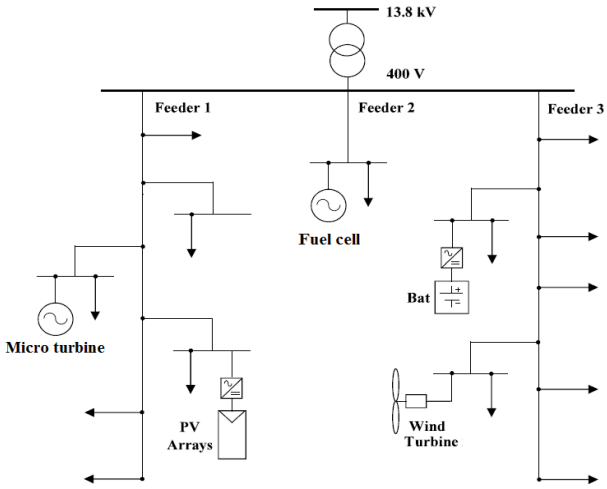


Figure 5. Test Microgrid

Energy scheduling of MG units is shown in Fig. 6. As shown in Fig. 6, in the hours with high wholesale price MGO using local units' energy production and sell back extra energy to upstream grid. The profit of MGO is 189 $ in this case.

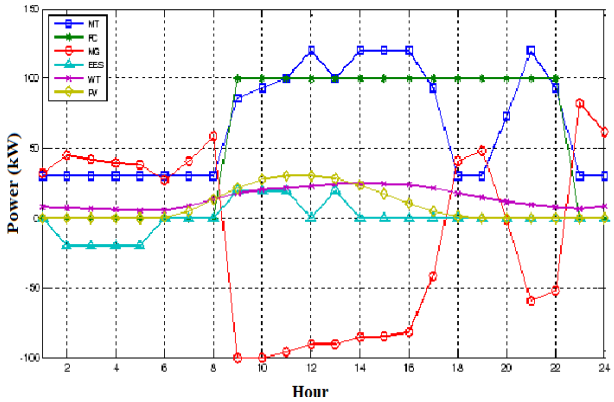


Figure 6. Energy scheduling of MG units

Reserve scheduling and available reserve of MG is illustrated in Fig. 7. As shown, with increasing energy production of RESs and load demand of MG, requirement reserve of MG is grown while available reserve of MG is reduced.

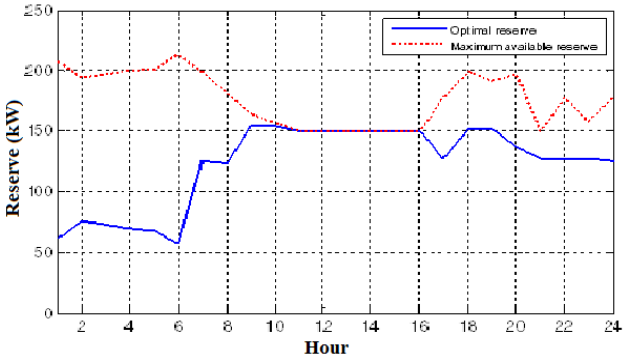


Figure 7. Reserve scheduling and available reserve of MG

Loss of load probability is presented in Fig. 8. As depicted, with increasing energy production of RESs and load demand of MG, uncertainty is grown. In this case study $LOLP^{max}$ is set to be 0.005.

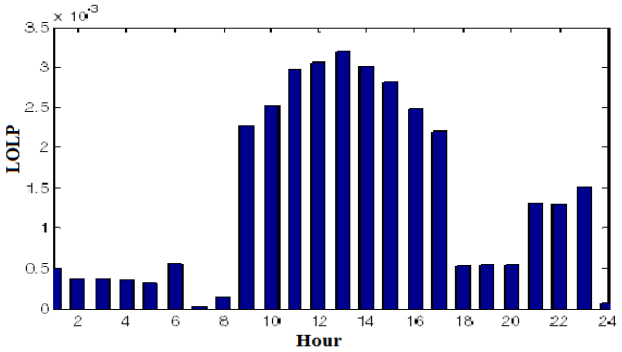


Figure 8. Loss of load probability

Reserve scheduling of MG with different FOR is shown in Fig. 9. As shown, with increasing of FOR, forced outage rate of units is grown and as a result requirement reserve of MGO is exceeded.

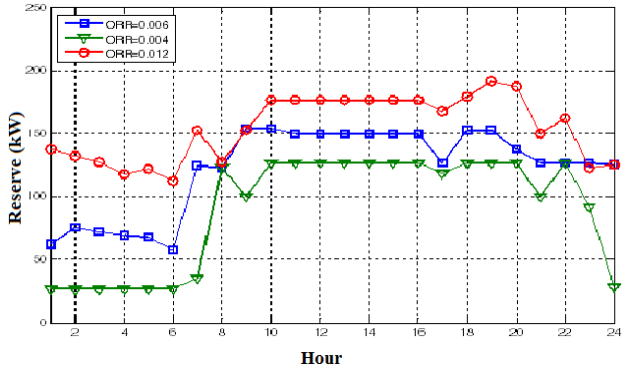


Figure 9. Reserve scheduling for different FOR

Reserve scheduling of MG with different wind penetration is shown in Fig. 10. As illustrated, with increasing energy production of wind unit, uncertainty is grown and as a result requirement reserve of MGO is exceeded.

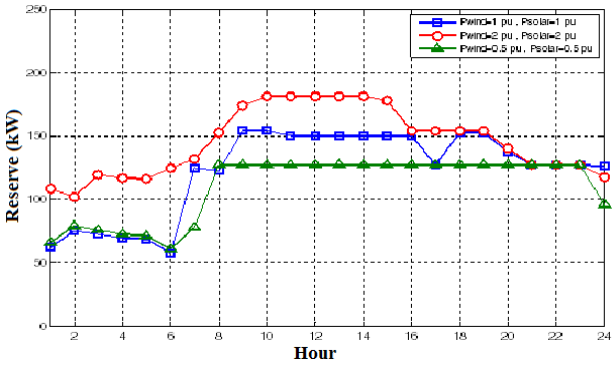


Figure 10. Reserve scheduling for different wind penetration

Reserve scheduling of MG with different VOLL is shown in Fig. 11. As shown, with increasing VOLL, cost of MGO caused by load shedding is grown and as a result requirement reserve of MGO is exceeded.

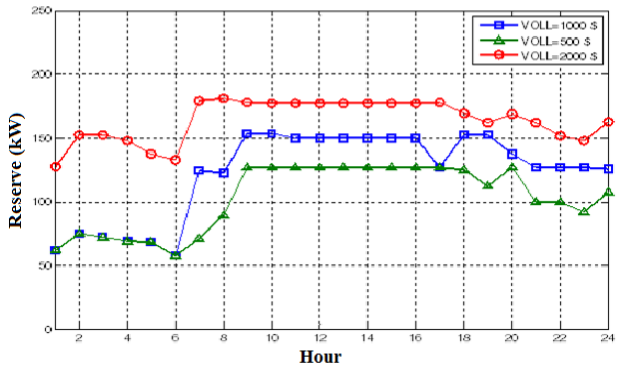


Figure 11. Reserve scheduling for different VOLL

DISCUSSION

Microgrids provide various benefits for consumers in terms of reliability and economy. However, the economic benefits of MG should be studied to justify high penetration of renewable resources while considering reliability constraints in the scheduling process. Specific features of proposed energy and reserve scheduling of MG are listed as follows:

1- Considering reliability constraints in energy and reserve scheduling of MG.

2- Proposed a unified and mixed integer linear mathematical formulation for modeling the problem.

3- A comprehensive uncertainty modeling using Monte Carlo simulation technic.

4- Doing a sensitivity analysis for evaluating reliability parameters.

CONCLUSION

In this paper, a probabilistic method for energy and reserve scheduling of MG is proposed. Effeteness of the proposed method is demonstrated by implemented on the MG and simulation results are analyzed. The results show that the MG profit is maximized while reliability constraints are met.

REFERENCES

[1] M.A. Matos, Bessa, R.J., "Setting the Operating Reserve Using Probabilistic Wind Power Forecasts," Power Systems, IEEE Transactions on , vol.26, no.2, pp.594,603, May 2011.

[2] A.G. Tsikalakis, Hatziargyriou, N.D., "Centralized Control for Optimizing Microgrids Operation," Energy Conversion, IEEE Transactions on , vol.23, no.1, pp.241,248, March 2008.

[3] S.X. Chen, Gooi, H.B., Wang, M.Q., "Sizing of Energy Storage for Microgrids," Smart Grid, IEEE Transactions on, vol.3, no.1, pp.142,151, March 2012.

[4] M.Q. Wang, Gooi, H.B., "Spinning Reserve Estimation in Microgrids," Power Systems, IEEE Transactions on, vol.26, no.3, pp.1164,1174, Aug. 2011.

[5] Y.S. Foo Eddy, Gooi, H.B.; Chen, S.X., "Multi-Agent System for Distributed Management of Microgrids," Power Systems, IEEE Transactions on , vol.30, no.1, pp.24,34, Jan. 2015.

[6] Wu Xiong Xiuli Wang; Chong Qu, "A Hierarchical Framework for Generation Scheduling of Microgrids," Power Delivery, IEEE Transactions on , vol.29, no.6, pp.2448,2457, Dec. 2014.

[7] D.T. Nguyen; Bao Le L., "Optimal Bidding Strategy for Microgrids Considering Renewable Energy and Building Thermal Dynamics," Smart Grid, IEEE Transactions on , vol.5, no.4, pp.1608,1620, July 2014.

[8] A. Khodaei, "Resiliency-Oriented Microgrid Optimal Scheduling," Smart Grid, IEEE Transactions on , vol.5, no.4, pp.1584,1591, July 2014.

[9] C. Cecati, Citro, C.; Siano, P., "Combined Operations of Renewable Energy Systems and Responsive Demand in a Smart Grid," Sustainable Energy, IEEE Transactions on , vol.2, no.4, pp.468,476, Oct. 2011.

[10] B. Zhao; Zhang X, Chen J., Wang C., Guo L, "Operation Optimization of Standalone Microgrids

Considering Lifetime Characteristics of Battery Energy Storage System," Sustainable Energy, IEEE Transactions on , vol.4, no.4, pp.934,943, Oct. 2013.

[11] D.T. Nguyen, Bao Le L., "Risk-Constrained Profit Maximization for Microgrid Aggregators With Demand Response," Smart Grid, IEEE Transactions on , vol.6, no.1, pp.135,146, Jan. 2015.

[12] N. Amjady, Keynia, F., Zareipour, H., "Short-Term Load Forecast of Microgrids by a New Bilevel Prediction Strategy," Smart Grid, IEEE Transactions on , vol.1, no.3, pp.286,294, Dec. 2010.

[13] M.A. Ortega-Vazquez, Kirschen, D.S., "Optimizing the Spinning Reserve Requirements Using a Cost/Benefit Analysis," Power Systems, IEEE Transactions on , vol.22, no.1, pp.24,33, Feb. 2007

Line Start Permanent Magnet Synchronous Motor Performance and Design; a Review

Prakash Chandra Sahu, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, pksahoo88@gmail.com*

Pratik Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, pratikmohanty92@hotmail.com*

Smruti Ranjan Nayak, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, smrutinayak@live.com*

Swadesh Ranjan Jena, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, swadesh.jena@yahoo.co.in*

Abstract *– Line-start PM synchronous motor can be used immediately instead of the conventional induction motor for many applications because of its advantages like high efficiency, high power factor and high power density compare to induction motor, especially for small motor sizes. This paper is a review on research about different aspects of LS-PMS motors presented in literature. LSPMS motor model and parameters, starting and synchronization, cogging torque, pm demagnetization and harmonics are the topics covered in the paper. Besides many researchers designed, simulated and prototyped different LSPMS motors which their results are summarized in this paper.*

Keywords*: Line Start Permanent Magnet Synchronous Motor, LSPMSM*

INTRODUCTION

Electric Motors in Industrial applications consume between 30% and 40% of the generated electrical energy worldwide. So higher efficient electric motors can lead to significant reductions in energy consumption and also reduce environmental impact. Despite the wide variety of electric motors available in the market, three-phase, squirrel-cage induction motors (IMs) represent, by far, the vast majority of the market of electric motors. For many applications, a permanent magnet (PM) synchronous motor can be designed smaller in size and more efficient as compared to induction motor. In particular, line-start PM synchronous motor can be used immediately instead of the conventional induction motors for applications in pumps, air conditioners and fans [1, 2].

LS-PMS motor model

An LS-PMS motor consists of a single or poly-phase stator as same as induction motor and a hybrid rotor involving electricity conducting squirrel cage and pairs of permanent magnet poles.

Different combinations of the cage, pole shapes and pole locations have been presented for the rotor so far.

The motor starts as an induction motor by the resultant of two torque components i.e. cage torque and magnet opponent torque (breaking torque). When the motor speed reaches near synchronous speed, a synchronization process begins and motor operation is transferred to synchronous state when no eddy current flows into the cage bars except harmonics field currents. In synchronous state two torque components i.e. a reluctance torque component and a synchronous torque component cause the rotor motion.

A two-axis dynamic model of three phase line start permanent magnet synchronous motor in rotor reference frame can be given by voltage, flux, and torque equations. The dynamic performance of LS-PMS motor in a stationary d-q reference can be described by this model. [1].

In Marcic et al. [3] study the magnetically linear two-axis LSIPMSM dynamic model with the flux linkage due to permanent magnets as a parameter is presented. In this work the parameters of the model are determined by the differential evolution (DE). The authors showed that the DE is a very suitable tool for determining parameters of the LSIPMSM dynamic model.

NdFeBr magnets are very temperature sensitive. Abbas et al. [4] studied the effect of temperature variation on the performance of an industrial LS-IPMSM is carried out by finite element method. Inductances sensitivity analysis of these motors is also studied and finally, the PM demagnetization due to abnormal operating conditions is presented.

Štumberger et al. [5] presented the usage of a lumped parameter dynamic model with current-dependant variable parameters in LSIPMSM dynamic performance

evaluation. For the identification of model parameters the finite element method procedure, is used. The variations of model parameters

(squirrel-cage resistances, stator and rotor self-inductances, and mutual inductances in d- and q-axis) are determined by post-processing of FEM output data.

Lu et al. [6] studied a simple experimental method to determine the magnetization characteristics of an LSPMSM are proposed. Numerical and experimental investigations have been performed to validate the determined characteristics. The results obtained from experiments are found to closely match with that of the numerical investigations. Also Lu et al. [7] studied exclusively a novel magnetic circuit model to design LSPMSM with improved starting performance is proposed. Further, a detailed procedure to deal with the design issues with the help of the developed magnetic circuit model is discussed and validated by developing a machine where the trade-off between their starting performance and efficiency post-synchronization has been reached.

Starting and synchronization

Starting and synchronization of LS-PMS motors have been challenging issues concerned by many researches so far. Magnet braking torque is not the only deficiency of LS-PMS motors starting. The motors also suffer from a sensitive dependency of starting process on input voltage, shaft inertia momentum and cages resistance. With a reduced input voltage, the motor starts more slowly and even may fail in synchronization. The settling time of speed curve also increases with falling of input voltage. There is an optimal value for cage resistance out of which the motor does not start properly.

By increasing the load inertia, the motor starting deteriorates. This may lead to high torque pulsations in induction mode, not reaching the synchronous operation at all. [1]

In Honsinger et al. [8] study, the permanent magnet machine during asynchronous operation has been treated using generalized machine theory. The main body of the paper is concerned with the calculation of torques and currents during run-up.

Miler [9] described the synchronizing process in line-start PM Ac motors, with particular emphasis on the analysis of factors that influence the synchronizing capability. He showed that of the two components of synchronous torque, the magnet alignment torque contributes much more to the synchronizing capability than the reluctance torque, and this leads to a better starting capability for PM motors than for pure reluctance motors of the same size.

Soulard and Nee [10] studied a simple method to define the maximum load torque that can be synchronized by LSPM motors has been presented. Authors defined a pull-in criterion by using a Lyapunov function. The model is derived and a Lyapunov function is defined using the Lagrange-Charpit method. Experiments and simulations are compared to check the validity of the model. Finally a criterion to define the capability of synchronization of LSPM motors is presented.

In Fengbo et al. [11] study, the starting process of high-voltage permanent magnet synchronous motor is calculated and analyzed by the time-stepping Finite Element Method coupling with field and circuits. The starting performance is analyzed by calculating the equations of transient electromagnetic field and mechanical movement equations. In this paper showed that increasing the number of cage bars can enhance Pull capacity of the HV-PMSM. The Pull-in torque was improved, the starting time was reduced. But the change will also increase the starting current, so the number of cage bar should be selected according to different requirements in practical application.

In Takegami et al. [12] study, the purpose of calculating asynchronous starting characteristics of an LSPMM, the method of the constant decision is described. In addition, the calculating results are confirmed by a comparison with an experimental result.

Stoia et al. [13] has proposed a study of synchronization capability of LSPMSM with very high cage resistance and small saliency ratio. The high value of the rotor resistance of the designed motor has beneficial effects on the early start, but the synchronization occurs at a large value of the slip and is relatively difficult. A minimal optimum value of synchronization energy has been found for the no-load voltage which maximizes the critical electromagnetic torque.

In Hassanpour Isfahani et al. [14] study, a simple analytical method for finding critical slip in synchronization assessment of LSPMS motors is presented. Appropriate approximations are employed in the method resulting in significant simplification of the method with reasonable accuracy. The proposed method is time efficient and can be easily included in iterative design procedures. Accuracy of the proposed method is evaluated by dynamic analysis of two different motors. Also, the experimental results are presented to support the

analytical and simulation results. Also in Hassanpour Isfahani et al. [15] study, different effects of magnetizing inductance of line start permanent magnet synchronous motors on the motors starting performance are investigated in three different ways. First, the relation between the magnetizing inductance value and the average and pulsating torques during asynchronous operation of motors is discussed. A critical load value is determined to assure the motor start-up. It is shown that this critical load significantly depends on the value of magnetizing inductance. It is also shown that an increase in the magnet flux and saliency reduces this critical load effectively, especially when the magnetizing inductance is small. A dynamic model of motors is then presented to study the starting performance of two motors with different magnetizing inductance values to support the discussions. Finally, the finite-element method is utilized to take into account the saturation, cross magnetization, asymmetrical rotor cage bar resistance, etc. The FEM results verify the results obtained from dynamic simulations.

In Nedelcu et al. [16] study, influence of various geometrical design parameters (i.e. geometry of PMs, flux barriers and rotor bar cross section) on the starting characteristics of a LSPMSM is carried out. The effect of modifying the geometrical configuration of PMs, rotor bars and flux barriers on the LSPMSM starting capability is analyzed using the FLUX software package. This study is proved to be useful for the optimal design of the machine with the purpose of improving its start-up capability, one of the most sensible points of LSPMSM.

Rabbi et al. [17] studied a simplified analytical method to determine the critical slip and the critical inertia of a line start IPM motor based on average torque analysis is presented. The synchronization process of a line start IPM motor has been explained in details. Also, the mathematical formulations of braking torque, cage torque and synchronous torque have been presented. An experimental investigation has also been carried out to determine the synchronization performances of a laboratory 1 hp, 3-phase, 4-pole IPM motor fed from a fixed 3-phase 60Hz ac supply.

Cogging torque

Cogging torque arises from interactions between permanent magnets mounted on the rotor and the anisotropy originated by stator windings slots. These cause variations of the magnetic field energy during the rotation. Cogging torque lowers torque quality and affect smooth running of the machine, producing vibrations and mechanical noise [18].

In Yang et al. [19] study, the analytical method is used to find the relationship between the pole arc coefficient and cogging torque. Then, according to the analytical result, the feasible region of the pole arc coefficient of the PM is derived. With the feasible region of pole arc coefficient, the improved domain elimination method and finite-element method are combined to optimize the pole arc coefficient of the PM to minimize the cogging torque of permanent magnet motor. The authors showed with this method, the computing time decreases notably, and the cogging torque is greatly reduced.

Bing-yi et al. [20] studied, a novel structure of motor which the number of slots per pole per phase q is less than 1 is proposed. The authors showed that the line-start PMSM with a new structure has good starting and running performance. The value of cogging torque is smaller than conventional PMSM. Simulation and prototype test results show that this novel structure is reasonable.

In Chu and Zhu [21] study, the influence of skewing on the torque ripples in PM machines with different magnet shapes and loads is investigated. Although the investigation is carried out on the SPM machines, the conclusions are also applicable to the interior PM machines and the electrically excited machines, where the influence of armature field and magnetic saturation is more significant.

In Bianchini et al. [22] study, an overview of the methods for cogging torque reduction in IPM synchronous machines is presented. The various methods are compared on a common reference machine (12-slot 4-pole IPM machine) by extensive FEM simulations. The results show that some techniques d2eveloped for SPM machines can be easily applied to IPM machines as well.

The authors suggest that care should be used when adopting design solutions for cogging torque reduction in IPM machines as, in some cases; they could negatively affect the torque quality at full load. For best results, during optimization it is advisable not to focus only on cogging torque reduction but to monitor the side effects as well.

In Lee et al. [23] study, torque ripple in the IPMSM according to the rotational speed is presented. The authors showed that torque ripple of the IPMSM tends to increase when flux weakening control is applied. They proposed a method to find harmonic injected current to achieve

minimization of torque ripple. This method can be used to improve the accuracy of speed and position when flux weakening control is applied.

Permanent Magnet

The irreversible demagnetization of permanent magnet due to the armature reaction during the starting process of line-start permanent magnet synchronous motor (LSPMSM) has tremendous influence on its performance, even makes the machine unable to work. In Kang et al. [24] study, the irreversible demagnetization characteristics of a ferrite-type permanent magnet in the line-start PM motor is carried out by using the two-dimensional finite element method. The demagnetizing currents are calculated from the transient analysis in combination of voltage equation and mechanical dynamic equation, and peak currents are applied to the irreversible demagnetization analysis computed by 2-D FEM. The nonlinear characteristic of the magnetic cores has been considered as well as that of a permanent magnet on the B-H curve in the analysis of irreversible demagnetization.

In Lu et al. [25] study, the armature reaction demagnetization during the starting process of LSPMSM has been calculated, and the variety of influencing factors has been studied. Analysis results shows that demagnetization is more prone to occur when start-up time is longer because of larger load or larger inertia or lower supply voltage, or because of a specific initial rotor position especially during the light-load starting process.

In Lu et al. [26] study, authors made an effort to study the causes and effects of permanent magnet demagnetization in permanent magnet machines and proposed an exclusive artificial neural network (ANN) based permanent magnet demagnetization detection scheme. A laboratory 2.8 kW line-start permanent magnet synchronous machine (LSPMSM) is used in the numerical investigations for initiating permanent magnet demagnetization and detecting the fault.

In Shen et al. [27] study, four rotor configurations are proposed to protect the magnets from demagnetization, and their effectiveness is comparatively studied. It is shown that the configuration with both dual cages and magnetic barriers performs the best to protect the magnets, and hardly deteriorates the normal operation performance. The line start permanent magnet motor is noted as an alternative to the induction motor because it offers a very high efficiency and unity power factor. However, a high manufacture cost as compared to an induction motor is disadvantage. Thus, the post-assembly magnetization of the NdFeB magnet is considered to reduce the material and the manufacturing costs. In Lee et al. [28] study, the magnetizing fixture that magnetizing the NdFeB in the rotor of LSPM is designed and its characteristics of magnetization are analyzed. The eddy current occurring in rotor bars disturbs the magnetization of NdFeB. To reduce the eddy current occurring in the rotor bars the authors have investigated the magnetization characteristics for various coil-turn and dimension of rotor bars by time stepping the FEM with model. The analysis results agree well with experiment results.

In Lee and Kwon [29] study, the design procedure and methods for designing the post-assembly magnetization system of the LSPM are proposed. The design procedure is focused on the design of the capacities of the magnetizer and the coil-turns. The methods for designing the post-assembly magnetization system are to increase the number of coil-turns and to reduce the rotor bar size in order to reduce the eddy current. In order to improve the efficiency, the rotor is redesigned and manufactured. In addition, the magnetization system for it is designed and manufactured.

In Stoia et al. [30] study, a graphical-analytical method for the size up procedure of PMs used in LSPMSM is proposed. Using this theoretical approach the amount of magnet for the required dynamic and steady state performances of this motor can be calculated. The designed operating point of the PM offers the advantage of a large magnetic energy density, near of its maximum.

HARMONICS

One of the drawbacks of LSPMSM is higher harmonic contents of flux density, as well as current and electromagnetic torque in comparison to an induction motor.

In Kurihara et al. [31] study, a method for analysis to obtain steady-state currents and torques of permanent magnet synchronous motors including space harmonics is presented. Time-stepping finite element techniques including the rotor movement are proposed, where both terminal voltage and load angle are given as the known values. The agreement between calculated and measured results of the synchronous performance in an experimental motor is good.

In Zawilak and Zawilak [32] study, on the basis of field-circuit calculations, investigation of higher harmonics of flux density, back emf, armature current and electromagnetic torque in a Line-Start Permanent Magnet Synchronous Motor have been conducted. The paper

presents a new construction of rotor with variable slot width. It is characterized by much lower amplitudes of magnetic field zonal harmonics. This new design has been compared with a typical LSPMSM construction.

In Rong and Manfeng [33] study, authors focus on the harmonic suppression function of the damper windings for air-gap magnetic field of line-start permanent magnet synchronous motors (LSPMSM) using finite element analysis. The simulations and the tests for the line-start PMSM designed by the author were compared, and the results shown in this paper have confirmed the validity of harmonic suppression function of the damper windings, and it is practical value for the reasonable design of line-start PMSM.

Design Aspect

In Kurihara and Azizur Rahman [34] study, a successful design of a high-efficiency small but novel line start PM motor using NdFeB magnets was developed and tested. It is designed to operate both at line and variable frequencies. The IPM motor can start and synchronize with large load inertia. Time-stepping finite-element analysis has been used to successfully predict the dynamic and transient performance of the prototype motors. It has been found that the proposed design has yielded successful simulation and experimental results. The maximum load inertia corresponding to the rotor-bar depth has been given from the simulation results.

In Bingyi et al. [35] study, the structural characteristics of the multi polar line-start PMSM for the low-speed and high torque gearless driver system has presented. The special structures of the stator and rotor have been analyzed. And how temperature influences the performance of the motor is also indicated. Moreover, how to select the length of air-gap and the size of permanent magnet is discussed. Also a high performance multi polar line-start PMSM is designed by MATLAB. The computer simulation results of starting process of the motor are given by ANSOFT base on FEM. The simulation results indicate the design scheme for the multi polar line-start PMSM is feasible.

Yang et al. [36] study, aims at optimal analysis and design of a three-phase line-start PMSM with simple structure, low cost and good performance. A prototype with 4 magnet poles is designed and manufactured. Simulation and experimental results are approximately the same and the prototype essentially satisfies the design request.

In Lu and Ye [37] study, a large capacity LSPMS motor which keeps the configuration as much as that of induction motor in order to reduce the manufacture cost is proposed. Its transient and steady-state performance is predicted by a useful dynamic FEM model that is validated by experiment on an induction motor. Compared with the induction motor, this LSPMSM not only has higher efficiency and power factor, but also has sufficient starting ability on full load.

In Xiaochen et al. [38] study, a solid rotor permanent magnet synchronous motor (PMSM) is developed for the electric propulsion part in electric vehicles (EV). The rotor in this kind of motor is composed of splits solid rotors, interior permanent magnets and starting bars. In this paper, a 30kW solid rotor PMSM with simply structure was taken as the analysis model, and the numerical calculation model under assumptions, as well as the solving regions of the derived model, is proposed. Starting performance and operation performance are analyzed, and some parameters as stator current, power factor and torque-speed characteristic are obtained. The calculated results show good agreement with the experimental data.

In Kim et al. [39] study, a comparison between three architectures of line-start PM motors for oil-pump application is presented. This paper is focused on the performances in synchronous operation as well as the self-starting operations. Effects of electrical parameters on the starting and steady performance characteristics are demonstrated to find a satisfying design to meet required performances.

In Peralta-Sánchez and Smith [40] study, a new form of line-start PM machine that uses a simple canned rotor construction with surface mounted magnets has described. The rotor can acts as the induction winding to provide the line-start capability and also provides a dual role as an environmental shield for applications where this is necessary. This paper has also developed a transient electromechanical model using a classical two-axis model combined with a layer model to determine certain motor parameters. This paper also includes experimental results of the dynamic starting and synchronization performance from a prototype 2.5-kW motor and examines the influence of certain key design features on synchronization.

In Fei et al. [41] study, a high-performance line-start permanent magnet synchronous motor which is developed by simple modifications of an off-the-shelf small industrial three-phase IM with minimized additional costs

is presented. Two-dimensional dynamic finite element analysis models are employed to assess the machine performances, which are validated by comprehensive experimental results. The experimental comparison between the amended LSPMSM and the original IM have indicated that significant improvements in efficiency and power factor can be achieved by the proposed motor.

In Stoia et al. [42] study, an analytical design method for the LSPMSM, considering the asynchronous starting and the synchronous steady state parameters is proposed. Using this theoretical approach, all the synchronous and asynchronous starting characteristics can be calculated.

In Yaojing and Kai [43] study, a two-pole three-phase high power line-start permanent magnet synchronous motor is introduced.

The authors described the operating principle and structural features of the line start PMSM . An interior PM rotor structure is presented, which adopts a radial-set of permanent magnets with multi-section for each pole. A motor model is then established by employing the finite element analysis software JMAG-Studio. The transient electromagnetic field is calculated and analyzed using the time-stepping FEM coupling with magnetic field analysis with electrical circuits. The starting process and steady-state performance are simulated.

A high-efficiency Line Start Permanent Magnet Synchronous Machine is designed in Jazdzynski and Bajek [44] work, to meet efficiency requirements, which can be expected for motors of a class IE4 in a new standard classification. A magnetically linear analytical model of the LSPMSM has been developed and investigated. The task to find a best design solution has been defined as a bi-criterial optimization problem, with criteria functions representing the interest of both the producer and end user. Calculation results were validated by means of a magnetically non-linear FEM model, before and after the optimization.

In Feng et al. [45] work, the supper premium efficiency LSPMSM is researched and developed. The challenges and key design techniques are introduced. Also the advanced digital simulation is used for designed performance evaluation. The prototypes have been built, tested, analyzed and compared with calculated data. The results show that the supper premium LSPMSM has much better efficiency, power factor, and power density, smaller frame size and less material consumption compared with Premium efficiency IM, leading to cost down and energy saving in various applications.

In Ruan et al. [46] study, a comparison between two architectures of line-start permanent magnet motors is presented .The authors focused on the performances in synchronous operation as well as the self-starting operations. Time stepping finite element analysis has been used to predict the dynamic and transient performances of the two prototype motors. It has been found that the motor with series magnetic circuit structure has yielded an impressive performance.

A high performance LSPMSM with consequent pole arrangement of magnets is proposed and studied in Ugale and Chaudhari [47] work. The proposed magnet arrangement results in improved air gap flux density when compared with other magnet configurations used earlier, for same magnet volume. The performance of proposed rotor is significant when benchmarked with the induction motor of the same rating and size. The proposed rotor has better power factor, less value of rated current and greater energy saving potential. Performance indicators such as no load power factor, open circuit induced emf, no load current, the rated current, the rated efficiency, rated power factor, torque angle, maximum torque ability are in favor of proposed motor. The proposed rotor can be used in 2 pole or 4 pole machines just by changing the magnet orientation for the arc magnets.

An optimal design of a new line-start permanent magnet synchronous shaded-pole motor (LSPMSSPM) is proposed in Shamlou S, Mirsalim [48] work. A genetic algorithm optimization method based on transient two-dimensional finite-element method (FEM) is applied to reach a global optimum design. Advantages and challenges of the proposed LSPMSSPM are investigated, and its performance characteristics are calculated. Efficiency, power factor, starting behavior and cost as important key factors are analyzed. FEM results are verified with experimental tests.

In Lu et al. [49] study, the electromagnetic parameters with different bar design parameters are calculated, and the influences of them on the starting performance are also studied. The authors showed with the increase of the bar width or material conductivity, the starting impedance decreases, and thus the starting current and the pull in torque increase. However, there is a maximal value of the starting torque with a specific bar design. Based on the analysis results, the rotor bar design can be optimized for the demand of the higher starting performance of LSPMSM.

CONCLUSION

In this paper research about different aspects of LS-PMS motors presented in literature, in last three decades, were reviewed. To make LS-PMS motor a real competitor to induction motor in a wide variety of applications, more investigations should be done.

In this respect, the starting performance must be considered the primary measure. Also demagnetization of permanent magnets should be considered in design of LSPMS motors. Generally in the design of LSPMS motors, a tradeoff between the motors parameters is needed. In fact the LSPMSM designer has to find many compromises in the design process. The compromise between the value of starting torque, which depends mainly on the squirrel-cage design and material, and the starting current. The compromise between the value of braking torques (due to the presence of PMs in the asynchronous operating region) which depends mainly on the placement, dimensions and the value of energy product of PMs, and motor's synchronization capability. The compromise between an adequate starting characteristic in the asynchronous operating region and the torque capability, power factor and efficiency in the motor's synchronous operating region.

REFERENCES

[1] Isfahani A. H. and Vaez-Zadeh S., "Line-start permanent magnet synchronous motors: Challenges and opportunities," J. Energy, vol. 34, pp. 1755–1763, 2009.

[2] De Almeida, A., Ferreira F .I. T. E., Fong I., "Standards for Efficiency of Electric Motors", IEEE Industry Applications Magazine, Vol 17, No. I, pp. 1 2- 1 9, Jan./Feb. 20

[3] Marcic T, Stumberger G, Stumberger B, Hadziselimovic M, Virtic P. Determining parameter of a line-start interior permanent magnet synchronous motor model by the differential evolution. IEEE Trans Magn vol. 44, no. 11 November. 2008

[4] Abbas A., Yousef H.A., and Sebakhy O.A.; FE Parameters Sensitivity Analysis of an Industrial LS Interior PM Synchronous Motor. Power and Energy Society General Meeting – Conversion and Delivery of Electrical Energy in the 21st Century , 2008 IEEE

[5] Štumberger B, Marčič T, Štumberger G, Hadžiselimović M, and Trlep M. Evaluation of Line-Start Interior Permanent Magnet Synchronous Motor Model Parameters Using Finite Elements. 14th Biannual IEEE Conference on Electromagnetic Field Computation (CEFC) , 2010

[6] Lu X, Lakshmi K., Iyer V, Mukherjee K, and Narayan C. Kar. A Novel Two-Axis Theory-Based Experimental Approach towards Determination of Magnetization Characteristics of Line-Start Permanent Magnet Synchronous Machines. IEEE Transactions on Magnetics, Vol. 49, No. 8, August 2013

[7] Lu X, Lakshmi K., Iyer V, Mukherjee K, and Narayan C. Kar. Development of a Novel Magnetic Circuit Model for Design of Premium Efficiency Three-Phase Line Start Permanent Magnet Machines With Improved Starting Performance. IEEE Transactions on Magnetics, Vol. 49, No. 7, July 2013

[8] Honsinger V.B., Permanent Magnet Machines: Sychronous Operation. IEEE Transactions On Power Apparatus And Systems, Vol. PAS-99, No. 4 July/Aug 1980

[9] Miller T.J.E. Synchronization of Line-Start Permanent-Magnet Ac Motors. IEEE Transaction on Power Apparatus and Systems, Vol. PAS- 103, No. 7, July 1984

[10] Soulard J, Nee H-P. Study of the Synchronization of Line-Start Permanent Magnet Synchronous Motors. IEEE Industry Applications Conference, 2000

[11] Fengbo Q, Z Li, Shukang Ch, Weili L. Calculation and Simulation Analysis on starting performance of the High-voltage Line-start PMSM. 2010 International Conference on Computer application and System Modeling (ICCASM 2010)

[12] Takegami T, Tsuboi K, Hasegawa M, Hirotsuka I, Nakamura M. Calculation Method for Asynchronous Starting Characteristics of Line-Start Permanent Magnet Synchronous Motor. International Conference on Electrical Machines and System (ICEMS) 2010

[13] Stoia D, Chirilă O, Cernat M, Hameyer K, Ban D, The Behaviour of The LSPMSM in Asynchronous Operation, 14th International Power Electronics and Motion Control Conference, EPE-PEMC 2010

[14] Hassanpour Isfahani A., Vaez-Zadeh S., Rahman M. A. Evaluation of Synchronization Capability in Line Start Permanent Magnet Synchronous Motors, 2011 IEEE International Electric Machines & Derives conference (IEMDC)

[15] Hassanpour Isfahani A, Vaez-Zadeh S, Effects of Magnetizing Inductance on Start-Up and Synchronization of Line-Start Permanent-Magnet Synchronous Motors, IEEE Transactions on Magnetics, Vol. 47, No. 4, APRIL 2011

[16] Steluta N, Tudorache T, Ghita C. Influence of design parameters on a line start permanent magnet machine characteristics,13th International conference on optimization of Electrical and electronic equipment (OPTIM) , 2012

[17] Rabbi S. F., Rahman M. A., Determination of the Synchronization Criteria of Line Start IPM motors, IEEE International Conference on Electric Machines and Derives (IEMDC) 2013

[18] Bianchini C, Immovilli F, Bellini A, Davoli M, Cogging Torque Reduction Methods for Internal Permanent Magnet Motors: Review and Comparison, XIX International Conference on Electrical Machines - ICEM 2010, Rome

[19] Yang Y, Wang X, Zhang R, Ding T, and Tang R, The Optimization of Pole Arc Coefficient to Reduce Cogging Torque in Surface-Mounted Permanent Magnet Motors ,

IEEE Transactions on Magnetics, Vol. 42, No. 4, April 2006
[20] Bing-yi Z, Cui G, Wang S, Feng G-H, Study of Cogging Torque in Line-Start PMSM With Proper Fractional Slot Winding , International conference on E- product, E-Service and E-Entertainment (ICEEE) , 2010
[21] Chu W. Q. and Zhu Z. Q., Investigation of Torque Ripples in Permanent Magnet Synchronous Machines With Skewing, IEEE Transactions on Magnetics, Vol. 49, No. 3, March 2013
[22] Bianchini C, Immovilli F, Lorenzani E, Bellini A, and Davoli M, Review of Design Solutions for Internal Permanent-Magnet Machines Cogging Torque Reduction, IEEE Transactions on Magnetics, Vol. 48, No. 10, October 2012
[23] Lee G-H, Kim S-I, Hong J-P, and Bahn J-H, Torque Ripple Reduction of Interior Permanent Magnet Synchronous Motor Using Harmonic Injected Current , IEEE Transactions on Magnetics, Vol. 44, No. 6, June 2008
[24] Kang G-H, HurHyuk Nam J, Hong J-P, and Kim G-T, Analysis of Irreversible Magnet Demagnetization in Line-Start Motors Based on the Finite-Element Method , IEEE Transactions on Magnetics, Vol. 39, No. 3, May 2003
[25] Lu W, Liu M, Luo Y, Liu Y, Influencing Factors on the Demagnetization of Line-start Permanent Magnet Synchronous Motor during Its Starting Process , International Conference on Electrical Machines and System (ICEMS) 2011
[26] Lu X, Lakshmi K., Iyer V, Mukherjee K and Narayan C. Kar, Study and Detection of Demagnetization in Line Start Permanent Magnet Synchronous Machines Using Artificial Neural Network, 15th International Conference on Electrical Machines and System (ICEMS) , 2012
[27] Shen J-X, Li P, Jin M-J, and Yang G, Investigation and Countermeasures for Demagnetization in Line Start Permanent Magnet Synchronous Motors, IEEE Transactions on Magnetics, Vol. 49, No. 7, July 2013
[28] Lee C. K., Kwon B. I., Kim B.-T., Woo K. I., and Han M. G., Analysis of Magnetization of Magnet in the Rotor Of Line Start Permanent Magnet Motor , IEEE Transactions on Magnetics, Vol. 39, No. 3, May 2003
[29] Lee Ch.K. and Byung Il Kwon, Design of Post-Assembly Magnetization System of Line Start Permanent-Magnet Motors Using FEM, IEEE Transactions on Magnetics, Vol. 41, NO. 5, May 2005
[30] Stoia D, Mihai CERNAT, Dănuţ ILEA, Educational Bench of Line-Start Permanent Magnet Synchronous Motors Part I: Operating Point Of Permanent Magnet , The 4th International Conference on Interdisciplinarity in Education ICIE'09, May21-22, 2009, Vilnius, Lithuania
[31] Kurihara K, Wakui G, and Kubota T, Steady-State Performance Analysis of Permanent Magnet Synchronous Motors Including Space Harmonics, IEEE Transactions on Magnetics, Vol. 30, No. 3. May 1994
[32] Zawilak T, Zawilak J, Minimization of Higher Harmonics In Line-Start Permanent Magnetsynchronous Motor, Studia i Materiały Nr 29, 2009
[33] Rong Fu, Dou M, Study on Harmonic Suppression of Damper Windings for Air-gap Magnetic Field of Line-start Permanent Magnet Synchronous Motors , International Conference on Electrical Machines and System (ICEMS) , 2011
[34] Kurihara K and Azizur Rahman M., High-Efficiency Line-Start Interior Permanent-Magnet Synchronous Motors, IEEE Transactions on Industry Applications, Vol. 40, No. 3, May/June 2004
[35] Bingyi Z, Wei1 Z, Fuyu Z, Guihong F, Design and Starting Process Analysis of Multipolar Line-Start PMSM , Proceeding of International Conference on Electrical Machines and Systems 2007, Oct. 8~11, Seoul, Korea
[36] Yang G, Ma J, Shen J-X, and Wang Y, Optimal Design and Experimental Verification of a Line-Start Permanent Magnet Synchronous Motor, International Conference on Electrical Machines and System (ICEMS), 2008
[37] Lu QF and Ye YY, Design and Analysis of Large Capacity Line-Start Permanent-Magnet Motor, IEEE Transactions on Magnetics, Vol. 44, No. 11, November 2008
[38] Xiaochen Z, Shukang C and Weili L, Development of Line-start PMSM with Solid Rotor for Electric Vehicles , IEEE Vehicle Power and Propulsion Conference (VPPC), September 3-5, 2008, Harbin, China
[39] Kim W-H, Kim K-Ch, Kim S-J, Kang D-W, Go S-Ch, Lee H-W, Chun Y-D, and Lee J, A Study on the Optimal Rotor design of LSPM Considering the Starting Torque and Efficiency, IEEE Transactions on Magnetics, Vol. 45, No. 3, March 2009
[40] Peralta-Sánchez E and Smith A. C., Line-Start Permanent-Magnet Machines Using a Canned Rotor, IEEE Transactions on Industry Applications, Vol. 45, No. 3, May/June 2009
[41] Fei W., P. C. K. Luk, J. Ma, J. X. Shen, and G. Yang, A High-Performance Line-Start Permanent Magnet Synchronous Motor Amended From a Small Industrial Three-Phase Induction Motor , IEEE Transactions on Magnetics, Vol. 45, No. 10, October 2009
[42] Stoia D, Cernat M, Jimor AA., Nicolae DV., Analytical Design and Analysis of Line-Start Permanent Magnet Synchronous Motors, IEEE AFRICON 2009 23 - 25 September 2009, Nairobi, Kenya
[43] Yaojing F, and Kai Y, A Two-pole High-power Line-start Permanent Magnet Synchronous Motor, International Conference on Electrical Machines and System (ICEMS), 2008
[44] Jazdzynski W., Bajek M., Modeling and Bi-Criterial Optimization of a Line Start Permanent Magnet Synchronous Machine to Find an IE4 Class High-Efficiency Motor , XIX International Conference on Electrical Machines - ICEM 2010, Rome
[45] Feng X., L. Liu, J. Kang, Y. Zhang , Super Premium Efficient Line Start-up Permanent Magnet Synchronous Motor, XIX International Conference on Electrical Machines - ICEM 2010, Rome
[46] Ruan T, Haipeng Pan Yongming Xia , Design and Analysis of Two Different Line-Start PM Synchronous Motors, 2nd International Conference on Artificial Intelligence Management Science and Electronic Commerce (AIMSCE) , Aug. 2011
[47] Ugale R. T. and Chaudhari B. N., A New Rotor Structure for Line Start Permanent Magnet Synchronous Motor , IEEE International Conference on Electric Machine & Drives (IEMDC), 2013
[48] Shamlou S, Mirsalim M, Design, optimisation, analysis and experimental verification of a new line-start permanent magnet synchronous shaded-pole motor , IET Electric Power Applications , 2012

[49] Lu W, Luo Y, and Zhao H, Influences of Rotor Bar Design on the Starting Performance of Line-start Permanent Magnet Synchronous Motor , 6th International Conference on Electromagnetic and field Problems and Applications (ICEF) , 2012

Optimum Economic Scheduling Strategy of Islanded Multi-Microgrid

Chinmaya Ranjan Pradhan, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, cr_pradhan@outlook.com*

Subhendu Sekhar Sahoo, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, subhendusahoo2@gmail.com*

Sunita Baral, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, sunita.baral95@gmail.com*

Sanjay Kumar Nayak, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, sk.nayak24@gmail.com*

Abstract *– The paper presents a new methodology of the multi-microgrid (MMG) system is developed for large-scale integration of microgrids (MGs) and distribution generation (DG) units. The different controllable MGs, DG units and loads for MMGs system requires an efficient dispatch strategy in order to balance supply demand for optimizing the total cost of the integrated system. This paper described an optimum economic dispatch strategy of islanded MMGs. To optimize the system operating and running cost, genetic algorithm have been used and searches the optimum value of the output parameters like power produced by the controllable DG. The objective function comprises input fuel cost, operation cost, as well as the cost of emissions subject to various system constraints. The proposed optimization process is applied to a newly designed MMG system that has been operated under various constraints. Simulation results guaranteed the validity the proposed optimization method.*

Keywords*: Multi-microgrid, Distributed Generation, Cost Function, Economic Dispatch, Genetic Algorithm.*

INTRODUCTION

Microgrid (MG) is a low voltage grid that integrates various distributed generation (DG) unit and energy storage devices for supplying power to load at distribution level. It can be operated both in on-grid or islanding mode. In islanding mode, MGs have restricted energy handling capability. A single MG can supply only a highest load capacity of approximately 10 MVA. However several MGs can be interconnected together to form larger power pool to meet the greater power demands. It also has more redundancy and ensures better supply reliability [1]. The interconnected MGs are generally called multi-microgrid (MMG) or integrated MG; it is a relatively new concept [2]. MMG system not only connect several individual MGs and but also other distributed generations to a medium voltage distribution grid. Normally, an MMG is operated to the high-voltage grid and for emergency purpose it may also be operated in islanded mode that is completely isolated from grid [3]. Several researches have been studied about the MMG system.

Gil and Pecas Lopes [2] proposed a robust frequency control methods for an MMG in islanded operation. Research [4] optimizing the operation cost of each individual MGs in a real grid. In work of Rua et al. [5] communications uncertainty is considered isolated mode in a MMGs operation. A novel methodology to implement a telephone line for communication and reliable control is presented in Arefifar et al. [6] work, considering the MGs building block methodology. The bi-level programming has been used for analysing the competitive state of major decision making among an energy services providers representing various MGs is presented in work of Georgia et al. [7]. The study of Rua et al. [8] analysed the impact of liaison in frequency and power control in multi-MGs systems in islanded mode. The allocation problem of numerous MGs by considering installation investment, optimal design and operation, and power losses is elaborately explained and suggest some way to improve those problems in work of Yang et al. [9]. The technical and commercial management strategy and state calculation has been developed for MMGs in Madureira et al. [10]. The paper of Li et al. [11] described an energy return plan methodology between a power grid and MMGs system. A multiobjective algorithm for improving the power flow controller performance of MMGs which minimizes MMGs operating cost, power loss, and all buses voltage profile fluctuation have been discussed in work of Kargarian et al. [12]. In Gregoratti and Matamoros [13]'s work, an arbitrary topology has been used in a distributed convex MGs optimization network for energy exchanging between islanded MGs that exchanged energy flows. An advanced control system can be used at medium to high voltage grid substations and can be used to manage micro-generation with load parameters in work of Vasiljevska et al. [14].

Optimal sizing estimation of distributed energy storage devices for integrated MGs is discussed in work of Logenthiran et al. [15]. For better utilization of

renewable energy, to reduce production cost (30%) a new algorithm for smart intelligent home energy management system with consumption shifting in demand response program considering various constraints has been proposed in work of Mirhosseini Moghaddam et al. [16] at a islanded mode. But the optimum sizes of energy storage system and real time implementations are not mentioned. In a MG, renewable sources, generators heat and output powers optimum combination is a major problem. To overcome multi-objectives optimization problems, the modified particle swarm optimization (PSO) called as neighbourhood re-dispatch PSO algorithm has been proposed in work of Si et al. [17]. The major objectives were to reduce electricity cost with minimum costs avoiding other parameters. In work of Chen et al. [18] a matrix perturbation theory based distributed optimization dispatch algorithm has been proposed to determine the optimal DG outputs and that also satisfied the supply, consumer demand constraints.

However, the researchers did not sufficiently study about the optimum economic dispatch strategy of islanded MMGs. In this paper genetic algorithm searches the optimum magnitude of the outturn power produced by the dispatchable distributed generator subject to minimization of operating cost. The objective function formulated the input fuel cost, operation cost, and emissions cost subject to various system constraints. The proposed optimization process is implemented on a typical MMG system.

CONTROL STRATEGY

There is no fixed or predefined structure of MGs control system; it depends on the types of configuration and MGs architecture [19]. Furthermore, the MMG system increases the system complexity. The centralized control architecture is discussed here [2, 10, 14, 19] with few modifications. The hierarchical control architecture comprises the following controller as shown in Fig.1: (i) Central power monitoring system (CPMS), (ii) Power management controller (PMC), (iii) Load management controller (LMC), and (iv) Microgrid controller (MGC).

The central power monitoring system (CPMS) is the decision maker and responsible for the economical optimization of the integrated system. It globally minimizes the total operating cost. It is aware about the characteristics of all the micro sources with their respective operational limits, controllable load and state of charge (SOC) of energy storage system (EES). The total load is observed and the supply demand is balanced according to operating strategy of the system.

The CPMS globally runs and sends the control signal to MGC, PMC and LMC for changing the power level under their controlling devices to balance the supply demand. PMC and LMC operate at medium voltage level. PMC controls the power of distributed generation units and EES. LMC monitors the controllable load by load shedding. Each of the MGCs will share out the power changes among its DG units and controllable loads at each MG.

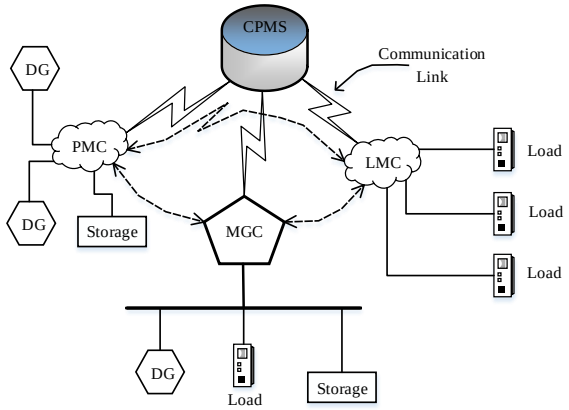


Figure 1. Hierarchical control structure of integrated microgrid

MODEL OF THE STUDIED SYSTEM

The schematic diagram of the studied MMGs system is shown in Fig.2. The adopted test network represents the architecture of a MV grid containing three MGs, several kinds of larger DGs and controllable loads. The MG1 consists with WTG, FC, ESS, MG2 with FC, DG, and MG3 with MTG and FC. Each MG has controllable load. MTG, WTG, ESS and some controllable loads is connected at medium voltage level. Two wind turbine generators is connected at MG1. At medium voltage level three wind turbine generators delivers power. The capacity and operational limits of distribution energy resources is given in Table 1.

The supply demand balance can be written by the following eq.(1) from Fig.2.

$$\sum P_{L} - \sum \left(P_{DEG} + P_{MTG} + P_{FC} + P_{WTG} + P_{PV} \pm P_{ESS} \right) = 0 \quad (1)$$

Table 1. Capacity and Limits of Microsources

Position	Micro-sources	Capacity (kW)	Lower Limit (kW)	Upper Limit (kW)
MG1	WTG	200	0	200
	FC	200	40	200
	ESS	50	-	-
MG2	FC	200	40	200
	DEG	250	45	250
MG3	MTG	200	30	200
	PV	300	0	300
MV	MTG	200	30	200
	WTG	300	0	300
	ESS	100	-	-

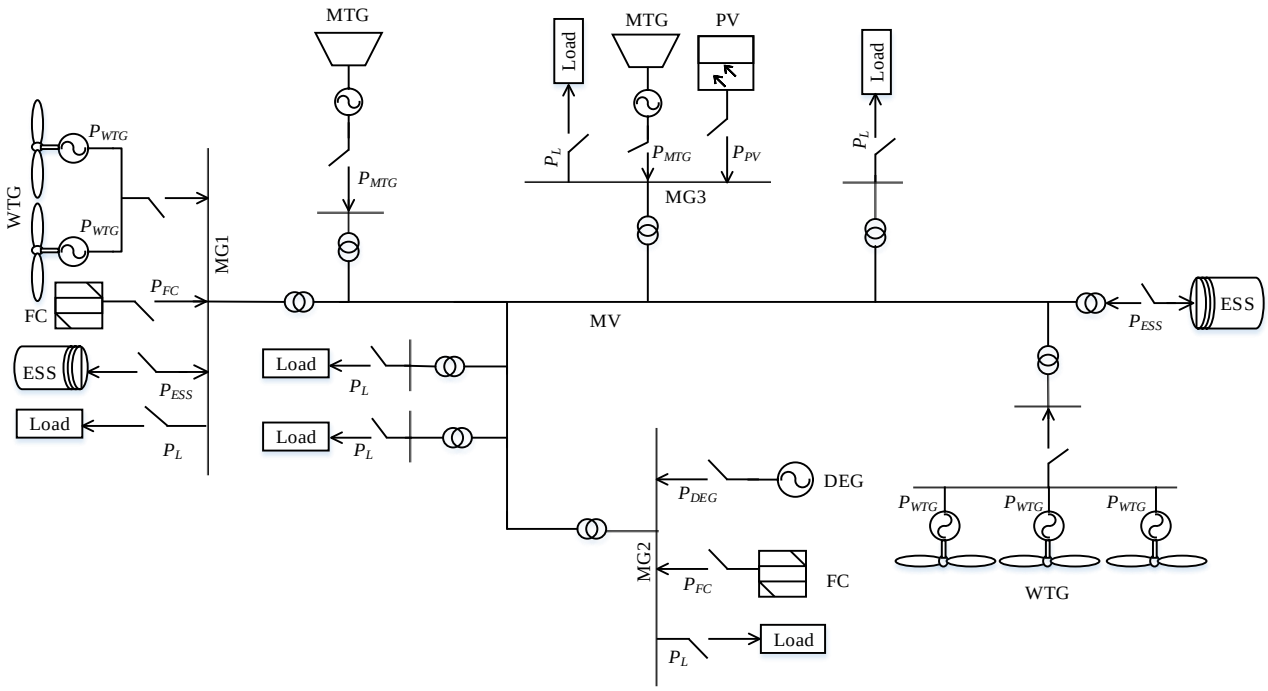


Figure 2. Schematic diagram of integrated microgrid

SYSTEM MODELLING

The modelling of different parts of the integrated MG system is presented in this section. The fuel input is provided only for the DEG, MTG and FC as PV energy comes from the nature. The energy storage system can be charged by producing electrical power from PV and WTG. Each part of the integrated system is designed separately based on its properties. The features of some equipment's are available from the manufacturer.

A. Diesel Engine Generator(DEG)

The diesel engine generator efficiency decreases at light load condition, and the fuel expenditure is almost full. Therefore, it is needed to fix up the minimum output power magnitude of DG [20]. The minimum loading capacity of a DG is limited to 30-50% [21] and the optimum operating scale is 70-89% from the rated power [22]. The fuel consumption rate (litre/hour) of a 250 kW DEG shown in Fig.3 is used for simulation [23].

$$F_{DEG} = \alpha P_{DEG}^2 + \beta P_{DEG} + \gamma \quad (2)$$

The DEG should be operated economically to control the governing system so that the generation costs will be lower.

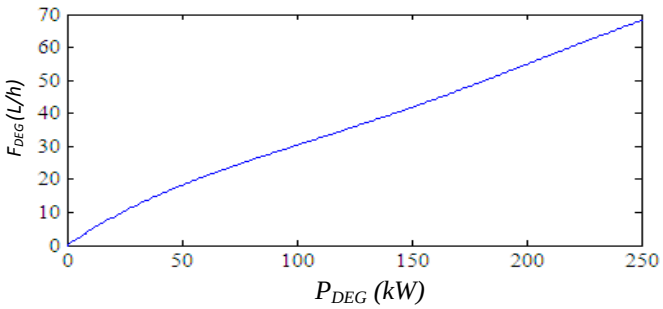


Figure 3. Output power vs. fuel consumption rate of DEG

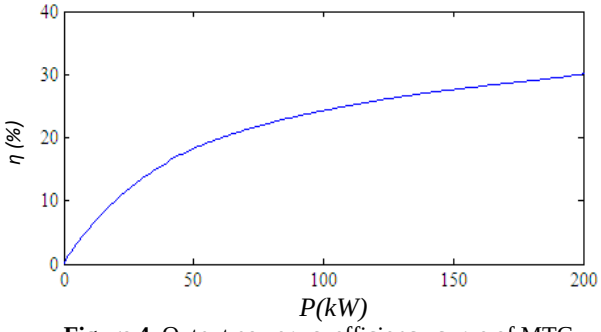


Figure 4. Output power vs. efficiency curve of MTG

A. Microturbine Generator (MTG)

The efficiency of the MTG increases with the increase of the supplied power. The typical efficiency curve of a 200 kW micro turbine is modelled as shown in Fig.4 [24].

The fuel input for a microturbine can be expressed as [25],

$$F_{MTG} = \sum_J \frac{P_J}{\eta_J} \quad (3)$$

P_J = Net power generated at interval J (kW).

η = Microturbine avail at interval J.

The minimum and maximum loading constraint of MTG is given by eq.(4).

$$P_{MTG}^{Min} \le P_{MTG} \le P_{MTG}^{Max} \quad (4)$$

B. ***Fuel Cell (FC)***

The competency of any fuel cell can be described as follows [26] where all unit must be in the same scale.

$$\eta_{FC} = \frac{Electrical\ Power\ Output\ (P_{FC})}{Fuel\ Input\ (F_{FC})} \quad (5)$$

The typical output power versus efficiency curve of a typical 200 kW fuel cells is shown in Fig.5.

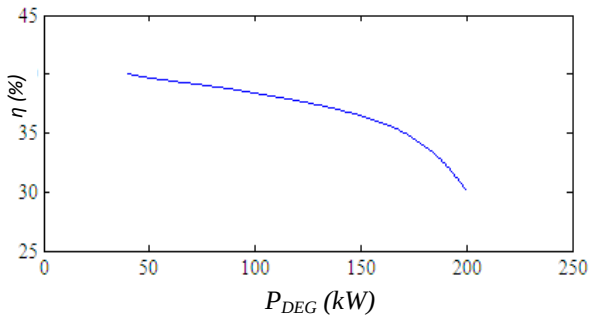


Figure 5. Fuel cell output power vs efficiency curve

The fuel input for the cell can be expressed as [25]:

$$F_{FC} = \sum_J \frac{P_J}{\eta_J} \quad (6)$$

Where the symbols represents their usual meaning

The minimum and maximum loading constraint of FC is given by eq.(7).

$$P_{FC}^{Min} \le P_{FC} \le P_{FC}^{Max} \quad (7)$$

C. ***Wind Turbine Generator (WTG)***

It is very common things that the speed of wind changes in every hours, days and seasons. For planning long term the wind distribution can be represents by by Weibull distribution functions as follows [27]:

$$f_v(v) = \begin{cases} \frac{\beta}{\alpha} \times \left(\frac{v}{\beta}\right)^{\rho-1} \times e^{\left(\frac{\gamma}{\alpha}\right)^{\beta}} & v \ge 0 \\ 0 & Otherwise \end{cases} \quad (8)$$

Where, $\alpha, \beta \& v$ are the shape parameter, scale parameters of Weibull function and wind speed, respectively.

The output of WTG The performance curve of WTG can be approximated as a function of wind speed (V_w). A third order polynomial function is used to fit the parameters on wind speed and wind turbine performance curve. By using the following expression, the generated output power of WTG can be determined.

$$P_{WTG} = \begin{cases} 0, & V_w < V_{Cut-in} \\ aV_w^3 + bV_w^2 + cV_w + d, & V_{cut-in} \le V_w < V_r \\ P_{Rated}, & V_r \le V_w \le V_{cut-out} \end{cases} \quad (9)$$

Where $V_{cut\text{-}in}$= Cut in speed, V_r=Rated speed and $V_{cut\text{-}out}$=Cut out speed.

The power curve of a 100kW turbine is shown in Fig.6 is used in this model [28]. The input wind speed is considered for this model is shown in Fig.7.

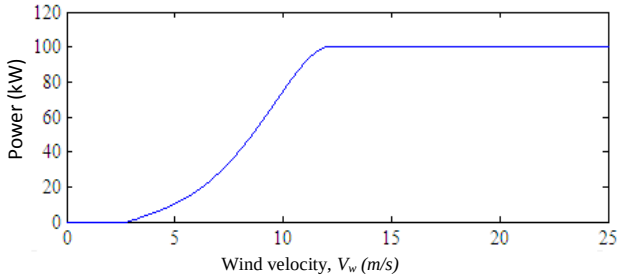


Figure 6. Power curve of a *100 kW* turbine

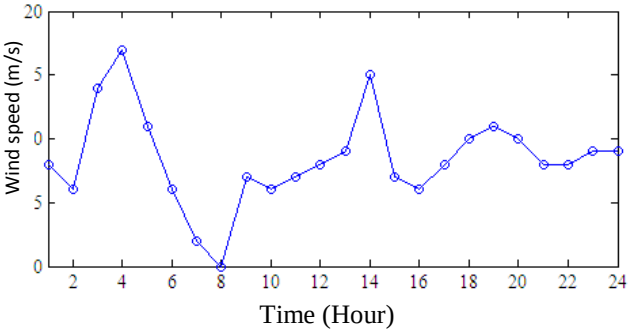


Figure 7. The input wind speed as used in the model

D. ***Solar Photovoltaic (PV) System***

The output power of solar photovoltaic depends on environmental conditions, such as solar radiation and temperature, resulting in a non-linear and time-variant power source. Depending in the solar radiation and load current, the output power of the PV module will be changed that follows the equation (10) [29],

$$P_{PV} = \eta A \phi [1 - 0.005(T_a + 25)] \quad (10)$$

Where, S is the area of PV array (m^2), φ presents the solar irradiation (W/m^2) and T_a is ambient temperature (^0C). Our Studied system has the following parameters: $A=2000m^2$, $\eta=20\%$ and we assume temperature is constant ($T_a=25^0C$). The solar irradiation data for 24 hour period used in this model is given in Fig.8.

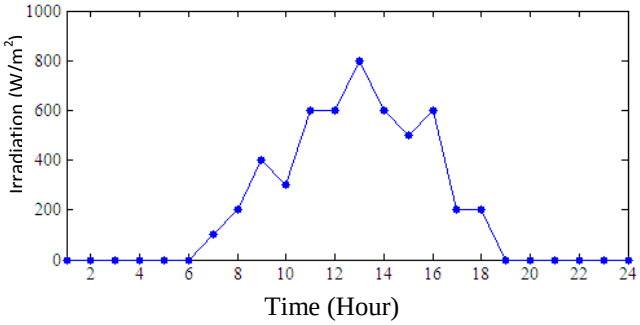


Figure 8. The input irradiation data for simulated model

*E. **Energy Storage System (ESS)***

Energy storage systems are an important part for the hybrid power system and effectively supply deficit power to maintain the system stability [29]. Energy can be stored in many ways like use of electro-chemical battery, super-capacitors, super conducting magnetic energy storage, flywheel storage system and many more. The state of charge (*SOC*) of ESS should be monitored and kept in acceptable ranges so that it will not be overcharged the battery [30]. The minimum and maximum *SOC* of ESS should be limited by eq. (11).

$$SOC_{ESS}^{Min} \le SOC_{ESS} \le SOC_{ESS}^{Max} \quad (11)$$

For safety operation of ESSs, battery management system can be applied. In this system, batteries voltage, current and temperature is controlled from centrally. To reduce the cost of output energies, proper allocation of EES batteries is an important factor. To design a smart MG, to keep balance or optimization between source & load demand, to maintain the limit of SOC and to design an optimum sizing of EESs are an important consideration.

OBJECTIVE FUNCTION FORMULATION

The main objective function includes operating fuel cost, maintenance cost and cost of emission released from each microsource [31].

$$CF = \sum_{i=1}^{N}\left(C_i F_i + OM_i\right) + \sum_{i=1}^{N}\sum_{j=1}^{M}\alpha_i EF_{ij} P_i \quad (12)$$

Where,

C_i Fuel costs, i in \$/L for the diesel, and \$/kWh for the gas.
F_i Rate of fuel expense, i in L/h for DG, and kW/h for the FC and MT.
OM_i Operation cost, i in \$/h
α_j Cost of emission, j
EF_{ij} Emission factor i, emission type j
M Emission types (NO_x or CO_2 or SO_2)
N Number of generating units i

Diesel is used in DEG but the fuel used in MTG and FC is natural gas. No fuel is required for WTG and PV. The operation cost of the generating unit *i* (OM_i) is considered be constant and proportional to the produced energy [32], where (K_{OM}) represents the proportional constant.

$$OM_i = K_{OMi} P_i \quad (13)$$

The magnitude of the K_{OM} for various generation units are listed in Table II.

The emission usually includes gases such as SO_2, CO_2 and NO_X. The costs and factors of emision of the DEG, FC, and MTG used here are listed in Table 3 [31].

TABLE 2. PROPORTIONAL CONSTANT

DG	DEG	MTG	FC	WTG	PV
K_{OM}(\$/kWh)	0.0125	0.0060	0.0050	0.0150	0.0010

TABLE 3. EXTERNALITY COSTS AND EMISSION FACTORS FOR NO_X, SO_2, AND CO_2

Emission Type	Externality cost (\$/kg)	Emission factors (Kg/MWh)		
		DEG	MTG	FC
CO_2	0.014	1.432	1.596	1.078
SO_2	0.99	0.454	0.008	0.006
NO_x	4.2	21.8	0.44	0.03

Constraints: To balance the real power, and the load demand the belows equation (1) is used.

The output power of generator unit *i* (P_i) is restricted to its maximum and minimum value.

$$P_i^{Min} \le P_i \le P_i^{Max} \quad (14)$$

Also the SOC of the energy storage system is properly controlled as Eq. (10).

OPTIMAL OPERATING STRATEGY

The genetic algorithm is applied to find the optimal output of dispatchable distributed generator for MMG system with minimum operating cost as described by objective function in the previous section. The implementation strategies are as follows:

1. The central power monitoring system calculates the total load demand (P_L).
2. Output power of WTG is calculated from the performance curve.
3. Output power of solar PV is determined from solar radiation .
4. Reduce the total load from WTG and PV power.

$$\Delta P_L = P_L - (P_{WTG} + P_{PV}) \quad (15)$$

If $\Delta P_L < 0$, the rest of the power will be given to the battery to charge the ESS. When the ESS is fully charged, the exceed power is unload

If $\Delta P_L > 0$, the remaining power will be given by the ESS or by the distributed generator (DEG, MTG, FC. Meanwhile, the charging and discharging of the ESS is properly monitored. The tuning of output power of DGs by genetic algorithm occurs in the following ways.

1. Initialization: The algorithm begins by creating an initial population. This population is normally randomly reproduced at any desired size.
2. Evaluation: The fitness value of chromosomes is now evaluated by calculating the cost function or objective function. Here, it is tried to find the minimum magnitude of the cost function.
3. Selection: Selection helps to discard the weak individuals and only keeps the best individuals called parents that contribute to the population at the next generation. There are a small number of selection methodologies but the primary concepts is the same, make it more likely the best adjuster individuals will be selected for our upcoming generation.
4. Crossover: In this stage, new individuals are created by combining prospective of chosen individuals. By combining two or more individuals it will create a fitter offspring from each of its parents.
5. Mutation: Mutation typically works by making very small changes at random to an individual's genome.
6. Termination: Now the next generation is started again from step two until it reaches a maximum number of generations.

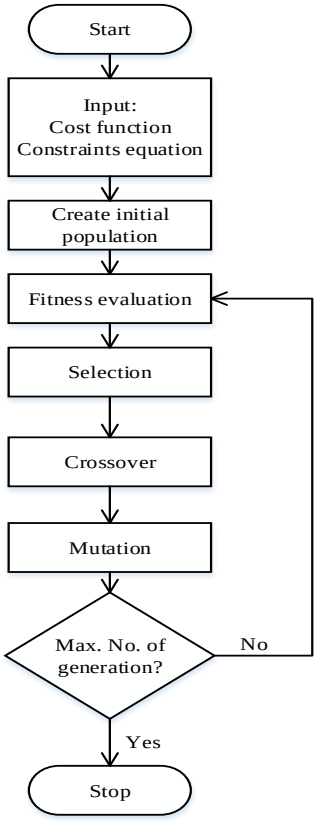


Figure 9. Flow chart for genetic algorithm

RESULTS AND DISCUSSION

The total load profile for the studied integrated system used for simulation purpose is shown in Fig.10. The scheduling time is 24 hours in a day with the scheduling time interim of 1 hour. The load demand varies between 430kW to 1500kW. The optimization model that is discussed in the past section is implemented to this time-varying load.

The total renewable power i.e. summation of total power from WTG and solar PV is shown in Fig.11. Also, WTG and solar PV power profile is given to the following figures according to their location in system. After observing the total load and renewable power of the system, the genetic algorithm is used to find out the optimum power generation of microsources subject to minimum operating cost according to control strategy. The various combination of produced power is shown from Fig.12 to Fig.15.

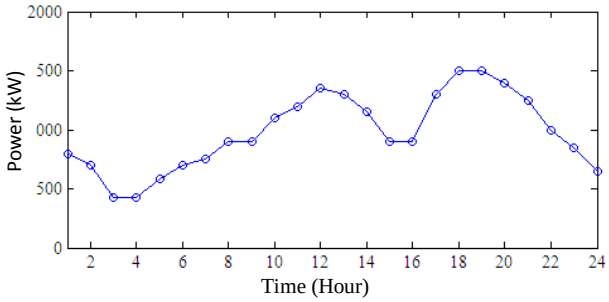


Figure 10. Total load profile of the simulated system

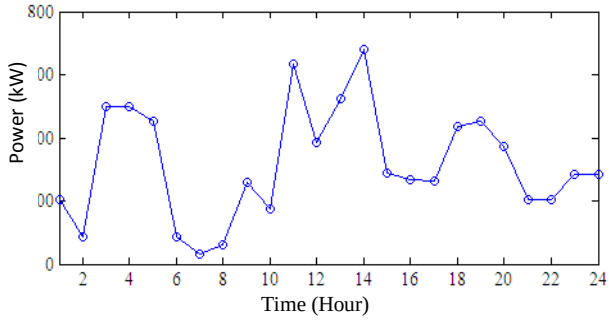


Figure 11. Total renewable power of the simulated system

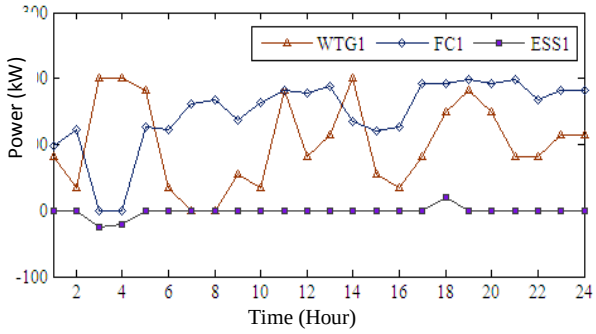


Figure 12. Power profile at MG 1

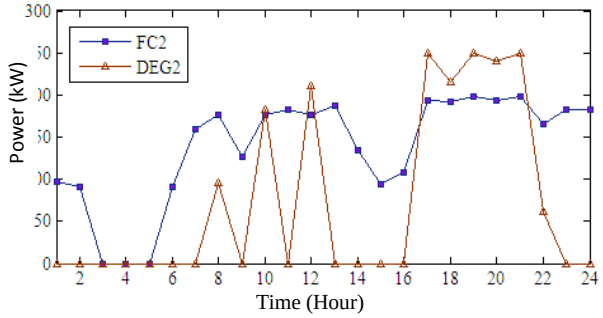


Figure 13. Power profile at MG 2

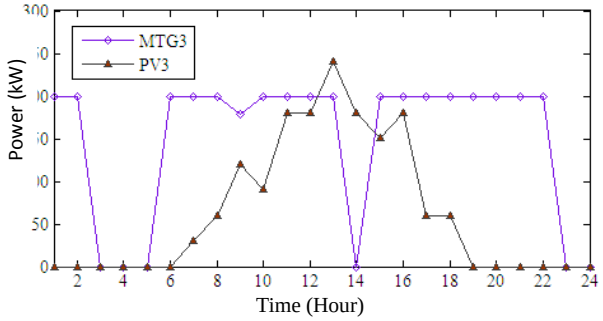


Figure 14. Power profile at MG 3

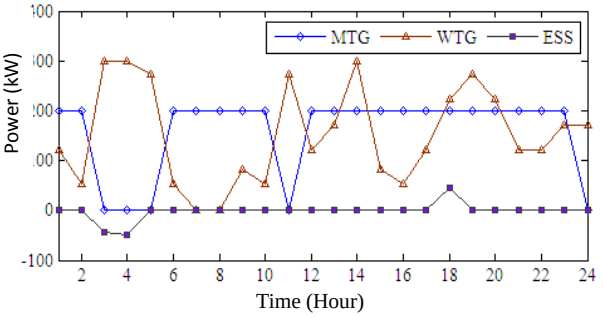


Figure 15. DGs and ESS at MV level

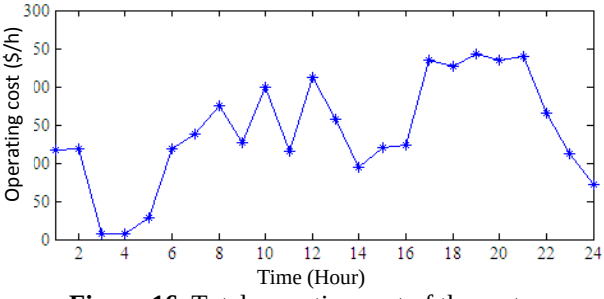


Figure 16. Total operating cost of the system

The Fig.12 to Fig.14, show the output power profile of MG1 to MG3 respectively. The energy supplied or absorbed by the ESS to or from the system is given in the figures. The total operating cost of the system is shown in Fig.16.Only controllable DGs such as DEG, FC and MTG can be tuned. The total renewable power generation of WTG and solar PV are given the first priority to meet the load demand. If the output power taking from PV and WTG is lesser than the load demand the genetic algorithm automatically finds various combination of DGs for producing power to minimize the operating cost and balance the load demand. Also the energy from ESS is properly exchanged. Fig.12 and Fig.15 show that the ESS is charged and can save the surplus power at time 3.00 and 4.00 due to available of wind power. During this period all the controllable DGs is shut down that is observed from Fig.12 to Fig.15. In this case the total operating cost is treated by only operation and maintenance cost of wind turbine generation. As a result the cost of the system is very low as shown in Fig.16. But at 18.00, the ESS is discharged the total load is balanced.

It can be seen from Fig.13 that the diesel engine generator is the least preferred generator for delivering power because of its higher cost. Due to low cost of MTG and FC, they are firstly selected for power production purpose to meet the load demand. When the DEG delivers power, the operating cost of the system increases. So if the load demand is low, the best selection of distributed generator in terms of operating cost is to switched off the diesel generator. It is used only when there are no other generation options is available. Due to time varying behaviour of load demand the start-stop cycles of DGs increases. Furthermore, the uncertain nature of wind and solar power increases the system complexity.

CONCLUSION

Optimum economic dispatch strategy for the islanded MMGs has been proposed in this paper. Genetic algorithm searches the minimum value of cost function by properly selecting the output power produced from controllable DGs. The objective function consists of operating fuel cost, maintenance cost and cost of emission released from each microsources. Practically, the power required for the connected loads can be effectively supplied with appropriate coordination among distributed generators and energy storage system for such type of system. But in the proposed method start-stop cycles of DGs increases. As a result the forecasted load, wind speed and solar irradiation data should be included to optimum dispatch strategy to reduce the start-stop cycle of DGs. The simulation results justify the correctness of the proposed algorithm. This research can be further modified by proposing new algorithm to optimize production cost that will also satisfy the supply and consumer demand constraints.

REFERENCES

[1] S. Chowdhury, S. P. Chowdhury, P. Crossley. *Microgrids and Active Distribution Networks: Renewable Energy Series 6.* London, United Kingdom: The Institution of Engineering and Technology, 2009.

[2] N. J. Gil, J. A. P. Lopes. Hierarchical Frequency Control Scheme for Islanded Multi-Microgrids Operation. *Lausanne Power Tech Lausanne*; 2007 July 1-5; Lausanne. IEEE; 2007. 473-478p.

[3] S. A. Gopalan, V. Sreeram, H.H.C. Iu, Z. Xu, Z.Y. Dong, K. P. Wong. Fault Analysis of an Islanded Multi-Microgrid. *Power and Energy Society General Meeting*. 2012 July 22-26; San Diego, CA.IEEE; 2012. 1-6p.

[4] W. Xi, Q. Xiaoyan, J. Runzhou, et al. Economic Operation of Multi-Microgrids Containing Energy Storage System. *International Conference on Power System Technology (POWERCON).* 2014 October 20-22; Chengdu. IEEE; 2014. 1712 – 1716p.

[5] D. Rua, J. A. Peças Lopes, J. Ruela. Communications Uncertainties in Isolated Multi-Microgrid Control Systems. *Power Systems Computation Conference (PSCC).* 2014 August 18-22; Wroclaw. IEEE; 2014. 1– 7p.

[6] S. A. Arefifar, Y. Abdel-Rady I. Mohamed, Tarek El-Fouly. Optimized Multiple Microgrid-Based Clustering of Active Distribution Systems Considering Communication and Control Requirements. *IEEE Transactions on Industrial Electronics*, vol. 62, no.2, pp. 711–723.

[7] Georgia E. Asimakopoulou, Aris L. Dimeas, Nikos D. Hatziargyriou. Leader-Follower Strategies for Energy Management of Multi-Microgrids. *IEEE Transactions on Smart Grid*. 2013 ; 4(4) : 1909 – 1916p.

[8] D. Rua, L. F. Moura Pereira, N. Gil, et al. Impact of Multi-Microgrid Communication Systems in Islanded Operation. *2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe)*; 2011 December 5-7; Manchester. IEEE; 2011. 1– 6p.

[9] Y. Yang, W. Pei, H. Xiao, et al. Comprehensive planning of Multiple Microgrids with Self-Healing Consideration. *International Conference on Power System Technology (POWERCON);* 2014 October 20-22; Chengdu. IEEE; 3275–3281p.

[10] A. G. Madureira, J. C. Pereira, N. J. Gil, et al. Advanced Control and Management Functionalities for Multi-Microgrids. *European Transactions on Electrical Power*. 2011; 21(2): 1159–1177p.

[11] P. Li, X. Guan, J. Wu, et al. An Integrated Energy Exchange Scheduling and Pricing Strategy for Multi-Microgrid System. *TENCON 2013-2013 IEEE Region 10 Conference (31194);* 2013 October 22-25; Xian. IEEE; 2013. 1-5p.

[12] A. Kargarian, B. Falahati, Yong Fu, et al. Multiobjective Optimal Power Flow Algorithm to Enhance Multi-Microgrids Performance Incorporating IPFC. *IEEE Power and Energy Society General Meeting*; 2012 July 22-26; San Diego, CA. IEEE; 2012. 1-6p.

[13] D. Gregoratti, J. Matamoros. Distributed Energy Trading: The Multiple-Microgrid Case. *IEEE Transactions on Industrial Electronics.* 2015; 62(4): 2551–2559p.

[14] J. Vasiljevska, J.A. Pecas Lopes, M.A. Matos. Integrated Micro-Generation, Load and Energy Storage Control Functionality under the Multi Microgrid Concept. *Electric Power Systems Research.* Elsevier. 2013; 95: 292–301p.

[15] T. Logenthiran, D. Srinivasan, A. M. Khambadkone et al. Optimal Sizing of Distributed Energy Resources for Integrated Microgrids Using Evolutionary Strategy. *IEEE Congress on Evolutionary Computation (CEC); 2012 June* 10-15; Brisbane, QLD. IEEE; 2012. 1-8p.

[16] M. Mirhosseini Moghaddam et al., Optimal Energy Management for a Home Microgrid Based on Multi-Period Artificial Bee Colony, *25th IEEE Iranian Conference on Electrical Engineering (lCEE2017)*, Iran, 2017,1446-1451p, DOI: 10.1109/IranianCEE.2017.7985270.

[17] F. Si1, J. Wang et al., A Multi-Objective Optimization Strategy for Combined Heat and Power Systems of the Energy Internet, *2017 29th Chinese Control And Decision Conference (CCDC),* 28-30 May 2017, China, **DOI:** 10.1109/CCDC.2017.7978735.

[18] G. Chen, Z. Li, Z. Liu, A Distributed Solution of Economic Dispatch Problem in Islanded Microgrid Systems, *2016 IEEE Chinese Control and Decision Conference (CCDC),* 28-30 May 2016, China, 6804-6809p, **DOI:** 10.1109/CCDC.2016.7532223.

[19] N. Hatziargyriou. *Microgrids Architectures and Control*. United Kingdom: John Wiley and Sons Ltd; 2014.

[20] Ch. Wang, M. Liu, L. Guo. Cooperative Operation and Optimal Design for Islanded Microgrid. *IEEE PES Innovative Smart Grid Technologies (ISGT)*; 2012 January 16-20; Washington, DC. IEEE; 2012. 1-8p.

[21] F. Katiraei, C. Abbey. Diesel Plant Sizing and Performance Analysis of a Remote Wind-Diesel Microgrid. *IEEE Power Engineering Society General Meeting*; 2007; Tampa, FL. IEEE; 2007. 1-8p.

[22] Said H. El-Hefnawi. Photovoltaic Diesel-Generator Hybrid Power System Sizing. *Renewable Energy, Elsevier Science Ltd.* 1998; 13(1) : 33-40p.

[23] Diesel Service & Supply website [Internet]. (cited: 15th march 2015) Available from: http://www.dieselservice andsupply.com/.

[24] Capstone turbine corporation website [Internet]. (cited: 15th march 2015) Available from: http://www.capstoneturbine.com/

[25] F. A. Mohamed, H. N. Koivo. System Modelling and Online Optimal Management of Microgrid Using Mesh Adaptive Direct Search. *Electrical Power and Energy Systems, Elsevier.* 2010. 32(5) : 398-407p.

[26] F. Barbir, T. Gomez. Efficiency and Economics of Proton Exchange Membrane (PEM) Fuel Cells. *International Journal of Hydrogen Energy*. 1997; 22(10/11): 1027-1077p.

[27] N. Nikmehr and S. Najafi Ravadanegh, Optimal Power Dispatch of Multi-Microgrids at Future Smart Distribution Grids, *IEEE Transactions on Smart Grid*. 2015; 6(4): 1648 – 1657p, **DOI:** 10.1109/TSG.2015.2396992.

[28] Polaris America Turbines website [Internet]. (cited 20th march 2015) Available from: http://www.polarisamerica.com /turbines /100kw-wind-turbines/.

[29] D. J. Lee, L. Wang. Small-Signal Stability Analysis of an Autonomous Hybrid Renewable Energy Power Generation/Energy Storage System Part I: Time-Domain Simulations. *IEEE Transactions on Energy Conversion.* 2008; 23(1) : 311-320p.

[30] S. Chanana, A. Kumar. Operation and control of BESS using frequency-linked pricing in real-time market with high wind penetration. *International Journal of Energy Sector Management, Emerald.* 2011; 5(4) : 585-602p.

[31] H. Vahedi, R. Noroozian, S. H. Hosseini. Optimal Management of Microgrid Using Differential Evolution Approach. *7th International Conference on the European Energy Market (EEM)*; 2010 June 23-25; Madrid. IEEE; 2010. 1-6p.

[32] A. M. Azmy and I. Erlich. Online Optimal Management of PEM Fuel Cells Using Neural Networks. *IEEE Transactions on Power Delivery*.2005; 20(2) : 1051–1058p.

Performance Evaluation of DFIG to Changes in Network Frequency

Somnath Mishra, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, somnathmishra@yahoo.co.in*

Satyajit Nayak, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, satyajit_nayak@gmail.com*

Balagoni Sampath Kumar, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, bs.kumar456@gmail.com*

Soumya Datta Mohanty, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, soumya_datta.mohanty@gmail.com*

ABSTRACT

Inertia response is the initial response of a generator to the frequency changes of a network and its presence in network is a matter of importance. Synchronous generators have high inertia response naturally. The more tendencies to renewable energies, the more usage of wind power plants in power systems. Due to advantages of doubly fed induction generator (DFIG), this kind of generator constitutes the most wind generators in power plants. Unlike the synchronous generators, DFIGs have negligible inertia response which is caused by generator controller operation. Thus, the evaluation of controller type on its performance is necessary. In this paper, a DFIG with two type of controllers (modified and unmodified) has been simulated in MATLAB/SIMULINK. Besides, the generator operation with network frequency droop conditions has been surveyed and compared as well as the effect of mentioned controllers' action speed on generator's inertia response. The Results showed that the inertia response can be improved significantly by modifying the generator controller system.

Keywords
DFIG,
Inertia Response,
Controller Performance,
Renewable Energy

INTRODUCTION

Generators and loads connected to power systems in the entire world rely on the strict regulation of system frequency in order to operate perfectly. For obtaining the standard operation in a power system, the frequency must be regulated within corroborated limits by adjusting the electrical supply to meet the demand. If supply and demand would not be balanced, the system frequency will change initially at a rate determined by the total inertia of the system. A generator or load has contribution of inertia of the system if a system frequency changing causes a change in its rotational speed and, thus, its kinetic energy [1]. The inertial response is the power associated with this change in kinetic energy is taken from or fed to the power system. The abrupt loss of supply is the typical initiator of a frequency event. The main factor of determination of the initial falling rate of frequency is the combined inertial response of all remaining electrical machines in the power system. High sensitivity of the frequency that related to the supply-demand imbalance is undesirable obviously. Therefore, it is critical that a large proportion of generation and load contribute to system inertia by providing an inertial response. Note that, the providing of inertial response has more importance for isolated networks or networks with weak interconnection [2, 3].

Analysis of effects of the system frequency changing on speed and kinetic energy of an electrical machine connected to the power system is needed for quantify its inertial response. In conventional synchronous generators, the rotor speed is sensitive to changes in system frequency, and therefore, an inertial response is naturally observable. Thus, if conventional synchronous generators have most contribution of total generation in a power system, the frequency variation will be lower.

In recent years, use of DFIG-based wind turbines has been increased considerably. Thus, a large number of conventional synchronous generators have been displaced by DFIG-based wind turbines.

Mentioned generators have a rotational speed that is decoupled from grid frequency. Reduction in system inertia is expectable, consequently. This is undesirable when there are a large number of DFIG-based wind turbines operating, especially in periods of low load and on smaller power systems. The frequency of a power system with low inertia will change rapidly for abrupt changes in generation or load. In this case, additional frequency response ancillary services must be provided to ensure that frequency limits are not exceeded [4, 5].

Based on above description, many authors have studied issues related to the inertia response of DFIG from different viewpoints. For instance, computer modeling of DFIG for better investigation of its inertia response and analysis of effects of the DFIG controller performance on its inertia response [4], employing a supplementary control loop to the DFIG controller to reintroduce inertia response [6], as well as investigation of effects of the pitch controller on the inertia response [7] are some contributions in this area.

In this paper, a DFIG connected to the infinite bus with simple and modified controller has been simulated in MATLAB®/Simulink® employing a developed model that proposed by Mullaneb et al. [4]. The inertia response of the DFIG with two types of controller (including simple and modified controller) has been analyzed and compared under the dropped frequency conditions. In addition, the effects of controller performance on their inertia response have been investigated. The Results prove that the inertia response can be improved significantly by modifying the generator controller system.

1. DFIG INERTIA RESPONSE WITH CONVENTIONAL CONTROLLER

A usual DFIG has wounded rotor that its winding ends are connected to the controller through the tumble rings. This controller applies a voltage to rotor with the specific domain and frequency, so that the generator slip is controlled.

With this type of control, when the network frequency drops during the constant wind speed and as a resultant, the speed of rotating stator field is going to decrease, the controller keeps the generator slip and electromagnetic torque stable by changing in the electrical speed of rotating rotor field. Thus, according to equation 3, rotor speed won't change in this conditions and its kinetic energy won't be released. Therefore, while frequency changes; this generator doesn't sense any difference from network so it won't have inertia response. However, due to the fact that the generator's control system has delay like all other control systems, it has negligible inertia response which depends on delay time; because during this delay, the generator will work as a normal electromagnetic generator [4, 5].

In normal electromagnetic generator, by decreasing in system frequency, speed of rotational field of stator will decrease. By decreasing in speed of rotational field of stator, slip that can be result by equation 1 will decrease. Cause of being our investigation in steady state situation, and in this state, slip is too low, relation of electromagnetic torque of this generator will be as equation that is given in the expression 2. It is clear by this equation that by increasing and decreasing in slip, electromagnetic torque will increase and decrease respectively.

In steady state and normal situation, mechanical torque of input and electromagnetic torque of generator are equal and by attention to equation 3, ($d\omega_r/dt$) will be zero. Therefore generator works without changing in rotor speed. But when electromagnetic torque increase which any reason, ($d\omega_r/dt$) will be negative and consequently, rotor speed will decrease. By this decreasing in rotor speed, kinetic energy releases that cause to momentary increase in output power of generator.

$$S = \frac{n_r - n_s}{n_s} \quad (1)$$

$$Te = \frac{3SV_{th}^2}{R_r \omega_S} \quad (2)$$

$$\frac{P_{mech}}{\omega_r} - T_e = j\frac{d\omega_r}{dt} \quad (3)$$

$$P_{out} = T_e . \omega_r \quad (4)$$

Where S, n_r, n_s, T_e, V_{th}, R_r, ω_s, ω_r, P_{mech} and P_{out} are slip, rotational field speed of rotor, rotational field speed of stator, electromagnetic torque, the Thevenin equal voltage, rotor resistance, stator electrical angular velocity, rotor electrical angular velocity, mechanical input power and output power of generator respectively.

A complete illustration about the case study and simulations can be found in works of Tayebi-Derazkolaie et al. [5, 8]. In simulation done for the generator, whose considered parameters are shown in table 1, the fault is considered in figure 1 as network frequency drop from 50 Hz to 49.75 Hz exponentially on 70th seconds. It should be noted that before the fault occurred, network frequency was 50Hz and generator delivers power of 2 MW to the infinite bus by having the rotor speed of 100 Rad/S.

By considering 1 second delay time of control system, as it is shown in figure 2, the generator electromagnetic torque increases from 20 KN.m to 20.00025 KN.m at fault time and consequently, the rotor speed decreases from 100 Rad/s to 99.988 Rad/s which is visible in figure 3.

During this slight speed decrease, rotor's kinetic energy decreases from 6950 KJ to 6948.332 KJ and consequently only 1.6KJ energy is released. By releasing this amount of energy, output power increases from 2 MW to 2.0005 MW. As it is seen in figure 4, the change is so low.

Therefore, it is observed that DFIG with conventional controller has negligible inertia response. Hence, because amount of inertia response is negligible, it is suggested to correct its controller by adding a new feedback [6].

TABLE 1. DFIG parameters

Parameter	Value	Unit
P_{out} (rated power)	2×10^{6}	W
R_S (stator resistance)	1.748×10^{-3}	Ω
R_r (rotor resistance)	3.253×10^{-3}	Ω
L_s (stator inductance)	2.589×10^{-3}	H
L_r (rotor inductance)	2.604×10^{-3}	H
L_m (mutual inductance)	2.492×10^{-3}	H
V_s (generator output voltage)	690	V
j (moment of inertia)	1.39×10^{3}	Kg/m
T_{in} (input mechanical torque)	2×10^{4}	N.m
P (number of pole)	6	----
f_s (frequency)	50	Hz

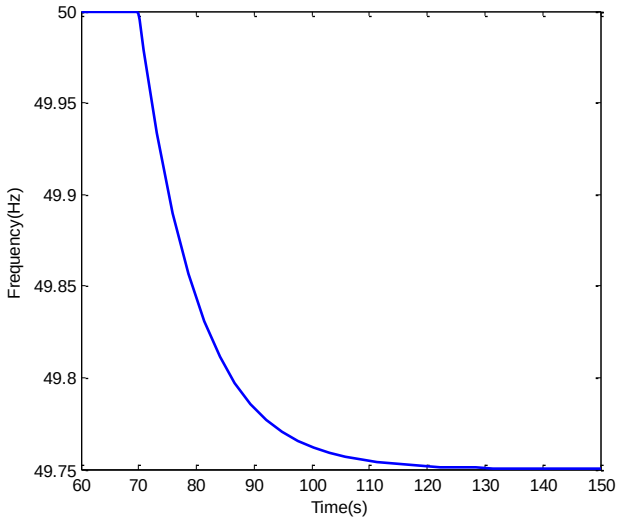


Figure 1: Network Frequency Drop

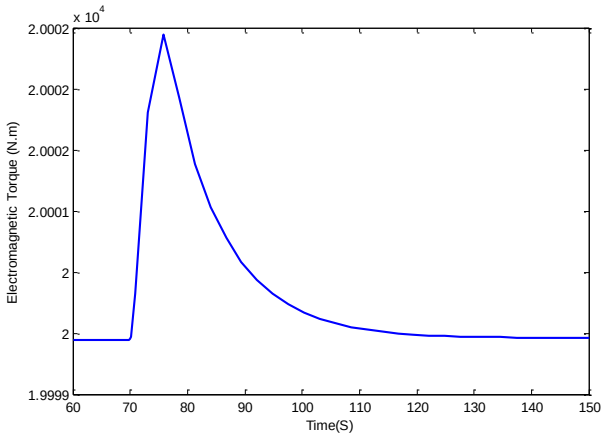


Figure 2: Electromagnetic Torque increasing of DFIG with conventional controller during Fault

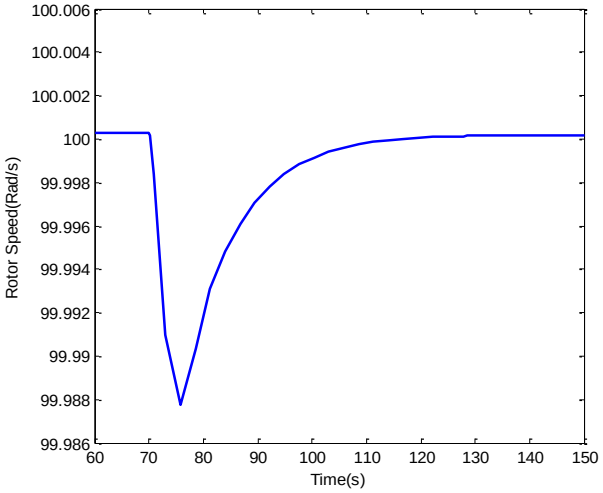


Figure 3: Rotor speed decreasing of DFIG with conventional controller during Fault

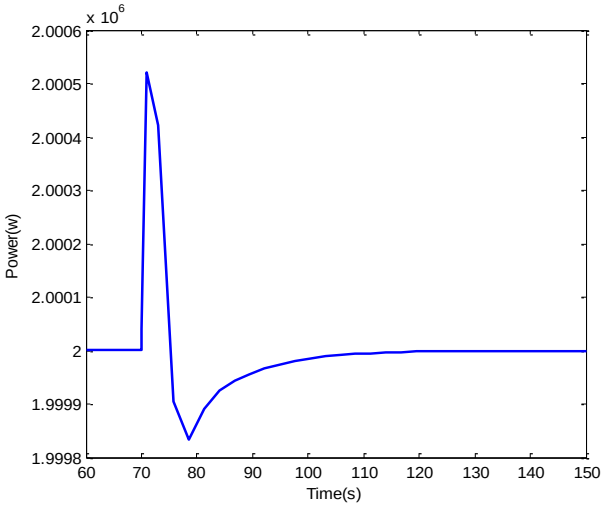


Figure 4: Variation of output power of DFIG with conventional controller during Fault

2. DFIG INERTIA RESPONSE WITH MODIFIED CONTROLLER

As it was observed in previous section the inertia response is near zero in DFIG with conventional controller because the slip is controlled. Conventional controller has a feedback according equation 5. To show inertia response during network frequency changes, according to equation 6 a special feedback is applied to the generator's reference electromagnetic torque in order to change rotor's speed by changing generator electromagnetic torque during network's frequency drop, and consequently generator's kinetic energy is released. So, in this generator, the reference torque in modified state is such as equation 7 [6-10].

$$T_e^* = K_1 \omega_r^2 \quad (5)$$

$$T_e = K_2 \frac{df_S}{dt} \quad (6)$$

$$T_e^* = K_1 \omega_r^2 + K_2 \frac{df_S}{dt} \quad (7)$$

In previous equations, T_e, K_1, K_2, ω_r, P and f_s were respectively reference electromagnetic torque, the main feedback coefficient in control system, increased feedback coefficient, rotor angular speed, mechanical power input and power network frequency.

By adding these feedbacks to reference feedback, when network frequency drops, amount of generator's reference torque increases based on amount of speed and how the network frequency decreases and amount of these feedbacks coefficient. By this increase, according to equation 3, rotor's speed decreases and rotor also releases kinetic energy causing temporarily increase in output power. When system frequency is fixed, derivative amount of frequency is zero, thus the added feedback is removed from circuit and generator is still in its normal operation state. Amount of the feedback's coefficient has direct relationship with amount of rotor's speed drop and inertia response. Therefore according to the limitation of amount of rotor's speed drop and amount of rotor's current increase, amount of this coefficient is limited [7].

In this paper by considering rotor's speed limitation of 85.5 Rad/s (for speed drop), the highest value of this feedback coefficient is considered 500 KN.m using trial and fault. As shown in figures 5 and 6, applying this amount of coefficient and frequency drop, electromagnetic torque increases to 28 KN.m and causes rotor's speed decreases from 100 Rad/s to 85.5 Rad/s. This amount of rotor's speed reduction according to figure 7, causes power increases to 2.66 MW which is better in comparison with the generator inertia response in conventional controller model.

Thus, in DFIG, inertia response can be reached from zero to expressed amount, by adding the feedback to control system and changing the feedback coefficient (depends on conditions).

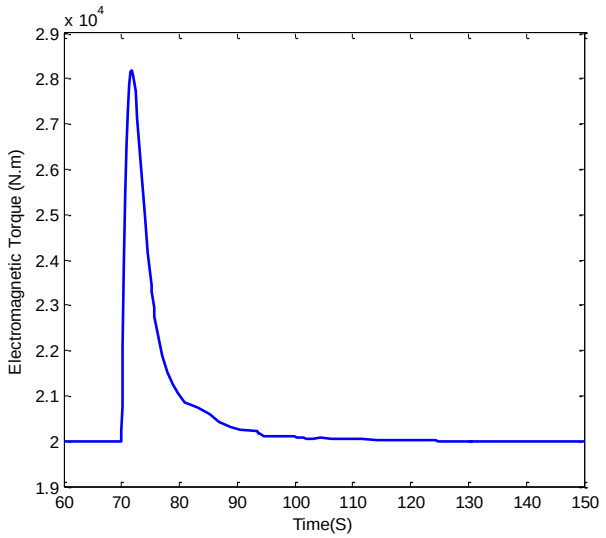


Figure 5: Electromagnetic torque increasing of DFIG with Modified controller during Fault

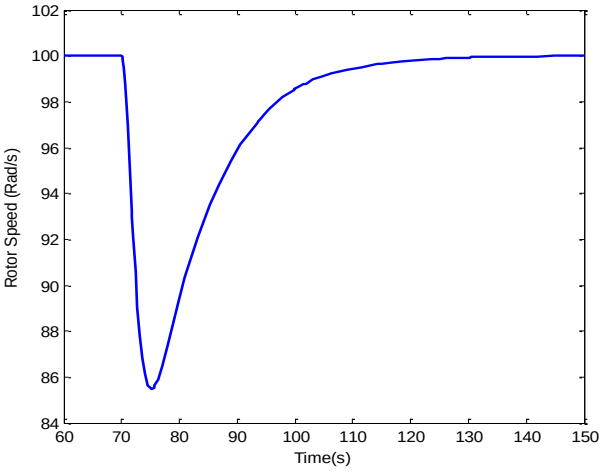


Figure 6: Rotor speed decreasing of DFIG with modified controller during fault

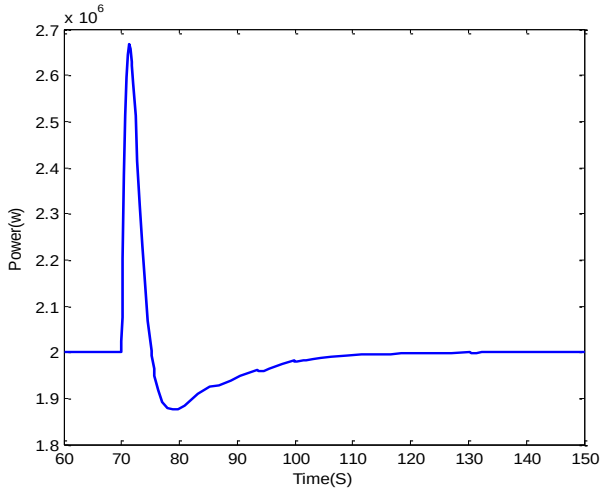


Figure 7: Variation of output power of DFIG with modified controller during fault

3. THE EFFECT OF DFIG WITH CONVENTIONAL CONTROLLER'S PERFORMANCE SPEED ON THE GENERATOR'S INERTIA RESPONSE

It was mentioned in section 2 that amount of delay time in generator control system affect amount of inertia response. In this part, effect of performance speed of DFIG with conventional controller on the generator's inertia response is analyzed. In conventional models, DFIG acts such as a usual electromagnetic generator till the control system has delay. In normal electromagnetic generator, with network frequency drop, speed of stator rotating field decreases. By decreasing in speed of stator rotating field, slip and electromagnetic torque (according to work in steady state) will decrease. At steady state, the input mechanical torque is equal to produced electromagnetic torque and according to equation 3, in this state amount of $d\omega_2/dt$ is zero. Therefore, generator works with a constant rotor speed. When electromagnetic torque increases for some reasons, this quantity will be negative and rotor's speed is reduced and because of this reduction, kinetic energy is released and appears as power in output. Therefore, when DFIG's controller system acts (performs) faster, it prevents the slip change faster and thus inertia response will be reduced [4, 5].

In performed simulation for this generator, to show the effect of controller system performance's speed on the generator's inertia response, the generator's inertia response is surveyed considering different time for controller system's speed. The result will be described and discussed in the following. The generator's electromagnetic torque has been shown in figure 8 with 0.1, 1, 2 second time delay in generator's control system. It is observed in this figure that when 2 seconds time delay were applied, electromagnetic torque increased 2 times more than when time delay was 1 second. But when the time delay was applied 0.1 second, electromagnetic torque almost didn't change. By delay time reduction in control system, electromagnetic torque increase becomes more and according to the mentioned content and equation 3, rotor speed drop will be lower and at this time, rotor releases less kinetic energy. Thus generator's output power increases less which is visible in figure 9. Therefore, it is observed that in DFIG without additional feedback, inertia response decreases by reducing the delay time in controller system performance.

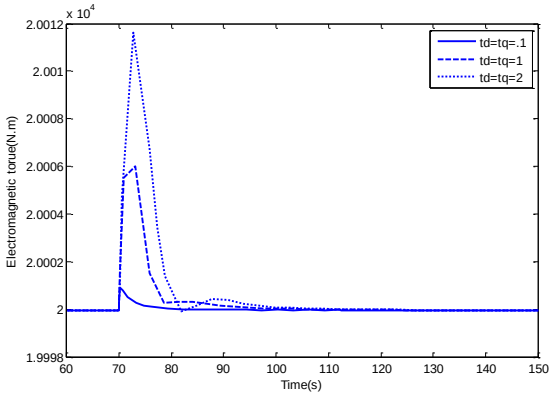


Figure 8: Variation of electromagnetic torque of DFIG with different time constants of controller under network frequency drop

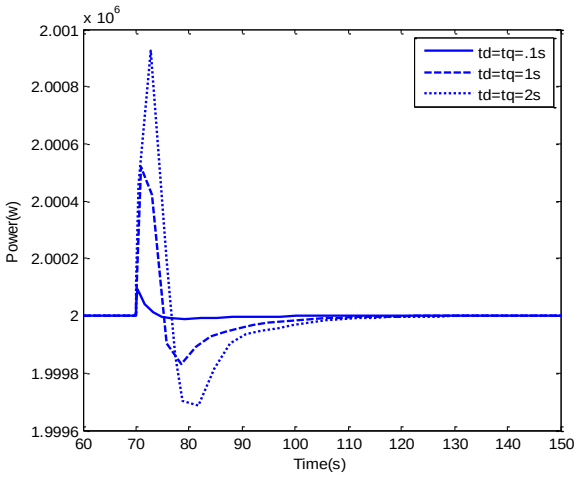


Figure 9: Variation of output power of DFIG with different time constants of controller under network frequency drop

4. THE EFFECT OF DFIG MODIFIED CONTROLLER'S PERFORMANCE SPEED ON THE GENERATOR'S INERTIA RESPONSE

In this section, the effect of modified controller performance speed on inertia response quantity is surveyed. When control system performs faster, effect of additional feedback on electromagnetic torque change appears faster. Because usually frequency drop in network is more in the first moment, the most effect on electromagnetic torque change is in first moment. This is clearly visible in figure 10 and as it is obvious in the figure, when delay time is applied for 1 second, electromagnetic torque caused by additional feedback performance increase to 2.8 KN.m (from 2 KN.m). But when the time is increased to 2 second, electromagnetic torque rate decreases to 2.7 KN.m and in case the delay time is applied for 0.1 second, electromagnetic torque increases to 3.2 KN.m.

Therefore, as it was observed, by reducing the control system's delay time, generator electromagnetic torque increases more with additional feedback existence. The more electromagnetic torque increases, the more rotor speed decreases according to equation 3. Consequently, by increasing controller speed, rotor releases more kinetic energy and generator's output power increases temporarily which is visible in figure 11. As it is obvious in this figure when the delay time was applied 0.1 second, generator's output power increased from 2 MW to 3 MW. It increased to 2.66 MW when the delay time was applied 1 second and when the delay time was applied 2 seconds the power increased to 2.55 MW. Thus in DFIG with additional feedback to improve inertia response, the inertia response increases by increasing control system's performance speed.

Another noticeable point in this part was control system performance speed in conventional model of DFIG wasn't important due to low inertia response of generator. But by applying additional feedback causes the generator's inertia response increase, the control system performance speed was very important. Because a small change in time changes the generator inertia response to high level. Thus this delay time reduction is very important.

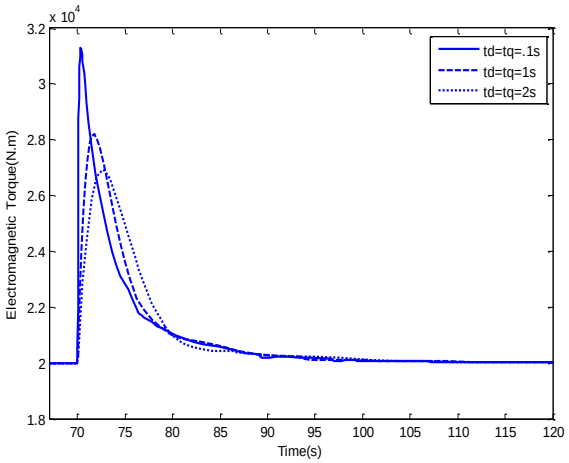


Figure 10: Variation of electromagnetic torque of DFIG with modified controller in different time constants of controller under network frequency drop

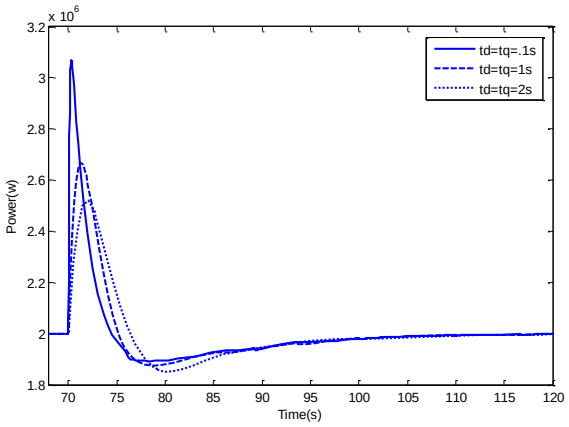


Figure 11: Variation of output power of DFIG with modified controller in different time constants of controller under network frequency drop

CONCLUSION

Conventional synchronous generators have a good inertia response due to coupling with power system frequency, therefore this type of the network which has many number of such generators is highly resistant to the change in frequency. In recent years, the numbers of wind power plants used in the network increases, and in some cases, were used instead of thermal power plants. Today DFIG becomes the dominant generator in wind power plants due to its high advantages. So analyzing the generator inertia is a matter of importance. It should be mentioned that this generator due to its own special control system, is not affected from network frequency and does not show the proper inertia response so specific feedback can be used to correct this situation. In this paper, by simulating a DFIG in MATLAB environment, inertia response with conventional and modified control models has been analyzed in presence of network frequency drop. So, the following results were obtained:

By adding the feedback, response of inertia increases. The noticeable point in this increase is that the response is controllable by controlling the feedback coefficient so by optimal controlling of this factor, to the

extent that the generator is not taken out of its operation, the rotor speed can be changed. Therefore, adding this feedback, the inertia response of the generator can be changed in a wide range.

In DFIG with conventional controller, when the network frequency is changing, while the control system doesn't act, this generator works such as a common induction generator and will have inertia response (of course it is negligible). Therefore, the more speed of control system, the faster compensation of frequency change and the less inertia response, but in modified model, because of the fact that inertia response optimization is caused by control system performance, the more speed of control system leads to increase in inertia response when the network frequency changes. Thus, from the viewpoint of inertia response, in conventional model of DFIG, increase in controller speed is not so important unlike the modified one which the increase in controller speed has a high importance.

REFERENCES

[1] R. Tayebi-Derazkolaie, H.A. Shayanfar, B. Mozafari, "Effects of Rotor Resistance Value of SCIG on its Output Power and Efficiency", International Journal of Pure and Applied Science and Technology, Vol. 4, No. 1, pp. 41- 48, 2011.

[2] J. Logan, S.M. Kaplan, "Wind Power in the United States: Technology, Economic, And Policy Issue", Published by Congressional Research Service, Missouri, USA, June 2008.

[3] H. Polinder, Sjoerd W.H. de Haan, Johannes G. (Han) Slootweg, "Basic Operation Principles and Electrical Conversion Systems of Wind Turbines", EPE Journal, Vol. 15, No. 4, pp. 43-50, 2005.

[4] A. Mullaneb, M. Malley, "The Inertial Response of Induction-Machine-Based Wind Turbines", IEEE Transactions on Power Systems, Vol 20, No. 3, pp.1496-1503, 2005.

[5] R. Tayebi-Derazkolaie, H.A. Shayanfar, B. Mozafari, "Effects of the Controller Performance of DFIG on its Inertia Response", Global Journal of Research in Engineering, Vol. 11,No. 3, pp. 21-24, 2011.

[6] J. B. Ekanayake, N. Jenkins, "Comparison of The Response of Doubly Fed And Fixed-Speed Induction Generator Wind Turbines To Changes In Network Frequency", IEEE Transactions on Energy Conversion, Vol. 19, No. 4, pp. 800 – 802, 2004.

[7] B. Rawn, Edward S. Rogers, "Wind rotor inertia and variable efficiency:fundamental limits on their exploitation for inertial response and power system damping", European Wind Energy Conference, Brussels, Belgium, April 2008.

[8] R. Tayebi-Derazkolaie, H.A. Shayanfar, B. Mozafari, "Investigation of the Inertia Response of the SCIG", Research Journal of Applied Sciences, Engineering and Technology, Vol. 4, No. 17, pp. 3078- 3082, 2012.

[9] J. Morren, J. Pierik, S. D. Haan, "Inertial Response of Variable Speed Wind Turbines", Electric Power Systems Research, Vol. 76, No. 11, pp. 980–987, 2006.

[10] A. Ebadi, M. Mirzaie and S. A. Gholamian, "Performance Evaluation of Three-Phase Induction Motor Fed by unbalanced voltage Combined with Over- or Under Voltage Using Finite Element Method", Journal of World's Electrical Engineering and Technology, Vol. 3, No. 1, pp. 18–25, 2014.

Optimization of renewable energy for buildings with energy storages and 15-minute power balance

Swarna Manjari Samal, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, swarnamanjari88@gmail.com*

Chinmaya Ranjan Pradhan, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, cr_pradhan@outlook.com*

Sandip Kar Mazumdar, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, sk_mazumdar1@gmail.com*

Achyutananda Panda, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, achutanada.panda23@gmail.com*

A B S T R A C T

When planning renewable hybrid energy solutions in buildings, it is important to consider both investment and operating costs. This study develops a novel building optimization model based on the coming 15 min power balance settlement. It utilizes multiple energy storages, including hot water tank and flow and lead-acid batteries. We apply the model to plan the retrofitting of an office building in Helsinki and a residential building in Tallinn, with photovoltaics and a ground source heat pump. The model is a large dynamic linear or mixed-integer linear programming model (LP/MILP) for an entire year. The results determine both the optimal dimensioning and the optimal operation of the different production and storage technologies for each building. The optimized configurations caused significant savings in energy costs for both buildings while reducing non-renewable primary energy consumption. Heat storage is highly cost-efficient, but power storages are not. Photovoltaics is cost-efficient in the Helsinki building but slightly unprofitable in the Tallinn building. Power and heat storages do not interact strongly, even in the presence of the ground source heat pump. The heat storage operates in concert with district heating and the ground source heat pump, while power storages operate together with photovoltaics and power trade.

Keywords:
Renewable energy optimization in buildings
15min power balance
Power storage
Heat storage
Ground source heat pump
Photovoltaics

1. Introduction

1.1. Background

Greenhouse gas emissions of buildings can be reduced by adopting renewable-based hybrid energy systems, including, for example, district heating (DH), district cooling, photovoltaics (PV), heat pumps (HPs), power storages (PS), cool storages, and heat storages (HS). Because different energy forms and technologies interact in a complex manner, configuring, dimensioning, and operating such systems optimally requires advanced modeling techniques [1].

In the power market, electricity production and consumption should always be balanced, and balance responsible parties must plan their operation to maintain this balance. However, there will always be imbalances. Balance (or imbalance) settlement means a financial settlement mechanism aiming at charging or paying balance responsible parties (BRPs) for their imbalances. Currently calculations carried out within imbalance settlement in most European countries are based on hourly energies which are obtained from hourly energy measurements, load profiles, production plans, and fixed deliveries. To make the power market more responsive to increasing amounts of intermittent renewable power production, the EU Commission Regulation 2017/2195 declares that all transmission system operators must apply a 15 min balance settlement period in all scheduling areas and power market trade periods must coincide with this period. This transition will come to effect on January 1, 2025, at the latest. This means that also building energy optimization models and methods must be adapted to handle power trade at 15 min intervals.

This study focuses on optimizing the configuration, dimensioning, and operation of a building hybrid energy system subject to 15 min power balance, emphasizing different types of power and heat storages.

Notation

Abbreviations

COP	Heat pump coefficient of performance
DH	District heating
HPH	Heat pump for heating
HS	Heat storage
LP	Linear Programming
MILP	Mixed Integer Linear Programming
nZEB	Nearly Zero-Energy Building
PP	Power purchase from grid
PP-out	Power sales to grid
PS	Power storage
PS-flow	Power storage, flow battery
PS-lead	Power storage, lead-acid battery
PV	Photovoltaics

Indices and index sets

t	Hourly period (1, …T) or 15 min period (1, …4T)
u	Energy supply or storage unit
H	Super- or subscript referring to heat
P	Super- or subscript referring to electric power
CONTR	Superscript for energy contracts
PROD	Superscript for production units
U	Set of energy supply units
S	Set of energy storages

Symbols

A_u	m^2 PV panel area
B_u	1 Maximum number of battery storage cycles in planning horizon
$c_{u,t}$	€/MWh Operating cost of unit u in period t
c_u^{CONST}	€ Constant cost term for energy supply or storage unit u
c_u^{MAX}	€/MW Fixed capacity cost of energy supply u; for storages €/MWh
$d_{H,t}$	MWh Demand of heat in period t
$d_{P,t}$	MWh Demand of power in period t
r_t	MW/m^2 Solar radiation intensity in period t
$s_{u,t}$	MWh Storage u level at end of period t
s_u^{MAX}	MWh Storage u maximum capacity
s_u^{DEEP}	1 Storage u deep discharge level as fraction of capacity
$s_{u,t}^{IN}$	MWh Storage u charge rate in period t
$s_u^{IN,MAX}$	MWh Storage u maximum charge rate per period
$s_{u,t}^{OUT}$	MWh Storage u discharge rate in period t
$s_u^{OUT,MAX}$	MWh Storage u maximum discharge rate per period
T	h Number of hours in planning horizon
$x_{u,t}$	MWh Unit u operating level in period t
x_u^{MAX}	MW Unit u maximum capacity
$x_{u,t}^H$	MWh Unit u heat production in period t
η^{PV}	1 PV efficiency, power production per radiation on panel
η_t^H	1 Heat pump heating efficiency, COP-factor
η_u^S	1 Storage u efficiency per period
η_u^{IN}	1 Storage u charging efficiency
η_u^{OUT}	1 Storage u discharging efficiency
z_u	1 Binary ON/OFF variable or parameter that determines if the energy system includes or excludes unit u

1.2. Related research

The related research field is broad due to the variety of hybrid energy systems, starting from PS technologies, ending with optimization model methodologies, countries/zones, and energy balancing multi-timescale settlements. Some studies focus on a single PS technology; others propose multiple types of batteries in one energy system. Table 1 presents studies and reviews on hybrid energy systems with various PS technologies in different buildings/facilities that have similar technologies in hybrid energy systems as in our study.

Table 1
Relative research on hybrid systems with power storages in different buildings and locations.

Type of storage in (hybrid) energy system	Type of building(s)/facility	Country/Zone	Year of publication	Reference
Fuel cell, hydrogen tank	University	Sharjah, UAE	2019	[9]
Fuel cell, hydrogen	Residential, nZEB	Test building in GAMS software, location unknown	2019	[10]
Fuel cell, hydrogen storage	University	Ekpoma, Nigeria	2020	[11]
Lead-acid, water storage tank	Residential, nZEB	3 locations in Italy	2017	[12]
Lead-acid, Li-ion, Sodium- sulfur	Hybrid power system with several PS	China	2020	[2]
Lead-acid (VRLA, two types)	Research center (laboratory)	Riyadh, Kingdom of Saudi Arabia	2020	[13]
Li-ion	12 detached houses, 1 apartment building	Tampere, Finland	2019	[3]
Li-ion	Office	Västerås, Sweden	2020	[14]
Li-ion (Tesla Powerwall)	Households	Munich, Germany	2016	[5]
Li-ion	Harbour	Åland Islands, Swedish-speaking region of Finland	2020	[15]
Li-ion (Tesla), hot water storage tank	Residential	12 locations in Sweden	2020	[16]
Battery (by parameters assume this is Li-ion)	Integrated energy system	China	2021	[8]
Battery (by the parameters assume this is Li-ion), Supercapasitor	Islanded DC microgrid	China	2021	[7]
Battery (assume this Li-ion or Lead-acid)	Single family household	Tehran, Iran	2018	[4]
Li-ion, Flow battery (VRFB)	Residential	Switzerland	2021	[6]
Flow battery (VRFB)	Energy station	Martigny, Switzerland	2018	[17]

1.3. Research gap and novelty

There is very little research on 15 min power balance settlement [18,19,20], and even less on applications for building energy optimization. The main target of this study is to develop a model for optimizing renewable energy solutions for a building subject to the future 15 min power balance settlement. In our previous study [1], we developed an hourly-based linear or mixed-integer linear programming (LP/MILP) model to determine simultaneously the optimal configuration, dimensioning, and operation of a hybrid energy system for a building.

This study extends the previous model to handle 15 min power balance and makes the storage model more detailed. In earlier research [1], PS in northern latitude buildings turned out to be grossly unprofitable. In this study, we evaluate if the introduction of the 15 min power balance improves the cost-efficiency of PS. We also study the optimal operation of multiple storages together. We apply the model to two buildings in different countries: an office building in Helsinki, Finland, and a residential building in Tallinn, Estonia. The included technologies are DH, PV, ground source heat pump for heating (HPH), HS, and different types of PS.

This paper is organized as follows. Section 2 describes the building energy system and all necessary parameters for the case study of the two buildings, one in Helsinki and another in Tallinn. In Section 3, we define the building optimization model. In Section 4, we present the results of applying the model to two buildings. In section 5, we conclude the study and give some directions for future research.

2. Case study

2.1. The building hybrid energy system

Fig. 1 shows the building energy system. In such a hybrid system, everything may depend on everything. The power balance combines different power-consuming units with power purchase, local PV production, and PS units. Similarly, the heat balance combines heat demand with heat sources and HS. HS combines hourly periods, and PS combines 15 min periods together. The HPH connects the 15 min power balance with the hourly heat balance. Interaction between energy forms and dynamics caused by storages imply that optimal operation and dimensioning can be found only by using an optimization model.

In this study, we develop a model that is based on hourly demand for heat and 15 min power demand and solar radiation in the target locations. The objective function minimizes the combined operating and fixed costs. The cost minimum defines which technologies should be included and how they should be dimensioned when optimally operating the energy system.

2.2. Hourly heat and power demand, and power price

Hourly demand for both buildings contains full-year data for 2019 [21,22]. Because hourly data for a full year does not display well, Fig. 2 shows the data for 30 weeks (Monday - Sunday) from February 4 to September 1. The first diagram in Fig. 2 presents the heat demand. The Helsinki building consumes more heat than the Tallinn building, in absolute terms and per building area. Also, the seasonal variation is somewhat larger in Helsinki than in Tallinn. This difference is mainly attributed to different building types, office in Helsinki vs. residential in Tallinn. The difference in heat load also comes from building's energy performance. The Tallinn building already satisfies Nearly Zero-Energy Building (nZEB) requirements [23]. The second diagram in Fig. 2 shows the power load for each building. The power loads of the two buildings differ significantly. The office building in Helsinki has a persistent baseload due to a data center, while the residential building in Tallinn demonstrates a more 'normal' power load with a typical daily, weekly and yearly variation. The third diagram in Fig. 2 shows the ElSpot power price in Finland. Power price in Estonia is omitted, because it is identical to Finnish power price for most of the time, except for 97 h in 2019 when transmission lines were congested, resulting in minor differences.

2.3. Building data

This research is applied to two different types of buildings in neighboring countries — an office building in Helsinki, the capital of Finland, and a residential building in Tallinn, the capital of Estonia. Currently, neither building has HPH, HS, or PS. The Helsinki building has no PV, while the Tallinn building has small-scale PV. Table 2 displays the properties of the target buildings [21,22].

2.4. Costs and technical parameters

In purpose to decrease the consumption of external electricity, 330 m^2 of PV panels can be installed on the roof of each building. The maximal radiation of 1 kW/m^2 and 15% PV panel efficiency [24] gives 49.5 kW peak production in Helsinki. The same data was applied to the building in Tallinn. In Estonia, the owner of a PV

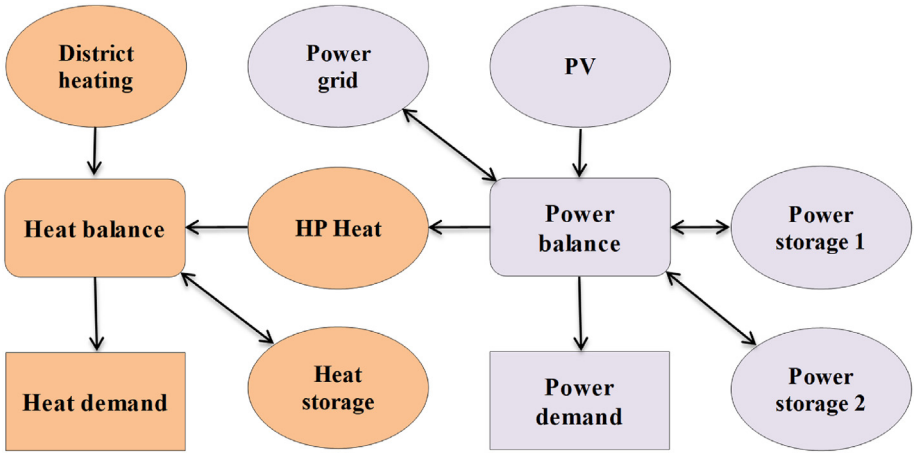


Fig. 1. Building energy system.

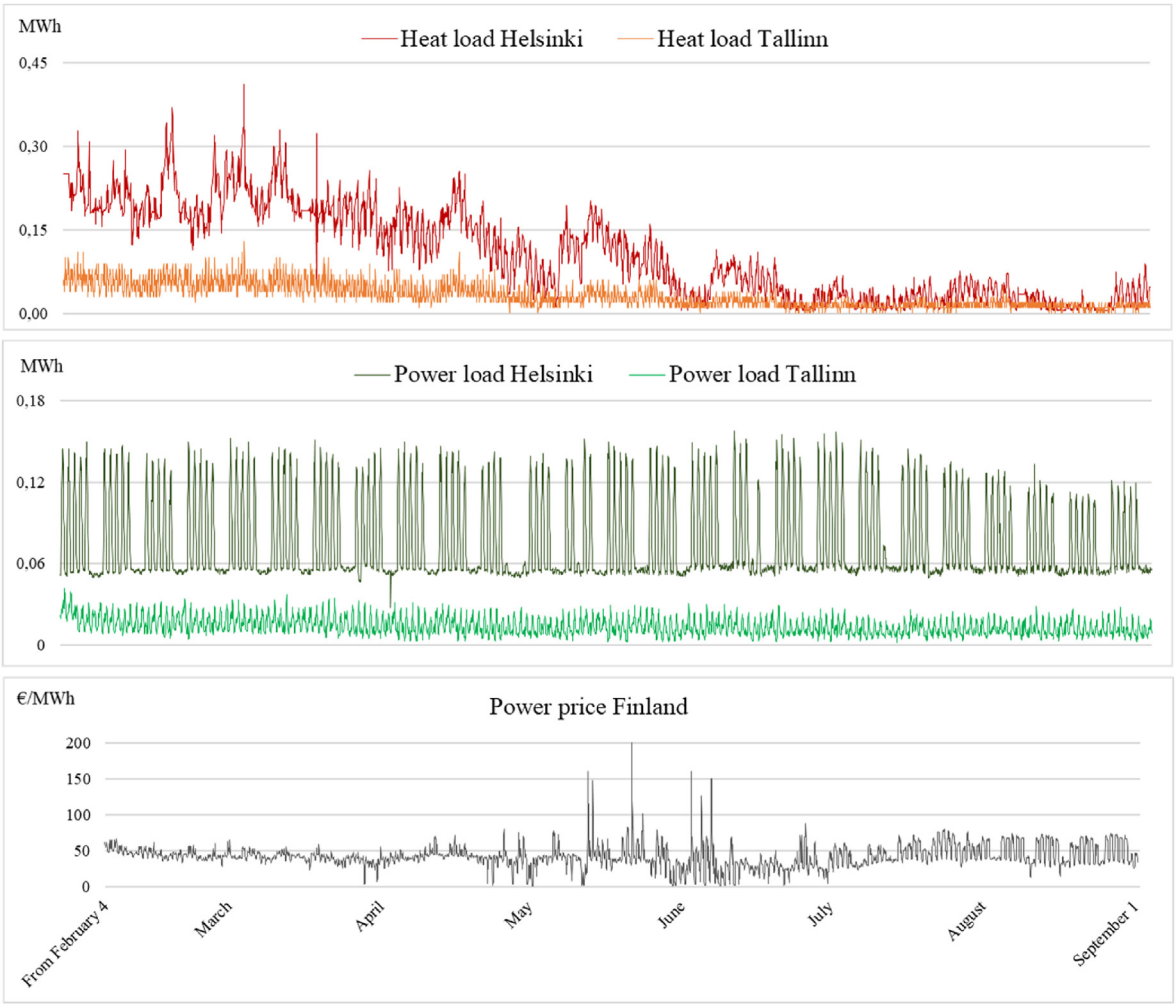


Fig. 2. Hourly heat and power load for two buildings and power price in Finland from February 4 to September 1, 2019.

Table 2
The parameters for both buildings.

Description	Building 1	Building 2
Building type	Office	Residential
Location	Helsinki, Finland	Tallinn, Estonia
Gross area	8341 m^2	5298 m^2
Net area	5878 m^2	4324 m^2
Annual heat demand	1083 MWh	331 MWh
Annual power demand	655 MWh	141 MWh

system under 50 kW can receive support for renewable energy production [25].

DH is an important and cost-efficient heating form in urban areas of Finland and Estonia. The market share of DH is over 50% of floor space in Finland and over 90% in Helsinki [26]. In Estonia, the market share of DH reached about 60% in 2018. According to Estonian District Heating Act and the Competition Act, the DH market in Estonia is regulated, and the Estonian Competition Authority approves the maximum prices that can be charged in various regions [27]. In addition to DH, both buildings can be equipped with ground source heat to supplement DH.

Table 3 presents prices for different energy sources [28–31] and storages [26,32] for both countries. All investment costs are computed as annuity using a 4% interest rate and technology-specific lifetime. PV investment cost is based on typical PV system price per peak power in Finland (1766 €/kWp) applying a 20-year lifetime [33]. The same system price is used for Estonia. For HPH, we applied a 15-year lifetime, 20 years for HS, 5-years for lead-acid battery, and 20 years for Flow battery [34–36]. Maintenance and replacement costs are excluded. Table 4 lists technical parameters for local heat production and heat storages.

2.5. Power storage data

Since not all generated PV can be consumed immediately, storing the excess energy during off-peak times may significantly improve the system efficiency and satisfy variable power demand at different timescales during the day [38]. Thus, PS can increase the self-consumption of PV power. Additionally, PS can lead to other benefits such as demand response or shaving high load peaks. The efficiency of the battery energy storage system (BESS) is mainly influenced by the battery efficiency, power conversion, and standby consumption of the different system components [39]. The economic value of BESS depends heavily on the load and generation profile: a considerable increase of self-consumption using storage results in higher savings [40]. Battery energy is primarily used for demand balancing and fast discharging purposes (less than an hour), but solutions for longer term battery operation are developed [41]. The number of charge/discharge cycles depends on the

Table 3
Prices for different energy sources and storages.

Unit	$c_{u,t}$ (€/MWh)	c_u^{MAX} (€/MW/a)	c_u^{CONST} (€/a)	Description
Helsinki				
DH	34.87/63.62	15052	2528	DH contract, energy price summer/winter
PP	ElSpot Finland	1591	125	PP contract (energy fee hourly changing)
Tallinn				
DH	60.17	0	0	DH contract, fixed maximum price for 2019
PP	ElSpot Estonia	764	308	PP contract (energy fee hourly changing)
Helsinki and Tallinn				
PV	–	1766	0	PV investment per kW peak power
HPH	–	116 923	0	Ground source heat pump investment
HS	–	396	0	Heat storage
PS-flow		23 914	0	Power storage, flow battery
PS-lead	–	33 694	0	Power storage, lead-acid battery

Table 4
Technical parameters for local heat production and heat storage in both locations.

Parameter	Value	Unit	Description	Reference/Comment
η_u	3.5	1	COP/efficiency ratio for heat pump	[37]
η_B^S	0.99	1	Storage efficiency, self-discharge per time step	[26]
η_B^{IN}	0.95	1	Efficiency for charging storage	[34]
η_B^{OUT}	0.95	1	Efficiency for discharging storage	[34]
$s_H^{IN,MAX}$	1.2	MW	Heat storage maximum charge rate	*calculation*
$s_H^{OUT,MAX}$	1.2	MW	Heat storage maximum discharge rate	*calculation*

capacity taken from the battery (a function of discharge rate and depth of discharge), operating temperature, and the charging method [42]. Selecting the appropriate short-term and longer-term PS technologies lay in a balance between the feasibility for buildings, reasonable costs, suitable technical performance parameters, and available now and near future.

Table 5 presents four different PS technologies. Before making the final decision on what kind of two technologies to include in the study, we considered a Vanadium Redox Flow Battery (VRFB) and a Proton Exchange Membrane (PEM) for (little) longer-term operation. VRFB has a very low daily self-discharge rate but a disadvantage is a relatively low round-trip efficiency (e.g., compared to the lead-acid battery) [35]. Our model uses a VRFB battery system with a capacity of 0.5 MWh and a daily self-discharge of 0.01% [43].

The fuel cell is similar to a battery where the electrochemical reaction occurs as long as fuel is available [44]. The proton exchange membrane, also known as PEM, uses a polymer electrolyte. PEM has a response time from seconds to minutes [44,45]. Without significant operational constraints, this battery operates over the quasi-entire power load range (10−100%) within seconds [45]. PEM has a lifespan of 10 years, but the high manufacturing costs and a complex water management system make this technology expensive compared to other considered PS technologies [46].

For short-term PS, we chose the lead-acid battery, which is known as the oldest storage device among all rechargeable batteries [47]. Lead-acid BESS has relatively low daily self-discharge rates ranging from 0.09% to 0.4%, which are somewhat higher than Li-ion batteries [35]. Our model uses a lead-acid battery with a capacity of 0.1 MWh and a daily self-discharge of 0.2% [38]. Another short-term PS is the Tesla Powerpack system. Each Powerpack contains 16 individual battery pods, each with an isolated DC converter [48]. The battery capacities start from 50 kWh. We tested

Table 5
Flow battery and lead-acid battery parameters (efficiency in 15 min settlement).

Flow	Lead-acid	Tesla Powerpack	Fuel cell	Unit	Description	Reference/Comment
0.99999	0.99998	0.99998	0.99999	1	Storage efficiency	*calculation*
0.83667	0.89443	0.94604	0.80623	1	Efficiency of charging storage	[16,40]
0.83667	0.89443	0.94604	0.80623	1	Efficiency of discharging storage	[36]
0.65	0.80	0.895	0.65	1	Round-trip efficiency	[35,48]
0.01	0.2	0.017	0.01	1	Self-discharge rate, daily loss	[38,43,48]
10 000	900	4000	3000	cycles	Full cycles (until 80% capacity)	[36,49]
20	5	10	10	years	Lifetime (until 80% capacity)	[35,36,49]
0	0.25	0.04	0	1	Depth of discharge (0 = empty)	[35,36]
0.5	0.1	0.1	0.5	MWh	Storage capacity	[36,49]
0.05	0.025	0.037	0.025	MW	Maximal charge rate	*calculation*
0.05	0.025	0.037	0.025	MW	Maximal discharge rate	*calculation*
325 000	150 000	385 000	410 000	€/MWh	Investment cost/capacity	[36,50]
23 914	33 694	47 467	50 549	€/MWh, a	Investment cost/capacity, annual	*calculation*
4	4	4	4	%	Interest rate, annual	[1,40]
0.07358	0.22463	0.1233	0.1233	1	Annuity factor	*calculation*

*Price in USD, currency rate 1 USD = 0.8299 EUR, end of day prices provided by Morningstar, January 21, 2021.

our model with a 100 kWh Powerpack with high efficiency and round-trip efficiency and a response time compared to the other three PS types. Due to the much higher price, the Powerpack did not appear superior to lead-acid batteries for our buildings.

2.6. Power price

Buying power from the grid means payments for the energy and distribution, subsidy and RES subsidy (in Estonia), electricity tax, retailer margin, and VAT. Finland does not have a feed-in tariff or feed-in premium for small-scale renewable power. This means that excess PV power must be sold to the grid at a low price that depends on the retailer. In Helsinki, this is the ElSpot power market price. Consequently, the sales price for electricity is much lower than the purchase price [1]. Therefore, it is generally best if local PV production can be used locally rather than sold to the grid.

Estonia is currently upgrading its transmission network to help integrate a higher share of PV generation. Under the current regulations, Estonia pays for renewable power production a feed-in premium [51] on top of the market price of electricity [52]. The premium has a sliding scale and is designed with a payment cap and floor. The full premium (53.7 €/MWh) is paid when the average market price of power in the previous month is below the floor value 39.3 €/MWh. The premium decreases linearly to reach zero at the cap value that equals the average market price 93 €/MWh. In 2019, the average monthly market price had mostly stayed between the floor and cap, which means that on average, the sales price for renewable power has been 93 €/MWh subtracted by retailer margin.

Table 6 presents the power parameters used in the model. Data for Helsinki is received from the distribution company Helen [30]. Data for Tallinn is obtained from distribution company Elektrilevi OÜ [31].

3. Building optimization model

The building optimization model has been developed as an extension of the model in Ref. [1]. The main extensions are the introduction of 15 min time interval for power-related components, construction of 15 min power demand and price data, introduction of multiple storages for energy forms, more detailed model for storages, and limits on selling power to the grid to gain feed-in premium for renewable power in Estonia. The current model omits cooling because neither of the target buildings has cooling. Cooling could be easily included in the model, but with multiple (power-based) cooling options, model size would grow significantly.

Table 6
Power price parameters for both locations (PV system 330 kW).

Value	Unit	Description
Helsinki (VAT 24%)		
44.04	€/MWh	Average ElSpot price 2019
40.67	€/MWh	Transmission fee
5.51	€/month	Fixed cost (220 V)
4.87	€/month	Basic fee
2.98	€/MWh	Retailer margin
22.53	€/MWh	Electricity tax (VAT 0%)
103.47	€/MWh	Energy price
124.55	€/year	Annual constant fee
Tallinn (VAT 20%)		
45.86	€/MWh	Average ElSpot price 2019
30.72	€/MWh	Transmission fee
21.44	€/month	Distribution integration fee
4.20	€/month	Current 100A (220V)
5.40	€/MWh	Subsidy
13.60	€/MWh	RES subsidy
3.06	€/MWh	Retailer margin
78.39	€/MWh	Electricity purchase price
39.21	€/MWh	Electricity sales price (< 50 kW)
307.73	€/year	Annual constant fee

3.1. Objective function

The model minimizes combined operating and fixed costs. Fixed costs depend linearly on the capacity of different energy supply and storage units. Purchase contracts for energy (DH, PP) have a monthly or yearly tariff for maximal capacity agreed on in the contract. For local energy production (PV, HPH), the fixed costs consist of the investment costs, which depend on the maximal production capacity (€/MW). Investment costs of storages depend on the storage capacity (€/MWh). In addition to the linear capacity-based components, the fixed costs may include a constant cost term (€). In the model, all fixed costs are represented as an annuity and scaled for the length of the planning horizon.

In the model, the set of energy supply units is U, and storages is S. The objective function minimizing the overall costs is:

$$\min \sum_{u\in U_H}\sum_{t=1}^{T} c_{u,t}x_{u,t} + \sum_{u\in U_P}\sum_{t=1}^{4T} c_{u,t}x_{u,t} + \sum_{u\in U}\left(c_u^{MAX}x_u^{MAX} + c_u^{CONST}z_u\right) + \sum_{u\in S}\left(c_u^{MAX}s_u^{MAX} + c_u^{CONST}z_u\right) \quad (1)$$

The first sum adds up operating costs of the hourly operating energy supply units for heat $u\in U_H$. The second sum adds up similarly operating costs for power-related units $u\in U_P$ that operate on 15 min level. Because the heat pump consumes power in 15 min periods, it is included in U_P. For heat units, the index t iterates over the hourly periods 1, …, T, and for power units over the 15 min periods 1, …, 4T. In a yearly model, $T = 8760$ h, and 4T = 35 040 quarters of an hour. For the supply units, the operating costs are the unit cost multiplied by the operating level of the unit $x_{u,t}$. For some supply units the unit costs may be zero. For sales contracts, the unit cost is negative. Storages have no operating costs in this formulation.

The fixed costs of the supply units and storages are computed by the third and fourth summations, correspondigly. The capacity variables x_u^{MAX} and s_u^{MAX} are multiplied by cost per capacity c_u^{MAX}. Constant cost terms c_u^{CONST} are multiplied by ON/OFF (0/1) variables z_u and added to the capacity-dependent costs. The ON/OFF variable equals 1 if corresponding unit is included and 0 if the unit is excluded from the energy system configuration.

Due to the ON/OFF variables, the model is a MILP model. If the ON/OFF variables are fixed to 0 or 1 or the constant cost terms are zero, then the model is reduced into an LP model.

3.2. Power demand and price at 15 min

Europe will transition from hourly power balance settlement into 15 min time intervals. Finland is planning to make this transition in 2023 [53]. Baltic countries, including Estonia, are also considering the same schedule, and will make this transition no later than 2025. It is still unknown how this change is reflected in the power demand and power price at 15 min level. We estimate that the market volatility will increase somewhat. In Ref. [54], the author tried to estimate the volatility based on the current Frequency Containment Reserve (FCR) market. The FCR market sets imbalance prices and is thus relevant to balance responsible parties [19]. In this work, we assume that the relative volatility (standard

deviation) of power demand and power price will be 1% point higher at the 15 min level than at the hourly level. To achieve this, we generate 15 min demand and price data from hourly data by distributing the hourly data on the corresponding four 15 min intervals and adding random noise to reach the desired volatility. In other words, each 15 min data point for power demand and spot price ($d_{P,t}$, $c_{P,t}$) is formed from the corresponding hourly data ($D_{P,t}$, $C_{P,t}$) using the following formulas:

$$d_{P,t} = (1 + a\rho)\cdot D_{P,\lfloor (t+3)/4 \rfloor} \Big/ 4, \quad t = 1, \ldots, 4T, \tag{2}$$

$$c_{P,t} = (1 + a\rho)\cdot C_{P,\lfloor (t+3)/4 \rfloor}, \quad t = 1, \ldots 4T. \tag{3}$$

Divisor 4 is needed in (2) to distribute hourly demand (MWh) to 15 min demand. In these formulas, ρ is a random number in range [−1,+1] and a is a scaling factor determined for each data series so that the 15 min data has 1%-point higher relative volatility. Fig. 3 illustrates the constructed 15 min power price data for two days in Finland.

3.3. Energy balances

In the following subscript H stands for heat and P for power. Energy balance constraints state that combined energy supply must match the demand ($d_{H,t}, d_{P,t}$) in each time interval. The hourly heat balance equations are formulated as:

$$\sum_{u \in U_H^{CONTR}} x_{u,t} + \sum_{u \in U_H^{PROD}} x_{u,t}^H + \sum_{u \in S_H} \left(s_{u,t}^{OUT} - s_{u,t}^{IN} \right) = d_{H,t}, \quad t = 1, \ldots, T. \tag{4}$$

The left-hand side of the heat balance adds up external heat through purchase contracts (set U_H^{CONTR}), production by local production units (set U_H^{PROD}), and discharge minus charge of heat storages (set S_H).

The power balance equations for 15 min periods are formulated as:

$$\sum_{u \in U_{+P}} x_{u,t} - \sum_{u \in U_{-P}} x_{u,t} + \sum_{u \in S_P} \left(s_{u,t}^{OUT} - s_{u,t}^{IN} \right) = d_{P,t}, \quad t = 1, \ldots, 4T. \tag{5}$$

The power balance adds up power from units that supply power (set U_{+P}: PP contract, PV), subtracts power from power-consuming units (set U_{-P}: power sales, HPH), and adds discharge minus charge of power storages (set S_P).

3.4. Contracts for power and heat

Purchase contracts for power (PP) and heat (DH) include an energy-based fee $c_{u,t}$ (€/MWh), a yearly capacity-based fee c_u^{MAX} (€/MW), and a constant fee c_u^{CONST}. The purchase contracts are then defined as follows:

$$0 \le x_{u,t} \le x_u^{MAX}, \quad t = 1, \ldots, T, \ u \in U_H, \tag{6}$$

$$0 \le x_{u,t} \le x_u^{MAX}, \quad t = 1, \ldots, 4T, \ u \in U_P, \tag{7}$$

$$0 \le x_u^{MAX} \le F z_u, \tag{8}$$

$$z_u \in \{0, 1\}, \quad u \in U_H^{CONTR} \cup U_P^{CONTR}. \tag{9}$$

Here (6) and (7) set the capacity limits for the contracts. When the contract is included in the configuration ($z_u = 1$), the right-hand side of constraint (8) equals a big number F that allows the capacity x_u^{MAX} to be optimized. When the contract is not included ($z_u = 0$), constraint (8) sets capacity to zero. To force inclusion or exclusion of the contract in the configuration, the binary ON/OFF variable (9) can be fixed to 1 or 0. Otherwise the inclusion or exclusion of the component is optimized using MILP. In practice, the power contract must always be included; PV cannot supply enough power for the target buildings without enormous seasonal power storages.

Selling power back to the grid (PP-out) is represented by variables $x_{u,t} \ge 0$ with negative unit cost. For a small-scale producer who has a power purchase contract, no capacity-based fee is applied.

Estonia has feed-in premium for renewable power sold into the grid. Power sales with premium cannot exceed the PV production each hour. Thus, the Estonian model includes the additional constraints

$$x_{u,t} \le x_{PV,t}, \quad t = 1, \ldots, 4T, \ u \in U_P^{CONTR}. \tag{10}$$

Here $x_{PV,t}$ is the hourly PV production (see section 3.6).

3.5. Ground source heat pump

The set of heat production units U_H includes a ground source heat pump (HPH) for a heating. The heat pump consumes power at 15 min time intervals (variables $x_{u,t}$), but the produced heat is fed into the heat balance at hourly intervals (variables $x_{u,t}^H$). The heat pump model is defined by the following constraints.

$$0 \le x_{u,t} \le x_u^{MAX}, t = 1, \ldots, 4T, \tag{11}$$

$$x_{u,t}^H = \eta_t^H \left(x_{u,4t-3} + x_{u,4t-2} + x_{u,4t-1} + x_{u,4t} \right), \quad t = 1, \ldots, T, \tag{12}$$

$$0 \le x_u^{MAX} \le F z_u, \tag{13}$$

$$z_u \in \{0, 1\} \tag{14}$$

Constraints (11) set the capacity limit for the operating level (input power) of HPH. Constraints (12) define the hourly heat output as the sum of production during four consecutive 15 min

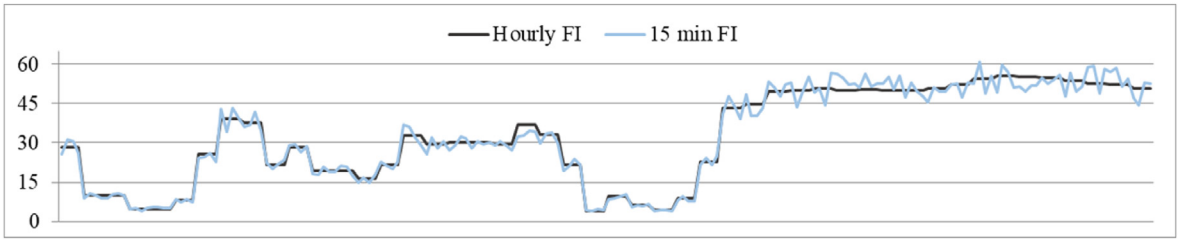


Fig. 3. Construction of 15 min power price data (€/MWh) from hourly ElSpot data in Finland.

periods using COP factors η_t^H. For an air-source heat pump the COP factor depends on outdoor temperature, but for a ground source heat pump the COP can be assumed constant. As for the energy contracts, F and binary variable z_u in (13)-(14) encode inclusion or exclusion of the heat pump. The direct operating costs for the heat pump are zero. Operating costs are caused indirectly by consumed electricity.

3.6. Photovoltaics

Production is computed for each 15 min time interval based on solar radiation intensity on the panel, the PV panel area, orientation of panels with respect to the sun, and panel efficiency factor η^{PV}. The PV model is:

$$x_{u,t} = \eta^{PV} A_u r_t \Big/ 4, t = 1, \ldots, 4T, \tag{15}$$

$$0 \le A_u \le z_u A_u^{MAX}, \tag{16}$$

$$z_u \in \{0, 1\} \tag{17}$$

In (15), variable $x_{u,t}$ is PV production, variable A_u is the panel area (m^2) and r_t is the intensity of the solar radiation (MW/m^2). This formulation assumes maximal PV production all times. Changing (15) into an inequality allows switching PV off. This is meaningful in situations where renewable power production exceeds demand and storage capacity, and a demand response request asks the building to reduce power production. As before, binary variable z_u in (16)-(17) encode inclusion or exclusion of PV.

3.7. Storages

The model for storages $u \in S$ is formalized as

$$s_{u,t} = \eta_u^S s_{u,t-1} + \eta_u^{IN} s_{u,t}^{IN} - s_{u,t}^{OUT} 1 \Big/ \eta_u^{OUT}, \tag{18}$$

$$s_u^{DEEP} s_u^{MAX} \le s_{u,t} \le s_u^{MAX}, \tag{19}$$

$$0 \le s_{u,t}^{IN} \le s_u^{IN,MAX}, \tag{20}$$

$$0 \le s_{u,t}^{OUT} \le s_u^{OUT,MAX}, \quad t = 1 \ldots, T \text{ when } u \in S_H \text{ and } t = 1, \ldots, 4T \text{ when } u \in S_P \tag{21}$$

$$0 \le s_u^{MAX} \le F z_u, \tag{22}$$

$$z_u \in \{0, 1\}, \tag{23}$$

$$s_{u.0} = s_u^{DEEP}, \quad u \in S. \tag{24}$$

Here dynamic constraints (18) link storage levels of subsequent hours together. Constraints (19) define deep discharge and capacity limits. The deep discharge limit is proportional to the variable capacity s_u^{MAX}. Constraints (20)–(21) limit charge and discharge rate. Heat storages (set S_H) operate hourly and power storages (set S_P) at 15 min periods. As before, F and binary variable z_u in (22)-(23) encode inclusion or exclusion of the storage unit. Constraint (24) sets the initial storage level to the minimal level. Alternatively, there could be a cyclic constraint defining the initial level equal to the level after the last period $s_{u,0} = s_{u,T}$. With a yearly model both approaches produce practically the same result. In operative use, the initial level is set to the current storage level.

To prevent power storage expiration before planned lifetime it is possible to restrict the number of full charge/discharge cycles during the planning horizon to a maximum of B_u. This is defined by an additional constraint

$$\sum_{t=1}^{4T} \eta_u^{IN} s_{u,t}^{IN} \le B_u \left(s_u^{MAX} - s_u^{DEEP} s_u^{MAX} \right) \tag{25}$$

4. Results and discussion

Solving the building energy model as a MILP problem gives the buildings the optimal configuration, dimensioning, and operation. However, to gain insight in how the different energy solutions interact, we also study several sub-optimal configurations for the two buildings. We have optimized the energy systems with different sizes of PV, flow battery, lead-acid battery, and heat storage. In each configuration, DH and PP contracts were included together with ground source HPH, and their capacity limits were optimized.

All models were solved using LP2, which is an efficient solver for large sparse LP/MILP problems using the product form of inverse and supporting upper and lower bounds for variables [55]. The models were validated hierarchically by testing each component first separately and then together. Sensitivity analysis was used to verify that the optimal solution reacted correctly to changes in input data.

4.1. Optimal solution and other configurations

Table 7 shows two optimized configurations for each building and corresponding annual total costs, operating costs, and fixed costs. Firstly, we observe that the optimal size of any kind of power storages was zero for both buildings in the optimal configurations and all other configurations that we have analyzed. Secondly, the heat storage was clearly cost-efficient in the optimal configurations and all other configurations. Thirdly, PV was cost-efficient in the Helsinki building but not in the Estonian building. The optimal PV panel area in the Helsinki building was 1033 m^2 when no space restriction for PV was set (first configuration). The persistent base load of the Helsinki building allows consuming PV locally even during the peak production hours in the summer. The second Helsinki configuration includes a space restriction of 330 m^2 PV panel area on the roof. The space restriction increased the annual total costs by 1271 €.

The first configuration for the Tallinn building shows that zero PV area was optimal. The unprofitability of PV in the Tallinn building is due to relative purchase and sales prices of power in Estonia, and the power load profile of the Tallinn building that does not coincide well with PV production. To introduce renewable power production, we included 330 m^2 PV on the roof of the Tallinn building in the second configuration. This is close to the maximal PV area that supports small-scale renewable energy production in Estonia [46]. The inclusion of 330 m^2 PV leads to only 290 € extra annual costs, so PV is close to being profitable.

Without local production and storages, the total energy costs are 159 322 € for the Helsinki building and 35 846 € for the Tallinn building. Thus, the optimal configurations cause yearly savings of 22 509 € in Helsinki and 7484 € in Tallinn.

4.2. Power and heat storage configurations

Next, we study selected energy storage configurations where

Table 7
Freely optimized and PV-constrained configurations for the two buildings. HPH capacity is for input power; heat output is COP (3.5) times larger.

Description	Unit	Helsinki		Tallinn	
		PV unrestricted	PV $\leq$ 330	PV unrestricted	PV = 330
DH capacity	MW	0.103	0.113	0.130	0.130
PP capacity	MW	0.209	0.215	0.061	0.060
HPH capacity (P)	MW	0.072	0.072	0.018	0.018
PV area	m^2	1033	330	0	330
HS size	MWh	1.059	0.703	0.180	0.161
PS size	MWh	0	0	0	0
Total costs	€/a	135 542	136 813	28 362	28 652
Operating cost	€/a	102 044	117 032	25 844	19 714
Fixed costs	€/a	33 498	19 781	2518	8938

different subsets of HS, PS-lead and PS-flow storages are included. In all models, 330 m^2 PV is included, while PP, and HPH capacities are optimized separately for each storage configuration. When included in a configuration, the capacity of PS-lead = 0.1 MWh, PS-flow = 0.5 MWh, and HS size is optimized. The considered configurations are:

- **PS-lead** included without other storages.
- **HS** included without other storages.
- **PS-lead & HS** included.
- **PS-flow & HS** included.
- **PS-flow & PS-lead & HS** included.

Note that the **HS** configurations are same as the PV-constrained optimal solutions in Table 7.

Table 8 reports the dimensioning and cost allocation for both buildings. First, we study the dimensioning and then the costs. For Helsinki, we see that without HS (the **PS-lead** configuration) the building needs a significantly larger DH contract than with HS (all other configurations). When HS is included, the optimal size of DH does not depend much on the presence of power storages. In Tallinn, the same DH size is used in all configurations because the DH contract does not include a capacity fee in Tallinn. The PP contract capacity is very similar across the different configurations in Helsinki. In Tallinn, when the large PS-flow battery is included, the system benefits from larger PP capacity. With larger battery capacity, the building can at times buy power into the storage and later sell more PV to the network while discharging the storage for its own use. Optimal HPH capacity is almost constant across different configurations, but a little larger in Tallinn when HS is excluded.

Optimal HS size is almost constant across the configurations in Helsinki (about 0.7 MWh). This suggests that HS does not interact significantly with PS. In Tallinn, optimal HS capacity is about 0.16 MWh without PS (**HS** configuration), but decreases as function power storage capacity to 0.14 MWh (**PS-flow & PS-lead & HS**). This indicates that HS and PS may compensate each other slightly in providing flexibility to the Tallinn building.

Looking at the overall costs, numbers show the unprofitability of both kinds of PS. Lowest total costs for both buildings are in the **HS** configuration, without PS. The configurations including PS together with HS (**PS-lead&HS, PS-flow&HS, PS-flow&PS-lead&HS**) have significantly higher total costs due to increase in fixed costs. Although, power storages cause savings in operational costs; these savings are not large enough to cover the PS investment. The annual savings in operational costs due to storages are actually very low, in the order of some hundreds of euros (maximally 914 € in Helsinki, 618 € in Tallinn).

By studying the breakdown of total costs by units, we observe that the greatest difference between the Helsinki and Tallinn

Table 8
Dimensioning and cost allocation for selected storage configurations for both buildings.

Description	Unit	PS-lead	HS	PS-lead & HS	PS-flow & HS	PS-flow & PS-lead &HS	PS-lead	HS	PS-lead & HS	PS-flow & HS	PS-flow & PS-lead & HS
		Helsinki					Tallinn				
Dimensioning											
DH	MW	0.182	0.113	0.113	0.113	0.113	0.130	0.130	0.130	0.130	0.130
PP	MW	0.205	0.215	0.204	0.207	0.204	0.059	0.060	0.057	0.078	0.101
HPH	MW	0.072	0.072	0.072	0.072	0.072	0.020	0.018	0.018	0.018	0.018
PV	m^2	330	330	330	330	330	330	330	330	330	330
HS	MWh	0	0.703	0.702	0.702	0.702	0	0.161	0.156	0.146	0.140
PS-flow	MWh	0	0	0	0.5	0.5	0	0	0	0.5	0.5
PS-lead	MWh	0.1	0	0.1	0	0.1	0.1	0	0.1	0	0.1
Overall costs											
Total	€	140 867	136 813	139 812	148 189	151 203	31 997	28 652	31 652	40 205	43 381
Oper	€	116 967	117 032	116 684	116 469	116 118	19 506	19 714	19 351	19 303	19 096
Fixed	€	23 900	19 781	23 128	31 720	35 085	12 492	8938	12 301	20 902	24 285
Total costs by unit											
DH	€	9116	7266	7260	7263	7264	607	412	413	414	420
PP	€	113 580	114 468	114 109	113 897	113 547	21 175	21 620	21 286	21 456	21 327
PP-out	€	0	0	0	0	−6	−1923	−1965	−1996	−2200	−2267
PV	€	6431	6431	6431	6431	6431	6431	6431	6431	6431	6431
HPH	€	8370	8369	8365	8363	8362	2338	2089	2087	2089	2087
HS	€	0	278	278	278	278	0	64	62	58	56
PS-flow	€	0	0	0	11 957	11 957	0	0	0	11 957	11 957
PS-lead	€	3369	0	3369	0	3369	3369	0	3369	0	3369

buildings is in power sales back to the grid (PP-out). PP-out is practically zero in Helsinki (highest revenue 6€). In Tallinn, PV is sold to the network for about 2000 € in each configuration. This is partly due to the Estonian renewable energy feed-in premium, partly because the power load profile of the Tallinn building makes it difficult to consume all PV locally. Larger PS in Tallinn supports a little higher PV sale, but the increase is far too small to make PS cost-efficient. The HS is quite cost-efficient in Helsinki because it enables savings in both DH energy and capacity fees. This happens by boosting the HPH operation: the HS is charged by HPH when power price or heat demand (or both) are low, and discharged to replace DH when power price or heat demand is high. In Tallinn, the benefit from HS is much smaller because it causes savings only in DH energy fee (no DH capacity fee in Tallinn).

4.3. Comparison of power storages for buildings

Different power storage technologies have different properties in terms of investment cost, lifetime, charge and discharge rate [56], capacity, storage losses, and roundtrip losses. Here we compare two common power storage technologies (the flow and lead-acid batteries) as part of building energy systems. While the flow battery is cheap per storage capacity (€/MWh), it suffers from low round-trip efficiency. The lead-acid battery has higher round-trip efficiency but also a somewhat higher price per storage capacity. For this reason, the lead-acid battery can be more suitable for short-term storage, while the flow battery may suit better as longer-term storage.

We wanted to study if long- and short-term power storages work in synergy together, or if only one of them is superior. Therefore, we optimized both building energy systems with flow battery sizes 0, 0.05, 0.1, 0.5 MWh, and lead-acid battery sizes 0, 0.05, 0.1 MWh. 330 m^2 PV was included in all these models, while PP, HPH, and HS capacities were optimized for each storage configuration.

Table 9 presents the increase in total costs, increase in investment costs, and decrease in operating costs for all storage combinations when compared to zero-size storages. Power storages are clearly non-profitable in both buildings. Although savings in operating costs increase monotonically as function of storage size, that is not enough to cover the increase in fixed (investment) costs. As a result, the total costs increase as a convex function of storage size. Smallest losses are obtained for both buildings by investing in the 0.05 MWh flow battery, which is the cheapest option considered in our study. Comparing the increase in investment costs with savings in operating costs allows us to judge how much investment costs must decrease (or savings in operating costs must increase) before the storages become cost-efficient. While the investment costs are the same for both buildings, the savings in operating costs depend on many local parameters. The ratio between fixed cost increase and operating cost decrease is in the range 9–21 in Helsinki, and in the range 8–29 in Tallinn. In both buildings, the smallest ratio is obtained by the smallest lead-acid battery of size 0.05 MWh. The ratio means that the investment price of lead-acid batteries is at least 9 and 8 times too high compared to savings in operational costs to make them cost-efficient in Helsinki and Tallinn buildings, correspondingly. To make power storages cost-efficient in buildings calls for lower price, extended lifetime, smaller losses, and lower interest rate. Increased power price volatility, greater imbalance between supply and demand, and smaller flexibility of the energy system in the absence of storages also work in the same direction by increasing the operational cost savings due to storage use.

We can use Table 9 to evaluate if there is synergy between a short-term and longer-term power storage. The savings in operative costs the combination of 0.05 MWh PS-lead and PS-flow batteries are 230 € in Helsinki, and 266 € in Tallinn. The average savings of a single 0.1 MWh PS-lead or PS-flow battery gives savings $(348 + 109)/2 = 228.5$ € in Helsinki, and $(363 + 139)/2 = 251$ € in Tallinn. So there is synergy, but it is minimal, worth 1.5 € in Helsinki and 15 € in Tallinn.

4.4. Optimal power and heat storage operation

Fig. 4 presents the optimal operation of storages for the Tallinn building in the selected storage configurations of Table 8. When included in the system, the size of PS-flow = 0.5 MWh, PS-lead = 0.1 MWh, and the size of HS is optimized separately in each configuration where it is included. The optimal storage operation for the Helsinki building was very similar to Tallinn, and same conclusions can be drawn. To save space, we omit the diagrams for Helsinki.

When only ***PS-lead*** is included (Fig. 4, first diagram), the storage operates throughout the entire year, but less intensively during the winter when PV production is negligible. PS-lead is charged to avoid wasting PV when the production is high, but power consumption for the HPH or the rest of the building is low. PS-lead is discharged as soon as the power can be used locally. It is rarely cost-

Table 9
Increase (€) in annual total costs = (fixed costs - operating costs) as function of flow and lead-acid power storages size (MWh). The cost increases are computed with the (0,0) storage size combination as reference. *The (0,0) reference cell shows the absolute costs.

Helsinki				
Lead\Flow	**0**	**0.05**	**0.1**	**0.5**
0	136 813 (19 781 + 117 032)*	1134 (1189–55)	2271 (2380–109)	11 376 (11 939–563)
0.05	1497 (1670–173)	2634 (2864–230)	3771 (4057–286)	12 889 (13 619–740)
0.1	3000 (3348-348)	4138 (4542–404)	5277 (5737–460)	14 391 (15 305–914)
Tallinn				
Lead\Flow	**0**	**0.05**	**0.1**	**0.5**
0	28 652 (8938 + 19 714)*	1120 (1194–74)	2250 (2389–139)	11 553 (11 964–411)
0.05	1470 (1681–211)	2610 (2876–266)	3760 (4070–310)	13 132 (13 654–522)
0.1	3000 (3363-363)	4155 (4560–405)	5318 (5757–439)	14 729 (15 347–618)

Fig. 4. Operative storage use in a hybrid system of a building in Tallinn. Data for 30 weeks (Monday-Sunday) (February 4 – September 1, 2019).

efficient to buy power into the storage when the price is low, and sell it back to the grid when the price is high. This is partly due to the relatively high losses in the charge/discharge cycle, partly due to limited storage capacity that can be used more cost-efficiently to optimize local PV use.

When only ***HS*** is included (Fig. 4, second diagram), the HS operates uniformly throughout the year to utilize heat demand and power price variations. When heat demand is low and power is cheap, the HS is charged using HPH, and occasionally using DH. When heat demand is high and power price is high, the HS is discharged to reduce the power consumption of HPH.

When ***PS-lead & HS*** are included (Fig. 4, third diagram), their respective usage patterns are surprisingly similar to the configurations where they were included individually. This means that there are not many synergies between heat and power storages in the building energy system. The HS operates in concert with DH and HPH to satisfy the heat demand as cheaply as possible. PP and PV dictate the PS-lead operation to optimize power consumption. While the HPH connects the building's power consumption and heat production, this connection does not cause significant dependencies between the HS and PS-lead operation.

Fourth and fifth diagrams in Fig. 4 show the operation of **PS-flow & HS** and for **PS-flow & PS-lead & HS**, respectively. To save space, we omit the diagram for PS-flow alone. The optimal operation of the PS-flow battery is quite different from PS-lead, because the charge/discharge cycle of PS-flow involves significantly larger losses. Regardless if PS-flow operates by itself, or together with HS, PS-lead, or both, it is charged and discharged much less frequently, only when power price or balance between power supply and demand changes significantly in time. PS-flow operation is also surprisingly similar, regardless in which configuration it is included.

4.5. Discussion

The optimized configurations with renewable HP and PV production and HS resulted in significant annual savings in total energy costs when compared to the original configuration of the buildings: about 22 500€ in Helsinki and 7500€ in Tallinn. While saving costs, the optimized configurations also improved significantly the energy efficiency of both buildings. Energy efficiency is measured as the *E-value*, which represents annual non-renewable primary energy consumption for HVAC (heating) per floor space [57]. The E-value is calculated by adding up consumption of different energy forms (DH, PP) multiplied by their *primary energy factors* and dividing the sum by heated floorspace. PV production can be subtracted from HPH power consumption even if it is used elsewhere in the building or sold back to the grid. The primary energy factors are 1.2 for PP and 0.5 for DH in Helsinki and 2.0 for PP and 0.65 for DH in Tallinn [58]. In Helsinki the initial E-value was 65 kW/m^2 and dropped to 37 kWh/m^2 in the optimal configuration. In Tallinn the initial E-value was 41 kW/m^2 and dropped to 16 kWh/m^2 in the optimized configuration with 330 m^2 PV.

5. Conclusions and future research

We have developed an LP/MILP model for optimizing dimensioning and operation of renewable hybrid energy solutions of buildings based on 15 min power balance settlement to be introduced soon. The model includes multiple energy storages, such as a heat storage and different kinds of battery power storages. We have applied the model to plan the retrofitting of an office building in Helsinki and a residential building in Tallinn, with PV, ground source heat pump, hot water tank, and flow and lead-acid batteries.

Overall, the optimized configurations caused significant annual savings in energy costs for both buildings (22 500 € in Helsinki, 7500 € in Tallinn) while reducing non-renewable primary energy consumption. PV was cost-efficient in the Helsinki building. The optimal PV size for the Helsinki building was over 1000 m^2, but only 330 m^2 could be fitted on the rooftop. The key to the cost-efficiency of PV in Helsinki was that practically all PV could be consumed locally in the building, without having to sell PV power to the grid at a very low price. This was facilitated by the power load

profile of the Helsinki building – power demand remained sufficiently high in the summer days when PV production was high.

In the Tallinn building, PV was, somewhat surprisingly, unprofitable. However, the deficit caused by 330 m^2 PV investment was small, about 290 €/a. There are primarily two reasons for the unprofitability: the power load profile of the Tallinn building and the level of the Estonian feed-in premium. The power load of the Tallinn building is not high enough during the PV peak days to allow all PV to be consumed locally, and a part of PV power must be sold to the grid. Although Estonia has a feed-in premium for small-scale renewable power production, it is not sufficiently high to make PV sales to the grid cost-efficient.

The HS was is highly cost-efficient, in both buildings, because it allows higher utilization ratio of the HPH causing savings in DH energy fees. In Helsinki, the HS also caused savings in the fixed DH contract costs by allowing smaller DH power limit.

It was disappointing to see that PSs were grossly unprofitable in both buildings even subject to 15 min power balance. The investment costs (annuity) of PS were between 9 and 21 times higher than the caused savings in yearly operational costs in the Helsinki building and 8 to 29 times higher in the Tallinn building. Based on these ratios, it seems unlikely that PS will reach the break-even point for cost-efficiency soon.

An interesting observation is that the power and heat storages do not interact strongly, even in the presence of the ground source heat pump. This observation applies both to the optimal dimensioning of the technologies and to the operation of the storages. The heat storage operates in concert with district heating and the ground source heat pump while power storages operate together with photovoltaics and power trade. However, the optimal operation of the PS-flow battery is quite different from PS-lead. Due to higher charge/discharge cycle losses, the PS-flow is charged and discharged less frequently than the PS-lead battery. Only significant variations in power price or in power demand can make it cost-efficient or necessary to use the PS-flow battery. Operation of both battery types is otherwise surprisingly similar, regardless in which configuration they are included.

Future research could involve extending the current model with other RES technologies, including cooling, and applying it to other buildings in different locations and in different local conditions. The cost-efficiency of PS can maybe be improved by participating in the frequency containment reserve (FCR) market. In the future, it would be interesting to evaluate if the FCR market can make PS cost-efficient. The current model could be modified to compute marginal power prices to form a basis for bids on the FCR market.

References

[1] Rikkas R, Lahdelma R. Energy supply and storage optimization for mixed-type buildings. Energy 2021;231:120839. https://doi.org/10.1016/j.energy.2021.120839.
[2] Jiang Y, Kang L, Liu Y. Optimal configuration of battery energy storage system with multiple types of batteries based on supply-demand characteristics. Energy 2020;206:118093.
[3] Koskela J, Rautiainen A, Järventausta P. Using electrical energy storage in residential buildings – sizing of battery and photovoltaic panels based on electricity cost optimization. Appl Energy 2019;239:1175–89.
[4] Mohammadi M, Ghasempour R, Astaraei FR, Aligholian EAA, Toopshekan A. Optimal planning of renewable energy resource for a residential house considering economic and reliability criteria. 2017. Electrical Power Energy Sys. 2018;96:261–73.
[5] Truong C, Naumann M, Karl R, Müller M, Jossen A, Hesse H. Economics of residential photovoltaic battery systems in Germany: the case of Tesla's powerwall. Batteries 2016. https://doi.org/10.3390/batteries2020014.
[6] Rinaldi A, Soini M, Streicher K, Patel M, Parra D. Decarbonizing heat with optimal PV and storage investments: a detailed sector coupling modelling framework with flexible heat pump operation. Appl Energy 2020;282(2021):116110.
[7] Zhao H, Guo W. Coordinated control method of multiple hybrid energy storage systems based on distributed event-triggered mechanism. Electrical Power Energy Sys. 2020;127:106637. 2021.
[8] Luo F, Shao J, Jiao Z, Zhang T. Research on optimal allocation strategy of multiple energy storage in regional integrated energy system based on operation benefit increment. 2020. Electrical Power Energy Sys. 2021;125:106376.
[9] Ghenai C, Bettaye M. Modelling and performance analysis of a stand-alone hybrid solar PV/Fuel Cell/Diesel Generator power system for university building. Energy 2019;171:180–9.
[10] Mehrjerdi H, Iqbal A, Rakhshani E, Torres JR. Daily seasonal operation in net-zero energy building powered by hybrid renewable energies and hydrogen storage systems. Energy Convers Manag 2019;201:112156.
[11] Okundamiya MS. Size optimization of a hybrid photovoltaic/fuel cell grid connected power system including hydrogen storage. Int J Hydrogen Energy 2020. https://doi.org/10.1016/j.ijhydene.2020.11.185.
[12] Prada A, Bee E, Grigiante M, Baggio P. On the optimal mix between lead-acid battery and thermal storage tank for PV and heat pump systems in high performance buildings. Energy Proc December 2017;140:423–33.
[13] Khana SU-D, Almutairi ZA, Al-Zaida OS. Development of low concentrated solar photovoltaic system with lead acid battery as storage device. Curr Appl Phys 2020;20:582–8.
[14] Azaza M, Eriksson D, Wallin F. A study on the viability of an on-site combined heat- and power supply system with and without electricity storage for office building. Energy Convers Manag 2020;213:112807.
[15] Kumar J, Parthasarathy C, Västi M, Laaksonen H, Shafie-Khah M, Kauhaniemi K. Sizing and allocation of battery energy storage systems in Åland Islands for large-scale integration of renewables and electric Ferry charging stations. Energies 2020;13:317. https://doi.org/10.3390/en13020317.
[16] Campana PE, Landelius T, Andersson S, Lundström L, Nordlander E, He T, Zhang J, Stridh B, Yan J. A gridded optimization model for photovoltaic applications. Sol Energy 2020;202:465–84.
[17] Bryans D, Amstutz V, Girault HH, Berlouis LEA. Characterisation of a 200 kW/400 kWh Vanadium redox flow battery. Batteries 2018;4:54. https://doi.org/10.3390/batteries4040054.
[18] Yan C, Wang F, Pan Y, Shan K, Kosonen R. A multi-timescale cold storage system within energy flexible buildings for power balance management of smart grids. Renew Energy 2020;161:626–34. https://doi.org/10.1016/j.renene.2020.07.079.
[19] Spodniak P, Ollikka K, Honkapuro S. The impact of wind power and electricity demand on the relevance of different short-term electricity markets: the Nordic case. Appl Energy 2021;283:116063.
[20] Sharma S, Xu Y, Verma A, Panigrahi BK. Time-coordinated multienergy management of smart buildings under uncertainties. IEEE Trans Ind Inf August 2019;15(8).
[21] Oy Granlund. Heating and power data for an office building in Helsinki. "Personal communication"; 2020.
[22] Tallinn Technical University. Heating power demand data for residential building in Tallinn. "Personal communication"; 2021.
[23] European Union Law. Directive 2010/31/EU of the European parliament and of the council of 19 may 2010 on the energy performance of buildings. Article 9. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:153:0013:0035:en:PDF. [Accessed 17 November 2020].
[24] Väisänen J, Kosonen A, Ahola J, Sallinen T, Hannula T. Optimal sizing ratio of a solar PV inverter for minimizing the levelized cost of electricity in Finnish irradiation conditions. Sol Energy 2019;185:350e62.
[25] Energy Policies of IEA Countries. Estonia. Review 2019, (pdf-file). [Accessed 19 May 2021].
[26] Kuosa M, Rahiala S, Tallinen K, Mäkilä T, Lampinen M, Lahdelma R, Pulkkinen L. Mass flow-controlled district heating with an extract air heat pump in apartment buildings: a practical concept study. Appl Therm Eng

2019;157:8.
[27] Volkova A, Latõšov E, Lepiksaar K, Siirde A. Planning of district heating regions in Estonia. Int. J. Sustain. Energy Plann. Manag. 2020;27:5–16.
[28] Helen Ltd. Heating and cooling. District heating and water flow price list from 2018. 2020. https://www.helen.fi/en/heating-and-cooling/district-heat/district-heat-prices. [Accessed 3 September 2020].
[29] Ministry of Economic Affairs and Communications. Estonia. District heating. Organisation and price of district heating. [Accessed 17 September 2020].
[30] Helen Ltd. Power distribution tariffs. Price list from July 2018. Available at: https://www.helen.fi/en/electricity/electricity-products-and-prices. [Accessed 3 September 2020].
[31] Elektrilevi OÜ. Estonian distribution system operator. Price list 2019. "Personal communication"; 2020.
[32] Huang J, Fan J, Furbo S. Demonstration and optimization of a solar district heating system with ground source heat pumps. Sol Energy 2020;202:171e89.
[33] Helen Ltd. PV system prices. The explanation of calculations to main author through customer service system. Modified calculation presented in the paper. "Personal communication"; 2020.
[34] Lahdelma R, Kayo G, Abdollahi E, Salminen P. Optimization and multicriteria evaluation of district heat production and storage. In: book: new Perspectives in multiple criteria decision making; 2019. 371e96. https://doi.org/10.1007/978-3-030-11482-4_15.
[35] Irena. Electricity storage and renewables: costs and markets to 2030. October 2017 (pdf-file), . [Accessed 19 November 2020].
[36] Hydroware, U.S. Department of Energy. Energy storage technology and cost characterization report. 2019 (pdf-file), . [Accessed 19 November 2020].
[37] Linsam Technology. Heat pump COP factor. 2020. Available at: https://linsamtech.com/documents-download/. [Accessed 5 September 2020].
[38] Olabi AG, Onumaegbu C, Wilberforce T, Ramadan M, Abdelkareem MA, Al-Alami AH. Critical review of energy storage systems. Energy 2021;214:118987.
[39] Munzke N, Schwarz B, Büchle F, Hiller M. Evaluation of the efficiency and resulting electrical and economic losses of photovoltaic home storage systems. 2020. J Energy Storage 2021;33.
[40] Truong CN, Naumann M, Karl RCh, Müller M, Jossen A, Hesse HC. Economics of residential photovoltaic battery systems in Germany: the case of Tesla's powerwall. Batteries 2016;2:14. https://doi.org/10.3390/batteries2020014.
[41] Oy Kiwatti. Energy storages increase independence. Available at: https://www.kiwatti.fi/en/. [Accessed 5 September 2020].
[42] Power Sonic Company. How to charge a Lead-acid battery. Available at: https://www.power-sonic.com/blog/how-to-charge-a-lead-acid-battery/. [Accessed 5 September 2020].
[43] Zakeri B, Syri S. Electrical energy storage systems: acomparative life cycle cost analysis. Renew Sustain Energy Rev 2015;42:569–96.
[44] Battery University. BU-210: how does the fuel cell work?. Available at: https://batteryuniversity.com/article/bu-210-how-does-the-fuel-cell-work. [Accessed 3 September 2020].
[45] Deloitte Centerfor Energy Solutions. Electricity storage technologies, impacts, and prospects. September 2015 (pdf-file), . [Accessed 5 September 2020].
[46] Staffell I, Scamman D, Abad AV, Balcombe P, Dodds PE, Ekins P, Shahd N, Warda KR. The role of hydrogen and fuel cells in the global energy system. Energy Environ Sci 2019;12:463.
[47] Khana S, Zeyad Ammar Almutairia ZA, Omer Salah Al-Zaida OS, Ud-Din Khane S. Development of low concentrated solar photovoltaic system with lead acid battery as storage device. Curr Appl Phys 2020;20:582–8.
[48] Tesla PowerPack1. Description, data sheet (pdf-file). [Accessed 3 September 2020].
[49] Clean Technica. Zachary Shahan (exclusive. 5.5.2020). Tesla megapack, Powerpack, & Powerwall battery storage prices per kWh.
[50] Mongird K, Viswanathan V, Balducci P, Alam J, Fotedar V, Koritarov V, Hadjerioua B. An evaluation of energy storage cost and performance characteristics. Energies 2020;13:3307. https://doi.org/10.3390/en13133307.
[51] Elering. Transmission system operator, Estonia. Renewable energy subsidy. Available at: https://elering.ee/en/renewable-energy-subsidy. [Accessed 5 April 2021].
[52] Bellini E. *PV Magazine*. Estonia to replace feed-in premium scheme for renewables and solar with auction mechanism [Online]. 15 March 2017. . [Accessed 1 July 2021].
[53] Nordic Balancing Model (NMB). Nordic TSOs: 15 minutes balancing period from 22 May 2023. Available at: Nordic TSOs: 15 minutes balancing period from 22 May 2023 – nordicbalancingmodel [Accessed 18.5.2021].
[54] Jaakamo N. Impact of the 15-minute imbalance settlement period and electricity storage on an independent wind power producer. Master Thesis. Aalto University, School of Electrical Engineering; 23.12.2019.
[55] Lahdelma R, Hakonen H. An efficient linear programming algorithm for combined heat and power production. Eur J Oper Res 2003;148:141–51. https://doi.org/10.1016/S0377-2217(02)00460-5.
[56] World Nuclear Association. Electricity and energy storage. Available at:] https://world-nuclear.org/information-library/current-and-future-generation/electricity-and-energy-storage.aspx, . [Accessed 19 May 2021].
[57] Decree of the Ministry of the Environment on the Energy Performance of New Buildings. Ympäristöministerion asetusuuden rakennuksen energiatehokkuudesta (pdf-file). 27.1. 2017. . [Accessed 17 November 2020].
[58] Kurnitski J. NZEB requirements in Nordic countries. REHVA European HVAC J 2019;6. Available at: https://www.rehva.eu/rehva-journal/chapter/nzeb-requirements-in-nordic-countries. [Accessed 19 May 2021].

Multistep electric vehicle charging station occupancy prediction using hybrid LSTM neural networks

Prajnadipta Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, prajnadipta.sahoo@yahoo.co.in*

Pratik Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, pratikmohanty92@hotmail.com*

Alekha Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, alekha.sahoo241@gmail.com*

Ajit Kumar Panda, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, akpanda555@hotmail.com*

Keywords:
Long short-term memory
Charging occupancy
Electric vehicle
Forecasting

A B S T R A C T

Public charging station occupancy prediction plays key importance in developing a smart charging strategy to reduce electric vehicle (EV) operator and user inconvenience. However, existing studies are mainly based on conventional econometric or time series methodologies with limited accuracy. We propose a new mixed long short-term memory neural network incorporating both historical charging state sequences and time-related features for multistep discrete charging occupancy state prediction. Unlike the existing LSTM networks, the proposed model separates different types of features and handles them differently with mixed neural network architecture. The model is compared to a number of state-of-the-art machine learning and deep learning approaches based on the EV charging data obtained from the open data portal of the city of Dundee, UK. The results show that the proposed method produces very accurate predictions (99.99% and 81.87% for 1 step (10 min) and 6 steps (1 h) ahead, respectively, and outperforms the benchmark approaches significantly (+22.4% for one-step-ahead prediction and +6.2% for 6 steps ahead). A sensitivity analysis is conducted to evaluate the impact of the model parameters on prediction accuracy.

1. Introduction

Electric vehicles (EVs) have been promoted as a widely accepted solution to reduce global CO_2 emissions and climate change. To make low-emission energy alternatives widely accepted, charging and maintenance infrastructure needs to be widely available across Europe [1]. While there has been an increase in recharging facilities installed in different countries, there is still a limited number of fast/rapid chargers (also referred to as charging points) due to their high investment cost. For example, at the end of 2020, there were only 51 public charging points in Manhattan, New York City (51 Level 2 and 0 Level 3 chargers).[1] The limited number of fast/rapid public chargers has become one of the major obstacles for widespread EV adoption [2]. As there are more and more EVs running on the road, it becomes a struggle to find a charging point before running out of battery. While existing EV station platforms like ChargePoint (www.chargepoint.com)http://www.chargepoint.com/or ChargeHub (https://chargehub.com) provide real-time charging point availability information for users, reservation in advance on public charging stations is still not available [3]. EV users might end up waiting in a queue when arriving at a charging station that was unoccupied a couple of minutes previously. A recent study shows that operating a fleet of EVs for transport network companies brings additional challenges as EVs need to charge several times a day and primarily rely on fast/rapid chargers [4]. Mitigating congestion at public fast/rapid charging stations has become an important issue for the efficiency of charging infrastructure management and for improving overall user experience and EV acceptance by the general public.

For this purpose, predicting charging occupancy patterns allows charging service platforms to better manage the limited charging resources available and reduce a customer's charging waiting time loss. For example, with predicted charging waiting times at charging stations, a real-time vehicle−charging station assignment/recommendation system could be developed to reduce the charging waiting time of EV fleets [5−8]. It can also support the development of apps for reducing vehicle idle time when terminating charging sessions [9] that are directly integrated into the vehicle's user interface (with cellular connectivity) or on the user's

smartphone. These smart charging service management applications rely on accurate EV charging pattern predictions for a short-term time horizon (e.g. from 30 min to several hours). However, there are still few studies that address this issue, and they are mainly based on conventional econometric or time series methodologies with limited accuracy [10–12].

Modelling the discrete EV charging occupancy state (i.e. a charger is occupied or not at a discrete-time index) at charging points is challenging as the target-dependent variable is a non-stationary binary time series, and there is very limited informa-tion available on public charging datasets [13]. Fig. 1 illustrates an example of such discrete charging occupancy states over each discrete time interval (10 min) of a rapid charger in the City of Dundee, UK. Irregular within-day and day-to-day charging occupancy patterns can be observed. In such a problem, it is not difficult to incorporate previous occupancy states to predict the occupancy state at the next time step. However, it is more complex to consider long-term tendencies and to predict the charging occupancy sequence over multiple time steps. In general, public EV charging session data contains limited information such as the start and end times of charging sessions, the amount of energy charged of each session, the charger type or power, the geographical location of the charger, and the charger or customer ID.

With more EV charging data made available publicly [13,14,16], several scholars have started modelling EV charging patterns [10,15]. Amara-Ouali et al. [13] provided a complete list of 60 publicly available EV charging datasets relevant for EV charging load modelling. The available data fields in these datasets are limited, as previously mentioned. Given the variables for which data is available, the occupancy and charging load can be calculated. Two categories of problems are generally studied using these public charging datasets.

The first category concerns modelling the charging load of individual EVs and charging stations. The objective is to predict the charging load profiles to evaluate the impact on the power grid or to develop algorithms for smart charging management [16]. Existing studies mainly apply statistical models to estimate the probability distribution of charging loads. Majidpour et al. [17] applied machine learning approaches to predict aggregate charging loads at EV charging stations. Lee et al. [16] applied Gaussian mixture models to estimate the distributions of charging arrivals, the duration of each charging session, and the amount of energy charged. Flammini et al. [14] estimated a mixture of multivariate normal distributions for the distribution of the number of charging sessions on the public charging network in the Netherlands.

The second category considers the problem of modelling and predicting the charging occupancy profile at chargers. The outcome allows estimating energy demand together with charging power or designing algorithms to assign EVs to chargers with the least charging waiting time loss [18,19]. For example, Gruoss et al. [20] proposed a Markov chain model to model the occupancy state of charging stations based on data collected for a fleet of car-sharing vehicles on about 40 public charging stations. However, the prediction accuracy of the proposed model is not reported. Bikcora et al. [10] applied an autoregressive logistic model for day-ahead charging station availability and charging rate prediction. Iversen et al. [15] proposed a state-space model to predict the status of EVs (driving or idle) over multiple discretised time steps. Verma et al. [21] analysed the load profile of household energy consumption and applied different classification methods, including classification and regression trees, random forest, and k-nearest neighbour methods, to identify the energy consumption profiles of households with or without EV charging. Motz et al. [11] applied the logistic regression model to predict the charging station occupancy using ACN charging data [16]. Soldan et al. [12] also applied the logistic regression model to predict the charging station occupancy using time-related and historical occupancy as input features. Existing studies are mainly based on econometric or machine learning approaches for charging load and occupancy prediction.

Recently, deep learning (DL) approaches have received increasing interest and have been successfully applied in speech recognition, natural language translation, computer vision, and medical image analysis, among many other fields. In particular, the long short-term memory (LSTM) network and its variants have

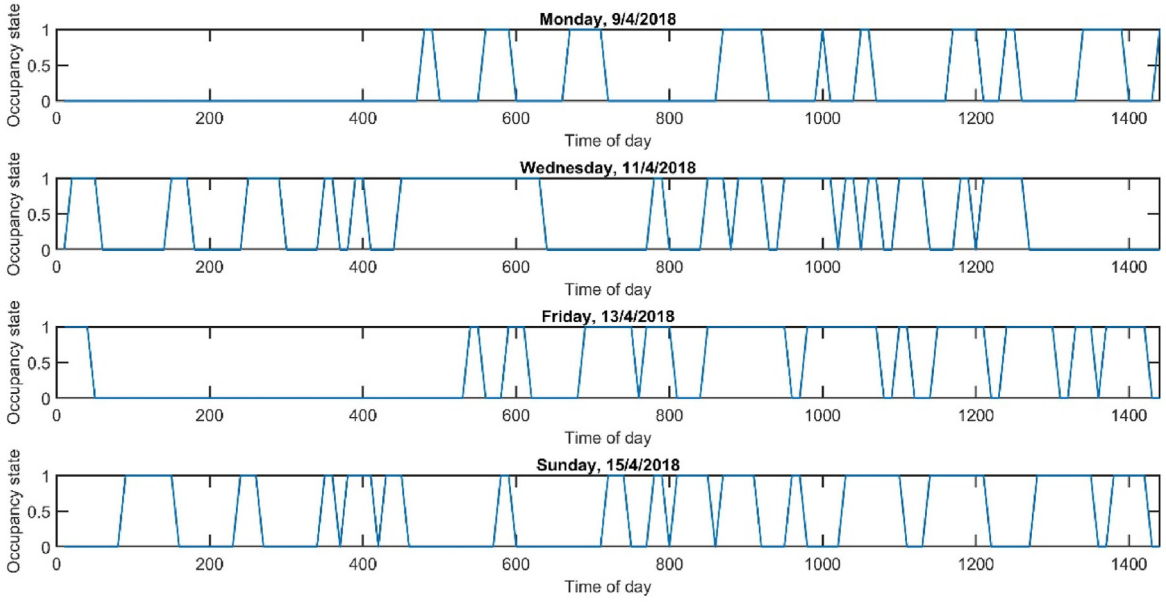


Fig. 1. Example of the charging occupancy profile of a rapid charger on different days of the week (0: unoccupied, 1: occupied by an EV).

Table 1
Descriptive statistics of EV charging sessions in the study area.

Charger type	Number of charging points	Number of charging sessions	Number of charging sessions per day per charging point	Charging duration (minutes)	
				Mean	S.D.
Slow (7 kW)	40 (70.2%)	5603	1.5	725.8	1338.5
Fast (22 kW)	8 (14.0%)	1357	1.9	413.3	819.9
Rapid (43 kW or above)	9 (15.8%)	8941	10.9	28.3	36.4
Total	57 (100%)	15,901	3.1	306.9	892.0

been successfully applied in various time series forecasts (see the recent review in Ref. [22]. An LSTM network is a kind of recurrent neural network allowing the modelling of complex temporal dependency in time series data and overcoming the vanishing gradient problem. In the energy field, DL has been successfully applied to residential load profile forecasting [23–25], energy consumption prediction [26,27], and gas consumption profile prediction [28]. However, the problems considered in these studies are regression problems. There are still few studies that tackle discrete EV charging occupancy state modelling problems.

In this study, we propose a hybrid LSTM neural network that combines both LSTM and forward neural networks to merge heterogeneous features for the prediction of EV charging occupancy over a planning horizon. The results show that the proposed approach outperforms the state-of-the-art machine learning approaches and benchmark deep learning networks using the public charging data from the City of Dundee, UK, in 2018. The main contributions of this paper are summarized as follows.

- We develop a new hybrid LSTM method to predict discrete EV charging occupancy sequences for multistep prediction. We generate new day-type tendency features from limited fields of public EV charging station data to increase significantly the prediction accuracy. The proposed mixed network structure allows merging heterogeneous data types, providing a generalized and flexible framework for time series forecasting based on LSTM neural networks.
- Our experiments show that the proposed approach outperforms significantly classical time series (logistic regression), machine learning approaches (support vector machine (SVM), random forest, and Adaboost), and benchmark deep learning networks (LSTM, Bi-LSTM, and GRU). The proposed method is easy to implement and can be applied for efficient charging infrastructure management.
- We also analyze the EV charging session data of the city of Dundee to understand the charging patterns of users and charging load profiles in the study area and evaluate the influence of different hyperparameters of the hybrid LSTM method on prediction accuracy.

This paper is organized as follows. Section 2 presents the dataset and the exploratory analysis of the data. Section 3 presents the proposed hybrid LSTM model, performance metrics, and alternative benchmark methods for multistep EV charging occupancy state prediction. In Section 4, we report the performance of the proposed method and compare it with the benchmark methods. A sensitivity analysis is conducted to evaluate to what extent different model parameters affect the prediction accuracy for multistep prediction. Finally, we discuss the results and offer some concluding remarks.

2. Data collection and pre-processing

2.1. Dataset

In this paper, we consider EV charging data from the open data portal of the city of Dundee, UK,[2] which provides datasets describing various EV charging sessions. Each session contains charger identifiers, start and end times of charging sessions, amount of energy charged, the power of chargers, and the geographical locations of chargers. There are three types of chargers in the study area: slow chargers (7 kW), fast chargers (22 kW), and rapid chargers ($\geq$ 43 kW).[3]

In the present study, a three-month charging session dataset from March 5, 2018, to June 4, 2018 (91 days) is used. Table 1 reports the descriptive statistics. There is a total of 40 slow, 8 fast, and 9 rapid chargers in the data; 56.2% of charging sessions are realized at rapid chargers, while 35.2% and 8.5% are at the slow and fast chargers, respectively. Due to the low charging occupancy of slow and fast chargers (1.5 and 1.9 sessions per day-charger for slow and fast chargers, respectively), we focus on the more challenging issue of modelling the occupancy of rapid chargers (10.9 sessions per day-charger, on average), which present very irregular within-day and day-to-day occupancy patterns as shown in Fig. 1.

We first delete the outliers: charging durations at rapid chargers that are more than three standard deviations from the median [21]. A total of 0.79% outliers are removed. Ultimately, 8870 rapid charger charging sessions are used for this study. The average charging duration of the rapid chargers is 28.3 min, with a standard deviation of 36.4 min (see Table 1). The geographical locations of the charging stations in the study area are shown in Fig. 2.

2.2. Characteristics of charging occupancy at rapid chargers

We further analyze the charging behaviour of users and the charging occupancy patterns at rapid chargers. The upper part of Fig. 3 shows the EV plug-in time distributions on weekdays and weekends. On weekdays, most charging sessions occur from 7:00 to 21:00, with a peak between 12:00 and 14:00. For weekends, the plug-in time profile is smoother compared to that of weekdays. The number of overnight (from 0:00 to 6:00) charges are doubled on the weekend. Regarding the charging duration on weekdays and weekends (the middle part of Fig. 3), almost all charging sessions are less than 60 min, and mostly between 10 and 40 min. The lower part of Fig. 3 shows the distribution of the charged energy on weekdays and weekends. The charging occupancy profiles are irregular and present a significant difference between weekdays and weekends, as shown in Fig. 4.

Fig. 2. The spatial distribution of chargers in the city of Dundee *(source:* https://data.dundeecity.gov.uk/dataset/ev-charging-data*).*

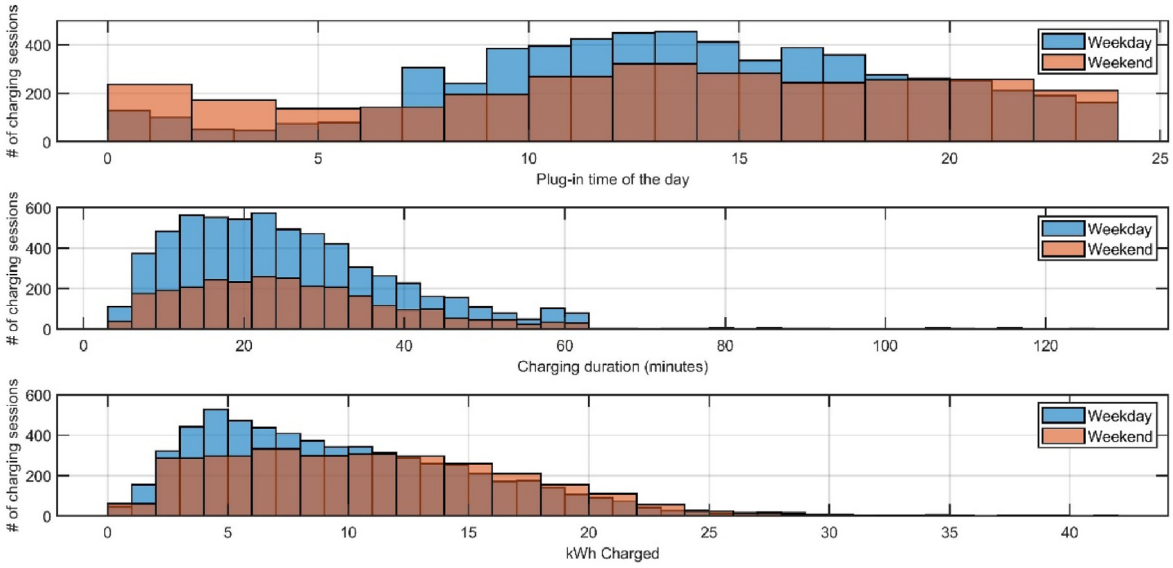


Fig. 3. Distribution of EV plug-in times, charging duration, and energy charged using rapid chargers on weekdays and weekends (March 5, 2018, to June 4, 2018).

2.3. Factors influencing charging occupancy profiles

The occupancy profile of chargers depends on various factors, including the time of the day, day of the week, weekday/weekend, charging power, remaining battery level of EVs, battery capacity, energy price, the geographical location of charging stations, and idle time during which a vehicle is fully charged but still occupies a charger [9]. Given the limited data fields available in our datasets, the features generated for modelling the charging occupancy state include the time of day, day of the week, weekend, and the past charging occupancy states and average charging occupancy rates on the type of the day (weekday or weekend), as shown in Table 2. The objective is to predict the charging occupancy state profile for each charger for multiple steps ahead (from 10 min to several hours). For this purpose, we discretised one day (24 h) into 144 discrete time slots with 10-min intervals. The entire dataset is divided into a training dataset (first 70% of the dataset) and a test dataset (remaining 30%of the dataset). The auto- and partial correlations of charging occupancy states show that the occupancy state at time t is correlated with its past states at times $t-1$ and $t-2$ (Fig. 5). Different from existing studies that consider time-related and historical occupancy state features only, we create a long-term charging occupancy tendency as an additional feature to help predict the charging occupancy state for multiple steps ahead. Fig. 4 illustrates an example showing that the charging occupancy profiles are volatile and distinct on weekdays vs weekends.

3. Occupancy state prediction models

3.1. Proposed hybrid LSTM model

Consider an EV charger occupancy state time series y_t for a

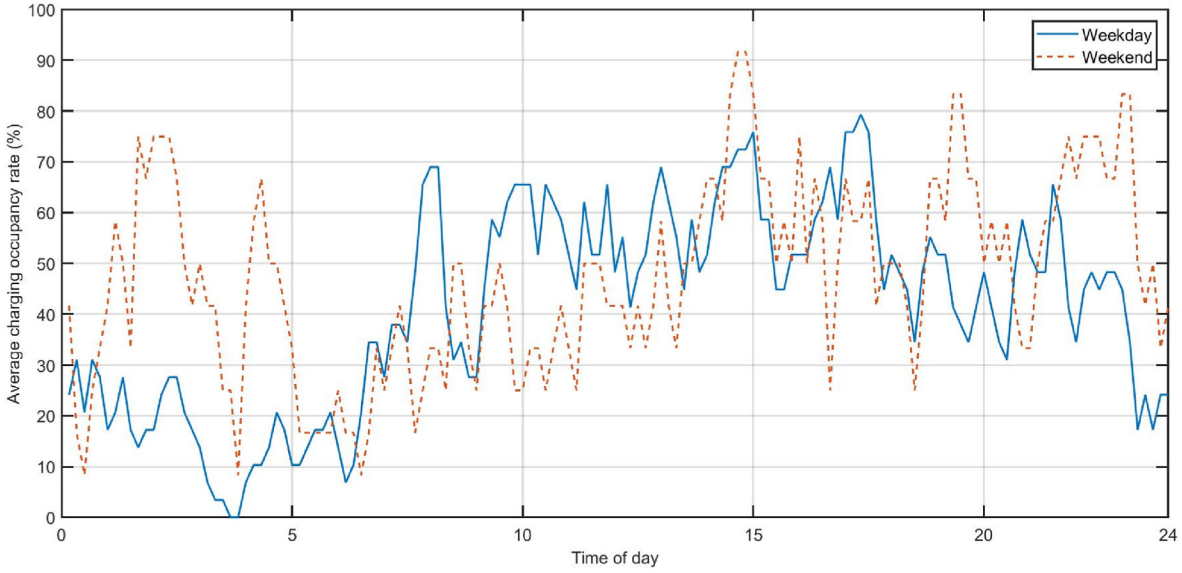


Fig. 4. Example of the average charging occupancy rate on weekdays and weekends.

Table 2
Features used to model the charging occupancy of rapid chargers.

Feature (Variable)	Meaning
Time of day (t)	Time index belonging to $t \in \{1, 2, \ldots, 144\}$ for 24 h
Day of week (d)	Sunday = 0, Monday = 1, …, Saturday = 6
Weekday/weekend (w)	1 if charging occurs on the weekend and 0 otherwise
Average charging occupancy rate profile for weekday/weekend ($\boldsymbol{p}$)	A vector of 144 continuous variables representing the tendency of the charging occupancy rate of a charger on weekdays or the weekend. Two constant vectors are generated, one for weekdays and another for the weekend, calculated from the 70% training dataset.
Past charging occupancy states ($\boldsymbol{y}$)	A sequence of historical k-step backward charging occupancy states from t (i.e. $t-1, t-2, \ldots, t-k$).

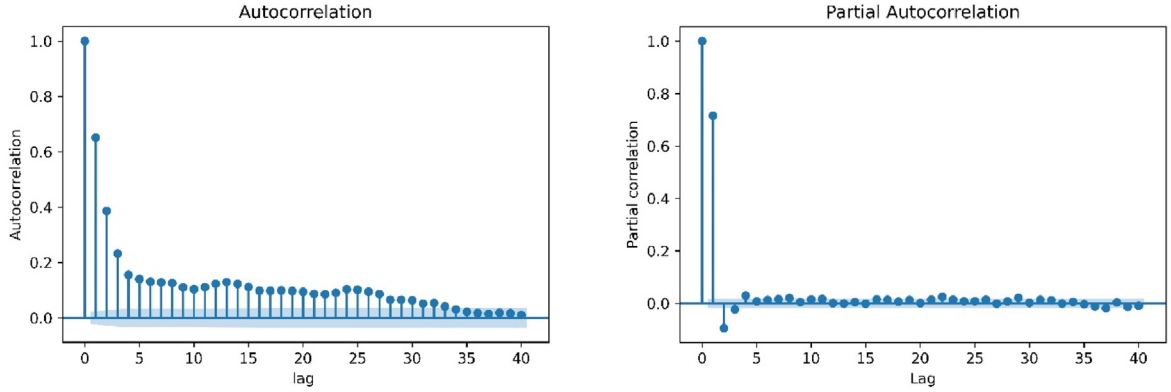


Fig. 5. Example of the auto- and partial correlations of the charging occupancy states at chargers.

charger, where t is a discrete-time index $t \in T = \left\{1, 2, \ldots, \frac{H}{\Delta}\right\}$, with Δ being the discrete time interval and H being 24 h. Let y_t be 1 if there is an EV charging event that occurs during $[(t-1)\Delta, t\Delta)$, and 0 otherwise. Given a sequence of past charging occupancy states $(y_{t-1}, y_{t-2}, \ldots)$ before t, we aim to predict the charging occupancy state sequence for a short-term time window ahead, i.e. $t+1, \ldots, t+k-1$.

We propose a hybrid LSTM model taking into account both local temporal dependency and long-term tendency to predict the charging occupancy state sequence. Fig. 6 shows the network structure of the model. The short-term charging state dependency is modelled by an LSTM block, while time-related information and the long-term tendency of charging state profiles are handled by a multilayer feedforward neural network. Let a sample of the input time series data for one-time step t be $X_t = \{X_{1t}, X_{2t}\}$, where $X_{1t} = \{y_{t-1}, y_{t-2}, \ldots, y_{t-m}\}$ and $X_{2t} = \{t, d, w, \boldsymbol{p}\}$ (Table 2). The LSTM block takes X_{1t} as input and generates a vector of hidden states and then combines this with the output of the feedforward neural network layers using X_{2t} as input (Fig. 6). The latter uses three fully connected layers to learn complex features from X_{2t}. The output of the LSTM block and the fully connected layers are concatenated and

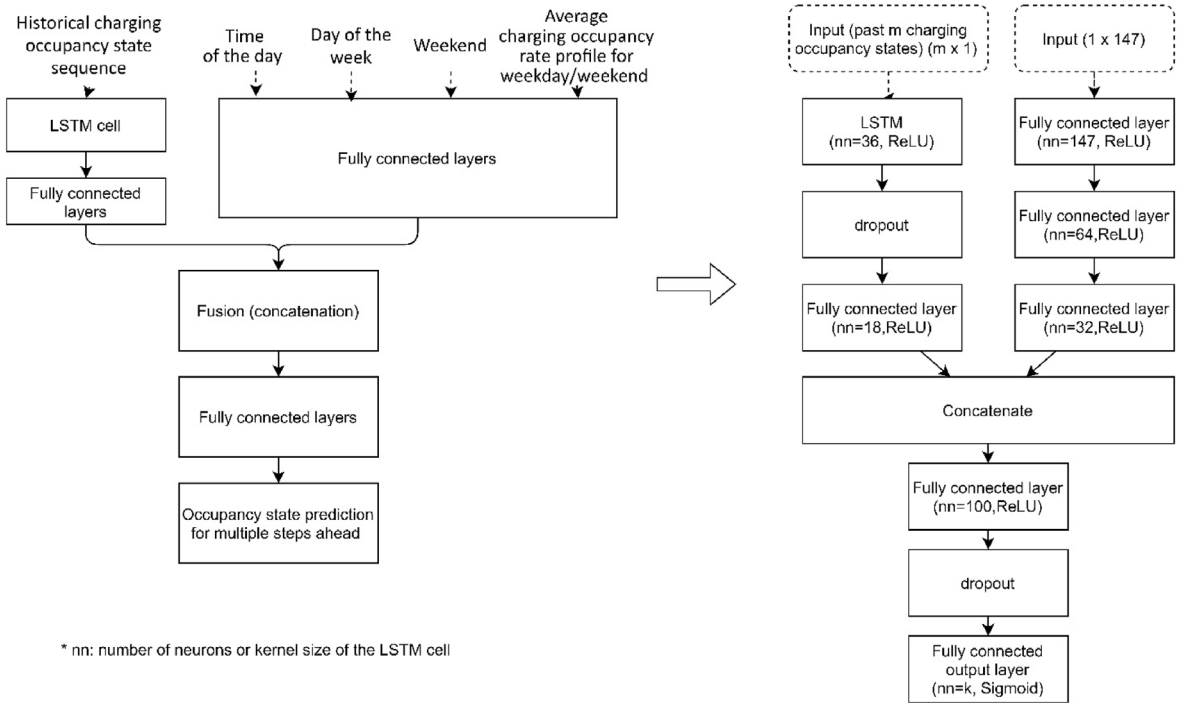


Fig. 6. Proposed hybrid LSTM network architecture.

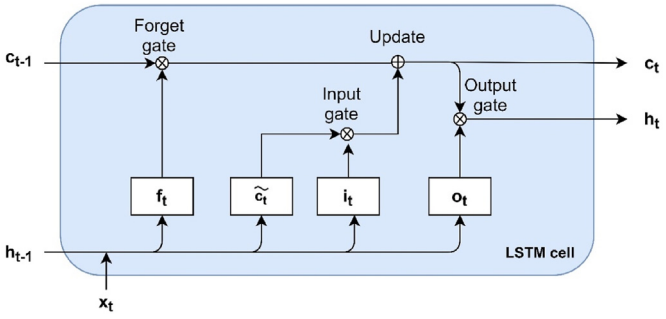


Fig. 7. The LSTM block architecture.

followed by a fully connected layer and an output layer with k neurons, each one corresponding to a charging occupancy state prediction for t, $t+1,\ldots,t+k-1$. Two dropout layers are used for regularization to avoid overfitting. Note that we also test incorporating the spatial correlation of charging occupancies at nearby charging stations (i.e. charging occupancies of nearby chargers) as input features, but the results show that this strategy did not improve the prediction accuracy. The detailed model structure is shown on the right side of Fig. 6.

For each time step, twelve past charging states are used as input for the LSTM block, representing a 2-h charging occupancy sequence. The choice of 12 past occupancy states is based on the experiments with different lengths of past states. Fig. 7 shows the structure of the LSTM block, which contains different cells (rectangles in Fig. 7), an input gate, a forget gate, and an output gate. The equations for the LSTM block are shown in Eqs. (1)–(6) [29]. The input gate (Eq. (1)) combines the input vector x_t ($x_t = X_{1t}$) and its hidden state vector h_{t-1} at $t-1$. The forget gate (Eq. (2)) filters the long-term information to be retained. The output gate (Eq. (3)) uses a sigmoid activation function to determine what information is to be output for the cell state. Equation (4) computes a temporary cell state based on the current input and hidden state at $t-1$. The cell state at t (Eq. (5)) is then updated by its previous state c_{t-1} and its temporal state $\tilde{c}_t$. The output gate (Eq. (6)) controls the final output of the cell at time t based on o_t and c_t.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right), \tag{2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \tag{3}$$

$$\tilde{c}_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{4}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t, \tag{5}$$

$$h_t = o_t \otimes \tanh(c_t), \tag{6}$$

where W_{x*} and W_{h*} are the set of weights connecting to the input vector x_t and the previous hidden state h_{t-1}, respectively. b_i, b_f, b_o, b_c are the corresponding bias terms. $\otimes$ is the element-wise multiplication. i_t is the input gate that decides what information should be retained. f_t Is the forget gate that decides what information is to be removed. o_t is the output gate controlling the information to be moved forward to the output at this time step. c_t is the updated cell-state vector at t. h_t is the output of the LSTM block at time step t. σ and $tanh$ are the sigmoid and hyperbolic tangent activation functions, respectively. The output of the LSTM block is followed by a dropout layer for regularization and then connected by a fully connected layer to further extract the temporal feature.

The weights of the hybrid LSTM model are learned by back-propagation through time to minimize a designed loss function. The activation function used in the fully connected layers is defined as Eq. (7).

Table 3
Hyperparameter settings for the hybrid LSTM model.

Hyperparameter	Value	Hyperparameter	Value
Size of hidden layers/LSTM block	See Fig. 6	Regularization	Dropout (0.2)
Number of hidden layers	See Fig. 6	Mini-batch size	30
Activation function	See Fig. 6	Number of training epochs	15
Learning rate	0.001	Optimizer	Adam

$$h_l = \mathrm{ReLU}(W_l h_{l-1} + b_l), \tag{7}$$

where ReLU is defined as $\mathrm{ReLU}(q) = max(q, 0)$ W_l , h_{l-1}, and b_l are the weight vector, hidden layer vector, and bias term of layer l, respectively. The output of the hybrid LSTM model is a vector of binary values representing the multistep forecasting of the charging occupancy states as in Eq. (8).

$$\hat{y}_t = r(W_{ho} h_o + b_o), \tag{8}$$

where W_{ho} is the hidden-to-output weights. h_o Is the input vector for the output layer o, and b_o is the bias term. r is a sigmoid function defined as $r(Z) = \frac{1}{1+e^{-Z}}$.

For hyperparameter settings, we adopt a manual coarse turning process by sequentially testing a set of hyperparameters based on the proposed hybrid LSTM network architecture [30–32]. This approach first sets up the reference values for the hyperparameters and tune a set of key hyperparameters (i.e. the size of hidden layers, kernel size of the LSTM block, number of hidden layers, dropout rate, mini-batch size, see Table 8) one by one in a sequential way. Schwemmle and Ma [32] showed that such a hyperparameter turning strategy could obtain very accurate results with considerable training-time savings. Applying the state-of-the-art hyperparameter search algorithms [33] to find the best-performing hyperparameters could further achieve around a 2%–5% performance gain, with the cost of high computational time [31]. Table 3 reports the retained final hyperparameters for the hybrid LSTM model. The total number of parameters to be trained for the final model is 45,152. The impact of the different hyperparameter settings is analysed in Section 4.3. The implementation of the hybrid LSTM model is based on Python's Keras Application Programming Interface of the TensorFlow package.

3.2. Benchmark methods and performance metrics

In addition to the proposed hybrid LSTM model, we consider four machine-learning models as benchmarks to compare the performance of our approach. These models include logistic regression [34], SVM [35], random forest [36], and Adaboost [37]. Logistic regression uses a logistic function to model the binary outcomes (classes) based on a set of features. SVM builds a kernel-based hyperplane within a high-dimensional feature space for classification tasks. Random forest builds a number of decision-tree classifiers by sampling and averages the predicted classes from each tree to improve the prediction accuracy and avoid overfitting. Adaboost is an ensemble of learning methods that combine a set of classifiers as a weighted sum to increase prediction accuracy. The weights are adaptively adjusted to increase the accuracy of difficult misclassified cases. These supervised learning methods (classifiers) model the charging occupancy state at one time step as a binary classification problem based on a feature vector as in Eq. (9).

$$\hat{y}_t = f(X_t), \tag{9}$$

where f represents an applied classifier. X_t is the feature vector for time step t. $\hat{y}_t$ is the predicted outcome for time step t. The considered feature space for the benchmark classifiers is the same as in the previous section. As having irrelevant features may reduce the prediction accuracy of classifiers, we test three feature sets as follows.

- Model 1: Consider 6 features as $X_t = \{t, d, w, y_{t-1}, y_{t-2}, y_{t-3}\}$.
- Model 2: Extend Model 1 by using 12 past charging occupancy states, i.e. $X_t = \{t, d, w, y_{t-1}, y_{t-2}, \ldots, y_{t-12}\}$.
- Model 3: Extend Model 2 with the additional features of average charging occupancy rates on the same type of day, i.e. $X_t = \{t, d, w, \boldsymbol{p}, y_{t-1}, y_{t-1}, \ldots, y_{t-12}\}$.

Our goal is to evaluate whether there are significant performance differences when incorporating more elaborated features and determine which feature setting to use. Table 4 shows that there are no significant gains when introducing more complex features for these classifiers. As a result, the feature set of Model 1 is used for the benchmark methods to predict multistep charging occupancy states in the next section. The walk-forward approach is used for multistep time series prediction [38]. This approach uses

Table 5
Performance metrics of the hybrid LSTM model and the benchmark methods.

k-step ahead	1	3	6	12	24	36
Accuracy						
Logistics	0.8837	0.8034	0.7518	0.7166	0.6927	0.6809
SVM	0.7511	0.7421	0.7366	0.7328	0.7295	0.7283
Random forest	0.8370	0.7710	0.7397	0.7240	0.7162	0.7137
Adaboost	0.8816	0.8042	0.7563	0.7257	0.7080	0.6999
Hybrid LSTM	**0.9999**	**0.8926**	**0.8187**	**0.7776**	**0.7593**	**0.7511**
F1 score						
Logistics	0.7760	0.7754	0.6687	0.5423	0.4110	0.3367
SVM	0.5987	0.6805	0.6031	0.5003	0.3918	0.3330
Random forest	0.6860	0.7247	0.6359	0.5452	0.4566	0.4088
Adaboost	0.7681	0.7736	0.6676	0.5425	0.4118	0.3366
Hybrid LSTM	**0.9999**	**0.8562**	**0.7305**	**0.6108**	**0.4896**	**0.4279**

Table 4
Prediction accuracy of the benchmark methods based on the different feature settings.

Classifier	Model 1 (6 features)	Model 2 (15 features)	Model 3 (159 features)
Logistics	0.8837	0.8833	0.8835
SVM	0.7511	0.7525	0.7509
Random forest	0.8370	0.8366	0.8299
Adaboost	0.8816	0.8818	0.8818

Remark: The results are based on the test dataset.

the prediction ($\hat{y}_t$) at time step t as input for the prediction at the next time step $t+1$ and moves forward for the prediction time window. The implementation of the benchmark classifiers is based on Python's Scikit-learn package.

A loss function is defined to measure the difference between predicted and observed outcomes. For our problem, the mean absolute error (MSE) of Eq. (10) is used as an accuracy metric to measure the performance of the predicted charging occupancy sequence on a predefined short-term time window with k time steps ahead.

$$MAE = \frac{1}{k}\sum_{s=t}^{t+k-1}\left|\hat{y}_s - y_s\right|, \tag{10}$$

where y_s is the observed value for time step s and $\hat{y}_s$ is the predicted value for time step s. k is the length of a prediction time window. For the considered problem, the performance of a model is measured as the average accuracy moving through each time step of the test dataset. Note that as certain chargers are unoccupied for most of the time (unbalanced data), the F1 score is calculated as a complementary metric. The F1 score is a weighted measure of precision (the number of correct positive predictions divided by the total number of positive outcomes) and recall (the number of correct positive predictions divided by the sample size). Given a vector of predicted outcomes for a time window, let TP denote the number of true positives (correct prediction for the outcome of 1) and FP the number of false positives. The F1 score is defined as in Eq. (11).

$$\text{F1 score} = \frac{2}{precision^{-1} + recall^{-1}} = \frac{TP}{TP + \frac{FN+FP}{2}}, \tag{11}$$

where TP, FN, and FP are the number of true positives, false negatives, and false positives, respectively. Note that in the case of zero division, the F1 score is set as 0.

4. Results analysis

4.1. Model performance metrics

The multistep prediction results for the hybrid LSTM model and the benchmark methods are shown in Table 5. The length of the prediction time window ranges from 1 (10 min) to 36 (6 h) time steps ahead. The reported results are the average of 10 runs on the test dataset for all rapid charging stations. Fig. 8 shows that the proposed hybrid LSTM model outperforms the benchmark machine learning methods significantly. The 1-step prediction of the hybrid LSTM model is very accurate (0.9999) compared to the benchmark methods (accuracy ranging from 0.7511 to 0.8837). The prediction accuracy decreases as the length of the prediction time window increases. For the prediction of 3- and 6-time steps ahead, the accuracy of the hybrid LSTM model remains satisfactory (0.8926 and 0.8187), outperforming the benchmark methods (0.8042 and 0.7563, respectively). As for the F1 score, its values drop significantly starting from 12-steps-ahead forecasting. We can conclude that the proposed approach is suitable for charging occupancy state prediction for time windows less than 60 min ahead.

Table 6 shows the detailed prediction result for each rapid charger for multiple time steps ahead. The performance of the hybrid LSTM model has a good prediction accuracy of 83.7–97.9% for prediction 3-time steps ahead (30 min). These numbers reduce gradually when the prediction time windows become longer, as shown on the left side of Fig. 9. From the right side of Fig. 9, we can observe that five chargers have an average charging occupancy rate of around 40% or more, while only charger 1 has a low occupancy rate of 6.3%. As an example, Fig. 10 compares the predicted and observed charging occupancy profiles for multiple time steps ahead. The results show that the proposed model predicts well the observed charging profiles over different prediction time horizons.

4.2. Performance comparison with other deep learning approaches

We further compare the performance of the proposed hybrid LSTM model with seven benchmark DL models considering the same feature space. The benchmark DL neural network architectures are shown in Fig. 11.

a. LSTM [29]: Use a classical LSTM network only to connect input sequences of feature data for multistep charging-state prediction.

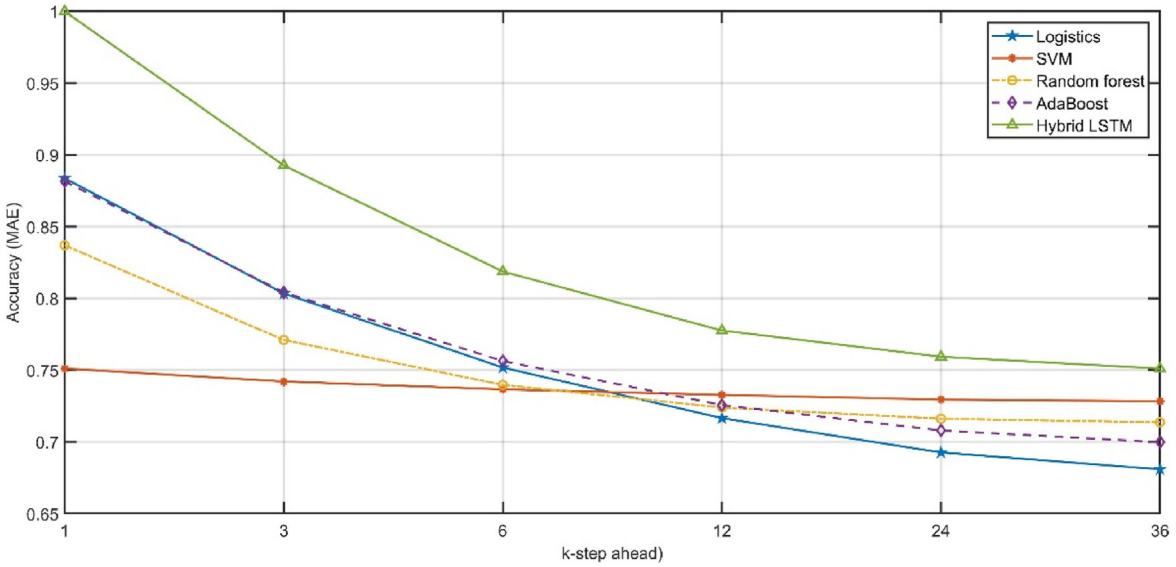


Fig. 8. The prediction accuracy of the hybrid LSTM and the benchmark methods.

Table 6
Prediction accuracy of the hybrid LSTM on the test dataset for different rapid chargers.

# of time steps ahead	Rapid charger ID								
	1	2	3	4	5	6	7	8	9
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.979	0.945	0.922	0.903	0.859	0.837	0.838	0.862	0.887
6	0.966	0.917	0.868	0.814	0.769	0.734	0.747	0.758	0.808
12	0.960	0.905	0.838	0.758	0.706	0.686	0.705	0.709	0.763
24	0.957	0.898	0.821	0.715	0.676	0.661	0.684	0.688	0.741
36	0.956	0.895	0.818	0.699	0.670	0.658	0.687	0.678	0.735

Remark: One-time step is 10 min.

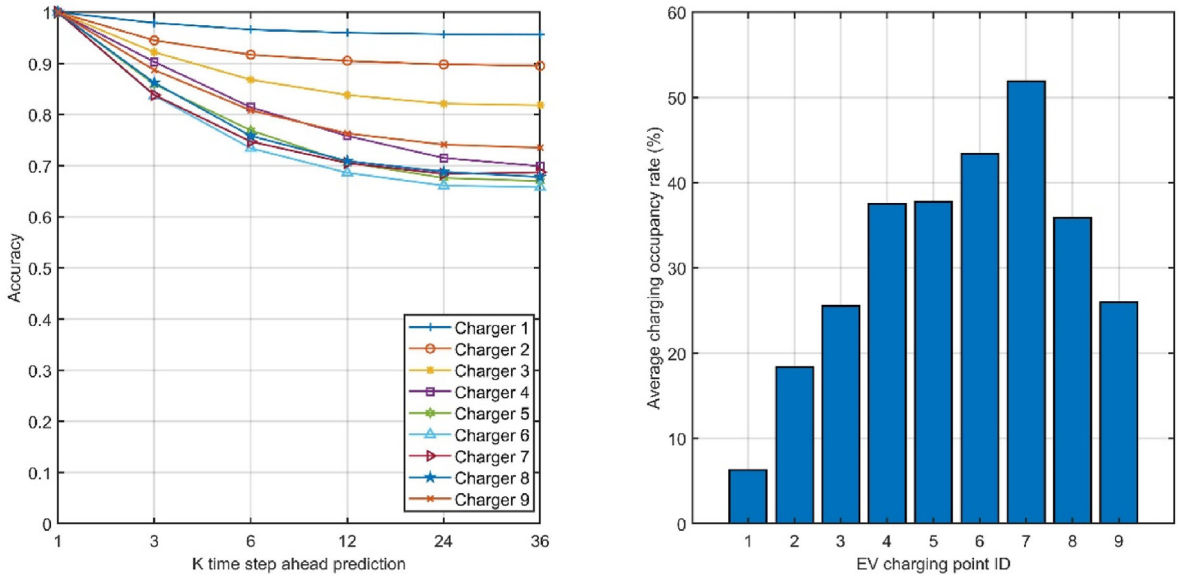


Fig. 9. Prediction accuracy of the charging occupancy states for each charger over multiple time steps (left) and the average charging occupancy rates for each charger (right).

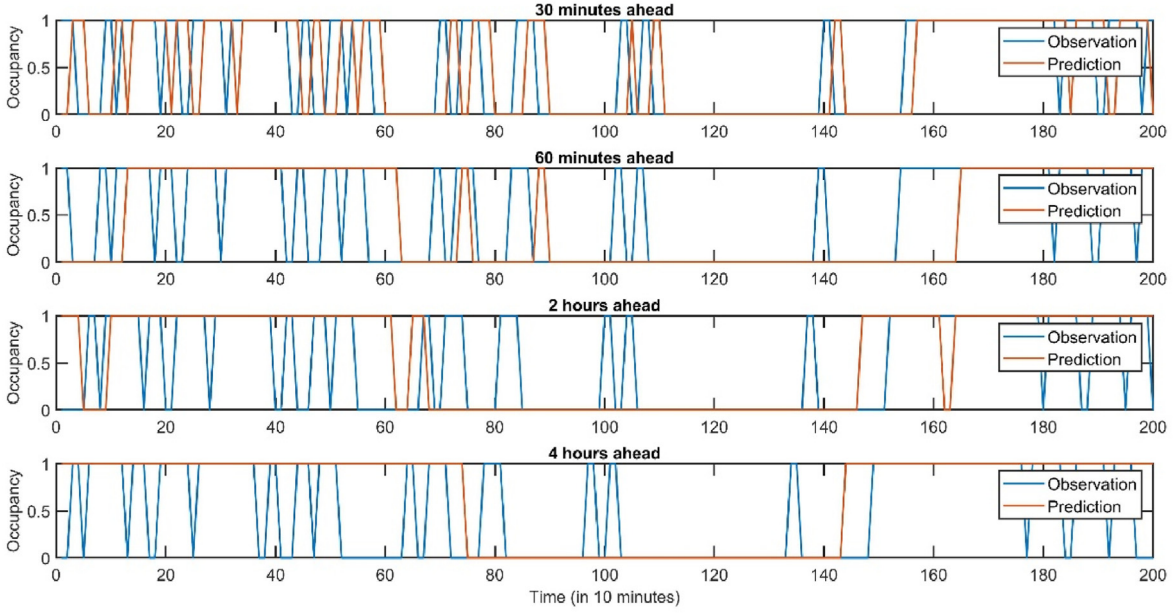


Fig. 10. Example of the observed and predicted charging occupancy profiles using the hybrid LSTM approach for multiple time steps ahead.

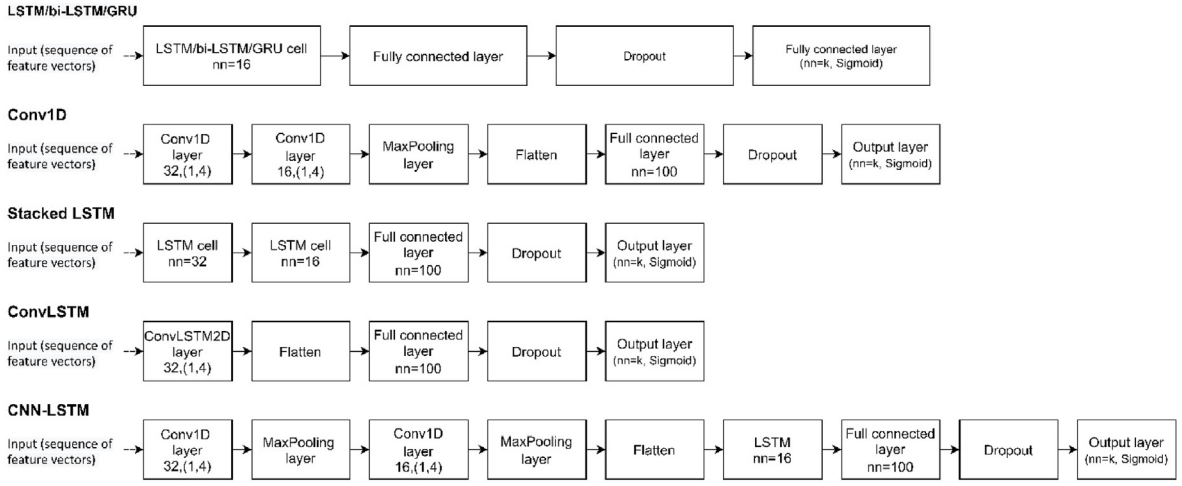


Fig. 11. The benchmark DL neural network architectures (kernel size is in parentheses).

b. Bi-LSTM [39]: Use a bidirectional LSTM network to consider both forward and backward information of input sequences of feature data sequences for multistep prediction.
c. GRU [40]: Use a gated recurrent unit (GRU) network, which is similar to LSTM but with a simpler architecture and fewer parameters to learn, for multistep charging-state prediction. For the above three network structures, the input sequences of features are connected with an LSTM/Bi-LSTM/GRU block, followed by a fully connected layer, a dropout layer, and a fully connected output layer with a sigmoid function.
d. Conv1D [41]: Use two 1-dimensional (1D) convolutional layers, one Max pooling layer, and one fully connected layer for sequential charging station occupancy state predictions. Input features are connected with two sequential 1D convolutional layers (kernel size = 4) to filter information and then followed by a Max pooling layer. The latter is flattened and then connected by a fully connected layer and an output layer.
e. Stacked LSTM [42]: Stack multiple LSTM layers on each other to learn deepened hidden-to-hidden state transitions for more complex pattern recognition. We connect the input features by two consecutive LSTM layers, followed by a fully connected layer, a drop-out layer, and an output layer.
f. ConvLSTM [43,44]: The convolutional LSTM is a combination of conventional networks and LSTM networks for spatiotemporally correlated data predictions by integrating convolutional filters into the LSTM structure. The input feature sequences are connected by a 2D ConvLSTM cell with a one-dimensional kernel size (1, 4) for handling one-dimensional time series data in our case. The output of the ConvLSTM cell is flattened and then connected by a fully connected layer, a dropout layer, and an output layer for multiple time-step predictions.
g. CNN-LSTM [45]: Different from the ConvLSTM, the CNN-LSTM uses multiple CNN layers to filter information and then connect their outputs by an LSTM cell for learning hidden temporal relationships. We connect the input features with multiple CNN layers and multiple Max pooling layers in-between. An LSTM cell is connected after flattening the CNN layers, then connecting to a fully connected layer, a dropout layer, and an output layer.

The feature vectors for these DL models are defined as $X_t = (V_t, V_{t-1}, \ldots, V_{t-m-1})$, with $V_t = \{t, d, w, \boldsymbol{p}, y_{t-1}\}$. To find well-performing feature settings, we first vary the number of historical steps m to find a good input sequence length and then refine the hyperparameters of the model. Finally, $X_t = (V_t, V_{t-1}, V_{t-2})$ (3 historical steps) is retained for the DL variants. The hyperparameters used for these models are shown in Fig. 11, more detailed settings can be found in the computational source codes below. Table 7 compares the prediction accuracies of the hybrid LSTM model and those of the benchmark DL models. The results show that the proposed hybrid LSTM model outperforms significantly the benchmark DL models, in particular for predicting 6 or fewer time steps ahead (see Fig. 12). The characteristics of the hybrid LSTM model for its outperformance can be highlighted as follows. First, time series data present different temporal regularities for which highly irregular ones are difficult to predict (e.g. customer arrival patterns and charging times might be random for certain charging stations). When applying the LSTM (or other time series/machine learning approaches) for highly irregular time series data prediction, it might achieve its maximum predictability by learning the local temporal patterns [46]. To augment the predicting accuracy, the hybrid LSTM incorporates the output of another predictor (expected charging occupancy trends on a longer horizon) using multiple fully connected layers to improve the limit of the predictability of the LSTM cells. Second, the proposed model provides a flexible framework to incorporate additional features as extended neural network branches for extracting information from different features or predictors to augment its predicting accuracy. Future extensions could consider this idea to combine predictions from other learning algorithms for better performance as is the case for ensemble learning approaches.

The data and Python codes used in this study are available at https://github.com/tym2021.

Table 7
The prediction accuracy of the benchmark DL models for multiple steps ahead.

Model	Prediction of k-time steps ahead					
	1	3	6	12	24	36
LSTM	0.8887	0.8107	0.7613	0.7528	0.7363	0.7294
Bi-LSTM	0.8888	0.8149	0.7751	0.7557	0.7351	0.7383
GRU	0.8835	0.8085	0.7681	0.7484	0.7414	0.7398
ConvlD	0.7708	0.7449	0.7388	0.7358	0.7376	0.7367
Stacked LSTM	0.8832	0.8100	0.7701	0.7504	0.7416	0.7393
ConvLSTM	0.8830	0.8102	0.7703	0.7488	0.7454	0.7390
CNN-LSTM	0.8834	0.8082	0.7711	0.7486	0.7450	0.7416
Hybrid LSTM	**0.9999**	**0.8926**	**0.8187**	**0.7776**	**0.7593**	**0.7511**

Remark: The results are based on the average of 10 runs for the test dataset of all rapid chargers.

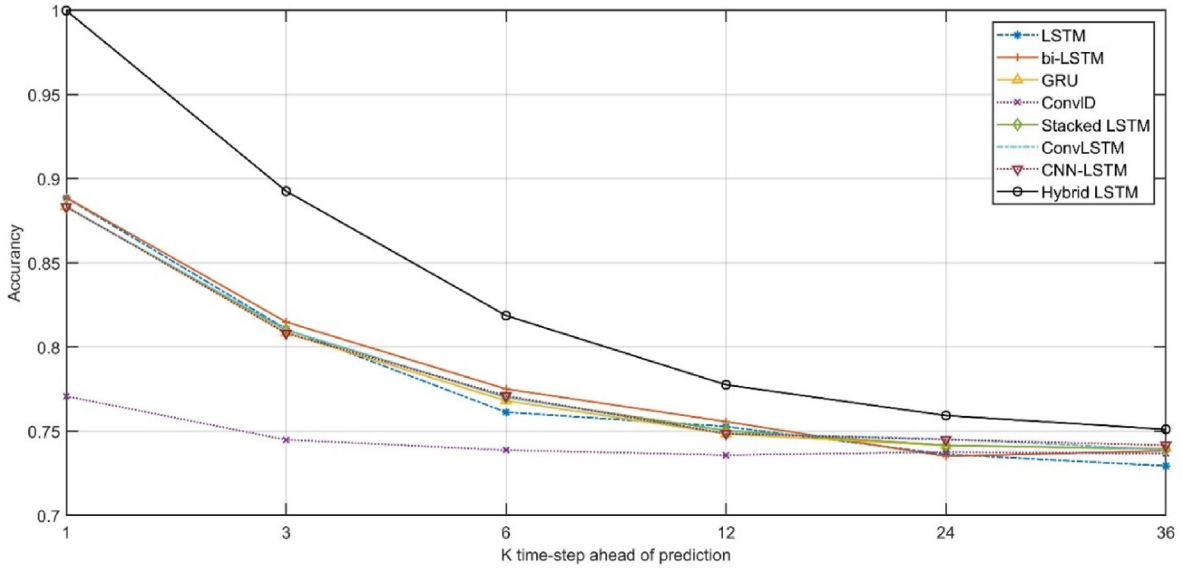


Fig. 12. Comparison of the prediction accuracies of different DL models for multiple time-step ahead predictions.

Table 8
Influence of the hyperparameters of the hybrid LSTM model on prediction accuracy.

Hyperparameter	Value	Accuracy on the test dataset
Learning rate	[0.0005,**0.001**,0.002,0.004,0.008,0.012]	0.8186 **0.8193** 0.8187 0.8178 0.8149 0.8110
Number of fully connected layers*	[1,2,**3**]	0.8156 0.8184 **0.8201**
Kernel size of the LSTM block	[18,**36**,54,72,108,144]	0.8187 **0.8201** 0.8199 0.8183 0.8185 0.8188
Dropout	[0.1, 0.2, 0.3, 0.4, 0.5]	0.8188 **0.8198** 0.8185 0.8188 0.8166 0
Number of training epochs	[10,**15**,20,25,30,40, 50,60,70,80,90,100]	0.8185 **0.8195** 0.8193 0.8191 0.8195 0.8187 0.8179 0.8166 0.8163 0.8156 0.8137 0.8123

Remark: *Top-right branch of the hybrid LSTM network architecture.

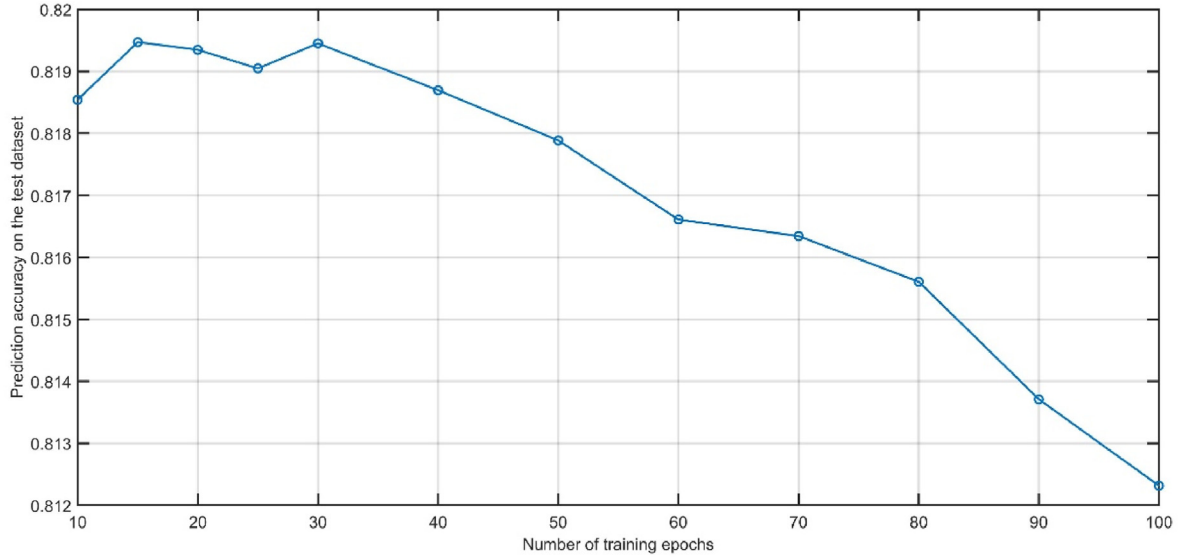


Fig. 13. Influence of the number of training epochs on the prediction accuracy of the test dataset.

4.3. Sensitivity analysis

To further explore the influence of the model parameters, we conduct a series of sensitivity analyses concerning five key model parameters, as shown in Table 8. Each experiment differs by varying the values of a tested hyperparameter while keeping other hyperparameters identical. The reported results are based on the average of 5 runs on the test dataset for all rapid chargers for 6-time-step prediction. Table 8 shows the summary of the accuracy performance for each tested hyperparameter value. The results show that

using different hyperparameter settings could improve marginally the prediction accuracy which is consistent with our previous study [32]. We find that using three fully connected layers allows the enhanced learning of non-linear relationships from the large (i.e.147) feature vector. For the kernel size of the LSTM block, we find it performs better with a size between 18 and 54. The best dropout rate is 0.2, and the best-performing number of training epochs is 15 to avoid overfitting on the test dataset (Fig. 13). For new charging occupancy datasets (e.g. newly available data from existing or new charging stations), one can tune the hyperparameters by the proposed sequential tuning approach using the values (Table 3) found in this study as a reference and test limited candidate values for each parameter around this reference point to adapt new scenarios quickly.

5. Discussion and conclusions

This paper proposes a new approach for predicting the occupancy of EV charging stations. This problem is really important for the management of EV fleets and has hardly been addressed in the scientific literature as a discrete EV charging occupancy-state modelling problem. To do so, we propose a hybrid LSTM neural network that considers both short-term and long-term charging occupancy states to model EV charging occupancy profiles at chargers. An open dataset provided by the city of Dundee, UK, is used as a basis to implement our approach and verify its performance. This method is compared with four other more conventional machine learning methods and three other DL networks. In all cases, the accuracy rate and F1 score show higher performance, both for short-term prediction (10 min: +22% improvement in F1 score over the best competing approach) and long-term prediction (6 h: +2%). Similar findings are obtained when comparing other state-of-the-art deep learning approaches. The computational codes and data are freely available for their potential applications and extensions.

These results show a strong potential for the improvement of charging station occupancy prediction methods, which allows EV-based mobility service operators to develop smart-charging scheduling strategies. Moreover, the proposed methodology could lead to a more advanced recommendation or allocation strategies than what exists today—for example, using multi-objective optimisation approaches to meet various constraints (e.g. which charging station should a user consider given waiting time, potential new arrivals, and the geographic position of that station). Similarly, the practical development of these new strategies would require a high-speed exchange of information and a full, low-latency interconnected network—which could involve distributed network issues or ones specific to the wireless and 5G communication literature. Future work might extend the proposed methodology for other time series forecasting involving continuous variables with heterogeneous (time series and cross-sectional) data. Other possible directions might involve different mixed architectures with additional spatiotemporal features for different applied fields such as EV energy consumption demand and taxi demand arrival pattern forecasting.

Data availability

The data and Python codes used in this study are available at https://github.com/tym2021.

References

[1] Eickhout B. European strategy for low-emission mobility. European Parliament report; 2017. https://www.europarl.europa.eu/doceo/document/A-8-2017-0356_EN.html.
[2] Engel H, Hensley R, Knupfer S, Sahdev S. Charging ahead: electric-vehicle infrastructure demand. McKinsey Center for Future Mobility; 2018.
[3] Sawers P. Google Maps will now show real-time availability of electric vehicle charging stations. 2019. https://venturebeat.com/2019/04/23/google-maps-will-now-show-real-time-availability-of-charging-stations-for-electric-cars/.
[4] Jenn A. Electrifying ride-sharing: transitioning to a cleaner future. UC Davis: National Center for Sustainable Transportation; 2019.
[5] Tian Z, Jung T, Wang Y, Zhang F, Tu L, Xu C, Tian C, Li XY. Real-time charging station recommendation system for electric-vehicle taxis. IEEE Trans Intell Transport Syst 2016;17:3098–109. https://doi.org/10.1109/TITS.2016.2539201.
[6] Yuan Y, Zhang D, Miao F, Chen J, He T, Lin S. P2Charging: proactive partial charging for electric taxi systems. In: Proceedings - international conference on distributed computing systems; 2019. https://doi.org/10.1109/ICDCS.2019.00074.
[7] Ma T-Y, Xie S. Optimal fast charging station locations for electric ridesharing with vehicle-charging station assignment. Transport Res Transport Environ 2021;90:102682.
[8] Ma T-Y. Two-stage battery recharge scheduling and vehicle-charger assignment policy for dynamic electric dial-a-ride services. PLoS One 2021;16(5). e0251582–e0251582.
[9] Eu Science Hub. Electric vehicles: a new model to reduce time wasted at charging points. https://ec.europa.eu/jrc/en/news/electric-vehicles-new-model-reduce-time-wasted-charging-points; 2019.
[10] Bikcora C, Refa N, Verheijen L, Weiland S. Prediction of availability and charging rate at charging stations for electric vehicles. In: 2016 int. Conf. Probabilistic methods appl. To power syst. PMAPS 2016 - proc; 2016. p. 1–6. https://doi.org/10.1109/PMAPS.2016.7764216.
[11] Motz M, Huber J, Weinhardt C. Forecasting BEV charging station occupancy at work places. In: Reussner RH, Koziolek A, Heinrich R, Hrsg, editors. Informatik 2020. Bonn: Gesellschaft für Informatik; 2021. p. 771–81. https://doi.org/10.18420/inf2020_68.
[12] Soldan F, Bionda E, Mauri G, Celaschi S. Short-term forecast of EV charging stations occupancy probability using big data streaming analysis. 2021. arXiv: 2104.12503.
[13] Amara-Ouali Y, Goude Y, Massart P, Poggi JM, Yan H. A review of electric vehicle load open data and models. Energies 2021;14(8):2233.
[14] Flammini MG, Prettico G, Julea A, Fulli G, Mazza A, Chicco G. Statistical characterisation of the real transaction data gathered from electric vehicle charging stations. Elec Power Syst Res 2019;166:136–50. https://doi.org/10.1016/j.epsr.2018.09.022.
[15] Iversen EB, Morales JM, Madsen H. Optimal charging of an electric vehicle using a Markov decision process. Appl Energy 2014;123:1–12. https://doi.org/10.1016/j.apenergy.2014.02.003.
[16] Lee ZJ, Li T, Low SH. ACN-data: analysis and applications of an open EV charging dataset. In: Proceedings of the tenth ACM international conference on future energy systems; 2019. p. 139–49.
[17] Majidpour M, Qiu C, Chu P, Pota HR, Gadh R. Forecasting the EV charging load based on customer profile or station measurement? Appl Energy 2016;163:134–41.
[18] Ma T-Y, Pantelidis T, Chow JY. Optimal queueing-based rebalancing for one-way electric carsharing systems with stochastic demand. In: Paper presented in transportation Research board 98th annual meeting; 2019. https://arxiv.org/abs/2106.02815.
[19] Pantelidis T, Li L, Ma TY, Chow JY, Jabari SE. Node-charge graph-based online carshare rebalancing with capacitated electric charging. 2020. arXiv: 2001.07282.
[20] Gruoss G, Mion A, Gajani GS. Forecasting of Electrical Vehicle impacton infrastructure: Markov chains model of charging stations occupation. eTransportation 2020;6:100083. https://doi.org/10.1016/j.etran.2020.100083.

[21] Verma A, Asadi A, Yang K, Maitra A, Asgeirsson H. Analyzing household charging patterns of Plug-in electric vehicles (PEVs): a data mining approach. Comput Ind Eng 2019;128:964–73.
[22] Van Houdt G, Mosquera C, Nápoles G. A review on the long short-term memory model. Artif Intell Rev 2020;53:5929–55.
[23] Kim TY, Cho SB. Predicting residential energy consumption using CNN-LSTM neural networks. Energy 2019;182:72–81.
[24] Yang Y, Hong W, Li S. Deep ensemble learning based probabilistic load forecasting in smart grids. Energy 2019;189:116324.
[25] Sajjad M, Khan ZA, Ullah A, Hussain T, Ullah W, Lee MY, Baik SW. A novel CNN-GRU-based hybrid approach for short-term residential load forecasting. IEEE Access 2020;8:143759–68.
[26] Ullah FUM, Ullah A, Haq IU, Rho S, Baik SW. Short-term prediction of residential power energy consumption via CNN and multi-layer bi-directional LSTM networks. IEEE Access 2019;8:123369–80.
[27] Wang JQ, Du Y, Wang J. LSTM based long-term energy consumption prediction with periodicity. Energy 2020;197:117197.
[28] Laib O, Khadir MT, Mihaylova L. Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks. Energy 2019;177:530–42.
[29] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[30] Zhang Q, Wang H, Dong J, Zhong G, Sun X. Prediction of sea surface temperature using long short-term memory. Geosci Rem Sens Lett IEEE 2017;14(10):1745–9.
[31] Schwemmle N. Short-term spatio-temporal demand pattern predictions of trip demand. Master Thesis. Katholieke Universiteit Leuven; 2021. https://zenodo.org/record/4514435#.YRZTNYgzbIU.
[32] Schwemmle N, Ma T-Y. Hyperparameter optimization for neural network based taxi demand prediction. In: Proceedings of the BIVEC-GIBET transport Research days 2021; 2021.
[33] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Proceedings of the 25th annual conference on neural information processing systems; 2011. p. 2546–54.
[34] Linoff GS, Berry MJA. Data mining techniques for marketing, sales and customer support. Wiley; 2011.
[35] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14(3):199–222.
[36] Ho T-K. Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1; 1995. p. 278–82. https://doi.org/10.1109/ICDAR.1995.598994. ISBN 978-0-8186-7128-9.
[37] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.
[38] Brownlee J. How to backtest machine learning models for time series forecasting. 2016. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/.
[39] Kiperwasser E, Goldberg Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations. Trans Assoc Comput Linguis 2016;4:313–27.
[40] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. arXiv preprint arXiv:1406.1078.
[41] Barkost PH. Detecting EV charging from hourly smart meter data. Master's thesis. UiT Norges arktiske universitet; 2020.
[42] Pascanu R, Gulcehre C, Cho K, Bengio Y. How to construct deep recurrent neural networks. 2013. arXiv preprint arXiv:1312.6026.
[43] Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-C. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of the 28th international conference on neural information processing systems, vol. 1; 2015. p. 802–10.
[44] Petersen NC, Rodrigues F, Pereira FC. Multi-output bus travel time prediction with convolutional LSTM neural network. Expert Syst Appl 2019;120:426–35.
[45] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 2625–34.
[46] Zhao K, Khryashchev D, Vo H. Predicting taxi and uber demand in cities: approaching the limit of predictability. IEEE Trans Knowl Data Eng 2019;33(6):2723–36.

MPPT Based on Adaptive Neuro-Fuzzy Inference System (ANFIS) for a Photovoltaic System Under Unstable Environmental Conditions

Debasish Mishra, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, debasishmishra1@gmail.com*

Subhendu Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srikant.p@yahoo.co.in*

Rajib Lochan Barik, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, rajib_barik543@gmail.com*

Anil Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, anil_sahoo342@gmail.com*

Abstract: Many algorithms have been used to track the MPP in a PV generator. Although these algorithms have proved their worth, the fact remains that they still have limits in terms of stability, response times and significant presence of oscillations, especially for sub-Saharan conditions where the climate variation is very sudden and has a considerable impact on the power delivered at the generator output. In this article, the objective is to develop a maximum power point tracking (MPPT) controller based on an Adaptive Neuro-Fuzzy Inference System (ANFIS) to improve the performance of the Felicity Solar photovoltaic module FL-M-160W submitted to varying environmental conditions. The specifications of the FL-M-160W module are used to analyze and model the PV generator and boost converter located between the panel and the load in Matlab / Simulink. After the experimental tests, a database was set up to develop the neurofuzzy controller. The proposed ANFIS model was tested and validated under the Matlab / Simulink environment and then inserted into the PV system. The optimum voltage Vopt provided by this model is compared to the reference voltage Vpv provided by the PV generator and the error obtained is used to adjust the duty cycle of the DC-DC boost converter. After simulations, the results obtained reveal a good performance of the ANFIS controller compared to conventional P&O, InC and HC controllers in terms of stability, convergence speed, accuracy, robustness, and response time even under unstable environmental conditions with an efficiency of about 98%.

Keywords: Photovoltaic System, Modeling, MPPT Controller, ANFIS, Converter, Unstable Environmental Conditions

1. Introduction

The sun is an inexhaustible source of renewable energy, generating little or no waste or polluting emissions. It is used to produce energy in two forms (thermal and photovoltaic). Photovoltaic solar energy is the transformation of part of the light from solar irradiation into electrical energy using a set of elements constituting a PV system whose basic phenomenon implemented is the photovoltaic effect. This form of energy has the advantage of stabilizing global warming, preserving our fossil fuel reserves and ensuring energy security for the planet. Compared to conventional energy resources, photovoltaic solar energy systems still presents a large area of competition due to its high installation cost and low power consumption due to the conversion efficiency of PV cells. In addition, during the operation of the PV generator, the P-V and I-V characteristics vary. Indeed, the maximum power point (MPP) changes position with the change of light intensity or temperature, and

therefore the optimal voltage changes value. An adaptation stage is generally introduced to operate the PV generator optimally to ensure its profitability. The voltage at the panel terminals must be continuously regulated to its optimum value to work at maximum power: this is the Maximum Power Point Tracking (MPPT). In a PV system, the MPPT command can be defined as an algorithm which associated with an adaptation stage allows the system to operate in its optimal operating point and this whatever the atmospheric conditions (temperature and global sunshine) and of load value [1]. Many algorithms have been used to track the MPP in a PV generator. Although these algorithms have proved their worth, the fact remains that they still have limits in terms of stability, response times and significant presence of oscillations especially for sub-Saharan conditions where the climate variation is very sudden and has a considerable impact on the efficiency of the solar generator. Unlike standard methods where the stability of the system cannot be ensured due to the fluctuations around the MPP that they cause, higher precision of systems, especially nonlinear systems, can be ensured with artificial intelligence methods. Fuzzy logic controller (FLC) and artificial neural networks (ANN) have been used successfully to track the peak power point of PV systems [2, 3]. Fuzzy controllers are fast converging and have minimal oscillations around the MPP but their effectiveness is highly dependent on the skills of the designer. On the other hand, neural networks allow to follow the MPP with precision [4, 5]. Nevertheless, the complexity of implementing this technique remains. To solve this problem, many MPPT controllers combining fuzzy logic and neural networks have been developed to establish a compromise between complexity and precision in the implementation of MPPT. In this work, the ANFIS controller is used to extract the maximum power in a Felicity Solar photovoltaic module FL-M-160W. This document is divided into 5 sections. After introducing this work, section 2 presents the PV system consisting of a panel and the adaptation stage. A review on the MPPT commands used for the maximum power point research is presented in section 3 and section 4 develops and models the neuro-fuzzy MPPT controller uses. Finally, after modeling and simulation, results and discussion are presented in section 5.

2. Modeling of the Photovoltaic System

A photovoltaic system is a set of elements that are used to produce solar energy [6]. Figure 1 illustrates the overall block diagram of the proposed system. The proposed model is a standalone PV system that includes a PV array use as a power generation source. This PV array is connected to the DC-DC boost converter that use ANFIS algorithm as MPP tracking technique to ensure the adaptation between the panel output voltage and the load.

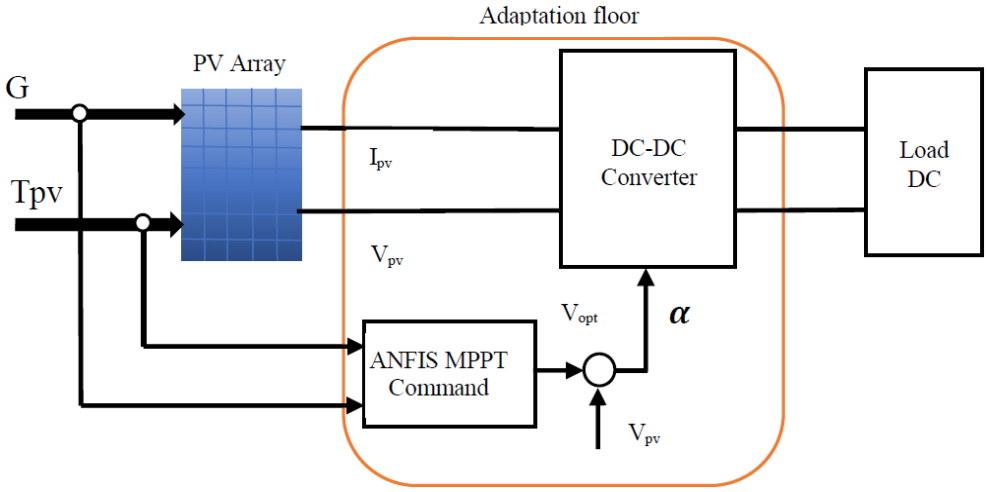


Figure 1. *Overall block diagram of the PV system proposed.*

2.1. Electrical Modeling of the Solar Panel

A photovoltaic (PV) cell can be represented by the equivalent circuit shown in figure 2. In the case of an ideal solar cell, the equivalent electrical circuit consists of a current source Iph, generated by light in parallel with a single-diode. But in practice, no solar cell is ideal. Therefore, a shunt and series resistance are added to the model in order to take into account all the phenomena present during the conversion of light energy. In practice, the maximum current is delivered to the load when the serie resistance *Rs* is very small and the shunt resistance *Rsh* is very large.

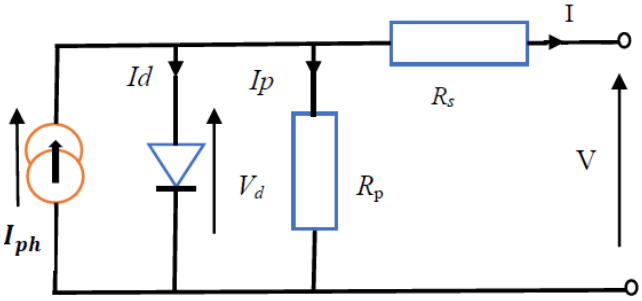


Figure 2. *Equivalent diagram of a PV cell with a diode.*

The power supplied to the output of a solar cell is generally very low. To increase the output power of solar PV systems, solar cells are connected in series and parallel configurations to form PV modules whose equivalent model is discribed in figure 3.

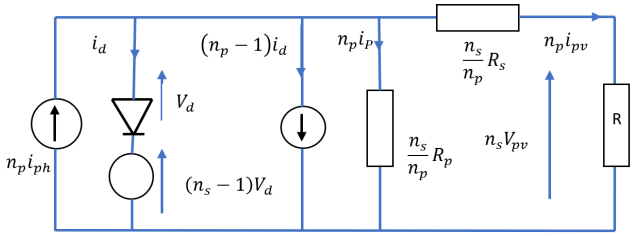


Figure 3. *Equivalent circuit of a PV module.*

Using the theory of semiconductors and photovoltaic, the non-linear relationship between current voltage of the PV module can be described mathematically using basic equations 1, 2, 3, 4, 5 and 6.

Saturation current:

$$I_S = I_{rs} * \left(\frac{T}{T_n}\right)^3 * exp\left[\frac{qE_{g0}}{kn}\left(\frac{1}{T_n}-\frac{1}{T}\right)\right] \quad (1)$$

Reverse saturation current:

$$I_{rs} = \frac{I_{sc}}{exp\left(\frac{V_{OC}}{nV_T}\right)-1} \quad (2)$$

Photocurrent:

$$I_{ph} = \frac{G}{G_n}(I_{sc} + K_i.\Delta T) \quad (3)$$

with $\Delta T = T - T_n = T - 298$.

Current through the shunt resistance:

$$I_{sh} = \frac{V_d}{R_{sh}} = \frac{V+R_S.I}{R_{sh}} \quad (4)$$

Output Current of a solar cell:

$$\mathrm{I} = I_{ph} - I_S\left[exp\left(\frac{(V+IR_S)}{nV_T}\right)-1\right] - \left[\frac{V+IR_S}{R_p}\right] \quad (5)$$

$$V_T = \frac{kT}{q} \quad (6)$$

The output current of the considered PV module is given by:

$$I_{PV} = n_P I_{ph} - n_P I_s\left[exp\left(q.\frac{V+R_S I}{n_s kTn}\right)-1\right] - \left[\frac{V+R_S I}{R_{sh}}\right] \quad (7)$$

where R_s is the series resistance of the solar cell and V is the output voltage of the cell and n_P is the number of cells in parallel.

Figure 4 represents the block diagram of the photovoltaic generator designed in Matlab/Simulink using previous equations.

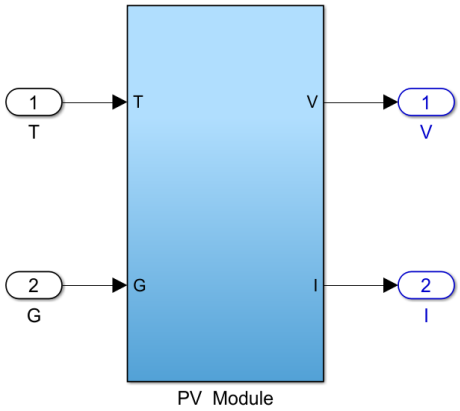


Figure 4. *Block diagram of the photovoltaic generator.*

At the Standard Test Condition (STC, AM = 1.5, G = 1KW / m² and Tc = 25°C), the characteristics of the FL-M-160W module chosen for modeling and simulation are shown in table 1.

Table 1. *Characteristics of the FL-M-160W solar module.*

Parameters	Values
Production tolerance	+/-3%
Maximum power (Pmpp)	160W
Maximum power voltage (Vmpp)	18.20V
Maximum power current (Impp)	8.80A
Short-circuit current (Isc)	9.33A
Open circuit voltage (Voc)	21.84V
Number of cells in series (Ns)	36
Number of cells in parallel (Np)	1

2.2. Analysis of PV Module Characteristics

The Characteristics of the solar module FL-M-160W are shown in figures 5, 6 and 7.

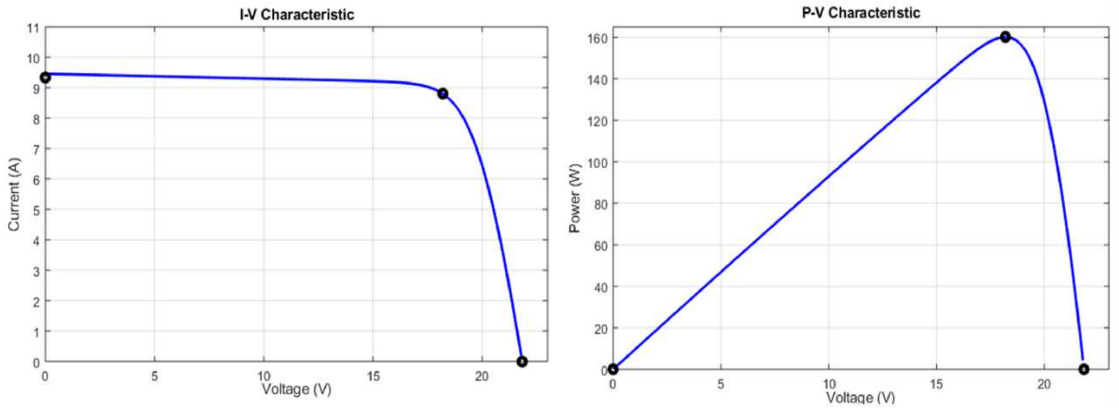


Figure 5. *I (V) and P (V) characteristics of the photovoltaic module.*

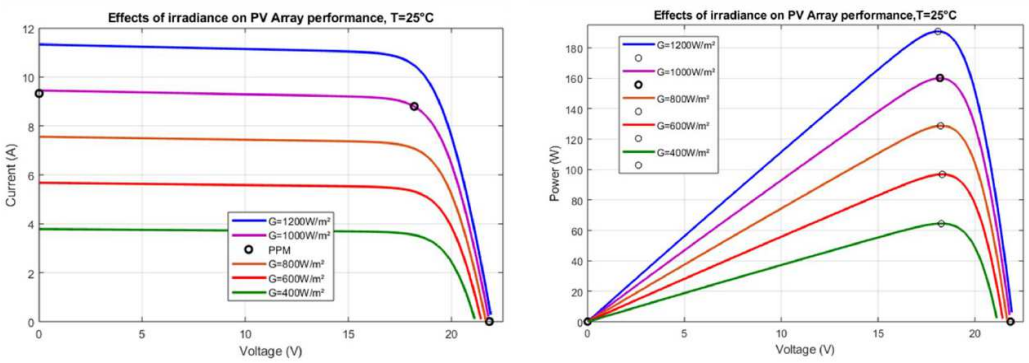


Figure 6. *Irradiation variation on I (V) and P (V) characteristics.*

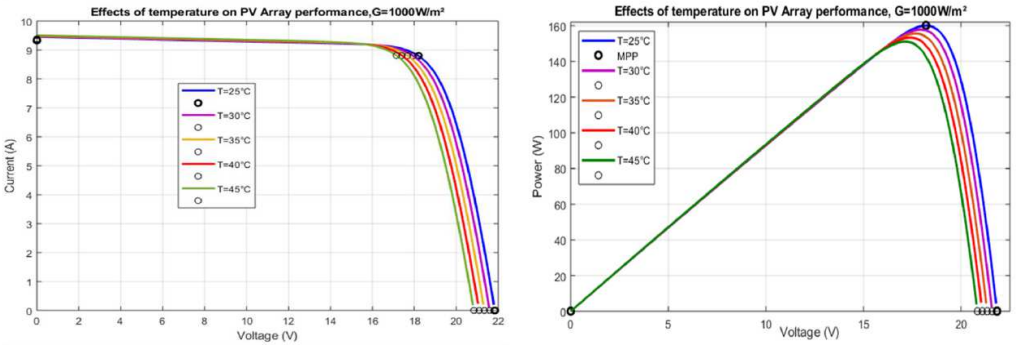


Figure 7. *Temperature variation on I (V) and P (V) characteristics.*

Figure 5 illustrates the non-linearity of the output power characteristic of a photovoltaic panel as a function of the voltage at its terminals. There is a point of the curve where the power is maximum called maximum power point (MPP). As we can see in figure 6 and 7, the I (V) and P (V) characteristics change with irradiance and temperature. A decrease of irradiation G causes a decrease in the current followed by a very slight decrease in the voltage Voc and therefore a shift of the maximum power (Pmax) of the solar panel towards lower powers. When the temperature increases, the open circuit voltage Voc considerably decreases while the current is almost unchanged. This has the immediate consequence of reducing the maximum power.

To take advantage of the maximum energy conversion, it is necessary to operate the PV panel around this MPP. To work at maximum power point, it is necessary to continuously adjust the voltage across the panel to its optimum value, this is called Maximum Power Point Tracking (MPPT).

2.3. DC-DC Boost Converter

The boost regulator is strongly recommended to follow the MPP because of its advantages over the buck converter [7]. This switching power supply enables a higher value variable DC voltage source to be fabricated from a fixed input DC voltage source. The principle is to change the duty cycle of a rectangular signal to create a variable average voltage called Pulse Width Modulation (PWM). Figure 8 illustrates the boost converter model produced on Simulink, the specifications of which are contained in table 2. The output voltage Vs is expressed by equation 8:

$$V_s = \frac{V_e}{(1-\alpha)} \tag{8}$$

With α the duty cycle such that $0 < \alpha < 1$.

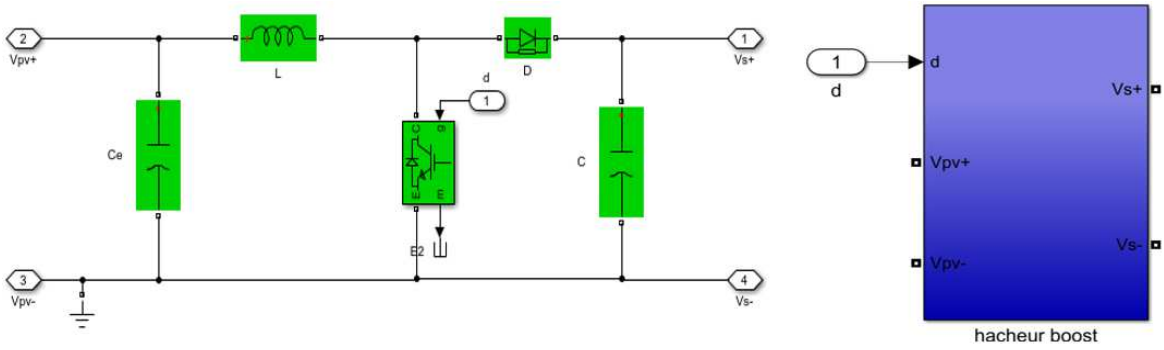


Figure 8. *Simulink model of the boost converter.*

Table 2. *DC-DC boost converter parameters.*

Parameters	Values
Switching frequency (khz)	20
Inductance L (μH)	10694.10^{e}-8
Input voltage (V)	18.20
Capacitors C and Ce (μF)	31575.10^{e}-8

3. Maximum Power Point Tracking

The direct connection of a PV source to a DC load poses the problem of transferring the maximum power when the irradiance changes suddenly [8]. To maximize the power produced by the PV generator, an impedance adapter stage is very often used and controlled by one or more control laws commonly referred in the literature as "Maximum Power Point Tracking". A MPPT command is a command associated with an adaptation stage which makes it possible to monitor the maximum power point of a photovoltaic module by making the PV module operate in its optimum operating point, whatever the atmospheric conditions (temperature and sunlight) and the value of the load [9].

3.1. Perturb and Observe (P&O) MPPT Command

The MPPT P&O algorithms are widely used to track MPP in PV systems. It allows to determine the point of maximum power for a solar irradiation and a temperature or for a level of degradation of PV system characteristics given. Figure 9 gives the flowchart of this algorithm [10].

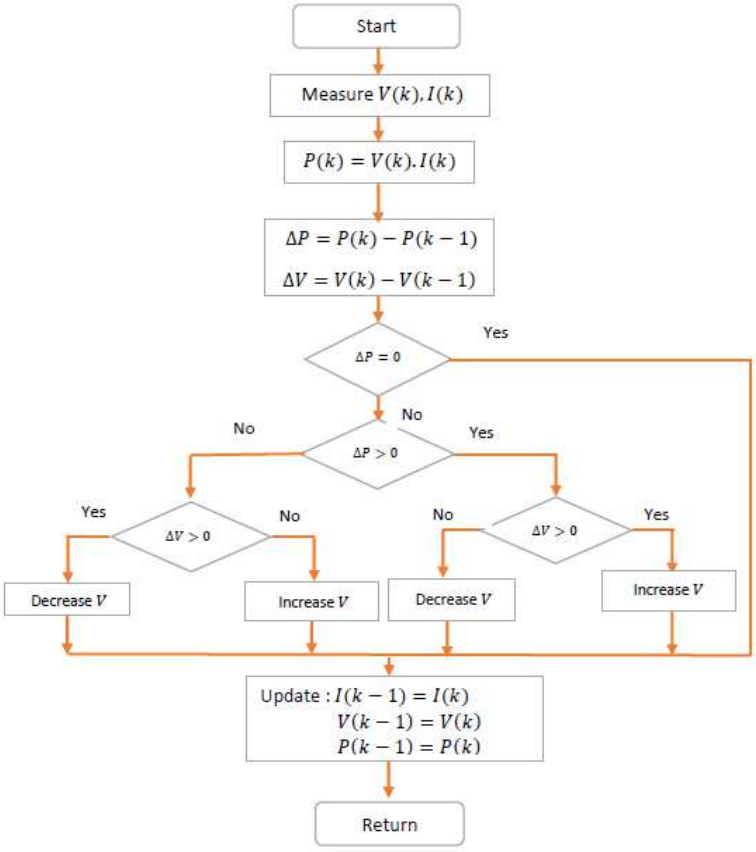


Figure 9. *Perturb and Observe (P&O) algorithm [10].*

The P&O algorithm is a classical algorithm widely used for its simplicity and ease of implementation, its precision, and its speed of reaction [11]. However, in the case of sudden variations in temperature and / or illumination (clouds), the poor convergence of the algorithm is noted [12]. This algorithm also presents some problems related to the oscillations around the MPP that it generates in steady state because the MPP search procedure must be repeated periodically, forcing the system to constantly oscillate around the MPP, once the latter is reached. These oscillations can be minimized by reducing the value of the disturbance variable. However, a low increment value slows down the search for MPP, so you have to find a compromise between precision and speed when choosing this update step that makes this command difficult to optimize.

3.2. Incremental Conductance (InC) MPPT Command

It is a widely used and easy to implement method. This technique addresses the problem of the P&O divergence in the case of a rapid change in sunlight. To calculate the MPP, the algorithm compares the conductance G with the incremental conductance ΔG, and this by looking for the point of cancellation of the derivative of the power. A schematic description of this algorithm is shown in figure 10 [8, 13].

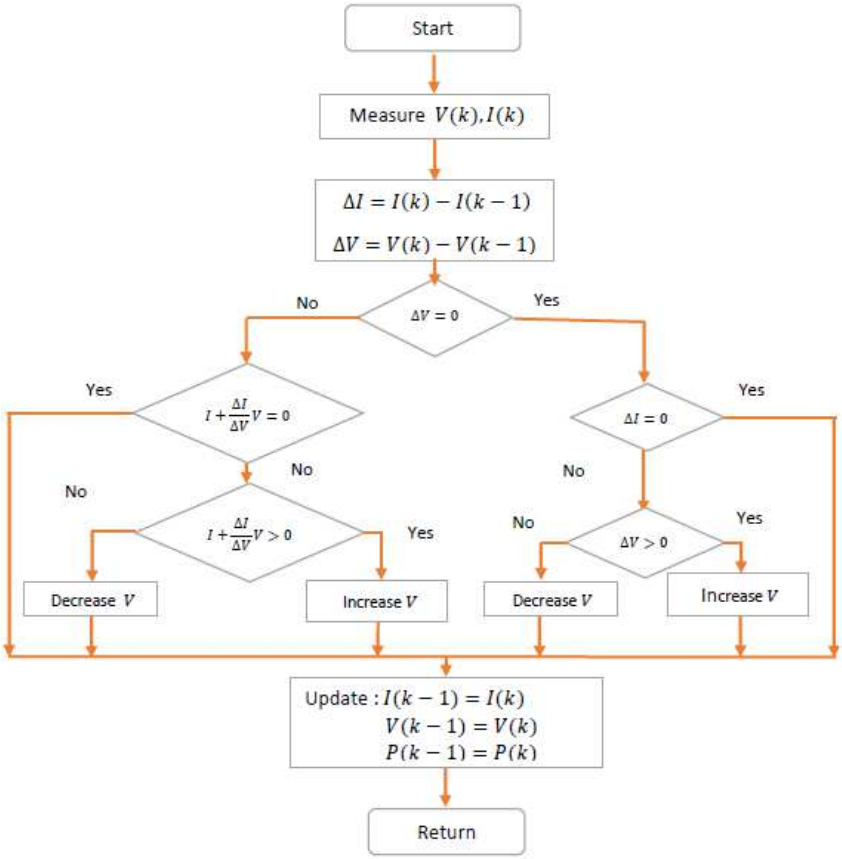


Figure 10. *Incremental Conductance (InC) algorithm [13].*

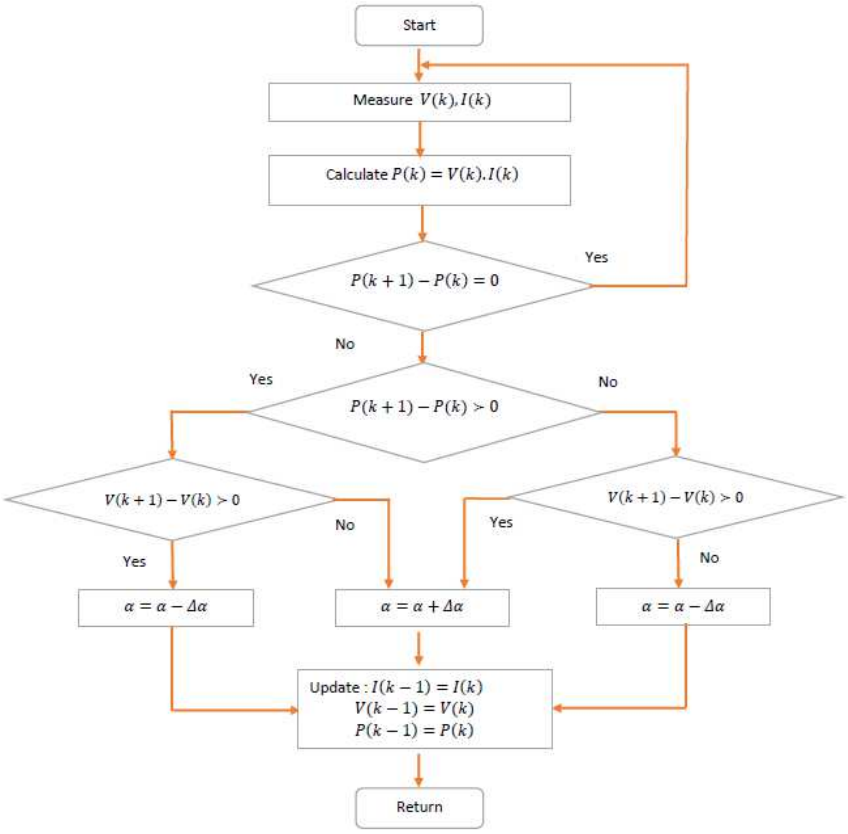


Figure 11. *Hill Climbing (HC) algorithm [14].*

The accuracy and speed with which the algorithm tracks the MPP depends on the size of the reference voltage increment or the duty cycle reference. Two main handicaps are reconciled with this method. The first is the oscillation of the operating point around the steady state MPP, the second is that the algorithm can easily lose track of the MPP if the solar radiation changes rapidly. When the irradiation varies instantly over time, the monitoring of the MPP evolves correctly. But, if the irradiation changes at a slope, the tracking will be poor. The algorithm is unable to determine whether the change in power is due to the voltage disturbance or the change in solar radiation. This principle is illustrated in Figure 9. Several authors, to verify the performance of this proposed method, choose an irradiation profile of different shapes for the simulations.

3.3. Hill Climbing (HC)MPPT Command

Hill Climbing (HC) method consists of making the operating point rise along a characteristic to reach the maximum of the power function of the GPV against the duty cycle of the converter. This is to give a disturbance on the duty cycle which results in a displacement of the operating point along the power-duty cycle characteristic of the photovoltaic generator. The Hill Climbing algorithm is developed in figure 11 diagram [14].

With an efficiency between 95.5% -99.1%, this technique is easy to implement. In addition, its main limitations are oscillations around the MPP in steady state and an occasional loss of the search for MPP during rapid change in weather conditions.

To remedy the various problems associated with the various classical algorithms, artificial intelligence techniques such as fuzzy logic and neural networks have been introduced.

3.4. Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS is an adaptive neuro-fuzzy inference system that uses a 5-layer MLP neural network to refine the fuzzy rules already established by human experts and readjust the overlap between the different fuzzy subsets to describe the input-output behavior of a complex system using a database for learning. It combines the advantages of two machine learning techniquesartificial intelligence techniques (Fuzzy Logic and Neural Network). In this system, fuzzy logic transforms input data into a desired output via a highly interconnected neural network, weighted to map the digital inputs to an output. As shown in figure 12, this network which integrates both the Takagi-Sugeno Kang fuzzy inference system (FIS) and an artificial neural network has a structure composed of five layers representing the network architecture artificial neural. Square nodes represent adaptive parts while circular nodes represent non-adaptive sections. The parameters of the adaptive nodes will be modified during the ANFIS learning process [15]. The learning is done by a hybrid technique based on the principle of backpropagation [16] and the least squares method. The role of learning is the adjustment of the parameters of this fuzzy inference system. This model provides very good approximation results for nonlinear functions.

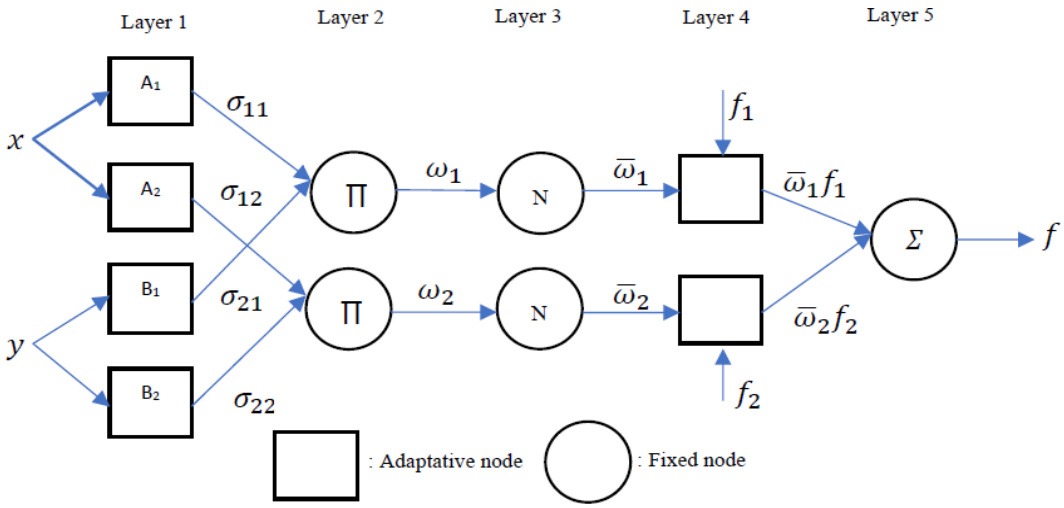


Figure 12. *Two-entry ANFIS architecture for two rules [17, 18].*

The functions of nodes and layers are (9) and (10):
Rule 1: If x is A_1 and y is B_1 then,

$$f_1(x,y) = P_1\,x + q_1\,y + r_1 \tag{9}$$

Rule 2: If x is A_2 and y is B_2 then,

$$f_2(x,y) = P_2\,x + q_2\,y + r_2 \tag{10}$$

Where x and y are the inputs, and A_1, A_2, B_1 and B_2 are the fuzzy sets that represent linguistic values such as small, medium, large. These fuzzy sets would be determined during the learning process. P_1, q_1, r_1, P_2, q_2, r_2 are design parameters also determined during the learning process.

The ANFIS structure consists of five layers:

Layer 1: The first layer represents the fuzzy membership functions. Each node of this layer has a Gaussian-type

membership function according to Jang's model (6):

$$\begin{cases} o_i^1 = \mu_{Ai}(x), i = 1,2 \\ o_i^1 = \mu_{Bi}(y), i = 1,2 \end{cases} \quad (11)$$

o_i^1 are the fuzzy membership levels used to specify the membership degree of net entries x and y in terms of linguistic valuesA_i and B_i.

Layer 2: Each neuron i of this layer is a circular fixed node with the label π which generates as output the product of its inputs (fuzzy rules) and generates the product w_i (12).

$$o_i^2 = w_i = \mu_{Ai}(x).\mu_{Bi}(y), i = 1,2 \quad (12)$$

Layer 3: Each neuron makes it possible to calculate the ratio between the weight of the rule and the sum of the weights of all the rules (13).

$$o_i^3 = \overline{w_i} = \frac{w_i}{w_1+w_2} \quad (13)$$

The obtained value represents the contribution of the fuzzy rule to the result.

Layer 4: Each neuron i of this layer is connected to a corresponding normalization neuron and to the initial inputs of the network. This layer calculates the coefficients of the first order equation of a Takagi-Sugeno rule for each fuzzy rule (14).

$$o_i^4 = \overline{w_i} f_i = \overline{w_i}(P_i\, x + q_i\, y + r_i), i = 1,2 \quad (14)$$

Where $\overline{w_i}$ is the output of layer 3, and $\{P_i, q_i, r_i\}$ is the set of output parameters of the first order rule i, which are called consequent parameters.

Layer 5: Includes a single neuron that represents the output layer that provides the output of ANFIS by calculating the overall weighted output of the system (10).

$$o_i^5 = f = \sum_1 \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (15)$$

Subsequently, a hybrid learning algorithm that combines the backpropagation learning algorithm and the least squares method makes is used to define the optimal values of the parameters of these membership functions and the consecutive parameters. These consecutives parameters are used to determine the ANFIS network output. An experimental database collected on a Felicity Solar PV module FL-M-160W is used to learn, test and check the neuro-fuzzy controller.

4. Modeling of ANFIS MPPT Controller

4.1. Experimental Features

In this work, ANFIS is used to track the maximum power point in a PV system under unstable environmental condition. The optimum voltage obtained at the outlet of the ANFIS network is used to build the duty cycle and allow the PV panel to deliver optimum power output. Figure 13 illustrates the data acquisition device. A database is built using a Benning Sun 2 type pyranometer and an acquisition card which serve as equipment for the acquisition of data through experimental tests which will be optimized to approximate the output which corresponds to the maximum power, depending on variation of irradiance and temperature.

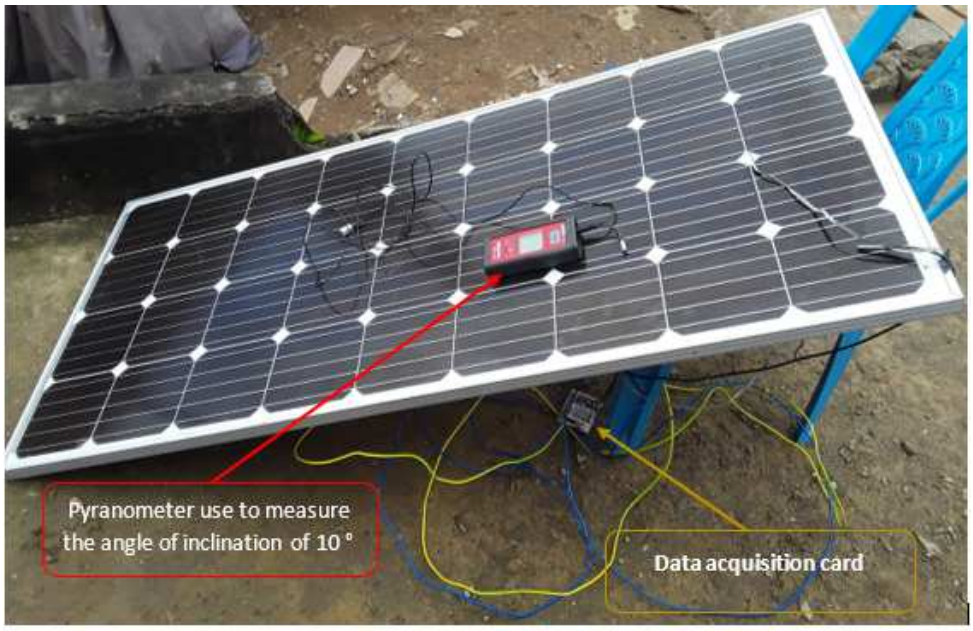


Figure 13. *Experimental data acquisition device.*

The experimental tests carried out for the day of November 04, 2020 made it possible to build a database including test, verification and learning data. Figure 14 illustrates the global irradiance G and the solar panel temperature TPV while figures 15 and 16 illustrate the evolution of current and voltage recorded in relation to the global irradiance.

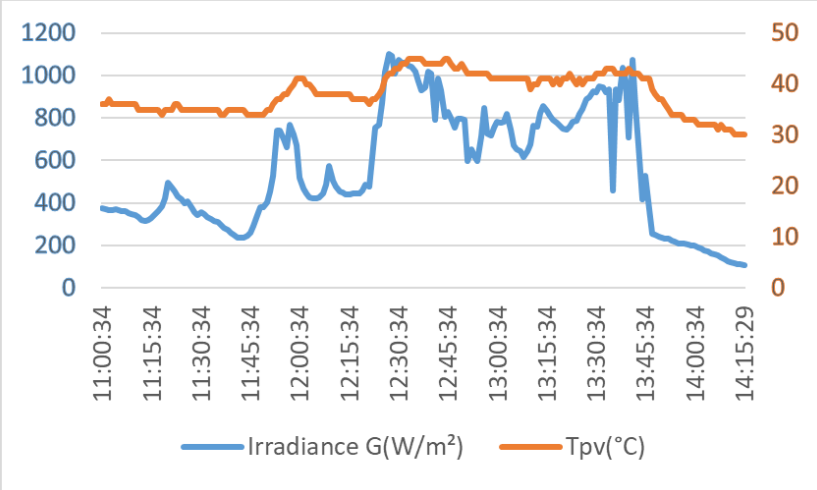


Figure 14. *Weather data (irradiance and temperature) recorded.*

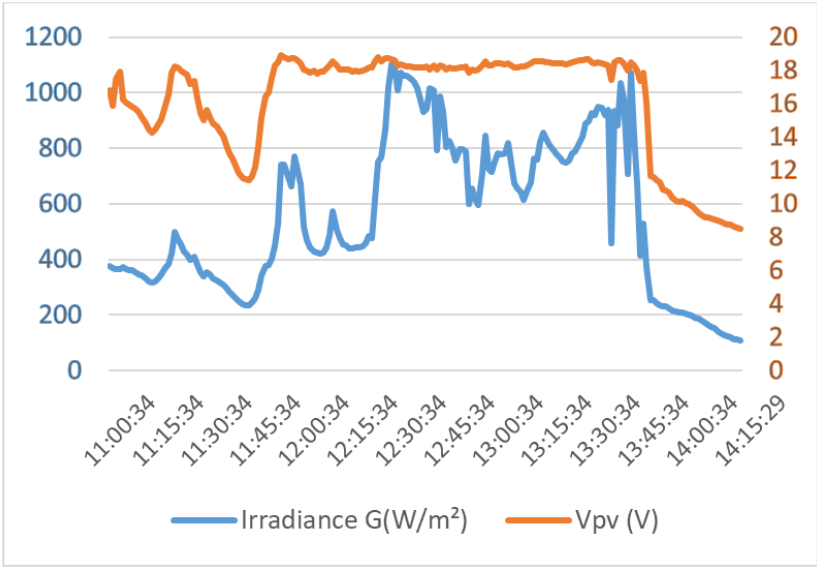


Figure 15. *Evolution of the Vpv voltage*

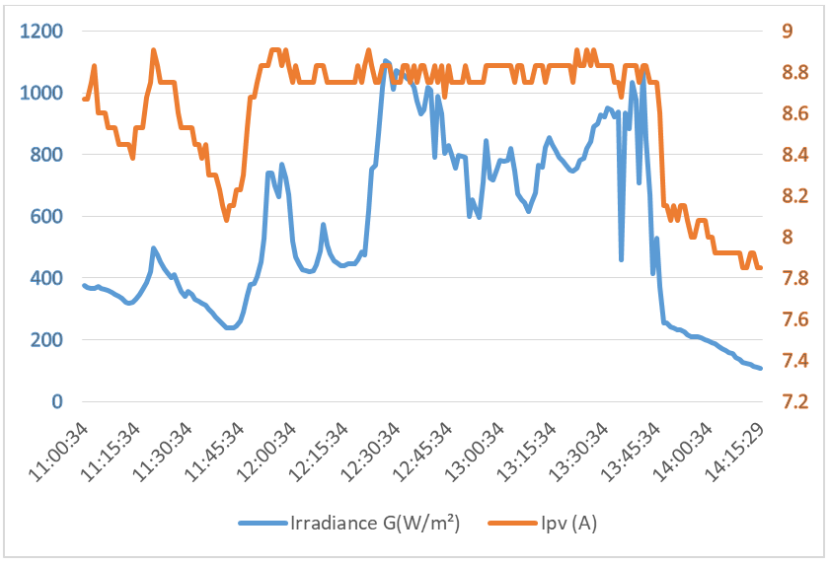


Figure 16. *Evolution of the Ipv current.*

As can be seen in figure 14, the weather conditions fluctuate sharply throughout the day. The solar irradiance fluctuates between 114 W / m² and 1072 W / m². The irradiance peak value of 1072 W / m² is obtained under an ambient temperature of 41°C and corresponded for a temperature of 37°C on the panel surface. The temperature fluctuates daily between 30°C and 41°C. Figure 15 illustrates the voltage evolution recorded in relation to solar irradiance. It appears that the measured Vpv voltage tends to follow the profile of the overall irradiance and the Ipv current. This variation is quite normal because, according to Ohm's law, the voltage across the resistive load is proportional to the current intensity.

Having our database, the ANFIS controller structure is developed as illustrated in figure 20 where irradiance (G) and temperature (T_{PV}) are the input variables. The output variable which is the PV generator voltage at which the maximum power point occurs is taken as the reference voltage. The optimum voltage produced by ANFIS is compared to the reference voltage of the PV generator and the error is given to generate operating signals. The operating signal is then given to the PWM generator. The generated PWM signals manage the DC - DC converter duty cycle to adjust the operating point of the PV module.

4.2. Modeling of ANFIS on Matlab / Simulink

To build the ANFIS structure, 509 elements are used as training data, 120 as checking data and 120 as testing data. Each data is made up of input (Irradiance, Temperature) and output (voltage) variables. Then, the number and type of input membership functions are defined to configure our fuzzy system for training. Seven membership functions are used for the two input variables and each of the membership use the 'trimf' type functions. To generate the initial membership functions, the genfis1 command is used. Figure 17 illustrates these initial membership functions.

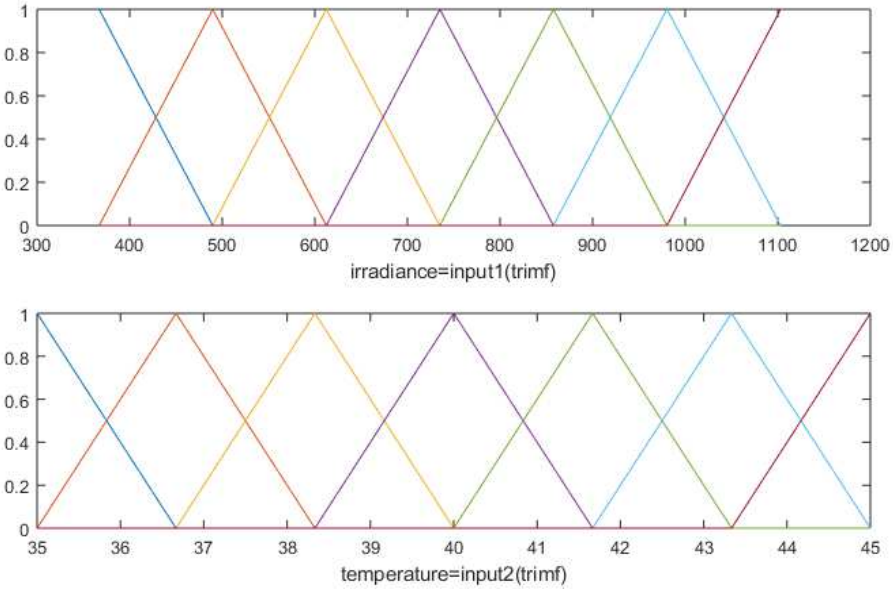


Figure 17. *Initial membership function of ANFIS.*

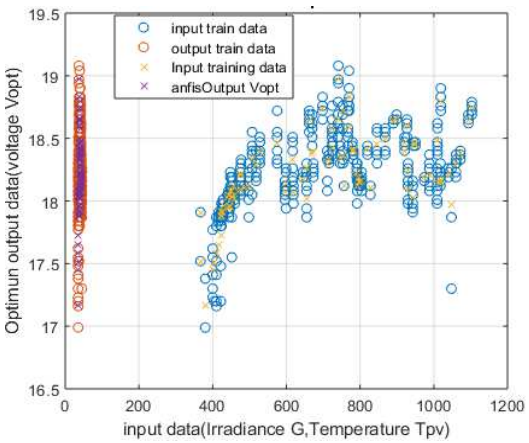


Figure 18. *Comparison between Vpv and ANFIS output Vopt.*

For the learning of the structure, a hybrid learning technique combining the backpropagation algorithm for the determination of the parameters of the premises (adjustment of the parameters related to the membership functions) and the method of least squares for the estimation substantial parameters is used. For an epoch number of 1000, the result of the training is checked. The syntax evalfis calculates the output of the fuzzy system where RMSE (Root Mean Squared Error) represents the root mean square error. Figure 18 shows the ANFIS output as a function of the system training variables (V_{opt}).

After training, completely unknown data parameters are presented to the model and the performance is tested. Figure 19 illustrates the new membership functions after training.

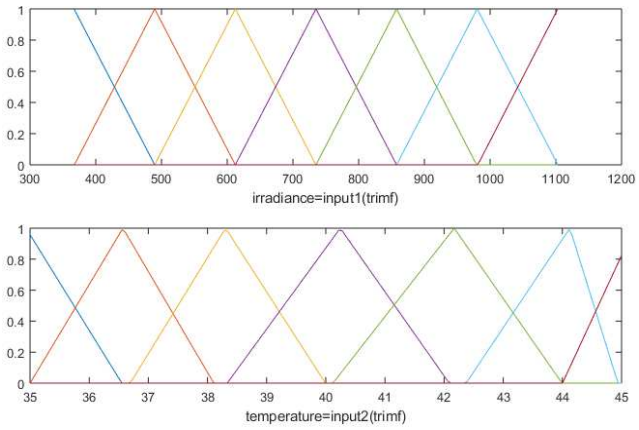


Figure 19. *Final membership function of ANFIS.*

The structure of ANFIS, generated by the Matlab code is a five layer network as shown in figure 20. It has two inputs (irradiance and temperature), one output and seven membership functions for each input. Fourty Nine fuzzy rules are derived from fourteen input membership functions. These rules are derived according to the input and output mapping, so as to construct maximum output power for each value of input temperature and irradiance level. Figure 22 shows output of fuzzy rule for a specific value of operating temperature and irradiance level. it is shown that the MPP voltage (V_{mpp}) vary with the changes of PV cell temperature and solar irradiance.

According to the geographical and meteorological position, an irradiance of 735w / m² and a panel surface temperature of 40°C are required to have an output voltage of 18.2V which corresponds to the manufacturer's V_{mpp}, we must, depending on our geographical and meteorological position, have an irradiance of 735w / m² and a panel surface temperature of 40°C. Figure 21 depicts the typical behavior of the ANFIS structure. It is three-dimensional plot between temperature, irradiance and maximum voltage. It is shown that with an increase in the irradiance level and a moderate temperature, the maximum available power of the PV module increases [19].

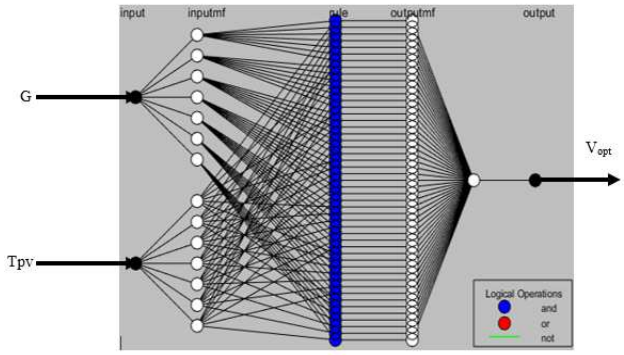


Figure 20. *ANFIS structure.*

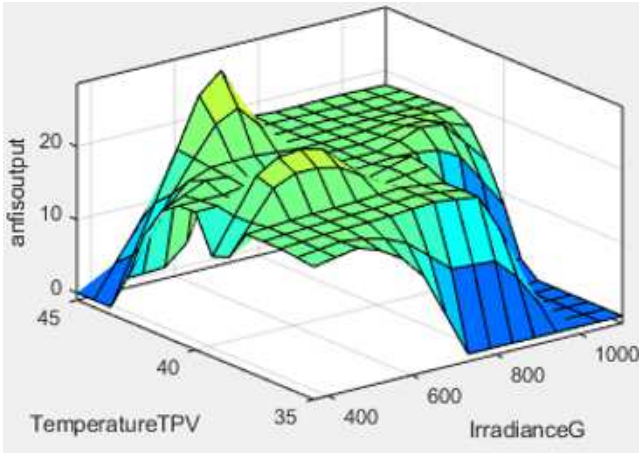


Figure 21. *Surface view created by ANFIS.*

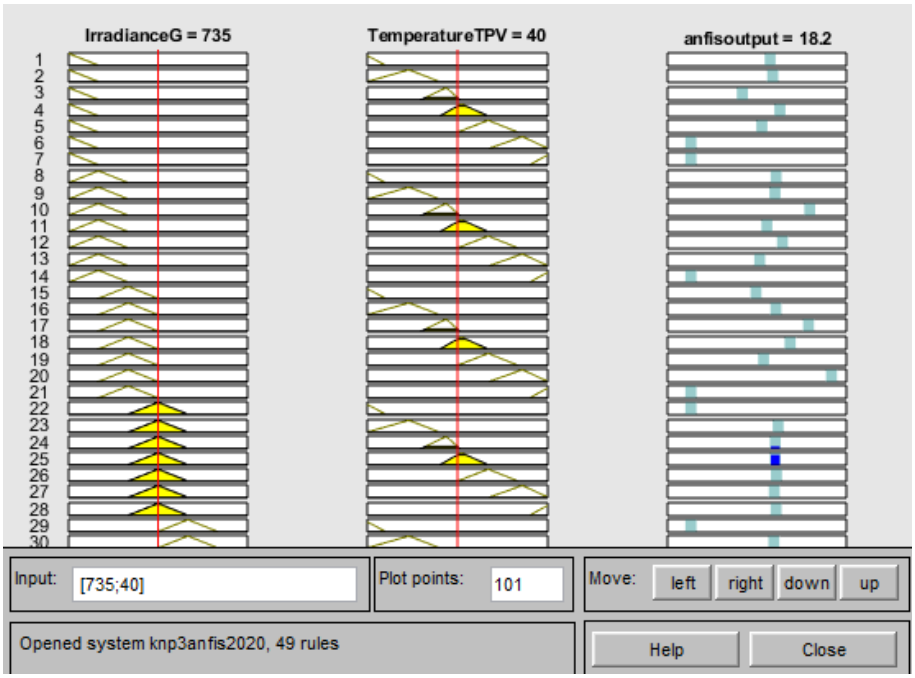


Figure 22. *Rule base of ANFIS controller.*

In the following section, the results presented show that the proposed ANFIS based MPPT is more stable and faster than the conventional MPPT algorithms.

5. Results and Discussion

The Simulink model of the standalone PV system is shown in figure 23. It consists of PV panel, DC-DC boost converter and load. The neuro-fuzzy MPPT control algorithm is associated to the converter to operate the generator at its maximum power point.

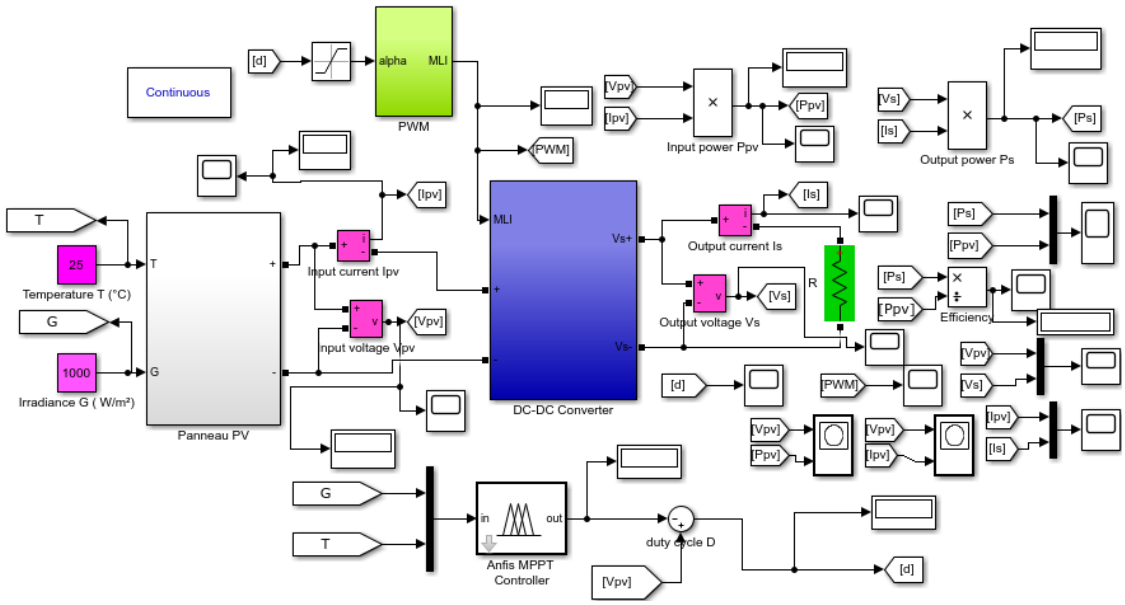


Figure 23. *PV system with ANFIS MPPT controller.*

5.1. Stable Environmental Condition

Simulations are carried out under fixed conditions of illumination G = 1000 w / m² and temperature T = 25°C. The output characteristics of the neural controller are illustrated by figures 24, 25 and 26.

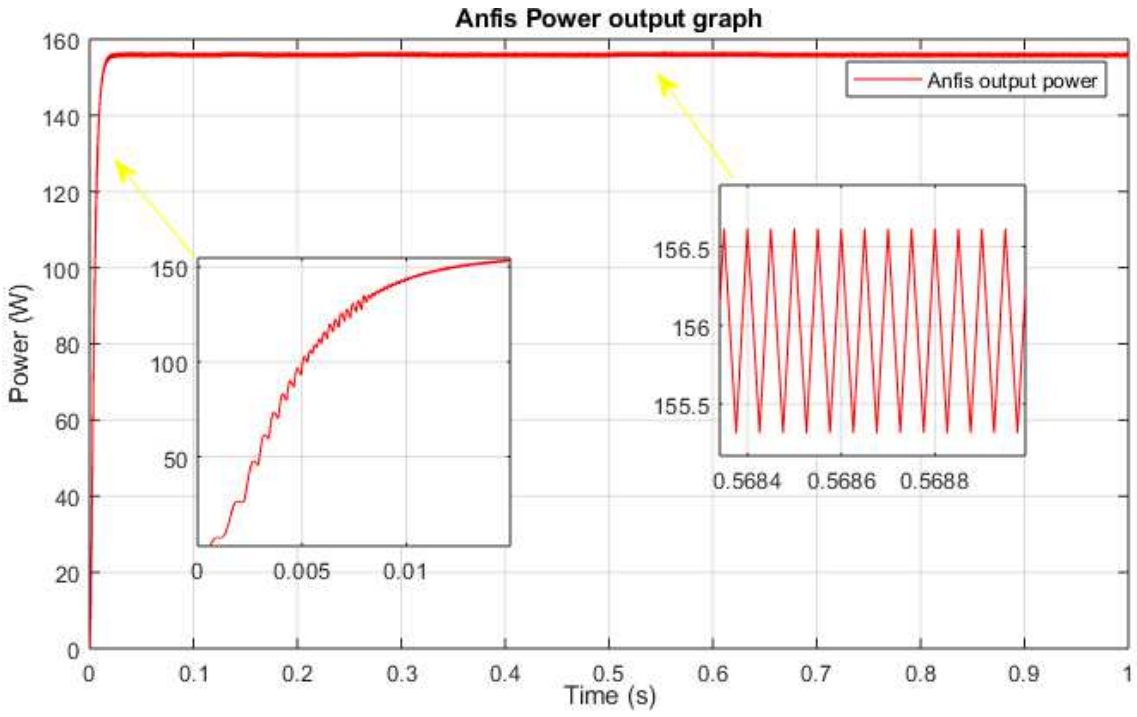


Figure 24. *Panel output power.*

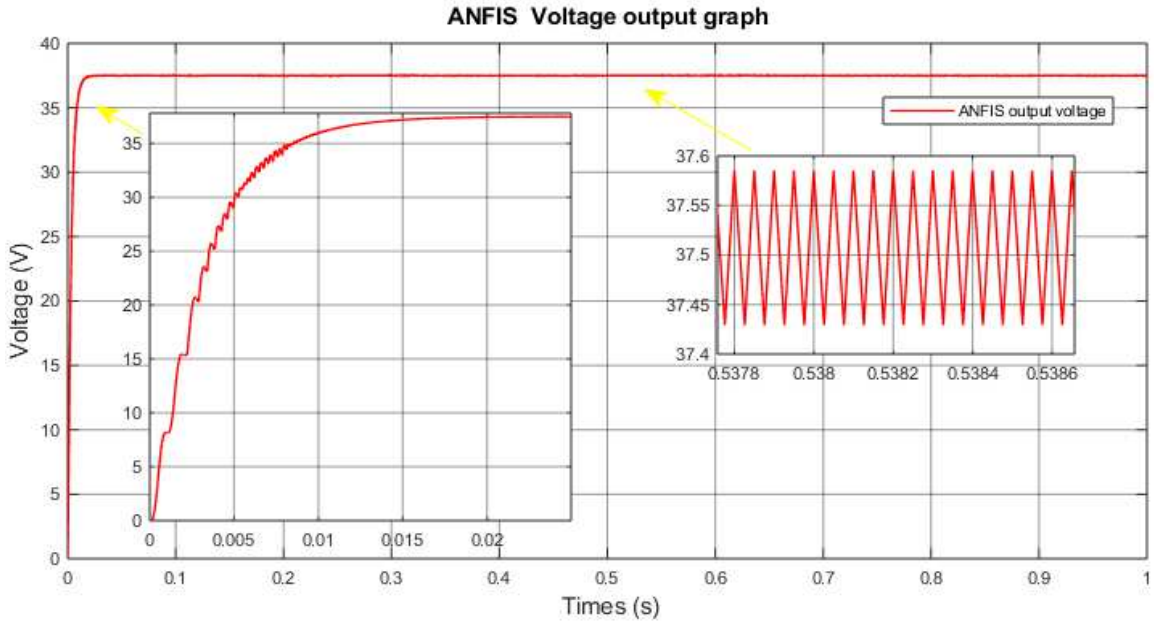


Figure 25. *Panel output voltage.*

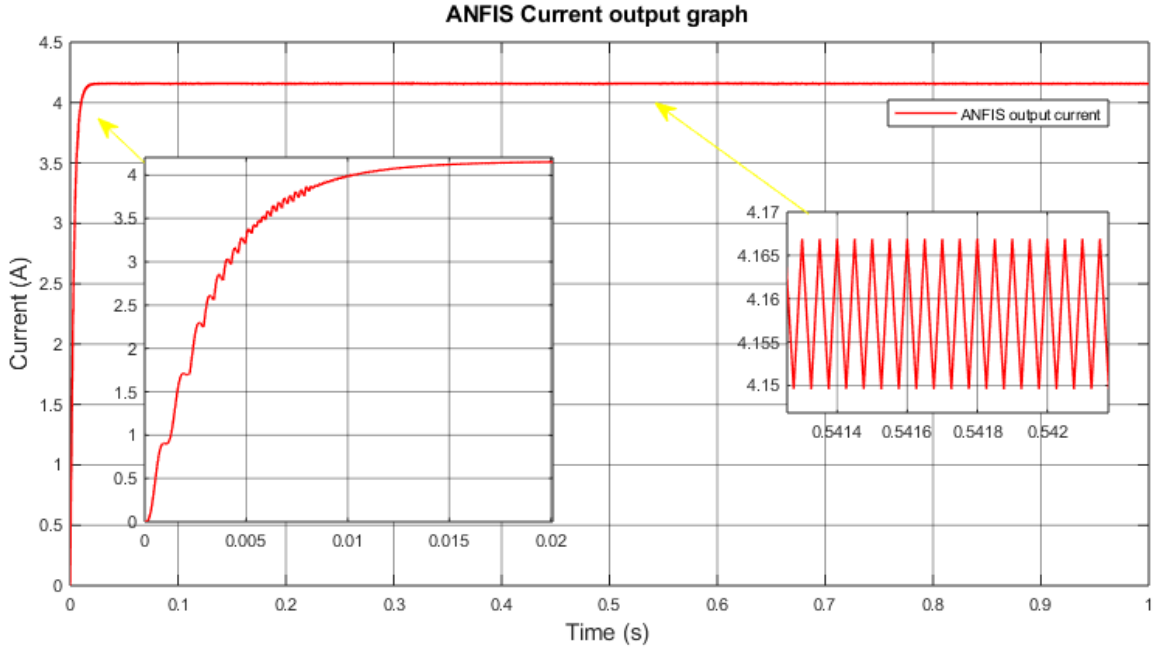


Figure 26. *Panel output current.*

As shown in Figure 24, the output power of the PV system using ANFIS MPPT controller has a significant stability with a very low response time of around 0.01s.

the important oscillations present in the transient phase disappear after 0.007s. However, there are weak oscillations that remain because this power varies between 155.3 and 156.7W. The output current is relatively low while the output voltage is very high with an average value of 37.5V.

Figure 27 compares neuro-fuzzy and conventional controllers. The ANFIS controller can accurately track the maximum power point of the PV generator.

Compared to conventional MPPT, the output power generated by ANFIS is more stable in both steady-state and transient conditions and closer to MPP. It converges quickly to the new MPP. Conversely, although the P&O, InC and HC controls follow the MPP perfectly, they are slower and only arrive at the MPP after a delay. The recovery time is approximately 0.018s for these techniques. The average values of the currents and voltages supplied by these 4 controllers are almost identical, ie approximately 4.15A and 37.5V. However, it should be noted that the currents and voltages supplied by conventional controllers exhibit significant oscillations. The ANFIS controller improved the transition state by reducing steady state oscillations and speeding up the tracking process.

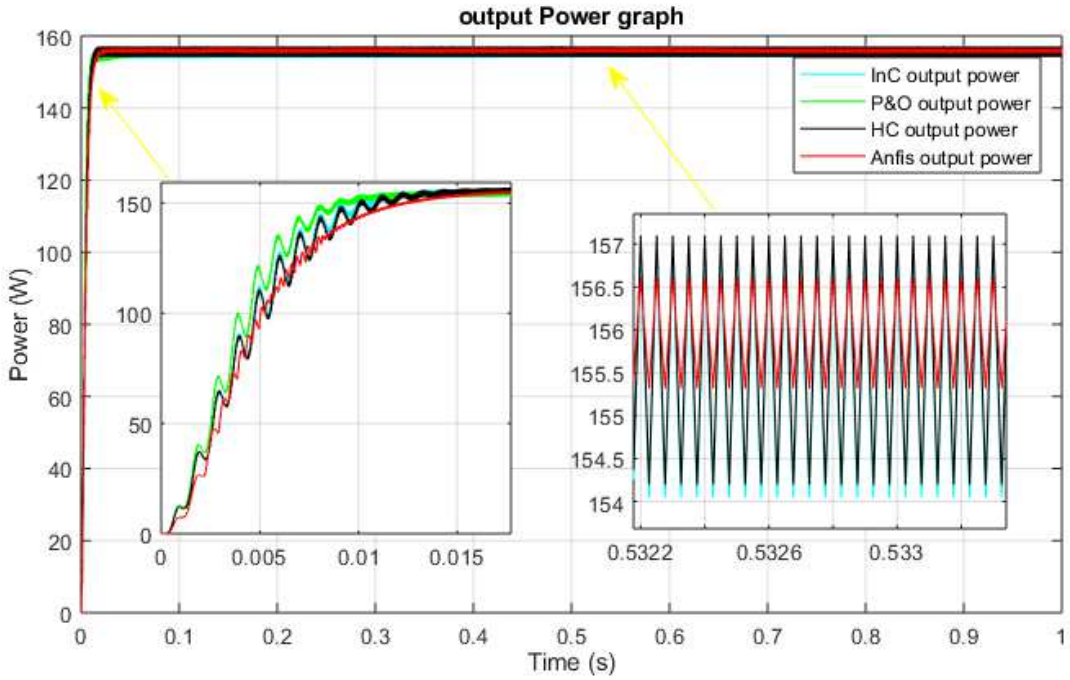


Figure 27. *Power output for various MPPT control in steady state.*

5.2. Variable Environmental Condition

Under rapidly change conditions, there are sharp variations in irradiance and / or temperature. However, the variation in temperature has little influence on the output power compared to the variation in sunlight. simulations are realised with a constant temperature equal to 25°C for a solar irradiation which suddenly deviates from 1000 to 700 W / m² then from 700 to 1200 W / m² and this for 1s. Figure 28 illustrates the output power of the generator in unstable conditions.

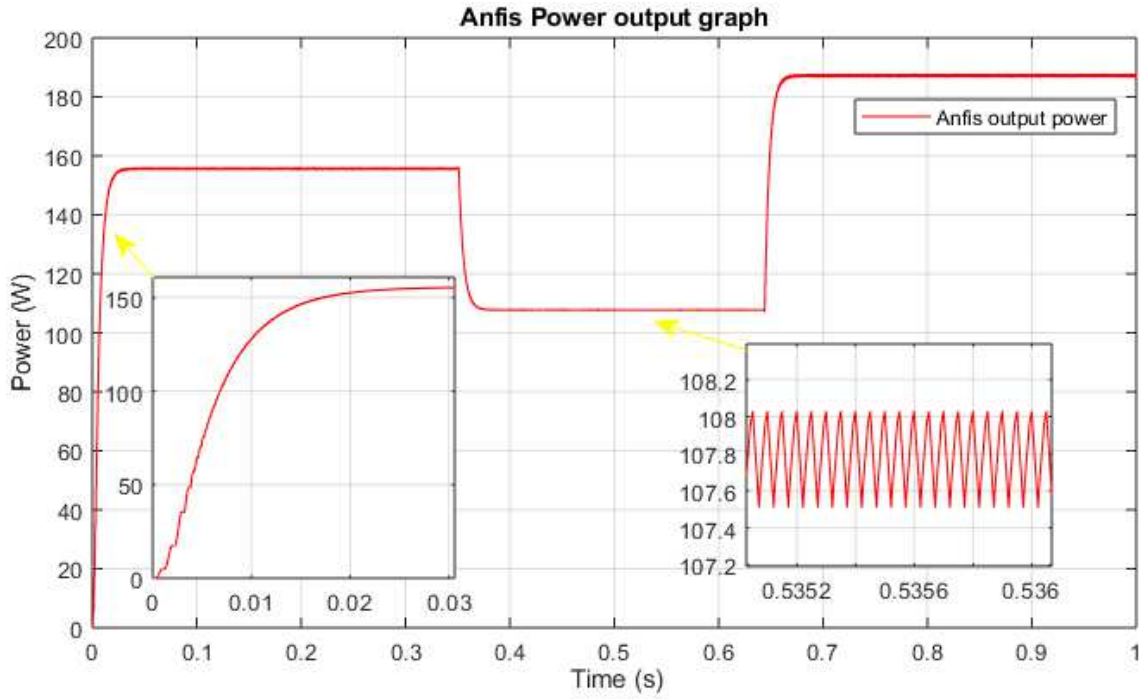


Figure 28. *Disturbed output power.*

In order to evaluate the robustness, the rapidity, the precision and the speed of convergence of the ANFIS technique developed as well as its capacity to follow the MPP under the conditions of abrupt variation of the environmental conditions, a comparison is made with the classical methods (figures 29, 30 and 31).

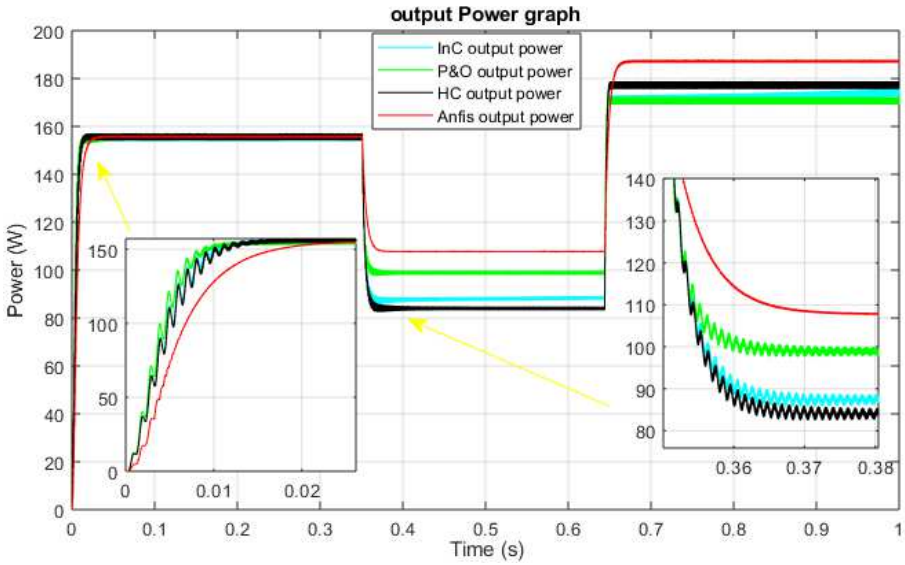


Figure 29. *Comparison of output powers in unstable conditions.*

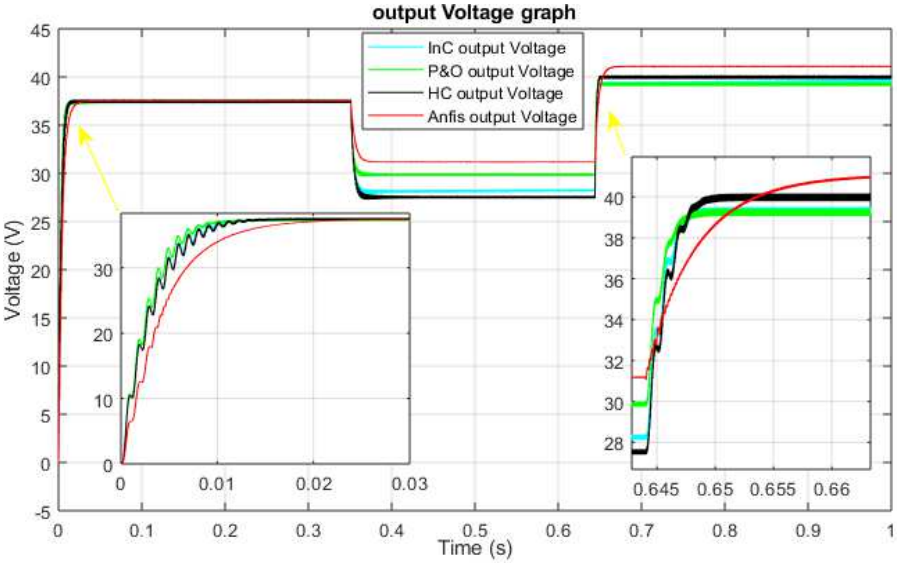


Figure 30. *Comparison of output voltages in unstable conditions.*

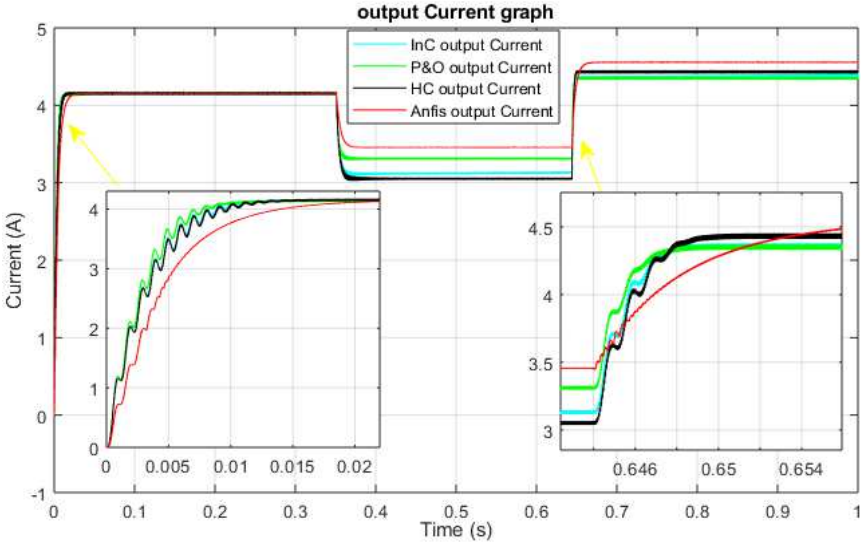


Figure 31. *Comparison of output currents in unstable conditions.*

During rapid changes in climatic conditions, the output characteristics (maximum power, voltage and current) provided by the PV generator varies proportionally with irradiation as shown in Figures 29, 30 and 31. When irradiation is 1000W / m², the maximum power supplied by the GPV stabilizes around 156 W for the different MPP tracking techniques. When the sun goes from 1000 to 700W / m², this maximum power is 110 W for ANFIS, 99W for P&O, 88W for InC and finally 85W for HC. Finally, when going from 700 to 1200 W / m², Pmax becomes equal to 190W for ANFIS, 170W for P&O and InC and 178W for HC. There are significant oscillations of conventional techniques compared to ANFIS first in transient and steady state as well. The sudden variation in sunlight greatly disturbs conventional controllers. In our case, for low irradiation, the HC controller is less cost effective compared to other conventional techniques, but it becomes better when the irradiation becomes greater. These results show that MPPT controllers allow adaptation of PV generator and load to MPP with optimal transfer of PV power.

6. Conclusion

In this work, a neurofuzzy controller is used to modelise the MPPT command for a PV generator in disturbed conditions. In order to achieve the goals, PV system elements have been modeled in Matlab / Simulink. A review on MPPT commands was developed and allowed to highlight the difficulties of classical MPPT commands in the MPP research and to give particular interest of using ANFIS as MPPT controller in PV system. From the experimental tests carried out on site, ANFIS controller has been developed and modeled using the temperature and solar irradiance as input variables and voltage as the output variable. The developed ANFIS model has been trained, tested and validated in Simulink/Matlab, then has been inserted into the system to regulate the DC-DC boost converter. Simulation of the behavior of the PV system under stable and disturbed environmental conditions has been carried out to analyze the characteristics obtained at the output of the panel. A comparative study between the ANFIS controller and conventional MPPT controllers reveals robustness, high stability and very low response time compared to conventional methods. The efficiency of the ANFIS MPPT controller is about 98%.

References

[1] A. Dandoussou, M. Kamta, L. Bitjoka, P. Wira and A. Kuitche (2016), "Comparative study of the reliability of MPPT algorithms for the crystalline silicon photovoltaic modules in variable weather conditions ", Journal of Electrical Systems and Information Technology (JESIT), Elsevier.

[2] Chaouachi A, Kamel RM, and Nagasaka K (2010), "A novel multi-model neuro-fuzzy-based MPPT for three-phase grid-connected photovoltaic system ", Solar Energy, 84 (12): 2219-2229.

[3] Hassan A. Yousef (1999), "Design and implementation of a fuzzy logic computer-controlled sun tracking system ", In the IEEE International Symposium on Industrial Electronics, IEEE, Bled, Slovenia, 3: 1030-1034.

[4] M. Veerachary, T. Senjyu, K. Uezato (2012), "Neural Network based Maximum Power Point Tracking of Coupled-Inductor Interleaved-boost Converter Supplied PV System using Fuzzy Controller", IEEE Transactions on Industrial Electronics, Vol. 50, pp. 749-758.

[5] Z. Cheng, Z. Pang, Y. Liu and P. Xue (2010), "An adaptive solar Photovoltaic array reconfiguration methods based on Fuzzy control", in 2010 8th World Congress on Intelligent Control and Automation (WCICA), pp. 176-181.

[6] W. Xiao, W. G. Dunford, and A. Capel (2004), "A novel modeling method for photovoltaic cells", in Proc. IEEE 35th Annu. Power Electron. Spec. Conf. (PESC), vol. 3, pp. 1950–1956.

[7] Syed Zulqadar Hassan, Hui Li, Tariq Kamal, Ugur Arifo glu, Sidra Mumtaz and Laiq Khan (2017), "Neuro-Fuzzy Wavelet Based Adaptive MPPT Algorithm for Photovoltaic Systems", MDPI, Energies, 10, 394.

[8] M. Kamta and O. Bergossi (2008), "Factors affecting the valorization of photovoltaic water pumping projects for irrigation in Adamawa province (Cameroon)", International Scientific Journal for Alternative Energy and Ecology (ISJAEE): Solar Energy, No. 6 (63), pp. 49-52.

[9] K. Kassmi, M. Hamdaoui et F. Olivié (2007), "Conception et modélisation d'un système photovoltaïque adapté par une commande MPPT analogique", Revue des Energies Renouvelables Vol. 10 N°4 451 – 462 451.

[10] Claude Bertin Nzoundja Fapi, Martin Kamta, Patrice Wira (2019), "A comprehensive assessment of MPPT algorithms to optimal power extraction of a PV panel", Journal of Solar Energy Research, Vol 4 No 3 172-179.

[11] Saravana Selvan, Pratap Nair and Umayal (2016), "A Review on PhotoVoltaic MPPT Algorithms", International Journal of Electrical and Computer Engineering (IJECE) Vol. 6, No. 2, pp. 567~582.

[12] Alok Kumar M., Amit Kumar M. Ranjana Arora (2015), "Overview of Genetic Algorithm Technique for Maximum Power Point Tracking (MPPT) of Solar PV System", International Journal of Computer Applications (0975 – 8887) Innovations in Computing and Information Technology.

[13] Hussein KH, Muta I, Hoshino T, Osakada M. (1995), "Maximum photovoltaic power tracking: an algorithm for rapidly changing atmospheric conditions", IEE Proc Gener, Trans Distrib; 142 (1): 59–64.

[14] M. Razzazan, Z. Mirbagheri, and A. Ramezani (2017), "Maximum Power Point Tracking Using Constrained Model Predictive Control for Photovoltaic Systems", Journal of Solar Energy Research (JSER), Spring 2 (2): p. 19-24.

[15] A. Bin-Halabi, A. Abdennour and H. Mashaly (2014), "An accurate ANFIS-based MPPT for solar PV System", Intl. J. Advanced Computer Research 4, 588–595.

[16] Semmah, H. Hamdaoui, A. Ayad, Y. Ramdani (2009), "Commande Floue et Neuro-Floue d'un Dispositif Facts", Rev. Roum. Sci. Techn – Électrotechn. et Énerg., 54, 2, 195–204.

[17] J.-S. R. Jang (1993), "ANFIS: Adaptive-Network-Based Fuzzy Inference System", IEEE Transaction on Systems, Man, and Cybernetics, vol. 23, No. 3, 665- 685.

[18] Long Zhang, Guoliang Xiong, Huijun Zou and Weizhong Guo (2010), "Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference", Expert Systems with Applications, 37, 6077–6085.

[19] Abdessamia Elgharbi, Dhafer Mezghani and Abdelkader Mami (2012), "A maximum power point tracking method based on artificial neural network for a PV system", International Journal of Advances in Engineering & Technology, ISSN: 2231-1963.

Detailed Dynamic Modeling, Control, and Analysis of a Grid-connected Variable Speed SCIG Wind Energy Conversion System

Sunil Kumar Mahapatro, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, skmahapatro78@outlook.com*

Anil Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, anil_sahoo342@gmail.com*

Subhendu Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srikant.p@yahoo.co.in*

Rajib Lochan Barik, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, rajib_barik543@gmail.com*

Abstract: Wind energy as a renewable energy source continues to be a better alternative to fossil-fuel based generation due to its low cost and environmental benefits. While considerable research efforts have been focused on modeling and control of grid connected variable speed squirrel cage induction generator (SCIG) wind energy conversion systems (WECS), comprehensive models for grid integration studies have been almost non-existent. This paper presents the detail modeling, control and analysis of a grid-connected 2.25-MW variable speed SCIG based WECS that can be utilized for grid integration studies. The presented WECS model consists of a pitch regulated wind turbine connected to a SCIG through a gear box. Then, a full-capacity power electronic converter with maximum power point tracking (MPPT), dc bus voltage regulation and power factor correction, connects the SCIG to the grid. The power converter system comprises of back to back two-level voltage source converters linked through a dc-link capacitor. The generator and grid-side converters are controlled using indirect rotor field-oriented control (IR-FOC) and voltage-oriented control (VOC) respectively. The overall system is simulated in MATLAB/Simulink for a varying wind speed. Results show that the presented model is adequate and efficient in the representation of a variable speed SCIG WECS. In addition, the model meets all of the system performance objectives while simultaneously meeting all of the control objectives.

Keywords: Modeling, Control, Simulation, Squirrel Cage Induction Generator, AC/DC/AC Converter, WECS

1. Introduction

TO DATE, the global penetration of renewable energy sources into the grid is in an all-time high due to falling costs, increases in investments, and advances in enabling technologies. Out of all the renewable energy sources, wind energy has the least-cost option for new power generation capacity making it one of the fastest growing renewable energy resources. 2019 saw 60 GW of annual global additions in installed wind capacity, which represented a 10% increase in global cumulative capacity [1]. It is expected that by 2025, there will be 469 GW of new wind capacity [2].

WECS convert energy stored in the wind to mechanical energy and finally to electrical energy. WECS are either fixed speed or variable speed. As the name implies, fixed speed WECS run at a fixed turbine rotor speed for different wind speeds. Fixed speed WECS use SCIG and are directly connected to the grid through their stator windings via a transformer and soft starter. The advantages of this configuration include low initial cost, reliable operation and simplicity. However, fixed speed WECS draw uncontrollable reactive power from the grid and are not always running at their maximum power point, as a result, they are less efficient. In addition, all fluctuations in the wind speed are further transmitted as fluctuations in the mechanical torque and then as fluctuations in the electrical power on to the grid [3-5]. On the other hand, variable speed WECS are connected to the grid through power electronic converter systems that enable MPPT operation at different wind speeds, dc bus voltage regulation and power factor correction. Furthermore, through the dc bus, the converter system fully decouples the generator from the grid thus allowing for a smoother grid output power. Due to

these advantages, the variable speed WECS has become the standard configuration nowadays.

There are various types of generators used in the variable speed WECS. The most favorable for MW-level applications are Doubly fed induction generators (DFIGs), permanent magnetic synchronous generators (PMSGs), wound rotor synchronous generators (WRSGs), and SCIG. DFIG holds the highest market share to date but future projects announced by the wind turbine manufactures indicate that variable speed WECS with PMSG, WRSG, and SCIG will take over the wind energy market in the coming years. Among the variable speed induction generators, SCIG based WECS is expected to be the most important in the near future due to the following facts: 1) the high efficiency and the low cost installation and maintenance of the SCIG; and 2) the continuing reduced costs of the power electronic devices even in higher power levels, since the SCIG connection to the grid is implemented using a full scale back to back AC/DC/AC frequency converter [6, 7].

The major components of a grid connected WECS can be broadly classified as mechanical, electrical, and control systems [7]. In order to perform grid integration studies, a comprehensive model that incorporates all of these features is essential for an accurate representation of a grid connected WECS. Several papers [6, 8-16], in literature have proposed models of grid connected variable speed SCIG based WECS. However, these models are not sufficient enough in the detail characterization of the whole WECS. Most of these papers [8-14] focus on control schemes that utilize the converters to achieve optimum operation of the WECS. Others exclude wind turbine control [6], and converter control formulation [15, 16] in their analysis. Moreover, these models are not comprehensive enough for integration in a large-scale power systems simulation software package. From the above, it can be concluded that there is a need for a comprehensive model of a variable speed WECS based SCIG that can be utilized for grid integration studies. Such a solution has not been reported in the open literature to the best of the author's knowledge.

Aiming on the aforementioned, this paper presents a comprehensive non-linear model of a grid connected variable speed SCIG based WECS that can be utilized for grid integration studies. The presented WECS model consists of subsystems namely: a wind profile, aerodynamic model, drive train model, pitch control model, SCIG model, AC/DC/AC converter model, and a phase locked loop (PLL) model. Also presented is a detailed converter control methodology that utilizes IR-FOC and VOC to achieve MPPT tracking, dc bus voltage regulation and power factor correction. The design of the converter control loops is aided by an input-output linearization method that eliminates the nonlinearity and coupling of system equations therefore making the design of the proportional integral (PI) controllers more convenient and the system is more easily stabilized. The Butterworth method is applied to design the PI controller gains and the full-scale back to back voltage source converters, are modulated by the sinusoidal pulse width modulation (SPWM) scheme for gate turn on of insulated-gate bipolar transistor (IGBT) switches. Simulations executed in MATLB/Simulink, are presented in order to evaluate the model's performance and control effectiveness for different wind speed values.

The remainder of the paper is organized as follows: Section II presents detailed modeling and control of the proposed WECS, while Section III provides the model's performance simulation results. Section IV concludes the paper.

2. SCIG Wind Generation System

The schematic diagram of the grid connected variable speed WECS with SCIG is shown in Figure 1. An AC/DC/AC converter is used to connect the SCIG to the grid. The AC/DC converter is linked to the DC/AC converter by a dc-link capacitor. A line inductor, L_g, representing the leakage inductance of the transformer and a line resistor, R_g, are assumed between the grid-side converter and the grid. The rest of this section aims to describe the control methodology and model the different subsystems of the above mentioned WECS.

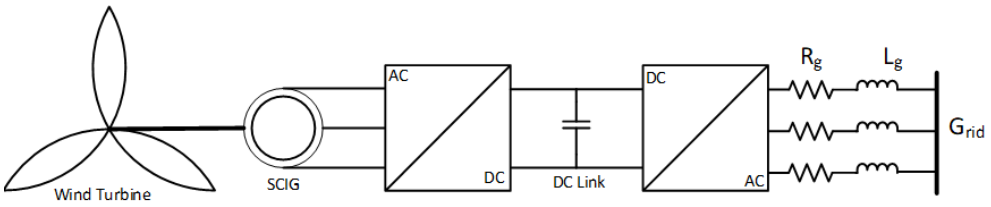


Figure 1. *A simple block diagram of a variable speed WECS with SCIG.*

A) Wind Profile

The wind profile is generated by an autoregressive moving average (ARMA) model [17] and is utilized in the simulation as shown in Figure 2. The wind profile, $v_w(t)$, is modeled as the sum of two components:

$$v_w(t) = v_m + v_t(t) \tag{1}$$

where v_m is the mean wind speed at hub height and $v_t(t)$ is the instantaneous turbulent part and is defined by the equation:

$$v_t(t) = \sigma_t \vartheta_t \tag{2}$$

where σ_t is the standard deviation and ϑ_t is the ARMA time series model:

$$\vartheta_t = \phi_1\vartheta_{t-1} + \phi_1\vartheta_{t-1} \cdots + \phi_n\vartheta_{t-n} + \alpha_t - \theta_1\alpha_{t-1} - \theta_2\alpha_{t-2} \cdots - \theta_m\alpha_{t-m} \tag{3}$$

where $\phi_i(i = 1,2,\dots,n)$ and $\theta_j(j = 1,2,\dots,m)$ are the autoregressive parameters and moving average parameters.

Their numerical values can be found in [17].

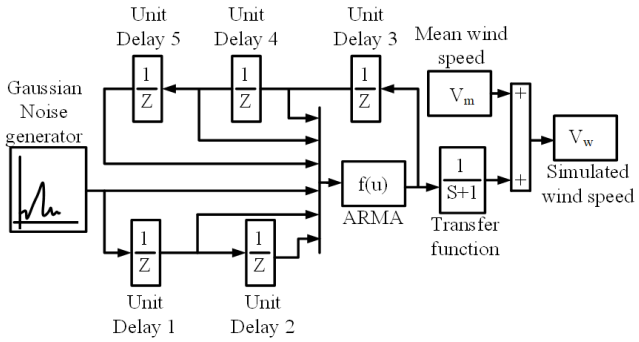


Figure 2. *Generation of wind profile using ARMA model.*

B) Aerodynamic Model

The wind turbine is the prime mover that converts the kinetic energy in the wind into mechanical power that produces electric power. The mechanical power is calculated as:

$$P_m = \frac{1}{2}\rho A {v_w}^3 C_p(\lambda, \beta) \tag{4}$$

where ρ is the air density (in kilograms per cubic meter), $A = \pi R^2$ is the cross-sectional area of the wind turbine through which the wind passes (in square meter), and R is the blade radius. v_w is the wind speed (in meters per second), and C_p is the power coefficient of the blade. The power coefficient is a function of the blade pitch angle, β (the angle at which the rotor blades rotate on their longitudinal axis), and tip speed ratio, λ. The numerical approximation of C_p used in this work is [18]:

$$C_p = 0.5(\lambda - 0.022\beta^2 - 5.6)e^{-0.17\lambda} \tag{5}$$

Tip speed ratio is defined as the ratio of the blade tip speed to the wind speed:

$$\lambda = \frac{\omega_t R}{v_w} \tag{6}$$

where ω_t, is the turbine rotor speed (in rad/s). The relationship between C_p and λ for different values of blade pitch angle ranging from zero to 26 degrees is shown in Figure 3. From this figure, it can be inferred, that at $\beta = 0$ and by choosing the optimal tip speed ratio, λ_{opt}, for the maximum power coefficient, C_{p-max}, the maximum power can be extracted from the wind. Therefore, the reference mechanical speed according to the optimal tip speed ratio, λ_{opt} is calculated by:

$$\omega_m^{ref} = \frac{\lambda_{opt} v_w n_g}{R} \tag{7}$$

where n_g is the gearbox ratio. The reference mechanical torque is:

$$T_m^{ref} = \left.\frac{\frac{1}{2}\rho A v_w^3 C_{p-max}(\lambda_{opt},\beta)}{\omega_m^{ref}}\right|_{\beta=0} \tag{8}$$

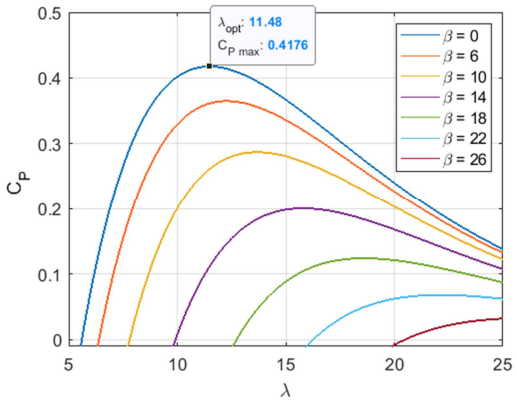


Figure 3. *Power coefficient and tip speed ratio for various values of pitch angle.*

C) Drive Train Model

The drive train of a WECS consists of, shafts, gearbox, and bearings. Some of the drive train components can be neglected depending upon the investigation being carried out. For example, for fixed speed WECS, a two-mass drive train model consisting of a low speed and high-speed shaft or a higher order model is required when transient stability and flicker are being investigated [19-20]. For variable speed wind WECS, the drive train dynamics have almost no effect on the grid side characteristics due to the decoupling effect of the power electronic converter system. Therefore, a one-lumped mass model consisting of a rigid low speed shaft is often considered in studies involving variable speed WECS [21]. Figure 4 shows the one-lumped model and thereafter, are the model's governing equations.

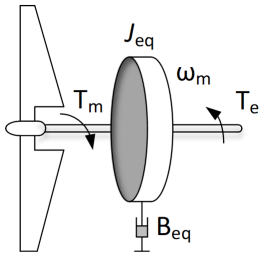


Figure 4. *One mass model of the wind turbine.*

$$J_{eq}\frac{d\omega_m}{dt} = \frac{T_m}{n_g} + T_e - B_{eq}\omega_m \tag{9}$$

$$n_g = \frac{T_m}{T_{m\_g}} = \frac{\omega_m}{w_t} \tag{10}$$

where

$$J_{eq} = J_g + \frac{J_t}{n_g^2} \tag{11}$$

$$B_{eq} = B_g + \frac{B_t}{n_g^2} \tag{12}$$

where J_{eq} is the total equivalent rotational inertia of the system and is derived from (11) where J_g and J_t are the generator and turbine rotor inertias respectively. B_{eq} is the total external damping coefficient, and B_g and B_t are the

generator and turbine damping coefficients respectively. The electromagnetic torque is T_e and the mechanical torque transferred to the generator side is T_{m_g}. The mechanical angular speed is ω_m.

D)Pitch Control Model

The pitch control model consists of two operating regions, namely the partial-load region and the full-load region. In the partial-load region, the wind speed is lower than the rated wind speed. As a result, the blade pitch angle is set at zero resulting to C_{p-max} and λ_{opt}, hence leading to maximum energy extraction from the wind turbine. In the full-load region, the wind speed exceeds its rated value and pitch control is activated to control the generator's output power. This control is accomplished by increasing the blade pitch angle to shed some of the aerodynamic power so that the generator's output power can be maintained at its rated value.

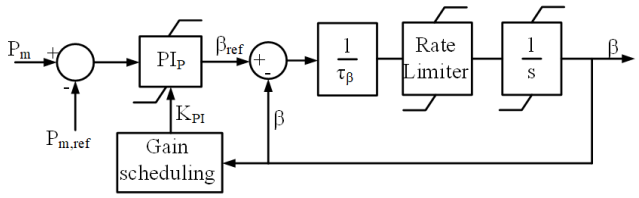


Figure 5. *Pitch angle control model.*

The block diagram of the implemented pitch angle control model is shown in Figure 5. The system consists of a pitch servo which is modeled as a first-order transfer function:

$$\dot{\beta} = \frac{1}{\tau_\beta}\beta_{ref} - \frac{1}{\tau_\beta}\beta \text{ (13)} \qquad (13)$$

For a practical response in the pitch angle control model, the pitch servo accounts for a servo time constant τ_β, and the limitation of both the pitch angle from 0 to 30 degrees and its gradient (± 10 degrees/s). The error between the measured mechanical power and rated mechanical power, ΔP_m, is sent to a PI controller with gain scheduling which outputs a reference pitch angle β_{ref}, which is then compared to the actual pitch angle, β and the resulting error $\Delta\beta$, is corrected by the pitch servo.

The PI controller is designed with gain scheduling to provide satisfactory control over different operating points of the WECS. To design the gains of this controller, the nonlinear wind turbine dynamics are linearized about a specific operating point ($w_{t,OP}$, $\beta_{,OP}$, $v_{w,OP}$). From (4) and expanding as a Taylor series at the operating point and neglecting higher order terms:

$$P_m = P_{m,OP} + \Delta P_m \qquad (14)$$

$$P_m - P_{m,OP} = A\Delta w_t + B\Delta\beta + C\Delta v_w \qquad (15)$$

$$A = \frac{\partial P_m}{\partial w_t}\Big|_{w_t = w_{t,OP}} \qquad (16)$$

$$B = \frac{\partial P_m}{\partial \beta}\Big|_{\beta = \beta_{,OP}} \qquad (17)$$

$$C = \frac{\partial P_m}{\partial v_w}\Big|_{v_w = v_{w,OP}} \qquad (18)$$

where the partial derivatives are evaluated at the operating points and $\Delta w_t = w_t - w_{t,OP}$ $\Delta\beta = \beta - \beta_{OP}$ $\Delta v_w = v_w - v_{w,OP}$ are small changes in w_t, β, v_w from the specified operating point. The pitch angle perturbation can be defined as a summation of the PI controller gains multiplied by the perturbed mechanical power:

$$\Delta\beta = K_{pp}\Delta Pm + K_{ip}\int \Delta Pmdt \qquad (19)$$

Substituting (19) into (15) and transforming to Laplace transform, a transfer function of the closed loop system is obtained.

$$\frac{\Delta P_m}{\Delta V_m} = \frac{sC}{s - sBK_p - BK_i} \qquad (20)$$

The denominator of the transfer function is compared to the second order butterworth polynomial, in order to determine the PI gains. This is further discussed in section F. When the operating point of the WECS changes, the PI controller gains will need to be re-designed to maintain satisfactory response from the pitch angle control model. For instant, at the turbine operating point $V_w = 12.5$, $w_t = 3.39$, and $\beta = 5.2638$ the tuned gains are $K_{pp} = 0.033$ and $K_{ip} = 0.169$. When this operating point changes, the PI controller gains will also change. The change in performance is caused by the variation of pitch angle due to a change in wind speed which in effect also causes a change in the pitch sensitivity $\partial P_m/\partial\beta$. The solution to this problem is to schedule the K_{pp} and K_{ip} gains as a function of pitch angle. Therefore, each PI gain is scaled by a gain scheduling constant $GK(\beta)$ to ensure suitable control loop performance is attained at all wind speeds for all pitch angle variations. The gain scheduling constant [22] is given as follows:

$$GK(\beta) = \frac{1}{(1 + \frac{\beta}{KK})} \qquad (21)$$

where KK is determined as the pitch angle where $\partial P_m/\partial\beta$ has increased by a factor of 2 [23]. β is the output of the pitch angle control model

E) Squirrel Cage Induction Generator Model

The induction generator qd equivalent circuit represented in the synchronous rotating reference frame is shown in Figure 6. Its corresponding qd model equations [24] are

$$V_{qs} = R_s I_{qs} + p\lambda_{qs} + w_e\lambda_{ds} \qquad (22)$$

$$V_{ds} = R_s I_{ds} + p\lambda_{ds} - w_e\lambda_{qs} \qquad (23)$$

$$0 = R_r I_{qr} + p\lambda_{qr} + (w_e - w_r)\lambda_{dr} \qquad (24)$$

$$0 = R_r I_{dr} + p\lambda_{dr} - (w_e - w_r)\lambda_{qr} \qquad (25)$$

$$\lambda_{qs} = (L_{ls} + L_m)I_{qs} + L_m I_{qr} = L_s I_{qs} + L_m I_{qr} \qquad (26)$$

$$\lambda_{ds} = (L_{ls} + L_m)I_{ds} + L_m I_{dr} = L_s I_{ds} + L_m I_{dr} \qquad (27)$$

$$\lambda_{qr} = (L_{lr} + L_m)I_{qr} + L_m I_{qs} = L_r I_{qr} + L_m I_{qs} \qquad (28)$$

$$\lambda_{dr} = (L_{lr} + L_m)I_{dr} + L_m I_{\mathrm{ds}} = L_r I_{dr} + L_m I_{ds} \qquad (29)$$

$$T_e = \frac{3PL_m}{2L_r}(I_{qs}\lambda_{dr} - I_{ds}\lambda_{qr}) \quad (30)$$

where $p = d/dt$ denotes the derivative operator. P is the number of pole pairs. The q-axis and d-axis stator voltages are V_{qs}, V_{ds}. The q-axis and d-axis stator and rotor flux linkages are λ_{qs} λ_{ds} and λ_{qr} λ_{dr} respectively. The q-axis and d-axis stator and rotor currents are I_{qs} I_{ds} and I_{qr} I_{dr}, respectively. L_s and L_r are the stator and rotor self-inductance. L_{ls} and L_{lr} are the stator and rotor leakage inductances whereas L_m is the magnetizing inductance. The rotor electrical speed is w_r and the rotating speed of the synchronous reference frame is w_e. Lastly, R_s and R_r are the stator and rotor resistances.

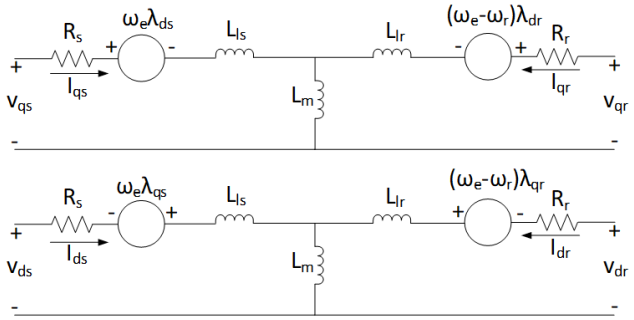


Figure 6. *SCIG Equivalent circuit.*

F) AC/DC Converter Model

A two-level three phase AC/DC converter is used after the induction generator. Its schematic diagram is shown in Figure 7. Writing Kirchhoff's voltage law (KVL) and Kirchhoff's current law (KCL) for the converter and transforming the equations into qd synchronous reference frame results in:

$$V_{qs} = \frac{1}{2}m_{qs}V_{dc} \quad (31)$$

$$V_{ds} = \frac{1}{2}\, m_{ds}V_{dc} \quad (32)$$

$$CpV_{dc} = \frac{3}{4}\left(m_{qs}I_{qs} + m_{ds}I_{ds} - m_{qn}I_{qg} - m_{dn}I_{dg}\right) \quad (33)$$

where I_{dg}, I_{qg}, are the d-axis and q-axis components of the grid currents. The dc-link voltage and dc-link capacitance are V_{dc} and C respectively. The q-axis and d-axis modulation components of the grid-side and generator-side converter are m_{qn}, m_{dn}, and m_{qs}, m_{ds} respectively.

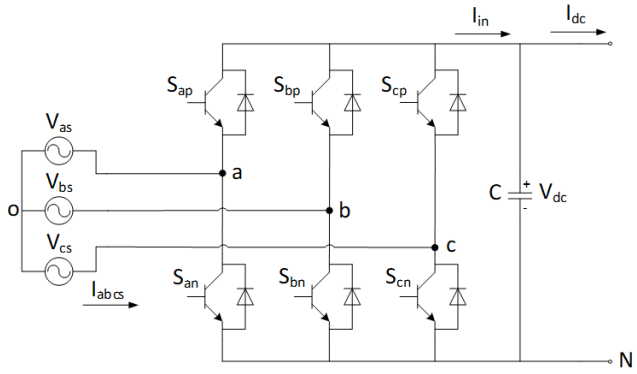


Figure 7. *Generator-side converter.*

1) Control of AC/DC Converter

This generator-side converter is used to control the speed of the generator with an MPPT scheme. As discussed in section B, the implemented MPPT scheme is based on maintaining the tip speed ratio at its optimum value for all wind speeds less than or equal to rated. Control is achieved by using one of the most popular control schemes used in both AC drives and wind energy systems namely the IR-FOC [25]. IR-FOC is achieved by aligning the d-axis of the synchronous reference frame with the rotor flux vector while the q-axis rotor flux linkage and its derivative are zero. The resultant qd axis rotor flux components are:

$$\lambda_{qr} = 0 \quad (34)$$

$$\lambda_{dr} = \lambda_r \quad (35)$$

where λ_r, is the magnitude of the rotor flux vector.

To apply the IR-FOC, the equations governing the induction generator, (22)-(30), are transferred to the rotor flux-model and then (34) -(35) are plugged in resulting into:

$$V_{qs} = RI_{qs} + L_o pI_{qs} + \omega_e L_o I_{ds} + \frac{L_m.\omega_r}{L_r}\lambda_r \quad (36)$$

$$V_{ds} = RI_{ds} + L_o pI_{ds} - \omega_e L_o I_{qs} - \frac{L_m R_r}{L_r^2}\lambda_r \quad (37)$$

$$0 = -\frac{R_r}{L_r}(L_m I_{qs}) + (\omega_{e-}\omega_r)\lambda_r \quad (38)$$

$$0 = p\lambda_r + \frac{R_r}{L_r}(\lambda_r - L_m I_{ds}) \quad (39)$$

$$T_e = \frac{3PL_m}{4L_r}(I_{qs}\lambda_r) \quad (40)$$

where

$$L_o = L_s - \frac{{L_m}^2}{L_r} \quad (41)$$

$$R = R_s + \frac{{L_m}^2}{L_r^2}R_r \quad (42)$$

$$\omega_r = P\omega_m \quad (43)$$

Substituting (40) into (9) and re-arranging gives the rotor electrical speed:

$$p\omega_r = \left[\frac{T_m}{n_g} + \frac{3PL_m}{4L_r}(I_{qs}\lambda_r) - \frac{B_{eq}\omega_r}{P}\right]\frac{P}{J_{eq}} \quad (44)$$

It can be observed that (38) yields the expression of the slip frequency required for IR-FOC.

$$(\omega_{e-}\omega_r) = \omega_{sl} = \frac{L_m I_{qs} R_r}{L_r \lambda_r} \quad (45)$$

The stator frequency is then determined by:

$$\omega_e = \omega_r + \omega_{sl} \quad (46)$$

The angle of the rotor flux vector is obtained from (46) as:

$$\theta_e = \int \omega_e \, dt \quad (47)$$

It should be noted that this angle is used to transform the

stator voltages and currents from abc variables to qd synchronous reference frame. Assuming IR-FOC is implemented in steady state operating conditions, λ_r is maintained at its rated value therefore the d-axis stator reference current is calculated from (39) yielding:

$$I_{ds,ref} = \frac{\lambda_r}{L_m} \tag{48}$$

To design the control scheme of the generator-side converter, the input-output linearization method with decoupling [26] is applied so that the non-linearity and coupling of the induction generator equations, (36)-(37) and (44) can be eliminated thus obtaining a linear relationship between the input control variables and the output-controlled variables. With this transformation, the classical linear control system theory is adopted to determine the structure of each controller as well as the constant gains of the PI controllers. The input control variables are m_{qs} and m_{ds} while the output-controlled variables are the rotor electrical speed w_r and the d-axis stator current I_{ds}.

From (36)-(37) and (44), the current and speed controller outputs are defined as:

$$L_o pI_{qs} + RI_{qs} = k_{qs}\left(I_{qs,ref} - I_{qs}\right) \tag{49}$$

$$L_o pI_{ds} + RI_{ds} = k_{ds}\left(I_{ds,ref} - I_{ds}\right) \tag{50}$$

$$p\omega_r = k_{\omega r}\left(\omega_{r,ref} - \omega_r\right) \tag{51}$$

where k_{qs}, k_{ds}, k_{wr} are the transfer functions of the PI controllers for the stator q-axis current, stator d-axis current, and speed controller respectively. They are defined as:

$$k_{qs} = k_{pqs} + \frac{k_{iqs}}{s},\ k_{ds} = k_{pds} + \frac{k_{ids}}{s},\ k_{\omega r} = k_{p\omega r} + \frac{k_{i\omega r}}{s} \tag{52}$$

Taking Laplace transform of (49)-(51) and re-arranging, the transfer functions are obtained as:

$$\frac{I_{qs}}{I_{qs,ref}} = \frac{(sk_{pqs}+k_{iqs})\frac{1}{L_o}}{s^2+\frac{s}{L_0}(R+k_{pqs})+\frac{k_{iqs}}{L_0}} \tag{53}$$

$$\frac{I_{ds}}{I_{ds,ref}} = \frac{(sk_{pds}+k_{ids})\frac{1}{L_o}}{s^2+\frac{s}{L_0}(R+k_{pds})+\frac{k_{ids}}{L_0}} \tag{54}$$

$$\frac{w_r}{w_{r,ref}} = \frac{sk_{p\omega r}+k_{iwr}}{s^2+sk_{p\omega r}+k_{iwr}} \tag{55}$$

The denominator coefficients of each transfer function are then compared to a second order Butterworth polynomial to determine the PI gains of the respective controllers. The second order Butterworth polynomial is expressed as:

$$s^2 + 2\zeta\omega_n s + \omega_n^2 \tag{56}$$

The PI gains are selected so that the roots of the characteristic equations appear in the left half of the s-plane, on a circle of radius ω_n, with its center at the origin. The damping coefficient, ζ, is chosen for underdamped condition $\left(\zeta = \sqrt{2}/2\right)$, while ω_n is chosen for a good dynamic response.

Combining (51) and (44), the q-axis stator reference current is defined as a function of the speed controller as follows:

$$I_{qs,ref} = \frac{\left[\frac{k_{wr}\left(\omega_{r,ref}-\omega_r\right)Jeq}{P} - \frac{T_m}{n_g} + \frac{Beq\omega_r}{P}\right]4L_r}{3PL_m\lambda_r} \tag{57}$$

The reference q-axis and d-axis modulation indices used in the PWM module to generate the control signals for the IGBT are obtained by combining (36)-(37) with (49)-(50).

$$m_{qs} = \frac{2}{V_{dc}}\left[k_{qs}\left(I_{qs,ref} - I_{qs}\right) + \omega_e L_o I_{ds} + \frac{L_m\omega_r}{L_r}\lambda_r\right] \tag{58}$$

$$m_{ds} = \frac{2}{V_{dc}}\left[k_{ds}\left(I_{ds,ref} - I_{ds}\right) - \omega_e L_o I_{qs} - \frac{L_m R_r}{L_r^2}\lambda_r\right] \tag{59}$$

The speed and current loops for the IR-FOC are shown in Figure 8 and Figure 9.

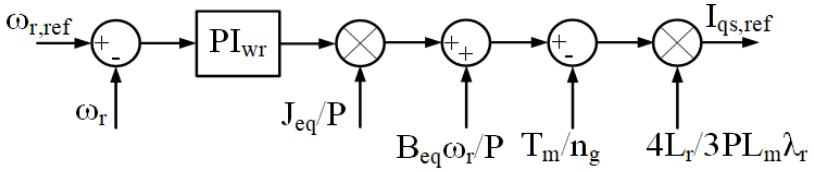


Figure 8. *Rotor electrical speed control loop for the generator-side converter.*

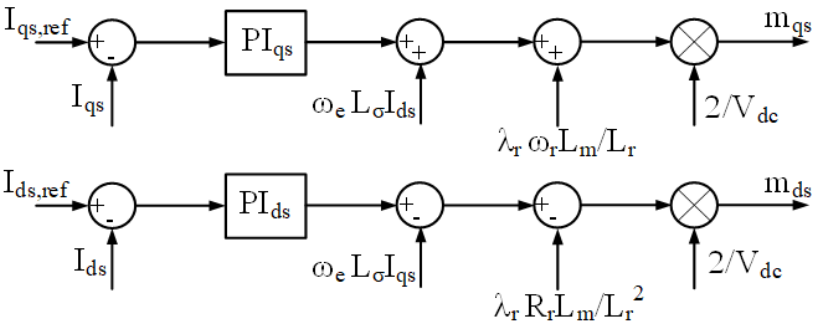


Figure 9. *Current control loops for the generator-side converter.*

G)DC/AC Converter Model

A two-level three phase DC/AC converter (inverter) is shown in Figure 10. It is linked to the generator-side converter via V_{dc}. Using KVL, the grid-side converter dynamic equations in qd synchronous reference frame are obtained as:

$$V_{qn} = L_g\omega_g I_{dg} + L_g pI_{qg} + R_g I_{qg} + V_{qg} \quad (60)$$

$$V_{dn} = -L_g\omega_g I_{qg} + L_g pI_{dg} + R_g I_{dg} + V_{dg} \quad (61)$$

$$V_{qn} = \frac{1}{2} m_{qn} V_{dc} \quad (62)$$

$$V_{dn} = \frac{1}{2}\, m_{dn} V_{dc} \quad (63)$$

where V_{qg}, V_{dg} are the q-axis and d-axis components of the grid voltages; ω_g is the grid frequency. V_{qn}, V_{dn} are the q-axis and d-axis components of the output voltage of the DC/AC converter.

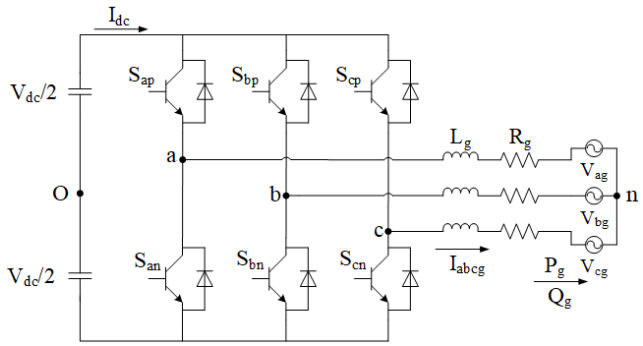


Figure 10. *Grid-side converter.*

1) Control of DC/AC Converter

The objective of the grid-side converter is to keep the dc-link voltage constant and to regulate the power factor at the point of common coupling (PCC) to the grid. Control is achieved by applying VOC [25] scheme. To realize VOC, the q-axis component of the synchronous reference frame, V_{qg}, is aligned with the magnitude of the grid voltage, $(V_{qg} = |V_g|)$ while the d-axis component of the synchronous reference frame, V_{dg} is regulated to zero $(V_{dg} = 0)$. Therefore, the active power and reactive power injected to the grid are obtained by:

$$P_g = \frac{3}{2}\left(V_{qg}I_{qg} + V_{dg}I_{dg}\right) = \frac{3}{2}|V_g|I_{qg} \quad (64)$$

$$Q_g = \frac{3}{2}\left(V_{qg}I_{dg} - V_{dg}I_{qg}\right) = \frac{3}{2}|V_g|I_{dg} \quad (65)$$

The power factor is regulated by controlling the d-axis grid current. From (65), the d-axis grid reference current $I_{dg,ref}$ is obtained as:

$$I_{dg,ref} = \frac{Q_g}{1.5|V_g|} \quad (66)$$

where Q_g can be set to zero for unity power factor operation or negative/positive value for a leading/lagging power factor.

To maintain the dc-link voltage, (33) is used to form the control loop. Multiplying both sides of (33) by V_{dc} and substituting for m_{qn} and m_{dn} using (60) – (63) and by assuming steady state operation, yields:

$$\frac{C}{2} pV_{dc}{}^2 = \left(P_{in} - P_{Loss} - \frac{3}{2}V_{qg}I_{qg}\right) \quad (67)$$

where

$$P_{in} = \frac{3V_{dc}}{4}\left[m_{qs}I_{qs} + m_{ds}I_{ds}\right] \quad (68)$$

$$P_{Loss} = \frac{3}{2}R_g\left[I_{qg}{}^2 + I_{dg}{}^2\right] \quad (69)$$

The dc-link voltage loop is shown in Figure 11. The reference dc-link voltage is chosen such that the maximum magnitude of the modulating index, M_i is 1. The modulating index is defined as the ratio of the fundamental component amplitude of the line-to-neutral inverter output voltage to one half of the dc-link voltage.

$$M_i = \frac{2V_m}{V_{dc}} \quad (70)$$

where M_i is the magnitude of the modulation signal, V_m is the magnitude of the fundamental inverter output phase voltage. From (70) the reference dc-link voltage is derived by:

$$V_{dc,ref} \geq 2(V_m) \quad (71)$$

The maximum rms value of the phase voltage at SCIG terminals as well as the rms value of the inverter output phase voltage is 398.4 V. Substituting this value in (71), $V_{dc,ref}$ should be equal or greater than 1126.8 V. Based on this assessment, the reference voltage of the dc-link is chosen as 1200V.

Using the input-output scheme discussed in section F, the q-axis grid reference current, $I_{qg,ref}$ and reference grid modulation indices, m_{qn}, m_{dn} are obtained as follows:

$$I_{qg,ref} = \frac{2}{3V_{qg}}\left(P_{in} - \left[k_{dc}\left(V_{dc,ref}{}^2 - V_{dc}{}^2\right)\right] - P_{Loss}\right) \quad (72)$$

$$m_{qn} = \frac{2}{V_{dc}}\left[L_g\omega_g I_{dg} + k_{qg}\left(I_{qg,ref} - I_{qg}\right) + V_{qg}\right] \quad (73)$$

$$m_{dn} = \frac{2}{V_{dc}}\left[-L_g\omega_g I_{qg} + k_{dg}\left(I_{dg,ref} - I_{dg}\right) + V_{dg}\right] \quad (74)$$

The current loops for the VOC are shown in Figure 12. The PI gains of the dc-link loop and current loops are determined by the method described in section F.

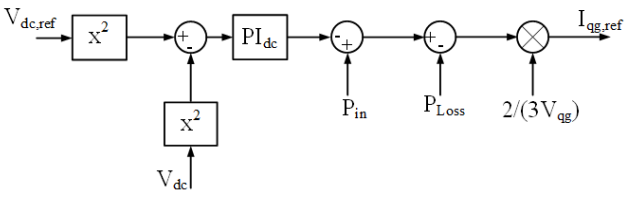


Figure 11. *Dc-link voltage control loop.*

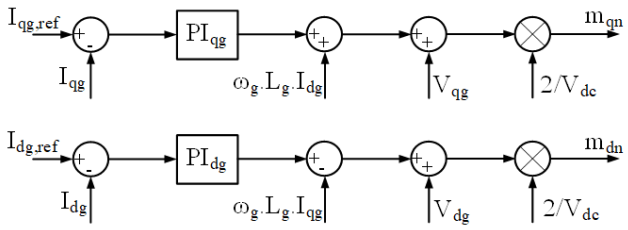


Figure 12. *Current control loops for grid-side converter.*

2) Sinusoidal Pulse Width Modulation

A continuous carrier based SPWM scheme is used to generate the switching functions, $S_{ip}, S_{in}, i = a, b, c$ for the two-level converters. The reference modulation signals, m_{qs}, m_{ds}, m_{qn} and m_{dn} are transformed to abc domain as m_{ap}, m_{bp}, m_{cp}, m_{an}, m_{bn}, m_{cn} and are then compared to a high frequency symmetric triangular carrier signal. When the modulation signal is greater than the triangular signal, then the respective switch turns on while its complimentary switch turns off and vice versa. This relation for a respective converter leg is represented by

$$S_{ip} + S_{in} = 1 \tag{75}$$

H)Phase Locked Loop

A PLL [27] as shown in Figure 13 is implemented to detect the grid's voltage angle, θ_g. This angle is used for the transformation of grid voltages and currents from the abc variables to qd synchronous reference frame. The process of detecting the grid's voltage angle is realized by setting the d-axis reference grid voltage to zero $(V_{dg,ref} = 0)$, and comparing it to the transformed d-axis grid voltage, V_{dg} which results in the lock in of the PLL output, θ to the grid's voltage angle θ_g.

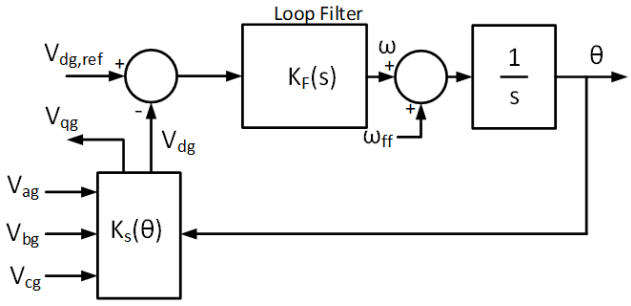


Figure 13. *A PLL block diagram.*

A proper design of the loop filter, $K_F = K_{pf} + \frac{K_{if}}{s}$ is needed to ensure the lock in of the PLL's output to the grid's voltage angle. To tune the filter's gains, a linear PLL model is first derived. This is accomplished by first transforming the grid phase voltages, V_{ag}, V_{bg}, V_{cg}, to qd synchronous reference frame:

$$V_{qg} = Vm\cos(\theta_g - \theta) \tag{76}$$

$$V_{dg} = -Vm\sin(\theta_g - \theta) \tag{77}$$

Assuming the difference between the grid's voltage angle and the PLL's output, is very small, then from (77), $\sin(\theta_g - \theta) \approx (\theta_g - \theta)$ and (77) is now linearized and so is the PLL. Vm now appears as a gain in the forward path and the error is now defined as:

$$e = \theta_g - \theta \tag{78}$$

The transfer function of the closed loop is found as:

$$\frac{\theta}{\theta_g} = \frac{V_m K_p s + V_m K_i}{s^2 + V_m K_p s + V_m K_i} \tag{79}$$

The PI gains are determined by the method described in section F.

3. Simulation Results

The proposed 2.25-MW variable speed SCIG based WECS model is implemented using MATLAB/Simulink. The WECS system is assumed to be connected to an infinite bus through its PCC bus. The infinite bus has a known voltage magnitude and angle, and represents a large power system.

The model's performance under a varying wind speed is evaluated. The rated wind speed for this study is 12 m/s. The reference rotor electrical speed for the generator-side converter controller is calculated from (7) and (43) according to the wind speed, whereas the d-axis stator reference current is set at $I_{ds,ref} = 600\ A$ based on (48). For the grid-side converter controller, the reference d-axis grid current is set at $I_{dg,ref} = 0$.

Figure 14(a) shows the wind speed profile generated by the ARMA model. It is shown from the simulation results, that when the wind speed is below rated speed, pitch angle, Figure 14(c), is not activated and is kept at zero. Meanwhile tip speed ratio λ, Figure 14(b), and the power coefficient, C_p, Figure 14(d), operate at their optimum values of 11.482 and 0.4176 in order to extract the maximum energy from the wind, i.e. MPPT operation. At above rated wind speeds, pitch angle, Figure 14(c), is activated and increases (the blades are pitched) until the excess extracted wind power is shed therefore limiting the generator's output power, Figure 14(l), to its rated value of 2.25 MW. Also, from Figure 14(l), one can see overshoots, for example at time=7 and time=27. This is due to the pitch response not being instantaneous and also due to the small dynamic variations in generator rotor speed which is allowed in order to absorb the fast wind gusts which results in the storage of rotational energy in the turbines inertial. It can also be noted, that in this control region, the wind turbine operates at a lower efficiency as seen in the decrease in performance of C_p and λ(are shifted downward).

Also, from the simulation results, it can be observed that the grid-side and generator-side modulation components, Figure 14(e) and Figure 14(f) operate in the linear modulation region as expected. The qd stator and grid currents are shown in Figure 14(g) and Figure 14(h) respectively. It can be seen that I_{ds} and I_{dg} are regulated to their reference values, whereas, I_{qs} and I_{qg} operate at their suitable steady state values. The dc-link voltage, Figure 14(i), is maintained at its reference value by the grid-side converter while the generator-side converter regulates the generator's electrical rotor speed, Figure 14(j), to its reference value, hence, achieving MPPT.

The reactive power is shown in Figure 14(k), and is set at zero for unity power factor operation as per (66).

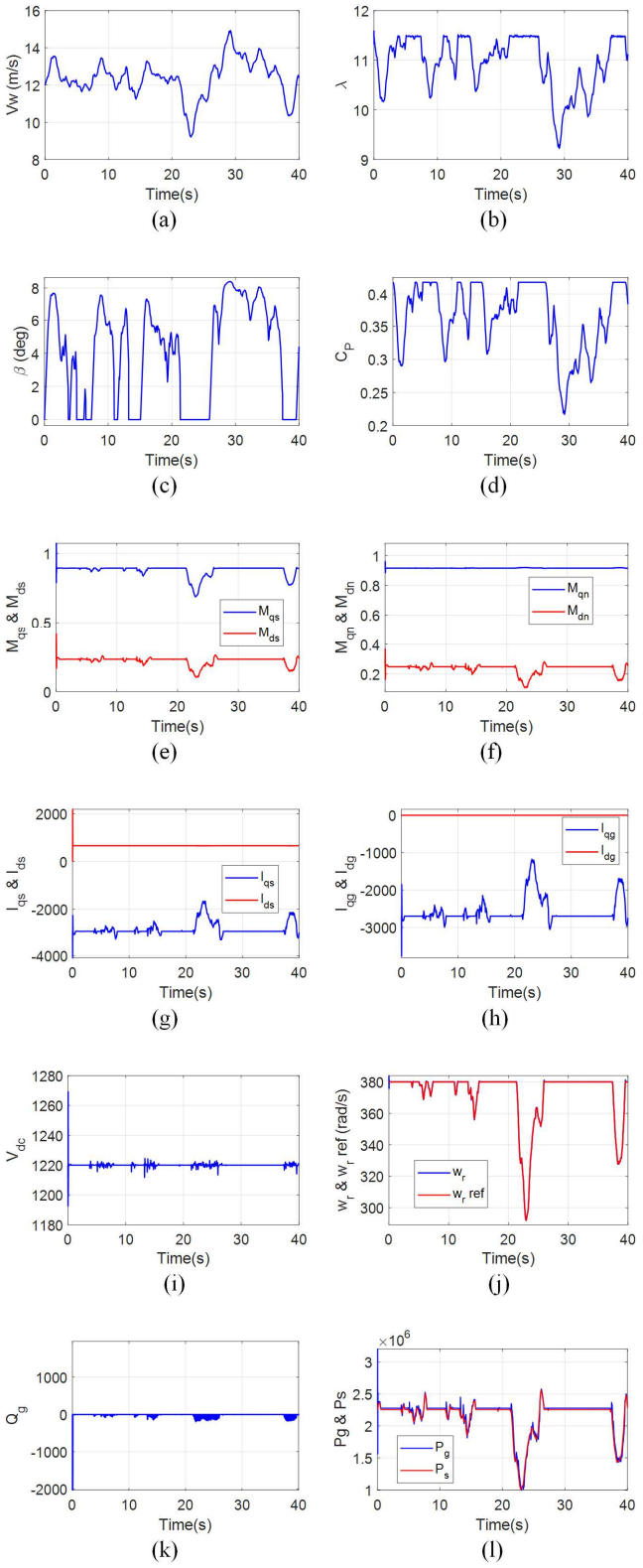


Figure 14. *Simulation results of the variable speed SCIG based WECS. (a) wind speed, (b) tip speed ratio, (c) pitch angle, (d), power coefficient, (e) generator-side modulation index, (f) grid-side modulation index, (g) stator q-axis and d-axis currents, (h) grid q-axis and d-axis currents, (i) dc-link voltage, (j) rotor electrical speed, (k) reactive power delivered to the grid, (l) generator stator power and real power delivered to the grid.*

4. Conclusion

This paper presented a comprehensive dynamic model of a grid connected variable speed SCIG based WECS that can be used for grid integration studies. In the WECS model, wind profile, aerodynamics, drive train, pitch control, SCIG, rectifier, inverter, and PLL have been considered. A detailed step-by-step control strategy was developed and implemented through the converter system. From simulation results, it has been shown that the presented control strategy is efficient for tracking MPPT, maintaining dc-link voltage and regulating the power factor. Also, simulation results for a fluctuating wind verify the fast and very good performance of the variable pitch control model. The developed WECS is detailed enough to capture all system performance objectives, thus, making the model suitable for inclusion in a multi-machine power system. Future work will focus on applying the model to a larger test system and investigating its performance in the presence of major disturbances such as three-phase faults, sudden load changes and line switching operation. In addition, an energy storage device with its associated controllers, will be added to the model, to smooth out the power delivered to the grid.

Appendix

Table 1. *WECS Data.*

$P = 2$	$\omega_e = 377\ rad/s$
$L_m = 2.13461\ mH$	$R_s = 1.102\ m\Omega$
$L_{lr} = 0.06492\ mH$	$R_r = 1.497\ mH$
$L_{ls} = 0.06492\ mH$	$\lambda_r = 0.9983\ Wb\ (rms)$
$\rho = 1.222\ kg/m^3$	$B_{eq} = 0.00015\ N.m.s/rad$
$J_{eq} = 18.7\ \mathrm{kg}.m^2$	$C = 60\ mF$
$R = 40.5987$ m	$n_g = 55.9835$
$R_g = 0.002\ m\Omega$	$L_g = 0.15\ mH$
$\omega_{ff} = 377\ rad/s$	

Table 2. *Controller Data for all control loops.*

$K_{pwr} = 5.8926$	$K_{iwr} = 17.3611$	$K_{pp} = 0.0412$
$K_{pas} = 0.0290$	$K_{ias} = 3.8730$	$K_{ip} = 0.2060$
$K_{pds} = 0.0290$	$K_{ids} = 3.8730$	$K_{pf} = 0.9463$
$K_{pdc} = 6.0670$	$K_{idc} = 613.47$	$K_{if} = 252.2481$
$K_{paa} = 0.3013$	$K_{iaa} = 306.735$	$\tau_\beta = 0.2$
$K_{pdq} = 0.3013$	$K_{idq} = 306.735$	

References

[1] REN21, "Renewables 2020-Global Status Report, "Renewable Energy Policy Network for the 21st Century, 2020.

[2] GWEC, "Global Wind Report 2021, 2021.

[3] P. W. Carlin, A. S. Laxson and E. B. Muljadi, "The History and State of the Art of Variable-Speed wind Turbine Technology," Wind Energy, vol. 6, no. 2, pp. 129-159, Feb. 2003.

[4] S. S. Murthy, B. Sing, P. K. Goel and S. K. Tiwari, "A Comparative Study of Fixed Speed and Variable Speed Wind Energy Conversion Systems Feeding the Grid," in *Proc. IEEE 7th Int. Conf. Power Electronics and Drive Systems* Bangkok, Thailand, Nov 27-30, 2007, pp. 736-743.

[5] T. Ackermann, *Wind power in power systems.* West Sussex, England: John Wiley & Sons Ltd., 2005.

[6] G. C. Konstantopoulos and A. T. Alexandridis, "Full-Scale Modeling, Control and Analysis of Grid-Connected Wind Turbine Induction Generators with Back-to-Back AC/DC/AC Converters," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 2, no. 4, pp. 739-748, December. 2014.

[7] V. Yaramasu, B. Wu, P. C. Sen, S. Kouro, M. Narimani, "High-Power Wind Energy Conversion Systems: State-of-the-Art and Emerging Technologies," in *Proc. IEEE*, vol. 103, no. 5, pp. 740-788, May 2015.

[8] A. Mesemanolis, C. Mademlis, and I. Kioskeridis, "Optimal Efficiency Control Strategy in Wind Energy Conversion System with Induction Generator," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 1, no. 4, pp. 238-246, December. 2013.

[9] V. D. Dhareppagol and S. Nagendraprasad, "Modelling and Simulation of WECS for Maximum Power Extraction and Optimal Efficiency Control using Squirrel Cage Induction Generator," in Proc. *IEEE Power, Communication and Information Technology Conference,* Bhubaneswar, India, October 15-17, 2015.

[10] M. Karrari, W. Rosehart, and O. P. Malik, "Comprehensive Control Strategy for a Variable Speed Cage Machine Wind Generation Unit," *IEEE Transaction on Energy Conversion*, vol. 20, no. 2, pp. 415-423, June. 2005.

[11] R. Pena, R. Cardenas, R. Blasco, G. Asher and J. Clare, "A cage induction generator using back to back PWM converters for variable speed grid connected wind energy system," in *Proc. IEEE Industrial Electronics Society,* Denver, CO, USA, 29 Nov-2 Dec, 2001, pp. 1376-1381.

[12] Manaullah, A. K. Sharma, H. Ahuja, G. Bhuvaneswari, R. Balasubramanian, "Control and Dynamic Analysis of Grid Connected Variable Speed SCIG Based Wind Energy Conversion System," in *Proc. IEEE 4th Int. Conf. Computational Intelligence and Communication Networks,* Mathura, India, Nov 3-5, 2012, pp. 588-593.

[13] B. Kedjar and K. Al-Haddad, "Optimal control of a grid connected variable speed wind energy conversion system based on squirrel cage induction generator," in *Proc. IEEE Industrial Electronics Society* Montreal, QC, Canada, Oct 25-28, 2012, pp. 3560-3565.

[14] M. G. Simoes, B. K. Bose and R. J. Spiegel, "Design and performance evaluation of a fuzzy-logic-based variable-speed wind generation system," *IEEE Transactions on Industry Applications*, vol. 33, no. 4, pp. 956-965, July-August. 1997.

[15] M. K. Paul, M. L. Doumbia and A. Chériti, "Modeling and control of induction generator applied to variable speed Wind Energy Systems Conversion," in *Proc. IEEE Electrical Power and Energy Conference,* London, ON, Canada, Oct 26-28, 2015, pp. 314-319.

[16] B. Bechir, B. Faouzi, and M. Gasmi, "Wind energy conversion system with full-scale power converter and squirrel cage induction generator," *Int. J. Physical Sciences,* vol. 7, no. 46, pp. 6093-6104, Dec. 2012.

[17] Endusa Billy Muhando *et al.*, "LQG Design for Megawatt-Class WECS With DFIG Based on Functional Models' Fidelity Prerequisites," *IEEE Transaction on Energy Conversion*, vol. 24, no. 4, pp. 893-904, Dec. 2009.

[18] P. M. Anderson and Anjan Bose, "Stability Simulation of Wind Turbine Systems," *IEEE Transaction on Power Apparatus and Systems*, vol. 102, no. 12, pp. 3791-3795, Dec. 1983.

[19] V. Akhmatova, H. Knudsena, and A. H. Nielsenb, "Advanced simulation of windmills in the electric power supply," *Int. J. Elect. Power Energy Sys.,* vol. 22, no. 6, pp. 421–434, Aug. 2004.

[20] S. M. Muyeen, T. Murata, and J. Tamura, *Stability Augmentation of a Grid-Connected Wind Farm*. London, UK: Springer-Verlag, 2008.

[21] J. G. Slootweg, S. W. H. de Haan, H. Polinder, and W. L. Kling, "General Model for Representing Variable Speed Wind Turbines in Power System Dynamics Simulations," *IEEE Transaction on Power Systems*, vol. 18, no. 1, pp. 144-151, Feb. 2003.

[22] A. D Wright and L. J. Fingersh, "Advanced Control Design for Wind Turbines," National Renewable Energy Laboratory, Golden, CO, USA, NREL Rep. TP-500-42437, March. 2008.

[23] M. H. Hansen *et al*, "Control design for a pitch-regulated variable speed wind turbine," Rise National Laboratory, Roskilde, Denmark, RISO Rep. R-1500, January. 2005.

[24] P. C. Krause, Oleg Wasynczuk and Scott D. Sudhoff, *Analysis of Electric Machinery and Drive Systems*. NY, USA: John Wiley & Sons Inc., 2002.

[25] B. Wu, Y. Lang, N. Zargari and S. Kouro, *Power Conversion and Control of Wind Energy Systems*. NJ, USA: John Wiley & Sons, 2011.

[26] Z. Wu, "An investigation of dual stator winding induction machines," Doctoral dissertation, Tennessee Tech University, Cookeville, TN, USA, December. 206.

[27] F. Blaabjerg, R. Teodorescu, M. Liserre and A. V. Timbus, "Overview of Control and Grid Synchronization for Distributed Power Generation Systems," *IEEE Transaction on Industrial Electronics*, vol. 53, no. 5, pp. 1398-1409, October. 2006.

Adaptive Balancing by Reactive Compensators of Three-Phase Linear Loads Supplied by Nonsinusoidal Voltage from Four-Wire Lines

J. Uday Bhaskar, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, j.udaybhaskar1@gmail.com*

Ajanta Priyadarshinee, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ajantapriyadarshinee10@gmail.com*

Srichandan Subhrajit Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, ss.sahoo93@outlook.com*

Pinaki Prasanna, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, pinakiprasanna91@gmail.com*

Abstract: A method of the design of an adaptive balancing reactive compensator in four-wire systems with linear loads and nonsinusoidal voltage is described in this article. The method of compensation is founded on the Currents' Physical Components (CPC) – based power theory of three-phase systems with nonsinusoidal voltages and currents. The compensator is built of two sub-compensators of Y and Δ structure, respectively. The Y compensator reduces the reactive current and the zero sequence symmetrical component of the unbalanced current. The Δ compensator reduces the negative sequence symmetrical component of the unbalanced current. The positive sequence symmetrical component of the unbalanced current and the scattered current remain uncompensated. It is because shunt reactive compensators do not have any capability for that. Thyristor Switched Inductors (TSIs) enable the susceptance control of the compensator branches, referred to in the article as Thyristor Controlled Susceptance (TCS) branches. Periodic switching of thyristors in these branches causes the generation of harmonic currents, in particular the third-order harmonic. Moreover, in the presence of the supply voltage harmonics, a resonance of the equivalent capacitance of the compensator with the distribution system inductance can occur. These two harmful phenomena in the compensator suggested were reduced by the selection of a special structure of the TCS branches and their *LC* parameters. The presented method of the adaptive compensator synthesis was verified in the article with a numerical example and results of computer modeling of the load with an adaptive compensator.

Keywords: Asymmetrical Systems, CPC, Currents' Physical Components, Unbalanced Loads, Power Definitions

1. Introduction

Degradation of the power factor in large manufacturing plants is a combined effect of the reactive power, the load imbalance, and harmonic currents generated by nonlinear or periodically switched loads. Electrical loads in such manufacturing plants are both three- and single-phase ones. The load imbalance is caused by the presence of single-phase loads in such plants. Such imbalance causes that an unbalanced current occurs in the supply lines of such plants, causing asymmetry of the distribution voltage. Therefore, the distribution system has to be built as a three-phase system with a neutral conductor. Consequently, a compensator needed for the power factor improvement should have the capability of compensating not only the reactive and unbalanced currents but also the neutral conductor‘s current. It should have, moreover, adaptive property.

A low power factor and the current asymmetry have economic and technical consequences, Therefore, they are the subject of long-lasting studies, initiated by Steinmetz in 1917 [1].

This research on methods of compensation is continued now. It includes studies on compensation with static reactive LC compensators, with thyristor controlled compensators and with hybrid compensators, composed of a static LC compensator and thyristor-controlled reactors [2-6]. Taking into account the very high power of manufacturing plants, switching compensators, SC, commonly known as active power filters, built of transistors, may not have sufficient power for compensation of such very high power loads.

This article is the next step in studies on compensation in

three-phase systems. These studies started with developing a power theory of three-phase, three-wire systems with nonsinusoidal voltages and currents, known now as the Currents' Physical Components (CPC) – based power theory (PT) [7]. This theory provided fundamentals for developing a method of design of reactive balancing compensators operating in the presence of the supply voltage distortion [8]. The CPC - based PT was extended for three-phase systems with a neutral conductor and nonsinusoidal supply voltage [9]. A method of synthesis of balancing reactive compensators for such systems was also developed by Czarnrcki [10]. Compensators as discussed in the above referenced articles were fixed-parameters devices, however. An approach to conversion into adaptive devices was suggested by Czarnrcki and Almousa in [11]. This article is a direct continuation of it, aimed at improvement of the previously obtained results. Since this article is built upon these results, it would be recommended that the reader, if needed, is acquainted with those details and results [7-11].

2. The Approach to the Compensator Development

The compensator under the development has to be controlled in real-time so that the time-consuming optimization approach to its design has to be excluded. Its parameters have to be specified by algebraic expressions. The method suggested includes the following five steps.

1. The LC parameters of a static reactive compensator, composed Δ and Y sub-compensators are calculated, based on the original method developed in the frame of the CPC – based power theory.
2. The complexity of the compensator branches is reduced to branches built of no more than two reactive elements by a branch.
3. The range of change of the reactive power and the load imbalance, as well as the voltage harmonics are evaluated and the range of changes of the compensator susceptances are calculated.
4. Branches of the static LC compensator are replaced by branches with thyristor-controlled susceptance (TCS).
5. A look-up table, that enables the selection of thyristor firing angle for needed susceptance of each of six TCS of the compensator is created.

The presented method of compensation was verified by a computer simulation (Matlab-Simulink) but not by a lab experiment. This was because there are major limitations in scaling down high-power equipment to low-power devices in a lab environment. Nonlinear devices, such as thyristors, cannot be scaled down to different voltages. Also, high power inductors have the *q*-factor, meaning the ratio of the resistance to reactance, of the order of 100 or above. Low power inductors, used in labs, have the *q*-factor at the level of only 10 or even below. Consequently, a lab model can have properties substantially different than the original system and the adaptive compensator.

3. A Rationale of the Approach

The power factor of electrical loads degrades not only because of the reactive power but also because of the load imbalance. The supply voltage and current harmonics could also contribute to its degradation.

The first balancing compensator, but only for a sinusoidal supply voltage, was developed by Steinmetz, and this compensator is known now as a Steinmetz circuit [1, 12]. Some other approaches to load balancing were also suggested by Mayer and Kropik [13]. To this moment, the major obstacles in the development of the reactive compensators were theoretical, mainly caused by a controversy as to powers in three-phase systems with unbalanced loads and nonsinusoidal supply voltage [14-17, 24]. Eventually, the fundamentals for load balancing at nonsinusoidal voltage were developed in the frame of the current orthogonal components, later referred to as the CPC – based power theory [7, 17-18].

The load power, power factor, and load imbalance usually change in time, so the compensators should have the adaptive property. Adaptive compensators can be built as reactive compensators, with thyristors as the controlling devices, or as switching compensators (SCs), known commonly as active power filters, which use power transistors for the compensator control. When the load power is in the range of hundreds MVA, as this is common for large manufacturing plants, in particular, metallurgic ones, the power of SCs could not be sufficient for their compensation [25]. Reactive compensation could be the only option.

There are methods of reactive compensator design based on optimization procedures but a lot of computation could be needed for such procedures. Having in mind adaptive, in real-time compensation, the method of compensator's parameters calculation from direct mathematical expressions, rather than from optimization procedures seems to be more appropriate.

Just such a method provides the CPC – based power theory, which interprets power-related phenomena in electrical systems but also creates fundamentals for their compensation [17, 18]. The CPC-based method of reactive compensation specifies the theoretical limits of such compensation and the LC parameters of a fixed-parameters compensator. It can be next converted to an adaptive device.

As was suggested by Steeper and Stratford, a thyristor switched inductor (TSI) with a shunt capacitor, shown in Figure 1, can be used for that [19-20].

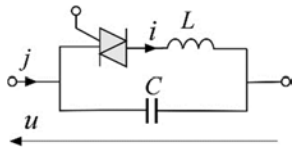


Figure 1. *A thyristor-switched inductor with a shunt capacitor.*

The TSI is a source of current harmonics of the level dependent on the thyristor firing angle. The harmonic of the third order is usually the dominating one. The adaptive compensators are built mainly as compensators of only the

reactive power, however, and consequently, they are symmetrical devices, with all thyristors switched at the same angle. Due to this symmetry, the third-order current harmonics do not leave the compensator to the compensated system. When the compensator is used as a balancing device, then the firing angles are different and the third-order harmonics in the compensator branches do not cancel mutually in the compensator but disturb the compensating system. Moreover, in the presence of harmonics in the supply voltage, the resonance of the supply system inductance with the compensator capacitance C can occur, causing substantial harmonic distortion in the compensated system. Countermeasures for that have to be provided. A TSI has to be integrated with resonant filters for that. Such filters should prevent, moreover, the system against the resonance of the compensator's equivalent capacitance with the supply system inductance.

Aggregates of single-phase computer-like loads as well as lightning, instrumentation equipment, and electrical transportation are the main cause of the load imbalance in manufacturing plants. Because a neutral conductor is needed for the supply of such loads, distribution systems in manufacturing plants are built as four-wire systems, shown in Figure 2.

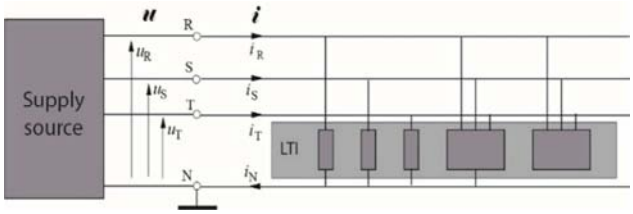


Figure 2. *A structure of a three-phase, four-wire system.*

Due to nonlinearity or periodic switching, the loads can generate current harmonics. Reactive compensators do not have any capability of compensating them, however. The load in Figure 2 is regarded therefore as a linear load.

Harmonic currents can be reduced, after the load for the fundamental frequency is compensated, by an additional switching compensator, known usually as an active power filter [21], of reduced power, because it does not have to compensate the harmful components of the supply currents' fundamental harmonic.

4. Currents' Physical Components

A distorted voltage in three-phase systems can be approximated by a sum of dominating harmonics of the order n a set N, expressed [7, 9] as a three-phase vector:

$$\boldsymbol{u}(t) = \sum_{n\in N} \boldsymbol{u}_n(t) = \sqrt{2}\mathrm{Re}\sum_{n\in N}\begin{bmatrix} U_{\mathrm{R}n} \\ U_{\mathrm{S}n} \\ U_{\mathrm{T}n} \end{bmatrix} e^{jn\omega_1 t} = \sqrt{2}\mathrm{Re}\sum_{n\in N} \boldsymbol{U}_n e^{jn\omega_1 t} \quad (1)$$

where U_{Ln} is the complex rms (crms) value of the n-th order supply voltage harmonic at L terminal, L = {R, S T}. Similarly, the load current can be approximated by the three-phase current vector

$$\boldsymbol{i}(t) = \sum_{n\in N} \boldsymbol{i}_n(t) = \sqrt{2}\mathrm{Re}\sum_{n\in N}\begin{bmatrix} \boldsymbol{I}_{\mathrm{R}n} \\ \boldsymbol{I}_{\mathrm{S}n} \\ \boldsymbol{I}_{\mathrm{T}n} \end{bmatrix} e^{jn\omega_1 t} = \sqrt{2}\mathrm{Re}\sum_{n\in N} \boldsymbol{I}_n e^{jn\omega_1 t}. \quad (2)$$

where I_{Ln} is the crms value of the n-th order load current harmonic at L terminal. The load current can be decomposed, according to the Currents' Physical Components CPC) – based power theory, into six physical components [8]

$$\boldsymbol{i} = \boldsymbol{i}_{\mathrm{a}} + \boldsymbol{i}_{\mathrm{s}} + \boldsymbol{i}_{\mathrm{r}} + \boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}} + \boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}} + \boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}}. \quad (3)$$

In this decomposition

$$\boldsymbol{i}_{\mathrm{a}} = \begin{bmatrix} i_{\mathrm{Ra}} \\ i_{\mathrm{Sa}} \\ i_{\mathrm{Ta}} \end{bmatrix} = G_{\mathrm{e}}\boldsymbol{u} = \sqrt{2}\ \mathrm{Re}\sum_{n\in N} G_{\mathrm{e}}\boldsymbol{U}_n e^{jn\omega_1 t}, \quad G_{\mathrm{e}} = \frac{P}{\|\boldsymbol{u}\|^2} \quad (4)$$

is the active current. It is the current component needed to supply the load at voltage $\boldsymbol{u}$ with the active power P. The symbol $\|\boldsymbol{u}\|$ denotes the three-phase rms value of the supply voltage, defined [7] for a three-phase quantity $\boldsymbol{x}(t)$ as

$$\|\boldsymbol{x}\| = \sqrt{\|x_{\mathrm{R}}\|^2 + \|x_{\mathrm{S}}\|^2 + \|x_{\mathrm{T}}\|^2}. \quad (5)$$

The component $\boldsymbol{i}_{\mathrm{s}}$ in decomposition (3) is the scattered current. It is defined as

$$\boldsymbol{i}_{\mathrm{s}} = \sqrt{2}\ \mathrm{Re}\sum_{n\in N}(G_{\mathrm{e}n} - G_{\mathrm{e}})\boldsymbol{U}_n e^{jn\omega_1 t}, \quad G_{\mathrm{e}n} = \frac{P_n}{\|\boldsymbol{u}_n\|^2} \quad (6)$$

is the load equivalent conductance for the n-th order harmonic.

The symbol P_n in formula (6) denotes the active power of the n-th order harmonic and $\|\boldsymbol{u}_n\|$ its three-phase rms value.

The component $\boldsymbol{i}_{\mathrm{r}}$ in decomposition (3) is the reactive current. It is defined as

$$\boldsymbol{i}_{\mathrm{r}} = \sqrt{2}\ \mathrm{Re}\sum_{n\in N} jB_{\mathrm{e}n}\boldsymbol{U}_n e^{jn\omega_1 t}, \quad B_{\mathrm{e}n} = -\frac{Q_n}{\|\boldsymbol{u}_n\|^2} \quad (7)$$

The three components in decomposition (3) with upper indices p, n, and z,

$$\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}} + \boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}} + \boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}} = \boldsymbol{i}_{\mathrm{u}} \quad (8)$$

stand for the positive, negative, and the zero sequence symmetrical components of the unbalanced current $\boldsymbol{i}_{\mathrm{u}}$ of the load. With symbols $\mathbf{1}^{\mathrm{p}}$, $\mathbf{1}^{\mathrm{n}}$, and $\mathbf{1}^{\mathrm{z}}$, explained in Figure 3

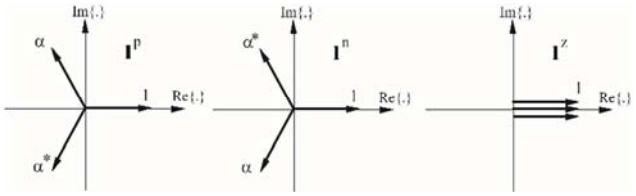


Figure 3. *Unit symmetrical vectors.*

the symmetrical components of the unbalanced current can be expressed as follows [8]

$$\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}} = \sum_{n\in N} \boldsymbol{i}_{\mathrm{u}n}^{\mathrm{p}} = \sqrt{2}\,\mathrm{Re}\sum_{n\in N} \boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{p}} \mathbf{1}^{\mathrm{p}} \boldsymbol{U}_{\mathrm{R}n} e^{jn\omega_1 t} \tag{9}$$

$$\boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}} = \sum_{n\in N} \boldsymbol{i}_{\mathrm{u}n}^{\mathrm{n}} = \sqrt{2}\,\mathrm{Re}\sum_{n\in N} \boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{n}} \mathbf{1}^{\mathrm{n}} \boldsymbol{U}_{\mathrm{R}n} e^{jn\omega_1 t} \tag{10}$$

$$\boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}} = \sum_{n\in N} \boldsymbol{i}_{\mathrm{u}n}^{\mathrm{z}} = \sqrt{2}\,\mathrm{Re}\sum_{n\in N} \boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} \mathbf{1}^{\mathrm{z}} \boldsymbol{U}_{\mathrm{R}n} e^{jn\omega_1 t} \tag{11}$$

Admittances $\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{p}}$, $\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{n}}$, and $\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}}$ in last formulas are unba- lanced admittances of the positive, negative, and the zero sequence of the load. All these admittances can be calculated, if the line-to-neutral equivalent admittances for harmonic frequencies $Y_{\mathrm{R}n}$, $Y_{\mathrm{S}n}$, and $Y_{\mathrm{T}n}$, of the equivalent load, shown in Figure 4, are known. These admittances can be obtained by a measurement of the crms values of the voltage and current harmonics at the load terminals, namely

$$\boldsymbol{Y}_{Ln} = G_{Ln} + jB_{Ln} = \frac{\boldsymbol{I}_{Ln}}{\boldsymbol{U}_{Ln}}, \qquad L = \mathrm{R, S\ or\ T}. \tag{12}$$

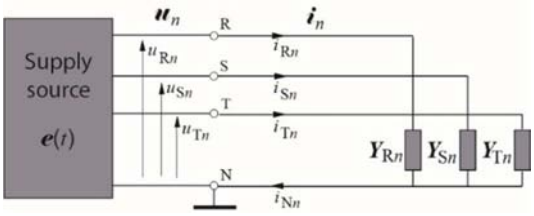


Figure 4. *An equivalent circuit of the LTI load.*

With these line-to-neutral equivalent admittances $Y_{\mathrm{R}n}$, $Y_{\mathrm{S}n}$, and $Y_{\mathrm{T}n}$, for harmonic frequencies, the unbalanced admittances of particular symmetrical sequences can be obtained, namely

$$\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{p}} = \frac{1}{3}[(\boldsymbol{Y}_{\mathrm{R}n} + \alpha\beta\boldsymbol{Y}_{\mathrm{S}n} + \alpha^*\beta^*\boldsymbol{Y}_{\mathrm{T}n}) - \boldsymbol{Y}_{\mathrm{e}n}(1 + \alpha\beta + \alpha^*\beta^*)] \tag{13}$$

where

$$\boldsymbol{Y}_{\mathrm{e}n} = G_{\mathrm{e}n} + jB_{\mathrm{e}n} = \frac{1}{3}(\boldsymbol{Y}_{\mathrm{R}n} + \boldsymbol{Y}_{\mathrm{S}n} + \boldsymbol{Y}_{\mathrm{T}n}) \tag{14}$$

$$\beta \overset{\mathrm{df}}{=} (\alpha^*)^n = \begin{cases} 1, & \text{for} \quad n = 3k \\ \alpha^*, & \text{for} \quad n = 3k+1, \\ \alpha, & \text{for} \quad n = 3k-1 \end{cases} \qquad \alpha = 1e^{j2\pi/3} \tag{15}$$

for the negative sequence harmonics

$$\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{n}} = \frac{1}{3}[(\boldsymbol{Y}_{\mathrm{R}n} + \alpha^*\beta\boldsymbol{Y}_{\mathrm{S}n} + \alpha\beta^*\boldsymbol{Y}_{\mathrm{T}n}) - \boldsymbol{Y}_{\mathrm{e}n}(1 + \alpha^*\beta + \alpha\beta^*)] \tag{16}$$

and for the zero sequence harmonics

$$\boldsymbol{Y}_{\mathrm{u}n}^{z} = \frac{1}{3}[(\boldsymbol{Y}_{\mathrm{R}n} + \beta\boldsymbol{Y}_{\mathrm{S}n} + \beta^*\boldsymbol{Y}_{\mathrm{T}n}) - \boldsymbol{Y}_{\mathrm{e}n}(1 + \beta + \beta^*)]. \tag{17}$$

The Currents' Physical Components in decomposition (3), are mutually orthogonal [9]. Thus, they contribute to the three-phase rms value of the supply current independently of each other, namely

$$\|\boldsymbol{i}\|^2 = \|\boldsymbol{i}_{\mathrm{a}}\|^2 + \|\boldsymbol{i}_{\mathrm{s}}\|^2 + \|\boldsymbol{i}_{\mathrm{r}}\|^2 + \|\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}}\|^2 + \|\boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}}\|^2 + \|\boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}}\|^2. \tag{18}$$

Only the active current $\boldsymbol{i}_{\mathrm{a}}$ is necessary for the permanent transmission of energy to the load. The remaining ones increase the supply current three-phase rms value. They reduce the load power factor thus, they are harmful.

5. Reactive Compensation

The power factor can be improved by the reduction of the harmful components of the supply current. the scattered current $\boldsymbol{i}_{\mathrm{s}}$ cannot be compensated by a shunt reactive compensator [7]. Such a compensator, shown in Figure 5, has to be composed of two sub-compensators with Y and Δ structures, which can reduce only the reactive and unbalanced currents [10]. The compensator in Figure 5 is specified by susceptances of its branches for the n^{th}-order harmonic.

As assumed in this article, the load is linear, similarly to the compensator, so that the whole system satisfies the Superposition Principle, so that, in the presence of the supply voltage distortion, it can be analyzed harmonic-by-harmonic.

Compensation of the zero-sequence symmetrical component of the unbalanced current $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}}$ is possible only when the compensator provides a pass for such a current. Thus, it has to be configured in Y. It compensates this current entirely on the condition that suceptance of the Y sub-compensator branches for harmonic frequencies, $T_{\mathrm{R}n}$, $T_{\mathrm{S}n}$, and $T_{\mathrm{T}n}$, satisfy for each n from the set N, the equation

$$\frac{1}{3} j(T_{\mathrm{R}n} + \alpha^* T_{\mathrm{S}n} + \alpha T_{\mathrm{T}n}) + \boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} = 0 \tag{19}$$

This equation has to be satisfied for the real and the imaginary parts of the complex coefficients of this equation thus, it has an infinite number of solutions with regard to three unknown susceptances of the compensator branches. It has only one solution if this Y sub-compensator has to compensate entirely also the reactive current $\boldsymbol{i}_{\mathrm{r}}$, so that its susceptances have to satisfy the additional equation

$$\frac{1}{3}(T_{\mathrm{R}n} + T_{\mathrm{S}n} + T_{\mathrm{T}n}) + B_{\mathrm{e}n} = 0. \tag{20}$$

Equations (19) and (20) result in the susceptances for harmo- nic frequencies of the Y sub-compensator branches, namely

$$\begin{aligned} T_{\mathrm{R}n} &= -2\,\mathrm{Im}\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} - B_{\mathrm{e}n} \\ T_{\mathrm{S}n} &= -\sqrt{3}\,\mathrm{Re}\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} + \mathrm{Im}\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} - B_{\mathrm{e}n}. \\ T_{\mathrm{T}n} &= \sqrt{3}\,\mathrm{Re}\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} + \mathrm{Im}\boldsymbol{Y}_{\mathrm{u}n}^{\mathrm{z}} - B_{\mathrm{e}n} \end{aligned} \tag{21}$$

Such a compensator eliminates components $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}}$ and $\boldsymbol{i}_{\mathrm{r}}$ from the supply current entirely, while the remaining ones are not changed. A second sub-compensator of Δ structure can be used for their reduction. Since the scattered current $\boldsymbol{i}_{\mathrm{s}}$ cannot be compensated by a shunt reactive compensator, only two

symmetrical components of the unbalanced current, namely $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}}$ and $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}}$ remain for compensation.

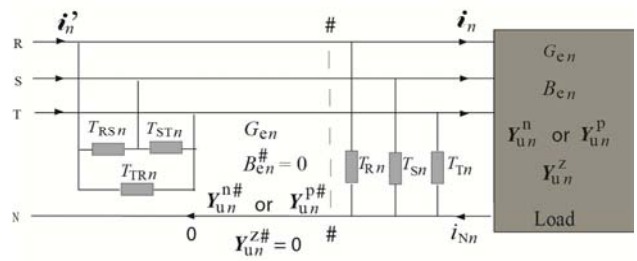


Figure 5. *Load with a reactive compensator for the n^{th}- order harmonic of the positive or negative sequence.*

The sub-compensator of Δ-structure, connected at the supply terminals, as shown in Figure 5, has to compensate the load with parameters modified by the sub-compensator of the Y structure.

The partially compensated load, as seen from the cross-section #–#, has the equivalent susceptance for harmonic of the order n from the set N, $B_{en}^{\#}$ and the unbalanced admittance of the zero-sequence $\boldsymbol{Y}_{un}^{z\#}$ reduced to zero, while two remaining ones are changed to $\boldsymbol{Y}_{un}^{p\#}$ and $\boldsymbol{Y}_{un}^{n\#}$, respectively.

A sub-compensator of the Δ structure compensates the unbalanced current of the negative sequence [10] on the condition that

$$j(T_{STn}+\alpha T_{TRn}+\alpha^{*}T_{RSn})+\boldsymbol{Y}_{un}^{n\#}=0\,. \tag{22}$$

and the positive sequence on the condition that

$$j(T_{STn}+\alpha^{*}T_{TRn}+\alpha T_{RSn})+\boldsymbol{Y}_{un}^{p\#}=0\,. \tag{23}$$

Because of complex coefficients, equations (22) and (23) stand for four equations. They cannot be satisfied for three unknown branch susceptances. Therefore, one of the equations (22) and (23) has to be abandoned, which means that only one symmetrical component of the unbalanced current, $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}}$ or $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}}$, can be compensated.

The positive sequence component of the unbalanced current $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}}$ occurs only [9] if the supply voltage has harmonics of the negative sequence, $n = 2, 5\ldots n = (3k–1)$, or the zero sequence, $n = 3, 6\ldots n = 3k$. The fundamental harmonic dominates usually in the negative sequence component $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}}$. Therefore, the component $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}}$ is usually much smaller than the component $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{n}}$ and consequently, the last one can be left uncompensated. It means that the branch susceptances of the sub-compensator of Δ structure should satisfy only equation (22). Since this sub-compensator should not load the supply source with the reactive current, these susceptances have to satisfy moreover, for each n from the set N, the condition

$$T_{STn}+T_{TRn}+T_{RSn}=0\,. \tag{24}$$

As to equation (22), we have to remember that the sub-compensator of Y structure modifies the unbalanced admittance of the negative sequence as seen from the supply source [10]. It is equal to

$$\boldsymbol{Y}_{un}^{n\#}=\boldsymbol{Y}_{un}^{z*}+\boldsymbol{Y}_{un}^{n}\,. \tag{25}$$

With this admittance, equations (22) and (24) results in the sub-compensator susceptances

$$\begin{aligned}T_{RSn}&=\frac{1}{3}(\sqrt{3}\,\mathrm{Re}\,\boldsymbol{Y}_{un}^{n\#}-\mathrm{Im}\,\boldsymbol{Y}_{un}^{n\#})\\ T_{STn}&=\frac{1}{3}(2\,\mathrm{Im}\,\boldsymbol{Y}_{un}^{n\#})\\ T_{TRn}&=\frac{1}{3}(-\sqrt{3}\,\mathrm{Re}\,\boldsymbol{Y}_{un}^{n\#}-\mathrm{Im}\,\boldsymbol{Y}_{un}^{n\#})\end{aligned}\,. \tag{26}$$

6. Compensator Complexity Reduction

Formulas (21) and (26) provide susceptances of all branches of a reactive compensator for harmonic frequencies needed for the entire compensation of the reactive as well as the zero and the negative symmetrical components of the unbalanced current. The structure and the *LC* parameters of such branches can be found using well-developed methods of reactance one-ports synthesis [22]. Unfortunately, with increasing number of supply voltage harmonics, the number of inductors and capacitors needed for the compensator construction increases. Approximately, one extra inductor and one extra capacitor per branch of the compensator are needed for each extra voltage harmonic. For example, if $N = \{1, 3, 5, 7\}$, then at least 40 reactance elements might be needed for the compensator construction. It would be too complex and consequently, too expensive to have a technical value.

The number of elements needed for the compensator construction can be reduced if the requirement of the entire compensation of the supply current components $\boldsymbol{i}_{\mathrm{r}}$, $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{z}}$, and $\boldsymbol{i}_{\mathrm{u}}^{\mathrm{p}}$, is abandoned for the only reduction of their three-phase rms value.

The most simple compensator has only one reactive element, capacitor or inductor, by a branch. Since capacitors in the compensator can result in a series resonance with the supply source inductance, purely capacitive branches are not acceptable. Therefore, the reduced complexity compensator cannot have branches other than those, shown in Figure 6.

Figure 6. *Acceptable branches of a reduced complexity compensator.*

The branch susceptances of the reduced complexity compensator are denoted by D_n, to distinguish them from the those calculated from formulas (21) and (26). They have for harmonic frequencies the values

$$D_n=-\frac{1}{n\omega_1 L}\quad\text{or}\quad D_n=\frac{n\omega_1 C}{1-n^2\omega_1^2 LC}\,. \tag{27}$$

A compensator of the reduced complexity minimizes the

supply current three-phase rms value on the condition, that the suceptance D_{kn} of each branch k is selected such that the following expression is minimized [10]

$$\sum_{n\in N}(T_{kn}-D_{kn})^2 U_{kn}^2 = \sum_{n\in N} A_{kn}^2 = \mathrm{Min}. \tag{28}$$

where susceptances T_{kn} of these branches are given by formulae (21) and (26), respectively.

Selection of the specific branch, meaning one of the two shown in Figure 6, can be based on the sign of the calculated susceptance for the fundamental frequency, T_{k1}. It is because the rms value of the supply voltage fundamental harmonic U_{k1} is usually much higher than this value for other harmonics. Consequently, the term

$$A_{k1} = (T_{k1}-D_{k1})^2 U_{k1}^2 \tag{29}$$

in formula (28) is much higher than such terms for other harmonics. To have the value of A_{k1} as close to zero as possible, branch k should be selected in such a way that its susceptance D_{k1} has the same sign as the susceptance T_{k1}. Thus, when T_{k1} is negative, then a purely inductive branch should be chosen. Its inductance should minimize the term

$$\sum_{n\in N}(T_{kn}+\frac{1}{n\omega_1 L_k})^2 U_{kn}^2 = \mathrm{Min}. \tag{30}$$

the optimum value of this inductance is

$$L_{k,\mathrm{opt}} = -\frac{1}{\omega_1}\frac{\sum_{n\in N}\frac{1}{n^2}U_{kn}^2}{\sum_{n\in N}T_{kn}\frac{1}{n}U_{kn}^2}. \tag{31}$$

When T_{k1} is positive, then LC branch should be chosen such that its LC parameters should minimize the term

$$\sum_{n\in N}(T_{kn}-\frac{n\omega_1 C_k}{1-n^2\omega_1^2 L_k C_k})^2 U_{kn}^2 = \mathrm{Min}. \tag{32}$$

The left side of this expression does not have minimum for finite values of the inductance L_k, however. Thus, any value can be chosen. The product L_kC_k specifies the approximate value of the frequency of the branch resonance. Therefore, inductance L_k should be selected such that this resonance will not occur for harmonic frequencies. It can be done, however, only in an iterative process, because the inductance L_k affects the capacitance C_k. Calculation of this capacitance is also possible only by an iterative process because expression (32) cannot be rearranged into an explicit formula with regard to this capacitance. It can be calculated as a limit of a sequence of capacitances obtained iteratively, namely

$$C_{k,s+1} = \frac{\sum_{n\in N}\frac{T_{kn}nU_{kn}^2}{1-n^2\omega_1^2 L_k C_{k,s}}}{\omega_1\sum_{n\in N}\frac{n^2U_{kn}^2}{(1-n^2\omega_1^2 L_k C_{k,s})^2}} \tag{33}$$

The method of synthesis of a compensator with reduced complexity was illustrated for the load shown in Figure 7, supplied with a symmetrical voltage of the fundamental harmonic rms value $U_1 = 240$ V [10]. It was assumed that the supply voltage was distorted by the 3rd, 5th, and 7th order harmonics of relative rms value $U_3 = 2\%U_1$, $U_5 = 3\%U_1$, and $U_7 = 1.5\%U_1$. There are also shown in Figure 7 the three-phase rms values of the load current physical components

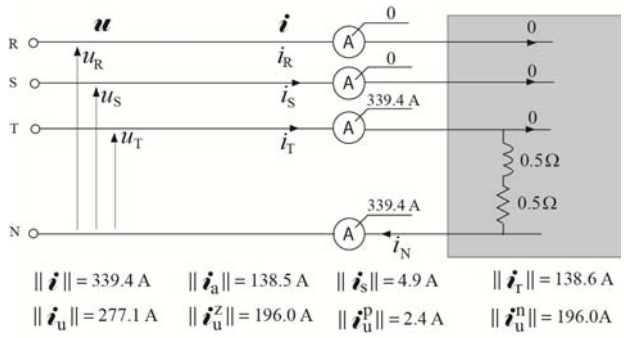


Figure 7. *An example of an unbalanced load and results of its analysis.*

The parameters of the reduced complexity compensator were calculated for the fundamental frequency normalized to $\omega_1 = 1$ rad/s, assuming that the resonant frequency of LC branches is 2.5 rad/s [10].

Table 1. *LC parameters of a reduced complexity compensator.*

	Line:	R	S	T	RS	ST	TR
L	mH	1730	770	444	0	2600	1155
C	mF	0	399	691	0	0	266

The results of compensation are shown in Figure 8. The power factor is improved by the compensator of the reduced complexity from $\lambda = 0.408$ to $\lambda = 0.994$.

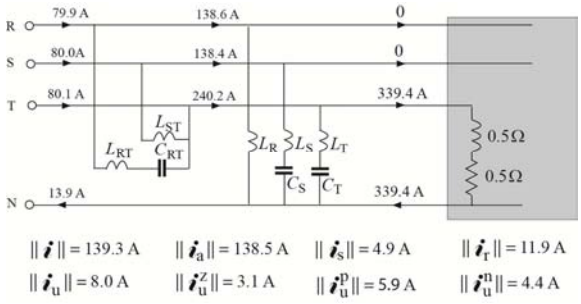


Figure 8. *Results of compensation with a compensator of reduced complexity.*

7. Adaptive Compensator

The compensator as developed above improves the power factor λ of the supply source when its load has fixed parameters. When these parameters are not constant, such a compensator is losing effectiveness. An adaptive compensator is needed instead.

A reactive compensator has adaptive properties if it can be adjusted to changes in the load power of individual lines. This can be done by switches or by using reactive elements with controllable parameters. Saturation of the ferromagnetic core

of an inductor enables a change of its inductance. A thyristor-switched inductor, shown in Figure 9a, is another and commonly used device of this category [19].

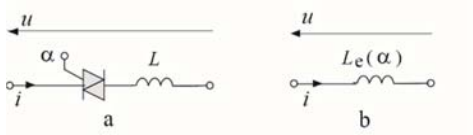


Figure 9. *A thyristor switched inductor (a) and its equivalent inductance (b).*

When it is connected in parallel with a capacitor, it serves as an adaptive compensator of the reactive power in systems with a sinusoidal supply voltage [19-20].

When the supply voltage is nonsinusoidal and the current harmonics generated by thyristors have to be reduced, the TSI is connected with a few reactive elements to shape the frequency properties of the compensator branches. Properties of such branches were studied by Czarnecki and Hsu [23]. Such branches, or reactive one-ports, will be referred to as thyristor-controlled susceptance (TCS) branches.

The current of a thyristor-switched inductor changes as shown in Figure 10. Symbol i_0 denotes current at $\alpha = 0$.

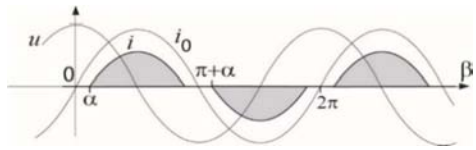


Figure 10. *The voltage and currents waveform of TSI.*

The equivalent admittance of the TSI for the fundamental frequency at a sinusoidal voltage at its terminal, expressed as a function of the firing angle α, is equal to

$$\boldsymbol{Y}_1 = \frac{\boldsymbol{I}_1}{\boldsymbol{U}_1} = jB_1 = (1 - \frac{2\alpha + \sin 2\alpha}{\pi}) \frac{1}{j\omega_1 L} = \frac{1}{j\omega_1 L_e(\alpha)} \qquad (34)$$

The current of thyristors is distorted from a sinusoidal waveform, so that TCS branches generate current harmonics. Usually, the 3rd order harmonic is the dominating one. When a compensator is used only for the reactive current compensation, it is built as a symmetrical device, usually in Δ structure. Each branch of such a compensator generates the 3rd order current harmonic of the same value and phase and consequently, it does not leave the Δ loop. This is no longer true in balancing compensators, which are unbalanced devices. They inject the 3rd order current harmonic into the compensated system, causing distortion. A parallel resonance of the TCS branch equivalent capacitance with the supply source inductance can occur as well.

The 3rd order current harmonic leaving a TCS can be reduced by an *LC* filter connected as shown in Figure 11, and tuned to the frequency of that harmonic.

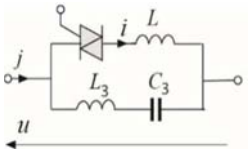


Figure 11. *A TCS branch with a filter of the 3rd order harmonic.*

8. Reduction Sensitivity to the Voltage Harmonics

The filter reduces that harmonic, unfortunately, it creates a short circuit path for the 3rd order harmonic in the supply voltage. To avoid it, an inductor denoted as L_0, can be added to the TCS branch as shown in Figure 12. It increases the impedance of the compensator as seen from the supply terminals for the 3rd voltage harmonic. It increases moreover this impedance for frequencies above the frequency of the 3rd order harmonic, where L_3C_3 branch has an inductive impedance.

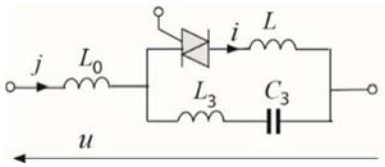


Figure 12. *A TSC branch with a series inductor L_0.*

The TCS branch is specified by four parameters, L, L_0, L_3, and C_3. There are only three conditions the TCS branch has to satisfy: the range of susceptance change, T_{min}, T_{max}, and resonant frequency of the L_3C_3 branch. Thus, one parameter can be selected at a designer's discretion [11].

After not being satisfied with harmonic distortion caused by the compensator, the authors concluded that there is one more phenomenon in the adaptive compensator that should be taken into account: namely, the voltage resonance of the whole TCS branch, at which its impedance approaches zero.

When the thyristor is in ON state, i.e., at firing angle $\alpha = 0$, then the susceptance $T(\omega)$ of such a branch changes with the frequency as shown in Figure 13.

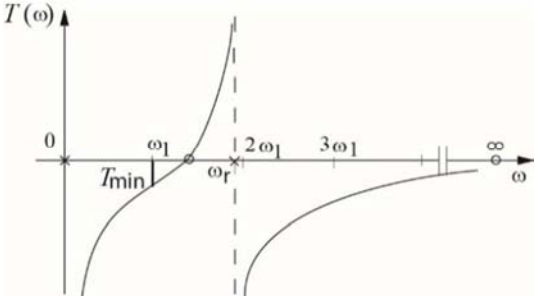


Figure 13. *Change of the TCS branch in the thyristor ON state.*

Its susceptance T for the fundamental harmonic has the minimum value, equal to

$$T_{\min} = \frac{9\omega_1^2 LC_3 - 8}{8\omega_1(L_0 + L) - 9\omega_1^3 L L_0 C_3}. \qquad (35)$$

When the thyristor is in OFF state, i.e., at $\alpha = 90°$, then the susceptance $T(\omega)$ of such a branch changes with frequency as shown in Figure 14. Its susceptance for the fundamental harmonic has the maximum value equal to

$$T_{\max} = \frac{9\omega_1 C_3}{8 - 9\omega_1^2 L_0 C_3}. \qquad (36)$$

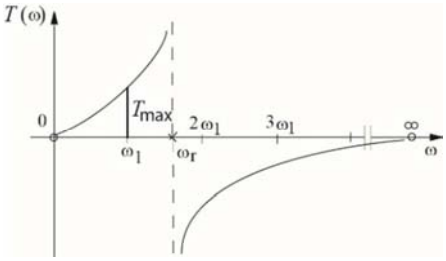


Figure 14. *Change of the TCS branch in the thyristor OFF state.*

At some frequency, denoted by ω_r, a voltage resonance of the whole TSC branch occurs. Its susceptance approaches infinity. Since the equivalent inductance of the TSI branch changes with the firing angle, the frequency ω_r changes as well. Its maximum value is in the thyristor ON state. Its relative value, referenced to the fundamental frequency, equal to

$$\frac{\omega_r}{\omega_1} = \Omega = \sqrt{8}\sqrt{\frac{L_3(L+L_0)}{L L_0}}\,. \tag{37}$$

To avoid resonance at the 2nd order harmonic, which can be present in the supply voltage, the parameters of the TCS branch should be select in such a way that the relative resonance frequency (37) is below 2.

The conditions (35), (36), (37), and the resonance frequency of the L_3C_3 branch, can be rearranged with regard to one of four parameters of the LSC branch. If inductance L_0 is such a parameter, then it has to satisfy the equation

$$a_3L_0^3 + a_2L_0^2 + a_1L_0 + a_0 = 0\,. \tag{38}$$

To compact symbols, let us denote $T_{\min} = T_a$, $T_{\max} = T_b$. With such symbols and the frequency ω_1 normalized to 1 rad/s, coefficients of eqn. (42) are

$$a_0 = 9 - \Omega^2 \tag{39}$$

$$a_1 = (27 - 11\Omega^2)T_b \tag{40}$$

$$a_2 = [2T_b(9 - 5\Omega^2) + 9T_a(1 - \Omega^2)]T_b \tag{41}$$

$$a_3 = 9(1 - \Omega^2)T_aT_b^2 \tag{42}$$

When eqn. (42) is solved, the parameters of the TCS branch can be expressed in terms of the inductance L_0 as follows

$$L_3 = (L_0 + 1/T_b)/8 \tag{43}$$

$$C_3 = 1/(9L_3) \tag{44}$$

$$L = [T_aT_bL_0^2 + (T_a + T_b)L_0 + 1]/[(T_b - T_a)]\,. \tag{45}$$

The needed range of the change of the branch's susceptance T depends, of course, on the load: its reactive power and possible level of imbalance. This relatively complex issue is, however, beyond the scope of this paper, which is to only demonstrate that adaptive balancing in four-wire systems in the presence of the supply voltage distortion is possible. Therefore, the circuit used in the numerical illustration before will be used again to illustrate an adaptive balancing. The adaptive compensator will be designed at the assumption that the supply voltage is identical as before, while the individual supply lines are loaded randomly but no more than to the degree as line T load in Figure 7.

The needed minimum and maximum values of the susceptance, $T_{\min}$, $T_{\max}$, of TCS branches of sub-compensators can be found having optimized LC values of the fixed-parameters compensator, previously calculated and compiled in Table 1. Thus, for the Y sub-compensator

$$T_{\min} = -\frac{1}{\omega_1 L_R} = -\frac{1}{1.730} = -0.578\ \text{S} \tag{46}$$

$$T_{\max} = \frac{1}{\frac{1}{\omega_1 C_T} - \omega_1 L_T} = \frac{1}{\frac{1}{0.691} - 0.444} = 0.997\ \text{S} \tag{47}$$

and for the Δ sub-compensator

$$T_{\min} = -\frac{1}{\omega_1 L_{ST}} = -\frac{1}{2.60} = -0.385\ \text{S} \tag{48}$$

$$T_{\max} = \frac{1}{\frac{1}{\omega_1 C_{TR}} - \omega_1 L_{TR}} = \frac{1}{\frac{1}{0.266} - 1.155} = 0.384\ \text{S}\,. \tag{49}$$

Assuming that $\Omega = 1.9$, coefficients of eqn. (42), calculated with (43) – (44), have the values compiled in Table 2.

Table 2. *Coefficients of the inductance L_0 equation.*

	a_0	a_1	a_2	a_3
Δ	5.39	– 4.88	– 0.004	1.34
Y	5.39	–12.66	8.98	13.49

Parameters of the Δ and Y sub-compensators TCS branches of the structure shown in Figure 12, calculated from formulas (47)–(49), are compiled in Table 3.

Table 3. *Parameters of the compensators' TCS branches.*

	L_0 [H]	L_3 [H]	C_3 [F]	L [H]
Δ	1.100	0.463	0.240	1.065
Y	0.370	0.172	0.640	0.683

The voltage and current at each TCS branch are in general, nonsinusoidal. Since the 3rd order harmonic is usually the dominating one in the distorted current of the TSI branch, the L_3C_3 filter reduces the harmonic distortion of the TCS branch current j substantially. Therefore, the TCS branch, as shown in Figure 12, can be approximated by an equivalent branch for the fundamental harmonic, as shown in Figure 15.

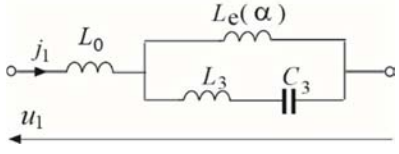


Figure 15. *An equivalent TCS branch for the fundamental harmonic.*

Its susceptance for the normalized fundamental frequency is

$$T = \frac{9\,L_e(\alpha)C_3 - 8}{8[L_0 + L_e(\alpha)] - 9\,L_e(\alpha)\,L_0\,C_3} = T(\alpha)\,. \qquad (50)$$

This formula, for a given value of the firing angle α, provides the branch susceptance T. It cannot be solved, however, with respect to the firing angle α. A look-up table, which for angles in the range from 0° to 90° specifies the susceptance T of TCS branches of both Y and Δ sub- compensators, is needed.

When the compensator has the structure and parameters as shown in Figure 8, then for the Y sub-compensator:

$$T_R = T_{min} = -\frac{1}{\omega_1 L_R} = -0.578\ \mathrm{S}; \qquad \alpha_R = 0^o \qquad (51)$$

$$T_S = \frac{1}{\frac{1}{\omega_1 C_S} - \omega_1 L_S} = 0.491\ \mathrm{S}; \qquad \alpha_S = 47.3^o \qquad (52)$$

$$T_T = T_{max} = \frac{1}{\frac{1}{\omega_1 C_T} - \omega_1 L_T} = 0.576\ \mathrm{S}; \quad \alpha_T = 90^o \qquad (53)$$

and for the Δ sub-compensator

$$T_{RS} = 0; \qquad \alpha_{RS} = 38^o \qquad (54)$$

$$T_{ST} = T_{min} = -\frac{1}{\omega_1 L_{ST}} = -0.385\ \mathrm{S}; \qquad \alpha_{ST} = 0^o \qquad (55)$$

$$T_{TR} = T_{max} = \frac{1}{\frac{1}{\omega_1 C_{TR}} - \omega_1 L_{TR}} = 0.384\ \mathrm{S}; \quad \alpha_{TR} = 90^o\,. \qquad (56)$$

The results of compensation are shown in Figure 16. These results confirm the possibility of an adaptive compensation of unbalanced linear loads supplied by a four-wire line in the presence of the supply voltage distortion. Similarly, as in the case of compensation by a fixed-parameters compensator of reduced complexity, some residual parts of the reactive and unbalanced currents remain.

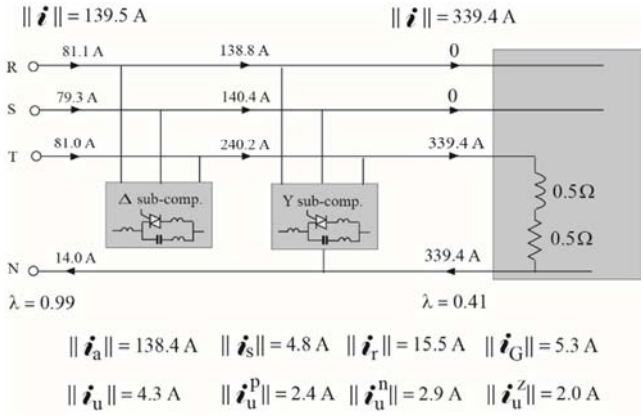


Figure 16. *Results of adaptive compensation.*

The waveform of the supply current in line R, after adaptive compensation, is shown in Figure 17. This current is in-phase with the supply voltage, but it is distorted. It is because the compensator does not compensate harmonics but some physical components of the supply current. All not compensated components, even the active current, are distorted.

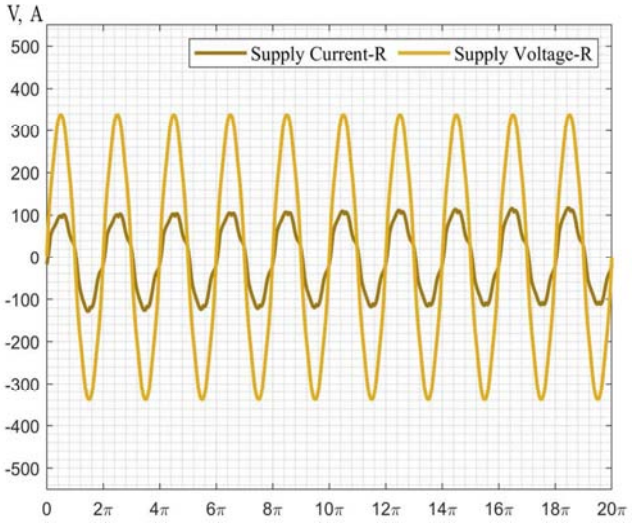


Figure 17. *Voltage and current waveforms in supply line R.*

The content of harmonics in the supply current is shown in Figure 18. It enables to compare this content when the compensator is designed based on an engineering intuition (blue) [11], with that based on the approach presented in this paper (red).

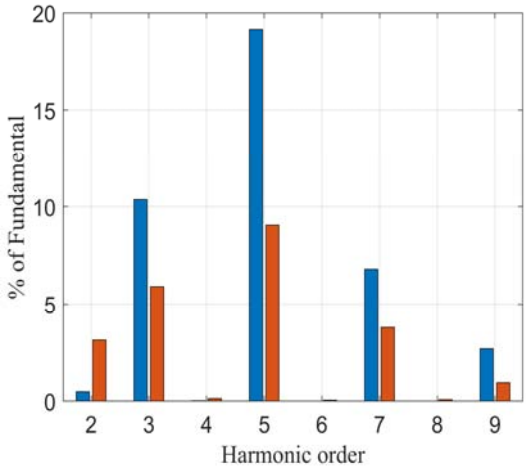


Figure 18. *The contents of harmonics in the supply current of a compensator designed by an engineering intuition [11] (blue) and in this paper (red).*

9. Compensator-Generated Harmonics

Distortion of the compensator current by thyristors means that apart from the supply voltage originated harmonics, also the compensator-originated harmonics can occur in the supply current, $\boldsymbol{i}_S$.

Let us denote the supply current in the system with the compensator but with removed thyristor branches, by $\boldsymbol{i}_{S0}$. The difference

$$\boldsymbol{i}_S - \boldsymbol{i}_{S0} = \boldsymbol{i}_G\,. \qquad (57)$$

approximates the compensator-generated harmonic current $\boldsymbol{i}_G$. The harmonics of the compensator-generated current $\boldsymbol{i}_G$

are created by thyristors switching. The effect of the voltage harmonics upon thyristors' switching is negligible so that harmonics of the currents $\boldsymbol{i}_{S0}$ and $\boldsymbol{i}_{G}$ are mutually random, thus these two currents are mutually orthogonal. Hence,

$$\|\boldsymbol{i}_{S}\|^2 = \|\boldsymbol{i}_{S0}\|^2 + \|\boldsymbol{i}_{G}\|^2 . \quad (58)$$

When the line-to-neutral magnitude of admittances for harmonics of the compensator without thyristors are denoted by Y with an apostrophe, then the three-phase rms value of such a compensator current can be expressed as

$$\|\boldsymbol{i}_{S0}\|^2 = \sum_{n\in N} \|\boldsymbol{i}_{S0n}\|^2 = \sum_{n\in N} [(Y'_{Rn} U_{Rn})^2 + (Y'_{Sn} U_{Sn})^2 + (Y'_{Rn} U_{Tn})^2] \quad (59)$$

and hence, the three-phase rms value of the compensator-generated current is

$$\|\boldsymbol{i}_{G}\| = \sqrt{\|\boldsymbol{i}_{S}\|^2 - \|\boldsymbol{i}_{S0}\|^2} \quad (60)$$

This value shown in Figure 16 was calculated just according to the above formula.

10. Conclusions

The paper shows that the presented method of synthesis of TCS branches of an adaptive compensator improves its performance in the presence of the supply voltage distortion. The adaptive compensator of unbalanced loads supplied with a nonsinusoidal voltage, effectively reduces the reactive and unbalanced currents, without any substantial distortion of the supply current. Nonetheless, this distortion still exists and it is not caused only by the thyristor switched inductors. This is because the active current, which is not the subject of compensation, reproduces the supply voltage distortion. Moreover, the scattered current is not affected by reactive compensators. Also, only two of three symmetrical components of the unbalanced current can be compensated by such a compensator. Despite that, the compensator is very effective in balancing even strongly unbalanced loads and in improving their power factor to almost unity value. Moreover, unlike switching compensators, its power is less confined.

References

[1] Ch. P. Steinmetz, Theory and calculation of electrical apparatur, McGraw-Hill Book Comp., New York, 1917.

[2] M. M. A. Aziz, at all, "LC compensator for power factor correction of non-linear loads", *IEEE Trans. on Power Delivery*, Vol. 19, No. 1, pp. 331-335, 2004.

[3] S.-J. Jeon, "Passive-component-based reactive power compensation in a non-sinusoidal multi-line system", *Electrical Engineering,* Vol. 102, pp. 1567-1577, 2020.

[4] D. Maiti, S. Mukhopadhyay, S. K. Biswas, "Three-phase thyristor controlled reactor using two sets of delta connected switches with low current harmonics", *IET Power Electronics*, DOI: 10.1049/iet-pel, 2019.

[5] S. Morello, T. J. Dionise, T. L. Mank, "Comprehensive analysis to specify a static var compensator for an arc furnace", *IEEE Trans. on Ind. Appl.,* Vol. 51, No. 6, pp. 1153-1160, 2015.

[6] A. Luo, Z. Shuai, W. Zhu, Z. J. Shen, "Combined system for harmonic suppression and reactive power compensation", *IEEE Trans. on Industrial Electronics*, Vol. 56, No. 2, pp. 418-420, 2009.

[7] L. S. Czarnecki, "Orthogonal decomposition of the current in a three-phase nonlinear symmetrical circuit with nonsinusoidal voltage", *IEEE Trans. on Instr. and Meas.,* Vol. M-37, No. 1, pp. 30-34, 1988.

[8] L. S. Czarnecki, "Reactive and unbalanced currents compensation in three-phase circuits under nonsinusoidal conditions", *IEEE Trans. on Instr. and Meas.,* IM-38, No. 3, pp. 754-459, 1989.

[9] L. S. Czarnecki, P. H. Haley, "Power properties of four-wire systems with nonsinusoidal symmetrical voltage", *IEEE Trans. on Power Delivery*, Vol. 31, No. 2, pp. 513-521, 2016.

[10] L. S. Czarnecki, "CPC–based reactive balancing of linear loads in four-wire supply systems with nonsinusoidal voltage", *Przeglad Elektrotechniczny,* R. 95, Nr. 95, pp. 1-8, 2019.

[11] L. S. Czarnecki, M. Almousa, "Adaptive balancing of three-phase loads at four-wire supply with reactive compensators and nonsinusoidal voltage, "2020 IEEE Texas Power and Energy Conf. (TPEC), A&M University, USA, pp. 1-6, DOI: 10.1109/TPEC48276.2020.9042572, 2020.

[12] O. Jordi, L. Sainz, M. Chindris, "Steinmetz system design under unbalanced conditions", *European Trans. on Electrical Power ETEP*, Vol. 12, No. 4, pp. 283-290, 2002.

[13] D. Mayer, P. Kropik, "New approach to symmetrization of three-phase networks", *Int. Journal of Electrical Engineering,* Vol. 56, No. 5-6, pp. 156-161, 2005.

[14] F. de Leon, J. Cohen, "A practical approach to power factor definitions: transmission losses, reactive power compensation, and machine utilization", *Proc. of the Power Eng. Soc. Meeting*, IEEE DOI: 10.1109/PES.2006.1709175, 2006.

[15] W. G. Morsi, M. E. El-Hawary, "Defining power components in nonsinusoidal unbalanced polyphase systems: the issues", *IEEE Trans. on PD.*, Vol. 22, No. 4, pp. 2428-2437, 2007.

[16] M. Malengret, C. T. Gaunt, "Active currents, power factor, and apparent power for practical power delivery systems", *IEEE Access*, Vol. 8, pp. 133095-133113, 2020.

[17] L. S. Czarnecki, "Currents' Physical Components (CPC) – based Power Theory. A Review, Part I: Power properties of electrical circuits and systems", *Przeglad Elektrotechniczny*, Nr. 95, pp. 1-11, Nr. 10/2019.

[18] L. S. Czarnecki, "Currents' Physical Components (CPC) – based Power Theory. A Review, Part II: Filters and reactive, switching and hybrid compensators", *Przeglad Elektrotechniczny,* R. 96, Nr. 4, pp. 1-11, 2020.

[19] D. A. Steeper, R. P. Stratford, "Reactive power compensation and harmonic suppression for industrial power systems using thyristor converters", *IEEE Trans. on Ind. Appl.,* Vol. IA-12, No. 3, 232-254, 1976.

[20] J. E. Miller, *Reactive power control in electric systems*, John Willey & Sons, 1982.

[21] H. Akagi, Y. Kanazawa, A. Nabae, “Instantaneous reactive power compensator comprising switching devices without energy storage components”, *IEEE Trans. on Ind. Appl.*, IA-20, No. 3, pp. 625-630, 1984.

[22] N. Balabanian, Network Synthesis, Prantice-Hall, Englewood Cliffs, New York, 1958.

[23] L. S. Czarnecki, M. S. Hsu, “Thyristor controlled susceptances for balancing compensators operated under nonsinusoidal conditions”, *Proc. IEE, B, EPA*, Vol. 141, No. 4, pp. 177-185, 1994.

[24] L. S. Czarnecki, “Do energy oscillations degrade the effectiveness of energy transfer in electrical systems,” IEEE Trans. on Ind. Appl. DOI: 10.1109/TIA.2021.3051314, pp. 1-10, 2021.

[25] L. S. Czarnecki, M. Almousa and V. M. Gadiraju, "Why the Electric Arc Nonlinearity Improves the Power Factor of Ac Arc Furnaces?," *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*, Palermo, Italy, 2018, pp. 1-6, doi: 10.1109/EEEIC.2018.8493860.

An Advanced Fuzzy Logic Based Method for Power Transformers Assessment

Sunita Baral, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, sunita.baral95@gmail.com*

Mahendra Kumar Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, mk_sahoo34@hotmail.com*

Dillip Kumar Nayak, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, dknayak225@yahoo.co.in*

Chinmaya Ranjan Pradhan, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, cr_pradhan@outlook.com*

Abstract: Dissolved Gas Analysis is an effective method for detecting faulty power transformers in their early stages. However, technical interpretation of results can be complex and highly dependent on the experience of experts. This paper presents an attempt to detect power transformer incipient fault via gas concentrations obtained from oil sampling and Dissolved Gas Analysis. The proposed method uses a sophisticated fuzzy logic system to perform fault type classification. Ratios and relative percentages of 5 key gases (Hydrogen, Methane, Ethane, Ethylene, and Acetylene) are taken as input variables, then the fuzzy system will try to generate an output vector that indicates six basic fault types, including partial, low, and high energy discharges as well as three ranges of thermal fault. This method can be easily implemented in any environment that supports basic mathematical operators. To demonstrate how the proposed fuzzy logic method works, the authors developed an offline MATLAB script and an online web-based application that can provide multiple assessments by various methods simultaneously. The set of membership functions and fuzzy rules presented in this paper allows the detection of multiple faults at once. Performance tests on many actual data sets show that the proposed method achieves better accuracy than the traditional ratio codes, even on a par with state-of-the-art graphical-based tools such as the Duval triangle or pentagon.

Keywords: Power Transformer, Incipient Fault Detection, Dissolved Gas Analysis, Fuzzy Logic

1. Introduction

In-service power transformers are frequently subjected to both potential internal defects and external stresses. Thermal stress caused by local overheating accelerates the aging process of oil and paper insulation. Electrical and mechanical stresses from external sources such as lightning strikes or short-circuit current greatly contribute to reducing the remaining lifetime of power transformers. Those stresses cause material decomposition and generate dissolved combustible gases in insulating oil, some of which are oxides of carbon, hydrogen, and hydrocarbons. Internal defects generate a particular amount of characteristic gases dissolved in insulating oil that can be used for early fault identification.

Proactive detection of faults helps minimize the risk of undesirable outage of power transformers from the power system network. Effective monitoring and diagnostic techniques must be adopted to improve the reliability of the equipment and to avoid any catastrophic failure. Among existing techniques, dissolved gas in oil analysis (DGA) is a powerful method to detect power transformer incipient faults [1-3].

Conventional DGA interpretation methods such as key gas inspection or gas ratios based methods [4-7] have been widely used, but they still have some limits and sometimes cannot give a proper diagnosis. Recently, the introduction of the Duval triangles and pentagons [8-10] solved the problem of unidentified faults. However, the analysis is not always straightforward as there may be more than one fault present at the same time. Precise DGA interpretation is still a hot topic in the power transformer fault diagnosis and condition assessment research area.

In this paper, a fuzzy logic-based method is developed to enhance the quality of existing DGA interpretation tools. While other methods can only detect a single fault, the fuzzy

method is feasible to address the classification of transformer faults in the case where multiple faults occur at once. Since fuzzy logic has an advantage in processing unclear states, it is also possible to implement the criticality alert in the fuzzy diagnostic system. At the end of this paper, the authors would like to compare the efficiency of the fuzzy logic-based approach to other conventional methods by surveying real cases in Vietnam.

2. Literature Review

Interpretation of DGA results is not always straightforward, as there are several possible causes of the presence of gas in a transformer. Some of those are related to real fault conditions, others are related to more benign conditions such as stray gassing. There is no direct and infallible method using DGA to obtain an exact evaluation of a transformer's condition. However, it is necessary to have a reliable DGA assessment tool to detect any possible fault that might occur inside a power transformer. In this section, the authors wish to provide a brief review of popular DGA interpretation techniques and point out the reason why a fuzzy logic-based method can provide a better solution.

2.1. Key Gas Inspection

The key gas method applies some basic rules for finding the fault pattern based on dominant gases. Hydrogen (H_2) is primarily generated from corona partial discharge; Acetylene (C_2H_2) is created from arcing in oil or paper at very high temperatures. Overheating and thermal faults give rise to Methane (CH_4), Ethane (C_2H_6), and Ethylene (C_2H_4) as well as Carbon Monoxide (CO), and Carbon Dioxide (CO_2) if the fault is related to solid insulation decomposition. By determining which gasses are dominating, one can speculate the existence of internal fault.

There is one big challenge in using this method, as it requires the users' experience. Furthermore, software implementation of the key gas method seems to be a challenge. Inconclusive or wrong fault identification occurs regularly even with sophisticated key gas rules. The reason for this problem is that it is not always clear which is the dominant gas, or the main gas formed may not be reliable enough for fault identification. However, observing key gases is essential for building an advanced interpretation system based on fuzzy logic or artificial intelligence.

Table 1. *Fault identification based on key gasses.*

Fault	Key gasses
Partial discharge	H_2
Arcing	C_2H_2
Thermal fault (oil)	CH_4, C_2H_4, C_2H_6
Thermal fault (paper)	CO, CO_2

2.2. Gas Ratio Methods

Gas ratio-based methods take correlation of ratio between some pairs of fault gas concentrations with certain fault types. These methods were introduced in the 1970s and remain popular until lately. There are several variations such as the Dornenburg ratio, the Rogers ratio, and the three gas ratio methods [3-7].

The Rogers ratio method [4] considers two of the four ratios CH_4/H_2, C_2H_2/C_2H_4, C_2H_4/C_2H_6, and C_2H_6/CH_4. However, later studies showed the ratio of C_2H_6/CH_4 did not correlate well with the faults, and thus it was removed in recent studies. The three ratio method is now recommended by both the IEEE and the IEC standards [3, 5]. The interpretation guide of this method is shown in tables 1 and 2, in which there are three ratio codes for each ratio and six fault types [6].

Dividing one small value of a fault gas by another small value of another fault gas will give a significant ratio, but the magnitudes of the fault gases in such cases are too small. For that reason, ratio methods are only applicable when a significant amount of the gas is present; otherwise, they may lead to misdiagnosis. The common weakness of ratio-based methods is that they sometimes are not capable of giving a result or may yield an incorrect one in others. Therefore, some researchers attempt to add or modify rules to achieve better accuracy [11].

Table 2. *The IEC ratio codes [6].*

Ratio	States		
	0	1	2
$r_1 = C_2H_2/C_2H_4$	< 0.1	0.1 – 3	>3
$r_2 = CH_4/H_2$	0.1 - 1	< 0.1	> 1
$r_3 = C_2H_4/C_2H_6$	< 1	1 – 3	> 3

Table 3. *Fault classification by using the IEC ratio codes [6].*

Fault	Fault Code	r_1	r_2	r_3
Normal	N	0	0	0
Partial discharge	PD	0 or 1	1	0
Low energy discharge	D1	1 or 2	0	1 or 2
High energy discharge	D2	1	0	2
Thermal fault with t < 150°C	T1	0	0	1
Thermal fault with 150°C < t < 300°C		0	2	0
Thermal fault with 300°C < t < 700°C	T2	0	2	1
Thermal fault with t > 700°C	T3	0	2	2

2.3. Graphical Methods

Several graphical methods have been developed to overcome the problem of having unidentified cases. Two of the most well-known graphical-based methods are the Duval triangles and pentagons [8-10]. Other approaches such as the Mansour diagnostic pentagon [12] or the heptagon developed by Gouda et al. [13] were introduced recently. In this section, the authors shall only briefly summarize the Duval triangles and pentagons, as they are used for comparison later in the research.

2.3.1. Duval Triangles

The original Duval triangle [8] uses a set of three characteristics gases: CH_4, C_2H_4, and C_2H_2. The sides of the triangle are expressed in triangular coordinates (x, y, z), where x, y, z are the relative percentage of CH_4, C_2H_4, and C_2H_2, respectively.

This method allows the identification of the six basic types of faults mentioned in the last section (PD, D1, D2, T1, T2, and T3), in addition to mixtures of electrical/ thermal faults in zone DT. Those regions are established through empirical inspection of DGA results from a specific liquid type and the observed equipment gassing source. A fault is identified based on which region the corresponding point (x, y, z) lies on.

To this day, there are several versions of the Duval triangle, in which the first, fourth and fifth are exclusive to mineral oil. The fourth and fifth triangles take a different set of gases and are only used for thermal fault inspection [9]. The first triangle, together with the pentagon counterpart, as depicted in figure 1, are widely used for the diagnosis of high voltage power transformers.

2.3.2. Duval Pentagon

The Duval pentagon [10] uses the percentages of five gases (H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2) to their sum. The vertices of this pentagon correspond to the maximum relative concentration of 40%. Inside this pentagon, the zones are corresponding to the basic types of faults just like that of the Duval triangle, as well as a stray gassing zone (S). The percentage of each gas is marked on the appropriate axis drawn from the center of the pentagon to one of the vertices. These points are then connected to form a small polygon in which the centroid always lies inside the Duval pentagon. The position of the designated centroid points to one of the seven fault zones as one can observe in figure 1.

Both Duval methods use relative gas percentages instead of ratios and thus avoid the problem of unidentified cases. In contrast, because the triangles and pentagons always give a diagnostic, they should only be used to identify a fault when a sufficient amount of combustible gas exists. Moreover, due to the nature of graphical-based interpretation methods, they cannot verify the existence of multiple faults at a time.

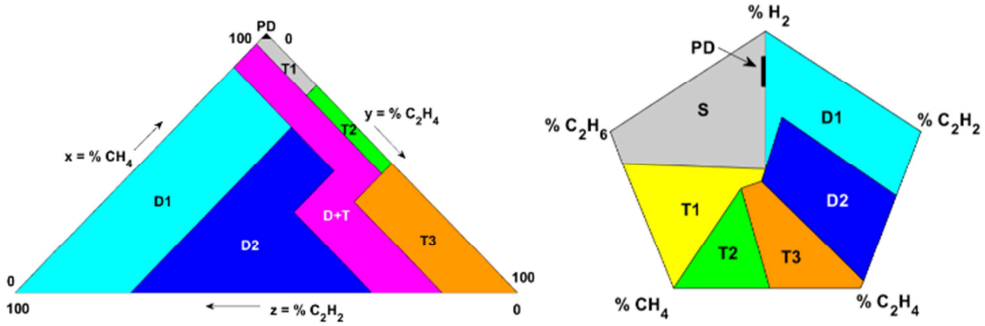


Figure 1. *The Duval triangle (left) and pentagon (right).*

3. Proposed Method

When one or more than one fault occurs in a transformer, multiple key gases with different concentrations exist and cause the ratio codes to overlap. Because of that, the relationship between various gases becomes too complicated and may not match the predefined values. In multiple-fault conditions, gases from different faults are mixed, resulting in confusing ratios between gas components. This problem can be overcome with the aid of more sophisticated analysis methods such as the fuzzy logic presented in this section. The proposed method is called Fuzzy Ratio and Percentage (FRP in short).

3.1. Fuzzy Logic Based Method

In the IEC ratio-based interpretation methods, the ratio codes 0, 1, or 2 can either be True or False, but not anything in between. The gas ratio boundary should be fuzzy, especially when more than one type of fault exists. Between different kinds of faults, the codes should not change sharply across their boundaries. Therefore, in the proposed fuzzy logic-based method, input variables are transformed into a set of states via fuzzy functions. Membership functions in the uppermost region are S-shape functions governed by (1), while counterparts in the lowest region are Z-shape curves represented by (2). The middle region is occupied by π-shape curves calculated by the minimum combinations of (1) and (2). In these equations, a, b, c, and d are parameters that affect the shape and boundary of the curve. They represent the boundary conditions so that the membership functions translate input values into intermittent fuzzy states.

$$\mu_s(x;a,b)=\begin{cases}0, & x\le a\\ 2\left(\frac{x-a}{b-a}\right)^2, & a\le x\le\frac{a+b}{2}\\ 1-2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2}\le x\le b\\ 1, & x\ge b\end{cases} \quad (1)$$

$$\mu_z(x;c,d)=\begin{cases}1, & x\le c\\ 1-2\left(\frac{x-c}{d-c}\right)^2, & c\le x\le\frac{c+d}{2}\\ 2\left(\frac{x-d}{d-c}\right)^2, & \frac{c+d}{2}\le x\le d\\ 0, & x\ge d\end{cases} \quad (2)$$

Conditional statements given by the rules in table 2 are combinations of conventional logics “AND” and “OR”, which can be converted into mathematical terms by the “MIN” and “MAX” operators. For example, the last rule is [r_1 is 0] AND [r_2 is 2] AND [r_3 is 2]; this statement is translated to $\min[\mu_s(r_1), \mu_z(r_2), \mu_z(r_3)]$. When a condition is fulfilled, either fully or partially, certain rules will be triggered. If the output is defined by a vector in which each

index corresponds to a fault code, then those rules will give values in the range between 0 and 1 to the indexes based on the trigger condition. This method allows the detection of multiple faults at once, a feature that neither the conventional gas ratio methods nor the Duval methods had.

3.2. Fuzzy Rules Based on Gas Ratio

Initially, membership functions and fuzzy rules were designed based on the IEC gas ratio method. However, this approach turned out to be the same as the conventional method; that means the weakness of having unidentified cases still exists. After inspecting various samples, the authors developed an improved set of ratio codes through membership functions shown in table 4. The corresponding fuzzy rules are described in table 5, in which F_r is the output vector consists of 7 indexes correspond to six basic types of faults (PD, D1, D2, T1, T2, T3) and a normal state (N).

Table 4. *Gas ratio membership functions.*

Membership function	Type	a	b	c	d
$\mu_{r10}(r_1)$	Z-shape	-	-	0.3	0.7
$\mu_{r11}(r_1)$	Π-shape	0.3	0.7	2.8	3.2
$\mu_{r12}(r_1)$	Π-shape	2.8	3.2	9.8	10.2
$\mu_{r13}(r_1)$	S-shape	9.8	10.2	-	-
$\mu_{r20}(r_2)$	Π-shape	0	0.4	0.8	1.2
$\mu_{r21}(r_2)$	Z-shape	-	-	0	0.4
$\mu_{r22}(r_2)$	Π-shape	0.8	1.2	2.8	3.2
$\mu_{r23}(r_2)$	S-shape	2.8	3.2	-	-
$\mu_{r30}(r_3)$	Z-shape	-	-	0.8	1.2
$\mu_{r31}(r_3)$	Π-shape	0.8	1.2	2.8	3.2
$\mu_{r32}(r_3)$	S-shape	2.8	3.2	-	-

Table 5. *Gas ratios fuzzy rules.*

Fault	Rule
Normal	$F_r(0) = \min[\mu_{r10}(r_1), \mu_{r20}(r_2), \mu_{r30}(r_3)]$
PD	$F_r(1) = \min[\max[\mu_{r10}(r_1), \mu_{r11}(r_1)], \mu_{r21}(r_2)]$
D1	$F_r(2) = \max[d_{11}, d_{12}, d_{13}, d_{14}]$ Where: $d_{11} = \min[\mu_{r11}(r_1), \mu_{r21}(r_2), \mu_{r30}(r_3)]$ $d_{12} = \min[\max[\mu_{r10}(r_1), \mu_{r11}(r_1)], \mu_{r21}(r_2), \mu_{r31}(r_3)]$ $d_{13} = \min[\mu_{r11}(r_1), \max[\mu_{r20}(r_2), \mu_{r22}(r_2), \mu_{r23}(r_2)], \max[\mu_{r30}(r_3), \mu_{r31}(r_3)]]$ $d_{14} = \min[\mu_{r12}(r_1), \max[\mu_{r20}(r_2), \mu_{r21}(r_2), \mu_{r22}(r_2)], \mu_{r31}(r_3)]$
D2	$F_r(3) = \max[d_{21}, d_{22}, d_{23}, d_{24}]$ Where: $d_{21} = \min[\max[\mu_{r10}(r_1), \mu_{r11}(r_1)], \mu_{r21}(r_2), \mu_{r32}(r_3)]$ $d_{22} = \min[\mu_{r11}(r_1), \max[\mu_{r20}(r_2), \mu_{r22}(r_2), \mu_{r23}(r_2)], \mu_{r32}(r_3)]$ $d_{23} = \min[\mu_{r12}(r_1), \max[\mu_{r20}(r_2), \mu_{r21}(r_2), \mu_{r22}(r_2)], \max[\mu_{r30}(r_3), \mu_{r32}(r_3)]]$ $d_{24} = \mu_{r13}(r_1)$
T1	$F_r(4) = \max[t_{11}, t_{12}, t_{13}]$ Where: $t_{11} = \min[\mu_{r10}(r_1), \mu_{r20}(r_2), \mu_{r31}(r_3)]$ $t_{12} = \min[\mu_{r10}(r_1), \max[\mu_{r22}(r_2), \mu_{r23}(r_2)], \mu_{r30}(r_3)]$ $t_{13} = \min[\mu_{r12}(r_1), \mu_{r23}(r_2), \mu_{r30}(r_3)]$
T2	$F_r(5) = \max[t_{21}, t_{22}]$ Where, $t_{21} = \min[\mu_{r10}(r_1), \max[\mu_{r22}(r_2), \mu_{r23}(r_2)], \mu_{r31}(r_3)]$ $t_{22} = \min[\mu_{r12}(r_1), \mu_{r23}(r_2), \mu_{r31}(r_3)]$
T3	$F_r(6) = \max[t_{31}, t_{32}, t_{33}]$ Where: $t_{31} = \min[\mu_{r10}(r_1), \mu_{r20}(r_2), \mu_{r32}(r_3)]$ $t_{32} = \min[\mu_{r10}(r_1), \max[\mu_{r22}(r_2), \mu_{r23}(r_2)], \mu_{r32}(r_3)]$ $t_{33} = \min[\mu_{r12}(r_1), \mu_{r23}(r_2), \mu_{r32}(r_3)]$

3.3. Fuzzy Rules Based on Gas Percentage

Performance test on the gas ratio-based fuzzy system shows an improvement in diagnostic accuracy. However, in some partial discharge cases, the Hydrogen contents are dominant, while other gases are insignificant. In those cases, the "Normal" rule is triggered instead of "PD", regardless of high H_2 concentrations. It is not uncommon to find increased levels of H_2 or C_2H_4 when C_2H_2 is detected, leading to a fuzzy boundary between low and high energy discharge faults. In that situation, the gas percentage method may be more effective and therefore, should be adopted to support the fuzzy gas ratio.

The relative percentages of H_2, CH_4, C_2H_6, C_2H_4, C_2H_2 in a sample are denoted as p_1, p_2, p_3, p_4, p_5. They are described in 3 levels: "Low", "Medium, and "High" by the Z, π, and S functions, just like their gas ratio counterparts. The boundary values of those functions the gas percentage fuzzy rules are mathematically described in the next two tables.

3.4. Output Calculation

By observation during performance tests, the authors realized that the gas ratio fuzzy system was more sensitive to thermal faults, while the gas percentage fuzzy system was more reliable in detecting partial discharge and low energy discharge faults. Therefore, the total fault vector should be

calculated by taking the normalized average of the gas ratio and the gas percentage fuzzy outputs by (3).

$$F_{total} = w_r \frac{F_r}{\sum_{i=0}^{6} F_r(i)} + w_p \frac{F_p}{\sum_{i=0}^{6} F_p(i)} \quad (3)$$

Table 6. *Gas percentage membership functions.*

Membership function	Type	a	b	c	d
μ_L(gas %)	Z-shape	-	-	0	10
μ_M(gas %)	Π-shape	0	10	15	25
μ_H(gas %)	S-shape	15	25	-	-

Table 7. *Gas percentage fuzzy rules.*

Fault	Rule
Normal	$F_p(0) = \min[\mu_L(p_1), \mu_L(p_2), \mu_L(p_3), \mu_L(p_4), \mu_L(p_5)]$
PD	$F_p(1) = \min[\mu_H(p_1), \max[\mu_L(p_3), \mu_M(p_3)], \mu_L(p_4), \mu_L(p_5)]$
D1	$F_p(2) = \min[\max[\mu_M(p_1), \mu_H(p_1)], \mu_M(p_5)]$
D2	$F_p(3) = \mu_H(p_5)$
T1	$F_p(4) = \min[\mu_H(p_3), \mu_L(p_4), \mu_L(p_5)]$
T2	$F_p(5) = \min[\max[\mu_M(p_4), \mu_H(p_4)], \mu_L(p_5)]$
T3	$F_p(6) = \min[\mu_L(p_3), \mu_H(p_4), \mu_L(p_5)]$

* p1, p2, p3, p4, p5 are the percentages of H_2, CH_4, C_2H_6, C_2H_4 and C_2H_2

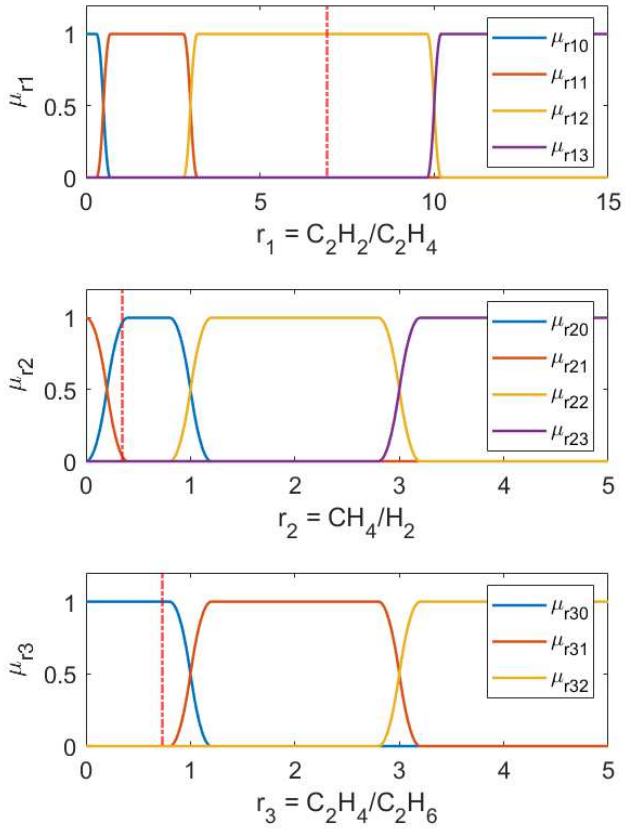


Figure 2. *Fuzzy gas ratio membership functions of the given example.*

This example illustrates in detail how the proposed method works. Consider a case where gas contents are: H_2 = 63 ppm, CH_4 = 22 ppm, C_2H_6 = 15 ppm, C_2H_4 = 11 ppm and C_2H_2 = 76 ppm. The gas ratios are r_1 = 6.9, r_2 = 0.35, and r_3 = 0.73. This combination results in a ratio code of "200" that does not belong to the original IEC guideline. The ratios are translated to fuzzy states (illustrated in figure 2):

1) $[\mu_{r10}(r_1), \mu_{r11}(r_1), \mu_{r12}(r_1), \mu_{r13}(r_1)] = [0,0,1,0]$;
2) $[\mu_{r20}(r_2), \mu_{r21}(r_2), \mu_{r22}(r_2), \mu_{r23}(r_2)] = [0.97,0.03,0,0]$;
3) $[\mu_{r30}(r_3), \mu_{r31}(r_3), \mu_{r32}(r_3) = [1,0,0]$.

This combination results in $F_r(3)$ equal to 0.97 and triggers the D2 fault rule. In this case, the fuzzy gas ratio rules in table 5 give a ratio diagnosis vector F_r = [0, 0, 0, 0.97, 0, 0, 0]. Similarly, the gas percentages are p_1 = 33.7%, p_2 = 11.76%, p_3 = 8.02%, p_4 = 5.88%, p_5 = 40.64%. They are translated to:

1) $[\mu_L(p_1), \mu_M(p_2), \mu_H(p_1)] = [0,0,1]$;
2) $[\mu_L(p_2), \mu_M(p_2), \mu_H(p_2)] = [0,1,0]$;
3) $[\mu_L(p_3), \mu_M(p_3), \mu_H(p_3) = [0.08,0.92,0]$;
4) $[\mu_L(p_4), \mu_M(p_4), \mu_H(p_4) = [0.08,0.66,0]$;
5) $[\mu_L(p_5), \mu_M(p_5), \mu_H(p_5)] = [0,0,1]$.

The condition of D2 is fulfilled and thus $F_p(3)$ equal to 1. The gas percentage laws in table 7 result in F_p = [0, 0, 0, 1, 0, 0, 0]. By using (3), the total output vector is [0, 0, 0, 1, 0, 0, 0], which indicates occurrence of high energy discharge (D2) with 100% certainty.

4. Results and Discussion

The authors built a script in the MATLAB software environment to perform DGA interpretations of several datasets. The code consists of multiple methods, including Rogers ratio, IEC ratio, Duval triangle and pentagon, and the proposed fuzzy logic system. Results from all of those methods are compared with one another to evaluate their efficiency in incipient fault classification.

4.1. Cases Study

To test the performance of the proposed Fuzzy Ratio and Percentage method, the dataset obtained from [14] was used. There are 20 samples described in table 9; results obtained from multiple analyses are also compared. The proposed method outperforms traditional ratio codes in fault diagnosis capability. With this dataset, the proposed method generally agrees with the Duval triangle and pentagon, even achieves better accuracy in the tricky cases of partial discharge.

Another sample dataset was obtained from the Long An Power Company in Vietnam to investigate the performance of the proposed method. The performances of multiple methods are compared in table 9. One can observe that with such low gas concentrations, fault classification, in this case, would be tricky. With this dataset, all method generally agrees with one another. There is one tricky case with sample number 4, which has dominant H_2 and CH_4 contents. This is an obvious sign of partial discharge. However, except for the proposed method, none of the others can classify this fault.

Multiple tests on different datasets [11, 15] were also performed but not fully show in this paper. In general, the proposed method generated highly reliable conclusions that agreed well with actual faults.

Table 8. *Dataset used for performance check [14] (gas contents in ppm).*

Sample	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Known fault [14]	IEC [5]	Rogers [4]	Duval triangle [8]	Duval pentagon [10]	Proposed method
1	200	700	250	740	1	T2	T2	-	T3	T3	T2: 84% T3: 16%
2	300	490	180	360	95	T2	-	T2	DT	T3	D1: 38% T2: 62%
3	56	61	75	32	31	D1	-	-	D2	T1	D1: 100%
4	33	26	6	5.3	0.2	N	N	N	T1	T1	N: 46% T2: 44%
5	176	205.9	47.7	75.7	68.7	D1	-	-	D1	T1	D1: 100%
6	70.4	69.5	28.9	241.2	10.4	T3	-	-	T3	T3	T2: 37% T3: 58% PD: 21%
7	162	35	5.6	30	44	D2	D2	D2	D2	D2	D1: 49% D2: 30%
8	345	112.25	27.5	51.5	58.75	D1	D1	D2	D2	D1	D1: 96%
9	181	262	210	528	0	T2	T2	T2	T3	T3	T2: 100%
10	172.9	334.1	172.9	812.5	37.7	T3	T3	T3	T3	T3	T2: 44% T3: 50%
11	2587.2	7.882	4.704	1.4	0	PD	PD	PD	T1	S	PD: 100%
12	1678	652.9	80.7	1005.9	419.1	D2	D2	-	DT	D2	D1: 50% T3: 41%
13	206	198.9	74	612.7	15.1	T3	-	-	T3	T3	T2: 39% T3: 59% N: 31%
14	180	175	75	50	4	T1	N	N	T2	T1	T1: 19% T2: 49% D1: 49%
15	34.45	21.92	3.19	44.96	19.62	D2	D2	-	DT	D2	D2: 12% T3: 38%
16	51.2	37.6	5.1	52.8	51.6	D2	D2	D2	D2	D2	D2: 100% PD: 19%
17	106	24	4	28	37	D2	D2	D2	D2	D2	D1: 37% D2: 44%
18	180.85	0.574	0.234	0.188	0	PD	PD	PD	T2	PD	PD: 100%
19	27	90	42	63	0.2	T2	T2	-	T2	T2	T2: 100%
20	138.8	52.2	6.77	62.8	9.55	D2	D2	-	T3	D2	T2: 21% T3: 71%

Table 9. *Dataset from Long An PC (gas contents in ppm).*

Sample	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Known fault	IEC [5]	Rogers [4]	Duval triangle [8]	Duval pentagon [10]	Proposed method
1	23.6	12.4	3.8	50.9	0	T3	N/A	T2	T3	T3	T3 (30%) T3 (70%)
2	5.6	32.4	10.1	13.1	0	T2	T2	N/A	T2	T2	T2 (100%)
3	43.3	50	8.9	11.2	0	T2	T2	T2	T1	T1	T2 (99%)
4	170.7	68.9	8.4	5.7	0	PD	N/A	N	T1	S	N (50%) PD (44%)
5	4.7	14.3	2	6	0	T2	T3	T2	T2	T2	T2 (69%) T3 (31%)
6	5.7	14.7	2.1	5.8	0	T2	T2	T2	T2	T2	T2 (90%) T3 (10%)
7	12.1	11.3	3.6	24.7	0	T3	N/A	T2	T3	T3	T2 (42%) T3 (58%)

4.2. Inspection on a Larger Dataset

The authors collected 240 samples from local utilities in the Southern region of Vietnam and performed multiple tests. The overall results are summarized in Table 10. In general, the proposed method generated highly reliable conclusions that agreed well with actual faults, with an accuracy of over 80%. Except for the Duval pentagon, none of the others can reach 50% accuracy in this dataset. A noteworthy feature of the proposed method is that it is more sensitive to partial discharge faults than other interpretation methods.

An online demo version of the method is also available for use. The IEC ratio, Roger ratio, Duval triangle, and pentagon are also included in this demo version; all of them are implemented using Javascript and HTML. However, the algorithm used for

developing graphical-based methods in this online version is not very accurate when the point lies on the edge of a fault zone. A better choice would be using a more mathematical-oriented environment such as MATLAB or OCTAVE.

Table 10. *Comparison of various methods over a large transformer fleet.*

Fault	Actual case	Number of correct diagnosis				
		IEC [5]	Rogers [4]	Duval triangle [8]	Duval pentagon [10]	Proposed method
PD	23	7	3	7	6	20
D1	32	6	1	22	20	24
D2	63	6	19	15	18	44
T1	76	44	68	35	74	65
T2	18	14	1	4	5	14
T3	28	20	5	28	28	28
Total	240	97	97	111	151	195
Percentage		40%	40%	46%	63%	81%

5. Conclusion

In this paper, a new algorithm to detect potential faults inside power transformers was introduced. The diagnostic system is made based on fuzzy logic that process ratios and percentages of key gases obtain from DGA results. In most cases presented throughout the paper, the use of fuzzy logic overcomes the limitations of traditional gas ratios based interpretation with high accuracy in fault diagnosis. Since the method allows the detection of multiple faults in one sample, it can provide comprehensive insights into the conditions of a power transformer. That feature might provide additional information and help condition assessment be more reliable. The proposed diagnosis algorithm is not only efficient but also very simple to implement. In short, the research contributes a useful tool for the condition assessment of power transformers.

Nowadays, alarm concentration values are set by independent experts, based on previous experience with equipment with similar characteristics [5]. Future research on this topic should examine the fault criticality to determine the normality percentages and critical concentrations. Again, using fuzzy logic would be a suitable approach to this task, since gas concentrations can vary from sample to sample.

References

[1] "IEEE Guide for Evaluation and Reconditioning of Liquid Immersed Power Transformers," in IEEE Std C57.140-2017 (Revision of IEEE Std C57.140-2006), 2017, doi: 10.1109/IEEESTD.2017.8106924.

[2] "IEEE Guide for Diagnostic Field Testing of Fluid-Filled Power Transformers, Regulators, and Reactors," in IEEE Std C57.152-2013, 2013, doi: 10.1109/IEEESTD.2013.6544533.

[3] "IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers," in IEEE Std C57.104-2019 (Revision of IEEE Std C57.104-2008), 2019, doi: 10.1109/IEEESTD.2019.8890040.

[4] R. R. Rogers, "IEEE and IEC Codes to Interpret Incipient Faults in Transformers, Using Gas in Oil Analysis," in IEEE Transactions on Electrical Insulation, vol. EI-13, no. 5, pp. 349-354, Oct. 1978, doi: 10.1109/TEI.1978.298141.

[5] "Mineral oil-filled electrical equipment in service - Guidance on the interpretation of dissolved and free gases analysis," in IEC 60599.

[6] S. Chakravorti, D. Dey, and B. Chatterjee, Recent trends in the condition monitoring of transformers theory, implementation and analysis. London: Springer, 2013.

[7] M. Duval and A. dePabla, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," in IEEE Electrical Insulation Magazine, vol. 17, no. 2, pp. 31-41, March-April 2001, doi: 10.1109/57.917529.

[8] M. Duval, "A review of faults detectable by gas-in-oil analysis in transformers," in IEEE Electrical Insulation Magazine, vol. 18, no. 3, pp. 8-17, May-June 2002, doi: 10.1109/MEI.2002.1014963.

[9] M. Duval, "The Duval triangle for load tap changers, non-mineral oils and low temperature faults in transformers," in IEEE Electrical Insulation Magazine, vol. 24, no. 6, pp. 22-29, November-December 2008, doi: 10.1109/MEI.2008.4665347.

[10] M. Duval and L. Lamarre, "The duval pentagon-a new complementary tool for the interpretation of dissolved gas analysis in transformers," in IEEE Electrical Insulation Magazine, vol. 30, no. 6, pp. 9-12, November-December 2014, doi: 10.1109/MEI.2014.6943428.

[11] B. M. Taha, S. S. M. Ghoneim and A. S. A. Duaywah, "Refining DGA methods of IEC Code and Rogers four ratios for transformer fault diagnosis," 2016 IEEE Power and Energy Society General Meeting (PESGM), 2016, pp. 1-5, doi: 10.1109/PESGM.2016.7741157.

[12] D. A. Mansour, "Development of a new graphical technique for dissolved gas analysis in power transformers based on the five combustible gases," in IEEE Transactions on Dielectrics and Electrical Insulation, vol. 22, no. 5, pp. 2507-2512, October 2015, doi: 10.1109/TDEI.2015.004999.

[13] E. Gouda, S. H. El-Hoshy, and H. H. El-Tamaly, "Proposed heptagon graph for DGA interpretation of oil transformers," IET Generation, Transmission & Distribution, vol. 12, no. 2, pp. 490–498, 2018.

[14] Hongzhong Ma, Zheng Li, P. Ju, Jingdong Han and Limin Zhang, "Diagnosis of power transformer faults on fuzzy three-ratio method," 2005 International Power Engineering Conference, 2005, pp. 1-456, doi: 10.1109/IPEC.2005.206897.

[15] Mang-Hui Wang, "A novel extension method for transformer fault diagnosis," in IEEE Transactions on Power Delivery, vol. 18, no. 1, pp. 164-169, Jan. 2003, doi: 10.1109/TPWRD.2002.803838.

Modelling and Simulation of Intelligent Master Controller Model for Hybridized Power Pool Deployment

Sanam Devi, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sanamdevi226@yahoo.co.in*

Manoj Mohanta, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, manoj.mohanta62@outlook.com*

Smruti Ranjan Panda, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, sr_panda@outlook.com*

Subhendu Sahoo, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, s.sahoo95@outlook.com*

Abstract: Every conceptual framework requires several developmental stages such as prototyping, preproduction and production stages. This paper considers prototype developmental stage which entails design, modelling and simulation for the conceptual system to determine suitable parameters and specifications before the production task is initiation. The inability to represent the conceptual control system with mathematical equivalence would hamper on the system operational efficiency, stability, controllability and observability; would not be guaranteed. This paper focuses on the modelling and simulation of intelligent master controller for hybridized power pool deployment. This is achieved using state space mathematical model, MATLAB/Simulink and proteus software. The state space model provides the mathematical equation for the system stability, controllability and observability criteria from the system transfer function. The MATLAB/Simulink software provides response trends and the Proteus software provides the virtual implementation platform for concept validation with its code written in Arduino (IDE). The system was demonstrated through simulation and the virtual results showed that the system capability in fostering intelligent control commands in the hybridized power pool scenario. The system stability was determined using Root locus, Nyquist and Routh Hurwitz criteria. Subsequent research efforts are being made towards implementing the design optimizable on the hardware using the design specifications.

Keywords: Deployment, Hybridized Power Pool, Intelligent Master Controller, Modelling, Simulation

1. Introduction

Transformation of any conceptual idea to physical reality in engineering product development is not possible without modelling and simulation [1]. This is where the fundamental concept detail would be unveiled, control loop parameter defined and transfer function formulated [2]. Intelligent master controller is a digital regulatory device developed for hybridized power system monitoring and control application. Providing a modelled-based environment for an engineering system development helps to understudy the system performance prior to the hardware building. The dynamic characteristic of the system is represented by several created models and these model aids in the evaluation of the proposed system performance from the different constituent of the subsystems [3]. Algorithms are developed for the system

simulations as a way of confirming the system performance [4, 5]. The intelligent master controller model is novel and it would be difficult to arrive at its model parameter and variable. This would in turn help in the system prototype development. The conceptual model for hybridized power pool deployment is complex and requires mathematical representation to assist in its complexity reduction [6, 7]. The created mathematical model would unfold the system parameters and variables. An ability to simulate the system variables and parameters from the mathematical expression facilitates the physical prototype model implementation [8]. The aim of this work is to model and simulate an intelligent master controller model for hybridized power pool deployment and the objectives are: to develop a mathematical model for the intelligent master controller; to simulate the developed mathematical model and to analyse the simulated model. The contribution to the body of knowledge in this work is the developed model and simulation parameters for an intelligent master controller, these parameters are deployable in hybridized power pool system.

2. Literature Review

The importance of modelling and simulation over the years has played vital roles in product development, one of whose first stage is prototype model behavioral observation. The system dynamic characteristic would be determined before the model implementation. Some modelling and simulation approach are reviewed; the model was validated by investigating multi-terminal direct current (MTDC) system with three or more converter stations. The simulation results showed that the proposed control strategy and the MTDC control protection system meet the requirements of the MTDC transmission system's operation [9].

Investigation on the notion of a hybrid power control system was introduced for voltage control in power systems, and the research establishes the static hybrid automatic voltage control system. The operating procedure was designed using a hybrid hierarchical voltage control system model based on hybrid theory. In order to drive the system, the stability and economic events were specified, by which the synthetic objects of safety, stability, and economy were attained in multi-power source system [10, 3]. The validity of the system and the methodologies proposed in the research were demonstrated by computer simulation results. Parallel time domain simulation was one of the most dependable and promising ways for performing real-time online power system transient stability study [11]. The research proposes a new parallel calculation approach for power system transient stability analysis based on the waveform relaxation method. The practical system's test results showed that the new parallel method completely achieves on-line real-time or even over-real-time calculation speed and can be applied to the practical system's on-line transient stability analysis [12]. The simulation of a command-and-control system makes use of computer simulation technology. In a virtual environment, evaluation of the performance of the designed command post system environment, served as the foundation for a review or optimization of system of command posts. This study examined the weaknesses and flaws in present command and control system modeling and recommended an entity-relationship-based command and control system modeling. This study uses command and control system models for power system scenario analysis, in combination with the Lanchester model that considers command efficiency [13].

Michaels, L. *et al*., carried out research on model-based control system design enhancing quality while also reducing development time, engineering costs, and rework. The time and money spent on hardware and software for each design iteration is saved by evaluating a control system's performance, functionality, and reliability in a simulation environment [4]. This work offers a software tool and approach that not only allows for a complete system simulation early in the design cycle, but also substantially simplifies model development by automatically integrating the components and subsystems that make up the model. The control system can be developed early in the vehicle or powertrain design cycle using this approach, which incorporates plant models, algorithm models, existing controller code, and architectural constructs to greatly speed up the creation of a system simulation that can be used for algorithm development, testing, and validation.

Naşcu, I. *et al*, carried out research on the model of a laboratory level control system. The model for each component in the system was based on both theoretical and experimental findings. This paper describes the steps involved in creating an accurate model of a laboratory level control system. A method to solve parameters of load model frequency characteristics was provided based on extensive research on frequency characteristics of power loads [14]. The weighted total of each static load component's power in the load station was used to determine the static load frequency factor. The load frequency parameters of power generators cluster in the entire load station were calculated using a combination of the statistic synthesis method and the fault fitting method. Simulation results demonstrated the efficiency of the proposed strategy [15].

Qing, K. *et al*., carried out a study on the permanent magnetic linear generator, generator side converter, grid side converter, controller, and grid as part of a directly driven wave power generating system that is connected to a power grid. The following control strategies were presented based on the back-to-back converter structure: The generator side converter was subjected to vector decoupling management in order to maximize power extraction from wave energy; Grid voltage-oriented control was employed to make the current sinusoidal and achieve unit power factor control on the grid side converter. Due to the voltage fluctuation of the DC link when using the standard method, a power feed forward method was proposed to keep the DC link voltage steady and increase the system's dynamic response [16].

In Matlab/Simulink, simulation model for the entire system was created, and simulation results confirmed that the proposed control approach is practical and successful. Due to

the high randomness of wind power, a higher demand for load frequency control in the power system was made. The load frequency control strategy is based on the wind power prediction by Kalman, filter was proposed to reduce the influence on the system frequency using the interconnected power grid with wind power as the research object. On the contrary, the Kalman filter technique was initially employed to estimate wind power. The load frequency controller was then designed using the expected wind power. A load frequency control model for an interconnected power system was also constructed [17]. This control strategy was applied to a four-area power system with integrated wind power (multi-energy injection for continuous energy harvest) in three areas [18]. The simulation results obtained using MATLAB/Simulink showed that the suggested load frequency control technique based on Kalman filter wind power prediction successfully reduced frequency fluctuation and kept the system frequency fluctuation within a narrow range. When compared to traditional PID-based load frequency control, simulation findings showed that it outperforms [19].

3. Methodology

This research considers modelling and simulation of an intelligent master controller model for deployment in hybridized power pool applications. The closed loop block diagram of the Intelligent master controller is formed and deduced into block format and mathematically represented (formular) in differential equation order to derive at the system parameters. This mathematical model uncovers the system parameter which is then simulated to validate the system internal behaviour. The suitable stability criteria of the system would also be considered.

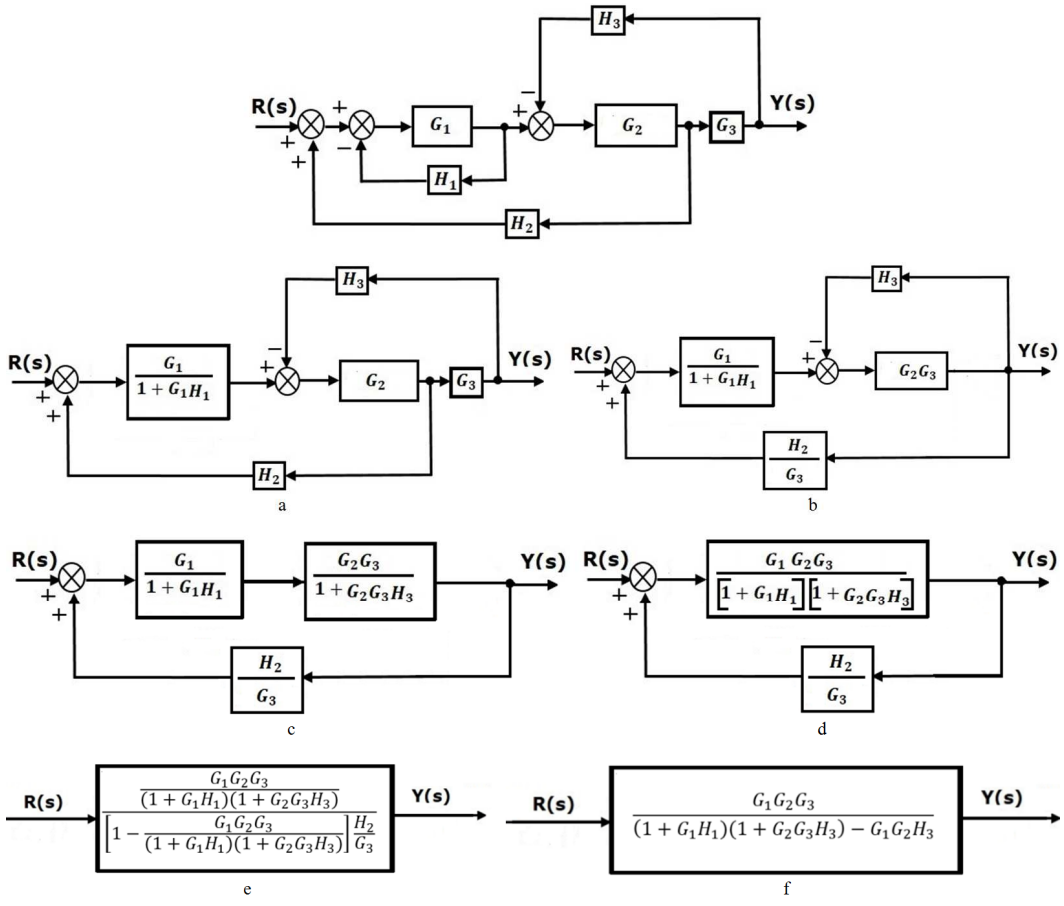


Figure 1. *Closed Loop Control Model for Intelligent Master Controller.*

G_1 represent input from Grid power supply
H_1 represent the loop gain from Grid power supply input
G_2 represent input from Renewable power supply
H_2 represent the loop gain from Renewable power supply input
G_3 represent input from Generating Set power supply
H_3 represent the loop gain from Generating Set power supply input

3.1. Modelling

Modelling entails the process of representing a system with block diagrams. The block helps to determine its sub-system parameters whereas differential equations are mostly used in modelling of control systems. This design takes into

consideration the general closed loop system which comprises of the input, output, feedback, controller and the plant to model the intelligent master controller system. This system is designed using multiple input single output (MISO) model. it has three-input system that is reducing to a single system model with the block represented in Figure 1.

The reduced block diagram gives the transfer function for the intelligent master controller model.

$$Transfer\ Function = \frac{Y(s)}{R(s)} = \frac{G_1G_2G_3}{(1+G_1H_1)(1+G_2G_3H_3)-G_1G_2H_3} \quad (1)$$

The input follows the unit step function U(t)

$$u(t) = 1; t \geq 0\ and\ u(t) = 0; t < 0 \quad (2)$$

The controller is in the ON/OFF (digital system) mode. The feedback is given by $H = 1$ which is in s-domain = $\frac{1}{s}$ and in first order system; the transfer function is given as

$$TF = \frac{G(s)}{1 \pm G(s)H(s)} \quad (3)$$

whereas in the second order system, the transfer function

$$G(s) = \frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2} \quad (4)$$

and is adopted from the system model.

Equation (4) is substituted into equation (1), and this gives equation (5)

$$\frac{\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)}{\left(1+\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\times\frac{1}{s}\right)\left\{1+\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\right\}\frac{1}{s}-\left(\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\times\frac{1}{s}\right)} \quad (5)$$

$$\frac{Y(s)}{R(s)} = \frac{\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)^3}{1+\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)\frac{1}{s}+\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)^3\times\frac{1}{s^2}}$$

$$= \frac{\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)^3}{\frac{1}{1}+\left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)s+\frac{(\omega_n^2)^3}{S^2(S^2+2\zeta\omega_n S+\omega_n^2)^3}}$$

$$= \frac{S^2(S^2+2\zeta\omega_n S+\omega_n^2)+S\omega_n^2+(\omega_n^2)^3(S^2+2\zeta\omega_n S+\omega_n^2)^3}{S^2(S^2+2\zeta\omega_n S+\omega_n^2)}$$

$$= \left(\frac{\omega_n^2}{S^2+2\zeta\omega_n S+\omega_n^2}\right)^3 \times \frac{S^2(S^2+2\zeta\omega_n S+\omega_n^2)}{S^2(S^2+2\zeta\omega_n S+\omega_n^2)+\omega_n^2 S+(\omega_n^2)^3(S^2+2\zeta\omega_n S+\omega_n^2)^2}$$

$$= \frac{(\omega_n^2)^3}{(S^2+2\zeta\omega_n S+\omega_n^2)^2} \times \frac{S^2}{S^2(S^2+2\zeta\omega_n S+\omega_n^2)+\omega_n^2 S+(\omega_n^2)^3(S^2+2\zeta\omega_n S+\omega_n^2)^2}$$

$$= \frac{(\omega_n^2)^2}{(S^2+2\zeta\omega_n S+\omega_n^2)^2} \times \frac{S^2}{S^2+2\zeta\omega_n S^2+\omega_n^2+\omega_n^2 S^2+\omega_n^2 S+(\omega_n^2)^3(S^2+2\zeta\omega_n S+\omega_n^2)^2}$$

$$= \frac{(\omega_n^2)^2}{(S^2+2\zeta\omega_n S+\omega_n^2)^2} \times \frac{S^2}{\omega_n^2(S^2+3\zeta\omega_n S^2+S)+(\omega_n^2)^2(S^2+2\zeta\omega_n S+\omega_n^2)^2}$$

$$= \frac{S^2(\omega_n^2)^2}{(S^2+2\zeta\omega_n S+\omega_n^2)^2(S^2+3\zeta\omega_n S^2+S)+(\omega_n^2)^2(S^2+2\zeta\omega_n S+\omega_n^2)^2}$$

The equation (5) in S-Domain gives equation (6)
Assuming the initial value of the $\zeta = 0.5\ and\ \omega_n = 1$

$$\frac{Y(s)}{R(s)} = \frac{S^2}{(S^2+S+1)^2(2.5S^2+S)+(S^2+S+1)^2} \quad (6)$$

$$\frac{Y(s)}{R(s)} = \frac{S^2}{(S^2+S+1)^2(2.5S^2+S+1)}$$

$$\frac{Y(s)}{R(s)} = \frac{S^2}{(S^4+2S^3+3S^2+2S+1)(2.5S^2+S+1)}$$

$$\frac{Y(s)}{R(s)} = \frac{S^2}{2.5S^6+6S^5+10.5S^4+10S^3+7.5S^2+3S+1}$$

$$R(s) = X(S)$$

The system from the transfer function in equation (6) provides the polynomial characteristics equation for the intelligent

master controller equation (7)

$$Y(S)[\,2.5S^6 + 6S^5 + 10.5S^4 + 10S^3 + 7.5S^2 + 3S + 1] = S^2X(S) \tag{7}$$

The higher order differential equation (7) for the model becomes equation (8);

$$2.5y^{'\,'\,'\,'\,'\,'} + 6y^{'\,'\,'\,'\,'} + 10.5y^{'\,'\,'\,'} + 10y^{'\,'\,'} + 7.5y^{'\,'} + 3y^{'} + y = \ddot{x} \tag{8}$$

Converting equation (8) to state space gives:

$$x_1 = y$$

$$x_2 = y' = \dot{x}_1$$

$$x_3 = y'' = \dot{x}_2$$

$$x_4 = y''' = \dot{x}_3$$

$$x_5 = y'''' = \dot{x}_4$$

$$x_6 = y''''' = \dot{x}_5$$

The state space is given by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{9}$$

$$\dot{y}(t) = Cx(t) + Du(t)$$

$$U(t) = u$$

$$2.5x_6 + 6x_6 + 10.5x_5 + 10x_4 + 7.5x_3 + 3x_2 + x_1 = U(\ddot{t})$$

$$\dot{x}_6 = \frac{6}{2.5}x_6 + \frac{10.5}{2.5}x_5 + \frac{10}{2.5}x_4 + \frac{7.5}{2.5}x_3 + \frac{3}{2.5}x_2 + \frac{1}{2.5}x_1 + u(t)$$

$$\dot{x}_5 = x_6 + 0 + 0 + 0 + 0 + 0 + 0$$

$$\dot{x}_4 = 0 + x_5 + 0 + 0 + 0 + 0 + 0$$

$$\dot{x}_3 = 0 + 0 + 0 + x_3 + 0 + 0 + 0$$

$$\dot{x}_2 = 0 + 0 + 0 + 0 + x_2 + 0 + 0$$

$$\dot{x}_1 = 0 + 0 + 0 + 0 + 0 + x_1 + 0$$

$$\dot{x} = 0 + 0 + 0 + 0 + 0 + 0 + x$$

State matrix gives

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \dot{x}_4\\ \dot{x}_5\\ \dot{x}_6\end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ -\frac{1}{2.5} & -\frac{3}{2.5} & -\frac{7.5}{2.5} & -\frac{10}{2.5} & -\frac{10.5}{2.5} & -\frac{6}{2.5}\end{bmatrix}$$

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \dot{x}_4\\ \dot{x}_5\\ \dot{x}_6\end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ -0.4 & -1.2 & -3 & -4 & -4.2 & -2.4\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6\end{bmatrix} + \begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}\ddot{U}(t)$$

$$\boldsymbol{y} = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}$$

3.2. Stability

Whenever a bounded input gives a bounded output, the system is stable. This is a function of the output/input relationship in the system, whenever there are disturbances in the system, this relationship coordinates the internal performance of the system and decides its stability status. The following are methods for determining the stability of a system.

i. Root Locus

$$General\ Characteristic\ equation = 1 \pm G(s)H(s) = 0$$

$$TF = G(s)H(s) = \frac{numerator\ (Nr)}{denominator\ (Dr)} = \frac{K(s+a)(s+b)\ldots\ldots\ldots}{s^n(s+a_1)(s+b_1)\ldots\ldots\ldots} \quad (10)$$

Zeros are the value of S at the numerator, $when\ Nr = 0;\ s = -a, -b \ldots\ldots\ldots$
while poles are the value of S at the denominator, $when\ Dr = 0;\ s = 0, -a_1, -b_2 \ldots\ldots\ldots$

The equation would be split into two to determine the equating angles and the magnitude.

$$|G(s)H(s)| = 1; the\ magnitude\ creteria$$

$$< G(s)H(s) = 1; the\ angle\ creteria$$

ii. Nyquist Criterion

$$TF = \frac{G(s)}{1 \pm G(s)H(s)} = \frac{G(s)}{1 \pm L(s)}$$

$$L(s) = loop\ gain = G(s).H(s)$$

$$1 \pm L(s) = 0 \quad (11)$$

For the system to attain stability the poles will be on the left half of the s-plane.

iii. Routh Hurwitz Criterion

The characteristics equation for equation (3) is required for the R-H stability determination.

$$General\ Characteristic\ equation = 1 \pm G(s)H(s) = 0$$

$$1 \pm G(s)H(s) = 0 = a_0 s^n + a_1 s^{n-1} \ldots\ldots\ldots + a_{n-1} s^1 + a_n$$

$$Characteristic\ equation = a_0 s^n + a_1 s^{n-1} \ldots\ldots\ldots + a_{n-1} s^1 + a_n \quad (12)$$

Routh Hurwitz Criterion is further evaluated using the Routh Array (12)

(a) (b) Controllability

$$(A, B) = [B|AB|A^2B| \ldots\ldots |A^{n-1}B] \quad (13)$$

(b) (c) Observability

$$(C, A) = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \quad (14)$$

3.3. Simulation

The model was simulated in MATLAB/Simulink to obtain and validate the system parameters emanated from the mathematical model of the intelligent master controller. The Simulation result is presented and discussed in section IV.

4. Result of Intelligent Master Controller Modelling and Simulations

The result obtained from the modelling and simulation environment, MATLAB/Simulink, is presented in this section. In a bid to ascertain the system stability status, the following stability criteria were considered in view of selecting the most suitable outcome. Furthermore, the internal behavior of the systems was ascertained through the observation of the controllability and observability of the modelled system. The mathematical representation unveiled that the model has a high order differential equation showing that it is a higher order control system. Time response of the system showed that it was critically damped in view of its unit step function.

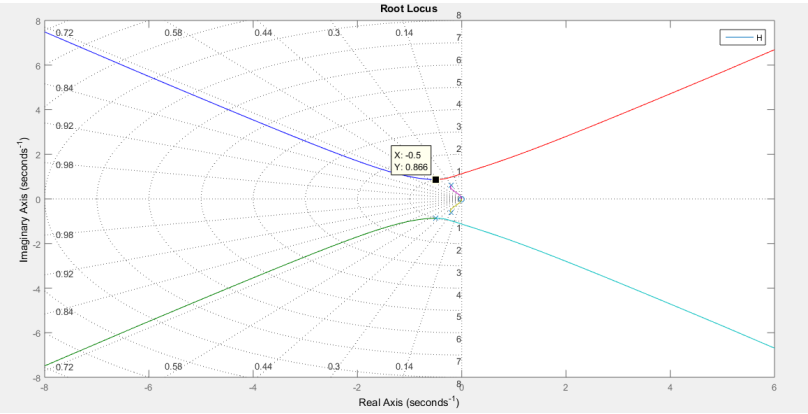
4.1. Intelligent Master Controller Stability Analysis

4.1.1. The Root Locus Plot

The stability of the intelligent master controller is described by the root locus plot in Figure 2. Condition for stability holds true when the poles of a system's characteristic equation lies on the negative half plane of the root locus to the magnitude of -0.5 and false for poles on the positive half. From the result obtained from Figure 2, the system can be said to be stable since the poles exist at the negative axis of the plot. The performance gain of the system can also be computed from the plot.

4.1.2. Nyquist Plot

Figure 3 shows the Nyquist plot of the complex margin gain (dB) of 8.32 and all frequencies of the phase (rad/s) of 1.13. This is an indication that the closed loop intelligent master controller is stable.



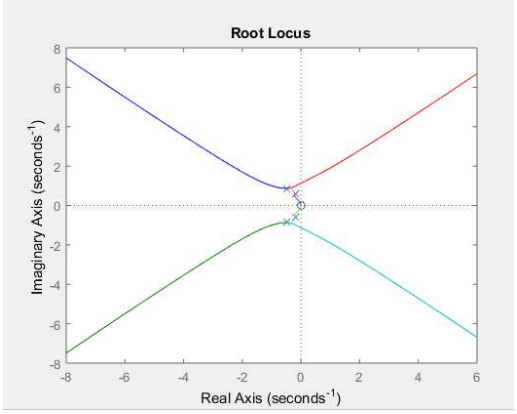
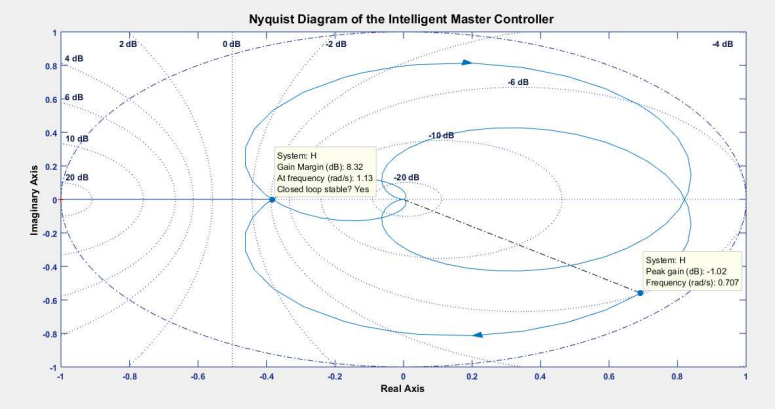


Figure 2. *Root Locus Plot of the intelligent master controller.*



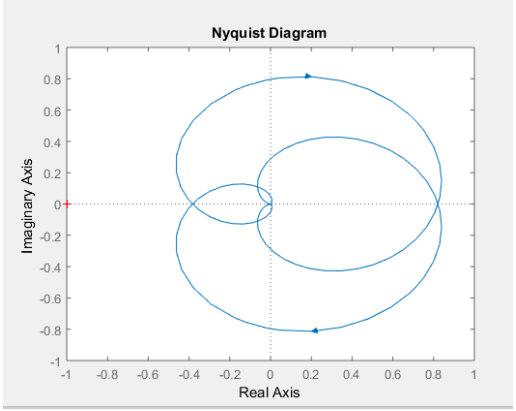


Figure 3. *Nyquist Plot of the intelligent master controller.*

4.1.3. Routh Hurwitz (RH) Stability Modelling

The denominator of the transfer function gives the higher order polynomial as its characteristic equation for the intelligent master controller model. The Routh Hurwitz (RH) stability test with the higher order polynomial equation (8) below gives the results in the Array in Table 1, this shows that the system is stable.

$$2.5S^6 + 6S^5 + 10.5S^4 + 10S^3 + 7.5S^2 + 3S + 1 = 0$$

All the coefficient of characteristic polynomial has same sign; thus, the equation has fulfilled the RH criterion for stability assessment.

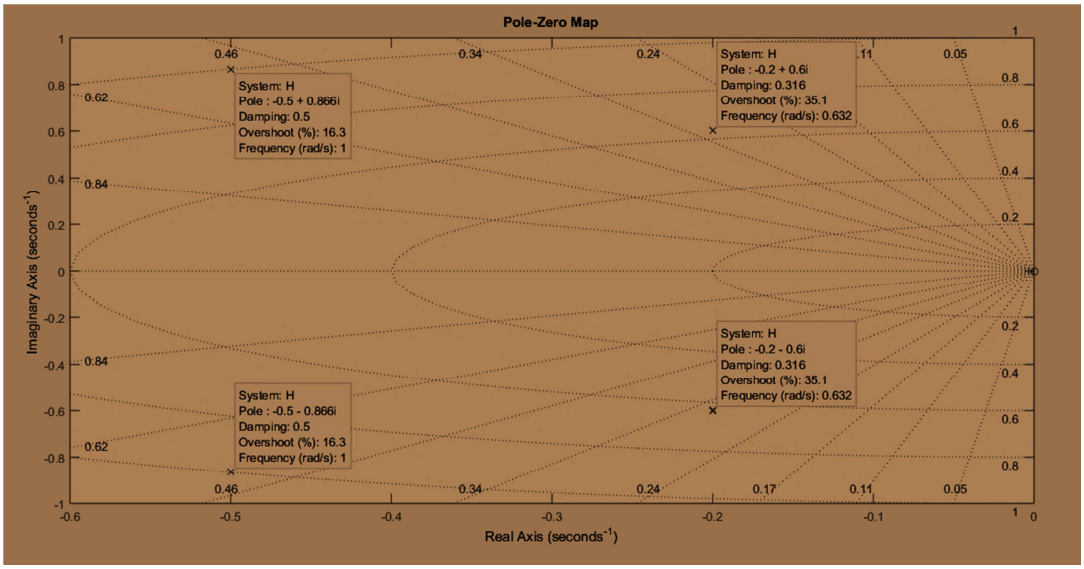


Figure 4. *Poles and Zero Plots of the intelligent master controller.*

Table 1. *The Routh Hurwitz Array.*

S^6	2.5	10.5	7.5	1
S^5	6	10	3	0
S^4	6.33	6.25	1	0
S^3	4.08	2.05	0	0
S^2	3.06	1	0	0
S^1	0.72	0	0	0
S^0	1	0	0	0

In Figure 4 the poles and zero plots for the intelligent master controller, the first two poles are located at the left-half S-Plane to the magnitude of -0.5 each with the damping of 0.5, the percentage overshoot of 16.3 and frequency of 1rad/s. The second two poles are located at the left-half S-Plane to the magnitude of -0.2 with the damping of 0.316, percentage overshoot of 35.1 and frequency of 0.632.

4.2. Controllability

The system Controllability test was carried out on the intelligent master controller model using MATLAB and the results shows that the develop system was stable. This result in Table 2, validates the controllability matrix condition in equation 13 and [17, 3].

Table 2. *Matrix Array of the Intelligent Master Controller System Controllability.*

Controllable Matrix is Co =					
0	0	0	0	0	1.0000
0	0	0	0	1.0000	-2.4000
0	0	0	1.0000	-2.4000	1.5600
0	0	1.0000	-2.4000	1.5600	2.3360
0	1.0000	-2.4000	1.5600	2.3360	-5.5584
1.0000	-2.4000	1.5600	2.3360	-5.5584	3.2890
Given System is Controllable					

4.3. Observability

The system observability test was carried out on the intelligent master controller model using MATLAB and the results shows that the develop system was observable. This result in Table 3. validates the observability Matrix condition in equation (14) and [3].

Table 3. *Matrix Array of the Intelligent Master Controller System Observability.*

Observable Matrix is Ob =					
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
Given System is Observable					

4.4. Step Unit Response of the Intelligent Master Controller with MATLAB

The unit step input response model in equation (7) presents the transient, steady state and disturbance status of the intelligent master controller. Figure 5 shows the output step response for a higher order system model, which is the time domain performance characteristic of the intelligent master controller model. The following parameters were deduced from the response, settling time of 29.5s, rise time of 0s, peak overshoot of infinite value at a time 6.91s, peak amplitude of -0.213 and a final steady state time of 0s. The stated data as collated from the figure 5, indicates the property of the system model to attain a settling time due to a delay in the transient response of the system. In conformity with the condition of critical dampness, the system roots of the intelligent master control provide stable result. This is in validation of the stability preposition by researcher in [11, 3, 7, 6] for the multiple power source scenario.

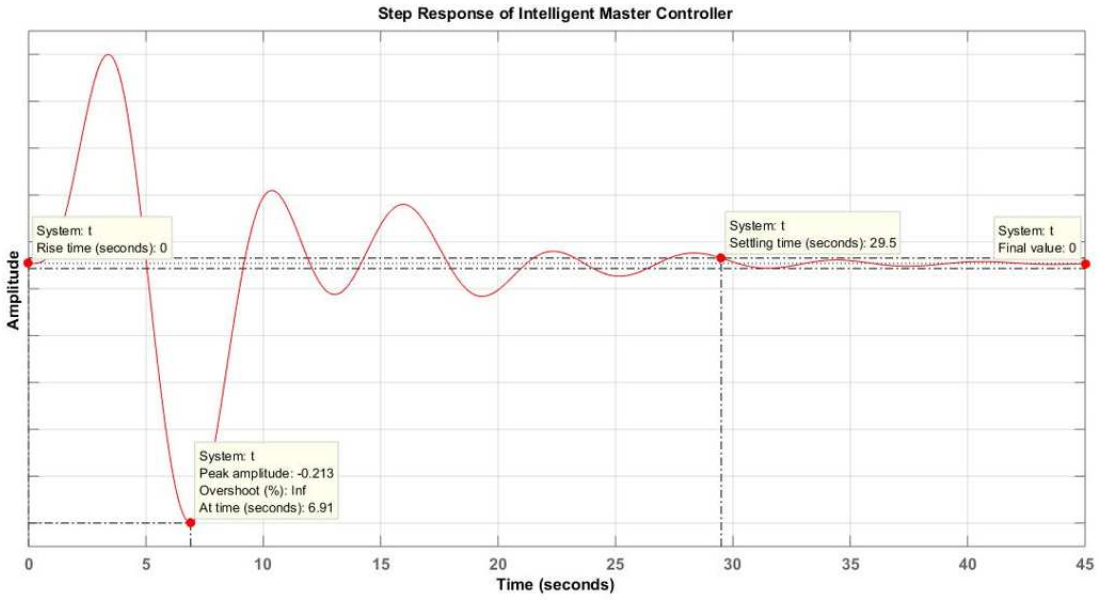


Figure 5. *System response of the intelligent master controller with MATLAB.*

4.5. The Intelligent Master Controller Model with Simulink

The unit step input response model in equation (7) was inputted into the Simulink model in figure 6 and the result from the scope is presented in Figure 7.

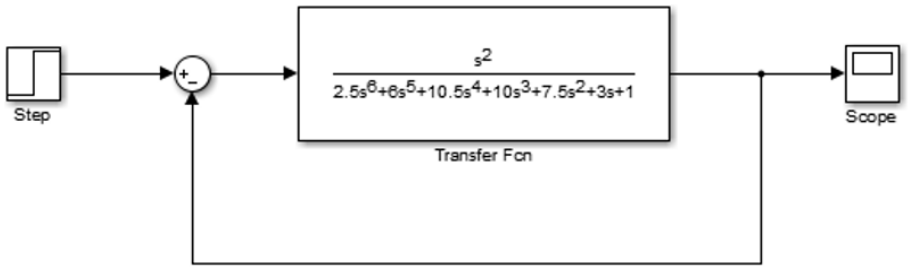


Figure 6. *System Model of the intelligent master controller with Simulink.*

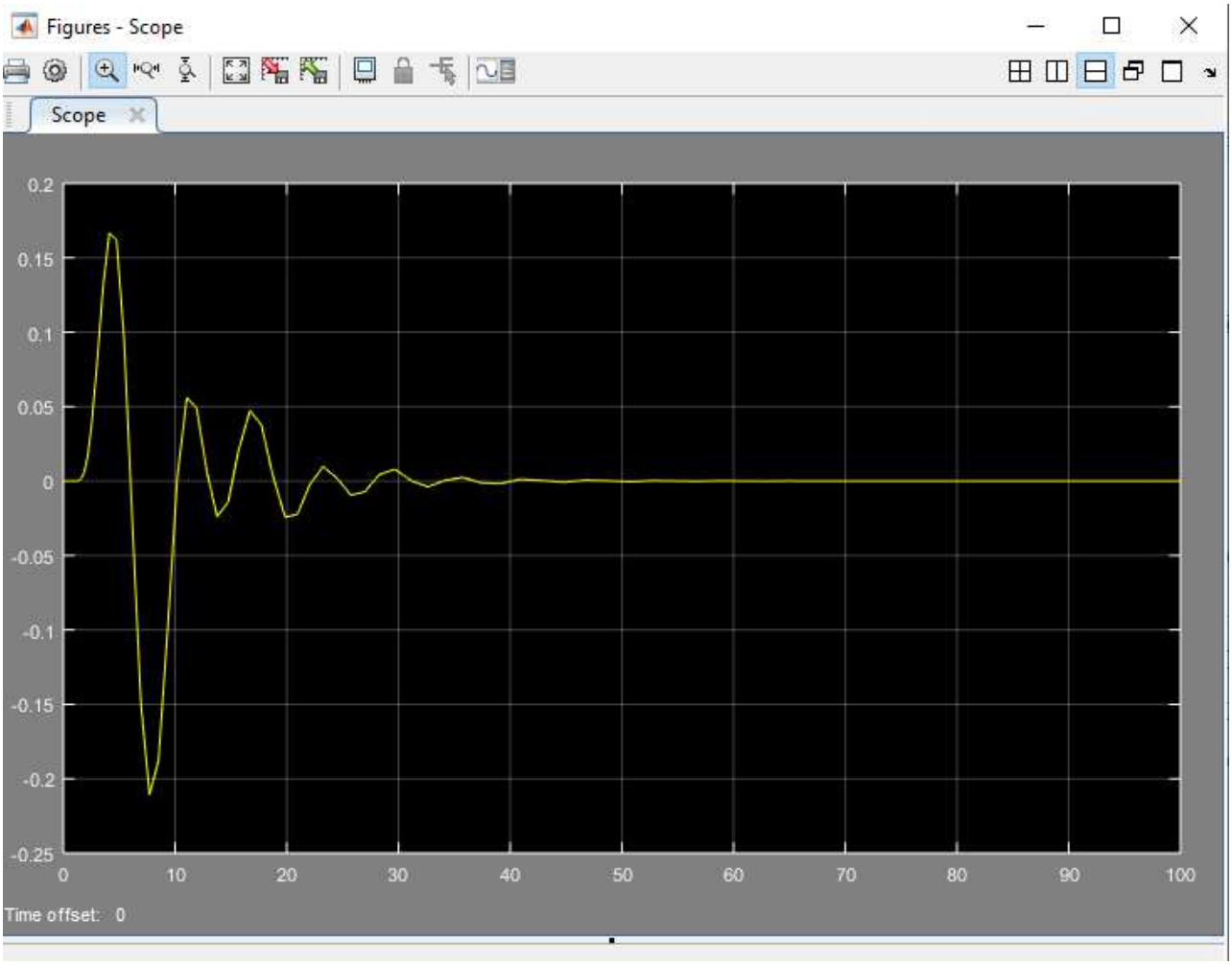


Figure 7. *System response of the intelligent master controller with Simulink.*

In Figure 5 and Figure 7, the System response of the intelligent master controller for both the MATLAB model and that of the Simulink are the same, this validates the performance of the system in term of its time response.

5. Conclusion

The intelligent master controller for hybridized power pool deployment was modelled. The mathematical model for the intelligent master controller to be integrated in the hybridized power pool deployment was derived, and simulations analysis were done using MATLAB/Simulink. The system stability was determined using Root locus, Nyquist criterion and Routh Hurwitz criterion. Furthermore, tests for observability and controllability that were carried out on the intelligent master controller model using MATLAB, shows that the system is stable.

References

[1] Heussen, K., Saleem, A. & Lind, M. (2009). Control architecture of power systems: Modeling of purpose and function. IEEE Power and Energy Society General Meeting, PES '09, 2009, 1–8.

[2] Zhang, W., Zhu, Q., Mobayen, S., Yan, H., Qiu, J. & Narayan, P. (2020). U -Model and U -Control Methodology for Nonlinear Dynamic Systems. Complexity, 1–13.

[3] Alexandridis, A. T. (2020). Modern Power System Dynamics, Stability and Control. Energies, 13 (15), 1–8.

[4] Michaels, L., Pagerit, S., Rousseau, A., Sharer, P., Halbach, S., Vijayagopal, R., Kropinski, M., Matthews, G., Kao, M., Matthews, O., Steele, M., & Will, A. (2010). Model-based systems engineering and control system development via virtual hardware-in-the-loop simulation. *SAE Technical Papers*, *October*. https://doi.org/10.4271/2010-01-2325

[5] Mu, S., Wang, H., Hou, J., & Wang, T. (2019). Fast Electromagnetic Transient Simulation Model of Photovoltaic Power System. *2018 International Conference on Power System Technology, POWERCON 2018 - Proceedings*, *201804270000731*, 286–291. https://doi.org/10.1109/POWERCON.2018.8602037

[6] Jack, K. E., Affia, N. J., Onu, U. G., Okekenwa, Ezugwu, E. E. O. & Udofa, E. S. (2021). Real time energy monitoring and control model for peer-to-peer integrated hybrid supply system," in 2021 IEEE PES/IAS PowerAfrica, PowerAfrica 2021, 1–5.

[7] Ezugwu, E. O., Aniagu, C. E., Okozi,, S. O., Jack, K. E. & Uzoechi, L. O. (2021). Design and Evaluation of a Hybridized Energy System for Remote Palm Oil Mill and Community. International Journal Mechatronics, Electrical Computer Technology, 11 (40), 4983–4988.

[8] Neumann, F. and Brown, T. (2020). The near-optimal feasible space of a renewable power system model. Electrical Power System Research., 190 (0), 106690.

[9] Hu, M., Fu, C., Wang, J. S., Rao, H., & Liu, H. B. (2012). Real time digital simulation of the parallel multi-terminal HVDC transmission system. *2012 IEEE International Conference on Power System Technology, POWERCON 2012*, 1–5. https://doi.org/10.1109/PowerCon.2012.6401417

[10] Dasu, B., Mangipudi, S. & Rayapudi, S. (2021) Small signal stability enhancement of a large scale power system using a bio-inspired whale optimization algorithm. Prot. Control Mod. Protection and Control of Modern Power System, 6 (35), 1–17.

[11] Olubiwe M.& Uzoechi, L. O. (2014). Modelling of Three Phase Short Circuit and Measuring Parameters of a Turbo Generator for Improved Performance," International Journal of Engineering Research Technology, 3 (4), 2264–2269.

[12] Lin, J., Tong, X., Wang, X., & Wang, W. (2008). Parallel simulation for the transient stability of power system. *3rd International Conference on Deregulation and Restructuring and Power Technologies, DRPT 2008*, *April*, 1325–1329. https://doi.org/10.1109/DRPT.2008.4523611

[13] Meng, H., & Song, X. (2012). The modeling and simulation of command and control system based on capability characteristics. *Communications in Computer and Information Science*, *327 CCIS* (PART 2), 255–261. https://doi.org/10.1007/978-3-642-34396-4_31

[14] Naşcu, I., De Keyser, R., Naççu, I., & Buzdugan, T. (2010). Modeling and simulation of a level control system. *2010 IEEE International Conference on Automation, Quality and Testing, Robotics, AQTR 2010 - Proceedings*, *1*, 181–186. https://doi.org/10.1109/AQTR.2010.55208

[15] Matej Krpan & Igor Kuzle (2018) Introducing low-order system frequency response modelling of a future power system with high penetration of wind power plants with frequency support capabilities. *The Institution of Engineering and Technology (IET): Renewable Power Generation,* 1-9.

[16] Qing, K., Xi, X., Zanxiang, N., Lipei, H., & Kai, S. (2013). Design of grid-connected directly driven wave power generation system with optimal control of output power. *2013 15th European Conference on Power Electronics and Applications, EPE 2013*. https://doi.org/10.1109/EPE.2013.6631811

[17] Dritsas, L., Kontouras, E., Vlahakis E., Kitsios, I, Halikias, G. & Tzes, A. (2020). Modelling issues and aggressive robust load frequency control of interconnected electric power systems. International Journal Control, 0 (0), 1–15.

[18] Mora, E. & Steinke, F. (2021). On the minimal set of controllers and sensors for linear power flow. Electr. Power Syst. Res., 190 (0), 106647.

[19] Yang-Wu, S., Xun, M., Ao, P., Yang-Guang, W., Ting, C., DIng, W., & Jian, Z. (2019). Load Frequency Control Strategy for Wind Power Grid-connected Power Systems Considering Wind Power Forecast. *2019 3rd IEEE Conference on Energy Internet and Energy System Integration: Ubiquitous Energy Network Connecting Everything, EI2 2019*, 1124–1128. https://doi.org/10.1109/EI247390.2019.906208

Calculation of Losses in Transmission System in Dependence on Temperature and Transmitted Power

Prativa Barik, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, p.barik213@gmail.com*

Romeo Jena, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, romeo_jena2@hotmail.com*

Balagoni Sampath Kumar, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, bs.kumar456@gmail.com*

Ajaya Kumar swain, *Department of Electrical and Electronics Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ajayaswain9@gmail.com*

Abstract: Innovations and trends has been significantly increased during modern theory and practical realizations in the field of energetic. In the Czech Republic, the research of predictions of technical losses on the transmission system can be considered as novel and important topic. Using software possibilities can be appropriately utilized in the frame of estimations of the technical losses. While they cannot be eliminated, they may be minimized. Losses can be measured or calculated using transmission-line parameters. This causality is considered in the form of the presented and proposed mathematical equations including real measured data of the atmospheric temperatures achieved on the substations selected in the Moravian-Silesian region. In this contribution, results of proposed calculations of technical losses based only on line parameters taking into account the ambient temperature are being compared in relation to a particular transmission system using prediction software. Particularly, technical losses caused by a configuration change of a selected part of a transmission system are considered related to the operation of Dlouhé Stráně pumped storage hydro power plant. As can be conclude, after the resulted comparisons using by the proposed mathematical models in software and the obtained real measured data, general minimization of the losses is necessary to create the most accurate models of the states that might occur in the future and to propose required modifications of the given part of the transmission system. A future bounded research can be focused on the sensors situated on the transmission lines instead of the substations.

Keywords: Transmission System, Technical Losses, Prediction, Software

1. Introduction

In the research area of the energetic [1-3], the modern approaches and trends has been frequently occurred with proposals of modifications in favor of the minimization of external influences as losses or noises. As near research areas of the solving these occurred problems, also, the technical cybernetics [4-6] and mathematical modelling of processes control [7-9] has been often considered.

Particularly, in this contribution, the calculation of predictions using by the software utilities with following evaluation of losses [10] that occur in the transmission system located in a certain part of the Czech Republic [11]. In the paper proposals, the authors' own realized software is utilized for purposes of the calculating the predictions.

The calculation has been performed with the program which inputs are the measured values obtained from databases of the transmission grid control system [12]. The results of the calculation can be then suitably compared with values of losses of a second program that calculates the technical losses based only on the line parameters. It is then possible to assess the impact of the losses on the examined transmission system in the area.

The specific area of the transmission system has been selected in view of the interesting states that can occur during its operation, especially greater variations of atmospheric temperatures and fluctuations of the transmitted power. [13-15]

It is an area in which provision of an optimal power to the Horní Životice substation posed problems in previous periods. The capacity of the area has been reinforced by the construction of a new Kletná substation and by the erection

of another V458 transmission line. An important aspect that played a role in selecting the examined area has been the commissioning of a new transmission line between the Horní Životice and Krasíkov substations. This resulted in the creation of a ring transmission system network boosted by power fed from the Dlouhé Stráně pumped storage hydro power plant. In terms of the transmission system operation management the power supplied by this power plant is variable. The paper also mentions the states under which this power plant is utilized with respect to its operation and the atmospheric temperature. [13-15]

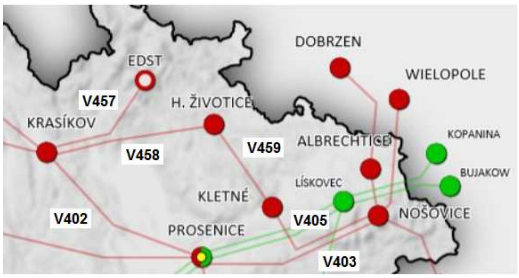


Figure 1. *The selected area of the transmission system used for the calculations and analysis.*

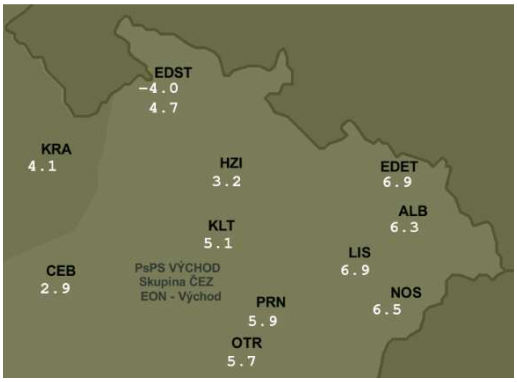


Figure 2. *A sample of measured atmospheric temperature data in the area under investigation.*

Losses can be measured or calculated using transmission-line parameters. This causality is considered in the form of the presented and proposed mathematical equations including real measured data of the atmospheric temperatures achieved on the substations selected in the Moravian-Silesian region. In this contribution, results of proposed calculations of technical losses based only on line parameters taking into account the ambient temperature are being compared in relation to a particular transmission system using prediction software. Particularly, technical losses caused by a configuration change of a selected part of a transmission system are considered related to the operation of Dlouhé Stráně pumped storage hydro power plant. [13-15]

2. Selected Transmission System Network

The calculations are based on real data and calculations using a program developed in previous years [13, 14]. The selected area of the transmission system is shown in Figure 1 and the atmospheric temperature data in Figure 2.

2.1. Description of the Selected Network

The selected area of the transmission system comprises six nodes, five of which are substations and the sixth the controlled power hydroelectric power plant. The area network forms a ring, which consists of six overhead lines, see Table 1 showing their length. The selected network is also connected by seven lines to a neighboring transmission system of Czechia, Slovakia and Poland. The default values used for the calculations are data measured in the power dispatching control system (Table 2) and include node voltages, information on the transmitted power, reactive power, line current and temperatures. An important input is the power contributed into the selected area by the pumped storage hydro power plant. An emphasis is placed on selected seasons and atmospheric temperature changes in the region, which are measured directly at power utilities. A database has been compiled based on all parameters of the lines and the measured data that are used to perform the calculations with the aid of the program. The program was developed by the staff of two Ostrava universities and has been published [13, 14]. The calculations have been performed due to the need to verify the accuracy of software calculations and to clarify changes in the magnitude of losses with respect to atmospheric temperature movements and the line transmitted power in the area where network configuration changes have been made. The changes included commissioning of a new V458 line, construction of a new 400 kV Kletné substation and creation of a V405 line as a result of splitting the V459 into two V405 and V459 lines terminated in the new Kletná substation. The results and evaluations are set out further in the article.

Table 1. *Lengths of the selected network lines.*

400 kV line	Substation 1	Substation 2	Line lengthí
V457	Dlouhé Stráně	Krasíkov	59.8 km
V458	Krasíkov	Horní Životice	107 km
V459	Horní Životice	Kletné	42.1 km
V402	Krasíkov	Prosenice	87.6 km
V403	Prosenice	Nošovice	79.6 km
V405	Nošovice	Kletné	53.5 km

2.2. Measured Values Database Analysis

The measured value database contains the values of transmitted active and reactive power (P; Q), technical losses for the given line (P_ztr), voltage (U), current (I) and temperatures from the power utilities (T_venk). All values, apart from technical losses, were measured at the start and end of the lines of the respective substation. Database data between August 2017 and February 2018 were used for processing. The measurement databases also include seasonal differences according to a given month and are divided into the summer season - L and the winter season - Z. The measured values are then divided into columns, where the respective measured quantity for the given line is shown in a separate column. The header of each column contains an abbreviation for the substation outlet, the code designation of the line and finally the symbol of the measured

quantity. Take for example the designation C: KRA4: V402: *P*, where KRA4 stands for the measurement taken at Krasíkov station. From the line designation of V402 it can be inferred that it is a 400 kV line, as this designation starts with the number 4. For the 220 kV line the designation then starts with the number 2. The last part consists of a letter. For example, *P* means active power values. The sample of the database section is shown in Table 2.

Table 2. *A part of the database of measured valuaes for v402 line.*

Time	C: KRA: 4: V402: *P*	C: KRA: 4: V402: *Q*	C: KRA: 4: V402: *U*	C: KRA: 4: V402: *I*	C: KRA: *T*_venk
10.12.2017 24:00:00 Z	-150.36	27.61	418.52	211.16	0
11.12.2017 00:15:00 Z	57.13	19.76	417.29	94.23	0.59
11.12.2017 00:30:00 Z	68.49	20.93	417.38	102.8	0.77
11.12.2017 00:45:00 Z	56.7	20.2	417.16	84.69	0.83
11.12.2017 01:00:00 Z	94.02	22.11	417.71	143.78	1.09
11.12.2017 01:15:00 Z	99.39	16.79	417.06	151.97	1.14

2.3. Analysis of Calculated Joule Losses of the Selected Transmission System Network

The calculations of the selected network are based on the real values measured by the sensors that are part of the energy dispatch control system in the given power utility. An example of the location of the sensors in the selected region is shown in Figure 2. The values have been selected from certain periods since 2017. These values serve as a basis for calculations and analysis of the selected area.

3. Model Statuses for Calculating Losses of Selected Transmission System Lines

The program for calculating Joule's losses works on the following principle: it takes the long-term measurements and uses it to assemble the predictive polynomial to calculate the losses in the selected temperature interval for the specified transmitted power [13]. The boundary temperatures of the temperature interval are chosen so that the one polynomial represents losses at the low temperatures and the second one losses at the higher temperatures. The transmitted power values are selected between 0 MW and 100 MW up to the maximum transmitted power [13, 15-18]. The V402 line Joule losses e.g. at a transmitted power of 500 MW taken from the line parameters and an assumed power factor of $\cos\phi = 0.95$ are calculated from the values given in Table 2 as follows - first we calculate the current then use it to arrive at the reactive power:

$$I = \frac{P}{\sqrt{3}\cdot U\cdot\cos\phi} = \frac{500\cdot 10^6}{\sqrt{3}\cdot 400\cdot 10^3\cdot 0.95} = 760\,\text{A} \tag{1}$$

$$\begin{aligned} Q &= \sqrt{3}\cdot U\cdot I\cdot\sin\phi \\ &= \sqrt{3}\cdot 400\cdot 10^3\cdot\sin(\arccos(0.95)) = 164\,\text{Mvar} \end{aligned} \tag{2}$$

The Joule's loses are then:

$$\begin{aligned} \Delta P &= R\cdot\frac{P^2+\left(Q+\frac{V^2\cdot B}{2}\cdot 10^{-6}\right)^2}{V^2} \\ &= 2.57\cdot\frac{500^2+\left(164+\frac{400^2\cdot 354}{2}\cdot 10^{-6}\right)^2}{400^2} = 4.029\text{MW} \end{aligned} \tag{3}$$

3.1. Analysis and Calculations for V402 Krasíkov – Prosenice Line

To calculate Joule's losses using the program, we select two temperature intervals, one for the winter and the other for the summer period

ΔT_1 between -14°C and 0°C and

ΔT_2 between 10°C and 35°C.

Then the resulting prediction polynomials for the selected temperature ranges are:

$$\Delta P_{T1} = 0.01203 + 0.00019\cdot P + 1.4\cdot 10^{-5}\cdot P^2 \tag{4}$$

$$\Delta P_{T2} = 0.014095 + 00011\cdot P + 1.5\cdot 10^{-5}\cdot P^2 \tag{5}$$

In Table 3 By program calculated Joule's loss results and the losses from the line parameters are then compared.

Table shows that the program-predicted losses in dependence on temperature are lower than the losses calculated by the program without taking into account the ambient temperature. These values are more realistic because they include the influence of the ambient temperature and the program is based on a comprehensive database of measured values, contrary to the losses calculated only from the line parameters that do not comprise the influence of the temperature [13, 14, 18-21].

Table 3. *The resulting joule's losses of the v402 line of the selected part of the transmission system network.*

Software calculated losses			**Losses from the line parameters**		
P	ΔP_{T1}	ΔP_{T2}	***I***	***Q***	**ΔP**
(MW)	(MW)	(MW)	(A)	(Mvar)	(MW)
0	0.012	0.015	0	0	0.013
100	0.176	0.173	152	33	0.174
200	0.629	0.626	304	66	0.655
300	1.371	1.373	456	99	1.459
400	2.402	2.414	608	131	2.583
500	3.722	3.751	760	164	4.029
600	5.331	5.381	912	197	5.795
700	7.229	7.307	1064	230	7.884
800	9.416	9.527	1215	263	10.293
900	11.893	12.041	1367	296	13.024
1000	14.658	14.85	1519	329	16.075
1100	17.713	17.954	1671	362	19.449
1200	21.056	21.352	1823	394	23.143

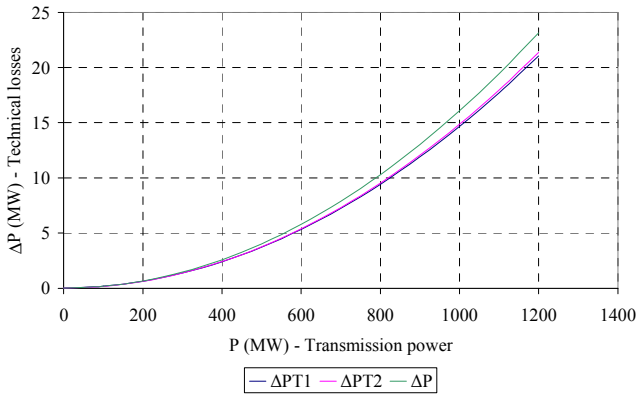


Figure 3. *Technickal loses of the V402 line.*

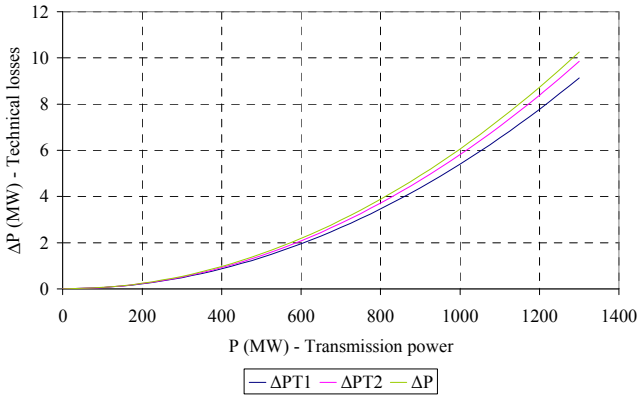


Figure 4. *Technical loses of the V459 line.*

The biggest losses in the course of the year were recorded in September, when the peak transmitted power was 800 MW, and the lowest one in February with the transmitted power of up to 600 MW. The graphical comparison of all three values of losses is shown in Figure 3 graph.

3.2. Horní Životice – Kletné V459 Line

For all subsequent lines, calculations are performed at the same temperature intervals as for the V402 line.

$$\Delta P_{T1} = 0.000312 + 10^{-5} \cdot P + 5.4 \cdot 10^{-6} \cdot P^2 \quad (6)$$

$$\Delta P_{T2} = 0.00454 - 6 \cdot 10^{-5} \cdot P + 5.9 \cdot 10^{-6} \cdot P^2 \quad (7)$$

The difference in predicted losses, taking into account the temperature and loss calculated only with respect to the transmitted power (ΔP - ΔP_{T1} and ΔP - ΔP_{T2}) is increasing. At the transmitted power of 1300 MW, the difference is 1.12 MW at low temperatures and 0.4 MW at high temperatures.

3.3. Dlouhé Stráně – Krasíkov V457 Line

This line is connected to the Dlouhé Stráně pumped storage hydro power plant with the installed capacity of 600 MW.

$$\Delta P_{T1} = 0.00762 + 5.6 \cdot 10^{-5} \cdot P + 8.9 \cdot 10^{-6} \cdot P^2 \quad (8)$$

$$\Delta P_{T2} = 0.01599 - 4.7 \cdot 10^{-4} \cdot P + 7.3 \cdot 10^{-6} \cdot P^2 \quad (9)$$

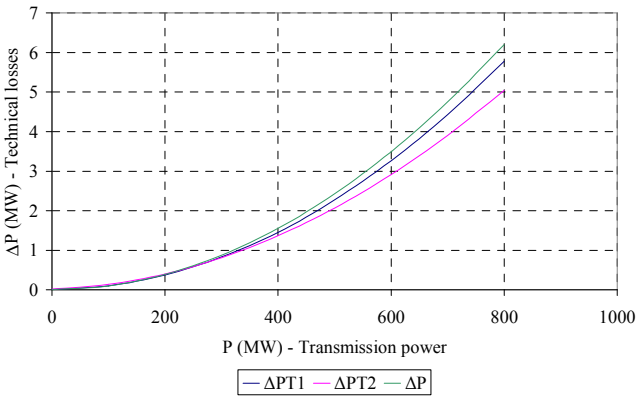


Figure 5. *Technical losses of V457 line.*

The V459, together with the V405 line formed the so-called radial network until the V458 line was connected. The Dlouhé Stráně power plant transmitted power at higher temperatures ranged around 300 MW; the maximum of 600 MW was rarely achieved. Its transmitted power was influenced by the needs of the transmission system. Therefore, the higher temperature prediction is more accurate only up to the transmission power of 300 MW, for higher transmitted power it is distorted because of the small data rate in the default database of measured values. At low temperatures, the prediction is accurate because the peak power values were much more frequent during this period, and therefore there were a sufficient number of data lines to calculate losses for the transmitted power above 300 MW.

3.4. Krasíkov – Horní Životice V458 Line

This is a new line erected in connection with the construction of a new 400 kV substation in Kletná.

$$\Delta P_{T1} = 0.01769 - 4 \cdot 10^{-5} \cdot P + 1.4 \cdot 10^{-5} \cdot P^2 \quad (10)$$

$$\Delta P_{T2} = 0.19254 - 9 \cdot 10^{-4} \cdot P + 1.6 \cdot 10^{-5} \cdot P^2 \quad (11)$$

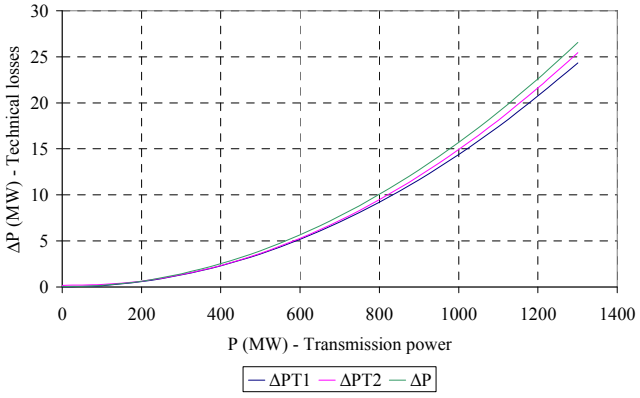


Figure 6. *Technical losses of V458 line.*

The difference in predicted losses, considering the temperature and loss calculated only for the transmitted power (ΔP - ΔP_{T1} and ΔP - ΔP_{T2}), is again rising. At the transmitted power of 1300 MW the difference is 1.2 MW at low temperatures and 1.1 MW at high temperatures.

4. Influences on the Dlouhé Stráně Power Plant Operation by Connecting the New V458 Line

In terms of the transmitted power over the respective lines, in relation to the technical losses and the atmospheric temperature the situation changed in the selected area of the transmission system after terminating the new V458 line connecting the Krasíkov substation to the Horní Životice substation. Prior to the line termination the substation formed the end of a radial network that comprised Nošovice - Kletné - Horní Životice substations. (V405, V459). The V402 and V403 lines experienced a significant drop in technical losses. The termination of the new V458 line also changed the direction and size of the transmitted power by the V405 and V459 lines from the Horní Životice substation to the Nošovice substation. With the increased power transmission in the area, the technical losses also increased. The transmitted power from the Krasíkov substation was split into two directions, along the V402 line and along the new V458 line. It can be stated that the technical losses have been divided and their total size has been reduced. The size of the technical loses is also notably influenced by the Dlouhé Stráně pumped storage hydro power plant operation. After connecting the new V458 line, the power plant started supplying more power to the selected area [16-18].

5. Conclusion

In conclusion we can state that many factors have changed in the selected area of the transmission system and those factors significantly influenced the size of the technical losses. The most important factors include the new network configuration, temperature influences, the power transmitted over the lines and the Dlouhé Stráně pumped storage hydro power plant operation. The technical losses calculated from the line parameters were compared and, in the latter case, calculated using a special program.

The losses calculated by using the program that takes temperature variations into account were lower on all lines. We can say that the results obtained by the program that takes into account the outdoor temperature are more realistic because they are based on the analysis of long-term measurements performed on the relevant lines and include the temperature influence. The program can be used for further analysis of losses in the transmission system networks and its use will result in correct analyzes and conclusions.

The computational program used could also find its use in practice. The analysis of the transmission system lines' technical losses is an important indicator for the economical evaluation of electricity transmission and also serves to evaluate changes taking place in certain parts of the system after construction of new substations and lines. In order to minimize the losses, it is necessary to create the most accurate models of the states that might occur in the future and to propose required modifications of the given part of the transmission system.

A future bounded research can be focused on the sensors situated on the transmission lines instead of the substations.

References

[1] Mahfouz, M. M. A., Alsumiri, M., & Althomali, R. (2021). Efficient Power Utilization Control Scheme for Hybrid Distribution Generation Grid. *Journal of Electrical and Electronic Engineering*, *9* (*1*), 26–32. doi: 10.11648/j.jeee.20210901.14.

[2] Dong, Y., & Shi, Y. (2021). Analysis of Losses in Cables and Transformers with Unbalanced Load and Current Harmonics, *Journal of Electrical and Electronic Engineering, 9* (*3*), 78–86. doi: 10.11648/j.jeee.20210903.13.

[3] Macků, L. (2019). Determination of exothermic batch reactor specific model parameters. *MATEC Web of Conferences, 292.01063*. doi: 10.1051/matecconf/201929201063.

[4] Vojtesek, J., & Spacek, L. (2019). Adaptive Control of Temperature Inside Plug-Flow Chemical Reactor Using 2DOF Controller. In: Machado J., Soares F., Veiga G. (eds) *Innovation, Engineering and Entrepreneurship. HELIX 2018. Lecture Notes in Electrical Engineering, Vol 505*. Springer, Cham. doi: 10.1007/978-3-319-91334-6_15.

[5] Vojtesek, J., Spacek, L., & Gazdos, F. (2018). Control Of Temperature Inside Plug-Flow Tubular Chemical Reactor Using 1DOF And 2DOF Adaptive Controllers. In Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (eds*) ECMS 2018 Proceedings* doi: 10.7148/2018-0239.

[6] Navratil, P., Pekař, L., & Matušů, R. et al. (2021). Experimental Investigation and Control of a Hot-Air Tunnel with Improved Performance and Energy Saving. *ACS Omega*, *6*, 16194–16215. doi: 10.1021/acsomega.1c02239.

[7] Pekař, L., Strmiska, M., & Song, M., et al. (2021). Numerical Gridding Stability Charts Estimation using Quasi-polynomial Approximation for TDS. In *23rd International Conference on Process Control (PC)*, pp. 290-295. doi: 10.1109/PC52310.2021.9447521.

[8] Korenova L., Vagova R., Barot T., & Krpec R. (2020). Geometrical Modelling Applied on Particular Constrained Optimization Problems. In: Silhavy R., Silhavy P., Prokopova Z. (eds) *Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing, Vol. 1295*. Springer, Cham. doi: 10.1007/978-3-030-63319-6_16.

[9] Barot T., Krpec R., & Kubalcik M. (2019). Applied Quadratic Programming with Principles of Statistical Paired Tests. In: Silhavy R., Silhavy P., Prokopova Z. (eds) *Computational Statistics and Mathematical Modeling Methods in Intelligent Systems. CoMeSySo 2019. Advances in Intelligent Systems and Computing, Vol. 1047*. Springer, Cham. doi: 10.1007/978-3-030-31362-3_27.

[10] Rego, A., Pereira, J. A., & Almeida, A. (2019). DEVELOPMENT OF MODELS FOR ASSESSING HYDRO-ENERGETIC LOSSES IN WATER SUPPLY SYSTEMS. *Journal of Urban and Environmental Engineering, 13*, 209–218. doi: 10.4090/juee.2019.v13n2.209218.

[11] Muzik, V., & Vostracky, Z. (2020). Communication and Intelligent Control in a Power Grid Using Open Source IoT Technology. In 2*1st International Scientific Conference on Electric Power Engineering (EPE)*, pp. 1–4, doi: 10.1109/EPE51172.2020.9269216.

[12] Ruppert M., Slednev V., & Finck R., et al. (2020). Utilising Distributed Flexibilities in the European Transmission Grid. In: Bertsch V., Ardone A., Suriyah M., Fichtner W., Leibfried T., Heuveline V. (eds) *Advances in Energy System Optimization. ISESO 2018. Trends in Mathematics*. Birkhäuser, Cham. doi: 10.1007/978-3-030-32157-4_6.

[13] Rudolf, L., Král, V, & Šamaj, A. (2017). Software Solution for Optimisation of Transmission Network Operation. In *Proceedings of the 18th International Scientific Conference on Electric Power Engineering (EPE)*. Ostrava: VSB-TU Ostrava, pp. 29–33. ISBN 978-1-5090- 6405-2.

[14] Křemen, O. (2018). Užití databází měřených hodnot k analýze provozu vedení přenosové soustavy. *Diploma thesis*, VŠB-TU Ostrava, 2018.

[15] ČEPS a. s. *Transmission Network in the Czech Republic and Central Europe in 2013-2015 in the context of EWIS* [online]. [cit. 2019-02-02]. Available at: https://www.ceps.cz/en/studies-and-analyses.

[16] ČEPS a. s. *Grid Code* [online]. [cit. 2019-02-02]. Available at: https://www.ceps.cz/en/grid-code.

[17] ČEPS a. s. *Transmission System Timeline* [online]. [cit. 2019-02-02]. Available at: https://www.ceps.cz/en/transmission-system-timeline.

[18] *Measured Data*. [Database export]. ČEPS, a. s. (Czech Transmission System Operator), 2018.

[19] ETAP. *Load flow analysis* [online]. [cit. 2019-02-02]. Available at: https://etap.com/product/load-flow-software.

[20] Bamigbola, O. M., ALI, M. M., & Awodele, K. O. (2014). Predictive Models of Current, Voltage, and Power Losses on Electric Transmission Lines. *Journal of Applied Mathematics, March 2014*. doi: 10.1155/2014/146937.

[21] Sankaramoorthy, M., & Veluchamy, M. (2017). A hybrid MACO and BFOA algorithm for power loss minimization and total cost reduction in distribution systems. *Turkish Journal of Electrical Engineering and Computer Sciences, January 2017, 25 (1)*, 337–351. doi: 10.3906/elk-1410-191.

Fast Parallel FDFD Algorithm for Solving Electromagnetic Scattering Problems

Himanshu Sekhar Moharana, *Department of Mechanical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, hsmoharana26@gmail.com*

Prativa Barik, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, p.barik213@gmail.com*

Achyutananda Panda, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, achyutananda.panda55@gmail.com*

Laxminarayan Mishra, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, laxminarayan.s@gmail.com*

Abstract: The finite difference frequency domain (FDFD) method is very suitable for working out narrowband problems and resonance problems. However, the FDFD method needs to solve a large complex sparse matrix equation. With the increase of computing scale, the dimension of matrix will increase rapidly, which is difficult to simulate. For improving the computational efficiency of solving the large complex sparse matrix equation and extend the application scope of the FDFD method, a fast parallel FDFD method on the basis of message passing interface (MPI) shared memory technology is proposed in this paper, which is used to solve the electromagnetic scattering problems of electrically large targets. Based on the conjugate gradient iterative algorithm, the large complex sparse matrix is reasonably distributed to each process according to the unequal row allocation scheme, so as to guarantee the load balancing of each process. In addition, the intermediate vectors utilized in total processes are stored in the shared memory of MPI, which reduces the communication time and the consumption of computer memory. The proposed parallel FDFD method is employed to solve the bistatic RCS of the PEC sphere, composite Von warhead and an automobile, compared with the serial FDFD method, the parallel FDFD method greatly improves the computational efficiency when the memory is not increased much.

Keywords: FDFD, Complex Sparse Matrix, Conjugate Gradient Iteration, MPI, Shared Memory

1. Introduction

The finite difference time domain (FDTD) method is very popular in computational electromagnetics. It solves Maxwell's equations in the time domain and is very suitable for solving broadband electromagnetic problems [1]. However, when dealing with the electromagnetic problem with resonant structure, the FDTD method requires lots of time iterations to guarantee the accuracy of the simulation. Besides, the time step used in the FDTD method is constricted by the CFL condition [2], it is difficult for FDTD method to solving long pulse electromagnetic problems. For example, the duration of the E3 (late-time) waveform for the high altitude electromagnetic pulse (HEMP) is about 400s, while the grid size of the target is meter-scale. Because of the restriction of the stability criterion, the time step of the FDTD method should be nanosecond, then hundreds of millions of time steps need to be iterated to complete the simulation [3].

The FDFD method applies the differential Maxwell equation to describe electromagnetic field relationships. The central difference is employed to discretize the spatial partial derivative, and the time harmonic factor is used to replace the time partial derivative. Afterwards, the electromagnetic scattering properties of targets at a certain frequency can be acquired by solving the sparse matrix equation [4]. The FDFD method is not limited by the stability condition, so it has a great advantage in dealing with electromagnetic problems such as resonance and long pulse. In recent years, FDFD method has also been widely used in periodic structures [4], microstrip structures and waveguides [10, 13], long-path propagation [12], the band gap of the photonic crystal [5], bioelectromagnetic

uncertainty analysis [15], etc. Researchers are also studying the absorbing boundary condition [11], fast algorithm of the FDFD method [14, 16].

However, the FDFD method needs to solve a large complex sparse matrix equation. With the increase of computing scale, the dimension of matrix will increase rapidly, which is difficult to simulate. For improving the computational efficiency of solving the large complex sparse matrix equation and extend the application scope of the FDFD method, the parallel FDFD method is proposed in this paper. Based on the conjugate gradient method, a shared memory technology of the message passing interface (MPI) is adopted in the FDFD method, and the large sparse matrix is allocated reasonably to each process. Then, the load balancing is ensured and the solution efficiency of the large complex sparse matrix equation is improved.

The rest of this paper is structured as follows. The FDFD algorithm is described in Section II. The parallel FDFD scheme is explained in Section III. Its accuracy and efficiency are proved in details in Section IV. Section V closes the paper with some conclusions.

2. Review of the FDFD Method

The 3D Maxwell equation in frequency domain is

$$\begin{cases} \nabla \times \boldsymbol{E} = j\omega\mu_r\mu_0\boldsymbol{H} \\ \nabla \times \boldsymbol{H} = -j\omega\varepsilon_r\varepsilon_0\boldsymbol{E} \end{cases} \tag{1}$$

where E, H represents electric field and magnetic field in the calculation region, respectively. ω is the angular frequency. μ_r, μ_0, ε_r, ε_0 represents relative permeability, permeability of vacuum, relative permittivity and permittivity of vacuum, respectively.

The Yee's cells are adopted to discretize the FDFD computation domain, and electric field sampling points and magnetic field sampling points are staggered with each other. Each electric field is surrounded by four magnetic fields, and each magnetic field is also surrounded by four electric fields.

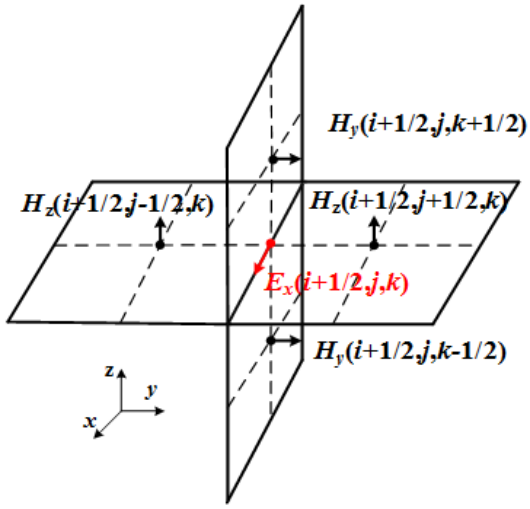
In the Cartesian coordinate system, the discretization equation related to the x direction's electric field $E_x(i+1/2, j, k)$ in (1) is

$$\begin{aligned} & j\omega\varepsilon_0\varepsilon_r(i+1/2, j, k)E_x(i+1/2, j, k) \\ &= \frac{H_z(i+1/2, j+1/2, k) - H_z(i+1/2, j-1/2, k)}{\Delta y} \\ &- \frac{H_y(i+1/2, j, k+1/2) - H_y(i+1/2, j, k-1/2)}{\Delta z} \end{aligned} \tag{2}$$

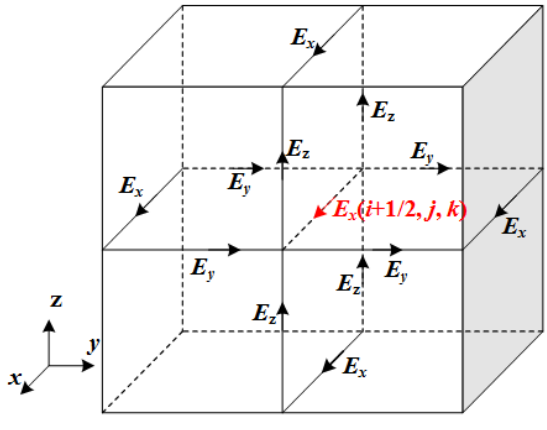
Similarly, we can get the other components of the electric field and magnetic field by discretizing (1) using central difference method. Equation (2) contains both electric field and magnetic field, for purpose of reducing the computational complexity, we substitute the difference equations for H_y and H_z into (2). Then, we can get an iterative equation containing only the electric field. For example, the FDFD iterative equation for electric field node $E_x(i+1/2, j, k)$ is as follows:

$$\begin{aligned} & cx0 \cdot E_x(i+1/2, j, k) \\ & + cx1 \cdot E_x(i+1/2, j-1, k) + cx2 \cdot E_x(i+1/2, j+1, k) \\ & + cx3 \cdot E_x(i+1/2, j, k-1) + cx4 \cdot E_x(i+1/2, j, k+1) \\ & + cy1 \cdot E_y(i, j+1/2, k) + cy2 \cdot E_y(i+1, j+1/2, k) \\ & + cy3 \cdot E_y(i, j-1/2, k) + cy4 \cdot E_y(i+1, j-1/2, k) \\ & + cz1 \cdot E_z(i, j, k+1/2) + cz2 \cdot E_z(i+1, j, k+1/2) \\ & + cz3 \cdot E_z(i, j, k-1/2) + cz4 \cdot E_z(i+1, j, k-1/2) = 0 \end{aligned} \tag{3}$$

As can be seen from (3), the equation of each electric field involves its own node and its twelve adjacent nodes. And the twelve nodes are four E_x nodes, four E_y nodes and four E_z nodes respectively, as shown in Figure 1.



(a) Difference equation of electric field node E_x.



(b) Basic equation of electric field node E_x.

Figure 1. *Peripheral nodes involved in different situations of electric field nodes.*

where the coefficient are as follows, which are related to the grid size and the medium parameters.

$$\begin{cases} cx0 = \omega^2\mu_0\varepsilon_0\Delta y\Delta z\varepsilon_r\left(i+1/2,j,k\right)-\left[\mu_r^{-1}\left(i+1/2,j-1/2,k\right)+\mu^{-1}{}_r\left(i+1/2,j+1/2,k\right)\right]\cdot\Delta z/\Delta y \\ \qquad -\left[\mu_r^{-1}\left(i+1/2,j,k-1/2\right)+\mu^{-1}{}_r\left(i+1/2,j,k+1/2\right)\right]\cdot\Delta z/\Delta y \\ cx1 = \mu_r^{-1}\left(i+1/2,j-1/2,k\right)\Delta z/\Delta y \\ cx2 = \mu_r^{-1}\left(i+1/2,j+1/2,k\right)\Delta z/\Delta y \\ cx3 = \mu_r^{-1}\left(i+1/2,j,k-1/2\right)\Delta y/\Delta z \\ cx4 = \mu_r^{-1}\left(i+1/2,j,k+1/2\right)\Delta y/\Delta z \\ cy1 = -cy2 = \mu_r^{-1}\left(i+1/2,j+1/2,k\right)\Delta z/\Delta x \\ cy3 = -cy4 = -\mu_r^{-1}\left(i+1/2,j-1/2,k\right)\Delta z/\Delta x \\ cz1 = -cz2 = \mu_r^{-1}\left(i+1/2,j,k+1/2\right)\Delta y/\Delta x \\ cz3 = -cz4 = -\mu_r^{-1}\left(i+1/2,j,k-1/2\right)\Delta z/\Delta x \end{cases} \tag{4}$$

Similarly, the FDFD iterative equations for the other electric field components can be obtained. By combining the FDFD equations corresponding to all electric field nodes, the solution of the field in the FDFD calculation domain can be simplified to the solution of the matrix equation $\boldsymbol{A}\cdot x = b$, where $\boldsymbol{A}$ is the coefficient matrix composed of the coefficients in (4), x is the vector formed by the electric field nodes to be solved in the computational domain, and b is the excitation source vector.

3. Fast Solution of Large Complex Sparse Matrix Equations

3.1. Conjugate Gradient Iteration

The FDFD method can be transformed into the form of the matrix equation $\boldsymbol{A}\cdot x = b$, where the coefficient matrix $\boldsymbol{A}$ is a complex sparse matrix. To reduce the consumption of computing resources, the format of Coordinate (COO) [6] is used to store the complex sparse matrix. This format stores only the row, column and value of each non-zero element for the sparse matrix, which is very efficient. Because $\boldsymbol{A}$ is not a special matrix such as triangular matrix and banded matrix, etc. the existing efficient iterative solvers are not suitable. Therefore, in this paper, the conjugate gradient method (CGM) of complex matrix is used to solve the matrix equation [7, 8], and the specific steps are as follows [9]:

(i) Initialization, set initial value r_0 and convergence accuracy *err*:

$$\begin{cases} r_0 = \boldsymbol{A}x_0 - b \\ p_1 = -\boldsymbol{A}^+ r_0 \end{cases} \tag{5}$$

where $\boldsymbol{A}^+$ is the conjugate transpose of matrix $\boldsymbol{A}$.

(ii) Iteration (n = 1, 2, 3, 4…)

$$\begin{cases} \alpha_n & = \left\|\boldsymbol{A}^+ r_{n-1}\right\|^2 \Big/ \left\|\boldsymbol{A}p_n\right\|^2 \\ x_n & = x_{n-1} + \alpha_n p_n \\ r_n & = \boldsymbol{A}x_n - b = r_{n-1} + \alpha_n \boldsymbol{A}p_n \\ \beta_n & = \left\|\boldsymbol{A}^+ r_n\right\|^2 \Big/ \left\|\boldsymbol{A}^+ r_{n-1}\right\|^2 \\ p_{n+1} & = -\boldsymbol{A}^+ r_n + \beta_n p_n \end{cases} \tag{6}$$

(iii) Make a judgment after each iteration. Once $\left\|\boldsymbol{A}x-b\right\|/\left\|b\right\| \le err$, output x_n and finish the iteration. Otherwise, continue with step (ii).

3.2. Parallel Algorithm of Conjugate Gradient Iteration Based on MPI

It can be seen from the above section that the conjugate gradient algorithm can greatly improve the computational efficiency of matrix equations by transforming the solution of matrix equations into simple sparse matrix vector multiplication, vector inner product and vector addition and subtraction. However, when the electrical size of the calculated target increases, the scale of the coefficient matrix also increases. On the one hand, it will lead to a sharp increase in the amount of computation and an increase in the computational resources consumed by each iteration. On the other hand, it makes the convergence of iteration worse, that is, it needs more iterations to complete the computation. In order to improve efficiency, parallel computing based on MPI is a better solution. However, due to the sparsity of the coefficient matrix, it is not feasible to use the parallel scheme of dense matrix partition by block. This is because the distribution of nonzero elements in the coefficient matrix is not uniform, and partition by block will lead to unbalanced load, which affects the parallel efficiency. In addition, because the number of nonzero elements in each row of the coefficient matrix is not equal, the common equal row allocation scheme will also lead to load imbalance. According to the characteristics of coefficient matrix, a scheme of unequal row allocation according to the number of elements is proposed in this paper. The coefficient matrix is evenly distributed to each process to ensure load balancing.

The idea of parallelization in this paper is:

(i) Matrix $\boldsymbol{A}$ is a large complex sparse matrix, whose most elements are zero, and only nonzero elements affect the calculation results. Thus, the total number of nonzero elements N_{tot} is counted first.

(ii) To guarantee the load balancing of each process, it is necessary to ensure that the number of nonzero elements allocated to each process is basically the same.

Therefore, the number of nonzero elements that should be allocated to each process is about $Aver = N_{tot} / n$, where n stands for the number of processes.

(iii) Following the principle of allocation by row, start to allocate nonzero elements to n processes. Because the number of nonzero elements in each row of the coefficient matrix is not equal, on the premise of ensuring that the number of nonzero elements contained in each process is greater than or equal to A_{ver}, the number of matrix rows allocated to each process may be inconsistent, but it must be ensured that the nonzero elements in each process are all nonzero elements of a row or several rows in the coefficient matrix.

The specific allocation steps are given below.

a) Traverse all nonzero elements N_{tot} and count the number of nonzero elements in each row.
b) An auxiliary array is allocated to store the start position and end position of nonzero elements in each row.
c) Start assigning rows to each process until the number of nonzero elements in the process is greater than or equal to A_{ver}.
d) Count the remaining rows of the matrix and assign them to a subprocess other than the main process.
e) Count the start and end positions of the matrix rows allocated by each process.

(iv) The above processes need to be allocated only once before the iteration, but the intermediate vectors r_n, p_n and x_n used by each process need to be assigned to each process during the iteration. Because the intermediate vectors are allocated during the iteration, the communication time is long. In addition, intermediate vectors need to be fully allocated to each process, so the computer memory consumption will increase with the increase of the number of processes. Therefore, the shared memory technology of MPI is introduced in this paper to optimize the allocation of intermediate vectors. In our work, the intermediate vectors needed by each process are stored in the shared memory, this shared memory technology eliminates the requirement for the main process to send intermediate vectors to each subprocess, which not only reduces the memory occupation, but also reduces the communication time between processes. Figure 2 shows the storage location of intermediate vectors in memory.

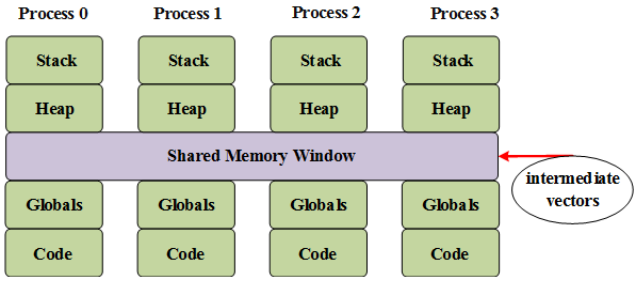


Figure 2. *The intermediate vector is stored in the shared memory area.*

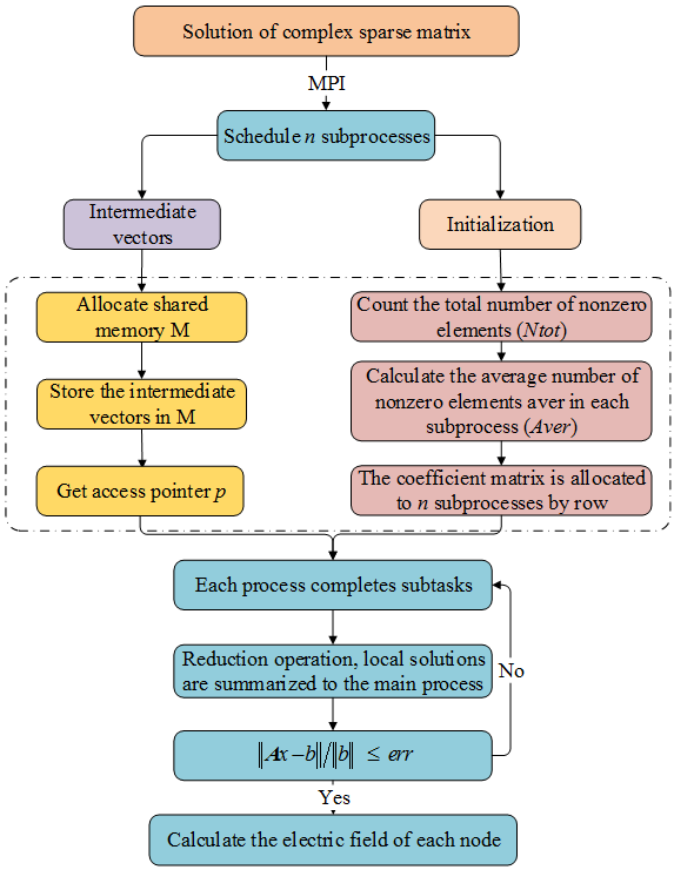


Figure 3. *MPI based conjugate gradient iterative parallel algorithm.*

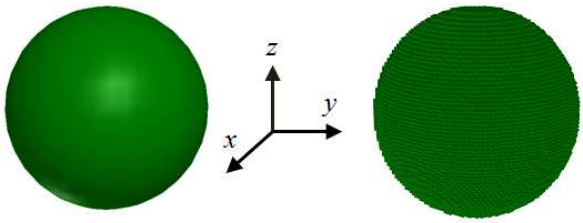
(v) After the assignment is completed, each process obtains the corresponding local solution according to the process of complex conjugate gradient method, and the final solution x_n can be obtained by gathering the local solution to the main process.

The flow chart of the parallel algorithm of conjugate gradient iterative based MPI is shown in Figure 3.

4. Numerical Results

4.1. Bistatic RCS of PEC Sphere

The geometric diagram and grid discretization of the PEC sphere are shown in Figure 4, with a radius $r = 1.0\text{m}$, the frequency of the plane wave is $f = 0.3\text{GHz}$ and its wavelength is $\lambda = c/f = 1.0\text{m}$. The plane wave is incident along the *z*-direction, and the electric field is in the *xoy* plane, i.e. the pitch angle is $\theta = 0^\circ$, the azimuth angle is $\varphi = 0^\circ$, and the polarization angle is $\alpha = 0^\circ$. The discrete grid size is taken as $\Delta x = \Delta y = \Delta z = \lambda / 20 = 0.05\text{m}$. In all directions, the target boundary is extrapolated to the truncated boundary by 10 grids. Finally, the whole calculation area is $(-30\Delta x : 30\Delta x, -30\Delta y : 30\Delta y, -30\Delta z : 30\Delta z)$. The dimension of the coefficient matrix is 703452, and there are 494844716304 elements in total, including 8928744 non-zero elements, accounting for 0.0018% of the total matrix elements. The calculation results of FDFD serial algorithm, parallel scheme and commercial software are shown in Figure 5. It is obvious that the calculation results of FDFD parallel scheme are in good agreement with FDFD serial algorithm and commercial software HFSS, which proves the correctness and feasibility of FDFD algorithm and parallel scheme. Table 1 shows the calculation time and computer memory occupied by FDFD parallel algorithm in solving PEC sphere when the number of processes is different. From Table 1, it is clear that the parallel efficiency of FDFD first increases and then decreases as the increase of the number of processes. This is because the calculation scale of this example is small, and the communication time between processes increases with the number of processes increases, which reduces the calculation efficiency. In addition, because the number of processes will also affect the number of conjugate gradient iterations, the parallel efficiency in Table 1 may be greater than 100%. In the case of same convergence accuracy, the number of iterations of serial algorithm and parallel algorithm is different.



(a) The geometric diagram. (b) The grid discretization.

Figure 4. *The model of the PEC sphere model..*

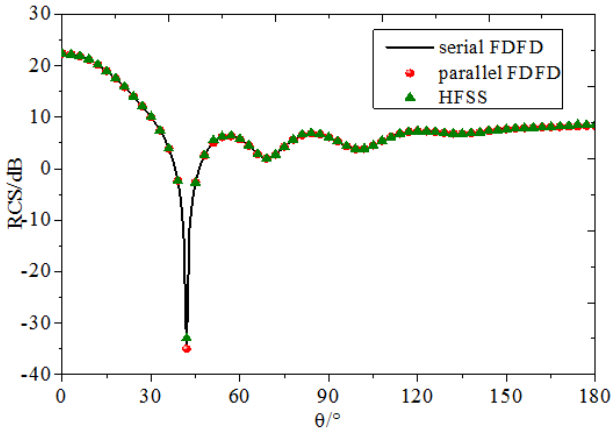
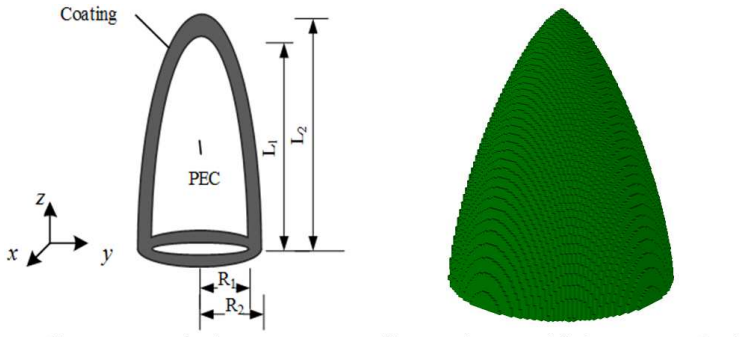


Figure 5. *The bistatic RCS of PEC sphere.*



(a) Composite warhead geometry (b) FDFD discrete model of composite warhead.

Figure 6. *Von warhead model.*

Table 1. *Performance comparison of PEC sphere in different number of processes.*

Number of processes	1 (serial)	2	4	8	12	16	24
Memory/MB	182	316	341	391	430	491	590
Time/s	1181	600	265	123	101	90	81
Speed up	1	1.96	4.45	9.6	11.69	13.1	14.6
Parallel efficiency /%	-	98	111	120	97	82	61

4.2. Bistatic RCS of Composite Von Warhead

The composite Von warhead is composed of inner and outer layers, the inner layer is PEC structure, and the outer layer is coated with absorbing materials to reduce its RCS. The geometric diagram and grid discretization of the composite warhead are shown in Figure 6. Its dimensions are $R_1 = 0.1$m, $R_2 = 0.15$m, $L_1 = 0.4$m, $L_1 = 0.4$m, $L_2 = 0.6$m. The frequency of the incident wave is f = 3GHz, and the plane wave is incident along the z-direction, that is, the incident angle are $\theta = 0°$, $\varphi = 0°$, and $\alpha = 0°$. The medium parameters of the outer coating are $\varepsilon_r = (2.0 - j0.5)$, $\mu_r = (1.24 - j0.2)$. The grid size is $\Delta x = \Delta y = \Delta z = \lambda/20$ = 5mm, the serial FDFD and parallel FDFD methods are adopted to calculate the bistatic RCS of the warhead, and the calculation results are shown in Figure 7. It can be seen that the calculation results of the two methods agree with well. Table 2 shows the results of time consumption and memory occupation of parallel methods with different process numbers. In this example, the parallel efficiency can reach 93% when the number of processes is $n = 24$.

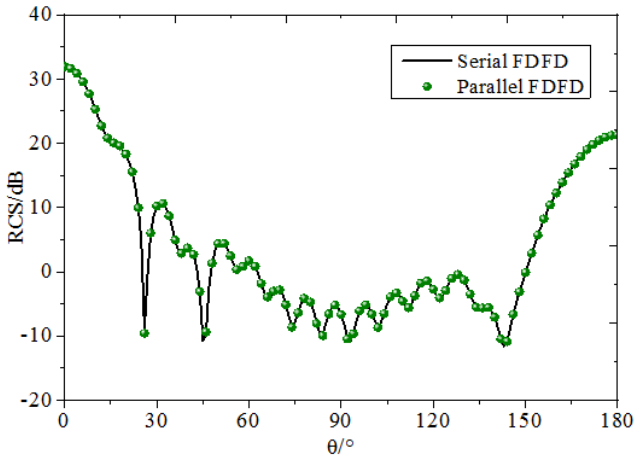


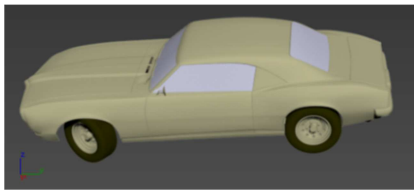
Figure 7. *Bistatic RCS of Von warhead.*

Table 2. *Performance comparison of Von warhead model in different number of processes.*

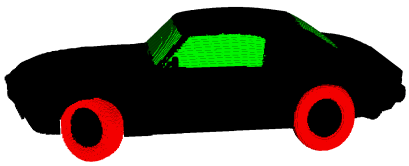
Number of processes	1 (serial)	2	4	8	12	16	24
Memory/MB	1233	2077	2104	2157	2204	2243	2347
Time/s	5867	2763	1428	600	376	330	263
Speed up	1.0	2.12	4.11	9.78	15.6	17.8	22.3
Parallel efficiency /%	-	106	103	122	125	111	93

4.3. Bistatic RCS of an Automobile

The geometric diagram and grid discretization of the automobile are shown in Figure 8. The length of automobile is $l = 4.2$m, the width is $w = 1.5$m and the height is $h = 1.2$m. The incident frequency in the case of lightning is $f = 100$MHz and the plane wave is incident from the automobile head direction, i.e. along the y-direction. The incident angle are $\theta = 0°$, $\varphi = 0°$ and the polarization angle is $\alpha = 0°$ and $\alpha = 45°$. Due to the complexity of automobile components, it is difficult to obtain the dielectric parameters of each component. In this example, the automobile components are mainly divided into three parts: automobile frame, tire and window glass. The material of automobile frame is regarded as PEC. The tire is mainly composed of rubber, and its corresponding dielectric parameters are $\mu_r = 1.0$ and $\varepsilon_r = 5.5$, The window glass is made of glass, and its dielectric parameters are $\mu_r = 1.0$ and $\varepsilon_r = 3.0$. In order to describe the fine structures of the automobile, the grid size is $\Delta x = \Delta y = \Delta z = \lambda/100 = 0.03$m. The target boundary is extrapolated to the truncated boundary by 30 grids, and the whole calculation area is ($-100\Delta x : 100\Delta x$, $-55\Delta y : 55\Delta y$, $-50\Delta z : 50\Delta z$). The serial FDFD and parallel FDFD methods are employed to calculate the bistatic RCS of the automobile with polarization angle of $\alpha = 0°$ and $\alpha = 45°$. The simulation results of the two polarization cases are shown in Figure 9 and Figure 10, respectively. It can be seen that the consistency between the parallel FDFD and the serial FDFD algorithms.



(a) Geometry of the automobile model



(b) Grid discretization of the automobile model

Figure 8. *The automobile model.*

Table 3. *Performance comparison of the automobile model in different number of processes.*

Number of processes	1 (serial)	2	4	8	12	16	24
Memory/MB	3645	6150	6171	6230	6277	6329	6424
Time/s	20890	10351	4246	2311	1488	1190	927
Speed up	1.0	2.01	4.92	9.04	14.0	17.6	22.5
Parallel efficiency /%	-	100	123	113	117	110	94

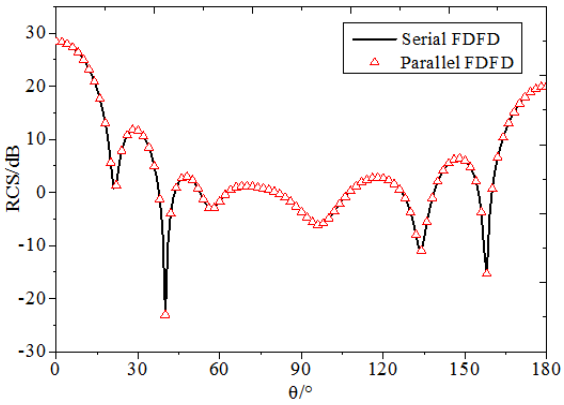


Figure 9. *Bistatic RCS of the automobile at polarization angle* $\alpha = 0^\circ$.

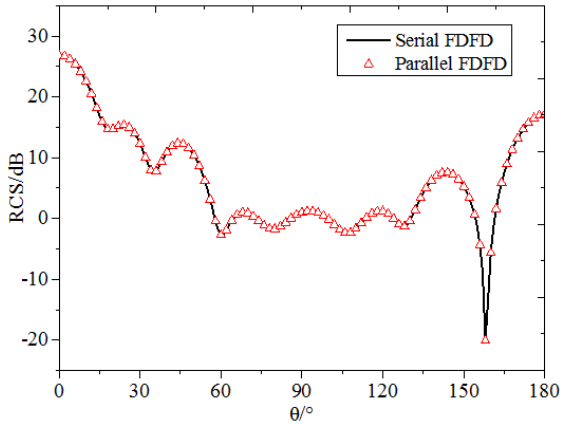


Figure 10. *Bistatic RCS of the automobile at polarization angle* $\alpha = 45^\circ$.

Table 3 shows the time and memory consumed by the parallel FDFD algorithm when the number of processes is different. It can be seen that the parallel efficiency can reach 94% when the number of processes is $n = 24$.

5. Conclusions

Aiming at the problem that the computational efficiency of FDFD method is low and difficult to solve electrically large targets, MPI is used to parallelize the FDFD algorithm, and the shared memory mechanism of MPI is introduced to store the intermediate vector, which reduces the computational resources and improves the computational efficiency of FDFD algorithm. Numerical results show that compared with the serial algorithm, the parallel algorithm occupies about two times as much memory as the serial algorithm, but its speedup ratio in 24 processes can reach 22.5, and the parallel efficiency can reach 94%. The parallel scheme greatly improves the computation efficiency and application scope of FDFD algorithm, and provides a certain reference value for further improving the computation performance of FDFD algorithm. In this paper, we only calculate the scattering problem of the target by FDFD method. In future work, the coupling effect of the HEMP E3 pulse and the rail, the propagation of low frequency electromagnetic wave in the ground-ionospheric waveguide will be studied.

References

[1] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Transactions on Antennas and Propagation, vol. 14, no. 3, pp. 302-307, May 1966.

[2] T. Namiki. A new FDTD algorithm based on alternating-direction implicit method. IEEE Transactions on Microwave Theory and Techniques, vol. 47, no. 10, pp. 2003-2007, Oct. 1999.

[3] T. An, M. Wei, S. Y. Li, et al. Study on the comparation of field-wire coupling effect for long wire in UWB with HEMP. Journal of Microwaves, vol. 26, no. 4, pp. 14-18, Apr. 2010.

[4] F. Xu, Y. L. Zhang, W. Hong, et al. Finite-difference frequency-domain algorithm for modeling guided-wave properties of substrate integrated waveguide. IEEE Transactions on Microwave Theory and Techniques, vol. 51, no. 11, pp. 2221-2227, Nov. 2003.

[5] A. G. Hanif, T. Arima, T. Uno. Finite-difference frequency-domain algorithm for band-diagram calculation of 2-D photonic crystals composed of Debye-type dispersive materials. IEEE Antennas and Wireless Propagation Letters, vol. 11, pp. 41-44, 2012.

[6] N. Neuss. A new sparse-matrix storage method for adaptively solving large systems of reaction-diffusion-transport equations. Computing, vol. 68, no. 1, pp. 19-36, Sep. 2001.

[7] D. A. H. Jacobs. A Generalization of the conjugate-gradient method to solve complex systems. IMA Journal of Numerical Analysis, vol. 6, no. 4, pp. 447-452, 1986.

[8] V. Demir, E. Alkan, A. Z. Elsherbeni, et al. An algorithm for efficient solution of finite-difference frequency-domain (FDFD) methods. IEEE Antennas and Propagation Magazine, vol. 51, no. 6, pp. 143-150, 2010.

[9] X. L. Li, B. Wei, X. B. He, et al. Parallel FDFD Algorithm Based on MPI and Its Application. 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), pp. 1-3, 2020.

[10] J. N. Hwang. A compact 2-D FDFD method for modeling microstrip structures with nonuniform grids and perfectly matched layer. IEEE Transactions on Microwave Theory and Techniques, vol. 53, no. 2, pp. 653-659, Feb. 2005.

[11] C. M. Rappaport, M. Kilmer, E. Miller. Accuracy considerations in using the PML ABC with FDFD Helmholtz equation computation. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, vol. 13, pp. 471-482, 2000.

[12] M. W. Chevalier, U. S. Inan. A technique for efficiently modeling long-path propagation for use in both FDFD and FDTD. IEEE Antennas and Wireless Propagation Letters, vol. 5, pp. 535-528, 2006.

[13] G. Zheng, B. Z. Wang. A hybrid MM-FDFD method for the analysis of waveguides with multiple discontinuities. IEEE Antennas and Wireless Propagation Letters, vol. 11, pp. 645-647, 2012.

[14] X. G. Xie, L. Wei, Ying L., et al. Using LU decomposition in FDFD for fast calculation of monostatic RCS, 2014 7th International Conference on Intelligent Computation Technology and Automation, pp. 887-889, 2014.

[15] K. Masumnia-Bisheh, K. Forooraghi, M. Ghaffari-Miab. Electromagnetic uncertainty analysis using stochastic FDFD method. IEEE Transactions on Antenna and Propagation, vol. 67, no. 5, pp. 3268-3277, May. 2019.

[16] X. Gu, X. L. Jin, J. X. Li, et al. Two-component compact 2-D FDFD method for waveguide structures with ARPACK. 2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, pp. 187-188, 2019.

About Calculation the Resistance of Two-dimensional Infinite Grid Systems

J. Uday Bhaskar, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, j.udaybhaskar1@gmail.com*

Sunil Kumar Tripathy, *Department of Mechanical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar,sktripathy1@gmail.com*

Rajib Lochan Barik, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, rajib_barik543@gmail.com*

Anil Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, anil_sahoo342@gmail.com*

Abstract: The paper considers the problem of calculating the resistance between nodes of infinite grid resistor systems with square and triangular cells. There has long been a question about the resistance between the nearest nodes of an infinite grid of resistances with square cells with the same resistance *r*. Here, earlier, by the method of symmetry and superposition, a result was obtained *r*/2 that is striking in its simplicity. However, this result is only approximate, although many physicists consider this result to be accurate. New examples are presented proving what the results obtained earlier by the superposition and symmetry method is only approximate. The result *r*/2 gives only the lower limit of the correct resistance value. In our work, the correctness of using the equivalent resistance method to calculate the resistance between nearest nodes of infinite grid systems is proved. Using this method, for the resistance between the nearest nodes of an infinite grid of resistances with square cells, a result is obtained about 0.5216 *r* that only slightly differs from *r*/2. The results differ from the previously obtained values by about 10%. The resistance between the diagonal points of an infinite grid of identical resistors *r* with square cells is calculated. For the value of this resistance, a value founded about 0.7071 *r* that differs from the value $2r/\pi$ obtained previously by the superposition and symmetry method.

Keywords: Calculation of Resistance, Infinite Two-Dimensional Grid of Resistances, Equivalent Resistance Method

1. Introduction

The calculation of the resistance of complex resistor compounds has always attracted the attention of physicists. Many different original methods for calculating endless circuits of resistance have been developed [1–7]. The tasks of finding the resistance of infinite grid systems also been included in Irodov's book [8] and were considered in our works [9, 10]. The problems of calculating the resistances of infinite resistor grids, using graphene [11, 12] and thin films [12] in connection with the development of nanotechnologies have become especially urgent.

Let us dwell in more detail on one problem reviewed in the works [2–8]. «There is a boundless wire grid with square cells (Figure 1). The resistance of each conductor between neighboring grid nodes is *r*. Find the resistance R_{AB} of this grid between two adjacent grids modes *A* and *B*».

This task first appeared in Irodov's problem book of 1979 yeas, where the method of « symmetry & superposition » was applied and the result was obtained:

$$R_{AB}=\frac{r}{2}. \qquad (1)$$

The same result (1) was also obtained in the works [2–4]. This opinion was especially strengthened in connection with work [4], as well as in works [14–16]. In this work, complex mathematics made it possible to obtain a general formula for the resistance between any points of the grid. However, here results is also based on the method the «symmetry & superposition», therefore, for the resistance between the nearest point of grid, this formula gives the same result $r/2$.

Many physicists consider this result to be accurate. Now this is a general misconception, which is very difficult to overcome. Of course, the result $r/2$ for resistances fascinates with its simplicity, but it is only approximate. Moreover, it gives only a lower bound for estimating the magnitude of the resistance.

In Figure 2 shows three cases of connecting voltage to the points of the grid. In case *a*) the voltage $+U/2$ is supplied only to point *A*, and in case *b*) the voltage $-U/2$ is supplied only to

point *B*. In these cases, the current distribution is symmetrical and they differ only in the direction of the currents. In this case, points with a potential equal to zero are at infinity.

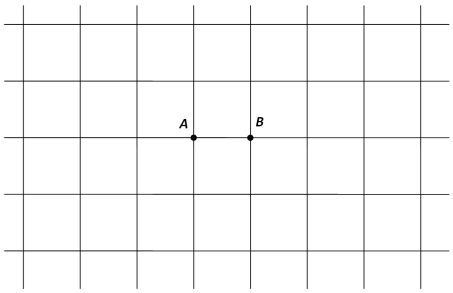


Figure 1. *Infinite wire grid with square cells.*

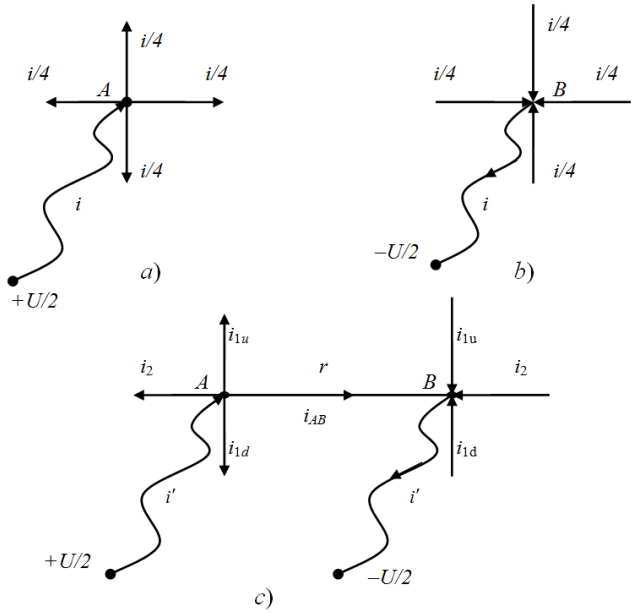


Figure 2. *a) The potential* $+U/2$ *only at point A (zero at infinity); b) The potential* $-U/2$ *only at point B (zero at infinity); c) The potential* $+U/2$ *at point A and the potential* $-U/2$ *at point B.*

In case c) voltage $+U/2$ is supplied to point *A*, and voltage $-U/2$ to point *B*. Here *A* and *B* are the nearest points of the grid, between which the resistance r is located. Current $i' \neq i$ approaches point *A*, and the same current i' emerges from point *B*. The current distribution at points *A* and *B* is shown, which corresponds to the symmetry of the problem. Now a line with a potential equal to zero passes in the middle between points *A* and *B*. A current i_{AB} flows from *A* to *B*. Then according to Ohm's law:

$$U = i' R_{AB} = i_{AB} r . \qquad (2)$$

Here R_{AB} is the desired grid resistance between the points *A* and *B*. If we make two assumptions that $i' = i$ and $i_{AB} = i/2$, then from formula (2) we get result (1): $R_{AB} = \frac{r}{2}$. In this case, if $i_{AB} = i/2$, $i_{1u} = i_{1d} = i_1 = i/4$, then according to the Kirchhoff rule: $i_2 = 0$.

Symmetry in case c) has changed. So the distribution of currents at points *A* and *B* has changed. Note that a line with a potential equal to zero runs in the middle between the points *A* and *B* and goes to infinity. Since the potential difference between points *A* and *B* and infinity is not equal to zero, the currents $i_2 \neq 0$ and $i_{AB} > i'/2$. These three factors indicate that the result $R_{AB} = r/2$ is an approximation, more precisely even $R_{AB} > r/2$. In our work [6] we used the equivalent resistance method and obtained the following result:

$$R_{AB} = \frac{2\left(\sqrt{2}-1\right)+\sqrt{2\sqrt{2}-1}}{2\sqrt{2}+\sqrt{2\sqrt{2}-1}} r \cong 0.521602\, r , \quad (3)$$

and the following values of the currents: $i_{AB} \cong 0.522\, i'$, $i_{1u} = i_{1d} = i_1 \cong 0.207\, i'$, $i_2 \cong 0.064\, i'$. It is easy to verify that in this case:

$$i' = i_{AB} + i_{1u} + i_{1d} + i_2 .$$

So, here everything is in order with the Kirchhoff rule in the points *A* and *B*.

But the result, which we obtained $R_{AB} \cong 0.522\, r$, is just only slightly superior to the result $r/2$, which confirms its correctness. The arguments we have presented, we think, prove that the result $r/2$ is only approximate.

2. Calculation of the Resistance Between Nearest Points of the Infinite Grid with Square Cells

We give below an alternative solution to this problem. First, we divide the entire plane into two identical half-planes, cutting the grid along a straight line passing through points *A* and *B*, to the left of point *A* and to the right of *B*. For this, each of the resistances *r*, lying to the left of point *A* and to the right of *B*, is replaced by two parallel-connected resistance of $2r$ each. We obtain the following picture, shown in Figure 3. In Figure 4 shows an equivalent scheme of a cut mesh, where the resistance of the half-planes obtained as a result of cutting is denoted by *R*:

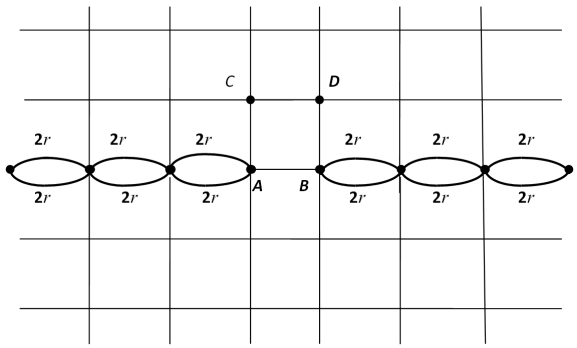


Figure 3. *Infinite wire grid, divided into two half-planes.*

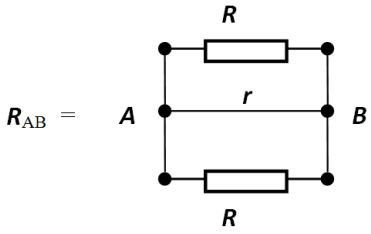


Figure 4. *Equivalent scheme of unlimited wire mesh.*

Then the resistance between adjacent nodes A and B of an infinite two-dimensional grid with square cells is equal to:

$$R_{AB} = \frac{r\,R}{2r+R}\,. \tag{4}$$

To do this, we similarly cut the grid along a straight line passing through points C and D. The resulting picture is shown in Figure 5, where you can see two infinite half-planes with resistance R, located below points A, B and above points C, D, and also two identical infinite chains going to the left of points A and C and to the right of points B and D. We denote the resistance of such an infinite chain by r^*. Then, comparing the schemes in Figure 4, we can draw up the equivalent circuit depicted in Figure 5.

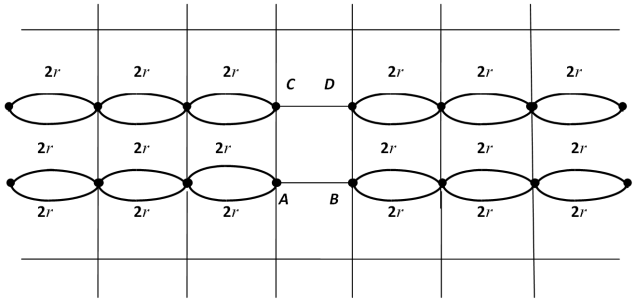


Figure 5. *Infinite wire mesh, cut along two straight lines.*

Using the equivalent circuit in Figure 6, we write the equation for determining the resistance of the half-plane R:

$$R = 2r^* + \frac{r\,R}{r+R}\,. \tag{5}$$

Solving this quadratic equation with respect to R, we find

$$R = r^* + \sqrt{r^{*2} + 2r^*r}\,. \tag{6}$$

The resistance of an infinite periodic chain r^* is found by the standard method, solving equation

$$r^* = \frac{(4r+r^*)r}{5r+r^*}\,. \tag{7}$$

Location

$$r^* = 2r\left(\sqrt{2}-1\right). \tag{8}$$

Substituting (8) into (6), we obtain

$$R = \left(-2+2\sqrt{2}+\sqrt{2\sqrt{2}-1}\right) r\,. \tag{9}$$

Substitution of this value of R into (4) leads to the final result:

$$R_{AB} = \frac{-2+2\sqrt{2}+\sqrt{2\sqrt{2}-1}}{2\sqrt{2}+\sqrt{2\sqrt{2}-1}}\,r \cong 0.52160212\ r\,. \tag{10}$$

This result, although not very strong, is still different from the result of 0.5 r obtained in the approximation of the principles of symmetry and superposition. In this connection it is interesting to look at the distribution of currents at point A. If current i' approaches the point A, then it will be distributed as follows: according to the resistance going up and down from point A, the same currents will go $i_{1u} = i_{1d} = i_1 \cong 0,207\,i'$, according to the resistance between nodes A and B, there will be a current $i_{AB} \cong 0,522\,i'$, and in the opposite direction there will be a current $i_2 \cong 0,064\,i'$.

It's a pity, the beauty is gone, the symmetry has disappeared, and everything has become very prosaic. Well, in life it often happens that beauty deceives us and then it is difficult to get rid of beautiful illusions.

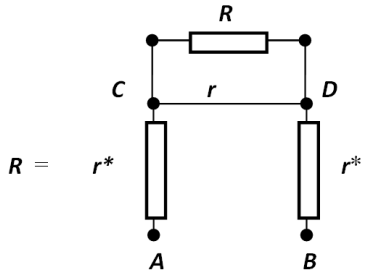


Figure 6. *An equivalent circuit for the resistance of the half-plane R in the case of a square grid.*

3. Calculation of the Resistance Between the Diagonal Points of the Infinite Grid with Square Cells

Here we give a solution to the problem of an infinite grid with square cells by the same resistances r, as shown in Figure 7. Suppose it is necessary to find the resistance R_{AC} of the lattice between the two points A and C. The solution of the problem by the method of superposition & symmetry in [4] leads to the following result:

$$R_{AC} = \frac{2r}{\pi} \cong 0.636620\ r\,. \tag{11}$$

Figure 3 shows a part of the grid, and the dashed lines show the directions along which it is necessary to make cuts. First, we cut along the rays issuing from the nodes A and C, thus breaking the entire grid into two half-planes with the same resistances R. As a result, we obtain for the resistance formula:

$$R_{AC} = \frac{R}{2} . \quad (12)$$

Now you need to find the half-plane resistance R. To do this, we draw a second section from the points A_1 and C_1, as shown in Figure 7. To determine R, we construct the following equivalent circuit shown in Figure 8.

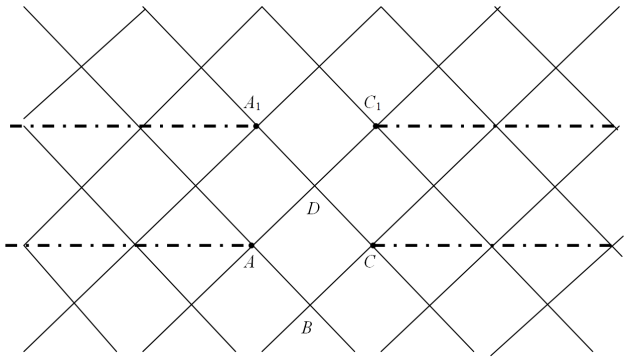


Figure 7. *Infinite grid with square cells and cutting line directions.*

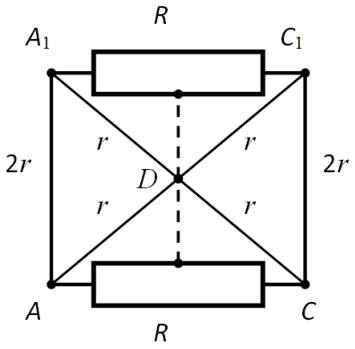


Figure 8. *Equivalent circuit for calculation of the half-plane resistance* R.

In Figure 4 shows a dashed vertical line passing through the point D and dividing the circuit into two symmetrical parts. Connecting the point D with the middle of the resistance R, we calculate the resistance of the resulting compound. First, find the resistance of the parallel connection $\frac{R}{2}$ and r get:

$$r' = \frac{\frac{R}{2} r}{\frac{R}{2} + r} = \frac{Rr}{R+2r} . \quad (13)$$

Then, connecting in series r' with $2r$, we find:

$$r'' = \frac{Rr}{R+2r} + 2r = \frac{3Rr + 4r^2}{R+2r} . \quad (14)$$

Now we make up the equation to determine the half-plane resistance R:

$$\frac{R}{2} = \frac{r'' r}{r'' + r} . \quad (15)$$

Substituting (14) into (15), we obtain:

$$\frac{R}{2} = \frac{3Rr^2 + 4r^3}{4Rr + 6r^2} . \quad (16)$$

Here will we find

$$R = \sqrt{2}\, r . \quad (17)$$

Substituting (17) into (12), we obtain the desired resistance between the diagonal points:

$$R_{AC} = \frac{R}{2} = \frac{r}{\sqrt{2}} \cong 0.707107\, r . \quad (18)$$

It is easy to verify that the difference with the result (11) obtained by the method of «superposition & symmetry» is 10%.

4. Calculation of the Resistance Between Points of the Infinite Grid with Triangles Cells

Here we also give a solution to the problem of an infinite net by the same resistances r forming regular triangles, as shown in Figure 9. Suppose it is necessary to find the resistance R_{AB} of the lattice between the two nearest nodes A and B. Using for the solution of the problem the principles of symmetry and superposition, by analogy with a square grid, it is easy to obtain the following simple result of $R_{AB} = r/3$, which is also only approximate.

In Figure 6 shows the part of the grid and the bold lines show the directions along which it is necessary to make cuts in order to break it into the same half-planes and endless chains. First, we cut along the rays issuing from the nodes A and B, thus breaking the entire grid into two half-planes with the same resistances R. As a result, just as in the case of a rectangular grid, we obtain the equivalent circuit shown in Figure 9, and formula (4) for the resistance of an infinite grid.

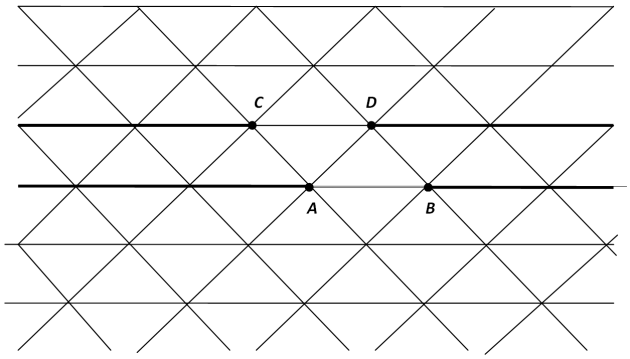


Figure 9. *Infinite grid of regular triangles.*

On Figure 9 each side of the triangle has a resistance r. The thick lines indicate the directions of the sections. To

determine the resistance of the half-plane R, we cut once more the lattice along the rays emanating from the nodes C and D. We obtain the equivalent circuit shown in Figure 10 and, respectively, equation (19):

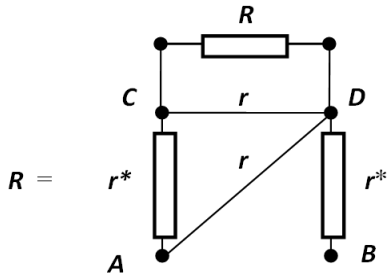


Figure 10. *An equivalent circuit for the resistance of the half-plane R in the case of a triangular grid.*

$$\frac{\left(\frac{Rr}{R+r}+r^*\right)r}{\frac{Rr}{R+r}+r^*+r}+r^*=R\,. \qquad (19)$$

After simple transformations, we obtain the following quadratic equation for the determination of R:

$$R^2-r^*R-rr^*=0\,. \qquad (20)$$

Location

$$R=\frac{r^*+\sqrt{r^{*2}+4rr^*}}{2}\,. \qquad (21)$$

Here, the resistance of the infinite chain r^* is found from the equivalent circuit shown in Figure 11, which leads to the equation (21).

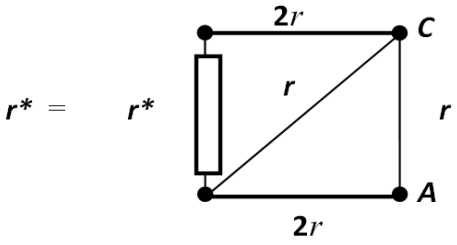


Figure 11. *Equivalent circuit for calculation of resistance r*.*

Infinite chain in the case of a triangular grid:

$$\frac{\frac{(r^*+2r)r}{r^*+3r}+2r}{\frac{r^*+2r}{r^*+3r}+3}=r^*\,. \qquad (22)$$

Solving this equation, we find

$$r^*=\left(\sqrt{3}-1\right)r\,. \qquad (23)$$

Substituting this value of r^* into formula (10), we obtain for R the expression:

$$R=\frac{\sqrt{3}+\sqrt{2\sqrt{3}-1}}{2}r\,. \qquad (24)$$

Substitution of this expression in (4) leads to the final result:

$$R_{AB}=\frac{\sqrt{3}+\sqrt{2\sqrt{3}-1}}{3+\sqrt{3}+\sqrt{2\sqrt{3}}}r\cong 0.39331989\,r\,. \qquad (25)$$

Thus, in this paper we develop a method for calculating infinite networks of resistances, which makes it possible to find the exact resistance between the nearest nodes of such networks. It is shown that the method «symmetry & superposition» used in [2-5], based on the principles of superposition and symmetry, gives only an approximate underestimated result for this resistance. And for a grid with square cells, the difference of results does not exceed 5%, and for a network with triangular elements it approaches 20%.

5. Conclusion

Thus, in work [6] and this article we have developed a method for calculating infinite resistance grids, which allows us to find the exact resistance between the nodes of such grids. It is shown that the calculation method used in [2–5], based on the principles of symmetry & superposition gives only an approximate underestimated result for this resistance. As we have shown in our works for the grid with square cells, the difference in results does not exceed 10%, and for the grid with triangular elements it approaches 15%.

The method of calculating the resistances of the infinite grids, which uses the principles of symmetry & superposition, is quite good, and its simplicity makes it very attractive for an approximate evaluation of the resistances of various infinite configurations of resistances. So, for example, for resistance R_{AB} between two nearest points of an infinite grid with hexagonal cells (a configuration of graphene) we get value $R_{AB}=2r/3$, and for resistance R_{AB} between two nearest points of an infinite 3D grid with cubic cells get value $R_{AB}=r/3$.

References

[1] Bessonov L. A. (2002) Theoretical bases of electrical engineering. Electric circuits. – Moscow: Gardariki, 2002. – P. 638.

[2] Venezian G. On the resistance between two points on a grid // Am. J. Phys. 62, 1000–1004 (1994).

[3] Van Steenwijk F. J. Equivalent resistors of polyhedral resistive structures // Am. J. Phys. 66, 90–91 (1998).

[4] Atkinson D. and F. J. van Steenwijk. Infinite resistive lattices // Am. J. Phys. 67 (6), 486–492 (1999).

[5] Q. Meng, J. He, F. P. Dawalibi and J. Ma, "A new method to decrease ground resistances of substationgrounding systems in high resisistivity regions", IEEE Trans. On New Power Delivery, Vol. 14, pp. 911-916, 1999.

[6] H. S. Lee, J. H. Kim, F. P. Dawalibi and J. Ma, "Efficient ground grid designs in layered soils", IEEE Trans. On Power Delivery, Vol. 13, No. 3, pp. 745-751, July 1998.

[7] Bairamkulov R., Friedman E. G. Effective resistance of finite two-dimensional grids based on infinity mirror technique // IEEE Transactions on Circuits and Systems I: Regular Papers. (2020/04/24) – DOI. 10.1109/TCSI.2020.2985652.

[8] Irodov I. E. Exercises in General Physics. Tutorial. 14 th ed. – S.-Pt. – Msk. – Krs.: Izd. "Lan", 2016. – 416 p.

[9] Spivak-Lavrov I. F, Kurmanbai M. S., Mazhit A. N. About one method of calculation of resistance of two-dimensional infinite grid systems. – Vestnik ARSU. – No. 1 (51), Aktobe, 2018. – P. 43-51.

[10] Spivak-Lavrov IF, Kurmanbai MS, Mazhit AN (2018) About One Method of Calculation of Resistance of Two-dimensional Infinite Grid Systems. Educ Res Appl: ERCA-157. DOI: 10.29011/2575-7032/100057/.

[11] Katsneison, M. I. Carbon in Two Dimensions. – New York: Cambridge University Press, 2012. – 366 p.

[12] Davydov S. A Chain Model of a Zigzag Contact of Lateral Graphene-Like Hetero-structures // Technical Physics Letters (2018).

[13] Bhattarai, S. P. Construction of Sheet Resistance Measurement Setup for Tin Dioxide Film Using Four Probe Method. – American Journal of Physics and Applications; 2017, 5 (5): 60-65.

[14] Dandekar R. and Deepak D. Proportionate growth in patterns formed in the rotor-router model // Journal of Statistical Mechanics: Theory and Experiment, Volume 2014 (2014) P. 11–30.

[15] Owaidat M. Q. Determining the resistance of a full-infinite ladder network using lattice Green's function // Advanced Studies in Theoretical Physics, Vol. 9, 2015, no. 2, 77-83. doi.org/10.12988/ astp.2015.412159.

[16] Owaidat M. Q. and Asad J. H. Resistance calculation of three-dimensional triangular and hexagonal prism lattices, The European Physical Journal Plus, 131 (2016), no. 9, 309.

Design and Implementation of an Automated Lighting System

Aravinda Mahapatra, Department of Mechanical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, aravindamahapatra92@outlook.com

Subhendu Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srikant.p@yahoo.co.in*

Alekha Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, alekha.sahoo241@gmail.com*

Pratik Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, pratikmohanty92@hotmail.com*

Abstract: A programmable logic controller is a digitally operated system used in an industrial environment that uses a programmable memory to implement specific functions to control through digital or analogue inputs and outputs of various types of machines. This report used a typical application for PLC's lighting control in a building for an office environment. The proposed design must define an appropriate office with various areas or rooms needing lighting with multiple control signals, the lighting controlled by a Siemens LOGO PLC (type 6ED1 052-1MD00-0BA5). The design has four relay outputs and eight digital inputs configured as analogue inputs. The scheme designed must be unique to meet the minimum requirements, including a provision for emergency evacuation if the radiation levels exceed a critical threshold. The system must use both analogue inputs but used for different purposes, and the analogue input level ranges must be significantly different. The relay logic has controlled by rules derived from an original relay logic using an embedded computer system. It optimises the control tasks to perform rugged designs to withstand vibrations, temperature, humidity, and electrical noise designs that allow a vast expansion. This circuit will be developed and simulated using the LOGO comfort software and then implemented on the PLC hardware. It includes diagrams and a description that clearly describes how the system should operate.

Keywords: Automated Lighting System, PLC Hardware, LOGO Comfort Software

1. Introduction

A programmable logic controller is a digitally operating system designed for use in an industrial environment that uses a programmable memory for internal storage of user-orientated instruction for implementing specific functions. Its widely used in various aspects such as logic, sequencing, timing, counting and arithmetic to control through digital or analogue inputs and outputs. Control systems based on relay logic and PLC's have limited knowledge of programming controlled by a series of rules derived from the original relay logic [1]. System optimised for control tasks and rugged design to withstand vibrations, temperature, humidity and electrical noise and available in a range of sizes and capabilities, modular design allows expansion. The Siemens Logo PLC performs a devised access control system for securing the laboratories. Typically, the PLC device has four relay outputs and eight digital inputs. Two of the inputs have functioned as analogue [2]. The system contains one analogue information with an external amplifier to obtain the maximum possible resolution of input of 0-1V DC. The block diagram clearly explains the outlook of this project. The system inputs are directly assigned to the reactor room and provided with output terminals. Whenever the switch was in on condition, the corresponding door would automatically open. The reactor room performs like when the threshold triggers radiation level exceeds the gates are open automatically. In this paper, the design was developed and implemented according to the blocks requirement [3].

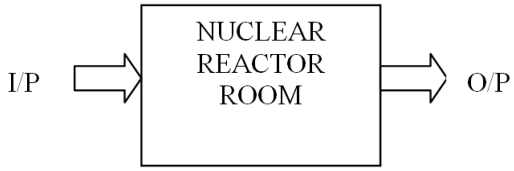


Figure 1. *Block Diagram.*

LOGO! Soft Comfort Software Features:

1) Method to delete user program and password from LOGO!
2) Additional Languages, resolution, and backlight function for the LOGO! Display.
3) It can perform online tests of LAD circuit programs.
4) Display of PI controller analogue output values in a trend view during simulation or online test.
5) Modem communication between a PC and LOGO!
6) USB cable communication between a PC and LOGO!
7) New memory card, battery card for LOGO! Basic Module.

2. Problem Analysis

In this Problem Analysis, The first step was to select the blocks for the circuit diagram by clicking on the icon group that contains the required blocks. According to this concept, some function blocks are analysed and discussed below, essential for this circuit diagram. Input blocks represent the input terminals of LOGO! It can be assigned an input block with a new input terminal. The LOGO! Versions Output blocks represent the output terminals of LOGO! It can be given an output block to a new terminal, provided not yet utilised in the circuit. The output block always carries the signal of the previous program cycle, and this value does not change within the current program cycle. AND Function tool was selected if a standard boolean logic block was placed on the programming interface [4]. The output of an AND function was only one if all inputs are 1 when they are closed. OR function was used when the output of an OR was 1 if at least one input was 1 (closed). Internal relays only exist in the internal memory of the PLC, and they are not associated with any real I/O. Timers are treated as configurable output devices with various available delays ON and delay OFF. The output of the timer appears as a switch elsewhere in the ladder diagram [5].

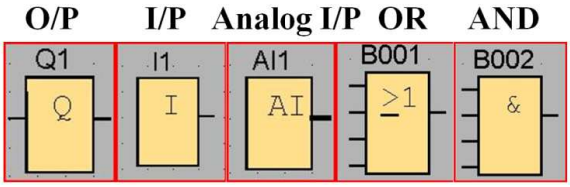


Figure 2. *Function Blocks.*

On delay, the output will not switch ON until the configured delay time has expired. The time Ta was triggered with the 0 to 1 transition at input Trigger. If the status at input stays one, at least for the configured time T, the output is set to 1 when this time has expired. The time reset if the status at input trigger changes to 0 again before T has passed. The output was reset to 0 when the input trigger was 0, and the threshold trigger will be switched ON and OFF depending on two configurable frequencies, and it measures the signals at the input frequency. It will capture the signal pulses during the configurable period of G_T. Q was set or reset according to the set threshold. An analogue signal is a physical quantity that can adopt continuous intermediate values within a given range. The opposite of analogue was digital analogue amplifier amplifies an analogue input value and returns it at the analogue output. The function reads the value of an analogue signal at the analogue input Ax. The value multiplied by the gain parameter A. Analog Comparator output was set and reset depending on Ax-Ay difference and on two configurable thresholds [6]. The function reads the value of the signal of the analogue input Ax. The value multiplies the value of parameter A (gain) and connects the product with Parameter B (offset). Output Q was reset or set depending on Axe-Ay's actual values and the set threshold [7].

3. Problem Solution

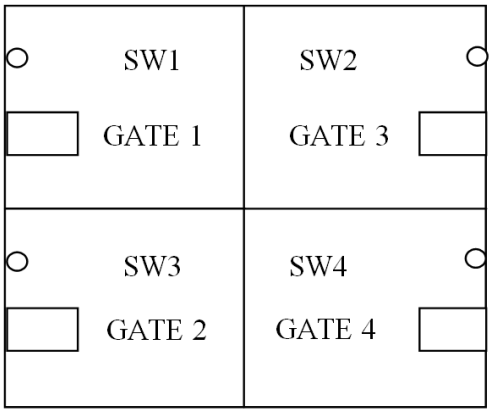


Figure 3. *Diagram of Nuclear Control Room Access.*

The above diagram explains the access in a nuclear reactor room. Whenever the switch is selected, the corresponding door will open automatically. The room was designed like whenever the threshold trigger radiation level exceeds the exit gats will open automatically. Developed the design using Logo comfort software and implemented it according to the blocks analysed in the problem analysis.

4. Problem Implementation

The functional block designing a control system for the final test stage of a surface-mounted PCB, and a delay of 2 seconds was allowed for the needle probes to contact the circuit and for the voltages of the PCB to stabilise. A test is initiated by a low to high output from the PLC. The PLC then checks the following results from the PCB (inputs to the PLC). Pin '1' of the PCB should give logic 1 and pin '2' logic 0. After 5 seconds, the analogue voltage on pin' 3'should is more significant than 4V. The PLC should then report a pass or fail (PLC output to go high for access). Noted the results of the system and described its operation. The blocks are placed in a functional block diagram and connected according to the ON delay input timer. It was connected to the OR gate with a message display and finally connected to the output.

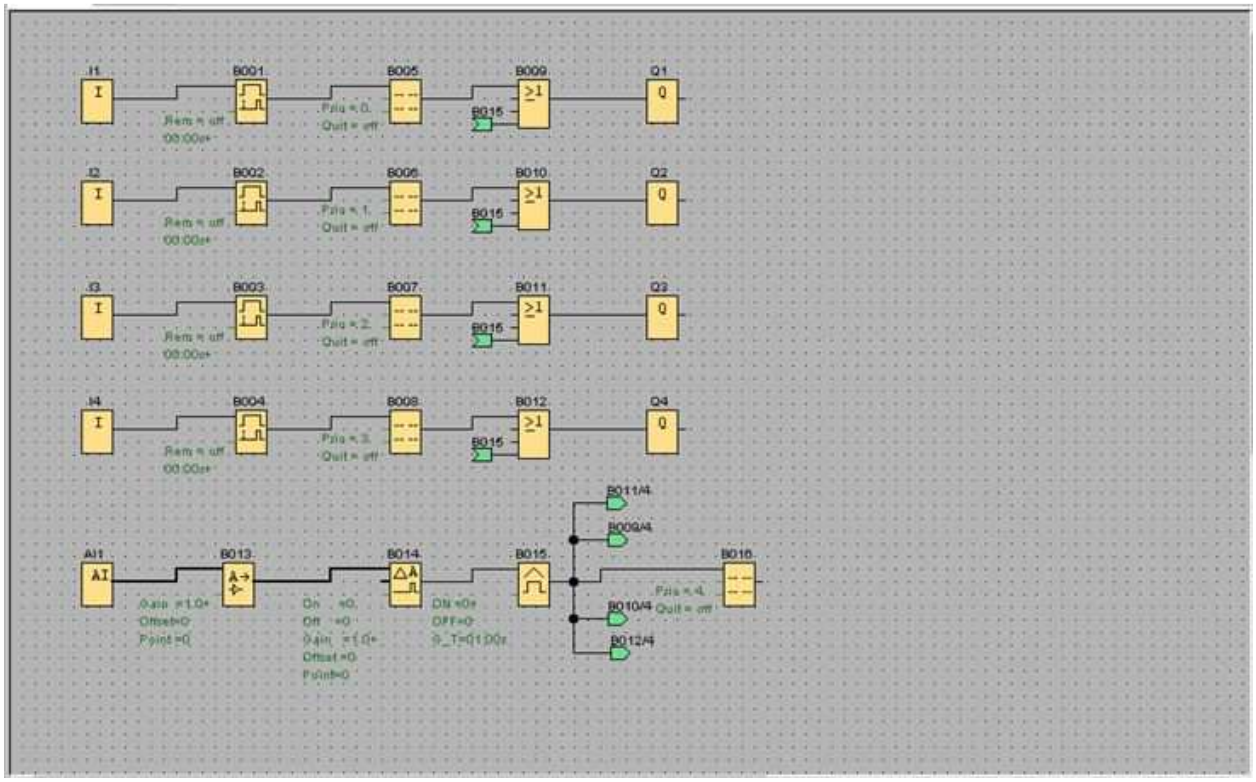


Figure 4. *Modified Blocks in FBD.*

The ladder design is simple relay wiring connections. Generally, vertical lines represent the power flow from left to right across a vertical rung. Each rung defines one operation in the control process. The ladder is read from left to right and top to bottom. When the rung has reached the endpoint, that control process starts again from the top. Each rung must begin with at least one input and end with one output [8].

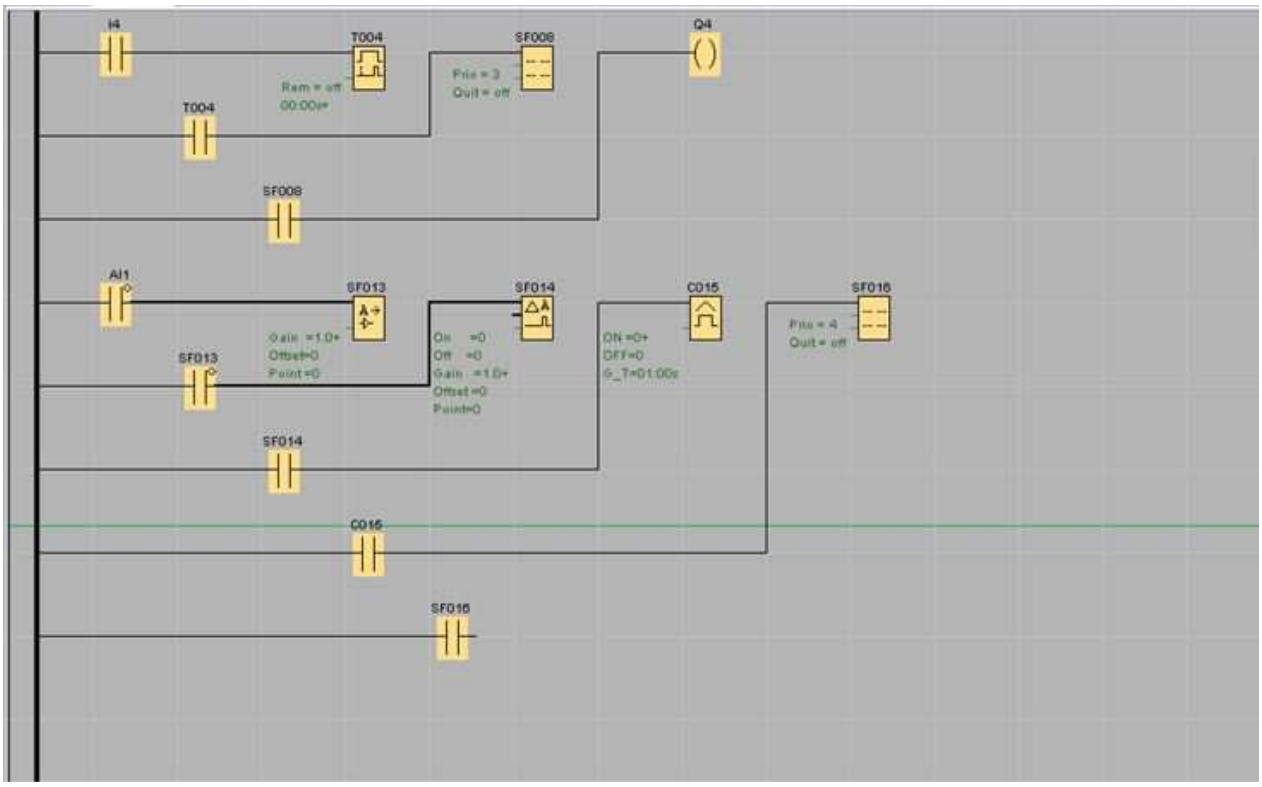


Figure 5. *Connecting Blocks in LAD.*

5. Results

The circuit was designed and analysed in the results section, and the functional block diagram and ladder logic diagram were performed and got the results.

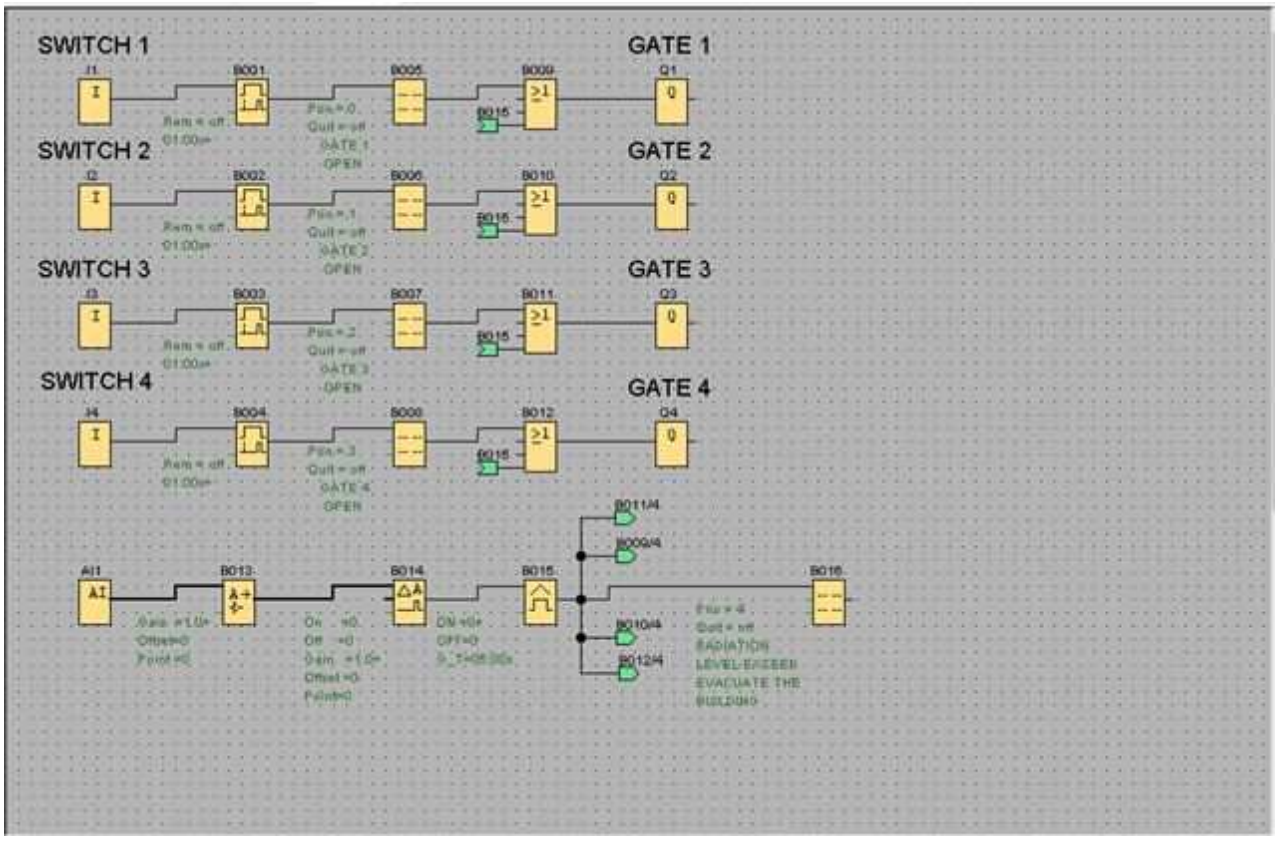


Figure 6. *Final Design of FBD.*

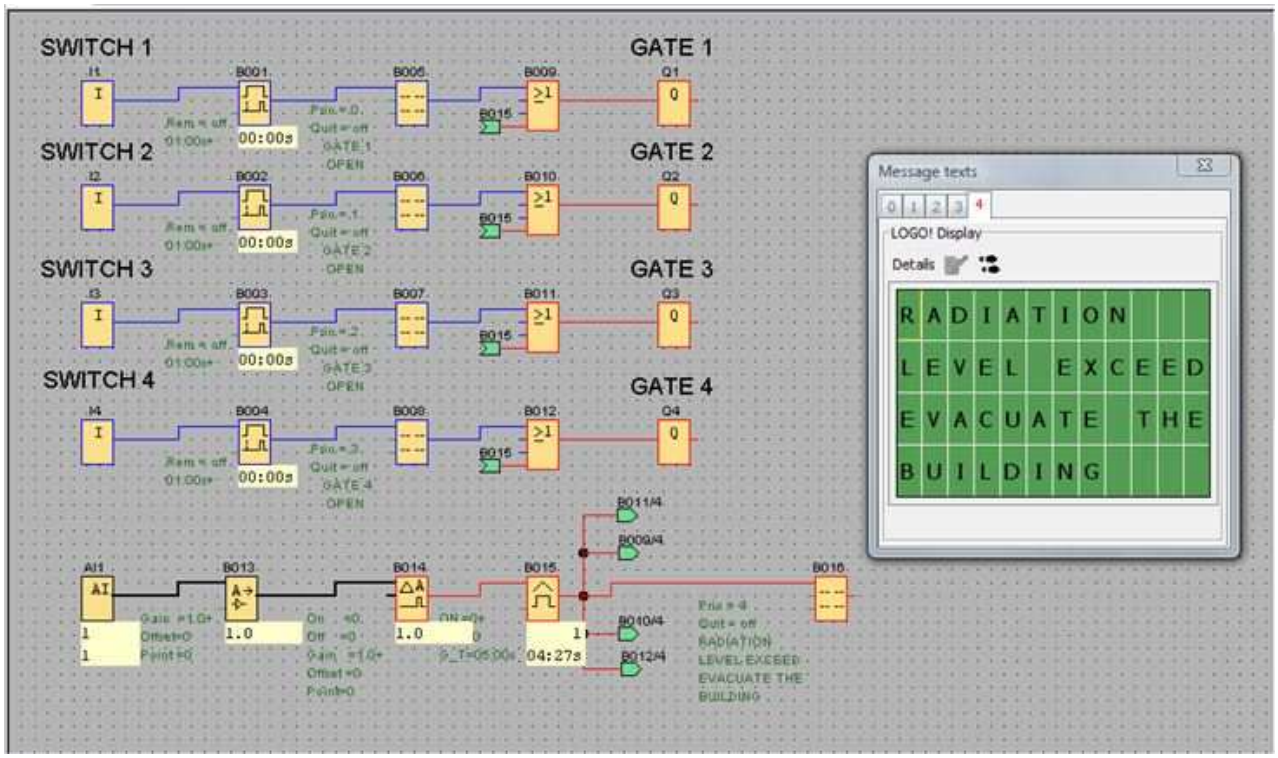


Figure 7. *All Gates are open when the radiation level exceeds FBD.*

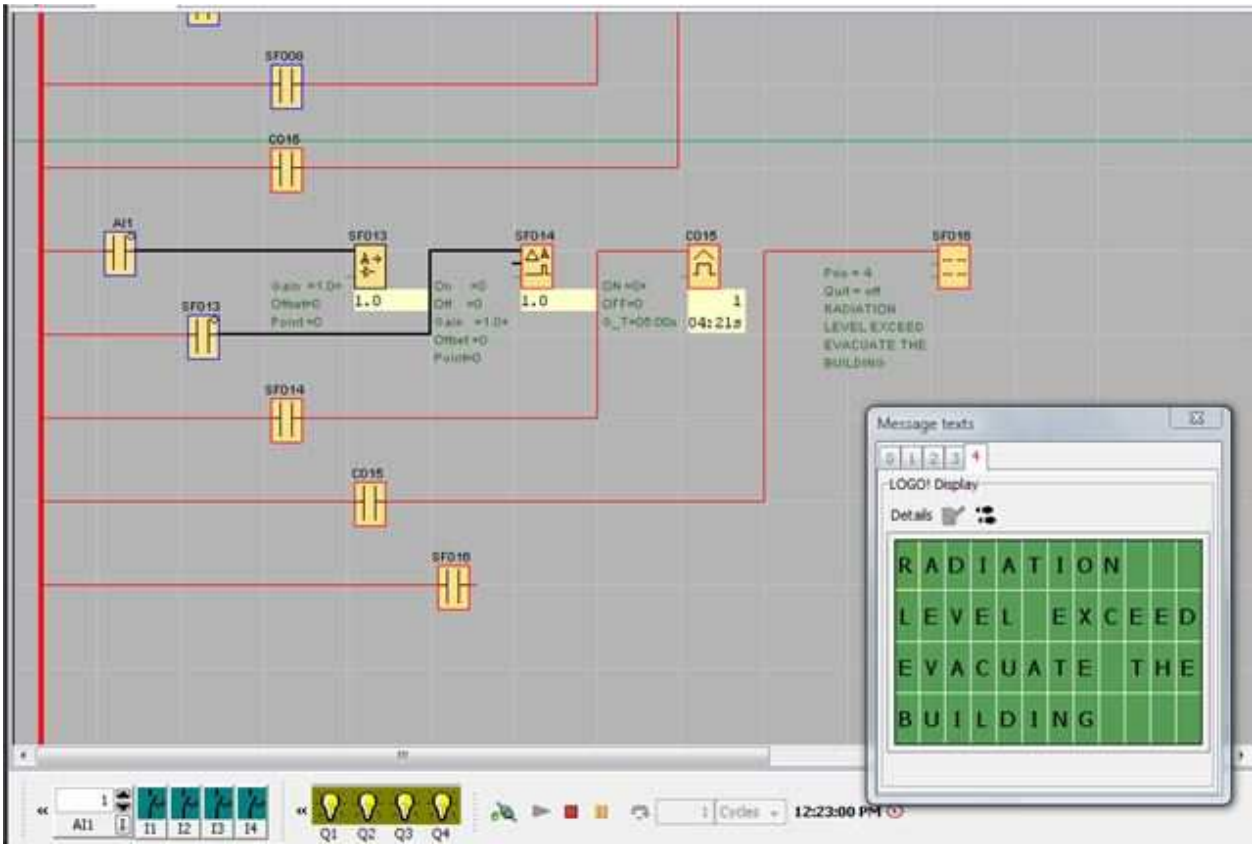


Figure 8. *All Gates are open when the radiation level exceeds in LAD.*

The diagram shows the PLC hardware, and the software was developed and implemented with the help of PLC. The output was obtained and produced in the below diagram.

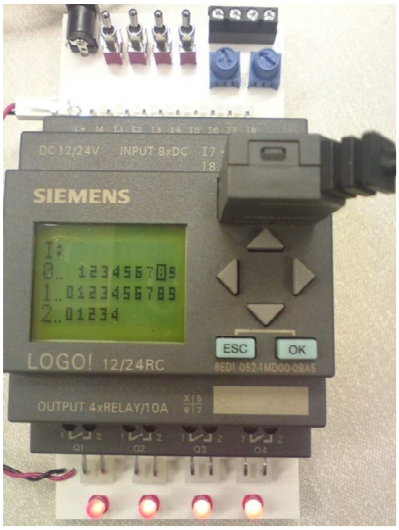


Figure 9. *All gates were open when the radiation levels exceeded.*

Table 1. *Tabular Colum of OR Gate.*

OR Gate:

I/P 1	I/P 2	I/P 3	I/P 4	O/P
0	0	0	0	0
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	1
1	0	0	1	1
1	0	1	0	1
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1

Table 2. *Tabular Colum of AND Gate.*

AND Gate:

I/P 1	I/P 2	I/P 3	I/P 4	O/P
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	1

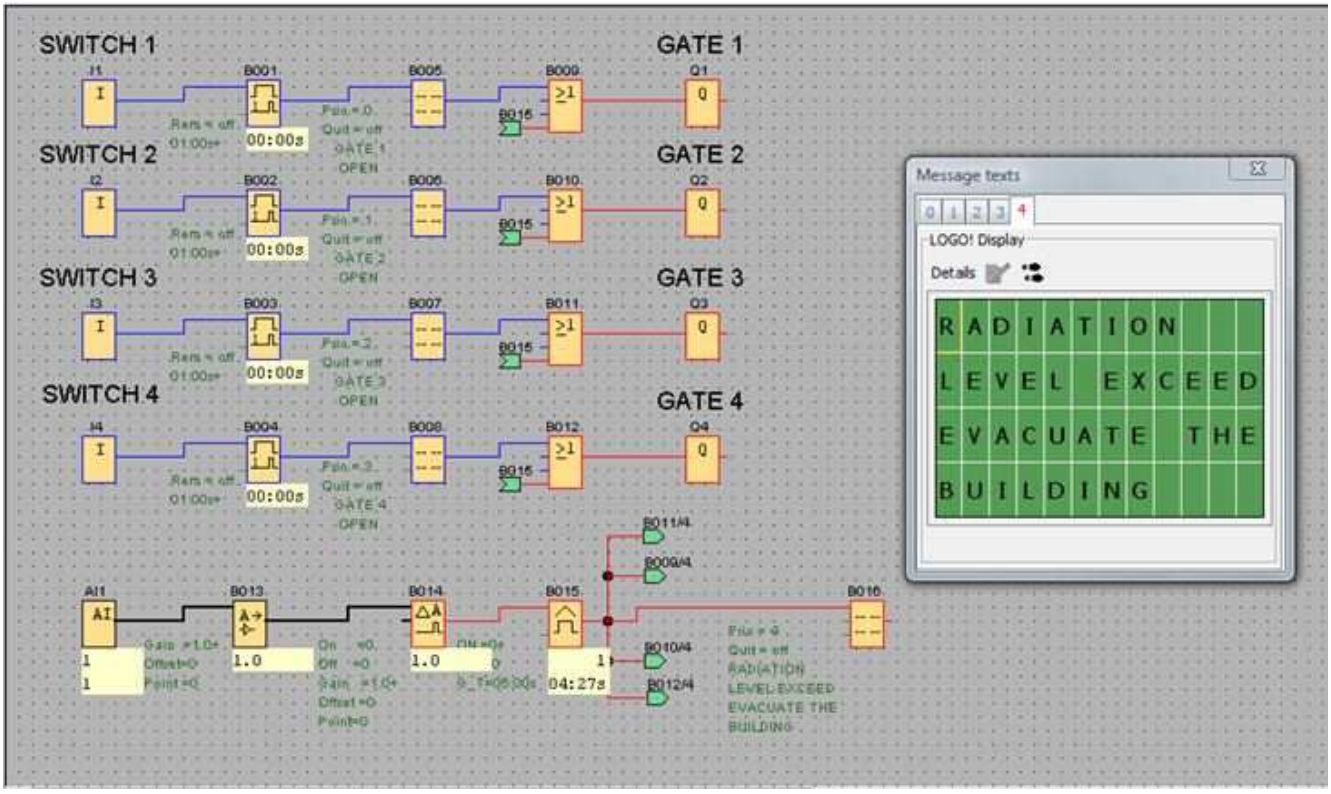


Figure 10. *All Gates are open when the radiation level exceeds FBD.*

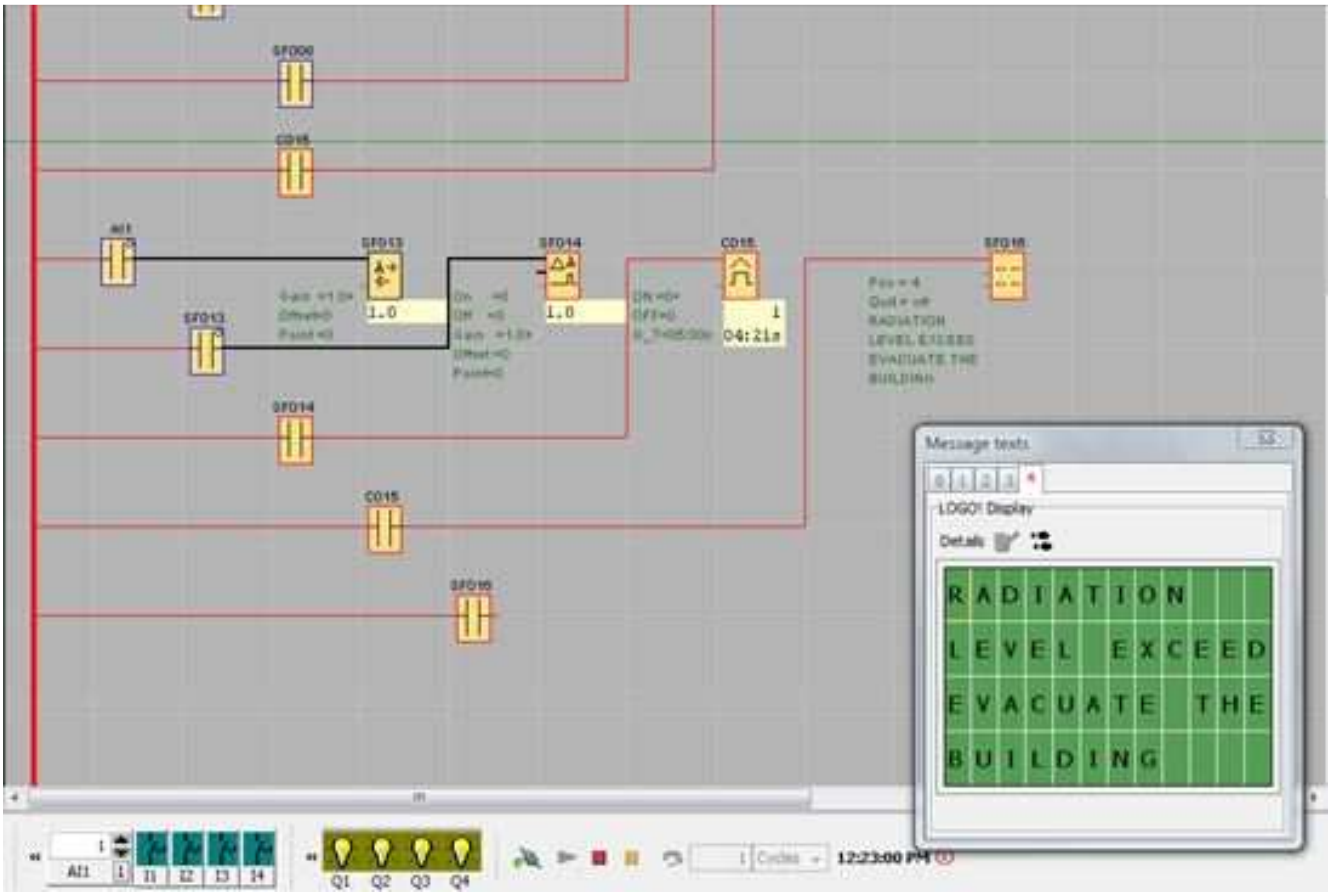


Figure 11. *All Gates are open when the radiation level exceeds in LAD.*

Figure 12. *All Gates are open when the radiation level exceeds PLC.*

6. Discussion of Results

The functional block diagram (FBD) graphical representation is used in functional blocks and consists of input contacts and output contacts, while the ladder diagram (LD) represents mainly relay logic operations. The network topologies of star and ring topologies utilised to design the system and timers are configurable output devices, a wide variety of delay-on and delay-off implemented [9]. The output of the timer appears as a switch elsewhere in the ladder diagram, and internal relays only exist in the internal memory of PLC. The proposed sensors will provide PLC inputs implemented on sensors and two different functions at different voltages. The communication path of PLC allows remote monitoring and control of intersystem communication. This section successfully achieved and discussed the results of functional block diagram (FBD) and ladder diagrams (LAD) outputs, along with PLC hardware results [10]. Expand the features of devices input and output, and analyses the internal architecture and various ladder programs and functional blocks. Programming the internal relays, timers, counter and fault diagnosis, and critically reflecting the design configuration of the circuits specific requirements of the system to interface into programmable devices and expand the synergy generated [11].

7. Conclusion

The system develops based on the hypothetical requirements for a lighting scheme for a building to have a floor plan showing where the lighting is automatic controlled. The Siemens PLC has a controller and sensors that manage the lighting. The design was developed for an access control system for a secure nuclear laboratory control and configuring the parameters like emergency evacuation and monitoring the temperature levels successfully with '3' relay outputs, and '2' analogue and '3' digital inputs with using of '1' timer and '1' counter. The circuit was developed using LOGO comfort software and implemented in PLC Hardware, and the produced results were accurate. All the contents like PLC hardware and LOGO comfort software were studied and implemented according to the procedure and observed the many new things while configuring the hardware.

References

[1] Fotouhi, M. and Eydgahi, Ali and Cavey, W. "Design of a programmable logic controller trainer". Vol 10, PP 17-20, 2000.

[2] W. Bolton, Programmable Logic Controllers, Fifth Edition, Newnes, 2009.

[3] M. G. Hudedmani, R. M. Umayal, S. K. Kabberalli, and R. Hittalamani, "Programmable logic controller (PLC) in automation", Adv. J. Grad. Res., Vol. 2, No. 1, PP 37-45, May 2017.

[4] Sadegh vosough and Amir vosough "PLC and its Applications" International Journal of Multidisciplinary Science and Engineering, November 2011, Vol. 2, No. 8.

[5] Programmable controllers- Part General information, international electrotechnical commission, IEC 1131-1, 1992 (BS EN 61131: 1994).

[6] W. Bolton "Programmable Logic Controllers" Elsevier Science & Technology, 2011, ISBN: 9780750681124.

[7] V. Pushpa Latha, K. R. Sudha, Swati Devabhaktuni, "PLC based Smart Street Lighting Control in Intelligent Systems and Applications" Vol 06, no 1, pp 64-72, 2013.

[8] Abueldahab, Yasser. "Installing PLC simulator for Ladder logic programming", November 2021.

[9] Wahlisch, Matthis, "Modeling the Network Topology", January 2010, PP 471-486, ISBN: 978-3-642-12330-6.

[10] Tasca, Laurence & Pingnton de Freits, Edison & Wagner, Flavio, "Enhanced Architecture for programmable Logic controllers targeting performance Improvement.", Vol 61, PP 306-315, June 2018.

[11] Wan, Hai & Chen, Gang & Song, Xiaoyu & Gu, Ming. "Formalisation and Verification of PLC Timers in Coq. 33rd Annual IEEE International Computer Software and Applications Conference. Vol 1, July 2009.

Polarization conversion in a polariton three-waveguide coupler

Ipsit Joshi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, ipsit.joshi29@gmail.com*

Madhusmita Mohanty, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, madhusmitamohanty@gmail.com*

Ankita Panda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ankitapanda21@yahoo.co.in*

Ashok Muduli, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, ashok.muduli29@yahoo.co.in*

ARTICLE INFO

Keywords:
Polarization
Waveguide
Waveguide coupler

ABSTRACT

We propose an approach to control of polarization of propagating light in a waveguide geometry. We suggest using a three-waveguide coupler etched in a planar semiconductor microcavity with embedded quantum wells. The structure allows converting a linear polarization of polaritons, photons hybridized with quantum well excitons, excited by a resonant optical pump, to circular polarization, herewith opposite circularly-polarized components are spaced apart from each other in different waveguides. The mutual effect of the evanescent coupling of waveguides and the TE-TM splitting of guided modes are in the basis of the predicted effect.

1. Introduction

Specially designed semiconductor heterostructures, optical microcavities with embedded quantum wells (QWs), support a specific regime of resonant interaction of light and matter, which is referred to as the strong coupling regime (Kavokin et al., 2017). New light-matter eigenstates known as exciton polaritons appear as a result of hybridization of cavity photon modes with excitons in QWs. Polaritons possess a spin degree of freedom linked to polarization of contributing photons. Recent achievements in the structure growth allow to highlight and enhance advantages of polaritons inherited from their components, including mobility of photons and controllability of excitons. In particular, in high-quality microcavities, polaritons possess lifetime of hundreds of picoseconds and are allowed to propagate in the cavity plane by distances of several millimeters (Steger et al., 2015; Myers et al., 2018).

Polariton structures form the basis of new optoelectronic devices including emitters of coherent light, interferometers, switches, etc. Particular attention is paid to the control of polarization of light (Liew et al., 2011). Our current study was inspired by the recent elegant work of Beierlein et al., 2021 devoted to a novel class of polariton devices for routing of polaritons. The devices represent etched in a microcavity polariton waveguides, waveguide clusters and circuits for directing and transforming polariton signals. In the paper, the authors report on the two-waveguide polariton coupler that enables controllable switching of the polariton flow between two exit ports. However, the polariton field is considered scalar, so the polarization effects are left beyond the scope of consideration. Polarization dynamics of polaritons and propagation of polarization domain walls in a single waveguide was performed in (Sich et al., 2018).

In this work, we expand the study of polariton polarization behavior on a three-waveguide coupler structure. We demonstrate theoretically a special case of such behavior when exciton polaritons injected by a resonant optical pump of linear polarization, when tunneling into adjacent waveguides, switch their polarization to circular. Herewith opposite circularly-polarized components are spaced apart in different waveguides. Thus the structure operates opposite to the polarization rectifier intended to converts circular polarization to linear (Sedov et al., 2019; Sedov et al., 2021).

2. Model

The considered structure is schematically shown in Fig. 1. It represents three polariton waveguides, which in a certain area pass in close proximity to each other. The waveguides are etched in a semiconductor microcavity with embedded QWs. The technique of partial etching of Bragg mirrors (Beierlein et al., 2021) enables photon coupling of the waveguides. Polaritons are excited by a resonant CW pump beam in the central waveguide.

To reveal polarization dynamics of polaritons, we solve numerically the generalized Pauli equation for the spinor $|\Psi\rangle = [\Psi_+(\mathbf{r},t), \Psi_-(\mathbf{r},t)]^{\mathrm{T}}$ with $\Psi_\pm(\mathbf{r},t)$ being the wave functions of the right- and left-circularly polarized polariton components:

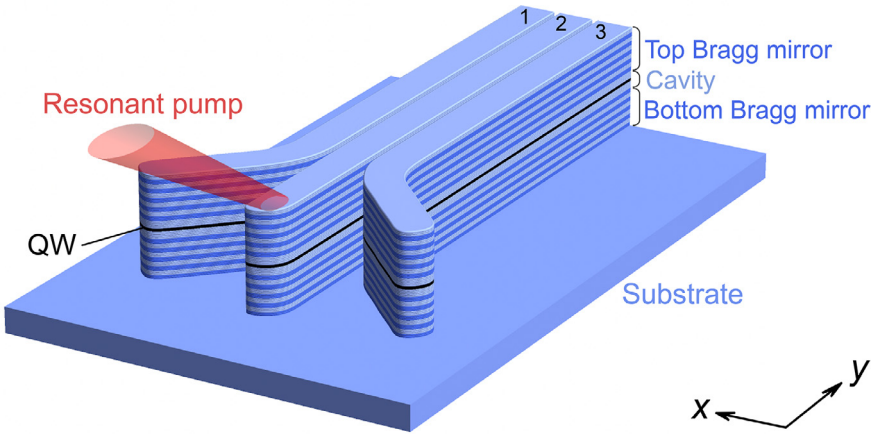


Fig. 1. Schematic of the polariton three-waveguide coupler with the excitation beam.

$$i\hbar\partial_t|\Psi\rangle = \left[\widehat{K} + \hat{\Theta} + U(\mathbf{r},t) - \hbar(\omega_p + i\gamma)\right] + |F\rangle, \quad (1)$$

where $\widehat{K} = \hbar^2\nabla^2/2m^*$ is the kinetic energy operator, m^* is the effective mass of polaritons taken as 10^{-34} kg. The operator $\hat{\Theta} = (\Delta/2)\left[\left(\partial^2_{xx} - \partial^2_{yy}\right)\hat{\sigma}_z + 2\left(\partial^2_{xy}\right)\hat{\sigma}_y\right]$ is responsible for the splitting of TE- and TM-polarized polariton modes, Δ is the splitting constant taken as $300\,\mu\text{eV}\,\mu\text{m}^2$. The effective potential $U(\mathbf{r},t) = V(\mathbf{r}) + (\alpha/2)[N_+(t)(\hat{\sigma}_0 + \hat{\sigma}_z) + N_-(t)(\hat{\sigma}_0 - \hat{\sigma}_z)]$ includes the stationary potential $V(\mathbf{r})$governed by the geometry of the waveguide coupler. Following Beierlein et al., 2021, we take it complex, $V(\mathbf{r}) = V_R(\mathbf{r}) - iV_I(\mathbf{r})$ of the following shape: $V_j(\mathbf{r}) \propto 1 - \exp\left[-(x/w_j)^{s_j}\right]$, where $j = R, I$, and for simulations we take $w_R = 4\,\mu\text{m}, s_R = 8$ and $w_I = 5\,\mu\text{m}, s_I = 16$. The second term in the effective potential describes spin-selective interaction of polaritons, α is the interaction constant taken $3\,\mu\text{eV}\,\mu\text{m}^2$. $N_\pm(t)$ are the occupations of the circularly polarized polariton states, $\hat{\sigma}_{0,x,y,z}$ are the Pauli matrices, γ is the polariton decay rate taken as $0.005\,\text{ps}^{-1}$. $|F\rangle \propto F(\mathbf{r})\exp(ik_p y)\,(f_+, f_-)^T$ is responsible for the resonant pump which excites polaritons in the central channel propagating in y direction with energy $\hbar\omega_p$ and quasimomentum k_p. The spatial distribution $F(\mathbf{r})$ is taken Gaussian of width $w_p = 2\,\mu\text{m}$. The rightmost parentheses describe polarization of the pump which is taken linear in the simulations, $f_\pm = 1/\sqrt{2}$.

3. Results and discussions

Fig. 2 shows the spatial distribution of the steady state density of polaritons $I(\mathbf{r}) = \Psi^\dagger\Psi$ as well as of the Stokes vector components $S_j(\mathbf{r}) = (\Psi^\dagger\hat{\sigma}_j\Psi)/I(\mathbf{r})$ $(j = x, y, z)$, that characterize linear (S_x), diagonal/antidiagonal (S_y) and circular (S_z) components of polarization of the polariton state. Clear oscillations of the density of polaritons in the waveguides are observed accompanied by redistribution of polaritons between the central and the peripheral waveguides. In the upper panel in Fig. 3 we show the variation in y direction of the density of polaritons in the central waveguide $\bar{I}^{(2)}(y)$ as well as in the side waveguides $\bar{I}^{(1)}(y) + \bar{I}^{(3)}(y)$ compared to the integral density in all waveguides, where $\bar{I}^{(j)}(y) = \int_{A_j} I(\mathbf{r})dx$ $(j = 1, 2, 3)$, A_j is the cross-section width of the jth waveguide. In the lower panel in Fig. 3 the variation of the circular polarization component in different waveguides $\bar{S}_z^{(j)} = \int_{A_j}(\Psi^\dagger\hat{\sigma}_z\Psi)dx/\bar{I}^{(j)}(y)$ is shown. The integral density in all waveguides exponentially decreases with y due to losses in the structure.

Somewhat expected result of our simulations is that the polariton polarization in the central waveguide remains linear (X) coinciding with the polarization of the optical pump over the entire length of propagation. Our supplementary simulations (not presented here) confirms preservation of polarization for the orthogonal (Y) linearly polarized pump.

The most remarkable peculiarity of the polariton polarization distribution in the considered structure is the appearance of a strong circular polarization in the side waveguides. Circular polarization degree experiences oscillations in y direction, herewith the oscillations in different waveguides are in antiphase with each other. One should mention that the oscillations of both the density and polarization are not regular and occur with different periods. This can be verified by relating the maxima of the density and the circular polarization distribution in the upper waveguide in the top and bottom panels in Fig. 2. While the position of the third density maximum (around $145\,\mu\text{m}$) nearly coincides with one of the circular polarization, the second maxima (around $70\,\mu\text{m}$) are shifted from each other by a distance of about $15\,\mu\text{m}$. The circular polarization in the first maximum in the side waveguides is weakly pronounced due to the proximity to the pump spot. This mismatch is described by different mechanisms causing oscillations of the density and polarization. In the case of the density, we deal with the Josephson-like oscillations emerging in waveguides coupled via transmissive barriers (Beierlein et al., 2021). In the basis of polarization oscillations is the optical spin Hall effect (Kavokin et al., 2005). The latter takes place in the presence of the splitting in TE- and TM-polarized optical (polaritonic) modes and manifests itself in the appearance of alternating domains of linear or circular polarization in the spatially distributed polariton field (Kammann et al., 2012; Cilibrizzi et al., 2016). In the paper of Cilibrizzi et al., 2016, the formation of two-dimensional quadruplet oscillating patterns in linear and circular polarizations in the microcavity plane under the nonresonant linearly polarized pump was demonstrated. Herewith, the particular directions, which coincide with the linear polarization axes (X and Y), remained free of oscillations. Analogously, in our case, the injection of polaritons in the central waveguide with the optical pump of the linear polarization parallel (Y) or orthogonal (X) to the axis of the waveguide does not lead to emergence of oscillations and allows polaritons confined in the central waveguide keeping their polarization inherited from the pump. This, however, does not apply to the side waveguides which shelter polaritons that deviate from the axis of the central waveguide. In the paper of Beierlein et al., 2021 the redistribution of polaritons in the reciprocal space is clearly explained in terms of the mutual influence of the ground and excited (symmetric and antisymmetric) modes in the coupler. The confinement in the narrow side waveguides modifies the manifestation of the optical spin Hall effect, which results in formation of alternating circular polarization domains in a quasi-one-dimensional geometry. The frequency of oscillations of polarization is determined by the magnitude of the

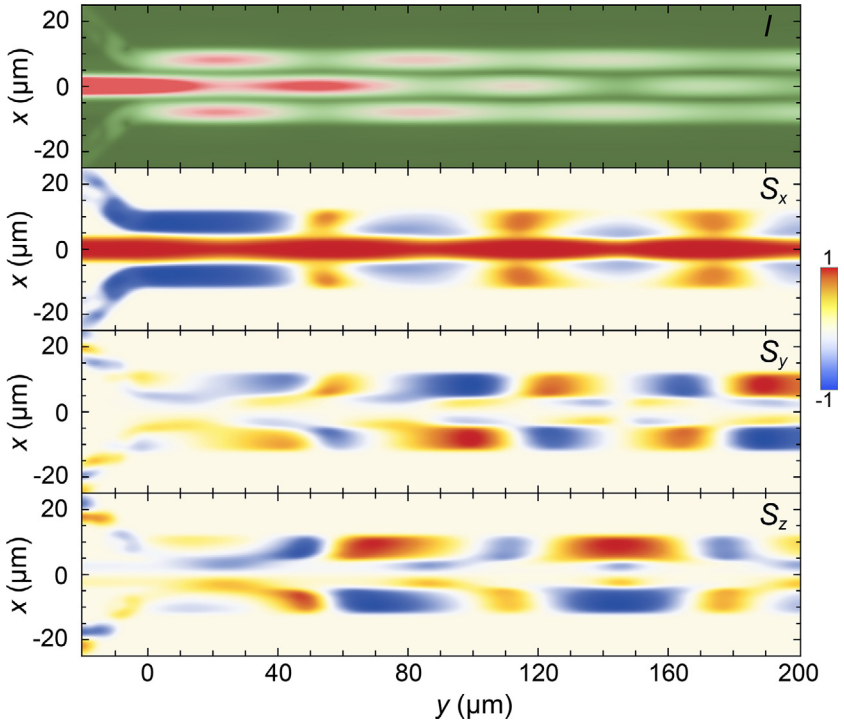


Fig. 2. Spatial distribution of the density $I(\mathbf{r})$ and the Stokes vector components $S_{x,y,z}(\mathbf{r})$ characterizing polarization of polaritons in the steady state. The wavenumber of the pump is $k_p = 0.6\mu m^{-1}$.

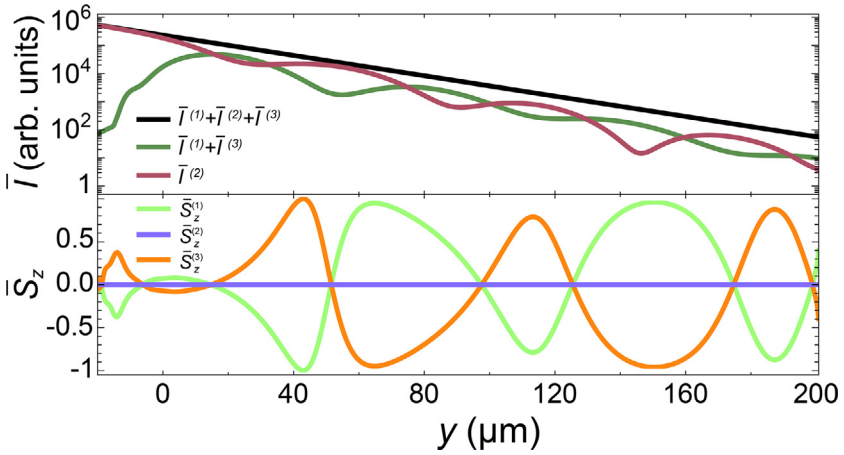


Fig. 3. Variation of the density of polaritons (upper panel) and the degree of circular polarization (lower panel) in y direction in different waveguides. The parameters are the same as in Fi.g. 2.

TE-TM splitting, Δk_p^2, which is further enhanced by confinement across the waveguide by an amount inversely proportional to the squared width of the later, see (Sich et al., 2018). The frequency of oscillations can be adjusted during the experiment by changing the quasimomentum k_p of the pump. In the unbounded structure of a planar microcavity, the polarization oscillations are accompanied by the oscillations of the center-of mass trajectory of a polariton wave packet in the direction perpendicular to the propagation direction (Sedov et al., 2018; Sedov et al., 2020; Sedova et al., 2020). However, in the presence of the waveguide confining potential these oscillations are suppressed.

In summary, we have demonstrated theoretically the possibility of a wide control over polarization of light propagating in novel semiconductor waveguide circuits. The proposed structure of a three-waveguide coupler enables transforming linear polarization of exciton polaritons to circular polarization and separation of opposite polarization components in space. This study was performed for the case of the strong light-matter coupling to pay tribute to the work of Beierlein et al., 2021. Nevertheless, a similar effect is expected to be observable in pure photonic systems, as both mechanisms of oscillations of the density and polarization are due to the photonic part.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Beierlein, J., Rozas, E., Egorov, O.A., Klaas, M., Yulin, A., Suchomel, H., Harder, T.H., Emmerling, M., Martín, M.D., Shelykh, I.A., Schneider, C., Peschel, U., Viña, L., Höfling, S., Klembt, S., 2021. Propagative oscillations in codirectional polariton waveguide couplers. Phys. Rev. Lett. 126,. https://doi.org/10.1103/PhysRevLett.126.075302 075302.

Cilibrizzi, P., Sigurdsson, H., Liew, T.C.H., Ohadi, H., Askitopoulos, A., Brodbeck, S., Schneider, C., Shelykh, I.A., Höfling, S., Ruostekoski, J., Lagoudakis, P., 2016. Half-skyrmion spin textures in polariton microcavities. Phys. Rev. B 94,. https://doi.org/10.1103/PhysRevB.94.045315 045315.

Kammann, E., Liew, T.C.H., Ohadi, H., Cilibrizzi, P., Tsotsis, P., Hatzopoulos, Z., Savvidis, P.G., Kavokin, A.V., Lagoudakis, P.G., 2012. Nonlinear optical spin hall effect and long-range spin transport in polariton lasers. Phys. Rev. Lett. 109,. https://doi.org/10.1103/PhysRevLett.109.036404 036404.

Kavokin, A., Baumberg, J., Malpuech, G., Laussy, F., 2017. In: Microcavities Series on Semiconductor Science and Technology. 2 ed.,. OUP Oxford.

Kavokin, A., Malpuech, G., Glazov, M., 2005. Optical spin hall effect. Phys. Rev. Lett. 95,. https://doi.org/10.1103/PhysRevLett.95.136601 136601.

Liew, T., Shelykh, I., Malpuech, G., 2011. Polaritonic devices. Physica E: Low-Dimensional Syst. Nanostruct. 43, 1543–1568. https://doi.org/10.1016/j.physe.2011.04.003. URL:https://www.sciencedirect.com/science/article/pii/S1386947711001160.

Myers, D.M., Ozden, B., Steger, M., Sedov, E., Kavokin, A., West, K., Pfeiffer, L.N., Snoke, D.W., 2018. Superlinear increase of photocurrent due to stimulated scattering into a polariton condensate. Phys. Rev. B 98,. https://doi.org/10.1103/PhysRevB.98.045301 045301.

Sedov, E., Sedova, I., Arakelian, S., Kavokin, A., 2021. Polygonal patterns of confined light. Opt. Lett. 46, 1836–1839. https://doi.org/10.1364/OL.418337. URL:http://ol.osa.org/abstract.cfm?URI=ol-46-8-1836.

Sedov, E.S., Rubo, Y.G., Kavokin, A.V., 2018. Zitterbewegung of exciton-polaritons. Phys. Rev. B 97,. https://doi.org/10.1103/PhysRevB.97.245312 245312.

Sedov, E.S., Rubo, Y.G., Kavokin, A.V., 2019. Polariton polarization rectifier. Light: Sci. Appl. 8, 79. https://doi.org/10.1038/s41377-019-0189-z.

Sedov, E.S., Sedova, I.E., Arakelian, S.M., Kavokin, A.V., 2020. Magnetic control over the zitterbewegung of exciton–polaritons. New J. Phys. 22,. https://doi.org/10.1088/1367-2630/aba731 083059.

Sedova, I.E., Sedov, E.S., Arakelian, S.M., Kavokin, A.V., 2020. Oscillating motion of exciton–polaritons in anisotropic microcavities. Bull. Russian Acad. Sci.: Phys. 84, 1453–1458. https://doi.org/10.3103/S1062873820120333.

Sich, M., Tapia-Rodriguez, L.E., Sigurdsson, H., Walker, P.M., Clarke, E., Shelykh, I.A., Royall, B., Sedov, E.S., Kavokin, A.V., Skryabin, D.V., Skolnick, M.S., Krizhanovskii, D.N., 2018. Spin domains in one-dimensional conservative polariton solitons. ACS Photonics 5, 5095–5102. https://doi.org/10.1021/acsphotonics.8b01410.

Steger, M., Gautham, C., Snoke, D.W., Pfeiffer, L., West, K., 2015. Slow reflection and two-photon generation of microcavity exciton–polaritons. Optica 2, 1–5. https://doi.org/10.1364/OPTICA.2.000001. URL:http://www.osapublishing.org/optica/abstract.cfm?URI=optica-2-1-1.

Intensification of noise tolerance against Rayleigh backscattering for bidirectional 10 Gbps WDM-FSO network by employing dual band of OFDM signal

Amit Kumar Jha, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, amitkumar.jha@outlook.com*

Ashis Acharya, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, ashisacharya12@gmail.com*

S. Sivasakthiselvan, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, s.sivasakthiselavan@live.com*

Maheshwari Rashmita Nath, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, maheswarinath1@outlook.com*

ARTICLE INFO

Keywords:
Rayleigh backscattering (RB) noise
Orthogonal frequency division multiplexing (OFDM)
Free space (FSO) communication
Doublet lens scheme

ABSTRACT

A noteworthy scheme to transport dual band 10 Gbps 16 quadrature amplitude modulated orthogonal frequency division multiplexed signal for downlink and uplink channel over 55 km single mode fiber as well as 650-m free space is proposed and demonstrated. Moreover, noise tolerance against Rayleigh backscattering, that arises in bidirectional transmission system is enhanced as different bands of subcarrier are used for uplink and downlink transmission in our proposed system. Negligible power penalty (<1 dB), good bit error rate value (under FEC limit i.e. 3.8×10^{-3}), proper error vector magnitude (<12.35%) and signal to noise ratio (>19 dB under FEC limit) marks the reliability of the proposed design. We observed that dual band subcarrier modulation scheme is able to decrease large power penalty (~8 dB) for total transmission link. The results from our proposed network makes it more potent to give an alternative platform with long reach, less noise, high data rate transmission which is the top most thirst to the upcoming generation.

1. Introduction

Wavelength division multiplexing passive optical networks (WDM-PONs) are deliberated to be most auspicious solution to attain exponentially growing bandwidth demand of end users for future optical access technology (Choudhury, 2015; Wong, 2012; Ni et al., 2015; Das and Patra, 2014). WDM-PON network is also upgraded to keep up the transmission of high data rate including 10 Gbps bidirectional transmission for both downlink and uplink (Straullu et al., 2017; Choudhury and Khan, 2016; Das and Patra, 2014). Wavelength reuse technique in WDM-PON network with the assistance of reflective semiconductor optical amplifier (RSOA) makes the system simpler and cost effective besides ensuring the matter of effective use of wavelength in bidirectional system for down and uplink (Cano et al., 2010; Parolari et al., 2014). Employment of reflective electro-absorption modulator (R-EAM), Fabry-Perot laser diode (FP-LD) and RSOA for uplink reuse in optical network unit (ONU) of the system ensues cost effectiveness and simpler network in structure (Lin et al., 2009; Chow et al., 2008; Girault et al., 2008). Now a days, besides WDM technology, an another advance technology with higher spectral efficiency named orthogonal frequency division multiplexing (OFDM) are used to support very high data rate in PON system (Li et al., 2020; Liaw et al., 2017; Mallick et al., 2018). Along with the free space optics (FSO), OFDM technique offers an incomparable solution to the bottleneck issues of radio frequency (RF) communication network and becomes a tempting technology for the next generation wireless communication world (Mallick et al., 2020). The multiple subcarrier modulation technique puts forward a conflict against the dispersion effect as channel impulse responses are lesser than the time period of OFDM symbols (Tsai et al., 2015; Cvijetic, 2012; Kabir, 2008; Djordjevic and Vasic, 2006; Yeh et al., 2012; Acar and Aldırmaz-Çolak, 2021). On the other hand, FSO network becomes voguish in communication world owing to its tremendous advantages like very large bandwidth for data transmission, independency on electromagnetic interference, greater speed than broadband and signal transmission flexibility even in the urban and rural areas without installation complexity (Khalighi and Uysal, 2014; Chaman-Motlagh et al., 2010; Chaudhary et al., 2014a, 2018b, 2018c; Upadhyay et al., 2019; Zhang et al., 2021). Use of OFDM-FSO technique together ensues the enhancement of spectral efficiency of the system and the bit error rate (BER) value (Balaji and Prabu, 2018; Wang et al., 2016; Sharma and

Sushank, 2014; Song et al., 2017; Sudheer and Mandloi, 2016). Employment of doublet lens scheme in FSO communication could be cause of enhancement of communication distance as this scheme supports to transmit signal of high data rate either by coupling or suppressing or expanding the beam size of laser light (Li et al., 2016a, 2016b; Tsukamoto and Komaki, 2007). However, all the aforementioned networks may suffer from performance degradation that arises due to interferometric crosstalk noise. There are so many techniques already have been proposed and demonstrated by several research groups to mitigate noise that is generated by RB (Kang and Han, 2006; Oh et al., 2008; Mandal et al., 2021). RB noise naturally introduced whenever the signals of same wavelength are allowed to transmit through a single fiber for both up and downlink. Crossed network technique also regarded as a fruitful alternative to minimize RB noise effect in transport system (Lin et al., 2011; Yeh et al., 2012; Chiuchiarelli et al., 2009; Mandal et al., 2021).

In this paper, we proposed and demonstrated an elevated scheme for transmission of RB noise mitigated 10 Gbps-16 QAM OFDM signal over 55 km as well as 650 m FSO for wireless users. To avoid the RB noise effect in our bidirectional set up, two dissimilar bands of OFDM sub-carriers are modulated for down and uplink transmission. A band of subcarriers different from downlink is remodulated by 10 Gbps-16 QAM signal in RSOA and transmitted over 55 km SMF and received at central office (CO) to maintain the bi-directionality of the system. Two doublet lenses are utilized for transportation of information over 650 m free space. A comparative study by changing the FSO distance also examined to check the maximum limit of reliable performance of this proposed architecture. Negligible power penalty, proper EVM value, clear constellation diagrams, BER and SNR performance present the proof of fruitful transmission of proposed set up.

2. Experimental setup

As a proof of concept, Fig. 1 indicates the experimental set up to check the feasibility of our proposed bidirectional scheme to transport RB noise eliminated information over 55 km SMF along with 650 m FSO link. The working principle of this architecture is explicated through three major sections namely CO, remote node (RN) and ONU. A carrier signal with central frequency at 1550.72 nm from a FPLD is externally modulated by 10 Gbps-16 QAM signal, which is created by an arbitrary waveform generator (AWG). Two different bands of subcarriers are modulated by 10 Gbps signal for downlink and uplink transmission to eschew RB noise. A subcarrier bandwidth of 2.8 GHz from 3.6 to 6.4 GHz is modulated by 10 Gbps signal by making use of/employing 16 QAM OFDM modulation along with 512-point fast Fourier transform (FFT) size and 1/64 cyclic prefix (CP) length. MATLAB programming is employed to generate OFDM signal via an AWG. Sampling rate of 10 GS/s and 8-bit DAC are applied here. Through 30 km SMF, the modulated 10 Gbps signal is launched to RN via local exchange (LE) and again transmitted over 25 km SMF. LE section simply consists of two erbium doped fiber amplifiers (EDFAs) and serve as an amplification point for both up and downlink signal. Finally, the power of the downlink information is split into two parts by optical coupler in ONU section. Among two, power of one part is boosted and optimized by employing EDFA and variable optical attenuator (VOA) which enhance the transmission performance too. After that, this amplified and optimized version of information is communicated through 650 m FSO link by utilizing two doublet lenses. Two doublet lenses may use to emit and received the laser light for 650 m FSO communication within two building. The natural expansion (Mallick et al., 2019) of laser light would cause an obstacle in coupling the laser light and fiber ferrule. By properly adjusting the field of view of fiber ferrule with the laser light, the doublet lens at the receiving end succours to optimize the performance of the system by improving BER and increasing received optical power increasing received optical power. Fig. 2 (a) and (b) gives the pictorial description of the relationship of fiber ferrule's field of view with doublet lens's field of view. Larger power will be amassed if the field of view of fiber ferrule is larger than that of doublet lens and lesser power will be accumulated whenever the field of view of fiber ferrule smaller than the field of view of doublet lens (Huang et al., 2020). For 650 m FSO communication two doublet lenses (AC 508–150-C) with focal-length1 (F_1), focal-length2 (F_2) and diameter of 117.7 mm, 150 mm and 50.8 mm respectively are employed here. So, diameter of the laser beams with numerical aperture (NA) of SMF of 0.14 is (Mallick et al., 2019)

$$d = 2*(F*NA) = 42 \text{ mm} \quad (1)$$

The diameter of the laser beam is lesser than that of the doublet lens1 (50.8 mm) that ensues the reduction of coupling loss induced by it and

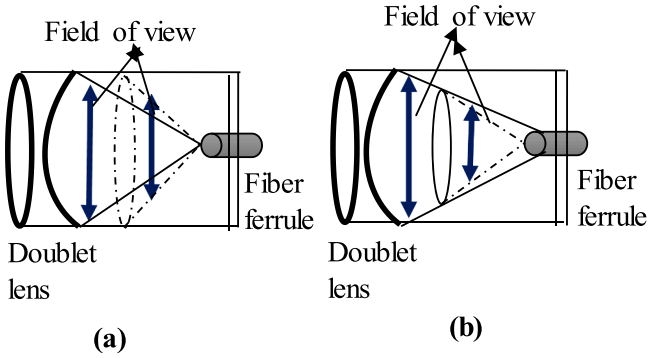


Fig. 2. (a) Doublet lens's field of view is lesser than fiber ferrule's field of view. (b) fiber ferrule's field of view is smaller than doublet lens's field of view.

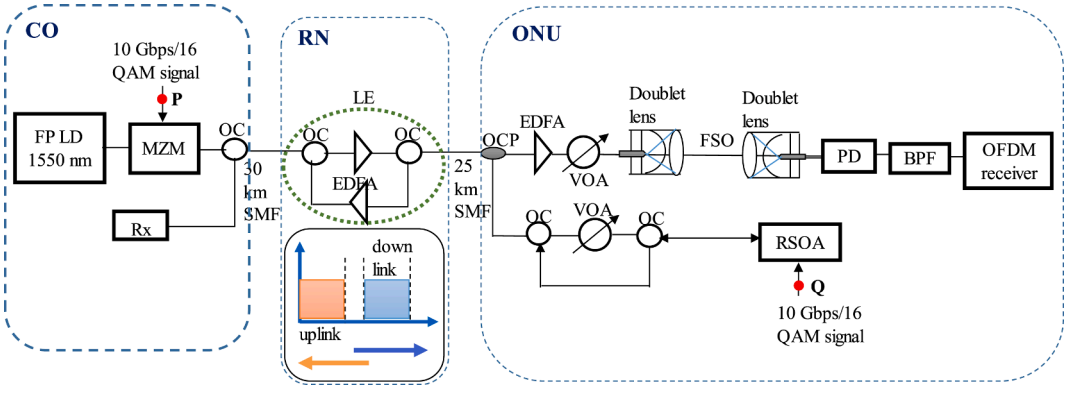


Fig. 1. Block diagram of proposed set up for intensification of noise tolerance against RB for bidirectional 10 Gbps WDM network by employing dual band of OFDM signal for transportation of data through FSO link utilizing doublet lens scheme.

express the feasibility of the doublet lens1 for FSO link. The beam radius (R) for this lens that is related with spatial frequency cut off (SFC) can be expressed as follows (Mallick et al., 2019):

$$R = 2.3^{*}\frac{1}{[2\pi^{*}SFC]} = 2.89\ \mu m \tag{2}$$

With SFC = $1/\{\lambda^{*}\text{focalratio}\}$ where λ is the wavelength of light. Now the divergence out of the objective lens (θ) is:

$$\theta = \frac{R}{150(mm)} = 19.26\ X10^{-6} \tag{3}$$

so, the spot size (R_0) over 650 m(l) FSO link is given by

$$R_0 = \left[d^2 + (2\theta l)^2\right]^{\frac{1}{2}} = 48.8\ mm \tag{4}$$

The diameter (50.8 mm) of the doublet lens2 is larger than the value of R_0 that leads to lessen the coupling loss generated by doublet lens2. For larger beam size in case of Gaussian laser beam, causes larger Rayleigh distance (Mallick et al., 2019). Doublet lens2 plays a paramount role in coupling the laser beam into the ferrule of SMF by reducing the beam size. The couple of doublet lenses used in our proposed architecture, are able to communicate the information over 650 m FSO link successfully. After FSO link, photodiode (PD) is employed to detect the signal and then a band pass filter (BPF) filtrates the noise, that generates owing to FSO communication. Finally, the signal is received by OFDM receiver through some sections namely 16 QAM demodulation format, equalizer, fast Fourier transform (FFT), serial to parallel converter (S/P), guard removal, analog to digital converter. Offline digital signal processing technique is employed to recover the OFDM data in the receiver section of ONU. And the other part of transmitted signal remodulated by 10 Gbps/16 QAM OFDM signal in RSOA with the frequency band of 0–2.8 GHz for uplink transmission. RSOA is operated in its saturation region with 65 mA operating current that leads to reduce the back-scattering effect. To avoid effectively the RSOA seeding power in downlink, a VOA is employed in between two circulators. This remodulated 10 Gbps OFDM signal then transmitted through RN over same 55 km SMF and communicated through 650 m FSO link and finally detected by similar process as earlier used in ONU.

3. Results and discussion

Spectra of different subcarrier bands of OFDM signal utilized for downlink and uplink transmission are depicted in Fig. 3 (a) and (b) [insert Fig. 1 (P and Q point)]. A subcarrier band from 3.6 to 6.4 GHz is modulated by 10 Gbps signal for downlink and a frequency band of 0–2.8 GHz is modulated by same data rate for uplink transmission to mitigate RB noise.

Measured BER value for 10 Gbps downlink OFDM signal and B2B under different received optical power of –22 dBm to −16.2 dBm along with constellation diagrams are indicated in Fig. 4. The BER based on the subcarrier SNRs (=$1/[EVM_{RMS}]^2$) is defined as (Mehedy et al., 2012).

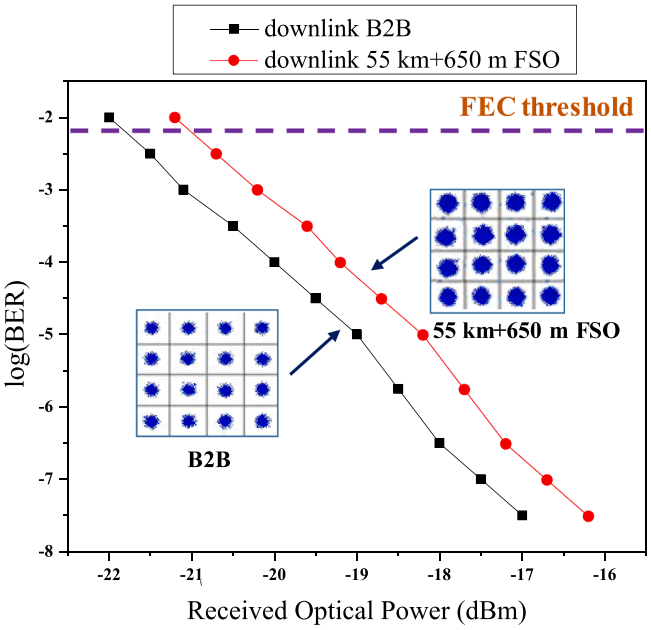


Fig. 4. BER curves and constellation diagrams for downlink 10 Gbps OFDM signal over 55 km SMF as well as 650 m FSO.

$$BER = \left(\frac{2\left(1-\frac{1}{L}\right)}{\log_2 L}\right) Q\left\{\sqrt{\left(\left(\log_2 L\right)\frac{1}{L^2-1}\right)(2SNR/(\log_2 M))}\right\} \tag{5}$$

where, L = 4 of 16 QAM constellation that represents the number of levels in each dimension and M = 16, which expresses the number of constellation point. Receiver sensitivity of −20.2 dBm with BER value of 3.8×10^{-3} (under FEC limit) is achieved by 10 Gbps/16 QAM downlink OFDM signal after transmitting through 55 km SMF along with 650 m FSO link. A very low power penalty of 0.8 dB is recorded between B2B and over total transmission link at a BER of 10^{-3}. Clear constellation diagrams for both B2B and over 650 m free space link along with 55 km SMF are also displayed in Fig. 4 that indicates the fruitfulness of transmission of information. The EVM value of 12.1% for B2B and 12.22% for 55 km SMF and 650 m FSO transmission are observed from constellation diagrams. Fig. 5 gives the measured BER curve vs received optical power along with constellation diagrams of uplink transmission for both B2B and over total transmission link. Receiver sensitivity of −18.4 dBm for transmission of 55 km SMF as well as 650 m free space is achieved by uplink OFDM signal at a BER of 3.8×10^{-3}(under FEC limit). Power penalty about 0.9 dB is measured between B2B and after transmission over SMF plus FSO link. Clear constellations are also depicted in Fig. 5

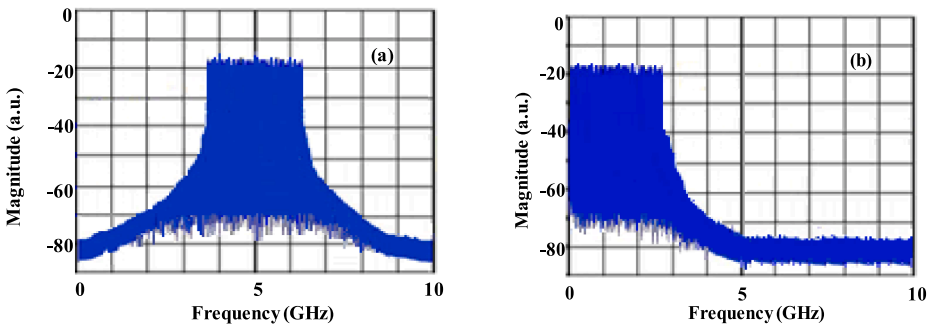


Fig. 3. (a) OFDM spectra at "P" for downlink and (b) OFDM spectra at "Q" for uplink transmission.

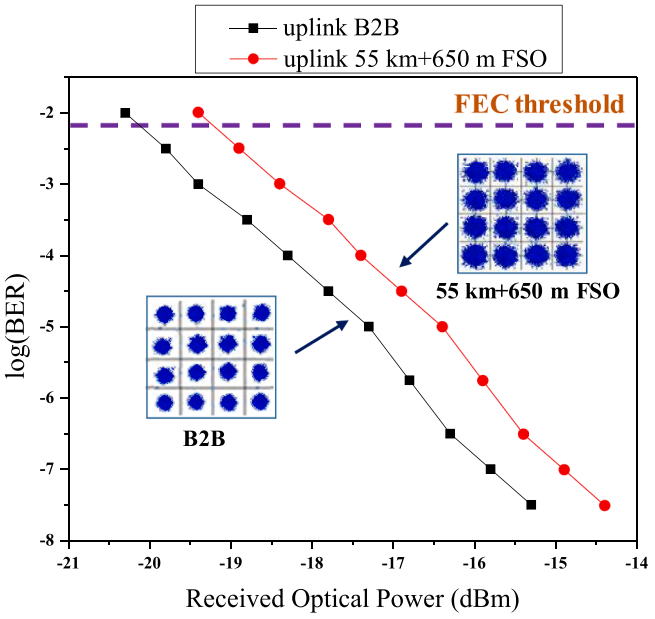


Fig. 5. BER curves and constellation diagrams for uplink 10 Gbps OFDM signal over 55 km SMF as well as 650 m FSO.

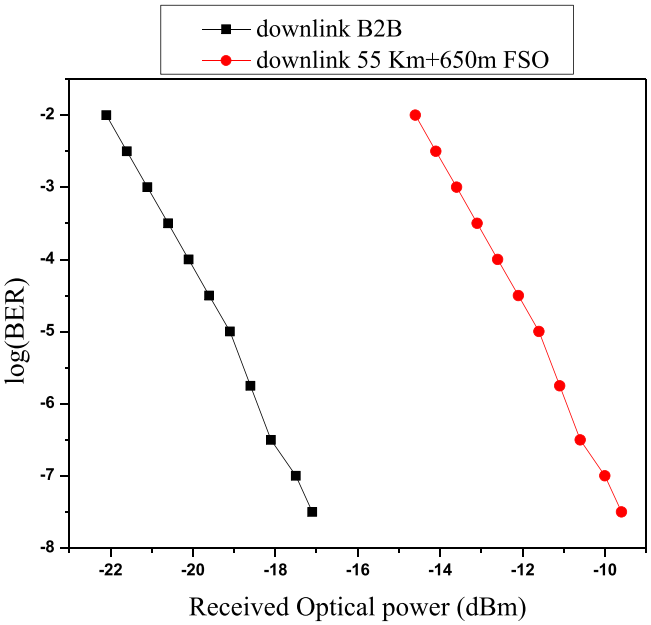


Fig. 6. BER for downlink 10 Gbps OFDM signal over 55 km SMF as well as 650 m FSO without employing dual band subcarrier scheme.

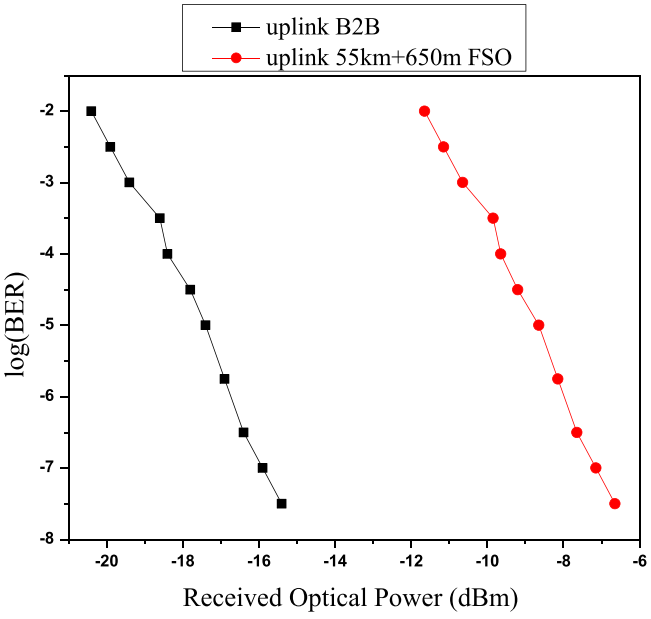


Fig. 7. BER for uplink 10 Gbps OFDM signal over 55 km SMF as well as 650 m FSO without employing dual band subcarrier scheme.

for both aforementioned cases of 12.18% and 12.30% EVM values, that are observed from constellations for 16 QAM OFDM signal.

Measured BER curve for 10 Gbps OFDM signal for downlink and uplink transmission over the total transmission link without employing dual band subcarrier modulation scheme are depicted in Figs. 6 and 7 respectively. Receiver sensitivity of −13.6 dBm and −10.65 dBm are achieved by the downlink and uplink signal respectively at the BER of 3.8×10^{-3}(under FEC limit). Larger power penalty is about 7.5 dB is recorded between B2B and over SMF and FSO link and that is recorded about 8.75 dB for uplink. The system performance degrades severely due to RB noise whenever same band of subcarriers are modulated by signal for both downlink and uplink transmission.

The SNR performance of OFDM subcarriers in the frequency band of 3.6 to 6.4 GHz by employing 16 QAM format for B2B and over total SMF-FSO link with fixed photodiode input power of −11 dBm are plotted in Fig. 8. The EVM is related to SNR of each subcarrier as follows (Mehedy et al., 2012):

$$EVM_{RMS} = \sqrt{\frac{1}{SNR}} = \sqrt{\{(\sum_{j=1}^{S_a}[\sum_{k=1}^{S_b}|X_{jk} - \overline{X_{jk}}|^2])/(S_a S_b P_{avg})\}} \quad (6)$$

where, $\overline{X_{jk}}$ and X_{jk} are the normalized ideal and normalized estimated received symbol of 16 QAM constellation respectively. $|X_{jk} - \overline{X_{jk}}|$ represents the error value, S_a and S_b denote total number of OFDM symbols and total number of data carrying subcarriers in each OFDM symbols respectively. the average power of the normalized constellation is denoted by P_{avg}. For B2B connection SNRs are obtained between 24.38 dB and 20.06 dB while, these are recorded 24.35 dB and 20 dB for 55 km SMF and 650 m FSO connection. Slight degradation of SNR performance is occurred in SMF-FSO link in comparison to B2B connection due to residual dispersion. The SNR value of each subcarrier should be>19 dB to meet the BER of $\leq 3.8 \times 10^{-3}$ (under FEC limit). This result clearly represents the effective mitigation of noise arises from data overlapping is achieved by using different band of subcarrier signals for up and downlink transmission.

A comparative study of BER values with detected optical power for different free space transmission distances by employing doublet lens scheme is graphically represented in Fig. 9. Measured BER values are plotted for 640 m, 650 m, 660 m, 663 m FSO transmission. Fig. 9 clarifies that with increasing transmission distance, BER values also increase. For our proposed system, up to 650 m FSO distance BER value is of the order of 10^{-3} and beyond that BER also greater than the 10^{-3}. With increase of transmission distance, photo current decreases gradually that ensues the reduction of power of the signal. As a whole, optical signal to noise (SNR) suppression and degradation of BER value occur with increasing transmission distance beyond 650 m free space.

So, the evaluated performance of our proposed system proves the ability to mitigate RB noise by employing the dual band subcarrier modulation scheme and transmit information successfully over 55 km SMF plus 650 m FSO link for wireless users.

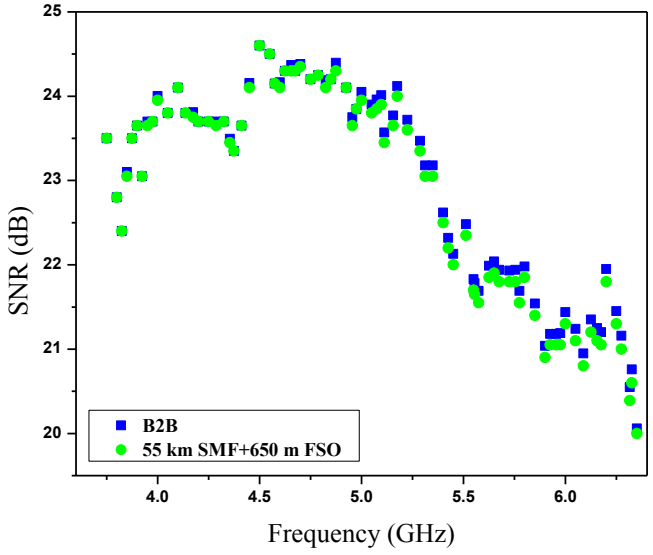


Fig. 8. SNR performance of OFDM subcarriers in the frequency band of 3.6–6.4 GHz by employing 16 QAM format for B2B and over 55 km SMF + 650 m FSO transmission.

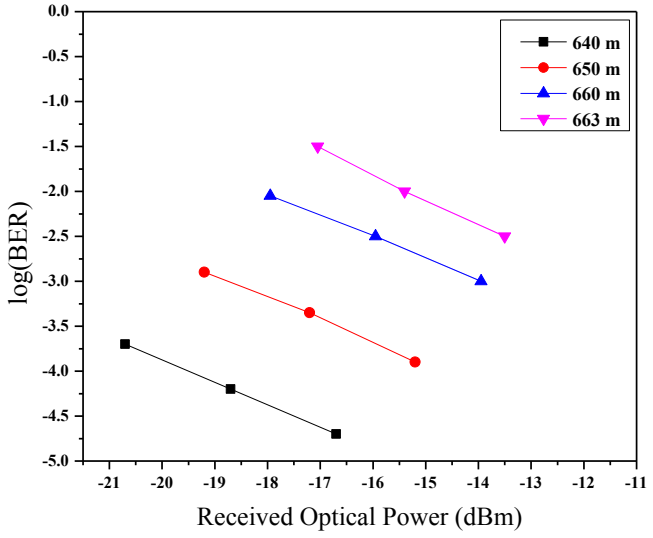


Fig. 9. BER curves measured of 10 Gbps OFDM signal for different FSO transmission distance (640 m, 650 m, 660 m, 663 m).

4. Conclusion

Enhancement of noise tolerance against RB for bidirectional 10 Gbps WDM network by employing dual band of OFDM signal for transportation of data over (30 + 25) km SMF via LE as well as 650 m FSO link utilizing doublet lens scheme is proposed and demonstrated successfully. Different band of subcarriers are modulated by OFDM signal for up and downlink channel to avoid RB noise. RSOA remodulated signal is also successfully transmitted through 55 km + 650 m FSO link for uplink transmission to maintain the bi-directionality. BER value under FEC limit i.e. 3.8×10^{-3}, SNR performance (>19 dB under FEC limit), <1 dB power penalty, clear constellation diagrams prove the feasibility of the proposed system. Power penalty is recorded ~ 9 dB without employing the dual band subcarrier modulation scheme. Proper EVM (<12.35%) value for 16 QAM signal is achieved that makes the network more reliable to the next generation communication system. Our proposed network is able to transmit less noisy information over 55 km SMF as well as 650 m free space. Such network makes itself special to the upcoming generation by not only mitigating the impairments due to RB noise but also communicates high-data rate over long haul SMF and above that it also provides information to the wireless users through FSO link.

References

Choudhury, P.K., 2015. In-band simultaneous transmission of baseband and broadcast signals in wavelength reused bidirectional passive optical network. Opt. Commun. 355, 296–300.
Wong, E., 2012. Next-generation broadband access networks and technologies. J. Lightwave Technol. 30 (4), 597–608.
Ni, C., Gan, C., Li, W., Chen, H., 2015. Bandwidth allocation based on priority and excess-bandwidth-utilized algorithm in WDM/TDM PON. Int. J. Electron. Commun. (AEÜ) 69 (11), 1659–1666.
Das, A.S., Patra, A.S., 2014. Simultaneous signal transmission of different data-rates in a DWDM system employing external injection locking technique. Opt. Laser Technol. 64, 23–27. https://doi.org/10.1016/j.optlastec.2014.04.011.
Straullu, S., Franco, G., Abrate, S., Forghieri, F., Ferrero, V., Gaudino, R., 2017. Symmetric 10 Gbps PON operating with a single laser over 31 dB of ODN Loss. IEEE Photonics Technol. Lett. 29 (12), 956–959.
Choudhury, P.K., Khan, T.Z., 2016. Symmetric 10 Gb/s wavelength reused bidirectional RSOA based WDM-PON with DPSK modulated downstream and OFDM modulated upstream signals. Opt. Commun. 372, 180–184.
Das, A.S., Patra, A.S., 2014. Bidirectional Transmission of 10 Gbit/s Using RSOA Based WDM-PON and Optical Carrier Suppression Scheme. J. Opt. Commun. 35, 239–243. https://doi.org/10.1515/joc-2013-0166.
Cano, I., Omela, M., Prat, J., Poggiolini, P. Colorless 10 Gb/s extended reach WDM PON with low BW RSOA using MLSE. OFC/NFOEC, San Diego, CA, USA, Paper OWG2 2010.
Parolari, P., Marazzi, L., Brunero, M., Martinelli, M., Brenot, R., Maho, A., Barbet, S., Gavioli, G., Simon, G., Saliou, F., Chanclou, P., 2014. 10-Gb/s Operation of a colorless self-seeded transmitter over more than 70 km of SSMF. J. Lightw. Technol. 26 (6), 599–602.
Lin, G.R., Wang, H.L., Lin, G.C., Huang, Y.H., Lin, Y.H., Cheng, T.K., 2009. Comparison on injection-locked Fabry-Perot laser diode with front-facet reflectivity of 1% and 30% for optical data transmission in WDM-PON system. J. Lightwave Technol. 27, 2779–2785.
Chow, C.W., Yeh, C.H., Wang, C.H., Shih, F.Y., Chi, S., 2008. Rayleigh backscattering performance of OFDM-QAM in carrier distributed passive optical networks. IEEE Photon Technol Lett 20 (22), 1848–1850.
Girault, G., Bramerie, L., Vaudel, O., Lobo, S., Besnard, P., Joindot, M., et al., 2008. 10 Gbit/s PON demonstration using a REAM-SOA in a bidirectional fiber configuration up to 25 km SMF. Proc. of ECOC P.6.08.
Li, X., Yu, J., Chang, G.-K., 2020. Photonics-aided millimeter-wave technologies for extreme mobile broadband communications in 5G. J. Lightwave Technol. 38 (2), 366–378.
Liaw, S.K., Hsu, K.Y., Yeh, J.G., Lin, Y.M., Yu, Y.L., 2017. Impacts of environmental factors to bi-directional 2x40 Gb/s WDM free-space optical communication. Opt. Commun. 396, 127–133.
Mallick, K., Mukherjee, R., Das, B., Mandal, G.C., Patra, A.S., 2018. Bidirectional hybrid OFDM based Wireless-over-fiber transport system using reflective semiconductor amplifier and polarization multiplexing technique. Int. J. Electron. Commun. (AEÜ) 96, 260–266.
Mallick, K., Mandal, P., Mukherjee, R., Mandal, G.C., Das, B., Patra, A.S., 2020. Generation of 40 GHz/80 GHz OFDM based MMW source and the OFDM-FSO transport system based on special fine tracking technology. Opt. Fiber Technol. 54, 102130.
Tsai, C.T., Chi, Y.C., Lin, G.R., 2015. Power fading mitigation of 40-Gbit/s 256-QAM OFDM carried by colorless laser diode under injection-locking. Opt. Express 23 (22), 29065. https://doi.org/10.1364/OE.23.029065.
Cvijetic, N., 2012. OFDM for next-generation optical access networks. J. Lightwave Technol. 30.
Kabir W. ORTHOGONAL FREQUENCY DIVISION MULTIPLEXING (OFDM). 2008 China-Japan Joint Microwave Conference, IEEE Xplore 2008:178–84. https://doi.org/10.1109/CJMW.2008.4772401.

Djordjevic, I.B., Vasic, B., 2006. Orthogonal frequency division multiplexing for high speed optical transmission. Opt. Express 14 (9), 3767. https://doi.org/10.1364/OE.14.003767.
Yeh, C.H., Chow, C.W., Chen, H.Y., Sung, J.Y., Liu, Y.L., 2012. Demonstration of using injection-locked Fabry-Perot laser diode for 10 Gbit/s 16-QAM OFDM WDMPON. Electron. Lett. 48, 940–942.
Acar, Y., Aldırmaz-Çolak, S., 2021. A new spectrally efficient pilot scheme for OFDM systems. Int. J. Electron. Commun. (AEÜ) 129, 153566. https://doi.org/10.1016/j.aeue.2020.153566.
Khalighi, M.A., Uysal, M., 2014. Survey on free space optical communication: a communication theory perspective. IEEE Commun Surveys Tuts 16 (4), 2231–2258.
Chaman-Motlagh, A., Ahmadi, V., Ghassemlooy, Z., 2010. A modified model of the atmospheric effects on theperformance of FSO links employing single and multiple receivers. J Modern. Opt. 57 (1), 37–42.
Chaudhary, S., Amphawan, A., Nisar, K., 2014. Realization of free space optics with OFDM under atmospheric turbulence. Optik 125 (18), 5196–5198. https://doi.org/10.1016/j.ijleo.2014.05.036.
Chaudhary, S., Tang, X., Wei, X., 2018. Comparison of Laguerre-Gaussian and Donut modes for MDM-WDM in OFDM-Ro-FSO transmission system. Int. J. Electron. Commun. (AEÜ) 93, 208–214. https://doi.org/10.1016/j.aeue.2018.06.024.
Upadhyay, K.K., Srivastava, S., Shukla, N.K., Chaudhary, S., 2019. High-Speed 120 Gbps AMI-WDM-PDM Free Space Optical Transmission System. J. Opt. Commun. 40, 429–433. https://doi.org/10.1515/joc-2017-0086.
Chaudhary, S., Lin, B., Tang, X., Wei, X., Zhou, Z., Lin, C., et al., 2018. 40 Gbps–80 GHz PSK-MDM based Ro-FSO transmission system. Opt. Quant. Electron. https://doi.org/10.1007/s11082-018-1592-z.
Zhang, H., Tang, X., Lin, B., Zhou, Z., Lin, C., Chaudhary, S., et al. Performance Analysis of FSO System with Different Modulation Schemes over Gamma-Gamma Turbulence Channel. 17th International Conference on Optical Communications and Networks (ICOCN2018) n.d.;11048. https://doi.org/10.1117/12.2519711.
Balaji, K.A., Prabu, K., 2018. BER analysis of relay assisted PSK with OFDM ROFSO system over malaga distribution including pointing errors under various weather conditions. Opt. Commun. 426, 187–193.
Wang, Y., Wang, D., Ma, J., 2016. Performance analysis of multihop coherent OFDM free-space optical communication systems. Opt. Commun. 376, 35–40.
Sharma, V., Sushank, 2014. High speed CO-OFDM-FSO transmission system. Optik 125 (6), 1761–1763.
Song, Y., Lu, W., Sun, B., Hong, Y., Qu, F., Han, J., et al., 2017. Experimental demonstration of MIMO-OFDM under water wireless optical communication. OptCommun 403, 205–210.
Sudheer, V.V., Mandloi, A., 2016. Enhanced coherent optical OFDM FSO link using diversity for different weather conditions. Information & Communication Technology (RTEICT).
Li, C.Y., Lu, H.H., Lu, T.C., Wu, C.J., Chu, C.A., Lin, H.H., et al., 2016. A 100 m/320 Gbps SDM FSO link with a doublet lens scheme. Laser Phys. Lett. 13 (7), 075201. https://doi.org/10.1088/1612-2011/13/7/075201.
Li, C.Y., Lu, H.H., Lin, C.Y., Chu, C.A., Chen, B.R., Lin, H.H., et al., 2016. Fiber-wireless and fiber-IVLLC convergences based on MZM-OEO-based BLS. IEEE Photon Technol J 8, 7902810.
Tsukamoto, K., Komaki, S., 2007. Radio on Fiber & Free Space Optics Systemsfor Broadband Wireless Access. ITU/BDT Regional Seminar on Broadband Wireless Access (BWA) for CIS. CEE and Baltic Countries, Moscow (Russian Federation 26–29.
Kang, J.-M., Han, S.-K., 2006. A novel hybrid WDM/SCM-PON sharing wavelength for up- and down-link using reflective semiconductor optical amplifier. IEEE Photonics Technol. Lett. 18, 502–504.
Oh, J.M., Koo, S.G., Lee, D., Park, S.-J., 2008. Enhancement of the performance of a reflective SOA-based hybrid WDM/TDM PON system with a remotely pumped erbium doped fiber amplifier. J. Lightwave Technol. 26 (1), 144–149.
Mandal, P., Mallick, K., Santra, S., Kuiri, B., Dutta, B., Patra, A.S. A bidirectional hybrid WDM-OFDM network for multiservice communication employing self-injection locked Qdash laser source based on elimination of Rayleigh backscattering noise technique. Optical and Quantum Electronics Volume 2021;53. https://doi.org/10.1007/s11082-021-02948-2.
Lin, S.C., Lee, S.L., Lin, H.H., Keiser, G., Ram, R.J., 2011. Cross-seeding schemes for WDM based next-generation optical access networks. J. Lightwave Technol. 29 (24), 3727–3736.
Yeh, C.H., Chow, C.W., Chen, H.Y., 2012. Simple colorless WDM-PON with Rayleigh backscattering noise circumvention employing m-QAM OFDM downstream and remodulated OOK upstream signals. J. Lightwave Technol. 30 (13), 2151–2155.
Chiuchiarelli, A., Proietti, R., Pres, M., Choudhury, P., Contestabile, G., Ciaramella, E., 2009. Symmetric 10 Gbit/s WDM-PON based on cross wavelength reuse to avoid Rayleigh backscattering and maximize band usage. Electron. Lett. 45, 1343–1345.
Mandal, P., Mallick, K., Dutta, B., Kuiri, B., Santra, S., Patra, A.S. Mitigation of Rayleigh backscattering in RoF-WDM-PON employing self coherent detection and bi-directional cross wavelength technique. Optical and Quantum Electronics 2021; 53. https://doi.org/10.1007/s11082-020-02720-y.
Mallick, K., Mandal, P., Mandal, G.C., Mukherjee, R., Das, B., Patra, A.S., 2019. Hybrid MMW-over fiber/OFDM-FSO transmission system based on doublet lens scheme and POLMUX technique. Opt. Fiber Technol. 52, 260–266.
Huang, X.-H., Li, C.-Y., Lu, H.-H., Chou, C.-R., Hsia, H.-M., Chen, Y.-H., 2020. A Bidirectional FSO Communication Employing Phase Modulation Scheme and Remotely Injection-Locked DFBLD. J. Lightwave Technol. 38 (21), 5883–5892.
Mehedy, L., Bakau, M., Nirmalathas, A., Skafidas, E., 2012. OFDM Versus Single Carrier Towards Spectrally Efficient 100 Gb/s Transmission With Direct Detection. J Opt. Commun. Netw. 4, 779–789.

Tapered multicore optical fiber probe for optogenetics

Bhagaban Sri Ramakrishna, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, maheswarinath1@outlook.com*

Abhishek Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, abhishekdas2256@gmail.com*

Manoranjan Sahoo, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, manoranjansahoo14@gmail.com*

Srimanta Mohapatra, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srimantamohaparta66@gmail.com*

ARTICLE INFO

Keywords:
Tapered multicore optical fiber
Optogenetics
Phased array antenna theory
Light delivery
Optical neural interface
Beam steering

ABSTRACT

Optical controlling and reading-out highly localized intracellular signaling events in network of neurons, single neurons and subcellular compartments is considered as the most groundbreaking innovation in the neuroscience field in recent years. New optical readout and manipulation tools for optogenetics are continuously required. Here we report on a tapered multicore optical fiber system able to achieve multipoint illumination and polychromatic lightening of a target with the resolution of single-cell and single organelles. The steering of the output beam can be achieved by controlling the wavelength and phase of the optical signals coupled to the fiber. Consequently, the change of the direction of the beam can be achieved without rotating the device, reducing the invasiveness of the final device. The possibility to steer the output optical beam can also allow to minimize the photoelectrical and photochemical effects caused by coupling the microelectrode and optical manipulation optoelectronics probes. Finally, our system can also collect light by exploiting multisite photometry approach.

1. Introduction

Nowadays, Optogenetics is placed at the center of the consideration realm of the neuroscience (Fenno et al., 2011; Bernstein and Boyden, 2011; Deubner et al., 2019). With this technology, scientists are able to investigate the neuronal system of the living organs from cell to system. Optogenetic tools comprise three main parts: gene delivery, light delivery and monitoring. (Knopfel, 2012). With the tremendous progress in the photonics field, new tools to boost these three steps have been developed. In particular, designing a tool able to illuminate and have control of this illumination remotely is a field that these days have become a very hot topic in the field of optogenetics (Kanneganti et al., 2011; Ronzitti et al., 2018). The important aspect in this illumination and light delivery process is that the sample or the animal in which the experiment is done should be compact and effective, bearing minimum damage.

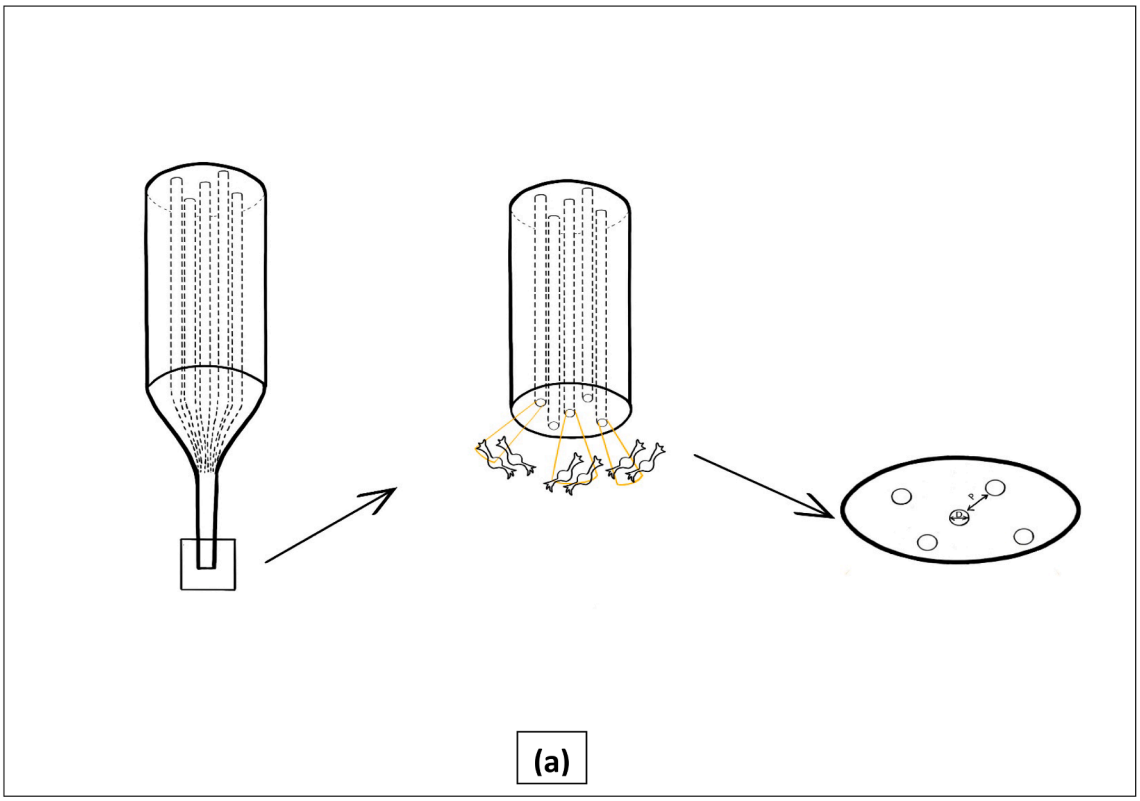
Optical fibers are currently used for light delivery and sometimes for photometry and fluorometry from the biological samples. The main advantages of using optical fibers is that they can image and illuminate the animals in the state of freely-behaving more deeply than with respect to the conventional imaging devices (Miyamoto and Murayama, 2016).

LeChasseur and co-workers showed that with the taper dual-core fiber they can illuminate and then electrically and optically measure the activity of neurons in the single-cell scale and resolution. (LeChasseur, et al., 2011). Pisanello et al., 2019 in a series of papers developed tapered fibers to illuminate multi spots of brain tissue and steer the beam. The same platform is also used for photometry from the fluorescent effect of genetically encoded neurons (F. Pisano, Depth-resolved fiber photometry with a single tapered optical fiber implant). Specifically, the device consists of tapered fiber (also coated by gold layer) where some optical windows were created by focused ion beam milling around the fiber tip (Pisanello, 2014). By playing on the geometry of the optical windows and the angle of the input light coupled to the fibers, they can illuminate the samples under test, site selectively. With these devices, they can scan the depth of the brain tissue, such as the cortex of the rat brain for illuminating the different layers of this sample (Pisanello, 2014; Pisanello et al., 2017; Pisano, 2018; Pisano et al., 2019). However, in order to steer the beam and illuminate the whole sample, it is necessary to move tapered part frequently with the risk to damage the biologic sample under study. Moreover, the manufacturing process, being based on focused ion beam milling, is quite time consuming and it is not suitable for high throughput production.

In this framework, here we report on an alternative device made of tapered multicore fiber (MCTF) and based on the theory of phased matching array antennas (Stark, 1974). By controlling the phase or wavelengths of the input light coupled in each core of the fiber we demonstrate that it is possible to dynamically steer the output beam

(from 0 to 180 degrees) by actively changing also the intensity, the beam width, without moving or changing the fiber position. Furthermore, in our case, the fiber probe can be easily fabricated also in high throughput fashion since tapering is well-assessed fabrication procedure that does not require complex nanomachining technologies.

With the aid of numerical simulations carried out with COMSOL Multiphysics, we study the dependence of the output beam characteristics such as angle, width and intensity as a function of the parameters



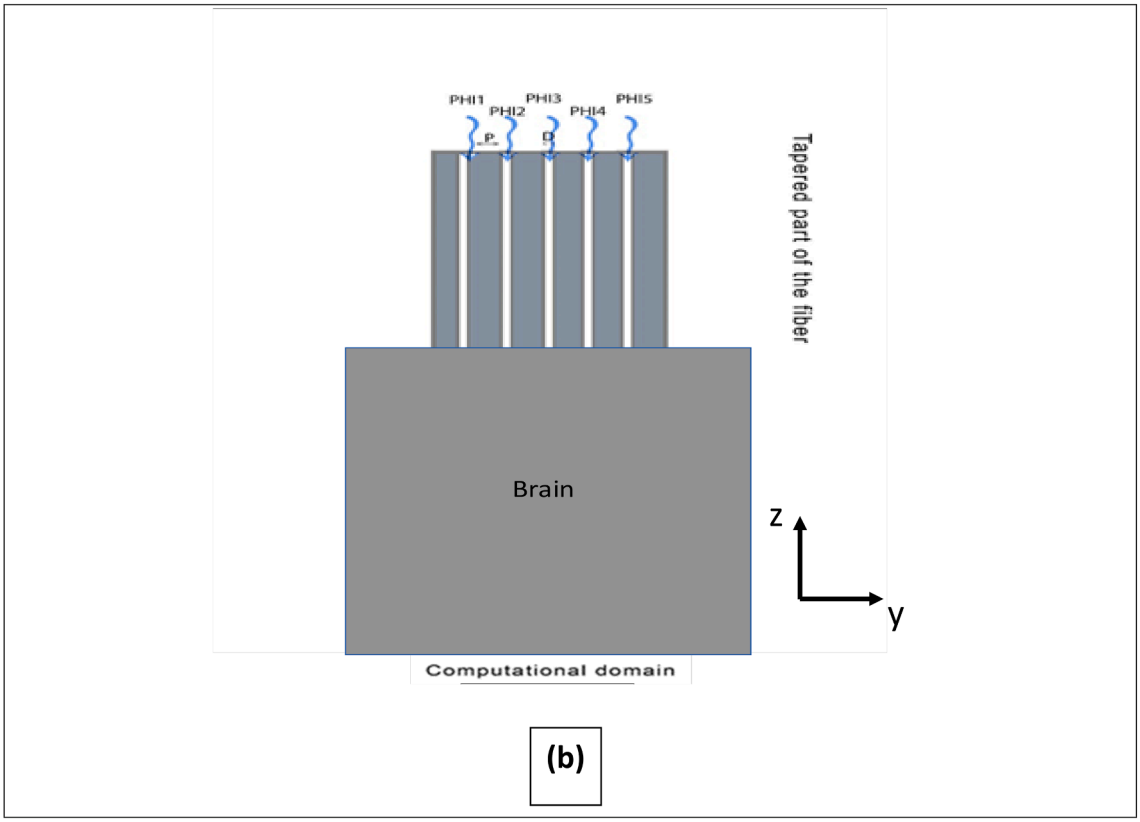


Fig. 1. (a) Schematic of the multicore tapered fiber with 5 cores; (b) 2D drawing (computational domain) used in our calculations (COMSOL Multiphysics).

of the input light coupled to the fiber cores. The steering the beam and the multi-site illumination can be made possible by exploiting all the degrees of freedom of our MCTF.

2. Results & discussion

We designed a multicore tapered fiber schematically shown in Fig. 1. Table 1 shows all the parameters considered in our simulations to tune the i) directivity ii) intensity iii) beam-width of the output optical beam coming out from the fiber tip.

Specifically, the geometrical parameters considered in our analysis are the distances between the adjacent cores (P), the diameter of the core (D) and the number of cores (N). In our simulations, only the final tapered part of the optical probe has been analyzed (the computational domain in our simulations is shown in Fig. 1b). From the experimental point of view, the geometrical parameters are determined and fixed once that the device is fabricated. It should be noted that tapering features in our study have been chosen on the basis of the dimensions of real common tapered fibers (Sakoda, 2005). In any case, it is important to underline that the fiber probe geometrical parameters can easily rescaled according to "scaling law", and the results here discussed remain consistent.

Indeed, there are several reasons for using the tapered fiber; first, fiber tapering allows us to create a miniaturized probe (overall diameter 20 μm) which avoids damaging biological tissue; secondly, thanks to the fiber tapering we can achieve a dimension of the probe that helps us to reach spatial beam widths of the order of around 5 μm that is comparable to the dimension of a single neuronal cell. Moreover, it is necessary to maintain distance between adjacent cores becomes of the order of 1 μm to create a significant interference effect.

The main tunable parameters of the optical system are thus the power, the wavelength and the phases of the light sources coupled to the fiber cores. In fact, by acting on these input parameters, and in particular on the last two, it is possible to "steer the beam" and form the "multisite illumination".

In our configuration, each core of the fiber can be seen as an optical antenna, so that the fiber tip forms a sort of antenna array. By exploiting basic physics of wave-optics, that is by changing the phase of the waves of these optical antennas, we observe the changing in the direction of the overall light beam according to the interfering and super positioning phenomena. (Visser, 2005).

Consistently with the theory of the phased array, we considered a constant phase difference (Phi) among the light sources coupled at each core; this means when moving from one core to the adjacent one, the phase of the light source increases by phi (for example in the case of 5 multicores tapered fiber, Phi (phase) in the ports will be: core 1: $Phi_1 = 0$, core 2: $Phi_2 = phi$, core 3: $Phi_3 = 2*phi$, core 4: $Phi_4 = 3*phi$, and core 5:$Phi_5 = 4*phi$). The refractive index of the brain which we used in our simulations is based on the scientific literature which is 1.38 A.U (Sun et al., 2012). Without loss of generality the refractive index of the fiber we considered is 1.8 and 1.5 for the cladding and the cores, respectively.

The size of the final part of the tapered multicore fiber considered in analysis is 15 μm. This size seems to feasible also from the experimental point of view (Chunxia et al., 2018; Tagoudi et al., 2016a, 2016b).

The calculations have been performed by using COMSOL Multiphysics (RF module). All of the simulations have been carried out in 2D geometry. Fig. 1 shows the drawing used in our simulations.

The first step is designing the device by choosing a suitable combination of the geometrical parameters. Then, we changed the input light parameters to study the effect that each parameter has on the characteristics of the output beam. Accordingly, in the following, we present several case studies and discuss the achieved results.

2.1. Case study 1: Effect of the fiber core diameters

We consider all the parameters as fixed quantities and change only the D parameter as described in table 2. We swept the D from 0.125 to 0.375 μm with a step of = 0.05 μm. The results are shown in Fig. 2. The intensity value is shown within the figure. As we see in the figure, just the intensity of the light and the beam width were changed. By passing from 0.125 to 0.375 μm, the intensity of output beam increases by a factor of 1.7, but the direction of the beam remains unaltered.

2.2. Case study 2: Effect of the distance between the fiber cores

As for the second case, we consider all the parameters fixed and we change only the P parameter, i.e. the distance between adjacent cores. In particular, we swept the P from 0.3 to 0.8 μm with a step of 0.1 μm as described in Table 3. The results are shown in Fig. 3. Analogously with the previous case, we observe that the intensity of the beam changes (there is also a slightly increase of the beam width). However, in this case, the side lobes changes from 35 degrees to 75 degrees moving from P = 0.3 to P = 0.8 μm.

2.3. Case study 3: Effect other parameters as fixed quantities and change only the f the phase difference and wavelength

After that, we changed the phi parameters from the 0 rad to 2*pi radians with the step of pi/3 Radian. The rest of parameters are fixed and the values are shown in table 4. The results are depicted in Fig. 4. When we change the parameter of Phi the beam direction changes from −75 degree to +75 degrees.

We repeated this calculation for different values of N. The results are summarized in Fig. 4. Note that in this case D = 0.1 μm (we chose this diameter because the results are clearer and more distinguishable). By increasing the number of cores, we increase the degree of complexity of our device. However, it possible to reach a fine tuning of the side lobes in terms of intensities and steering angles. In all of the cases the optical system can steer almost the angle between −75 degree to 75 degrees. The changing of the intensity of output beam between the system with 3, 5, 7 and 9 cores are 35, 55,100 and 130 respectively.

Also, we evaluated the effect of changing the wavelength of the light sources coupled to each core from 400 nm to 900 nm with a step of 100 nm. We would like to clarify that all the wavelengths considered in our study (400–800 nm) could be useful for optogenetics experiments (Prigge, et al., 2012; Oda, et al., 2018). The calculations are repeated for both Phi = 0 and phi = 2*pi/3 rad (we chose these two values because the results are clearer and distinguishable), while the remaining parameters are fixed (see table 4). The results are shown in Fig. 5 (a) and (b).

In the case of Phi = 0 rad, by changing the wavelengths of the input light sources coupled to the fiber cores, only the intensity of the output

Table 1
Parameters which are important for designing and optimizing our device.

Geometrical parameters of MCTF	N Number of cores	D Core diameter	P Distance between the cores
Input light parameters	P_{in} Power of incident light	Phi_{in} The phase difference between light coupled into adjacent cores	λ(lambda) The wavelength of light coupled to each core

Table 2
Input parameter for calculating results shown in Fig. 2.

Geometrical parameters of MCTF	N N = 5	D varies from 0.125 to 0.375 μm	P P = 0.3 μm
Input light parameters	P_{in} Power = 1 W	Phi_{in} Phi = 0 Radians	λ(lambda) λ = 600 nm

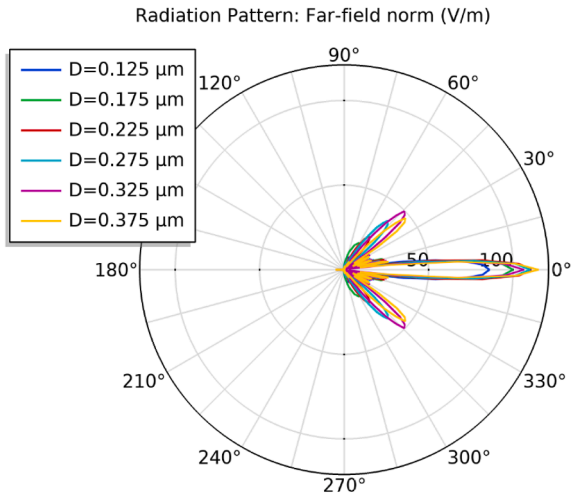


Fig. 2. Radiation pattern (far field) calculated at the output of the MCTF as a function of D.

Table 3
Input parameter for calculating results shown in Fig. 3.

Geometrical parameters of MCTF	N Number of cores = 5	D or d = 0.125 μm	Changing P from 0.3 to 0.8 μm
Input Light Parameter	P_{in} Power = 1 W	Phi_{in} Phi = 0 rad	λ(lambda) λ = 600 nm

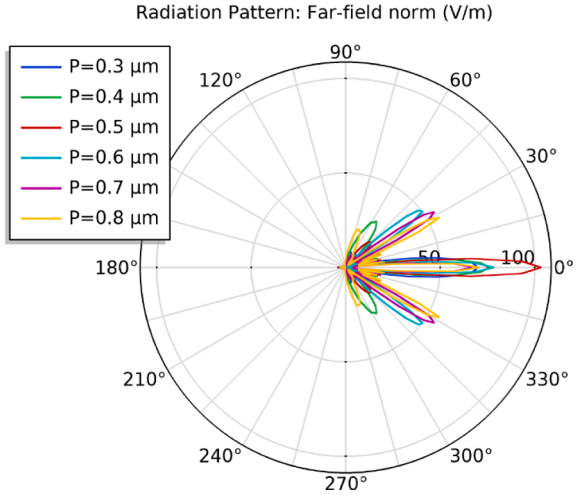


Fig. 3. Radiation pattern (far field) calculated at the output of the MCTF as a function of P.

Table 4
Input parameter for calculating results shown in Figs. 4 and 5.

Geometrical parameters of MCTF	N Number of cores = 5	D = 0.1 μm	P = 0.3 μm
LASER Parameters	P_{in} Power = 1 W	Phi_{in} Phi between 0 and 2*Pi Radian	λ(lambda) λ = 400 to 900 nm

beam changes drastically. Specifically, by increasing the wavelength, the intensity decreases and the angle of the output beam remains the same. In the case of Phi = 2*pi/3 rad, we can observe that the direction of the output light changes as a function of the wavelength. This effect can be used for achieving multicolor illumination that is very common in the optogenetics experiments (Bugaj and Lim, 2019; Han and Boyden, 2007; Packer et al., 2016). This technique consists of illuminating different part of the brain with different wavelengths to better control the cells (neurons) to inhibit and excite them at the same time. As a matter of fact, this functionality in our device can be achieved by coupling different wavelengths of the sources in different cores.

Finally, we plot the steering angle θ as a function of phi (Fig. 6 (a)) and λ (Fig. 6(b)). Clearly, the value of θ does not depend on the geometrical parameters that only affects the intensity and the width of the beam. For the fitting process we used MATLAB curve fitting toolbox (As for clarifying better we also compared three different values of λs and Phis in one plot to show better the dependency of the independent and dependent values to each other).

Overall, the proposed configuration offers many degrees of freedom to be exploited in order to reach a high level of functionality. The degrees of freedom are both geometrical (number of cores, diameters and distances among different cores defined at the fabrication phase) but also physical (intensity, wavelength and phased of the light signals coupled to the different cores, chosen during the testing phase). By acting on these parameters, it could be possible to steer the beam at different wavelengths in an efficient way. Beam steering efficiencies are estimated to be, on average, about 14% and 24% for steering angles larger and smaller than 35° respectively. In principle, with our device it could be possible to control and stimulate different areas of the brain (different neurons or different part of the same neuron) with different colors (wavelengths). This specific characteristic of our device is very useful in optogenetics and could help to perform experiments in a novel intriguing manner (Vierock et al., 2020; (Klapoetke, et al., 2014; Yizhar, et al., 2011; Akerboom et al., 2013; Erbguth et al., 2012).

3. Remarks on the beam width

Concerning the spatial resolution, values of beam width up to 5 μm could be achieved with our probe. Fig. 6 shows the energy density (in the far field) as a function of the spatial coordinate (y) for the combination of the parameters reported in the Table 5. This beam resolution value is comparable with the size of single neuron cell, and in particular with the soma of the cells that has overall the size between 5 and 10 μm in the cortex of the brain.

4. Conclusions

In conclusion, we have demonstrated that with this new device we can steer the beam just by changing the phase of the sources which are coupled with the core of the multi-core fibers without moving the place of the fiber tip. In fact, we can steer the output beam from −75 to 75 degrees by simply changing the phases of the optical sources coupled to the different fiber cores. Moreover, we can change the intensity of the beam by changing the different parameters of the optical system, especially, the core diameters D, the distances between cores P, and the number of the cores N and power of the incident light P_{in}; besides, we are also able to tune the beam-width (to change the illumination spot size).

The geometrical parameters are useful *ab initio,* during the probe design, to change both the spatial resolution (beam width) and the penetration depth. Specifically, probe with larger/smaller diameters can be used if we want to achieve larger/smaller penetration depths.

All in all, with our device we have more control in the process of the light delivery. Our device allows, in principle, to achieve multicolor illumination of neurons, providing more degree of freedom to control the brain during illumination without rotating the probe. In fact, since different wavelength do not interfere with each other (Fowles, 1989), this could help us to couple different wavelengths in different cores with different phases and consequently one will be able to achieve the multipoint illumination with different wavelengths. This could help the experimenters who are working on the Bidirectional Pair of Opsins for

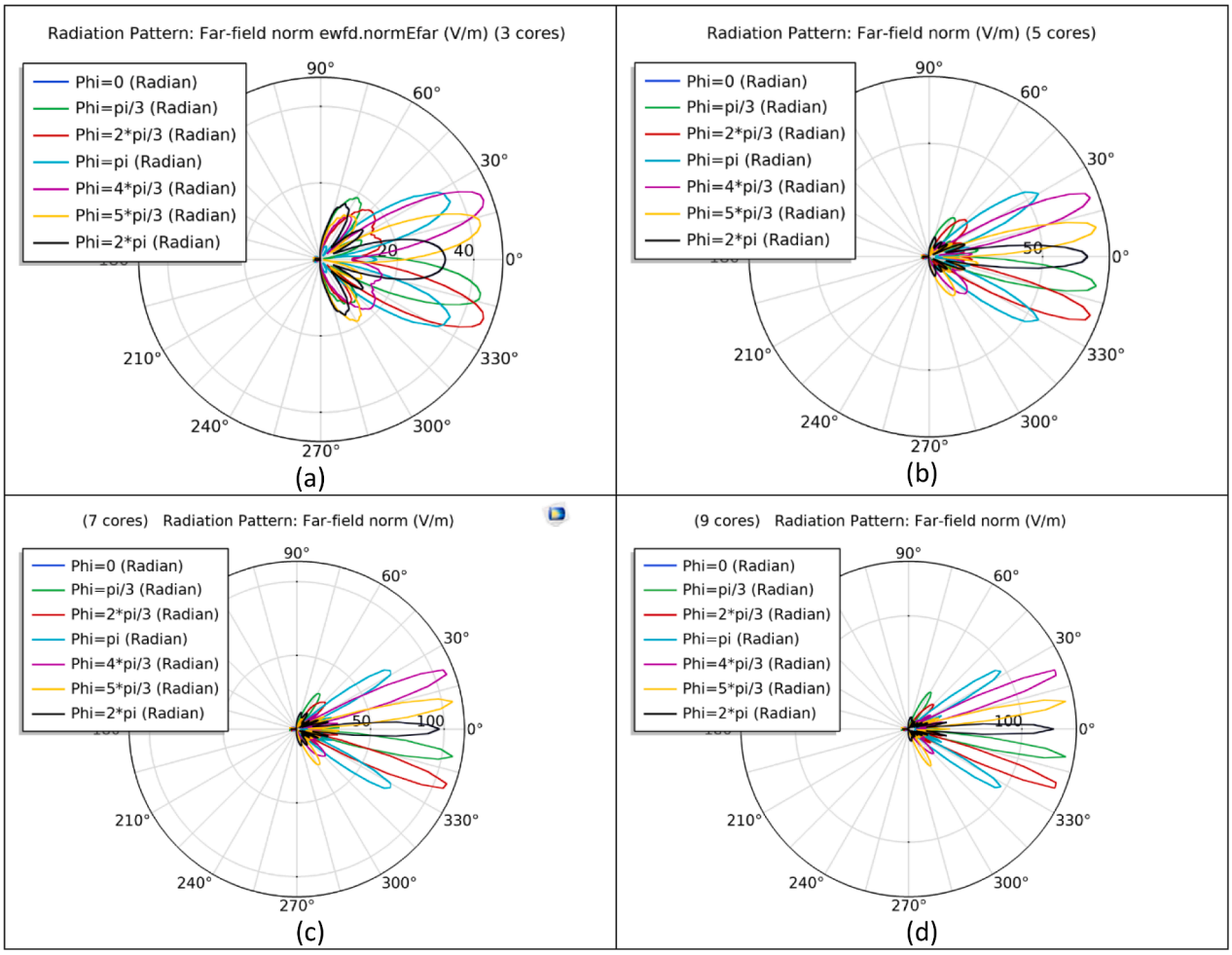


Fig. 4. a) Radiation pattern (far field) calculated at the output of the MCTF with 3 cores as a function of Phi incidence (b) Radiation pattern (far field) calculated at the output of the MCTF with 5 cores as a function of Phi incidence (c) Radiation pattern (far field) calculated at the output of the MCTF with 7 cores as a function of Phi incidence (d) Radiation pattern (far field) calculated at the output of the MCTF with 9 cores as a function of Phi incidence.

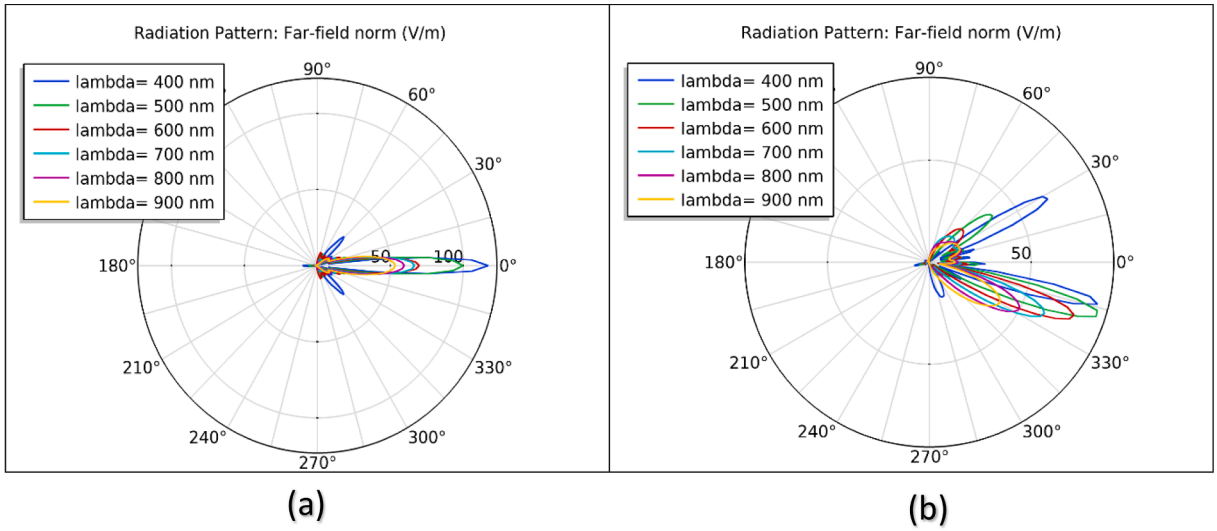
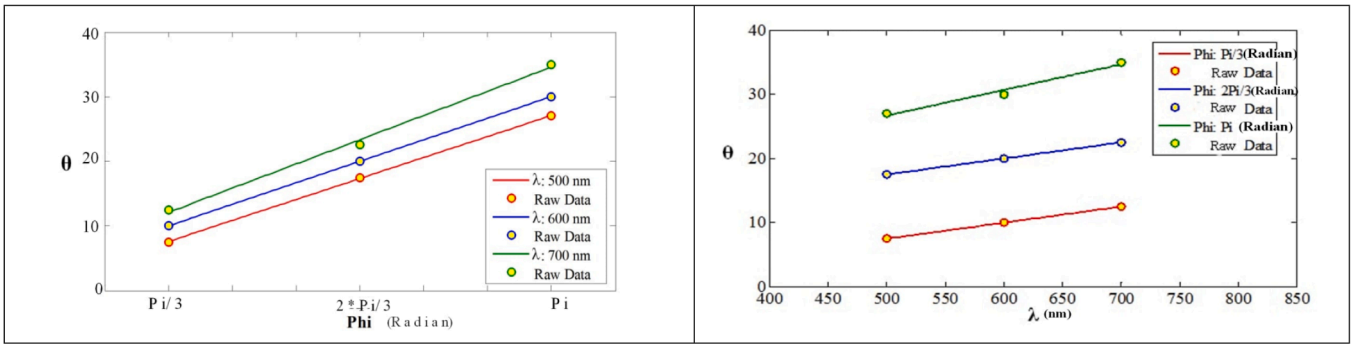


Fig. 5. a) Radiation pattern (far field) calculated at the output of the MCTF as a function of wavelength of incident light (With Phi = 0 rad) b) Radiation pattern (far field) calculated at the output of the MCTF as a function of wavelength of incident light (With Phi = 2*pi/3 Radian).



(a) (b)

Fig. 6. a) The plot of θ against the parameter of Phi; the R-squared for the fitting (coefficient of determination) is 0.98 for the green line(λ:500 nm) and is 1 for the blue(λ:600 nm) and red(λ:700 nm) line: b) The plot of θ against the Parameter λ of incident (wavelength of incident light); the parameter of R-squared (coefficient of determination) is 1 for fitting curve green line (Phi: pi rad.) and is 0.98 for blue line(Phi: 2*pi/3 rad.) and red line(Phi: pi/3 rad.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

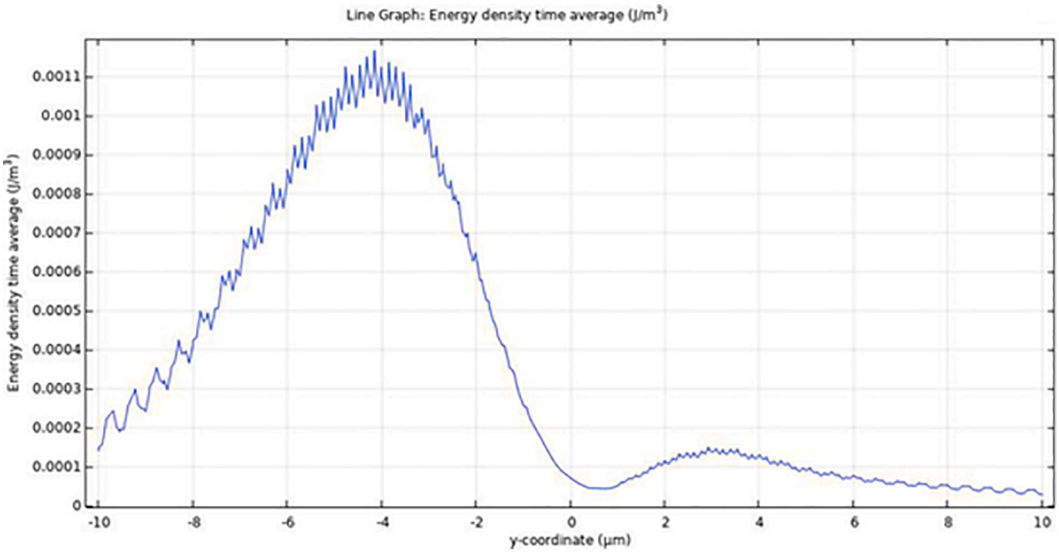


Fig. 7. The wave front of the output-beam in front of the fiber tip.

Table 5
Input parameter for calculating results shown in Fig. 7.

Geometrical parameters of MCTF	N Number of cores = 5	D or d = 0.1 μm	P = 0.3 μm
Input Light Parameter	P_{in} Power = 1 W	Phi_{in} Phi = pi/3 rad	λ(lambda) λ = 600 nm

Light-induced Excitation and Silencing (BiPOLE) (Vierock et al., 2020) For the future studies, thanks to the exploitation of different kinds of optimizing toolboxes like artificial neural networks, it could be possible to set the parameters so that the optical system be able to target every biological organ in specific places in front of the multicore tapered fiber.

References

Akerboom, J., Carreras Calderón, N., Tian, L., Wabnig, S., Prigge, M., Tolö, J., L. Looger, L. (2013). Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics. Front. Mol. Neurosci.

Bernstein, J.G., Boyden, S.E., 2011. Optogenetic tools for analyzing the neural circuits of behavior. Cell Press-Trends Cognitive Sci.

Bugaj, L.J., Lim, W.A., 2019. High-throughput multicolor optogenetics in microwell plates. Nat. Protoc. 14 (7), 2205–2228.

Chunxia, Y., Hui, D., Wei, D., Chaowei, X., 2018. Weakly-coupled multicore optical fiber taper-based high-temperature sensor. Phys. Sensors Actuators A.

Deubner, J., Coulon, P., Diester, I., 2019. Optogenetic approaches to study the mammalian brain. ScienceDirect 57, 157–163.

Erbguth, K., Prigge, M., Schneider, F., Hegemann, P., Gottschalk, A., Nitabach, M.N., 2012. Bimodal activation of different neuron classes with the spectrally red-shifted channelrhodopsin chimera C1V1 in Caenorhabditis elegans. PLoS ONE 7 (10).

Pisanello, F., 2014. Multipoint-emitting optical fibers for spatially addressable in vivo optogenetics. Neuron.

Pisanello, F., Mandelbaum, G., Pisanello, M., Oldenburg, I.A., Sileo, L., Markowitz, J.E., Peterson, R.E., Della Patria, A., Haynes, T.M., Emara, M.S., Spagnolo, B., Datta, S.R., De Vittorio, M., Sabatini, B.L., 2017. Dynamic illumination of spatially restricted or large brain volumes via a single tapered optical fiber. Nat. Neurosci. 20 (8), 1180–1188.

Pisano, F., 2018. Focused ion beam nanomachining of tapered optical fibers for patterned light delivery. Microelectr. Eng.

Pisano, F., Pisanello, M., Lee, S.J., Lee, J., Maglie, E., Balena, A., Sileo, L., Spagnolo, B., Bianco, M., Hyun, M., De Vittorio, M., Sabatini, B.L., Pisanello, F., 2019. Depth-resolved fiber photometry with a single tapered optical fiber implant. Nat. Methods 16 (11), 1185–1192.

Fenno, L., Yizhar, O., Deisseroth, K., 2011. The development and application of optogenetics. PubMed.
Fowles, G. R. (1989). Coherence and interfernce. In G. R. Fowles, Introduction to modern optics. Dover.
Han, X., Boyden, E.S., Rustichini, A., 2007. Multiple-color optical activation, silencing, and desynchronization of neural activity, with single-spike temporal resolution. PLoS ONE 2 (3), e299.
Kanneganti, A., Shivalingaiah, S., Gu, ling, Alexandrakis, G., Mohanty, S., 2011. Deep brain optogenetic stimulation using bessel beam. Biophys. J . 100 (3), 95a. https://doi.org/10.1016/j.bpj.2010.12.725.
Klapoetke, N.C., Murata, Y., Soo Kim, S., Pulver, R.S., Birdsey-Benson, A., Ku Cho, Y., Boyden, S.E., 2014. Independent optical excitation of distinct neural populations. Nature Methhod.
Knopfel, T. (2012). In E. Boyden, Optegenetics.
Miyamoto, D., Murayama, M., 2016. The fiber-optic imaging and manipulation of neural activity during animal behavior. Neuroscience Res.
Oda, K., Vierock, J., Oishi, S., Rodriguez-Rozada, S., Taniguchi, R., Yamashita, K., Nureki, O., 2018. Crystal structure of the red light-activated channelrhodopsin Chrimson. Nat. Commun.
Packer, A.M., Roska, B., Häusser, M., 2016. Targeting neurons and photons for optogenetics. Nat. Neurosci.
Prigge, M., Schneider, F., Tsunoda, S.P., Shilyansky, C., Wietek, J., Deisseroth, K., Hegemann, P., 2012. Color-tuned channelrhodopsins for multiwavelength optogenetics. J. Biol. Chem. 287 (38), 31804–31812.
Ronzitti, E., Emiliani, V., Papagiakoumou, E., 2018. Methods for three-dimensional All-optical manipulation of neural circuits. Front. Cell. Neurosci.
Sakoda, K. (2005). Scaling law and time reversal symmetry. In K. Sakoda, Optical Properties of Photonic crystals. Springer Science & Business Media.
Stark, L., 1974. Microwave theory of phased-array antennas—a review. IEEE.
Sun, J., Lee, S.J., Wu, L., Sarntinoranont, M., Xie, H., 2012. Refractive index measurement of acute rat brain tissue slices using optical coherence tomography by. Opt. Express 20 (2), 1084. https://doi.org/10.1364/OE.20.001084.
Tagoudi, E., Milenko, K., Pissadakis, S., 2016a. Intercore coupling effects in multicore optical fiber tapers using magnetic fluid out-claddings. IEEE. 34 (23), 5561–5565.
Tagoudi, E., Milenko, K., Pissadakisa, S., 2016b. Power Coupling in Multicore Optical Fiber Tapers Utilizing Out-Cladding Ferrofluids. Photo-Optical Instrumentation Engineers (SPIE).
Vierock, J., Rodriguez-Rozada, S., Pieper, F., Dieter, A., Bergs, A., Zeitzschel, N., Wiegert, J., 2020. BiPOLES: a tool for bidirectional dual-color optogenetic control of neurons. BioRxiv.
Visser, H.J., 2005. Array and Phased Array Antenna Basics. John Wiley & Sons Ltd, West Sussex.
Yizhar, O., Fenno, L.E., Prigge, M., Schneider, F., Davidson, T.J., O'Shea, D.J., Sohal, V. S., Goshen, I., Finkelstein, J., Paz, J.T., Stehfest, K., Fudim, R., Ramakrishnan, C., Huguenard, J.R., Hegemann, P., Deisseroth, K., 2011. Neocortical excitation/inhibition balance in information processing and social dysfunction. Nature 477 (7363), 171–178.
LeChasseur, Y., Dufour, S., Lavertu, G., Bories, C., Deschênes, M., Vallée, Réal, De Koninck, Y., 2011. A microprobe for parallel optical and electrical recordings from single neurons in vivo. Nat. Methods 8 (4), 319–325.

A proposal of depth of focus equation for an optical system combined a digital image sensor

Sudhansu Sekhar Khuntia, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sskhuntia88@gmail.com*

Major Das, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, major.das556@gmail.com*

Swaha Pattnaik, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, swaha.pattanayak@gmail.com*

Smruti Samantray, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, smrutisamantray23@hotmail.com*

ARTICLE INFO

Keywords:
Depth of focus
Digital image sensor
Microscope

ABSTRACT

When an optical system and a digital image sensor are combined, it is known that pixel size of the sensor affects digital images. In particular, the DOF was not represented in a simple formula due to the influence of pixel size. To consider the effect of pixel size on digital images, we set the ratio, f_C/f_N, between the cut-off frequency of the optical system, f_C, and the Nyquist frequency of the imaging device, f_N, as the parameter expressing the effect of the pixel size. By using the parameter, we derived that the DOF can be expressed as a simple formula. For confirming the DOF formula, we measured the DOF of a microscope combined a digital image sensor. The values calculated from the DOF formula that we derived were consistent with the experimentally measured results.

We propose our formula expressing the DOF when an optical system and a digital image sensor are combined

1. Introduction

The resolution and depth of focus (DOF) of optical equipment with optical performance at the diffraction limit are $0.61\lambda/NA$ and $n\lambda/NA^2$, respectively, as determined by the Rayleigh criterion.

However, the Rayleigh criterion is expressed using changes in the point spread function (PSF) on the image plane and characteristics (element size, density, thickness) of the medium that receives the optical images are not reflected.

The resolution and the DOF are known to deviate from the values obtained using the Rayleigh criterion depending on the characteristics of the light-receiving medium.

For example, in semiconductor lithography equipment that projects circuit patterns onto thick photoresists, the resolution ε and the DOF Δ_L are different from the Rayleigh values because of effects of thickness and the developing process of the photoresist; two parameters considering the effects, k_1 and k_2, are used in their formulations: (Levinson, 2005; Mack, 1988)

$$\begin{aligned} \varepsilon &= k_1 \cdot n\lambda/NA \\ \Delta_L &= \pm k_2 \cdot n\lambda/NA^2 \end{aligned} \tag{1}$$

Another example, when visually observing using a microscope, the DOF is known to depend on a combination of objective and ocular lenses and deviates from the Rayleigh values $n\lambda/NA^2$. The reason for this deviation is the resolution of the eye. Berek showed, that when the resolution of the eye is considered, the DOF is

$$\Delta_B = \left\{ \frac{n}{NA^2}\left(\frac{\lambda}{2}\right) + \frac{n \bullet \omega_B}{\beta \bullet NA} L \right\} \tag{2}$$

Here, ω_B is the resolution of the eye, as defined by Berek (5 min), β is the total magnification of the objective and ocular lenses, and L is the distance of distinct vision (Martin, 1966; Murphy and Davidson, 2013).

It can be thought that the Berek's DOF equation (Eq. (2)) was formed in consideration of the size and density of photoreceptors in the retina.

Using a digital image sensor, such as a CCD (charge couple device) or CMOS (complementary metal oxide semiconductor) that pixel size cannot ignore comparing with diameter of the PSF, in conjunction with an optical system results in the resolution and the DOF values that depend on the pixel size of the digital image sensor; thus, they differ from the Rayleigh values.

When the resolution of the monitor displaying the image is higher than that of the image sensor, the resolution is limited by the worse constraint of the cut-off frequency of the optical system on the image plane, f_C, and the Nyquist frequency of the imaging device, f_N (ISO, 2016).

However, the DOF is not formulated in a simple form, as in Eq. (1) and Eq. (2); thus, empirical methods and numerical calculations are used to obtain the DOF.

One empirical method is a geometrical optics method that geometrically calculates the amount of defocusing in which the diameter of a light flux incident on an image sensor becomes smaller than the pixel size multiplied by an empirically determined coefficient. A problem with this geometrical optics method is how to obtain the empirical coefficient.

Yamamoto (2014) proposed a method to numerically calculate the relation between the pixel size and the defocus wave front coefficient when the reduction in total amount of light is within a given tolerance. This method alleviated the problem in the geometric optical methods, and shows that the DOF converges to the Rayleigh criterion when the pixel size is small enough for the diameter of the PSF.

Another numerical calculation method is to obtain the modulation transfer function (MTF) when the optical system is defocused; an amount of defocus is derived where the decrease in MTF is less than a predetermined tolerance.

These two methods require the determination of an empirical coefficient and numerical calculations to obtain the MTF or intensity distribution on the image surface. Therefore, a simple formulation that can adapt to changes in the optical system or the image sensor with convenient calculations, as in the Berek's DOF formula, is called for.

It is well known that the spatial frequency in the digital image is expressed using pixel numbers, and the unit is different from the spatial frequency used in the optical system.

By using f_C/f_N as the conversion parameter between the spatial frequency of the optical system and the spatial frequency of the digital image, we showed that the resolution and the DOF in the digital image are represented with simple equation in plural analysis model, and we derived that the DOF can be expressed with an simple formula similar to the Berek's DOF formula when a digital image sensor is combined with an optical system.

Comparing the value calculated by the DOF formula and the experimental value from real DOF measurements, results showed the values were consistent with each other.

Therefore, we propose that our formulated DOF equation be used to obtain the DOF when an image sensor is combined with an optical system.

2. Theory

The coordinates in the digital image represent the pixel number as units and are different from the coordinate units used in the optical image. Therefore, when representing the spatial frequency, a conversion parameter between the digital image and the optical image was required.

Focusing on this conversion parameter, we considered the effect of pixel size on the DOF in four analytical models shown below. After that, a formula was derived to obtain the DOF when a black-and-white image sensor is combined with an optical system, considering the effect of the pixel size.

2.1. An analogy from Berek's DOF equation

By replacing the distribution of the photoreceptors on the retina with an array of digital image sensors, we consider the resolution and depth of focus of visual observation of a microscope.

Assuming the Nyquist frequency of the eye, $f^{eye}{}_N$, Eq. (2) is rewritten using the cut-off frequency of the optical system, f_C, as

$$\Delta_B = \frac{n\bullet\lambda}{2NA^2}\bullet\left\{1+\frac{2NA\omega_B}{\beta\lambda}L\right\} = \frac{n\bullet\lambda}{2NA^2}\bullet\left\{1+\frac{\omega_B L}{\beta}f_C\right\} = \frac{n\bullet\lambda}{2NA^2}\bullet\left\{1+\frac{f_C}{f_N^{eye}}\right\} \quad (3)$$

The ratio, $f_C/f^{eye}{}_N$, between the cut-off frequency of the optical system, f_C, and the Nyquist frequency of the eye, $f^{eye}{}_N$, can be regarded as a parameter of how the resolution of the eye affects the DOF.

From Eq. (3) when the conversion parameter is $f_C/f^{eye}{}_N \le 1$, the resolution and the DOF of visual observation of a microscope are obtained from the Rayleigh criterion.

When $1 < f_C/f^{eye}{}_N$, the resolution is determined by the eye, and the DOF is different from the Rayleigh criterion.

2.2. A formulation of the DOF equation by geometric optics

The DOF of a digital camera can be obtained using geometrical optics, as shown in Fig. 1. Here, a light flux of uniform intensity is incident on the image sensor.

The diameter of the circle of confusion, C, is less than m-times the pixel size at the focus point. The value 'm' is typically obtained empirically. If the pixel size is large compared to the diameter of the PSF formed on the imaging surface, the DOF is expressed – using a geometrical optics method – as the range where the PSF forms a circle of confusion with uniform light intensity by defocusing.

When f_C and f_N are equal (the conversion parameter $f_C/f_N = 1$), most of the PSF intensity is within an area of 4×4 pixels.

The diameter of the circle of confusion, C, and the pixel size are not compared simply by the lengths; instead, the DOF is calculated using the change in total amount of light in 4×4 pixels.

Assuming a constant light intensity in the circle of confusion, the total amount of light incident on the image sensor can be described by the area, CS, of the circle of confusion, C, through the equation,

$$CS = \pi r^2 = \pi(\ \delta_G tan\theta)^2 \approx \pi(\ \delta_G sin\theta)^2$$

Denoting the pixel size as PP and the tolerance of the DOF as γ, the following relation holds between the area of the circle of confusion, C, that increases by defocusing, and the area of the 4×4 pixels:

$$(4PP)^2 = \gamma\ CS\ \approx\ \gamma\ \pi\ (\delta_G sin\theta)^2$$

The DOF Δ_{GO} is given by

$$\Delta_{GO} = 2\delta_G = \frac{4}{\sqrt{\gamma\pi}}\frac{PP}{sin\theta} = \frac{4}{\sqrt{\gamma\pi}}\frac{nPP}{NA} = \frac{2}{\sqrt{\gamma\pi}}\frac{n\lambda}{NA^2}\frac{f_C}{f_N}$$

Taking the tolerance as $\gamma = 0.81$, which is the same as the Rayleigh criterion, the DOF Δ_{GO} becomes

$$\Delta_{GO} = 1.25\frac{n\lambda}{NA^2}\frac{f_C}{f_N} \quad (4)$$

Therefore, the effect of pixel size on the DOF is expressed using the conversion parameter f_C/f_N.

The light intensity in the PSF becomes a light flux that is almost uniform when the wave front coefficient is larger than $\pm 3\lambda/4$ at the defocused wave front (Wyant and Creath, 1992); thus the DOF is, from Eq. (4),

$$3\bullet\frac{n\lambda}{NA^2} < \Delta_{GO} \quad (5)$$

The pixel size is within the range

$$2.4 < f_C/f_N \quad (6)$$

from Eqs. (4) and (5)

However, the intensity in the light flux is not uniform when the defocus wave front coefficient is less than $\pm\lambda/2$, meaning the intensity distribution approaches a PSF instead (Wyant and Creath, 1992).

In order to convert the distribution of the PSF to the circle of confusion C within the defocus range, we assumed that 90 percent light flux in the CS from the average value of encircled energy in the diameter λ/NA enter into 4x4 pixels. Thus, in this range Eq. (4) can be rewritten as

$$\Delta_{GO} = 0.9\bullet 1.25\frac{n\lambda}{NA^2}\bullet\frac{f_C}{f_N} = 1.12\frac{n\lambda}{NA^2}\bullet\frac{f_C}{f_N} \quad (7)$$

The DOF is given by Eq. (7), according to geometrical optics, when

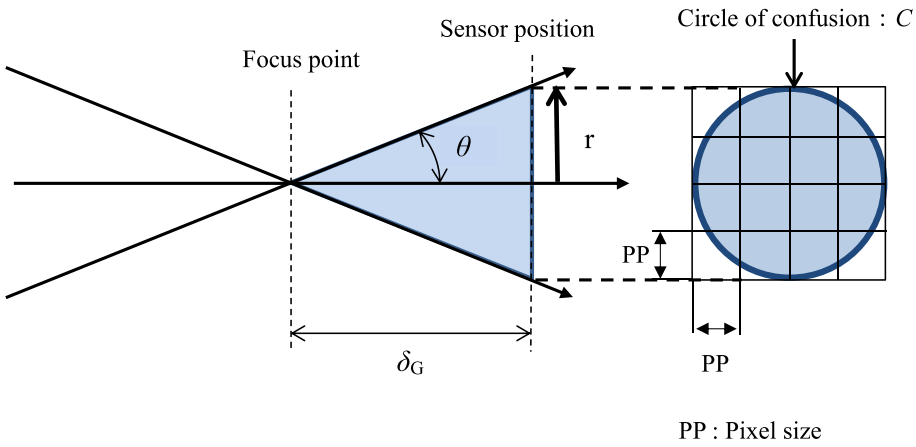


Fig. 1. Schematic of obtaining the DOF by geometrical optics.

f_C/f_N is smaller than 2.4.

2.3. Formulations of DOF equation by numerical calculation methods

2.3.1. A formulation of DOF equation from a method calculating light intensity in pixels within a PSF (Evolving Yamamoto's method (Yamamoto, 2014)

In the geometrical optics method, the DOF was obtained by making the light flux on the imaging surface form a circle of confusion with uniform intensity distribution and then comparing the areas of the circle of confusion and the number of pixels. However, the error when considering the pixel size is large because the intensity of the PSF near the focal point is approximated from the encircled energy calculation.

Yamamoto's method (Yamamoto, 2014) should be more precise because the change in the PSF from defocusing is numerically calculated.

We show here that the DOF obtained by the method can be replaced by a simple formula, using the conversion parameter f_C/f_N as a parameter to express the effect of the pixel size on the DOF.

The method numerically calculates the wave front coefficient of the defocus corresponding to the DOF by setting a receiving region of 2×2 pixels in the PSF.

The wave front coefficient of the defocusing from setting a receiving region of 4×4 pixels in the PSF, as in the geometrical optics method, is calculated using the method in Fig. 2 (a).

Fig. 2a can be rewritten as Fig. 2 (b) using the following information: the amount of defocusing is $\lambda/2NA^2$ when the wave front coefficient of the defocus is $\lambda/4$; the wave front coefficient of the defocusing is proportional to the amount of defocusing; the wave front coefficient of the defocus is given as a coefficient of λ/NA^2 when the DOF is expressed with its full width; and the pixel size is normalized with f_C/f_N instead of f_N/f_C.

When the calculated result in Fig. 2(b) is approximated with a linear function, the DOF calculated with the method, Δ_{YM}, is approximated as

$$\Delta_{YM} = \frac{n \bullet \lambda}{2NA^2} \bullet \left\{ 1.0 + b \bullet \frac{f_C}{f_N} \right\} \tag{8}$$

$1.0 \leqq b \leqq 1.25$

The method claims that the DOF converges to $n\lambda/NA^2$ in the range $f_C/f_N \leq 1$; however, if the pixel size becomes small, including that of the monitor, the DOF will be determined by the resolution of the eye that looks at the monitor.

This effect means the DOF would be expressed with Berek's DOF equation (Eq. (3)) when the pixel size becomes smaller.

2.3.2. A formulation of DOF equation from calculating MTF at defocus points

Denoting the intensity distribution of an object as I_0 (x,y), the point spread function of the optical system as *PSF* (λ, *NA*), the pixel function expressing a pixel (PP) as *REC* (PP), and the function representing the periodicity of pixels as *Calm*(PP), the intensity distribution of an image recorded by an imaging device is given as

$$I_S(x,y) = I_0(x,y) \otimes PSF(\lambda, NA) \otimes REC(PP) \otimes Calm(PP) \tag{9}$$

Here, $\otimes$ represents convolution.

Writing the Fourier transformation of *PSF* as *OTF* (f_C), the Fourier transformation of the function expressing a pixel (PP), *REC* (PP), as *SINC* (f_S), and the Fourier transformation function as $F\{\ \cdot\ \}$, the Fourier transformation of Eq. (8) results in

$$F\{I_S(x,y)\} = F\{I_0(x,y)\} \bullet \mathrm{OTF}(f_C) \bullet \mathrm{SINC}(f_S) \bullet \mathrm{Calm}(f_S) \tag{10}$$

The effect of the pixel size (PP) is expressed using *SINC* (f_S). Here, $f_S = 1/PP$ is the sampling frequency of the imaging device, and $f_S = 2\,f_N$.

MTF (f_C) and *SINC* (f_S) may be considered as MTFs that consider the pixel size; thus, the DOF can be obtained from changes with defocusing.

MTF (f_C) and *SINC* (f_S) were calculated with different pixel sizes, assuming that the defocused wave surface manifests in the pupil as an ideal optical system with optical magnitude of 10 and NA of incident light of 0.3, a light source wavelength, λ, of 550 nm, and the pixels of the image sensor are arranged two-dimensionally as 600×600 pixels squares.

The coefficients of the defocused wave front were obtained when the rate of change of *MTF* (f_C) and *SINC* (f_S) at $f_N/2$ is within a tolerance. The tolerance was the rate of change of *MTF* (f_C) at $f_N/2$ when a defocused wave plane of $\lambda/4$ was applied to the pupil. Fig. 3 shows the DOF calculated through a similar change of variables, as in the results in Fig. 2 (b).

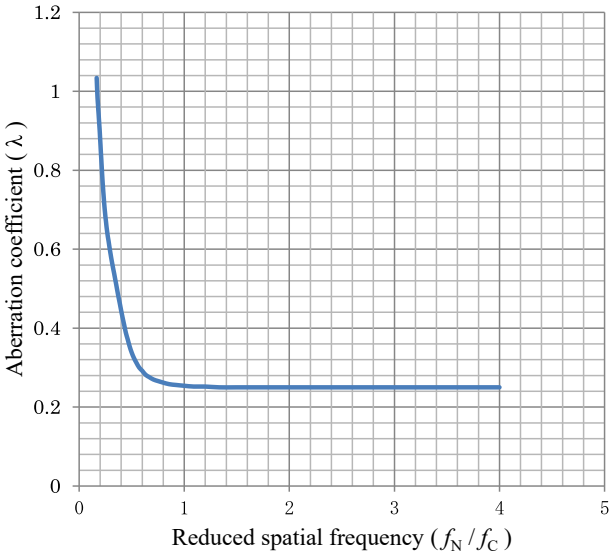
The DOF obtained through the rate of decrease in *MTF*, Δ_{MTF}, is given as

$$\Delta_{MTF} = \frac{n \bullet \lambda}{2NA^2} \bullet \left\{ 1.0 + \frac{f_C}{f_N} \right\} \tag{11}$$

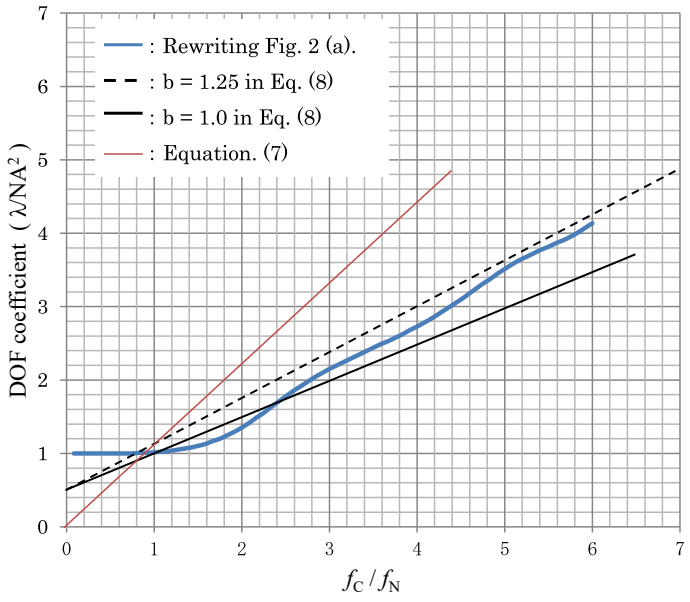
through a linear approximation of the result in Fig. 3.

2.4. A formulation of DOF equation for a microscope combing an image sensor with considering effect of pixel size

From the results of the analytical models in Sections 2.2 and 2.3, Eqs. (7), (8), and (11) showed, that when an image sensor is combined with an optical system, the DOF can be obtained, as in Eq. (1), by multiplying a coefficient by Rayleigh's DOF, $n\lambda/NA^2$ when the conversion parameter f_C/f_N is used as a parameter reflecting the effect of pixel size.



(a) Calculated results shown using the defocusing wave front coefficient and reduced spatial frequency, f_N / f_C. [6]



(b) Calculated results of (a) rewritten by using the DOF coefficient and the reduced spatial frequency of f_C / f_N.

Fig. 2. Results of DOF obtained from the total amount of light of 4 × 4 pixels positioned in the point image intensity distribution. (a) Calculated results shown using the defocusing wave front coefficient and reduced spatial frequency, f_N/f_C. (Yamamoto, 2014). (b) Calculated results of (a) rewritten by using the DOF coefficient and the reduced spatial frequency of f_C/f_N. The blue solid line is a rewriting of the calculated values in Fig. 2a, the black dotted line is Eq. (8) with b = 1.25, the black solid line is Eq. (8) with b = 1.0, and the red line is Eq. (7). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

However, the expression for DOF is considered slightly different whether the intensity distribution on the sensor is approximated by the PSF or by a confusion circle. For examples, in the range where the conversion parameter f_C/f_N is represented by Eq. (6), Eq. (7) is represented by Eq. (4). Similarly, it is considered better to use a coefficient b = 1.25 in Eq. (8).therefore, the DOF equation is expressed below

$$\Delta = \frac{n \bullet \lambda}{2NA^2} \bullet \left\{ A + B \bullet \frac{f_C}{f_N} \right\} \tag{12}$$

Within the defocused range where the light intensity distribution on the sensor is approximated by the PSF, it is considered that the coefficients A and B of Eq. (14) are expressed by the following values.

The coefficient A is the common value of Eqs. (8) and (11), A = 1.0, and B is defined as the intermediate value, B = 1.1, between the upper limit of b in Eq. (8) and the value in Eq. (11); therefore, the DOF is formulated as

$$\Delta = \frac{n \bullet \lambda}{2NA^2} \bullet \left\{ 1.0 + 1.1 \frac{f_C}{f_N} \right\} \tag{13}$$

Using the DOF coefficient, K (f_C/f_N), which represents the effect of the pixel size, the relation

$$\Delta = K(f_C/f_N) \bullet \frac{n \bullet \lambda}{NA^2}$$

$$K(f_C/f_N) = \{1.0 + 1.1 \bullet f_C/f_N\}/2 \tag{14}$$

holds similar to Eq. (1).

Focusing the conversion parameter f_C/f_N, when f_C nearly equal f_N, the resolution and the DOF are nearly equal values of the Rayleigh criterion.

When f_C is larger than f_N, the resolution is determined by the Nyquist frequency of the imaging device f_N and the DOF differ from values of the Rayleigh criterion.

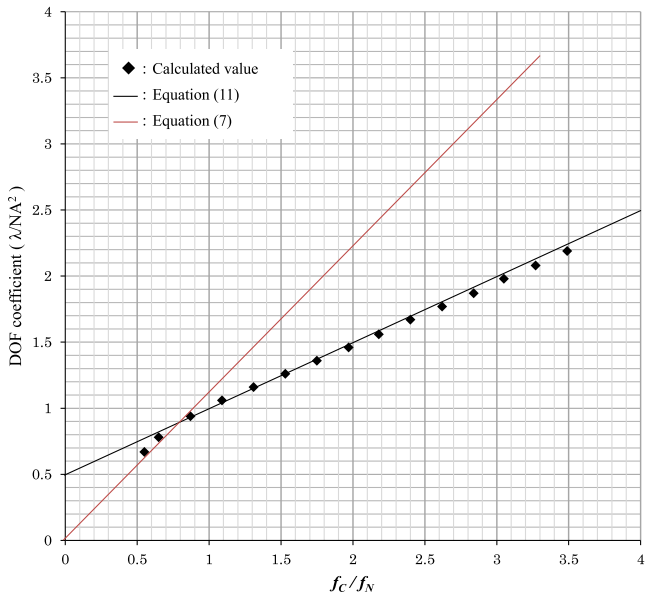


Fig. 3. The DOF, calculated from decreasing the MTF from defocusing and shown using the DOF coefficient, and the reduced spatial frequency expressed using f_C/f_N. Black squares ◆ show calculated values. The black and red solid lines indicate the values from Eqs. (11) and (7), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, since the DOF coefficient K (f_C/f_N) of Eq. (14) can be replaced with the wavefront coefficient of defocus as shown in Section 2.3.2, it is possible to associate the intensity distribution change of the PSF with the wave front coefficient of the defocus using the DOF coefficient K (f_C/f_N).

3. Experiments

We showed the DOF when an image sensor is combined with an optical system, Δ, can be expressed by Eq. (13) when the conversion parameter f_C/f_N is used as a parameter that takes the effect of the pixel size into account.

When a microscope, which is an optical equipment with diffraction-limited optical performance, is combined with a conventional image sensor, the conversion parameter f_C/f_N can be easily selected from less than 1 to values at which digital noises like moiré patterns might appear. Therefore, we then used a microscope as an experimental tool to verify the DOF equation that we formulated.

An object to be observed was moved in the optical axis (z) direction and images were captured. The DOF was measured using the following two experimental methods and the measured DOF values were compared with the DOF values calculated using Eq. (13).

Experimental method 1 calculated the image contrast at each z-step position and the DOF was measured from changes in the image contrast.

Experimental method 2 built an image-set of captured images with z-step position information. The image-set was displayed on a monitor. An observer measured the DOF through visual inspection by changing the z-step position.

3.1. Experimental method 1

In the experiment where the DOF was obtained from the contrast changes of captured images with different object (z) positions, the microscope system used was an IX81 (Olympus), with an IX2-LWUCD condenser lens (NA 0.55, Olympus), a black-and-white CCD camera –BM-150GE (2/3″, pixel size 6.45 μm × 6.45 μm, and number of pixels 1392 × 1040, JAI), and a quasi-monochromatic light source using an interference filter of 530 ± 20 nm.

The f_C/f_N value was adjusted by changing the total magnification to control the cut-off frequency on the image plane, f_C; the objective lens was chosen from one of LMPLFLN10x (10×, NA 0.25, Olympus), LMPLFLN20x (20×, NA 0.4, Olympus), and UMPLFL10x (10×, NA 0.3, Olympus); the TV adapter was either TV-1x (1×, Olympus) or TV-0.5x (0.5×, Olympus); and the intermediate magnification changer of the IX81 was switched between 1× and 1.6×.

The changes to the z-step position of the objective lens and the capturing of images were controlled simultaneously using a laptop computer ThinkPad X280 and LabView. The pattern chart in Fig. 4 was observed while changing the z-step position of the objective lens at step δz in the measurement range shown in Table 1a. Images were recorded in BMP format and the DOF was obtained from changes in the image contrast between images (within a predetermined region where the DOF was to be derived).

Three regions were defined for image contrast calculations: line widths W1, W2, and W3, where the line widths corresponded to 2- to 3-times, 3- to 4-times, and 4- to 5-times the pixel size, respectively.

DOF measurements were conducted four times for each combination of objective lens, TV-Adapter, and intermediate magnification changer. Table 1a shows the average of the measured DOFs for each combination.

Firstly, Fig. 5 (a) shows a comparison against theoretical results. The calculated values using Eq. (11) and measured values in Table 1a are shown to compare the DOF obtained from changes in the MTF, Δ_{MTF}, and the measured results. Next, to compare values from our proposed DOF equation and measured values, Fig. 5b compares calculated values using Eq. (13) and measured values in Table 1a. Moreover, to investigate the effect of pixel size on the DOF, Fig. 5c compares the DOF coefficient K (f_C/f_N) in Eq. (14) and the measured values divided by λ/NA_2.

3.2. Experimental method 2

This experiment determined the DOF by visually observing an image set shown on a monitor. As in the experiment in Section 3.1, an IX81 (Olympus) was used as the microscope system, and the changes in the z-step position of the objective lens and capturing of images were controlled with a desktop computer OptipPlex 960 and MetaMorph.

The image sensor was a black-and-white CCD camera CoolSnapHQ (2/3″, pixel size 6.45 μm × 6.45 μm, number of pixels 1392 × 1040, PHOTOMETRICS), with a LUCPLFLN20x objective lens (20×, NA 0.45, Olympus), IX2-LWUCD condenser lens (NA 0.55, Olympus), and quasi-monochromatic light source using an interference filter IF550 (550 ± 30 nm, Olympus). A striped pattern with uniform line width, as shown in Fig. 6, was observed. Images were captured while changing the z-step position of the object to obtain an image set.

Four types of TV-Adapters (0.25×, 0.5×, 1×, 2×, Olympus) were exchanged to switch the total magnification. The value of f_C/f_N was changed by changing the cut-off frequency, f_C, through this procedure.

During the capturing of images, the line width of the striped pattern in the observed image was changed for each total magnification such that the spatial frequency of the displayed image was the same; this was to suppress differences in image recognition between observers which arise from changes in line widths in the observed image.

Two types of images, with line widths of 10 lp/mm and 2 lp/mm on the image plane, were displayed in the observations for this study. Four types of striped patterns, namely 40 lp/mm, 100 lp/mm, 200 lp/mm, and 400 lp/mm, were used for 10 lp/mm on the image plane. Similarly, striped patterns with 10 lp/mm, 20 lp/mm, 40 lp/mm, and 100 lp/mm were used as the observation sample for 2 lp/mm on the image plane.

The observation region of the sample was set within a certain distance on both sides from the focus position. The interval, δz, splits the range of $\pm\lambda/NA^2$ calculated from the NA of the objective lens, into 40 segments. Here, the image does not change much when the measurement conditions are changed. Forty-one images were captured in the BMP format by changing the z-step position of the objective lens and an

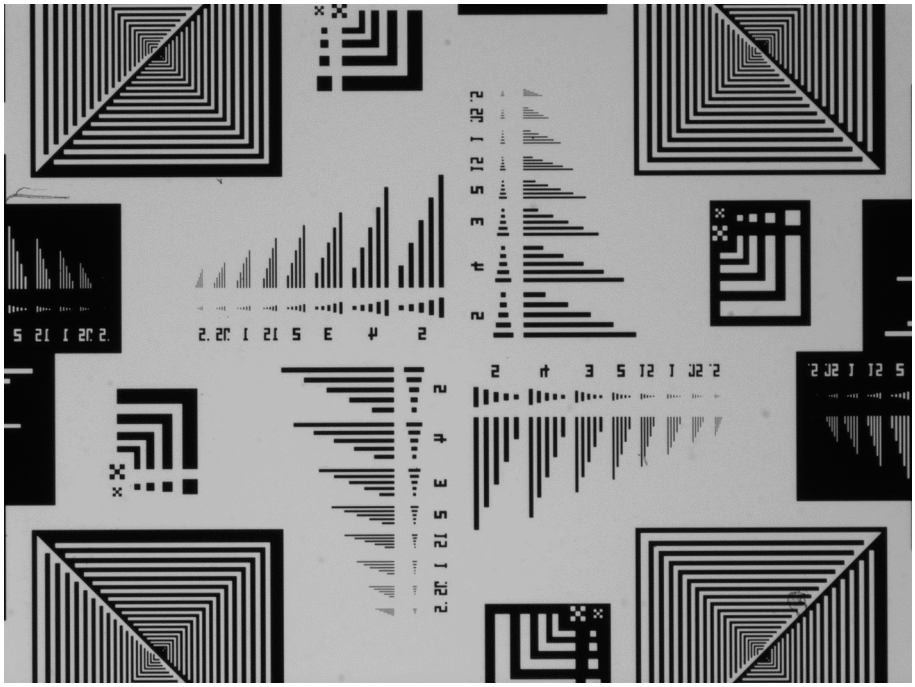


Fig. 4. Image of the pattern chart used in experiment 3–1.

Table 1
Results of the DOF measurement in experiment 3–1.

(a) Results of Calculation and measurement.

Observing optics		Calculation results			Measurement results					Measuring condition	
Objective lens	Total Mag.	λ/ NA^2(μm)	f_c / f_n	Δ by Eq. (13) (μm)	Mean value (μm)	Deviation σ (μm)	Results in Line Width (μm)			δz (μm)	Range (μm)
							W1	W2	W3		
LMPLFLN10x (NA0.25)	10×	8.48	1.217	9.92	9.50	0.84	8.38	9.50	10.63	0.5	20
	16×	8.48	0.779	7.87	8.67	0.17	8.38	8.38	9.25	0.5	20
	5×	8.48	2.434	15.59	14.71	1.06	13.38	14.88	15.88	0.5	20
UMPLFL10x(NA0.3)	10×	5.89	1.460	7.67	7.87	0.88	6.70	7.90	9.00	0.4	16
	16×	5.89	0.925	5.94	6.77	0.15	6.40	6.60	7.30	0.4	16
	5×	5.89	2.921	12.40	12.63	0.67	11.50	13.00	13.40	0.4	16
LMPLFLN20x(NA0.4)	10×	3.31	1.947	5.20	5.45	0.51	4.60	5.40	6.35	0.2	8

(b) The results of Table 1 (a) rewritten using with DOF coefficients.

Observing optics		Calculation results			Measurement results / λ/ NA^2				
Objective lens	Total Mag.	λ/ NA^2 (μm)	f_C / f_N	K f_C / f_N	Mean value	Deviation σ	Results in sub-domain		
							W1	W2	W3
LMPLFLN10x (NA0.25)	10×	8.48	1.217	1.1693	1.1203	0.0117	0.9876	1.1203	1.2529
	16×	8.48	0.779	0.9284	1.0220	0.0024	0.9876	0.9876	1.0908
	5×	8.48	2.434	1.8387	1.7345	0.0147	1.5772	1.7541	1.8721
UMPLFL10x (NA0.3)	10×	5.89	1.460	1.3032	1.3358	0.0254	1.1377	1.3415	1.5283
	16×	5.89	0.925	1.0087	1.1491	0.0043	1.0868	1.1208	1.2396
	5×	5.89	2.921	2.1064	2.1453	0.0193	1.9528	2.2075	2.2755
LMPLFLN20x (NA0.4)	10×	3.31	1.947	1.5709	1.6453	0.0466	1.3887	1.6302	1.9170

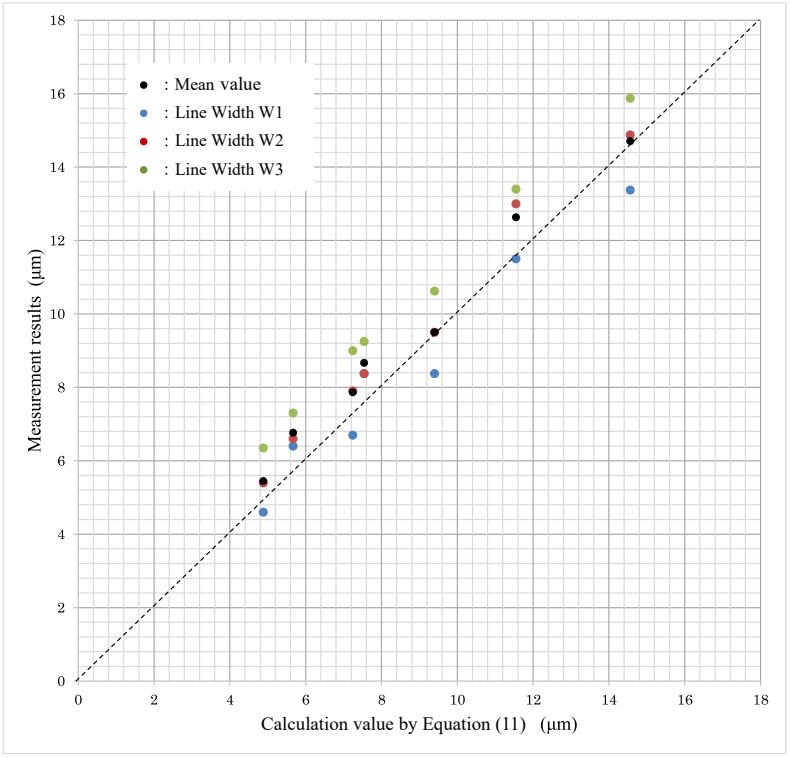
image set for the DOF measurement prepared in TIF format.

The obtained TIF-format image sets were shown on two types of monitors of size 20″ (SyncMasterXL20) and 12″ (ThinkPadX220). Measurements were conducted after displaying images such that each image pixel matched a monitor pixel. The distance between the observer and the monitor was controlled to keep the size of the object to be observed on the monitor constant.

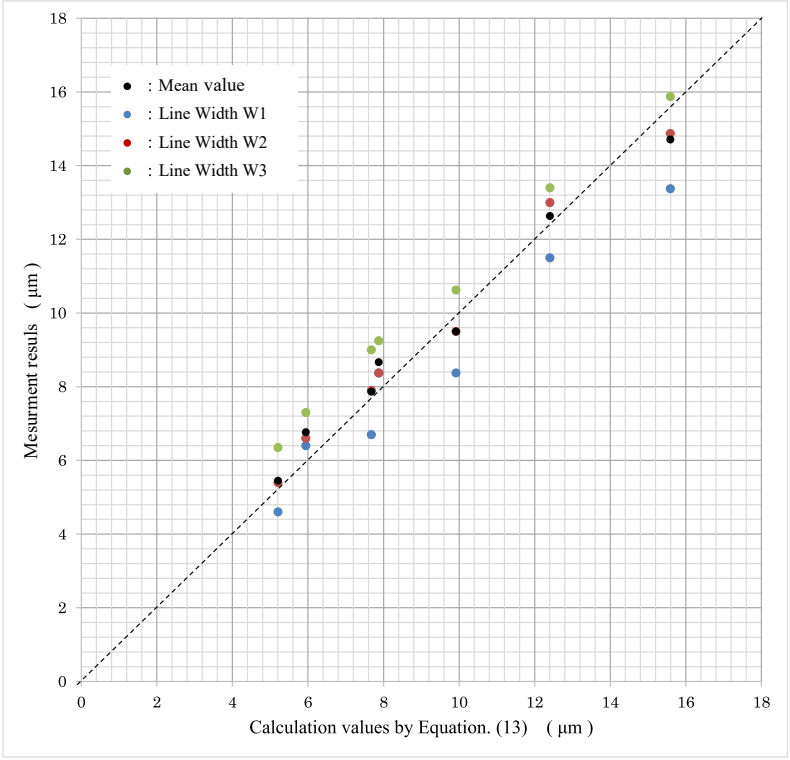
Images with different z-step positions in the image position were sequentially displayed. The DOF was determined, as shown in Fig. 7. The position in 41 images where the image became focused (page number of the TIF-format image) is denoted as ZS, and the position, after the focus position, where the image became defocused (page number of the TIF-format image) is defined as ZE. The number of image steps (pages) between ZS and ZE, Zn, and the change in z in one step, δz, was multiplied to obtain the DOF Δ.

$$\Delta = Zn \cdot \delta z \tag{15}$$

DOF measurements with different total magnification were conducted by 12 observers for the two observation samples with two different monitors each. The mean DOF of the 12 observers was taken as

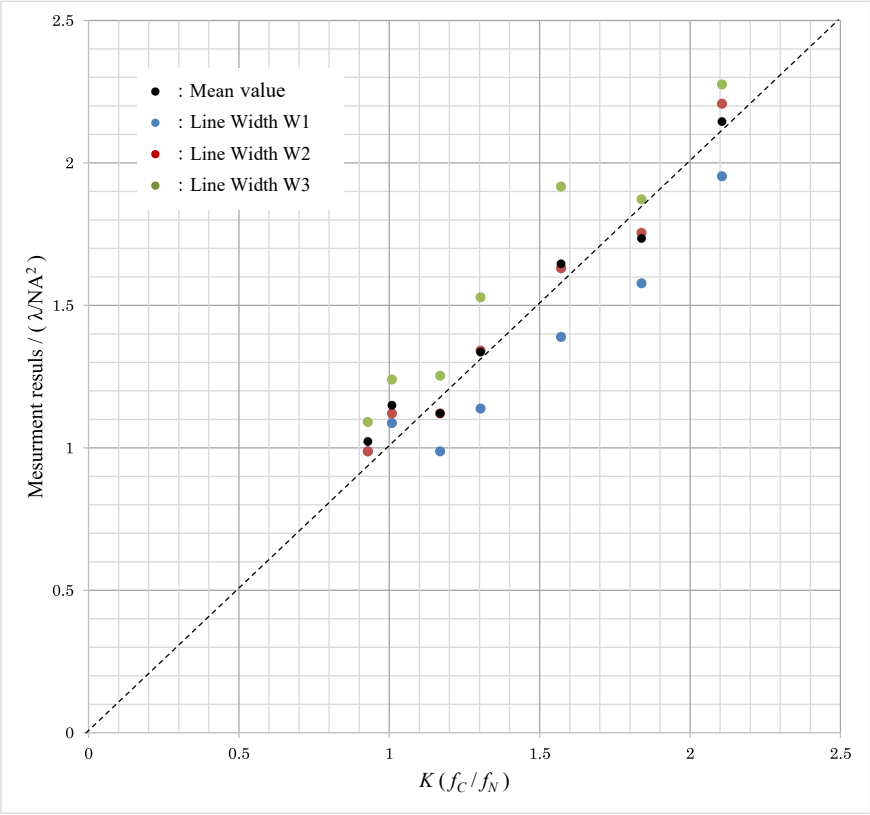


(a) Comparison of measured results and calculated results using Eq. (11)



(b) Comparison of measured results and calculated results using Eq. (13)

Fig. 5. Results measured in experiment 3–1. (a) Comparison of measured results and calculated results using Eq. (11) (b) Comparison of measured results and calculated results using Eq. (13). (c) Measured results divided by λ/NA^2 and results of (b) rewritten using the DOF coefficient, K (f_C/f_N). Blue, brown, and green dots (●, ●, and ●) indicate the measured values from line width W1, W2, and W3 patterns, respectively. Black dots (●) represent the mean values of the measured patterns, and the measured and calculated values are the same on the black dotted line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(c) Measured results divided by λ/NA^2 and results of (b) rewritten using the DOF coefficient, K (f_C / f_N).

Fig. 5. (*continued*).

the measured value.

Table 2 (a) shows the mean value, Ap, and its standard deviation, σ, for measurements where the spatial frequency of the object to be observed on the image plane, monitor size, and total magnification of the observation optical system were changed.

Table 2(b) shows a comparison between the measured values derived using Eq. (15) and the average values in Table 2(a). The results in Table 2(b) are plotted as Fig. 8(a).

To consider the effect of pixel size on the DOF, Fig. 8b compares the DOF coefficient K (f_C/f_N) in Eq. (14) and the measured value divided by λ/NA^2 in the same manner as in experiment 3–1.

4. Results and discussion

The values from the proposed equation and the measured values in experiment 3–1 are compared first.

Fig. 5a shows a comparison of measured results in Table 1(a) and the DOF from Eq. (11), Δ_{MTF}. The observed object is the pattern chart shown in Fig. 4; there is a good linear relation between Δ_{MTF}, which is obtained from changes in the MTF and measured values. In particular, the line width of the pattern chart for W1 is close to $f_N/2$, which is the condition used to calculate changes in the MTF; thus, the linear relation between

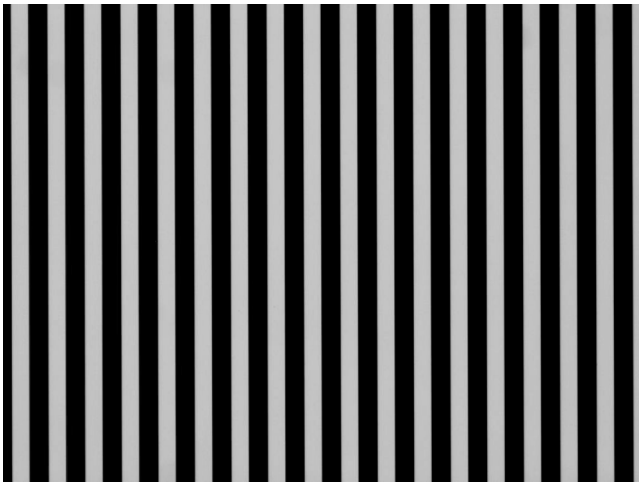


Fig. 6. Image of a striped pattern used in experiment 3–2.

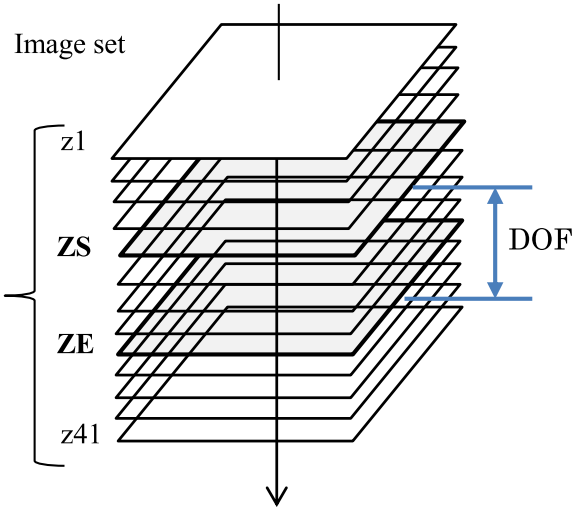


Fig. 7. Schematic of the measurement method in experiment 3–2.

Table 2
Results of the DOF observation in experiment 3–2.

(a) Observation results by twelve parsons

	Total magnification5 ×				Total magnification10 ×				Total magnification20 ×				Total magnification40 ×			
Sample	10 lp/mm		2 lp/mm		10 lp/mm		2 lp/mm		10 lp/mm		2 lp/mm		10 lp/mm		2 lp/mm	
Monitor	20″	12″	20″	12″	20″	12″	20″	12″	20″	12″	20″	12″	20″	12″	20″	12″
Average Ap (μm)(12 parsons)	9.15	8.93	7.05	7.35	4.08	4.5	3.75	3.71	2.75	2.65	2.3	2.55	1.85	1.70	1.88	1.65
Deviation σ (μm)(12 parsons)	3.77	2.79	2.47	2.71	1.72	1.29	1.42	1.66	1.25	1.09	1.01	1.06	0.93	0.84	0.93	0.84
Mean Ap (μm)	8.12				4.01				2.56				1.77			
Mean σ (μm)	2.935				1.523				1.103				0.885			

(b) Results of calculation and observation

	Calculation by Eq. (14)			Observation (μm)					Observation value / (λ/ NA^2)				
TotalMag.	*fc* / *fn*	*K*(*fc* / *fn*)	DOF Δ (μm)	Meanvalue	10 lp/ mm20″	10 lp/ mm12″	2 lp/ mm20″	2 lp/ mm12″	Meanvalue	10 lp/ mm20″	10 lp/ mm12″	2 lp/ mm20″	2 lp/ mm12″
5×	3.753	2.564	6.96	8.12	9.150	8.925	7.050	7.350	2.99	3.369	3.286	2.596	2.706
10×	1.876	1.532	4.16	4.01	4.083	4.500	3.750	3.708	1.476	1.503	1.657	1.381	1.365
20×	0.938	1.016	2.76	2.56	2.750	2.650	2.300	2.550	0.943	1.013	0.976	0.847	0.939
40×	0.469	0.758	2.06	1.77	1.850	1.700	1.875	1.650	0.652	0.681	0.626	0.69	0.608

the measured and calculated values is better than W2 and W3.

In contrast, the measured values for line width W1 in Fig. 5(a) are slightly smaller than the calculated values. A possible reason for this result is that the image contrast changed because spatial frequency components exceeding f_N in the pattern chart image are affected by noise removal during the developing process.

The effect of noise removal during the development process should be small for measurements of line widths W2 and W3 as the line width of the pattern chart increases and the ratio of spatial frequency components exceeding f_N becomes small.

Fig. 5b compares the measurement results in Table 1(a) and the calculated results using Eq. (13). The relation between the calculated values and measured values for line widths W2 and W3 is better than the results in Fig. 5(a). A good linear relation was found between the mean of line widths W1 to W3 and the calculated values, suggesting that Eq. (13) is more suitable than Eq. (11) for DOF calculations.

To consider the effect of pixel size on the DOF, Fig. 5c compares the DOF coefficient K (f_C/f_N) and the measured value divided by λ/NA^2. The measured value divided by λ/NA^2 can be replaced with the wavefront coefficient of defocus as shown in section 2.3.2. When this value K (f_C/f_N) is less than 2.5, the wavefront coefficient of defocus is less than $3\lambda/4$, and the light distribution on the sensor can be approximated by the PSF.

In the condition, the relation between the calculated values and measured values has a good linear relation as the DOF can be calculated by Eq. (13).

Next, the calculated and measured DOF values were compared when images displayed on a monitor were visually observed in experiment 3–2.

There is much scattering in the measurement results of the 12 observers, as shown in Table 2(a); however, a good linear relation is found between the mean value of the 12 observers and the calculated value in Eq. (13), as shown in Table 2b and Fig. 8a.

To consider the effect of pixel size on the DOF, Fig. 8b plots the DOF coefficient K (f_C/f_N) and the measured value divided by λ/NA^2, as in the results given in Fig. 5(c).

There is some discrepancy compared to the calculated result when the 10 lp/mm sample was observed near K (f_C/f_N) = 2.5 but the other measured values have a good linear relation with the calculated results.

As mentioned above, when K (f_C/f_N) is more than 2.5, the wavefront coefficient of the defocus approaches $3\lambda/4$, and the light intensity distribution on the sensor changes from the PSF to a uniform confusion circle. As shown in Section 2.4, when the light intensity distribution on the sensor is considered in a uniform confusion circle, the coefficient B of Eq. (12) can be 1.25.

Therefore, when K (f_C/f_N) is more than 2.5, it is thought that the DOF value calculated by the following formula is closer to the measured value than the value calculated by Eq. (13).

$$\Delta = \frac{n\cdot\lambda}{2NA^2}\cdot\left\{1.0 + 1.25\frac{f_C}{f_N}\right\} \tag{16}$$

From Eq. (10), moiré patterns might form when the pixel size expressed by the conversion parameter is in the range $3 < f_C/f_N$, because the spatial frequency components near f_C of an optical image is folded back by the sampling frequency $f_S = 2f_N$ and is mixed with the image components lower than the Nyquist frequency f_N.

When K (f_C/f_N) is near 2.5, the conversion parameter f_C/f_N is 3.7, the resolution of digital image is determined by f_N, and the 10 lp/mm sample included higher spatial frequency components than line width in an optical image may have had moiré effects on the image. Therefore, we think that the difference has widened between the DOF value calculated by Eq. (13) and the measured value.

When the pixel size is in the range $4 \leq f_C/f_N$, the resolution is determined by f_N, and the spatial frequency components of an optical image from lower than 0.75 f_C to f_C could be mixed with the image component low than the f_N as moiré patterns.

The DOF can be thought of as an unrecognizable range of changes in the intensity distribution of the image (less than 20 percent, as is the Rayleigh criterion), so the effects of moiré patterns need to be removed.

Since the effect of moiré depends on the periodicity of the observed object, we determined Eq. (13) as the DOF formula in the range $f_C/f_N < 4$ where the influence of moiré is small as shown in experiment 3–2.

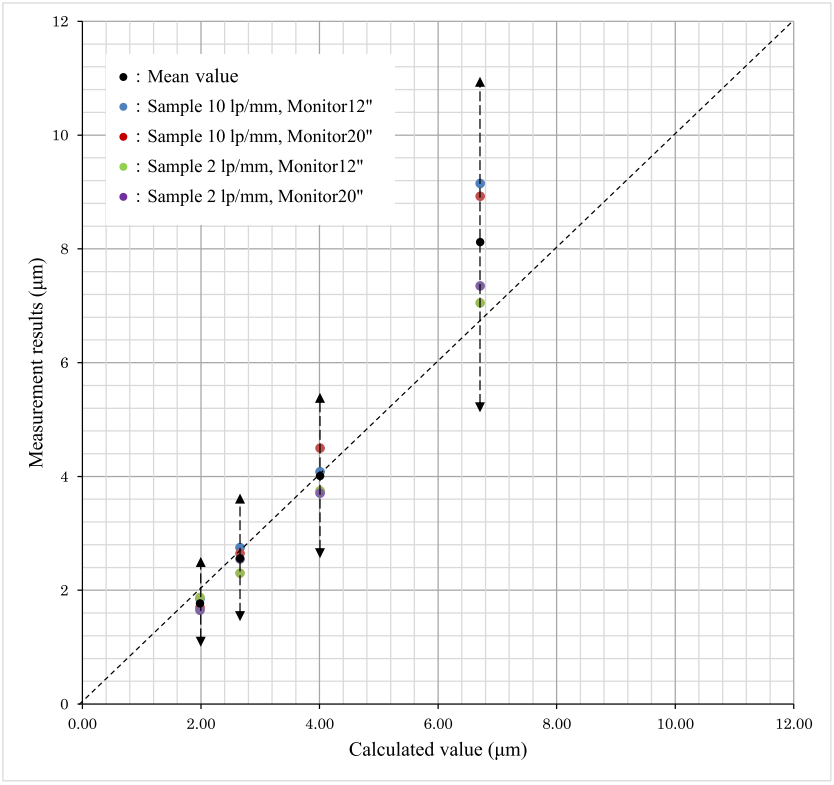
Recent digital image sensors reduce moiré patterns and noise using processes such as averaging during developing instead of an optical low-pass filter. Therefore, the developing process might have affected taken images.

The DOF of the object side and the DOF of the image side can be obtained separately using Eq. (13) by defining the cut-off frequency, f_C, of the microscope optical system and the Nyquist frequency, f_N, of the image sensor, as well as the NA on each side (object and image side).

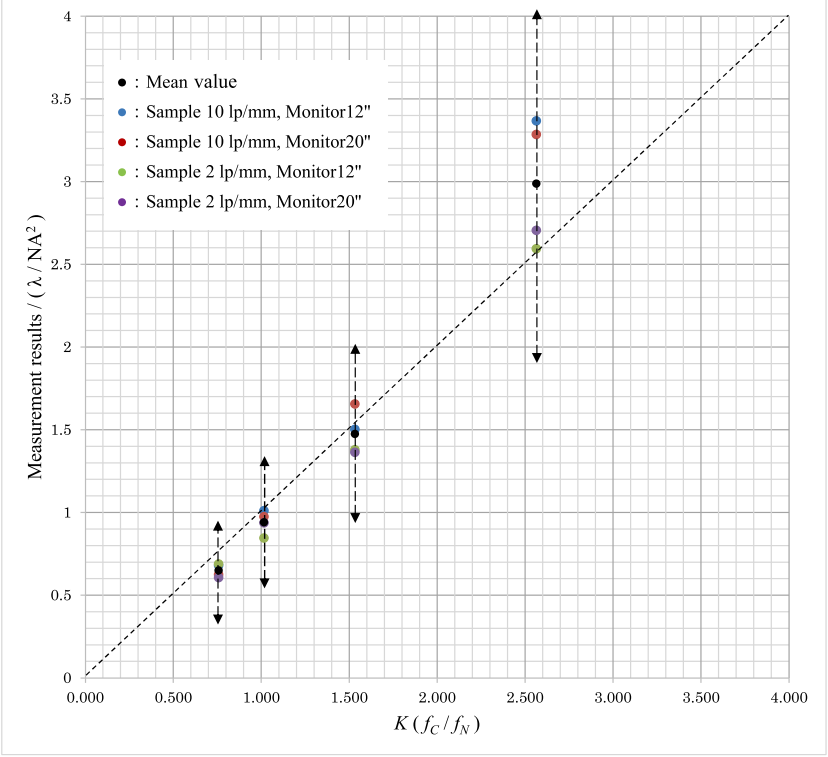
However, when considering the DOF of the object side, the value for just the optical system may be different from λ/NA^2, obtained using the Rayleigh criterion, when the NA is large (Young et al., 1993). This study does not consider the range when NA is large enough that the DOF of just the optical system is no longer λ/NA^2 (from the Rayleigh criterion). Therefore, the DOF on the object side is within the NA range where the Rayleigh criterion is applicable.

This study assumed a black-and-white image sensor with no effect of monitor resolution.

The resolution when an imaging device is combined with a microscope is defined as the lowest value of the cut-off frequency of the optical



(a) Comparison of measured results and calculated results using Eq. (13).



(b) Measured results divided by λ/NA^2 and results of (a) rewritten using the DOF coefficient, K (f_C / f_N).

Fig. 8. Measurement results from experiment 3–2. (a) Comparison of measured results and calculated results using Eq. (13). (b) Measured results divided by λ/NA^2 and results of (a) rewritten using the DOF coefficient, K (f_C/f_N). Blue dots (●) are the results of a striped pattern with 10 lp/mm on the imaging surface observed with a 12″ monitor. Brown dots (●) are results of observing a 10 lp/mm striped pattern with a 20″ monitor. Green dots (●) are results of observing a 2 lp/mm striped pattern with a 12″ monitor. Purple dots (●) are results of observing a 2 lp/mm striped pattern with a 20″ monitor. Black dots (●) are the mean of the measured values from blue to purple dots (● to ●), the black dash bars are the deviations (±σ) from each mean, and the measured and calculated values are the same on the black dotted line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

system and the Nyquist frequencies of the image sensor and monitor (ISO, 2016). An analogy to the resolution could be made for the DOF. The effect of the monitor can be obtained by choosing the Nyquist frequency, f_N, in the parameter considering the effect of pixel size, f_C/f_N, as the lower value of the Nyquist frequencies of the image sensor and monitor.

When a color image sensor is combined, a three devices type can be treated the same as a black-and-white image sensor thus, the DOF can be obtained using Eq. (13).

The image resolution information in a Bayer filter type image sensor depends on the interpolation processing of green (G) pixels. The DOF could be obtained with Eq. (13) if the image of green (G) pixels, after interpolation processing, is equivalent to images from a black-and-white image sensor.

Recently image processing methods that can extend or reduce the depth of focus are being investigated. The effect of pixel size must be considered when performing these methods because digital images are created by a sum of light intensity within a pixel.

In particular, the pixel size gives the depth of focus reduction method large effects, and in order to reduce the depth of focus below value of the Rayleigh criterion, from Eq. (13) we think the pixel size of a digital image sensor should satisfy the following condition.

$$f_C \leqslant f_N$$

In addition, in the non-staining evaluation of pluripotent cells that formed colonies, we measure phase distribution of cells in the colonies and evaluate them (Nishimura et al., 2019).

Since the phase amount of a three-dimensionally spread object such as a cell in a colony is proportional to the DOF on the object side, the DOF of an observing optical system is considered to plays an important role in quantitative phase distribution measurement.

Thick objects, such as large colonies or tissue fragments, scatter light that has passed through and reduce the image component, including the phase distribution, and display low signal-to-noise ratio images. In order to improve the signal-to-noise ratio, it is necessary to increase the light intensity that can be taken in by the optical system without reducing the DOF on the object side.

We think Eq. (13) is a good way to find the relationship between the DOF, the numerical aperture, and the pixel size, as one possible solution is to increase the numerical aperture without reducing the depth of focus.

5. Conclusions

When an optical system and a digital image sensor are combined, pixel size of an image sensor affects the resolution of digital images, the depth of focus, forming moiré patterns and etc.

In order to consider the effect of pixel size on digital images, we set the ratio, f_C/f_N, between the cut-off frequency of the optical system, f_C, and the Nyquist frequency of the imaging device, f_N, as the parameter expressing the effect of the pixel size and called it conversion parameter in this paper.

By using the conversion parameter, we derived that the effect of pixel size on digital images can be expressed in a simple formula.

For resolution, it can be expressed by the cut-off frequency of the optical system f_C when it is $f_C/f_N \leq 1$, and the Nyquist frequency of the imaging device f_N when it is $1 < f_C/f_N$.

In this paper, it is shown that the depth of focus within the range $f_C/f_N < 4$ without the influence of moiré can be expressed by equation Eq. (13), and that the results of two different measurement experiments using the microscope agreed with the calculated results from Eq. (13).

Therefore, we propose Eq. (13) as simple formula expressing the depth of focus when an optical system and a digital image sensor are combined.

Appendices

A.1 A method by calculating MTF at each defocus position

A method is shown to obtain the DOF from the decrease in MTF when defocusing is observed in the observation optical system.

The optical image of the observed object is expressed as a convolution of the intensity distribution of the object, I_0 (x,y), and the PSF of the optical system. The image intensities are integrated in each pixel when the image sensor receives this optical image, and the image is converted into a discrete image signal, as shown in Eq. (9).

Fourier transformation of the image signal in Eq. (9) is expressed as in Eq. (10); thus, the characteristics of the image signal captured at the imaging element can be viewed as the product of *OTF* (f_C) and a function expressing the effect of the pixel size (PP), *SINC* (f_S).

The characteristics of the observed image can be considered using *MTF* (f_C) when not combined with an image sensor. The DOF can be obtained from *MTF* (f_C) when there is defocusing in the optical system.

On the other hand, when combined with an image sensor, the image signal I_S(x,y) is discussed with *MTF* (f_C) • *SINC* (f_S).

The effect of pixel size on the DOF could be considered by calculating *MTF* (f_C) • *SINC* (f_S) when there is defocusing.

A method to calculate the DOF from *MTF* (f_C) • *SINC* (f_S) is shown using the following example. An ideal optical system with magnification 10 × and NA on the object side 0.3 is observed using light with a wavelength of 550 nm. The imaging device has square pixels arranged in a 600 × 600 two-dimensional array, and all pixels receive light. The pixel size was changed.

The tolerance of the DOF is the value of *MTF* (f_C) when defocusing of λ/4 happens in the optical system using the same defocusing wave front coefficient as the Rayleigh criterion.

The spatial frequency for evaluation is set when deriving the DOF using MTF. Here, $f_N/2$ is chosen as the spatial frequency when $f_N \leq f_C$.

As shown in Fig. A-1(a) and A-1(b), the *MTF* (f_C, *0, f*) at the focal position (def = 0) of the optical system only and the *MTF* when there is defocusing of λ/4, *MTF* (f_C, λ/4, *f*), also of the optical system only, are calculated. The difference at the focal position between MTF at $f_C/2$, *MTF* (f_C, 0, $f_C/2$), and the corresponding value with defocusing of λ/4, *MTF* (f_C, λ/4, $f_C/2$), is defined as *ΔMo* (λ/4). The value of *ΔMo* (λ/4), normalized by dividing with *MTF*

(f_C, 0, f_C/2), is obtained as $Po(\lambda/4)$.

$$\Delta Mo(\lambda/4) = MTF(f_C,\ 0, f_C/2) - MTF(f_C,\ \lambda/4, f_C/2)$$
$$Po(\lambda/4) = \Delta Mo(\lambda/4)/MTF(f_C,\ 0, f_N/2) \quad \text{(a.1)}$$

$Po(\lambda/4)$ represents the rate of change in the spatial frequency of the MTF at $f_N/2$ when there is defocusing of $\lambda/4$; thus, this can be considered the tolerance of the DOF.

Similarly, when there is a defocus wave front of an arbitrary amount, $MTF(f_C) \cdot SINC(f_S)$ is considered by defining ΔMp (*Defocus*) and Pp (*Defocus*) that correspond to ΔMo ($\lambda/4$) and Po ($\lambda/4$), respectively.

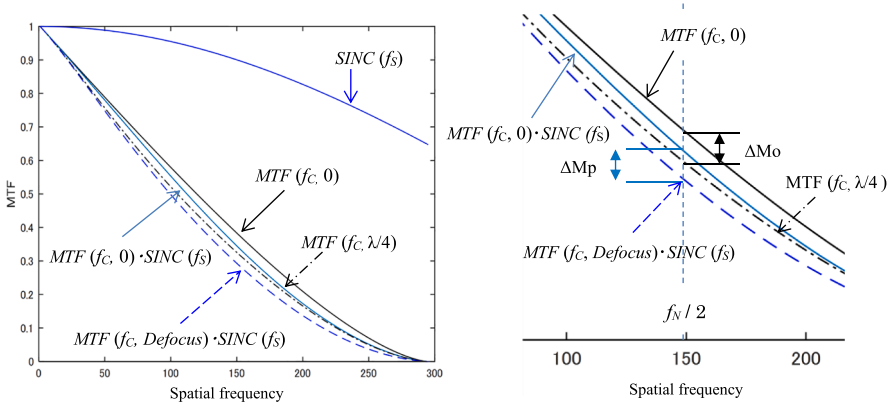
$$\Delta Mp(Defocus) = \{MTF(f_C, 0, f_N/2) - MTF(f_C, Defocus, f_N/2)\} \cdot SINC(f_S, f_N/2)$$
$$Pp(Defocus) = \Delta Mp(Defocus)/\{MTF(f_C, 0, f_N/2)\, SINC(f_S, f_N/2)\} \quad \text{(a.2)}$$

Po ($\lambda/4$) is the tolerance of the rate of change of MTF when an image sensor is combined. Therefore, the DOF is calculated using a defocusing wave front coefficient of

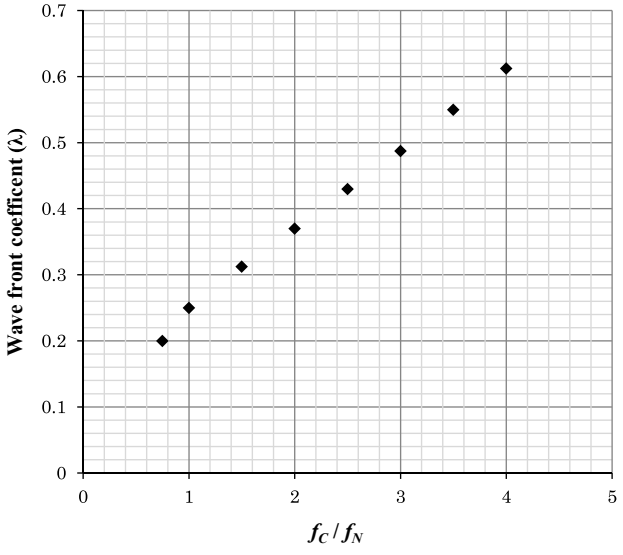
$$Pp(Defocus) = Po(\lambda/4) \quad \text{(a.3)}$$

Fig. A-1(c) shows the defocusing wave front coefficient becomes less than the tolerance when the pixel size (PP) is changed. As in Fig. 2(a), Fig. A-1 (c) shows the pixel size (PP) using the ratio, f_N/f_C, between the Nyquist frequency of the imaging device, f_N, and the cut-off frequency of the optical system, f_C.

The result in Fig. 3 shows the data from Fig. A-1(c), redrawn by expressing the defocusing wave front coefficient as a coefficient of λ/NA^2.



(a) Change in MTF by defocusing and setting the tolerance. (b) Enlargement of (a) near f_N / 2.



(c) Results of the defocusing wave front coefficient within the tolerance plotted against the normalized spatial frequency, f_C / f_N.

Fig. A-1. Schematic of a calculation to obtain the DOF from a decrease in the MTF and calculated results. (a) Change in MTF by defocusing and setting the tolerance. (b) Enlargement of (a) near $f_N/2$. (c) Results of the defocusing wave front coefficient within the tolerance plotted against the normalized spatial frequency, f_C/f_N.

A.2 A method by calculating contrasts at each z position images and DOF

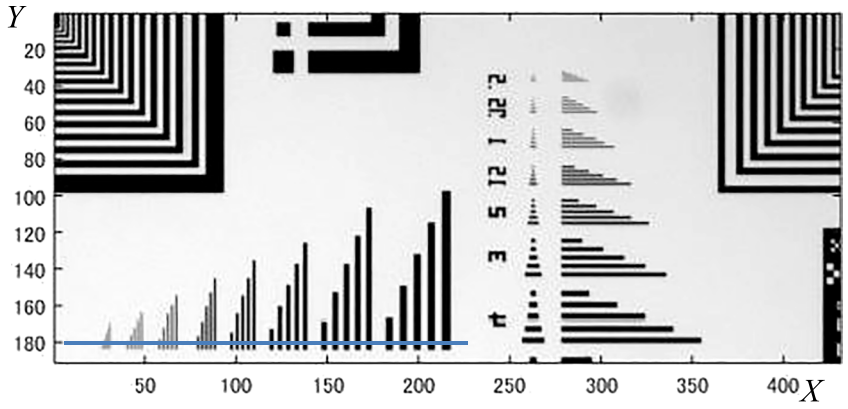
A method to calculate the DOF from images captured in experiment 3–1 is given below.

Six images are captured in BMP format at each z position, and an average image with lower electric noise is generated for each z position by averaging the six images. A calculation domain, shown in Fig. A-2(a), is cut out from the average image, and a line domain is defined where contrast calculations are conducted. Further averaging is conducted over three lines, which includes the lines immediately above and below the designated line domain. The image intensity distribution for each line width is calculated, as in Fig. A-2(b).

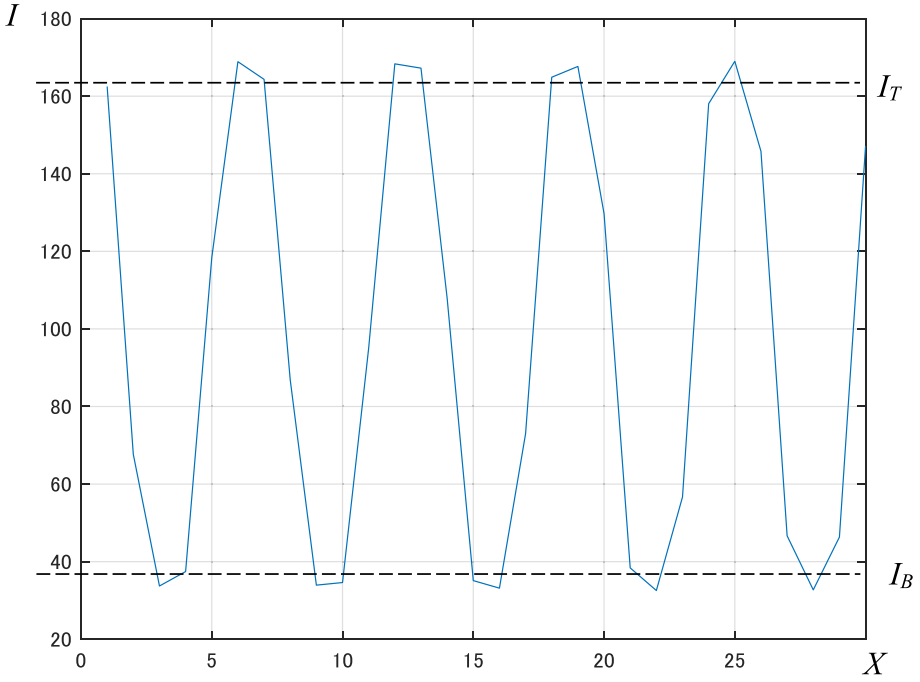
This image intensity information is sorted and the mean of a certain number of values (7, 12, and 20, in W1, W2, and W3, respectively) from the maximum value is taken as the maximum image intensity, I_T. Similarly, the minimum image intensity, I_B, is calculated by averaging a certain number of values from the minimum. The image contrast, Cont (z), for each z position of each line width is defined as

$$Cont(z) = (I_T - I_B) / (I_T + I_B) \quad \text{(a.4)}$$

The image contrasts are calculated for images captured at all z positions for DOF calculations and the relation between image contrast and z-step

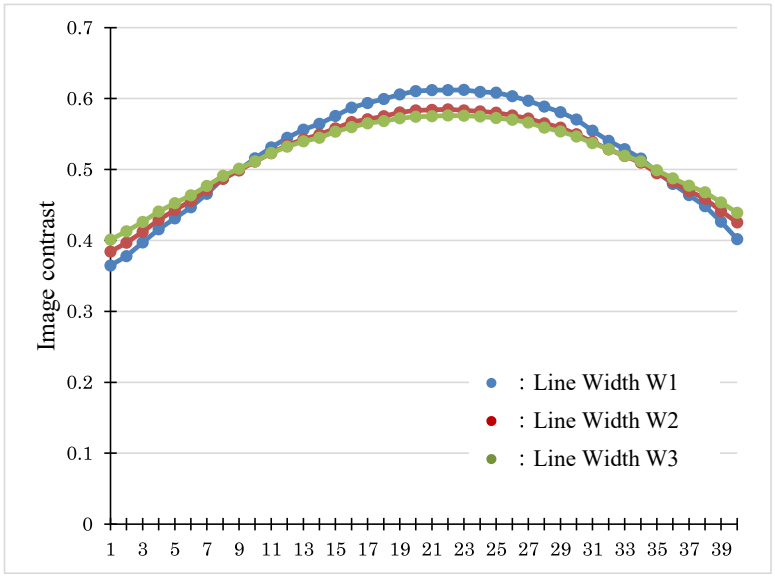


(a) Domain used for image contrast calculations.

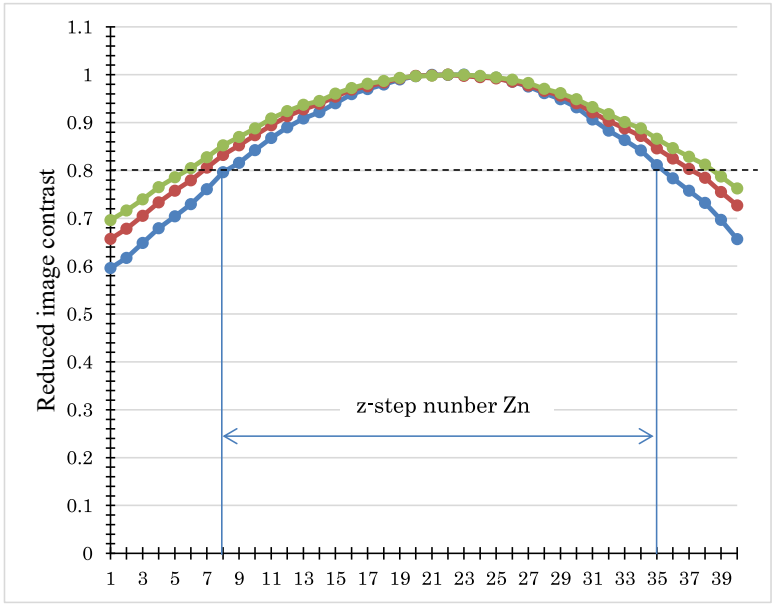


(b) Intensity of each line width in the cross-section along the domain.

Fig. A-2. Image contrast calculation in experiment 3–1. (a) Domain used for image contrast calculations. (b) Intensity of each line width in the cross-section along the domain.



(a) Image contrast at each z-step position.



(b) Calculating the DOF by normalizing the results in (a) by the image contrast at the focal position.

Fig. A-3. Method to calculate the DOF from image contrast calculations in experiment 3–1. (a) Image contrast at each z-step position. (b) Calculating the DOF by normalizing the results in (a) by the image contrast at the focal position. Blue, brown, and green dots (●, ●, and ●) indicate the measured values from line width W1, W2, and W3 patterns, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

position is calculated for each line width, as shown in Fig. A-3(a).

The maximum value in each line width in Fig. A-3(b) is extracted from the obtained relation between the image contrast and the z-step position, which is then used to obtain the relation between the normalized image contrast and the z-step position.

The tolerance in the relation between the normalized image constant and z-step position, which is shown in Fig. A-3a), is set to 0.8. The difference in z-step positions where the value of the normalized image constant dips below 0.8 is obtained as Zn.

The DOF Δ is calculated using the following equation and values of Zn and δz,

$$\Delta = (Zn - - 1) \delta z \tag{a.5}$$

Experiment 3–2 is a measurement to obtain the DOF through visual observation. The tolerance is not explicitly set but instead is implied from the judgment by the observer; thus, the DOF is obtained by Eq. (15).

In contrast, experiment 3–1 derives the DOF by setting a tolerance value and using z-step positions that dip below the tolerance, meaning an error of, at most, 2 z-steps could occur.

The DOF was calculated using Eq. (a-5) to reduce this measurement error. The measurement error can be decreased to δz by using Eq. (a-5). The value of δz is set to be $< 1/10$ of λ/NA^2 so the measurement error is negligible.

References

ISO 18221:2016, Microscopes - Microscopes with digital imaging displays - Information provided to the user regarding imaging performance.

Levinson, H.J., 2005. Principles of Lithography Second Edition, SPIE PRESS Bellingham Washington USA.

Mack, C.A., 1988. Understanding focus effects in submicron optical lithography, SPIE Vol. 922 Optical /Laser Microlithography.

Martin, L.C., 1966. 'Chapter IV Incoherent Illumination', The Theory of The Microscope, London BLACKIE Glasgow.

Murphy, D.B., Davidson, M.W., 2013. Fundamental of light microscope and electronic imaging, Wiley-Blackwell.

K. Nishimura H. Ishiwata Y. Sakuragi Y. Hayashi A. Fukuda K. Hisatake Live-cell imaging of subcellular structures for quantitative evaluation of pluripotent stem cells Sci. Rep. 9 2019 1777.

J.C. Wyant and K. Creath, 'Chapter 1 Basic Wavefront Aberration Theory for Optical Metrology' Appl. Opt. Opt. Eng., Vol. Xl, (1992), Academic Press, Inc.

Yamamoto, K., 2014. Depth of diffraction-limited imaging system incorporating electronic device, Opt. Rev., 21, No. 6 (2014) 795-799.

I.T. Young, R. Zagers, L.J. van Vliet, J. Mullikin, F. Boddeke and H. Netten, Depth-of-Focus in microscopy, in: SCIA'93, Proc. of the 8th Scandinavian Conference on Image Analysis, Tromso, Norway, (1993), 493-498.

ARYAN INFRA PROJECTS

To Provide Homes and Happiness for Generation

To deliver "HOMES" not Houses through, Innovative ideas Partnership & Technology. with speed commitment & accuracy for the ultimate "CUSTOMER DELIGHT".

WELCOME TO Aryan Infra Projects

ABOUT US

Aryan Infra is an established & renowned ISO 9001-2008 Certified. Within a short span of time it has set up many branches across different parts of Odisha and also even outside Odisha-Kolkata, Delhi, Surat & Bangalore. Our Projects includes independent Housing Scheme, Apartments, Flats, Residential plotted schemes, Farm Houses, etc. We have been crowned with numerous awards like Utkal Sanman, Jewel of India Award, and Bhartiya Nirman Ratna Awards.

PROJECTS

(The Prestige)
Block-C Front side view

(The Prestige)
Block-G,H,I Front side view

(The Prestige)
Block-B Back side view

(The Prestige)
Block-E Front side view

Address:
Plot No-540,Saheed Nagar
Near Saheed sporting club,BBSR-7

Phone:
+91 9438567803

Email:
info@aryaninfraprojects.com

Website
www.aryaninfraprojects.com