



Iranian Association of Electrical
and Electronics Engineers



Shahid Chamran
University of Ahvaz

Journal of **Applied Research in Electrical Engineering**

Vol. 2, No. 1, Winter and Spring 2023



PUBLISHER: SHAHID CHAMRAN UNIVERSITY OF AHVAZ

E-ISSN: 2783-2864

P-ISSN: 2717-414X



Journal of Applied Research in Electrical Engineering

E-ISSN: 2783-2864

P-ISSN: 2717-414X



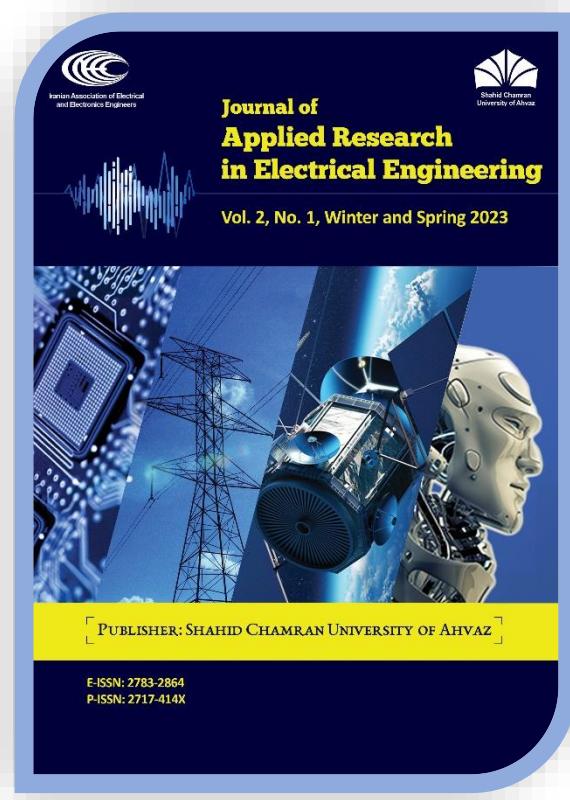
Journal of Applied Research in Electrical Engineering (JAREE), Volume 2, Number 1, Winter and Spring 2023

Publisher: Shahid Chamran University of Ahvaz, Iran

This magazine is the result of a formal partnership of **Shahid Chamran University of Ahvaz** and **Iranian Association of Electrical and Electronics Engineers**

Website: <https://jaree.scu.ac.ir>

E-mails: jaree@scu.ac.ir; jaree.scu@gmail.com



Address: Department of Electrical Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Golestan Street, Ahvaz, Iran

P.O. Box: 61357-85311

Tel: +986133226600-14, Ext. 5632 & 5630, +989122876375

Fax: +986133226597

Editorial Board

| | |
|--|---|
|  Editor-in-Chief Prof. Mahmood Joorabian Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Power System Planning, Renewable Energy and Smart grid, FACTS Devises</i> mjorabian@scu.ac.ir |  Director-in-Charge Prof. Seyed Ghodratolah Seifossadat Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Power System Protection, Power Electronics, Power Quality</i> seifossadat@yahoo.com |
|  Managing Editor Dr. Alireza Saffarian (Associate Professor) Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Power System Protection, Power System Stability, Power Quality, Distribution Systems</i> a.saffarian@scu.ac.ir |  Executive Assistant Dr. Mohammad Nabipour (Assistant Professor) Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Power Electronics and Drive</i> mo.nabipour@scu.ac.ir |
|  Associate Editor Dr. Yousef Seifi Kavian (Associate Professor) Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Electronics and Telecommunications</i> y.s.kavian@scu.ac.ir |  Associate Editor Dr. Mohsen Saniei (Associate Professor) Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Power and Control</i> m.saniei@scu.ac.ir |
|  Editorial Board Prof. Saeedallah Mortazavi Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Intelligent Control Systems, Power system Control, Automation, Fuzzy logic, Neural Networks</i> mortazavi_s@scu.ac.ir |  Editorial Board Prof. Abdolnabi Kosarian Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Solid State Electronic Devices, Solar Cell Fabrication Technology</i> a.kosarian@scu.ac.ir |
|  Editorial Board Prof. Ebrahim Farshidi Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Analog and Digital Integrated Circuits, Data Converters, Microelectronics</i> farshidi@scu.ac.ir |  Editorial Board Prof. Naser Pariz Ferdowsi University of Mashhad, Mashhad, Iran <i>Nonlinear Control, Hybrid Systems, Aeronautics, Industrial Control, Applied Mathematics</i> n-pariz@um.ac.ir |
|  Editorial Board Dr. Yousef Seifi Kavian (Associate Professor) Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Digital Circuits and Systems, Communication Networks and Distributed Systems</i> y.s.kavian@scu.ac.ir |  Editorial Board Dr. Mohsen Saniei (Associate Professor) Shahid Chamran University of Ahvaz, Ahvaz, Iran <i>Power System Dynamics, High Voltage Engineering, Electricity Market, Microgrid</i> m.saniei@scu.ac.ir |
|  Editorial Board Prof. Abbas Zarifkar Shiraz University, Shiraz, Iran <i>Optical Electronics</i> zarifkar@shirazu.ac.ir |  Editorial Board Dr. Edris Pouresmaeil (Associate Professor) Aalto University, Espoo, Finland <i>Integration of renewable energies into the power grid</i> edris.pouresmaeil@aalto.fi |
|  Editorial Board Prof. Reza Ghaderi Shahid Baheshti University, Tehran, Iran <i>Control Theory, System Identification, Control Systems, Fuzzy Engineering</i> r_ghaderisbu.ac.ir |  Editorial Board Prof. Fushuan Wen Tallinn University of Technology, Tallinn, Estonia <i>Power systems and power economics</i> fushuan.wen@taltech.ee |
|  Editorial Board Prof. Majid Sanaye-pasand University of Tehran, Tehran, Iran <i>Power Systems Protection, Digital Protective Relays, Power Systems Automation, Power System Transients</i> msanaye@ut.ac.ir |  Editorial Board Prof. Zabih (Fary) Ghassemlooy Northumbria University, Newcastle upon Tyne, United Kingdom <i>Optical Communications, Visible Light, Communication Systems</i> z.ghassemlooy@northumbria.ac.uk |

Editorial Board (Continued)

| | | | |
|---|---|---|--|
|  | Editorial Board Prof. Mahdi Tavakoli University of Alberta, Alberta, Canada <i>Robotics and Telerobotics, Haptics and Teleoperation Control, Surgical and Therapeutic Robotics, Image-Guided Surgery</i> mahdi.tavakoli@ualberta.ca |  | Editorial Board Prof. Mohamad Hassan Modir Shanechi Illinois Institute of Technology, Chicago, USA <i>Nonlinear and intelligent systems, Power system dynamics and security, Power system planning and maintenance scheduling</i> shanechi@iit.edu |
|  | Editorial Board Prof. Mohammad Shahidehpour Illinois Institute of Technology, Chicago, USA <i>Power Systems, Microgrids, Power System Operation, Power System Planning</i> ms@iit.edu |  | Language Editor Majid Sadeghzadeh Hemayati <i>English Language Editing</i> m_s_hemayati@yahoo.com |
|  | Page Designer Dr. Mahyar Abasi Shahid Chamran University of Ahvaz, Ahvaz, Iran mahyarabasi1368@yahoo.com | | |

About Journal

Journal of Applied Research in Electrical Engineering (J. Appl. Res. Electr. Eng.) is a single-blind peer-reviewed, [open access](#) and [free of charge](#) international journal published by Shahid Chamran University of Ahvaz in cooperation with Iranian Association of Electrical and Electronics Engineers (IAEEE). The JAREE is a medium for global academics to exchange and spread the latest discoveries and advances in their applied research in electrical engineering. The JAREE aims at presenting important results of analytical, computational and experimental works within all specialized fields of electrical engineering (electronics, power, control and telecommunications). It welcomes high quality original research papers from contributors throughout the world. All papers are subject to a peer reviewing procedure. Submission, processing and publication of the papers in JAREE is [free of charge](#).

Types of accepted papers include:

- Research articles
- Review articles
- Applied articles

Research papers are expected to present innovative solutions, novel concepts, or creative ideas that can help to address existing or emerging technical challenges in electrical engineering.

Application papers are expected to share valuable industry experiences on dealing with challenging technical issues, developing/adopting new standards, applying new technologies or solving complex problems. JAREE welcomes application papers that can have a significant impact on industry practices in the coming years.

Review papers are expected to provide insightful and expert reviews, tutorials, or study cases on an important, timely and widely-interested topic in electrical engineering.

All researchers in the fields of electrical science are invited to publish their scientific and research achievements in this journal. Interested authors can submit their manuscripts in the journal's website. More information is available in the website on how to prepare and submit the manuscripts.

Amis and Scope

The *Journal of Applied Research in Electrical Engineering* aims to provide a medium for dissemination of innovative and consequential papers that present analytical, computational and experimental works within all specialized fields of electrical engineering (electronics, power, control and telecommunications). The scope of the journal includes, but is not limited to, the following topic areas:

Electronics:

- Optical electronics, photonics and solar cells
- Integrated analog circuits and mixed signals
- Integrated radio frequency circuits
- Digital electronics (VLSI)
- Semiconductor devices
- Sensor technology

Power:

- Dynamics and stability of the power systems
- Power system protection
- Electric power quality
- Operation and planning of the power systems
- High voltage insulation technology
- Flexible AC Transmission Systems (FACTS)
- Electric power distribution systems
- Smart grids, micro-grids, renewable energies and distributed generation
- Reliability of electrical energy systems
- Energy management and electricity market
- Electric machines and transformers
- Power electronic and electric drives

Control:

- Linear and non-linear control systems
- Adaptive, optimal and stochastic control
- Fuzzy systems, neural networks and intelligent control
- Robotic and mechatronic
- Modeling, Identification and optimization of systems
- Guidance and navigation systems
- Automation, industrial control and instrumentation

Telecommunications:

- Signal and image processing
- Wireless and cellular communication systems
- Telecommunication networks
- Radar and sonar
- Information theory and coding
- Cognitive radio
- Antenna design
- Microwave devices
- Wave propagation and electromagnetic compatibility

Indexing Databases and Social Networks

Directory of Open Access Journals (DOAJ):

<https://doaj.org/toc/2783-2864>

Google Scholar:

<https://scholar.google.com/citations?user=F7KQPtYAAAAJ&hl=en&authuser=1>

Directory of Open Access Scholarly Resources (ROAD):

<https://portal.issn.org/resource/ISSN/2783-2864>

LinkedIn:

<https://www.linkedin.com/in/journal-of-applied-research-in-electrical-engineering-jaree-7540871b2/>

Academia:

<https://shahidchamranahwaz.academia.edu/JournalAppliedResearchinElectricalEngineering>

Mendeley:

<https://www.mendeley.com/profiles/journal-of-applied-res-in-electrical-engineer/>

Twitter:

https://twitter.com/jaree_scu

Facebook:

<https://www.facebook.com/jaree.scu>

Researchgate:

https://www.researchgate.net/profile/Jaree_Engineering

Telegram:

<https://t.me/jareescu>

Instagram:

<https://www.instagram.com/jaree.scu/>

Journal homepage:

<http://jaree.scu.ac.ir>

Journal emails:

jaree@scu.ac.ir, jaree.scu@gmail.com

Guide for Authors

How to submit a manuscript

For the initial submission, the authors have to just send the main manuscript file and the signed [Copyright Form](#) of the journal. While preparing manuscripts for initial submission, authors are kindly requested to follow the guidelines, described below:

- The manuscript should be written in a Microsoft Word file (.doc or .docx).
- The file should include text (preferably in 10 points, “Times New Roman” font) and all figures (figures can be placed within the text at the appropriate point or at the end of the text).
- The manuscript pages should be prepared either using a double-column single-line spacing layout or a single-column double-line spacing layout. A margin of at least 1.5 cm on each side is required.
- All papers should be composed of Title, Author Name, Affiliation, Corresponding author email, Abstract, Keywords, Body, and References.
- The manuscript should be written in good English. It should have been carefully checked for clarity, conciseness, the correctness of grammar, and typographical errors.
- The corresponding author should sign the journal copyright form on behalf of any and all co-authors and upload it to the Journal’s Submission System when submitting the manuscript. The journal copyright form can be downloaded from [here](#).
- The corresponding author can use the [JAREE Template for Cover Letter](#) as a default text for the cover letter when submitting the manuscript.
- It is recommended that the title of the paper does not contain abbreviations or formulae.
- The abbreviations used in the abstract should be introduced both in the abstract and again on first use in the body.
- References should be numbered in the order they are mentioned in the text.

Manuscript Submission

Submission to this journal proceeds totally online and you will be guided stepwise through the creation and uploading of your files. All correspondence, including notification of the Editor's decision and requests for revision, takes place by e-mail. To submit your manuscript, click on the [Submit Manuscript](#) link on the journal's homepage. Then, click on [Register](#) to create an author account. A message is sent to your email address containing your username and password. Then, login to the Journal’s Submission System at the [User's login](#) page using the username and password to submit your new manuscript. Once you have logged in, you can change your password by clicking on the My Home link at the top menu.

Copyright and Open Access License

An author submitting a paper should ensure that he or she has the right to publish the paper and that it contains nothing defamatory. The JAREE will assume that all co-authors have agreed to the submission of any paper received. The corresponding author should sign the journal copyright form on behalf of any and all co-authors and upload it to the Journal’s Submission System when submitting the manuscript.

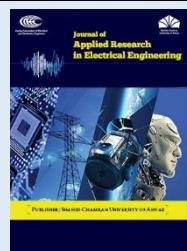
Contents

| Article Title and Authors | Page No. |
|---|-----------|
| Analyzing the Inference Process in Deep Convolutional Neural Networks using Principal Eigenfeatures, Saturation and Logistic Regression Probes Mats Leon Richter, Leila Malihi, Anne-Kathrin Patricia Windler, and Ulf Krumnack | 1 |
| Improving Stochastic Computing Fault-Tolerance: A Case Study on Discrete Wavelet Transform Shabnam Sadeghi, and Ali Mahani | 11 |
| Investigation of the Operation of Active Superconducting Fault Current Limiters in Distribution Networks Connected to Microgrids Ahmad Ghafari, Mohsen Saniei, Morteza Razzaz, and Alireza Saffarian | 19 |
| Multi-Objective Optimal Power Flow Based Combined Non-Convex Economic Dispatch with Valve-Point Effects and Emission Using Gravitation Search Algorithm Nabil Mezhoud, and Mohamed Amarouayache | 26 |
| Smart AI-based Video Encoding for Fixed Background Video Streaming Applications Mohammadreza Ghafari, Abdollah Amirkhani, Elyas Rashno, and Shirin Ghanbari | 37 |
| Improving the Quality of ECG Signal Using Wavelet Transform and Adaptive Filters Amir Hatamian, Farzad Farshidi, Changiz Ghobadi, Javad Nourinia, and Ehsan Mostafapour | 45 |
| Effect of Changes in the Parameters of a Modular Converter in Its Controllability Range in Fuel Cell Applications Mohammad Afkar, Parham Karimi, Roghayeh Gavagsaz-Ghoachani, Matheepot Phattanasak, and Serge Pierfederici | 54 |
| Robustness Analysis of Model Reference Adaptive Controller in The Presence of Input Saturation Using Describing Function Method Fatemeh Tavakkoli, Alireza Khosravi, and Pouria Sarhadi | 62 |
| Investigating the Effect of Geometric Design Parameters on the Mutual Inductance Between Two Similar Planar Spiral Coils With Inner and Outer Diameter Limits Ata Ollah Mirzaei, Amir Musa Abazari, and Hadi Tavakkoli | 70 |
| Partial Discharge Pattern Recognition in GIS Using External UHF Sensor Reza Rostaminia, Mehdi Vakilian, and Keyvan Firouzi | 75 |
| A Feedforward Active Gate Voltage Control Method for SiC MOSFET Driving Hamidreza Ghorbani, and Jose Luis Romeral Martinez | 87 |
| Design of Low-Power Approximate Logarithmic Multipliers with Improved Accuracy Mojtaba Arab Nezhad, and Ali Mahani | 95 |

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution -Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Analyzing the Inference Process in Deep Convolutional Neural Networks using Principal Eigenfeatures, Saturation and Logistic Regression Probes

Mats Leon Richter , Leila Malihi* , Anne-Kathrin Patricia Windler, and Ulf Krumnack

Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

* Corresponding Author: lemalih@uni-osnabueck.de

Abstract: The predictive performance of a neural network depends on the one hand on the difficulty of a problem, defined by the number of classes and complexity of the visual domain, and on the other hand on the capacity of the model, determined by the number of parameters and its structure. By applying layer saturation and logistic regression probes, we confirm that these factors influence the inference process in an antagonistic manner. This analysis allows the detection of over- and under-parameterization of convolutional neural networks. We show that the observed effects are independent of previously reported pathological patterns, like the “tail pattern”. In addition, we study the emergence of saturation patterns during training, showing that saturation patterns emerge early in the optimization process. This allows for quick detection of problems and potentially decreased cycle time during experiments. We also demonstrate that the emergence of tail patterns is independent of the capacity of the networks. Finally, we show that information processing within a tail of unproductive layers is different, depending on the topology of the neural network architecture.

Keywords: Convolutional neural networks, logistic regression probes, saturation, eigenfeatures, tail pattern.

Article history

Received 16 February 2022; Revised 13 June 2022; Accepted 29 June 2022; Published online 9 July 2022.

© 2022 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

M. L. Richter, L. Malihi, A. -K. P. Windler, and U. Krumnack, "Analyzing the inference process in deep convolutional neural networks using principal eigenfeatures, saturation and logistic regression probes," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 1-10, 2023. DOI: [10.22055/jaree.2022.40073.1049](https://doi.org/10.22055/jaree.2022.40073.1049)



1. INTRODUCTION

The problem of the opaqueness of neural networks is one of the key challenges in deep learning. This has led to a primarily trial-and-error driven mode of development, based on the comparison of abstract metrics that capture model-agnostic concepts like predictive performance, demand in computational resources, and capacity [1-7]. To move towards a more efficient, principle-based design process, a more profound understanding of the model’s state is required. This understanding does not have to be necessarily complete in regard to fully understanding the relation of the input and output of the model. Comparative analysis methods based on singular vector canonical correlation analysis (SVCCA) [8] are good examples of such non-holistic approaches. The information extracted from the model using SVCCA is highly aggregated, condensing each layer pair to a single value, but allows for useful insights into the converged model and the training process, by comparing different layers within that model, as well as the states of one layer at different training epochs. Another example close to our approach is the intrinsic dimensionality used for PCA-based pruning by [9, 10]. This

method studies the inference process with spectral methods to determine unproductive layers and unnecessary filters. This method, originally only operable for simple sequential architectures, was later expanded on in [11] to ResNet-style architectures. Logistic regression probes [12] and saturation [13] aggregate a single layer to a number, which allows for easy and intuitive analysis, similar to measuring with a thermometer. While logistic regression probes measure the intermediate solution quality very directly by training logistic regressions on the output of a layer, saturation is more task agnostic. Richter et al. [13, 14] have shown that for visual classification tasks, the dimensionality of the subspace of features responsible for data processing varies significantly depending on the input resolution, leading to model and training inefficiencies. While saturation is easy to define, the exact properties of this metric are yet to be discovered. In this paper, we will consider the following questions to gain a better understanding of hidden layer saturation:

- How do model capacity and problem difficulty influence the saturation value?

- Is the organization of the inference process influenced by the capacity of individual layers or the capacity of the entire network?
- How do saturation patterns evolve in the training phase?

We address these questions in a series of experiments. This paper is an extended version of the conference contribution [15], introducing new results and adding some details. It is structured as follows: After introducing the concepts relevant for our work (Section 2), we demonstrate the idea of principal eigenfeatures using an autoencoder (Section 3). We then present further experiments conducted to address the questions raised above (Sections 4, 5, and 6). The paper concludes with a summary of the results (Section 7).

2. CONCEPTS AND RELATED WORK

2.1. Logistic Regression Probes

Logistic regression probes, in this work abbreviated as probes, are a tool proposed by Alain and Bengio [12] used as a "thermometer-style"-scalar metric for analyzing the intermediate solution quality during the forward pass. They are obtained by training a simple logistic regression model on the same task as the original neural network, however, using the layer's output values as input data for training. Hence, the probe performance can be considered as a measure of the linear separability of the target classes in the layer's output representations. As the neural network's softmax layer and the logistic regression probe both minimize cross-entropy, both solve effectively the same task. Therefore, we can use the test accuracy of the logistic regression relative to the model's predictive performance to judge the intermediate solution quality. The logistic regression probe performance should increase monotonically from early to later layers of the network, approaching the predictive performance of the model towards the final layers. Such a development implies that all layers contribute qualitatively to the inference process. Logistic regression probe performance is visualized as a curve with individual measuring points (network layers) arranged in the same order as the data flows through the network during a forward pass. An example of this can be seen in Fig. 1, which displays the probe performances measured on the convolutional layers of VGG19 trained at a low resolution on the ImageNette dataset. From the example we can observe how the intermediate solution quality measured by the probes improve from layer to layer until reaching the same level of accuracy as the model.

2.2. Saturation

The saturation s_l of a layer l is a simple scalar metric that was first introduced by Shenk et al. [16] and Richter et al. [13]. It can be computed for any layer in a neural network based on the layer's output values. It measures how many of the available dimensions in the output space Z_l of the layer l are relevant for the inference process:

$$s_l = \frac{\dim E_l^k}{\dim Z_l} \quad (1)$$

Saturation is computed by approximating the ratio of the dimensionality of the relevant eigenspace $\dim E_l^k$ of layer l and the extrinsic dimensionality of that layer's activation

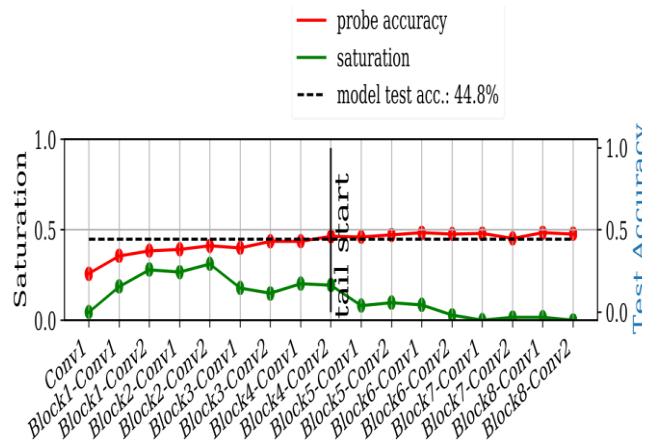


Fig. 1: An example of a tail pattern on a trained ResNet18 model. The tail is starting on the layer with the black border. Tail patterns can be identified by low saturation and stagnation of the logistic regression probe performance. Layers of the tail are no longer improving on the intermediate solution quality. For this reason, these layers can be considered a parameter-inefficiency. The model is trained on ImageNette at 32×32 pixel input resolution.

values $\dim Z_l$. The relevant eigenspace E_l^k is a subspace of Z_l in which the information is processed. This space is referred to as "relevant" because a projection of the data into the relevant eigenspace will not lead to a loss of predictive performance [13]. The relevant eigenspace can be considered the subspace in which the information processing is happening. The approximation of E_l^k is done using principal component analysis (PCA) [17], where the largest eigendirections are kept in order to explain 99% of the data's variance in the output of layer l . This technique allows computing saturation on-line during training. In contrast, evaluating logistic regression probes may take significant extra time, as it requires the additional training of the probes from a complete set of activation values, which can easily take more time than training the network. In this work, we use our implementation prepared in the context of the Delve-Framework [18].

2.3. The Semantics of Saturation

A sequence of low saturated layers (< 50% of the average saturation of all other layers) is referred to as a "tail pattern" and indicates that these layers are not contributing qualitatively to the prediction. The example in Fig. 1 displays the saturation values of VGG19 trained on ImageNette alongside the logistic regression probe performances extracted from the same layers. We observe that layers improving the logistic regression probe performances are significantly higher saturated than layers that do not improve the probe accuracy relative to the previous layer.

This suggests that solving a problem saturates the layer more than simply passing through information. However, this does not mean that the absolute saturation value is indicative of the activity within a layer. So far, saturation has been only explored as a quicker on-line computable alternative for logistic regression probes. As such, saturation has always been viewed relative to other layers within a neural network.

In this work, we will explore how the absolute saturation value changes in different scenarios. We further explore how saturation evolves during training, and we will explain the low intrinsic dimensionality observed in the tail pattern and we will explain the low intrinsic dimensionality observed in the tail pattern.

3. EXPERIMENT I: PRINCIPAL EIGENFEATURES

The results and experimental work of this paper heavily rely on the analysis of the eigenspace of neural network layers. Since neural networks are feature extractors, we refer to a principal eigendirection inside the feature space of a neural network layer as principal eigenfeature. We first demonstrate the effectiveness of principal eigenfeatures and their relation to the orthogonal feature space using an autoencoder.

3.1. Methods

We choose a convolutional autoencoder since the output is easy to visualize and differences in predictive performance are intuitive to understand with the human eye. The exact architecture of the autoencoder is depicted in [Table 1](#). We train the autoencoder for 30 epochs using the Adam optimizer [\[19\]](#) and a batch size of 128 images on the Food101 dataset [\[20\]](#); the hyperparameters can be seen in [Table 2](#). We also use random cropping, horizontal flipping, and random rotations for data augmentation purposes, to increase the difficulty of the reconstruction.

3.2. Results

In [Fig. 2](#) we visualize a randomly chosen example from the test set. During training time, we evaluate the autoencoder as normal. However, during inference time, we only keep the k largest eigenfeatures that are needed to explain a percentage δ of the data's variance in that layer by using a linear retraction generated from the reduced k -dimensional eigenspace E^k using the following formula: $P_{E^k} = (E^k)^T E^k$. By choosing various values for δ , we can observe the ablation caused by the removal of eigenfeatures. As we can see in [Fig. 2](#), the images are recognizable until 99% explained variance. However, it is worth noting that even at 99.99% variance, 4374 of 8192 eigenfeatures were used in the bottleneck of the autoencoder. This is apparent by the dimensionality of the reduced eigenfeature space $E^{K_{enc}}$ of the encoding layer. At 99% variance, the principal eigenfeatures of the bottleneck layer are only 597-dimensional, demonstrating that over-parameterization results in underutilization of the feature space, even in the bottleneck of an autoencoder.

4. EXPERIMENT II: CAPACITY AND PROBLEM DIFFICULTY BEHAVE PROPORTIONALLY

In this section, we analyse the relationship between problem difficulty and model capacity in two experiments, exploring how this relationship is reflected in the saturation values. In our experiments, we train the entire VGG-network family (VGG11, 13, 16 and, 19) on Cifar10 [\[21\]](#) and reduce their capacity evenly over the entire architecture to observe how these reductions affect the saturation values. Our first hypothesis states that the average saturation s_μ increases proportionally with a reduction in capacity while the model performance decreases. We then move on to investigate how

Table 1: Convolutional Autoencoder.

| Encoder | Decoder |
|---------------------------------------|---------------------------------------|
| $512 \times 512 \times 3$ Input | (3×3) conv, 8 ReLU |
| (3×3) conv, 16 filters, ReLU | upsampling, nearest, scale-factor 2 |
| (2×2) max pooling, strides 2 | (3×3) conv, 8 filters, ReLU |
| (3×3) conv, 8 filters, ReLU | upsampling, nearest, scale-factor 2 |
| (2×2) max pooling, strides 2 | (3×3) conv, 16 filters, ReLU |
| (3×3) conv, 8 filters, ReLU | upsampling, nearest, scale-factor 2 |
| (2×2) max pooling, strides 2 | (3×3) conv, 3 filters, ReLU |

Table 2: Hyperparameters for the convolutional autoencoder.

| Parameter | Parameter |
|---------------------|------------------|
| Input Resolution | 224×224 |
| Epoch | 50 |
| Batch size | 128 |
| Optimizer | Adam |
| Adam: beta1 | 0.9 |
| Adam: beta2 | 0.999 |
| Adam: epsilon | $1e-8$ |
| Adam: learning rate | 0.0001 |

the problem difficulty changes the saturation emerging in a neural architecture. Since the relevant eigenspace is generally larger when the layer is contributing to the quality of the solution [\[13\]](#), we further hypothesize that more processing in a layer requires a larger relevant eigenspace. If this assumption holds true, the overall saturation level should increase with an increase in the difficulty of the task. If both working hypotheses are true, we can conclude that the difficulty of the problem and the capacity of the layers influence saturation in an antagonistic way.

4.1. Methodology

We test our working hypotheses by conducting two experiments. We first train the VGG-family of networks on Cifar10. We further train 4 additional variants of each model, which have the respective number of filters (and thus capacity) reduced by a factor of 12, 14, 18 and 116. We choose Cifar10 for its manageable size, which allows for a larger number of model training runs to be conducted with our available resources, which is necessary for this experiment. We choose the VGG-family of networks for its architectural simplicity and because we can test different depths of convolutional neural networks by experimenting on the entire family of networks. The training itself is conducted using a stochastic gradient descent (SGD) optimizer with a learning rate of 0.1, which is decaying after 10 epochs with a decay factor of 0.1. The models are trained on a batch size of 64 for 30 epochs in total.

The second experiment is conducted on ResNet18. However, we are using a standardized input resolution of

224×224 pixels, to avoid artifacts caused by the input resolution. We train the model on multiple datasets of different difficulties (in ascending order of complexity): MNIST, Cifar10, TinyImageNet, and the ImageNet dataset [21-24]. While it is hard to precisely define the complexity of the task, we think that the selected datasets can be regarded as increasingly difficult based on the number of classes and the complexity of the visual information provided as data points to the model. The resolution of MNIST binary images is 28×28 pixel. That is suitable for the 10-class classification problem. Cifar10 comprises RGB images with a 32×32 resolution. That is suitable for the 10-class classification problem as well. TinyImageNet consists of RGB images of size 64×64 with 200 classes, and ImageNet is made up of RGB images of various sizes belonging to 1,000 classes.

4.2. Results

When the capacity of the model is reduced, the average saturation s_μ increases, and the predictive performance decreases. The exponential reduction in capacity is reflected in a logarithmic relation between the increasing s_μ and predictive accuracy measured on the test set (see Fig. 3). From these observations, we can conclude that reducing the capacity of the architecture results in an increase in saturation. We further observe in Fig. 3 that saturation also increases with problem complexity. The saturation levels of all layers increase when the model is trained on a more difficult problem. The overall shape of the saturation curve only deviates slightly, with no tail pattern or similar anomalous shapes emerging. Since we know from the works of [13] that a resolution of 224×224 pixels results in an even

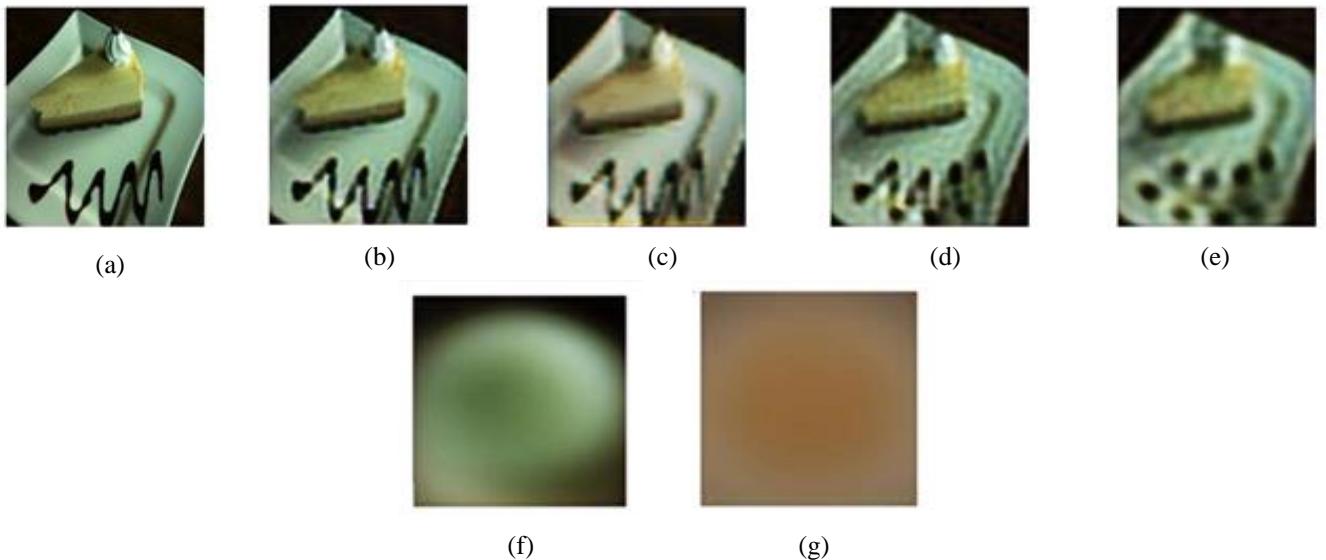
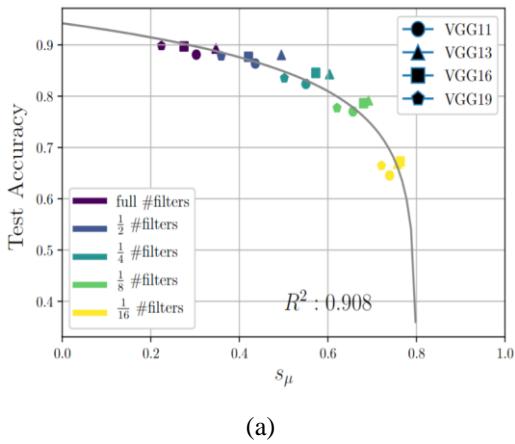
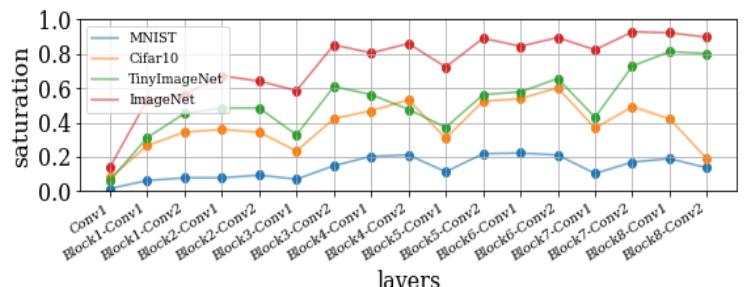


Fig. 2: Reconstructions of a single sample image, with the network being restricted to a percentage of its eigenfeatures: (a) Original: $\text{dimEncoding} = 8192$, (b) $\delta = 99.99\%$: $\text{dimE}^K_{\text{enc}}=4374$, (c) $\delta = 99.9\%$: $\text{dimE}^K_{\text{enc}}=1626$, (d) $\delta = 99.5\%$: $\text{dimE}^K_{\text{enc}}=1332$, (e) $\delta = 99.0\%$: $\text{dimE}^K_{\text{enc}}=59$, (f) $\delta = 95.0\%$: $\text{dimE}^K_{\text{enc}}=17$, and (g) $\delta = 90.0\%$: $\text{dimE}^K_{\text{enc}}=1$. Note how the visualized reconstructions degenerate with decreased explained variance,



(a)



(b)

Fig. 3: Relationship of network saturation to model capacity and data complexity, (a) Accuracy and saturation with varying model capacity, and (b) ResNet18 saturation curves for different datasets. Reducing the number of filters and thus reducing the model capacity leads to an increase in the average saturation and a decrease in performance. Training a model on more difficult datasets also increases the overall saturation level. This indicates that saturation can measure the load on a ResNet18 model.

distribution of the inference process for the trained model, we can conclude that less processing is required for less complex problems. Combined with the insights gained from training the VGG-variants on Cifar10, we can conclude that for the pairs of dataset and model in our experiments, a saturation “sweet spot” exists between $s_u = 0.2$ and $s_u=0.4$, which yields good predictive performance without being too excessively over-parameterized. This sweet spot allows us to empirically formulate the algorithm for optimizing the network width proposed in our previous work [13]. This algorithm is depicted in Fig. 4. Since the experiment suggests a roughly linear relationship between the saturation values and the width scaling, the scaling parameter can be approximated from the average saturation.

5. EXPERIMENT III: ON THE EMERGENCE OF SATURATION PATTERNS

The tail pattern that we discussed earlier in this work allows for the identification of inefficiencies caused by mismatches between the neural architecture and the input resolution. However, since saturation can be computed life during training with little overhead [13], we think that it might be interesting to see how these patterns emerge during the training process.

5.1. Methodology

We first examine how the saturation levels evolve in a layer under different conditions. We train a set of multilayer perceptrons (MLP) with 3 fully connected layers. The first layer has 256 units, and the size of the second layer varies for each network, being in the range of 8, 16, 32, 64, 128, 256, 512, and 1024 units. We train these networks using the Adam optimizer and a batch size of 128 on Cifar10 using the native resolution of the dataset. The training is conducted twice. Once using 8 epochs, which is enough for all models to converge, whereas the second experiment is run for 30 epochs, which results in the loss increasing again due to overfitting. The hidden layer saturation is calculated after each epoch for observing the evolution of the architecture. We also calculate the cross-entropy loss of the model to observe a possible relationship between loss and saturation convergence. Based on these observations, we repeat the experiment on VGG11 and VGG19 as well as on sparse (low capacity) versions of these models with 1/8 of the original number of filters. We do this to understand if saturation patterns depend on the depth, architecture, and capacity of the network.

5.2. Results

In Fig. 5, we observe that an increase in the number of units in the fully connected layer will result in a decreased saturation. However, the saturation does not change substantially during training, indicating that the inference process is not changed or shifted substantially inside the layer.

In Fig. 6, we can see that the increase in validation loss does not affect saturation. The fact that overfitting is not reflected in saturation values indicates that the changes to the way the data is processed when the model starts to overfit are subtle and thus are not reflected in changes to the relevant

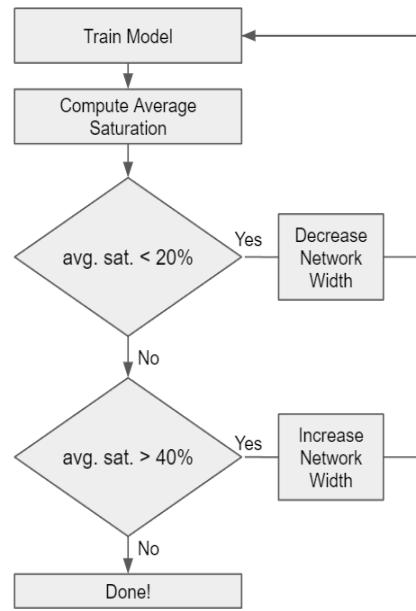


Fig. 4: This flow-chart depicts the basic procedure of optimizing the width of a neural architecture based on the average saturation of the model. The width of the network is increased to decrease saturation and vice versa until the model has an average saturation in the “sweet-spot”-range of 20-40%.

eigenspace and therefore saturation. This also means that saturation patterns in fully connected networks are independent of the training progress, which could allow for early detection of over- and under-parameterization during training. However, it also means that overfitting and convergence of the model need to be taken into consideration when analyzing saturation on fully connected neural networks, as these are not reflected by the saturation patterns.

In Fig. 7, we can see that saturation behaves substantially differently in convolutional neural networks, which exhibit a converging behavior towards a final pattern. This converging behavior is independent of the position of the layer in the network, the number of layers, and the capacity of the network, as Fig. 7 illustrates. Another interesting observation is that the tail pattern seems to be observable rather early during training, which indicates that an online analysis during training allows the data scientist to detect inefficiencies early, before the training has concluded.

6. EXPERIMENTS IV: PREDICTABILITY OF TAIL PATTERNS REGARDING COMPLEXITY

In the following, we examine how overall saturation affects the predictability of tail patterns. Richter et al. [14] show that the tail patterns in sequential convolutional neural networks can be predicted by computing the receptive field of all convolutional layers. The receptive field can be considered the field of view of a convolutional layer. Everything contained in the area spanned by the receptive field can hypothetically influence the value on a single position on the output feature map. In Section 5, we showed that changing the number of filters in a convolutional layer results in the changing of the global saturation level.

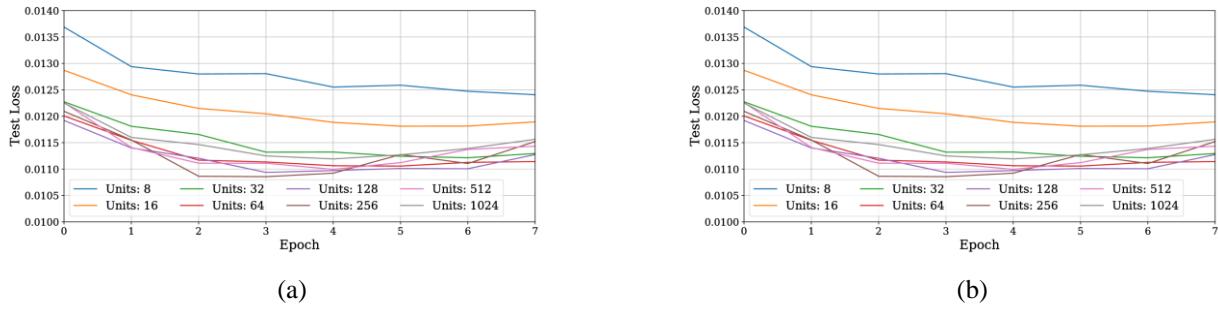


Fig. 5: (a) Saturation of layer 2 during training, and (b) Validation loss during training. Saturation of a 3-layer MLP does not change substantially during training while the loss is converging.

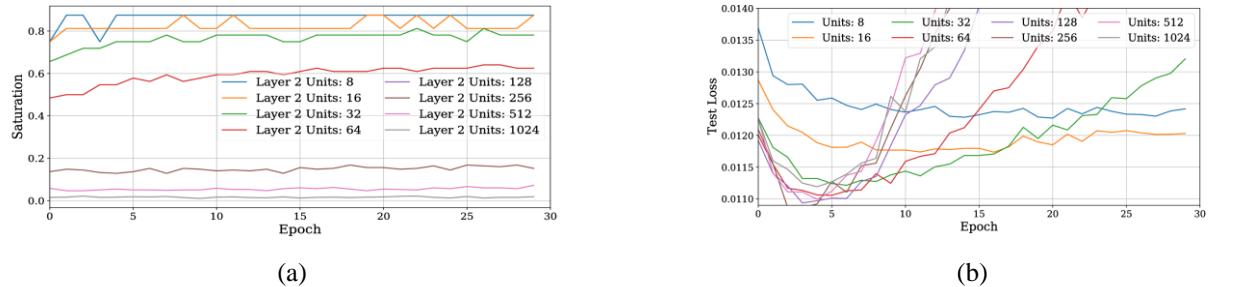


Fig. 6: (a) Saturation of layer 2 while overfitting, and (b) Saturation of layer 2 while overfitting. Saturation patterns of a 3-layer MLP are unaffected by overfitting, which indicates that overfitting is a process not affecting the overall dimensionality of the data inside the feature space.

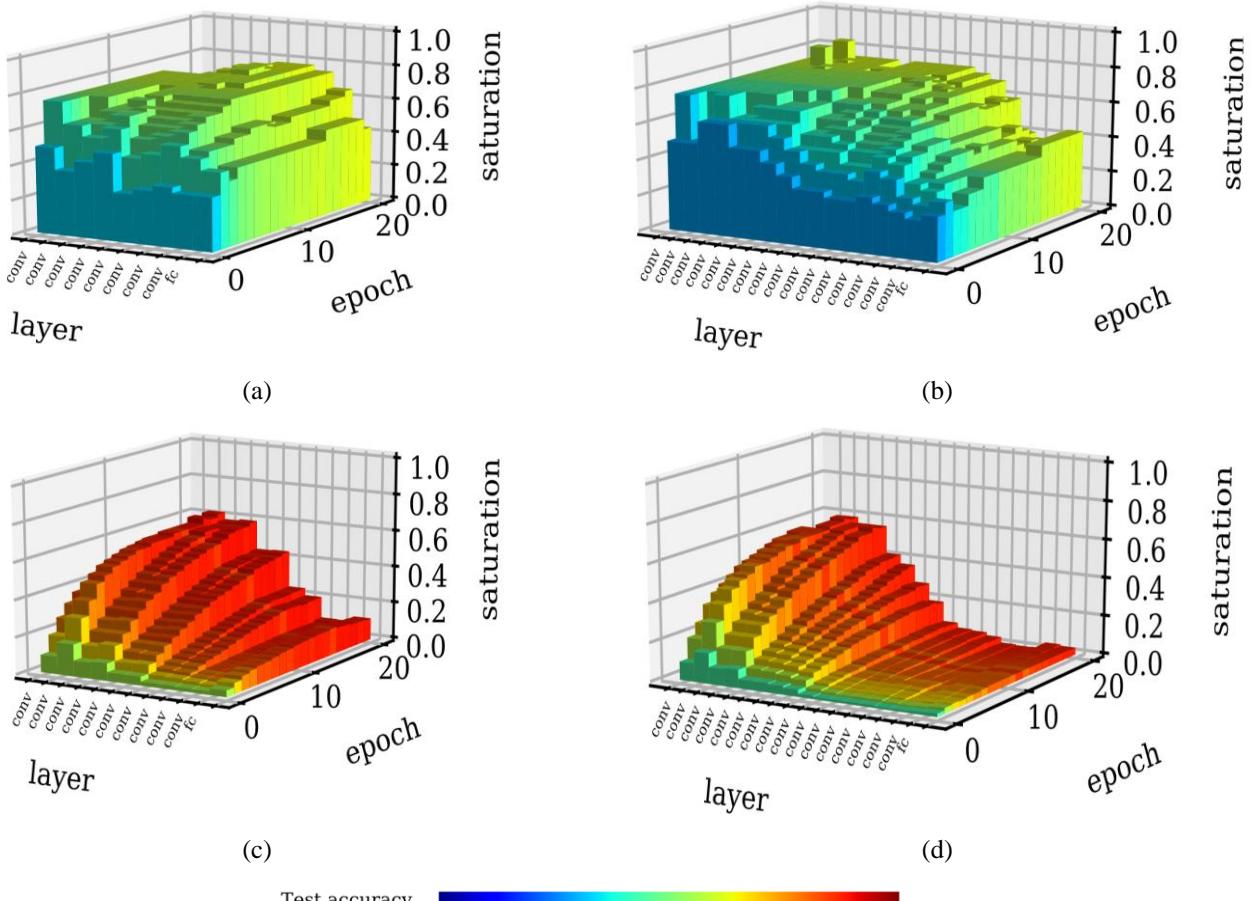


Fig. 7: Saturations of convolutional neural networks, (a) VGG 11, (b) VGG 19 (Sparse), (c) VGG 13, and (d) VGG 19. There is a converging behaviour regarding saturation in contrast to previous observations in Fig. 5.

However, if the receptive field expansion is determining the number of unproductive layers, we will observe a tail pattern of unproductive layers starting at the ‘border layer’ [14].

6.1. Methodology

We test the hypothesis by repeating the experiments conducted by [14] regarding the prediction of unproductive layers. The authors of [14] were able to predict unproductive layers by computing the border layer for VGG11, 13, 16, and 19 on Cifar10. We reduce the capacity of these models by reducing the filter size to 1/8 of the original size to see whether a drastic loss in capacity changes how the inference is distributed. The models are trained for 30 epochs using the SGD-optimizer with a learning rate of 0.1, decaying by a factor of 0.1 every 10 epochs. The batch size is 64, each batch is channel-wise normalized, each image is randomly cropped during inference time as well as randomly horizontally flipped with a probability of 50%. The receptive field and the border layer are computed using the formulas provided by [14].

6.2. Results

Even though the capacity of the networks has been significantly reduced in every layer, the networks do not spread the inference process among significantly more layers (see Fig. 8). Based on these results, we conclude that the inference dynamics of the tested networks did not change substantially by reducing their capacity. This means that the capacity of layers primarily interacts with the difficulty of the problem, while the presence and absence of tail patterns interact with the receptive field, as exemplified by [14].

6.3. Different Types of Tail Patterns - A Brief Explanation

We find that saturation is subject to noise induced by certain features of the neural architecture. The increase or decrease in a number of filters from layer to layer, the use of 11 convolutions and downsampling layers are common culprits for zig-zag-like behavior or sudden dips and spikes in saturation. An example for the latter is DenseNet18 in Fig. 9b. It has to be stressed that these factors are not random nor create non-reproducible perturbations. Instead, they usually result in anomalous patterns that are very stable over multiple runs (which is exemplified by [13]). Logistic regression probes are considerably more robust against the aforementioned properties. However, they are influenced by the path that the information takes during the forward pass, revealing different types of tail patterns that can be differentiated based on the processing in the tail-layers. The three examples found commonly are exemplified in Fig. 9. These examples also give insights into how neural networks process information differently, which is the main reason why we dedicate an additional section to these findings in this paper. All the networks are trained on Cifar10 using a 32x32 pixel input resolution. In Fig. 9a we find a pass-through tail, where the layers process the information but do not advance the quality of the intermediate solution. The second type of tail, depicted in Fig. 9b, is caused by the multiple pathways inside the DenseBlock of DenseNet. Information can pass from any previous layer to the current layer within the DenseBlock, effectively allowing the information to skip layers. When layers are skipped, the intermediate solution quality degrades and instantaneously recovers after the skipped section is over. The latter is apparent in the depicted example by the high model performance relative to the probe performance of the last DenseBlock layers. This phenomenon

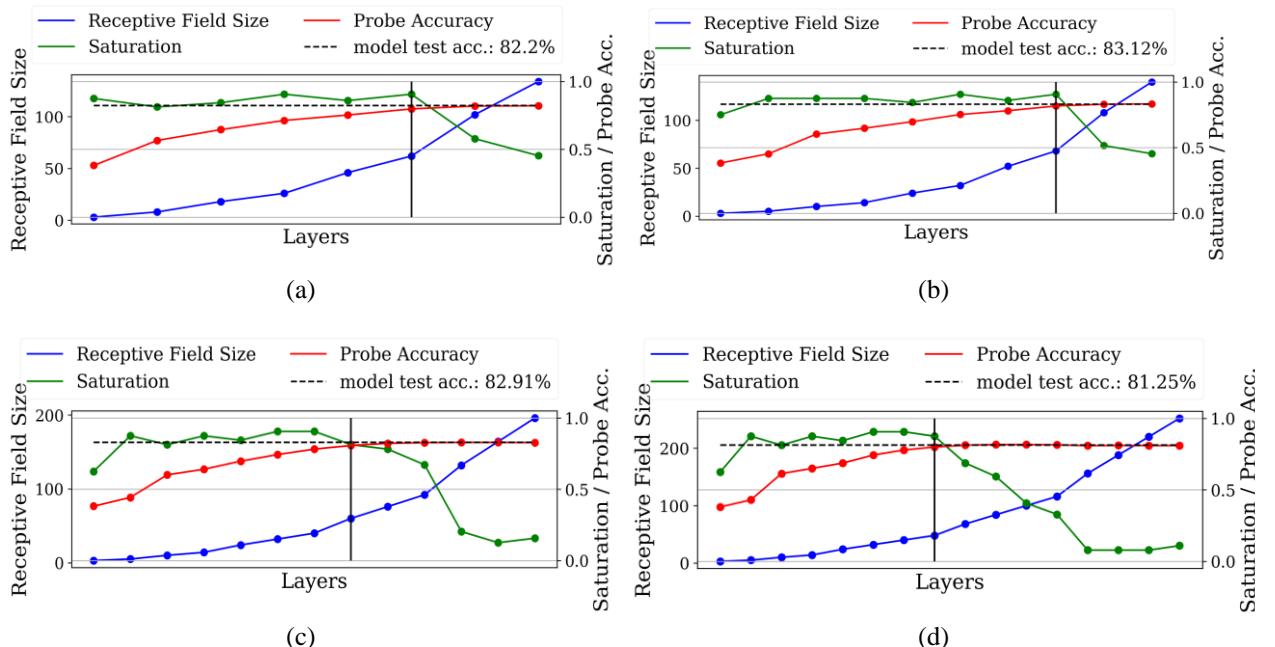


Fig. 8: Performance of the logistic regression probes past the border layer are minuscule, (a) VGG11 with $\frac{1}{8}$ filters per layer, (b) VGG11 with $\frac{1}{4}$ filters per layer, (c) VGG16 with $\frac{1}{8}$ filters per layer, and (d) VGG19 with $\frac{1}{8}$ filters per layer. The performance is improved even though the capacity of each layer is reduced to $\frac{1}{8}$ of the original capacity. This indicates that the networks are unable to shift processing to otherwise unused layers even if the capacity is limited. This is consistent with observations made by Richter et al. [14].

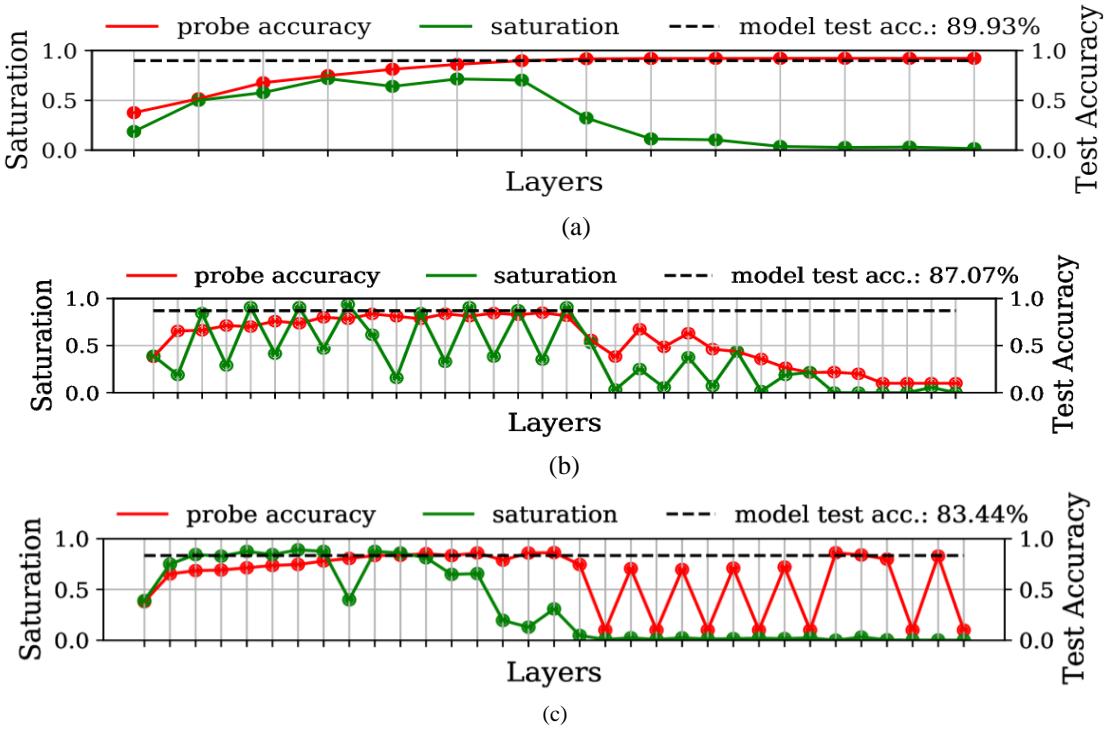


Fig. 9: (a) VGG16 tail layers maintain the quality of the intermediate solution, (b) The tail of DenseNet18 shows decay in probe performance, indicating that the last DenseBlock is skipped entirely [12], and (c) ResNet34 skips most residual blocks in the tail, which is apparent by the zig-zag pattern in probe performances caused by the starts and ends of skip-connections [12].

Depending on the neural architecture, tail patterns may deviate in their appearance in probe performance. In sequential architectures (a) the layers maintain the quality of the intermediate solution. If shortcut connections exist in the architecture, layers may be skipped. Skipped layers are apparent by their decaying probe performance [12]. This is apparent in DenseNet18 (b).

was initially observed in a simple MLP example by Alain and Bengio [12]. If necessary, the signal may jump more than a single building block in the architecture. An example of this can be seen in Fig. 9c on a ResNet34 architecture. This jumping is indicated by the zig-zag-pattern in the probe performance, where the higher performing layer resembles the first and the lower performing layer the second layer of a residual block. This shows that architecture decisions, which are influencing the potential pathways that the information can take from input to output, can have a significant influence on the way the model processes (or chooses not to process) information. In any case, the semantic of the tail-pattern remains unchanged, since a skipped layer and an unproductive layer can both be considered a parameter and computational inefficiency.

7. CONCLUSION

In this work, we explored the properties of the saturation metric in more detail and integrated this knowledge with insights from [13] and [14]. We have shown that model capacity and problem difficulty have opposite effects on the saturation value, as could be expected. A more surprising observation concerns the influence of individual layer capacity on the inference process: the tested models seem to be unable to shift processing to other layers, when some layers have substantially higher or lower capacity. An analysis of the evolution of saturation patterns during training revealed that they converge at a similar pace as the loss of the model, with saturation increasing during training. The way

saturation evolves also gives hints on the properties of the dataset, but it is not influenced by the model over-fitting.

These insights allow us to expand upon the optimization strategies for neural architectures, proposed in our previous work. We demonstrate quantitatively that the average saturation of a model is indicative of over- and under-parameterization. This allows us to adjust the width of the model effectively. We further show that this property is independent of the tail pattern. The tail pattern is a symptom of a different design flaw, related to the depth of the neural network. Hence, we show that multiple axes of neural network design (depth and width) can be optimized in an informed manner using saturation. Our optimization strategies still require one or multiple training runs of the model, which could be seen as a disadvantage compared to pruning techniques like PCA-pruning [9-11]. On the other hand, our approach is architecture-independent, which we demonstrate on multiple experiments, while also being able to detect and resolve under-parameterization. The latter cannot be addressed by pruning algorithms. Furthermore, we demonstrate that saturation converges early during training, greatly reducing the cycle time of experiments, since pathological inefficiencies can be diagnosed before training has finished. However, the current approach is still too noisy to allow narrowing design decisions on a layer-by-layer basis.

While we demonstrate that tail-patterns are similar for different types of architectures, some architectural properties like down sampling and skip-connections induce artefacts

into the saturation values, making a layer-by-layer analysis harder to read. Therefore, we seek to combine this approach with the analysis of the receptive field, which was shown to greatly impact the presence of tail patterns [14], to make the diagnosis of inefficiencies more precise and robust.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Mats Leon Richter: Software, Writing - original draft. **Leila Malihi:** Writing - review & editing. **Anne-Kathrin Patricia Windler:** Writing - review & editing. **Ulf Krumnack:** Writing - review & editing.

DECLARATION OF COMPETING INTEREST

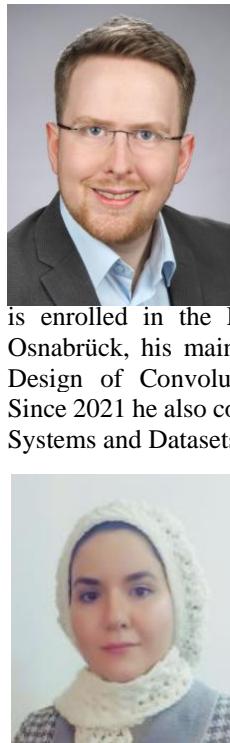
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy have been completely observed by the authors.

REFERENCES

- [1] M. Tan *et al.*, “MnasNet: Platform-aware neural architecture search for mobile,” CoRR, vol. abs/1807.11626, 2018. Available: <https://arxiv.org/abs/1807.11626>
- [2] M. Tan, and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: <http://proceedings.mlr.press/v97/tan19a.html>
- [3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CoRR, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [4] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” 2016. [Online]. Available: <https://arxiv.org/abs/1602.07261>
- [6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” CoRR, vol. abs/1512.00567, 2015. [Online]. Available: <https://arxiv.org/abs/1512.00567>
- [7] C. Szegedy *et al.*, “Going deeper with convolutions,” in Computer Vision and Pattern Recognition (CVPR), 2015. [Online]. Available: <https://arxiv.org/abs/1409.4842>
- [8] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, “SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [9] I. Garg, P. Panda, and K. Roy, “A Low Effort Approach to Structured CNN Design Using PCA,” *IEEE Access*, vol. 8, pp. 1347–1360, 2020.
- [10] W. Ahmed, S. Ansari, M. Hanif, A. Khalil, “PCA driven mixed filter pruning for efficient convNets,” *PLoS ONE*, vol. 17, no. 1, article e0262386, 2022.
- [11] I. Chakraborty, D. Roy, I. Garg Constructing, A. Ankit, and K. Roy “Energy-efficient mixed-precision neural networks through principal component analysis for edge intelligence”, *Nature Machine Intelligence*, vol. 2, pp. 43–55, 2020.
- [12] G. Alain, and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” ICLR 2017 workshop submission. [Online]. Available: <https://arxiv.org/abs/1610.01644v4>
- [13] M. L. Richter, J. Shenk, W. Bytner, A. Arpteg, and M. Huss, “Feature space saturation during training,” in *32st British Machine Vision Conference (BMVC)*, 2021.
- [14] M. L. Richter, W. Bytner, U. Krumnack, L. Schallner, and J. Shenk, “(Input) size matters for CNN classifiers,” in *Artificial Neural Networks and Machine Learning – ICANN 2021*. Springer International Publishing, 2021.
- [15] M. L. Richter, L. Malihi, A.-K. P. Windler, and U. Krumnack, “Exploring the properties and evolution of neural network eigenspaces during training,” in *2022 International Conference on Machine Vision and Image Processing*, 2022.
- [16] J. Shenk, M. L. Richter, A. Arpteg, and M. Huss, “Spectral analysis of latent representations,” 2019. [Online]. Available: <http://arxiv.org/abs/1907.08589>
- [17] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *Journal of Educational Psychology*, vol. 24, no. 6, pp.417–441, 1933.
- [18] J. Shenk, M. L. Richter, W. Bytner, M. Marcinkiewicz “Delve: Neural Network Feature Variance Analysis”, *Journal of Open Source Software*, vol. 7, no. 69, article 3992, 2022.
- [19] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations*, San Diego, CA, USA, May 7-9, 2015.
- [20] L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – mining discriminative components with random forests,” in *Computer Vision– ECCV 2014*, Springer International Publishing, 2014, pp. 446–461.
- [21] A. Krizhevsky, “Learning multiple layers of features from tiny images,” MIT and NYU, Tech. Rep., 2009.
- [22] Y. LeCun, and C. Cortes, “MNIST hand written digit database,” 2010. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [23] Y. Le and X. Yang, “Tiny ImageNet Visual Recognition Challenge,” 2015. [Online]. Available: http://cs231n.stanford.edu/reports/2015/pdfs/yle_projet.pdf

- [24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009.

BIOGRAPHY



Mats Leon Richter studied in the University of Osnabrück, where he received a Bachelor in Cognitive Science in 2017, a Bachelor in Computer Science in 2018 and a Master in Cognitive Science in 2019 with distinction. He worked full-time in the industry between 2017 and 2020 while pursuing his respective studies. Since 2019 Mats is enrolled in the PhD program in the University of Osnabrück, his main topics are Explainable AI and the Design of Convolutional Neural Network Classifiers. Since 2021 he also conducts research on Face Recognition Systems and Datasets as part of the KLIX Project.

Leila Malihi studied at the University of Shahid Chamran, Khuzestan, Iran, where she received a Bachelor and Master in Electrical engineering. Malihi's research interests include deep learning, transfer learning, and feature visualization. She has published 10 papers in Journals

and conferences. She has 4 years of industrial experience and collaborations. Since 2021 she has also conducted research on Face Recognition Systems and Datasets as part of the KLIX Project.



Anne-Kathrin Patricia Windler studies at the University of Osnabrück, where she received a Bachelor in Cognitive Science and is finishing her Master in Cognitive Science in 2022. Her research interests include deep learning, face recognition and transfer learning.



Ulf Krumnack studied computational linguistics, artificial intelligence, and mathematics and received his PhD for work on the logical formalization of analogical reasoning. He currently works as a postdoc in the group of biologically inspired computer vision at the institute of cognitive science, University of Osnabrück, Germany.

His research aims at the analysis of representations in deep networks, with a special focus on the visual domain.

Copyrights

© 2022 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Improving Stochastic Computing Fault-Tolerance: A Case Study on Discrete Wavelet Transform

Shabnam Sadeghi* , and Ali Mahani

Department of Electrical Engineering Shahid Bahonar University of Kerman, Kerman 76169133, Iran

* Corresponding Author: sadeghi.shabnam@eng.uk.ac.ir

Abstract: The stochastic computing (SC) method is a low-cost alternative to conventional binary computing that processes digital data in the form of pseudo-random bit-streams in which bit-flip errors have a trivial effect on the signal final value because of the highly redundant encoding format of this method. As a result, this computational method is used for fault-tolerant digital applications. In this paper, stochastic computing has been chosen to implement 2-dimensional discrete wavelet transform (2-D DWT) as a case study. The performance of the circuit is analyzed through two different faulty experiments. The results show that stochastic 2-D DWT outperforms binary implementation. Although SC provides inherent fault tolerance, we have proposed four structures based on dual modular redundancy to improve SC reliability. Improving the reliability of the stochastic circuits with the least area overhead is considered the main objective in these structures. The proposed methods are applied to improve the reliability of stochastic wavelet transform circuits. Experimental results show that all proposed structures improve the reliability of stochastic circuits, especially in extremely noisy conditions where fault tolerance of SC is reduced.

Keywords: Stochastic computing, fault-tolerant computation, image processing, discrete wavelet transform.

Article history

Received 28 September 2021; Revised 19 January 2022; Accepted 25 February 2022; Published online 10 December 2022.
 © 2022 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

S. Sadeghi, and A. Mahani, "Improving stochastic computing fault-tolerance: A case study on discrete wavelet transform," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 11-18, 2023. DOI: [10.22055/jaree.2022.38634.1036](https://doi.org/10.22055/jaree.2022.38634.1036)



1. INTRODUCTION

When the size of electronic components shrinks to nanoscale dimensions, the number of expected errors in a circuit increase, and since most error tolerance methods in binary systems increase cost, power consumption, and area, modern computing methods that directly address reliability issues should be considered an alternative to conventional computing. Recently, stochastic computing (SC) has gained attention due to its fault-tolerant capacity and less area requirement as these features are very much attractive for nano-scale CMOS technologies [1]. Therefore, the SC method is used in this paper as a fault-tolerant method.

In SC, a number x is encoded by a random bit stream of 0s and 1s with equal weight for every bit. Irrespective of the length, the ratio of the number of 1s to the length of the bit-stream, i.e., $P(X = 1)$, determines the data value. Error tolerance in stochastic circuits is based on the fact that the occurrence of a single bit error in a bit-stream of length N

changes the value of stochastic number (SN) by $1/N$ because all bits in a stream have the same weight. The larger the bit-stream length is, the more insignificant and smaller the change will be. For example, consider bit stream 00101010 containing three 1s denotes $x=p(X=1)=3/8$, a single bit-flip changes its value from $3/8$ to $4/8$ or $2/8$, which are the representable numbers closest to the correct result. But, if we consider the same number $3/8$ in conventional binary format 0.011, a single bit-flip causes a huge error if it affects the most significant bit. A change from 0.011 to 0.111, for example, changes the result from $3/8$ to $7/8$.

Additionally, multiple bidirectional errors (i.e., one-to-zero and zero-to-one conversion errors) may even cancel each other out, while the occurrence of errors in binary numbers changes the signal value according to the faulty bit weight [2].

In addition, arithmetic units of SC are very simple, so they have very low power requirements. For instance, the multiplication of two N -bit SNs A and B to form the arithmetic product $A \times B$ can be performed using a single AND gate in

N clock cycles, so SC permits complex computations to be realized using low-cost units in terms of hardware complexity. Fig. 1 shows an example of multiplying two input values, $1/4$ and $3/4$.

SC provides an accuracy-energy trade-off. Both the accuracy and the energy consumption of the circuit increase with the length N of the stochastic numbers [3], which introduces one of the major drawbacks of SC, i.e., long computing time. While a long stochastic stream may introduce long computational latency, parallel computations can be massively performed [4]. Another new solution to reduce the response time is the SC implementation scheme based on a memristive system [5].

Due to its unique features, the SC method is used to implement randomized algorithms and applications that require large amounts of data. Since small fluctuations are tolerable in such applications but many errors are ruinous, they are suitable to be implemented by stochastic logic [6].

Since image processing operations face severe design limitations in terms of power and area and they do not require high precision [7], several circuit designs have been proposed for different image processing applications in [6,8], including edge detection and gamma correction, which shows that stochastic designs can be significantly smaller, more power-efficient, and noise-tolerant.

One of the most important applications of SC is the implementation of LDPC decoders that mainly require a large number of parallel, fast, and relatively simple operations. In [9], SC has been used to implement an LDPC decoder in that low power requirements, error tolerance, and probabilistic aspects of SC have been exploited to achieve high power efficiency and throughput.

Another area in which SC has been applied with significant progress is artificial neural networks. Due to the resemblance of spike sequences in a biological neural network to stochastic numbers, implementing artificial neural networks continues to be a major application of SC [10]. In [11], authors have shown that stochastic neural networks (SC NN) achieve better area overhead and power consumption than state-of-the-art works by slightly sacrificing accuracy. With recent improvements in SC, the result of SC NNs have become comparable with conventional NNs.

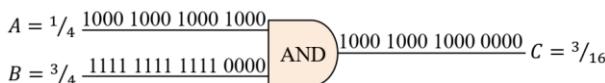


Fig. 1: An example of multiplication using stochastic logic.

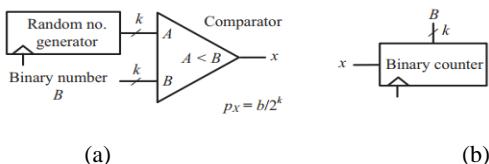


Fig. 2: (a) The binary-to-stochastic and b) stochastic-to-binary converters.

Since hardware implementation of low-cost and fault-tolerant architecture has recently received tremendous attention in modern IC applications, we utilized SC as a fault-tolerant method to implement 2-dimensional discrete wavelet transform (2-D DWT) and improved its reliability by four proposed methods based on error correction and dual modular redundancy.

The remainder of this paper is organized as follows: A brief overview of SC is given in Section 2. The architecture to implement stochastic 2-d wavelet transform is presented in Section 3. Section 4 provides proposed methods to enhance the reliability of stochastic circuits. The proposed methods are implemented and their performance is compared in Section 5, and finally, some conclusions are given in Section 6.

2. STOCHASTIC COMPUTING

In the SC method, operations occur on bit-streams that are interpreted as probabilities. The value of each bit-stream in the unipolar coding format is expressed as the probability of seeing a 1 along the bit-stream. For example, an N-bit stochastic number (SN) X containing N_1 1s and N_0 0s has the value $x=p(X=1)=N_1/N \in [0,1]$. Since the probability of seeing a 1 is a value between 0 and 1, this encoding format is used to display unsigned numbers. In addition to the unipolar format, several alternative SN formats have been proposed in [7] one of which is the bipolar encoding format that deals with the positive and negative numbers in the range [-1,1]. In the scenario of bipolar coding format, the relationship between x and $P(X=1)$ becomes $P(X=1)=(x+1)/2$, which enables the stochastic representation for negative numbers and the stochastic value is defined as $P(X=1)=(N_1-N_0)/N$. Notice that for either unipolar or bipolar coding format, the represented number ranges in [0, 1] or [-1,1]. To represent a number beyond this range, a pre-scaling operation or integer bit-stream-based representation [12] can be used to overcome this limitation. Stochastic circuits consist of three main parts: the stochastic number generator (SNG) of the kind shown in Fig. 2a, which produces a stochastic number. It consists of a pseudo-random number source such as a linear feedback shift register (LFSR) and a magnitude comparator and converts an unsigned k -bit binary number B to an N -bit stochastic bit-stream X . The comparator produces a 1 if the random number is less than B and otherwise a 0. The central unit contains the conventional logic gates and processes the bit-streams, and the last unit converts the output bit-stream to binary values. Fig. 2b shows the structure of the stochastic-to-binary (S2B) converter [7].

Figure 3 illustrates some basic computing units of SC, including multiplication, normalized subtraction, and addition. For instance, multiplication can be performed with an XNOR gate in bipolar coding format since $c = 2P(C=1)-1 = 2(P(A=1)P(B=1) + P(A=0)P(B=0))-1 = (2P(A=1)-1)(2P(B=1)-1) = ab$. Another example is addition, which can be simply implemented with a multiplexer in the SC method for $c = (C=1) = 1/2((A=1) + 1/2(B=1)) = 1/2(a+b)$. Additionally, the addition in the bipolar form uses this multiplexer as well since $c = 2P(C=1)-1 = 2(1/2(P(A=1) + 1/2P(B=1))-1 = 1/2(2P(A=1)-1) + (2P(B=1)-1)) = 1/2(a+b)$, and finally subtraction is easily implemented by

combining a multiplexer and a not gate. Normalization in addition and subtraction aims to make sure that the result is in the range [0,1] or [-1,1] so that it can be treated as a probability [6].

Besides the basic operations of addition and multiplication, the stochastic mean circuit and absolute value subtraction are presented in [13]. SC has also been applied to division [14] and some trigonometric and polynomial arithmetic functions [15, 16].

3. PROPOSED STOCHASTIC CIRCUIT DESIGN

3.1. Wavelet Transform Overview

Due to the wide application of wavelet transform and its computational complexity, the study of VLSI implementation of discrete wavelet transform (DWT) has become significant and unavoidable. A small and straightforward architecture will be advantageous, especially in image processing applications. Furthermore, since error and noise rates are high in most image processing applications, error resistance is also essential for these structures. For example, a fault-tolerance method is discussed in [17] to deal with silent data corruption errors on DWT. Therefore, having the advantages of SC, an error-resistant low-cost design is presented for two-dimensional DWT.

Wavelet transform is a mathematical tool that can decompose signals into different sub-bands of well-defined time-frequency characteristics. This conversion uses various methods to analyze the signal and adjust the accuracy in both time and frequency domains [18]. These benefits lead to extensive use of DWT in different areas such as medicine for processing medical images and diagnosing disorders using a computer [19], data transmission through the internet [20], and noise detection in data collected with a sensor [21].

Compared with the standards JPEG and JPEG-LS, the JPEG2000 standard not only offers a superior image compression ratio but also benefits from better image reconstruction performance [18]. One of the most widely used types of DWT is the 5/3 method used in the JPEG2000 standard to implement lossless image compression. The 5/3 wavelet transform can be implemented by using mathematical notations as follows [22]:

$$H(n) = -\frac{1}{2}[X(2n) + X(2n + 2)] + X(2n + 1) \quad (1)$$

$$L(n) = \frac{1}{4}[H(n - 1) + H(n)] + X(2n) \quad (2)$$

where $H(n)$ and $L(n)$ represent the high-frequency (detail coefficients) and low-frequency (approximation coefficients) components of the input signal, respectively and $X(n)$ represents the n th input sample.

The data flow related to these equations is shown in Fig. 4. To implement them in binary mode, we need a structure similar to Fig. 5. As shown in Fig. 5, this structure requires four adders and two multipliers for each decomposition level.

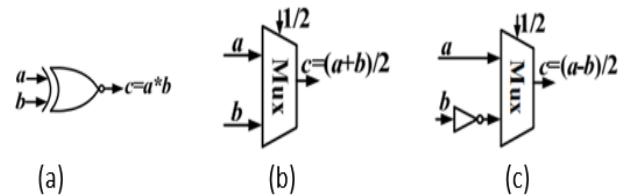


Fig. 3: a) The multiplication of bipolar format, b) scaled addition, b) scaled subtraction of bipolar format.

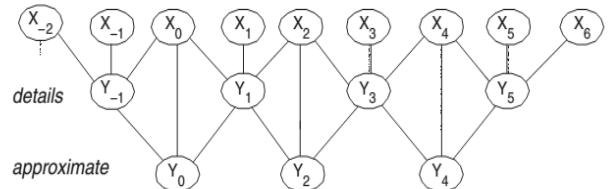


Fig. 4: The data flow of 5/3 wavelet transform.

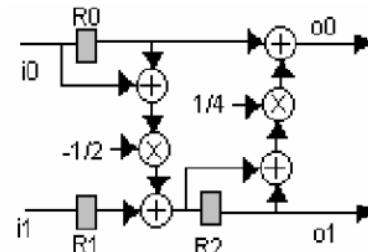


Fig. 5: The one-dimensional structure of binary DWT.

3.2. Stochastic Architecture for the Lifting Structure

As mentioned earlier, operations occur on bit-steams that are interpreted as probabilities in SC. Since we are dealing with positive and negative numbers in wavelet transform, the bipolar coding format is used in this article.

The blocks used in the lifting-based DWT include multipliers and adders, which are easily implemented in the BP stochastic method using an XNOR gate and a two-way multiplexer. Because the coefficients in the 5/3 wavelet transform equations are all multiples of 0.5, they are implemented using the coefficient of 0.5 in the stochastic adder, and there is no need to use a stochastic multiplier (XNOR gate). Since stochastic addition and subtraction are normalized, the implementation of equations (1) and (2) by stochastic logic leads to equations (3) and (4).

$$H'(n) = \frac{-\frac{1}{2}[X(2n) + X(2n + 2)] + X(2n + 1)}{2} = \frac{H(n)}{2} \quad (3)$$

$$\begin{aligned} L'(n) &= \frac{\frac{1}{2}[H'(n - 1) + H'(n)] + X(2n)}{2} = \\ &= \frac{\frac{1}{2}\left[\frac{H(n - 1)}{2} + \frac{H(n)}{2}\right] + X(2n)}{2} = \\ &= \frac{\frac{1}{4}[H(n - 1) + H(n)] + X(2n)}{2} = \frac{L(n)}{2} \end{aligned} \quad (4)$$

By implementing equations (3) and (4) with stochastic addition blocks, the structure of Fig. 6 is obtained for a one-dimensional 5/3 wavelet transform.

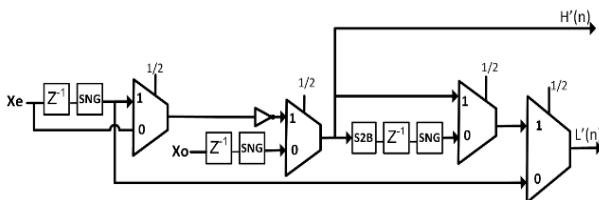


Fig. 6: The stochastic one-dimensional 5/3 wavelet transform structure.

In the proposed one-dimensional architecture, input signals pass through the delay elements (Z^{-1}) in 2's complement format and then each signal from the delay element is converted to a stochastic bit-stream. Each delay requires an 8-bit shift register instead of using a 256-bit delay element used in [23]. Also, for the intermediate delay element, an 8-bit memory element is used so the mux2 output is converted from stochastic to binary format before being stored in the delay element and then back to stochastic again. Since stochastic addition is correlation-insensitive [24] and only one LFSR has been used for these SNGs to reduce the hardware cost. In each clock cycle, one bit of high-frequency output coefficients and one bit of low-frequency output coefficients are generated. As it can be seen in Fig. 6, instead of using complex binary multiplier and adder blocks, only a multiplexer is used in the SC design.

To run wavelet transform on the images, it is necessary to apply the 1-D DWT in both vertical and horizontal directions of the image. The proposed 2-D architecture is based on three 1-D DWT structures, which operate in parallel and communicate through shared memory. The input image is fed to the architecture bit-by-bit using row-by-row scanning. In each clock cycle, a single bit is fed. In the row module, 1-D DWT of each row is computed to yield the low- and high-frequency components of each row, and the results are stored in memory to utilize in column transform.

To reduce the memory required, the row transform results are converted to binary format before being stored in memory. We employ shift registers for intermediate data storage. Low-frequency outputs and high-frequency outputs are stored in two distinct register files. Since three data are required to compute equation 3, the column transform process

starts as soon as the first data in the third row of memory is available.

Simultaneously with the column transform, the rest of the data required for the column operation are prepared by the row transform and overwritten in the memory. So, the capacity of the register file is considered $3 \times N/2$ coefficients for the $N \times N$ image.

The column operation begins by converting the stored binary numbers to stochastic ones, which can also eliminate autocorrelation [24] and improve the accuracy of stochastic computation, then performing a step of computation on these three rows in the column direction.

Two column transform modules are also employed and work in parallel to increase the design speed. One module performs the column direction wavelet transform on high-frequency coefficients of the row transform module and the other on low-frequency coefficients. Fig. 7 illustrates the stochastic 2-D 5/3 wavelet transform structure. The section marked with a dotted line in Fig. 7 shows the structure of a 1-D DWT. The other two modules perform the column transforms. Although this architecture is presented for the 5/3 case, the method can be applied to all lifting schemes that rely on single or double lifting steps.

4. METHODS TO IMPROVE FAULT TOLERANCE OF STOCHASTIC CIRCUITS

Although the SC method is inherently fault-tolerant, so far no action has been taken to improve its reliability and its error tolerance. Therefore, in this paper, four methods based on dual modular redundancy (DMR) are proposed to improve the reliability of stochastic circuits. Furthermore, since one of the significant advantages of SC is the small size of the stochastic circuits, in order to maintain this feature, the goal has been set to improve reliability with minimum area overhead in all proposed methods.

As the most important computational part of stochastic wavelet transform is multiplexers, these methods are only applied to multiplexers so that improved reliability does not lead to much hardware overhead. These methods are explained in the next subsections.

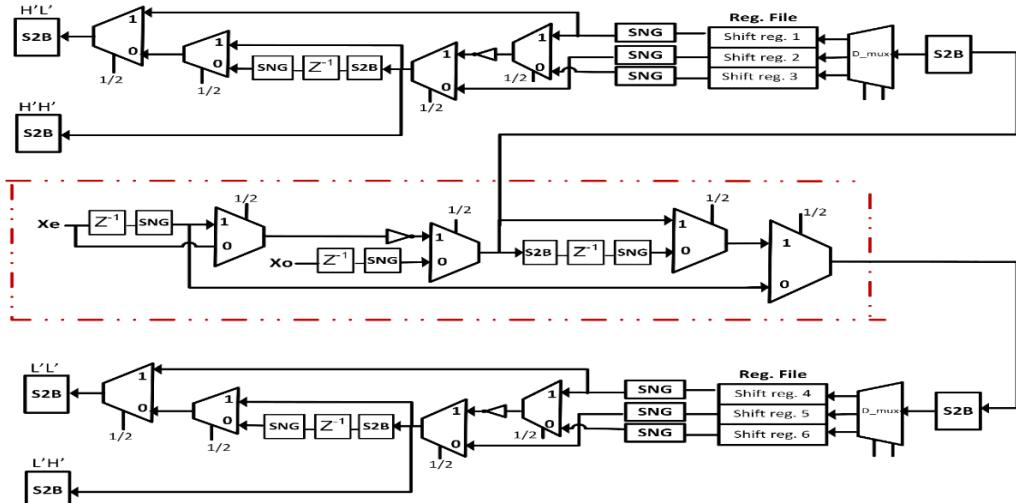


Fig. 7: Stochastic two-dimensional 5/3 wavelet transform structure.

4.1. Multiplexer-Based Error Correction

In the first method, DMR and multiplexer-based error correction structures are proposed to increase the reliability of stochastic circuits. Fig. 8 shows the structure of the error detection and correction circuit for a multiplexer. This multiplexer has a selection input with a constant value of 0.5 in which case 50% of the mux1 output and 50% of the mux2 output are randomly transferred to the final output.

4.2. C-Element Based Error Correction

The second method proposed to improve the reliability of stochastic circuits is based on DMR and a C-element. In this case, when the two main modules have an equally logical value, the same value is placed on the output, but if the two values are different, the output retains its previous value. The proposed circuit structure is shown in Fig. 9.

4.3. Error Correction Method based on Repeating the Operation for Faulty Bits

In this method, the multiplexers are placed in the form of DMR and an XOR gate is used for error detection. The clock signal in this circuit is blocked by an AND gate. When there is no error in the circuit, the output of the two multiplexers is equal and as a result, the XOR gate has a value of 0, which causes the next edge of the clock signal to be seen by the system and the operation to be performed on the next bit, but when an error occurs on one of the multiplexers, the output of the XOR gate becomes 1, the edge of the clock signal is not passed, and the operation is repeated on the same bit until the system has no error. Fig. 10 shows a part of the wavelet transform circuit optimized using this method.

Since the critical path in SC circuits is short, which means that it can be run with extremely high clock frequency [7], this method will be more efficient in stochastic circuits than in binary ones and the overall latency can be greatly mitigated by exploiting a special property of SC, which is called progressive precision [8] or by using parallel computational elements.

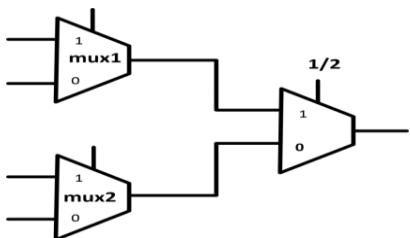


Fig. 8: The mux_based error correction circuit structure.

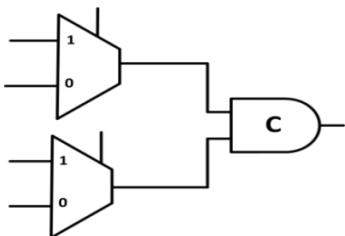


Fig. 9: The C_element-based error correction circuit structure.

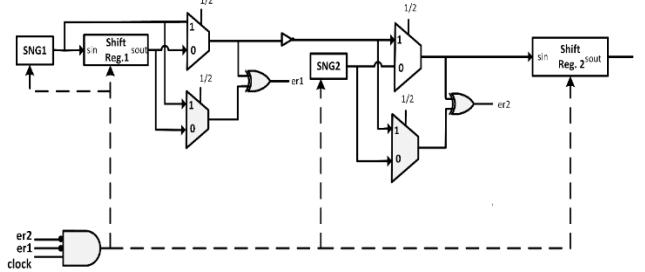


Fig. 10: The partial structure of a wavelet transform optimized by the fourth method.

4.4. Error Correction Method based on Stochastic Value of the Signal

As mentioned in Section 2, the value of each bit-stream in SC is defined as the number of 1s in the bit-stream divided by the total number of bits. In other words, the probability of occurrence of a logical one in the bit-stream indicates the stochastic number. Using this definition, the probability of the output bit of each module is clear. For example, the probability of output of the stochastic adder being one is obtained $P_y = \frac{1}{2}(P_{x_1} + P_{x_2})$ and the probability of being zero is $1 - P_y$. We suggest an algorithm to correct the bit-flip errors based on this SC feature. Consider an error occurs in the adder block. After the error detection to correct the error, that bit becomes one with the probability of P_y and zero with the probability of $1 - P_y$.

5. EXPERIMENTAL RESULTS

In this section, the results of implementing the proposed stochastic circuit designs are presented. These results include the estimated hardware cost and analyzed performance under two different faulty situations.

The classic Lena image was used as the input source of the wavelet transform circuit. The input is a 256×256 -pixel grey-scale image. Each pixel is represented by an 8-bit binary number. The length of the stochastic bit-stream is 256, which corresponds to 8-bit precision for conventional binary design.

The wavelet transform circuit presented in Section 3 was coded in VHDL and synthesized with Xilinx ISE on Virtex5 (XC5VLX110T) FPGA device. The results obtained in Table 1 were compared with the binary wavelet transform circuits [25-27].

As shown in the previous section, in the design of stochastic DWT, a multiplexer is only used instead of using complex multipliers and adder blocks, and as Table 1 shows, the proposed 5/3 stochastic DWT has a lower area than the other existing conventional binary architectures. Since new computing methods must be able to meet severe constraints such as very small size and low power consumption, SC can be a suitable alternative to conventional binary computing to design circuits that require fault-tolerant procedures on large amounts of data (e.g., various image processing operations).

Table 1: Area comparison.

| architecture | Device | Slice LUTs | Slice registers |
|--------------|-----------------|------------|-----------------|
| [25] | XC5VLX110T | 494 | 633 |
| [26] | XC5VLX110T | 9424 | 301 |
| [27] | 6VLX760FF1760-2 | 361/433 | 411/511 |
| stochastic | XC5VLX110T | 215 | 408 |

Table 2: The comparison of fault tolerance.

| Method | average error | | | | | |
|-------------------------|----------------|-------|-------|-------------------------------|-------|-------|
| | Noise at input | | | Noise at computational blocks | | |
| | 5% | 10% | 15% | 5% | 10% | 15% |
| SC method | 0.030 | 0.038 | 0.043 | 0.082 | 0.139 | 0.174 |
| Convention al method | 0.329 | 0.353 | 0.390 | 0.315 | 0.350 | 0.397 |

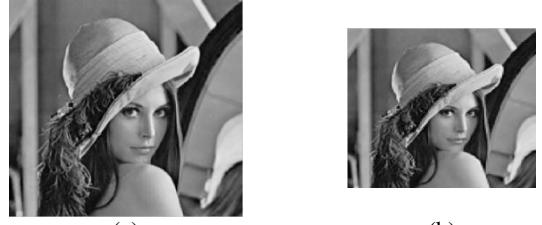
Fig. 11 is an image obtained from a stochastic 2-D DWT applied to the Lena image, which shows approximation coefficients for one decomposition level. The result of the wavelet transform processing by stochastic structure has some errors mainly due to the random fluctuations of SC [24] and the output MSE will be 0.0031. Since digital circuits are easily affected by manufacturing defects and transient errors, this inherent error is negligible compared to the high error rate in digital circuits.

To demonstrate the fault tolerance of the proposed architecture, we performed two kinds of experiments. In the first one, the inputs contain noise and in the second one, the circuit computational components are considered noisy and unreliable. Experiments were performed by injecting soft errors including flipping the bits by different injected noise ratios and evaluating the output average error in pixel values. **Table 2** shows these experimental results for conventional implementations compared to stochastic 2-D DWT.

As can be seen in **Table 2**, the stochastic method shows high resistance to bit-flip errors in both faulty experiments and gives acceptable results even in the case of 15% noise, but the performance of the binary method decreases dramatically with increasing error rate. Also, the average error drop rates for these experiments in the stochastic method are less than the binary one while the error rate is increasing from 5% to 15%.

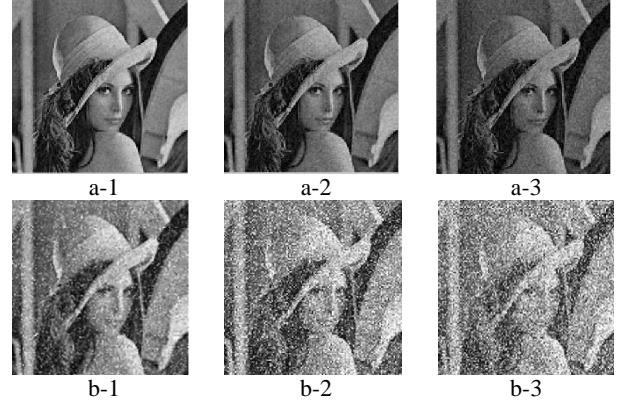
Fig. 12 illustrates the output image of the wavelet transform implemented by the stochastic method and the binary method with different noise ratios. As can be seen in **Fig. 12** when errors are injected at the rate of 15%, the image transformed by stochastic DWT is still recognizable and becomes greyer, while the image generated by the conventional method is full of noisy pixels.

In the following, the hardware area overhead and latency overhead are investigated for proposed error correction methods using stochastic DWT and to evaluate the robustness of our methods we analyze the effect of randomly injected errors by measuring the corresponding average output error for each implementation. Area overhead and latency overhead are given in **Table 3**. As in the third proposed method, the delay overhead changes according to the error rate, so the delay for this method is reported for each error rate in **Table 3**. The latency overhead is calculated for operations on 256 bits.



(a)

(b)

Fig. 11: The result of applying stochastic wavelet transform to Lena image, (a) Original image, (b) Approximate coefficients of stochastic DWT.**Fig. 12:** The output of (a) stochastic and (b) conventional DWT with noise level of (1) 5%, (2) 10%, and (3) 15%.**Table 3:** Area and latency overheads of error correction methods to stochastic DWT circuit.

| Method | Area overhead | Latency overhead |
|--|---------------|------------------|
| multiplexer-based error correction | 8.14 | 34.50 |
| C-element based error correction | 21.46 | 55.71 |
| Error correction method based on repeating the operation | 15.59 | Noise level |
| | 5% | 10% |
| | 15% | |
| | 25.75 | 31.85 |
| | | 37.95 |

Table 4 summarizes the performance of the proposed error correction methods at various noise levels of the inputs. As shown in **Table 4**, all the proposed methods improve the reliability of stochastic DWT according to average output error. In these methods, the average error is significantly improved and their efficiency increases with increasing error rate. In the multiplexer-based error correcting method, the least overhead (only 8%) is applied to the circuit and has acceptable error correction performance. The use of C_element to improve the reliability of stochastic circuits will improve average error so that it is more effective than the previous method but it has more latency and area overheads.

By the error correction method, which is based on repeating the operation, with an area overhead of 13.6%, errors are completely corrected. The latency overhead of this method varies according to the error rate, but the correction in the other methods is performed with constant overhead. Since the clock period of stochastic circuits is small, this method will be more efficient in stochastic designs than in conventional architectures. In addition, because of stochastic circuits' simplicity, parallel processing approaches and some methods based on the progressive precision property [28, 8] can be used to further reduce the latency.

Table 4: Fault tolerance comparison of error correction methods.

| Method | Improved average error (%) | | | | | |
|---|----------------------------|-------|-------|-------------------------------|-------|-------|
| | Noise at input | | | Noise at computational blocks | | |
| | 5% | 10% | 15% | 5% | 10% | 15% |
| Multiplexer-based error correction | 36.94 | 41.86 | 49.06 | 39.59 | 46.69 | 55.03 |
| C-element based error correction | 39.28 | 48.9 | 57.21 | 46.10 | 54.15 | 63.8 |
| Error correction method based on stochastic value of the signal | 53.25 | 58.73 | 66.19 | 60.78 | 66.09 | 71.99 |
| Error correction method based on repeating the operation | 100 | 100 | 100 | 100 | 100 | 100 |

The error correction method based on the stochastic value of the signal improves reliability better than other methods because it performs the correction according to the actual value of the signal, but the other methods work randomly. Similar to other methods, the efficiency of this method is improved by increasing the error rate, which makes these methods more suitable for environments with very high noise levels where the inherent fault tolerance of SC is reduced.

6. CONSULTATION

This study presents a fault-tolerant and low-area architecture for 2-D lifting-based DWT based on SC. The proposed architecture not only is much more tolerant of soft errors but also requires less area than the conventional implementation of this algorithm. To make the SC designs more robust to soft errors, we introduced four error correction methods based on DMR. Our experimental results show that the proposed methods had low hardware costs and all the proposed methods improved the reliability of the stochastic circuits according to average error. The remarkable note about all the proposed methods is that in all the methods, the performance of the proposed methods improves by increasing the error rate, which in turn makes these methods suitable for highly noisy environments where the inherent fault tolerance of SC is reduced. Future work will focus on how to generate low-cost probabilities to use in the error correction method based on the stochastic value of the signal.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Shabnam Sadeghi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Roles/Writing - original draft, Writing - review & editing. **Ali Mahani:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] V. V. Mahesh, and T. K. Shahana, "Design and synthesis of FIR filter banks using area and power efficient Stochastic Computing," in 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability, 2020, pp. 662-666.
- [2] T.-H. Chen, A. Alaghi, and J. P. Hayes, "Behavior of stochastic circuits under severe error conditions," *Information Technology*, vol. 56, pp. 182-191, 2014.
- [3] A. Alaghi, W. T. J. Chan, J. P. Hayes, A. B. Kahng, and J. Li, "Optimizing stochastic circuits for accuracy-energy tradeoffs," in 2015 IEEE/ACM International Conference on Computer-Aided Design, 2015, pp. 178-185.
- [4] K. Papachatzopoulos, C. Andriakopoulos, and V. Palioras, "Novel Noise-Shaping Stochastic-Computing Converters for Digital Filtering," in 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 2020, pp. 1-5.
- [5] T. Li, S. Duan, J. Liu, and L. Wang, "Memristive combinational logic circuits and stochastic computing implementation scheme," *Circuit World*, 2021.
- [6] P. Li, D. J. Lilja, W. Qian, K. Bazargan, and M. D. Riedel, "Computation on stochastic bit-streams digital image processing case studies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 3, pp. 449-462, 2014.
- [7] A. Alaghi, W. Qian, and J. P. Hayes, "The promise and challenge of stochastic computing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 8, pp. 1515-1531, 2018.
- [8] R. Seva, P. Metku, K. K. Kim, Y. -B. Kim, and M. Choi, "Approximate stochastic computing (ASC) for image processing applications," in 2016 International SoC Design Conference (ISOCC), 2016, pp. 31-32.
- [9] A. Hussein, M. Elmasry, and V. Gaudet, "On the fault tolerance of stochastic decoders," in 2017 IEEE 47th International Symposium on Multiple-Valued Logic (ISMVL), 2017, pp. 219-223.
- [10] J. P. Hayes, "Introduction to stochastic computing and its challenges," in *Proceedings of the 52nd Annual Design Automation Conference*, 2015, pp. 59.
- [11] Y. Liu, S. Liu, Y. Wang, F. Lombardi, and J. Han, "A survey of stochastic computing neural networks for machine learning applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 2809-2824, 2021.

- [12] M. Hasani Sadi, and A. Mahani, "Accelerating deep convolutional neural network base on stochastic computing," *Integration*, vol. 76, pp. 113-121, 2021.
- [13] M. H. Najafi, and M. E. Salehi, "A fast fault-tolerant architecture for sauvola local image thresholding algorithm using stochastic computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, pp. 808-812, 2016.
- [14] T. Chen and J. P. Hayes, "Design of division circuits for stochastic computing," in *2016 IEEE Computer Society Annual Symposium on VLSI*, 2016, pp. 116-121.
- [15] W. Qian, and M. D. Riedel, "The synthesis of robust polynomial arithmetic with stochastic logic," in *2008 45th ACM/IEEE Design Automation Conference*, 2008, pp. 648-653.
- [16] Y. Liu, and K. K. Parhi, "Computing complex functions using factorization in unipolar stochastic logic," in *2016 International Great Lakes Symposium on VLSI*, 2016, pp. 109-112.
- [17] Ch. Bao, and Sh. Zhang. "Algorithm-based fault tolerance for discrete wavelet transform implemented on GPUs." *Journal of Systems Architecture*, vol. 108, pp. 101823, 2020.
- [18] B.-F. Wu, and C.-F. Lin, "A high-performance and memory-efficient pipeline architecture for the 5/3 and 9/7 discrete wavelet transform of JPEG2000 codec," *IEEE Transactions on Circuits and Systems for Video Technoogy*, vol. 15, pp. 1615-1628, 2005.
- [19] R. Starosolski, "Hybrid adaptive lossless image compression based on discrete wavelet transform," *entropy*, vol. 22, no. 7, pp. 751, 2020.
- [20] S. H. Farghaly, and S. M. Ismail. "Floating-point discrete wavelet transform-based image compression on FPGA," *AEU-International Journal of Electronics and Communications*, vol. 124, pp. 153363, 2020.
- [21] T.-B. Dang, D.-T. Le, M. Kim, and H. Choo, "DWT-PCA combination for noise detection in wireless sensor networks." in *Proc. of the Korea Information Processing Society Conference*, 2020, pp. 144-146.
- [22] C. Xiong, J. Tian, and J. Liu, "Efficient architectures for two-dimensional discrete wavelet transform using lifting scheme," *IEEE Transactions on Image Processing*, vol. 16, pp. 607-614, 2007.
- [23] S. A. Salehi, and D. D. Dhruba, "Efficient Hardware Implementation of Discrete Wavelet Transform Based on Stochastic Computing," in *2020 IEEE Computer Society Annual Symposium on VLSI*, 2020, pp. 422-427.
- [24] A. Alaghi, P. Ting, V. T. Lee, and J. P. Hayes, "Accuracy and correlation in stochastic computing," Gross W., Gaudet V. (eds) *Stochastic Computing: Techniques and Applications*, pp. 77-102, 2019.
- [25] A. D. Darji, S.S. Kushwah, S. N. Merchant, and A. N. Chandorkar, "High-performance hardware architectures for multi-level lifting-based discrete wavelet transform," *EURASIP Journal on Image and Video Processing*, vol. 2014, pp. 47, 2014.
- [26] S. S. Bhairannawar, S. Sarkar, and K. B. Raja, "Implementation of fingerprint based biometric system using optimized 5/3 DWT architecture and modified CORDIC based FFT," *Circuits, Systems, Signal Process*, vol. 37, no. 1, pp. 342-366, 2018.
- [27] MR. Lone, "A high speed architecture for lifting-based 2-D Cohen-Daubechies-Feauveau (5, 3) discrete wavelet transform used in JPEG2000". *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, vol. 24, no. 9, pp. 9-24, 2017.
- [28] T. Chen, P. Ting, and J. P. Hayes, "Achieving progressive precision in stochastic computing," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 1320-1324.

BIOGRAPHY



Shabnam Sadeghi received her B.Sc. degree in electronic engineering from Shahid Bahonar University, Kerman, Iran in 2013. Since then, she was with the RSS Lab at Shahid Bahonar University for 3 years. Her research interests

are fault-tolerant designs, stochastic computing, and data analysis.



Ali Mahani received his B.Sc. degree in Electronic Engineering from Shahid Bahonar University, Kerman, Iran in 2001, and his M.Sc. and Ph.D. degrees both in Electronic Engineering from the Iran University of Science and Technology (IUST), Tehran, Iran in 2003 and 2009, respectively. Since then, he has been with the Electrical Engineering Department of Shahid Bahonar University where he is currently an associate professor. His research interests focus on fault-tolerant designs, FPGA-based accelerators, approximate digital circuits, stochastic computing, and networked systems.

Copyrights

© 2022 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Investigation of the Operation of Active Superconducting Fault Current Limiters in Distribution Networks Connected to Microgrids

Ahmad Ghafari* , Mohsen Saniei , Morteza Razaz , and Alireza Saffarian

Department of Electrical Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz 6135785311, Iran

*Corresponding Author: ghafari.ieee@gmail.com

Abstract: Increasing the penetration level of distributed generation (DG) units in radial power distribution systems can increase the short-circuit level in these networks, which can, in turn, have destructive effects such as exceeding the tolerable current of the equipment and disrupting the protective coordination in the network. The active superconducting fault current limiter (ASFCL) is a new device that can limit fault current using voltage series compensation. This paper discusses the modeling of ASFCL and control strategies including fault detection and converter performance in normal and fault modes. Initially, its performance in limiting the fault current is investigated by simulating a sample three-phase system with ASFCL. In the next step, three operating modes including normal mode, upstream fault mode, and downstream fault mode are proposed to achieve an adaptive FCL that solves these problems in grid-connected microgrids. The simulation results confirm the proper performance of the ASFCL modes in both fault current limiting and protective coordination of overcurrent relays in the network.

Keywords: Fault current limiter, active superconducting current controller, grid-connected microgrid, protective coordination.

Article history

Received 25 January 2022; Revised 20 March 2022; Accepted 09 April 2022; Published online 10 December 2022.

© 2022 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

A. Ghafari, M. Saniei, M. Razaz, and A. Saffarian, "Investigation of the operation of active superconducting fault current limiters in distribution networks connected to microgrids," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 19-25, 2023.

DOI: [10.22055/jaree.2022.39862.1046](https://doi.org/10.22055/jaree.2022.39862.1046)



1. INTRODUCTION

The recent growth of electrical energy demand and the rapid development of power systems have increased short-circuit phenomena, which can damage circuit breakers and other equipment. The deployment of current limiting equipment can be regarded as a useful solution for this issue [1]. Several studies have introduced and evaluated various types of FCLs. For example, the resistive, magnetic-shield, high-temperature superconducting, saturated iron-core, and shunt superconducting FCL types have been presented and examined [2-4]. These FCLs generally create a small impedance in the normal state and a large limiting impedance in the event of a fault.

When connecting a microgrid to the main grid, an FCL can be located between the upstream grid (main grid) and the downstream grid (micro-grid). The conventional type of FCLs generally performs a current limiting operation for both the upstream and downstream faults. Such operation of the conventional FCL can be useful when a short-circuit

fault occurs in the main grid, but during a fault in the microgrid, the limiting impedance of the FCL may distort the coordination between the upstream and downstream OCRs [5].

The active superconducting fault current limiter (ASFCL) is a new generation of series compensations that combines superconducting transformers and series voltage converters [6]. This type of superconducting FCL can limit current limiting different levels.

In [12-14], some functional modes have been defined for the ASFCL only with the aim of fault current limiting in the main grid. The performance of the functional modes defined in these papers may disrupt the coordination of existing OCRs in the network. In [16], an ASFCL has been used to limit the fault current and to coordinate the existing OCRs in the main grid. However, the defined functional modes in [16] are not used for connecting the microgrid to the main grid. Also, in this reference, the relay coordination method is performed by changing the setting parameters of all of the OCRs in the network.

In [4], a unidirectional fault current limiter (UFCL) has been used to maintain the coordination between the upstream and the downstream OCRs. For this purpose, the FCL is deactivated for the downstream fault state. However, this may cause problems if the fault current exceeds the tolerable range of microgrid devices.

The main contribution of this paper is the protective coordination of all OCRs in the main grid and microgrid by defining appropriate operating modes in the event of upstream and downstream faults and without changing the relay setting parameters. In fact, by applying appropriate limiting impedances in different states of the network including upstream fault, downstream fault, and normal mode, the fault current is controlled and the coordination of all OCRs is maintained without changing the setting parameters of OCRs. The simulation results obtained using MATLAB confirm the effectiveness of the presented method.

2. DESCRIPTION AND MODELING

Fig. 1 shows the structure of a three-phase ASFCL employed in a typical three-phase circuit. The ASFCL consists of three superconducting transformers and a three-phase voltage source inverter. C_1 and C_2 are the split DC link capacitors. L_d and C_d are used to filter the harmonics generated by the PWM converter. The air-core superconducting transformer has some advantages compared to the conventional ones, such as the absence of iron losses and magnetic saturation, and lower transformer size and weight [12].

where A , B , and P are constants that are determined depending on the characteristics of OCRs. In this paper, the OCRs are assumed to have a very inverse characteristic. So, the corresponding constant values are 3.922, 0.0982, and 2, respectively [1]. $TDS_{primary}$ and $TDS_{back-up}$ are time dial settings of the primary and backup relays, respectively. The value of these parameters is calculated such that the primary and backup OCRs are coordinated. Also, M represents the plug setting multiplier (PSM) of the relay, which depends on the fault current and the current setting $I_{pick-up}$ of OCR.

The coordination time interval is defined as:

$$\Delta t = t_{back-up} - t_{primary} \quad (1)$$

The acceptable range of this parameter is normally a value between 0.2 and 0.5 seconds. **Fig. 2** shows the flowchart of the coordination of OCRs by calculating the ASCC converter settings for the fault modes.

3. DESCRIPTION OF THE PROPOSED METHOD

To describe the proposed method, a typical distribution system connected to a microgrid is shown in **Fig. 3**. To provide the same performance in terms of fault current limiting for all the DG units, the ASFCL is placed between the upstream and downstream grid. With the occurrence of a short circuit in the downstream network, the FCL operation can lead to the loss of protective coordination of the downstream OCRs and the OCRs between the upstream and downstream networks.

To solve these problems, three operating modes are proposed for ASFCL regarding the location of the fault in the overall system. The fault direction at the ASFCL

location (upstream or downstream) is detected using a directional relay. The operating modes are defined as follows:

Mode 1: Normal Operation Mode

As mentioned in **Section 2**, to neutralize the effect of the ASFCL in the main network, the output current and voltage of the converter must be set as:

$$i_{2a} = \frac{L_{S1}}{M_s} i_{1a} \quad (2)$$

$$u_{2a} = j \omega \frac{L_{S1} L_{S2} - M_s^2}{M_s} i_{1a} \quad (3)$$

where (u_{1a}, i_{1a}) and (u_{2a}, i_{2a}) are the primary and secondary voltage and current of the superconducting transformer, respectively.

Mode 2: Upstream Fault Mode

With the occurrence of a short circuit in the upstream network, the fault current without ASFCL and with ASFCL can be calculated as:

$$I_F = \frac{U_S - U_G}{Z_{T1} + Z_{T2}} \quad (4)$$

$$I_{F-withASCC} = \frac{U_S - U_G + j \omega M_s i_{2a}}{Z_{T1} + Z_{T2} + j \omega L_{S1}} \quad (5)$$

where U_S , Z_{T1} , U_G and Z_{T2} represent the equivalent source voltage and impedance of the upstream and downstream networks at the ASFCL location, respectively. According to (5), by adjusting the amplitude and angle of the converter output current (i_{2a}), the fault current can be adjusted to a suitable value so that the effect of increasing the current due to the application of new DG units is compensated.

Mode 3: Downstream Fault Mode

In this case, to reduce the voltage sag and thereby improve the power quality of the microgrid loads, ASFCL must operate in such a way that the minimum limiting impedance is applied to the network. It should be noted that the protective equipment of the microgrids is usually designed with high cut-off powers considering the future development of the microgrids. Therefore, in this case, the primary side voltage of the superconducting transformer of ASFCL can be set as:

$$u_{1a} = j \omega L_{S1} I_F - j \omega M_s i_{2a} = 0 \quad (6)$$

where I_F is equivalent to the fault current when the short circuit fault occurs in the downstream grid. The output current and voltage of the converter are calculated by (7) and (8):

$$i_{2a} = \frac{L_{S1}}{M_s} I_{FD} \quad (7)$$

$$u_{2a} = j \omega \frac{L_{S1} L_{S2} - M_s^2}{M_s} I_{FD} \quad (8)$$

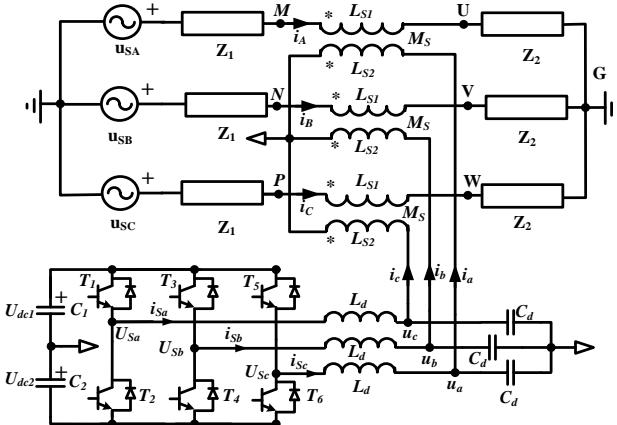


Fig. 1: The structure of a three-phase ASFCL.

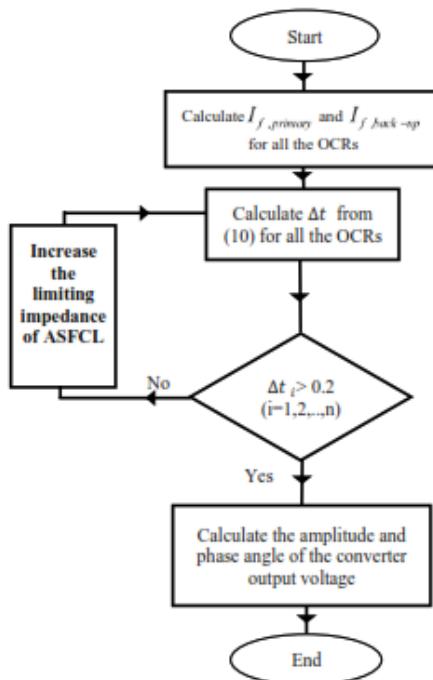


Fig. 2: The flowchart of the OCRs coordination by setting the ASFCL converter in faults mode.

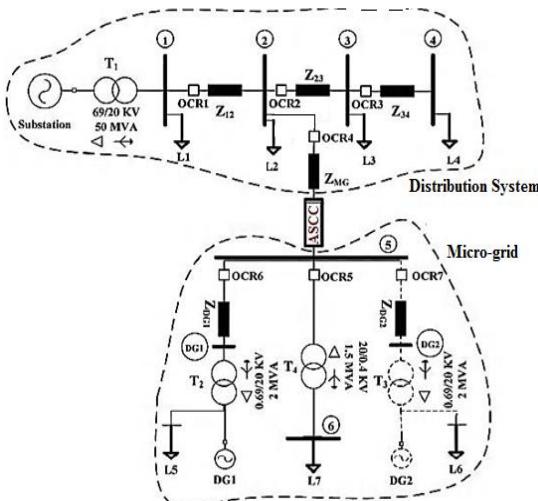


Fig. 3: A typical distribution system connected to a microgrid.

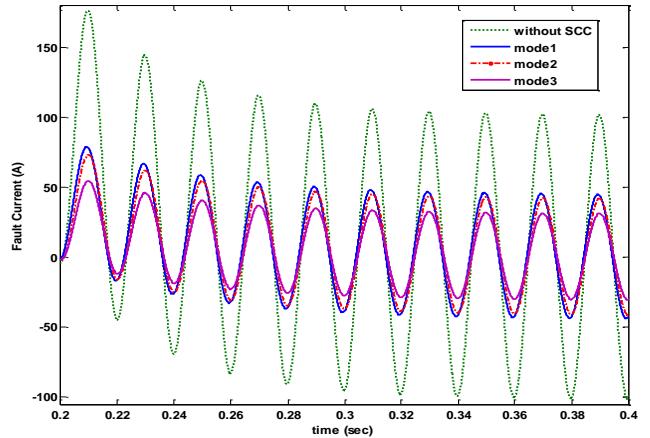


Fig. 4: The fault current without and with ASFCL.

4. SIMULATION RESULTS

This section simulates the ASFCL in different systems to test the fault current limitation and the coordination of the overcurrent relays.

4.1 Current Limiting Test

To test the operation of ASFCL on current limiting, the three-phase system shown in Fig. 1 with the parameters listed in Table 1 is simulated.

Fig. 4 shows the fault current without and with ASFCL. According to Fig. 4, the fault current is reduced to a suitable value in the presence of ASFCL. In addition, by adjusting the phase angle of the secondary current of the transformer to 90° (i.e. mode 3), the highest effectiveness of the ASFCL in limiting the fault current is obtained.

The ASCC converter reference signals in the normal and fault modes are shown in Fig. 5. In the case of the single-phase fault, the AC components of U_{dc1} and U_{dc2} are opposite to each other, so the total DC voltage is kept at the level of 600 V.

Fig. 6 depicts the current and voltage waveforms of the superconducting transformer in the presence of the ASFCL. It is worthwhile to note that once a fault occurs, the fault current is suddenly reduced to a suitable level since, for the first cycle, the ASFCL with its original setting operates in mode 1. After fault detection, based on the control strategy of the converter, the ASFCL operates in mode 3, as it is the most effective in current limiting in this mode. In other words, the operating modes of ASFCL are selected based on the reference signals.

4.2 Investigating the Effect of ASFCL on the Protective Coordination of OCRs

In this section, the power system shown in Fig. 3 is simulated as a test system with the system data listed in Table 2 [1].

In this section, the IEC Standard 60909 [18] is used to calculate the short-circuit level, and the simulation results are analyzed in four different cases to investigate the coordination of over-current relays.

Case 1) Before Adding DG2

For the base case (before adding DG2), the values of setting parameters of OCRs are calculated as shown in Table

3. Fig. 7 illustrates the time-current curves (TCC) of all main and backup OCRs for Case1. The currents measured by the main and backup OCRs are calculated for the fault in front of the main relay. As shown in Fig. 7, by adjusting the relay parameters in accordance with Table 3, all the Δt_i s are in an acceptable range. Therefore, the protective coordination of all OCRs has been carried out.

Case 2) DG2 addition and without FCL

In this case, it is assumed that the relay settings are the same as those shown in Table 3. The operating times of the relays for this case are shown in Fig. 8. As shown in Fig. 8, in this case, the coordination time of the upstream OCRs (R1 and R2, as well as R2 and R3) are out of the acceptable range ($0.2 < \Delta t < 0.5$). Thus, the coordination between these relays is disrupted, but the coordination of the OCRs between the main grid and microgrid (R4 and R5, as well as R4 and R6) is preserved.

Case 3) After DG2 addition with conventional FCL

In this case, a conventional FCL with the limiting impedance $Z_{FCL} = 16 + 0.8j \Omega$ is used in the tie feeder [1]. As shown in Fig. 9, the main grid OCRs coordination (R1 and R2, as well as R2 and R3) is preserved. However, due to the significant decrease in the fault current on the downstream side, the coordination of the OCRs between the main grid and microgrid (R4 and R5, as well as R4 and R6) is lost.

Case 4) After DG2 addition and with ASFCL

In this case, the effect of the ASFCL operating modes on the coordination of the OCRs is demonstrated. As shown in Fig. 10, with the occurrence of a short-circuit fault in the main grid, the ASFCL acts in mode 2 (the upstream fault mode) and the coordination between R1 and R2 and between R2 and R3 is preserved.

Furthermore, when a short-circuit fault occurs in the downstream network, the performance of the ASFCL in mode 3 preserves the coordination between the downstream OCRs by adjusting the fault current reduction, unlike the conventional FCL.

5. CONCLUSION

In this paper, an Active Superconducting Current Controller (ASFCL) was utilized as a voltage compensator type fault current limiter. It is placed between the main grid and microgrid to preserve the fault current level when a new DG unit is added to the microgrid. Various operating modes were defined for the ASFCL, including normal mode, upstream fault mode, and downstream fault mode. The performance of the ASFCL operation modes was compared to that of a conventional FCL for both upstream and downstream fault conditions. The simulation results show that with the occurrence of a short-circuit fault in the main grid, both ASFCL and conventional FCL have a positive effect on the coordination of the overcurrent relays and power quality of microgrid loads. On the other hand, when a short-circuit fault occurs in the microgrid and a conventional FCL is used, the coordination of the OCRs in the downstream network is violated and the power quality of the microgrid loads is reduced due to an increase in the voltage sag of these loads. The results also confirm that the application of the ASFCL with the proposed operating

modes for this case resolves the mentioned problems. Thus, the ASFCL with the proposed method outperforms the conventional FCL.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Ahmad Ghafari: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software. **Mohsen Saniei:** Funding, Supervision, Validation, Roles/Writing - original draft, Writing - review & editing. **Morteza Razaz:** Supervision, Writing - review & editing. **Alireza Saffarian:** Supervision, Writing - review & editing.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

Table 1: The parameters of the simulated system

| Parameter | Value |
|--------------------|------------------------------|
| $[U_{SA}, U_{dc}]$ | [220,600] (V) |
| Z_1 | $0.19 + 2.16 i$ (Ω) |
| Z_2 | $15 + 2 i$ (Ω) |
| F | 50 (Hz) |
| $L_{S1}=L_{S2}$ | 10 (mH) |
| $[M_s, L_f]$ | [9, 6] (mH) |
| $C_1=C_2$ | 2000 (μF) |
| C_f | 30 (μF) |

Table 2: Data of the test system

| The network components | Data |
|------------------------|---|
| Main substation | $U_{nQ}=69\text{KV}$, $S''_{kQ}=1000\text{MVA}$ |
| Transformer (T_1) | $S=50\text{MVA}$, $69/20\text{KV}$, $u_k=20.5\%$ |
| $Z_{12}\text{-}Z_{34}$ | $2.75+4.15j$ |
| ZMG | $2.15+3.24j$ |
| L1-L4 | $S=20\text{MVA}$, $\text{PF}=0.94$ |
| DG ₁ | $S_{rG}=1.5\text{MVA}$, $U_{rG}=690\text{V}$ |
| T_2 and T_3 | $S=2\text{MVA}$, $0.69/20\text{KV}$, $u_k=6\%$ |
| T_4 | $S=1.5\text{MVA}$, $20/0.4\text{KV}$, $u_k=6.5\%$ |
| Z_{DG1} | $0.081+0.057j$ |
| Z_{DG2} | $0.162+0.114j$ |
| L_5 and L_6 | $S=1.2\text{MVA}$, $\text{PF}=0.95$ |
| L_7 | $S=0.9\text{MVA}$, $\text{PF}=0.97$ |

Table 3: Setting values for each OCR for the base case

| Relay unit | Max. Load current (A) | CT ratio | Pick-up Current | TDS |
|------------|-----------------------|----------|-----------------|-----|
| OCR1 | 800 | 1000/5 | 5.496 | 0.4 |
| OCR2 | 488 | 500/5 | 7.636 | 0.2 |
| OCR3 | 220 | 300/5 | 5.5 | 1.1 |
| OCR4 | 60 | 100/5 | 4.5 | 1 |
| OCR5 | 15 | 100/5 | 1.5 | 1.9 |
| OCR6 | 75 | 100/5 | 5.62 | 2.7 |

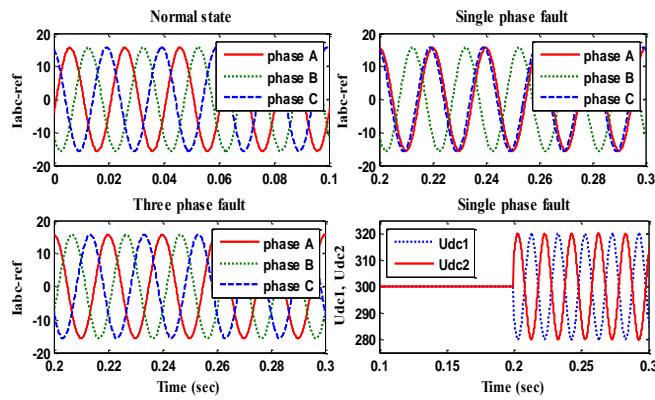


Fig. 5: The reference signals of the ASFCL converter in normal and fault states

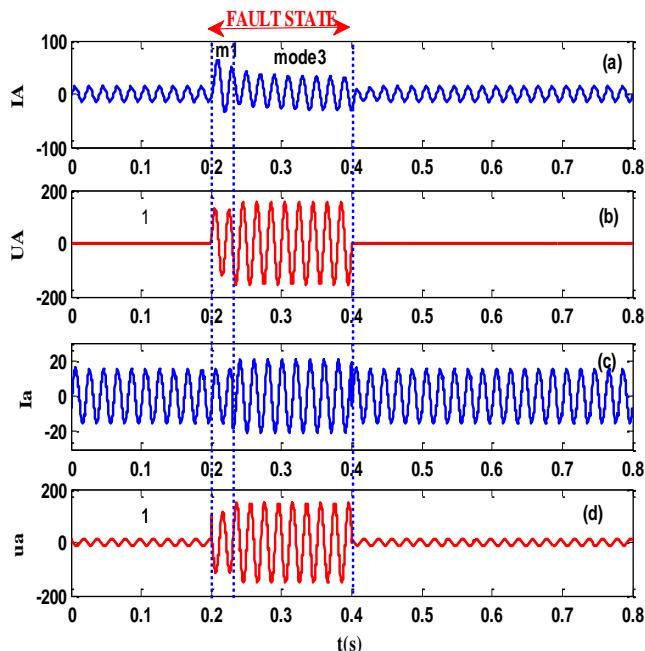


Fig. 6: The waveforms of the primary and secondary currents and voltages of the superconducting transformer

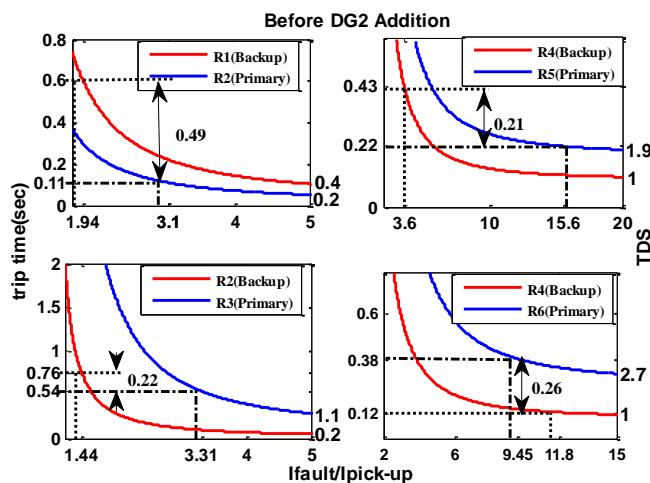


Fig. 7: The time-current curves of OCRs before adding DG2

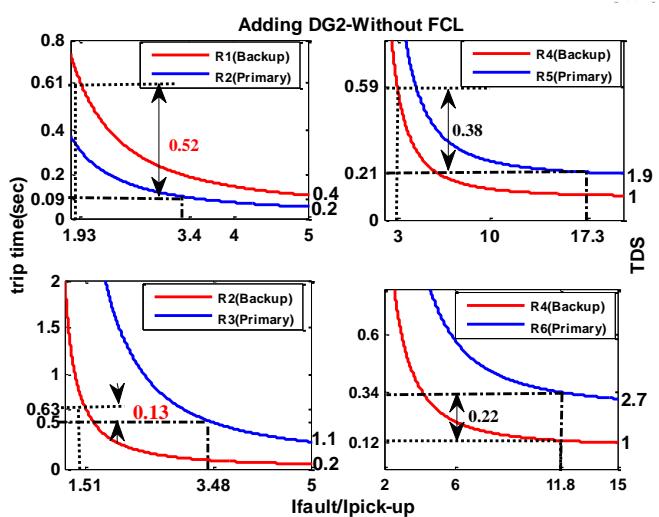


Fig. 8: The time-current curves of OCRs after DG2 addition and without FCL

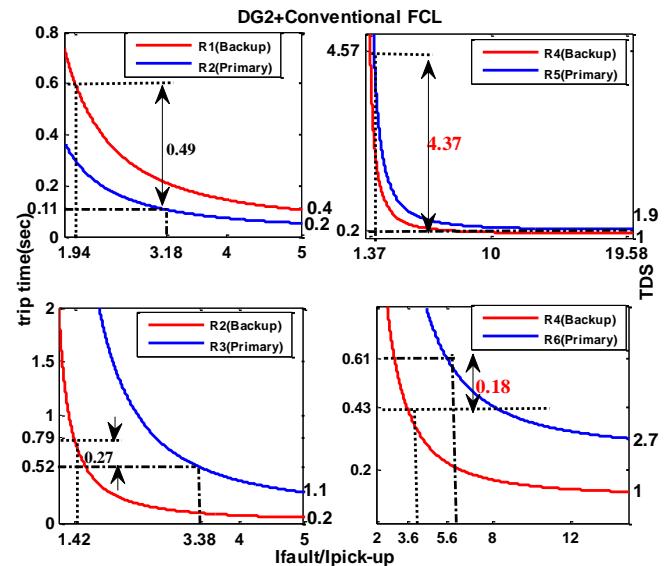


Fig. 9: The time-current curves of OCRs after adding DG2 and with conventional FCL

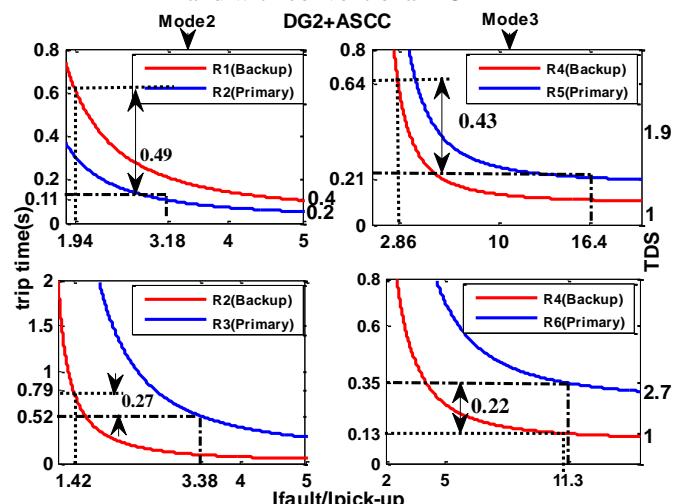


Fig. 10: The time-current curves of OCRs after adding DG2 and with ASFCL

REFERENCES

- [1] T. Ghanbari, and E. Farjah, "A multiagent-based fault-current limiting scheme for the microgrids," *IEEE Transactions on Power Delivery*, vol. 29, no. 2, pp. 525-533, 2014
- [2] S. T. Lim, and S. H. Lim, "Analysis on protective coordination between over-current relays with voltage component in a power distribution system with SFCL," *IEEE Transactions on Applied Superconductivity*, vol. 30, no. 4, pp. 5601706, 2020.
- [3] M. Yang, X. Wang, W. Sima, T. Yuan, P. Sun, and H. Liu, "Air-coreTransformer-based solid-state fault-current limiter for bidirectional HVDCsystems," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 5, pp. 4914-4925, 2022.
- [4] T. Ghanbari, E. Farjah, "Unidirectional fault current limiter: An efficient interface between the microgrid and main network," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1591-1598, 2013.
- [5] A. G. Pronto, F. Vale, N. Vilhena and J. Murta-Pina, "Electromechanical analysis of core- and shell-type inductive superconducting fault current limiters under general fault conditions," *IEEE Transactions on Applied Superconductivity*, vol. 32, no. 1, pp. 1-5, 2022.
- [6] J. Sheng *et al.*, "Field test of a resistive type superconducting fault current limiter in distribution network," *IEEE Transactions on Applied Superconductivity*, vol. 31, no. 8, pp. 1-4, 2021.
- [7] B. Li, F. Guo, J. Wang, C. Li, "Electromagnetic transient analysis of the saturated iron-core superconductor fault current limiter," *IEEE Transactions on Applied Superconductivity*, vol. 25, no. 3, pp. 1-5, 2015.
- [8] M. Song, Y. Tang, Y. Zhou, L. Ren, L. Chen, S. Cheng , "Electromagnetic characteristics analysis of air-core transformer used in voltage compensation type active SFCL," *IEEE Transactions on Applied Superconductivity*. vol. 20, no. 3, pp. 1194-1198, 2010.
- [9] O. Naeckel, and M. Noe, "Design and test of an air coil superconducting fault current limiter demonstrator," *IEEE Transactions on Applied Superconductivity*, vol. 24, no. 3, pp. 5601605, 2014.
- [10] S. Lim, J. Moon, J. Kim, "Improvement on current limiting characteristics of a flux-lock type SFCL using E-I core," *IEEE Transactions on Applied Superconductivity*. vol. 19, no. 3, pp. 1904-1907, 2009.
- [11] Y. Zhou, C. Ji, Z. Dong and S. Zhang, "Cooperative control of SFCL and SMES-battery HESS for mitigating effect of ground faults in DC microgrids," *IEEE Transactions on Applied Superconductivity*, vol. 31, no. 8, pp. 1-5, 2021.
- [12] J. Wang, L. Zhou, J. Shi, and Y. Tang, "Experimental investigation of an active superconducting current controller," *IEEE Transactions on Applied Superconductivity*, vol. 21, no. 3, pp. 1258-1262, 2011.
- [13] L. Chen, Y. Tang, J. Shi, and Z. Sun, "Simulations and experimental analyses of the active superconducting fault current limiter," *Physica C: Superconductivity and Its Applications*, vol. 459, no. 1, pp. 27-32, 2007.
- [14] L. Chen, Y. Tang, J. Shi, Z. Li, L. Ren, and S. Cheng, "Control strategy for three-phase four-wire PWM converter of integrated compensation type active SFCL," *Physica C: Superconductivity and Its Applications*, vol. 470, no. 2, pp. 231-235, 2010.
- [15] H. Yamaguchi, K. Yoshikawa, M. Nakamura, T. Kataoka, K. Kaiho, "Current limiting characteristics of transformer type superconducting fault current limiter," *IEEE Transactions on Applied Superconductivity*, vol. 14, no. 2, pp. 815-818, 2004.
- [16] A. Ghafari, M. Razaz, S.G. Seifossadat, and M. Hosseinzadeh, "Protective coordination of main and back-up overcurrent relays with different operating modes of active super-conducting current controller," *Maejo International Journal of Science and Technology*, vol. 8, no. 3, pp. 319-333, 2014.
- [17] Short-circuit currents in three-phase AC systems - Part 4: Examples for the calculation of short-circuit currents, IEC 60909-4, 2021.

BIOGRAPHY



Ahmad Ghafari Gusheh was born in 1987 in Shahrekord, Iran. He received his B.Sc. degrees in electrical engineering from Islamic Azad University of Najafabad, Najafabad, Iran in 2009 and his M.Sc and Ph.D. degrees in electrical engineering from Shahid-Chamran University of hvaz, Iran in 2012 and 2021, respectively. His main research interests are power system protection, microgrid, and power distribution systems.



Mohsen Saniei was born in 1966 in Dezful, Iran. He received his B.Sc. degree in electrical engineering from the Ferdowsi University of Mashhad, Iran, in 1989, his M.Sc. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran in 1992 and his Ph.D. degree from the University of Strathclyde, Glasgow, UK, in 2004. Currently, he is an Associate Professor with the Department of Electrical Engineering, the Shahid Chamran University of Ahvaz, Ahvaz, Iran. His research interests include power system operation, control and stability, microgrid, and electricity market.



Morteza Razaz was born in Dezful, Iran. He received his B.Sc. and M.Sc. degrees in electrical engineering and applied mathematics from Texas University in 1977 and 1979, respectively, and his Ph.D. degree from Sharup University in 1993. Currently, he is an Associate Professor with the Department of Electrical Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran. His research interests include transformers and power systems.

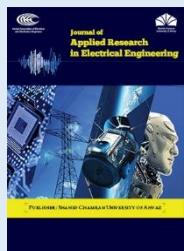


Alireza Saffarian was born in 1981 in Ahvaz, Iran. He received his B.Sc. and M.Sc. degrees in electrical engineering from the Amirkabir University of Technology, Tehran, Iran in 2003 and 2005, respectively and his Ph.D. degree in Iran in 2011. Currently, he is an Associate Professor with the Department of Electrical Engineering, the Shahid Chamran University of Ahvaz, Ahvaz, Iran. His research interests include power system protection, power system stability, from the University of Tehran, Tehran, and power quality assessment.

Copyrights

© 2022 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Multi-Objective Optimal Power Flow Based Combined Non-Convex Economic Dispatch with Valve-Point Effects and Emission Using Gravitation Search Algorithm

Nabil Mezhoud* and Mohamed Amarouayache

Electrical Engineering Department, LES Laboratory, University of August 20th, 1955, Skikda, Algeria

* Corresponding Author: mezhounab@yahoo.fr

Abstract: This paper presents a solution to the Optimal Power Flow (OPF) problem combined economic dispatch with valve-point effect and Emission Index (EI) in electrical power networks using the physics-inspired optimization method, which is the Gravitational Search Algorithm (GSA). Our main goal is to minimize the objective function necessary for the best balance between energy production and its consumption which is presented in a nonlinear function, taking into account equality and inequality constraints. The objective is to minimize the total cost of active generations, the active power losses, and the emission index. The GSA method has been examined and tested on the standard IEEE 30-bus test system with various objective functions. The simulation results of the used methods have been compared and validated with those reported in the recent literature. The results are promising and show the effectiveness and robustness of the used method. It should be mentioned that from the base case, the cost generation, the active power losses, and the emission index are significantly reduced to 823 (\$/h), 6.038 (MW), and 0.227 (ton/h), which are considered 5.85%, 61.61%, and 44.63%, respectively.

Keywords: Optimization, optimal power flow, emission index, gravitational search algorithm.

Article history

Received 14 April 2022; Revised 09 June 2022; Accepted 29 June 2022; Published online 28 February 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

M. Nabil, and A. Mohamed, "Multi-objective optimal power flow based combined non-convex economic dispatch with valve-point effects and emission using gravitation search algorithm," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 26-36, 2023. DOI: [10.22055/jaree.2022.40549.1055](https://doi.org/10.22055/jaree.2022.40549.1055)



1. INTRODUCTION

Electric power plants that operate on fossil-fuels are among the most prominent sources of air pollution and contribute to causing great harm to the environment due to the burning of raw fuels such as coal, gas, and oil [1].

Electric power systems engineering has the longest history of development compared to the various fields of engineering. In electrical supply systems, there are a wide range of problems involved in system optimization [2]. Among these problems, power system scheduling is one of the most important in system operation, control, and management.

Power plants Coal-fired contribute a large quantity of polluting gases to the Atmosphere, as they produce large amounts of Carbon oxides CO₂, and some toxic and dangerous gases such as emissions of Sulfur oxides SO_x, and Nitrogen oxides, NO_x [1-2].

After implementation of the 1990 amendment to the United States Clean Air Act and increasing public awareness of environmental protection and public utilities, electricity production companies were obligated to adapt their designs and making strategy to reduced pollution rate and emissions of electric power plants [2-3]. Several efforts and strategies have been proposed and devoted to reduce atmospheric of pollutant emissions [2].

The OPF problem has a long history in its development for more than 60 years. Since the OPF problem was first discussed by Carpenter in 1962, then formulated by Dommel and Tinney in 1968 [4]. The OPF are non-linear and non-convex very constrained optimization problems.

The ED problem is one of the concerns of statistical optimization in the planning, control and operation of electric power; he is a sub-problem of OPF [5].

The OPF is an important criterion in today's power system operation and control due to scarcity of energy

resources, increasing power generation cost and ever-growing demand for electric energy [2].

The main purpose of an OPF is to determine the optimal operating state of a power system and the corresponding settings for economic operation of control variables by optimizing a particular objective while meeting the constraints of economics and security, such as equality and inequality constraints [1, 5-6].

In the past, various optimization methods have been applied, and some of them have been implemented into practice. Over the past few years, many methods have been used to solve the OPF and EI problems like; Quadratic programming method (QP) [7], Newton and Qassi-Newton methods [8-9], linear and non-linear programming methods [10-11] and interior point methods (IPM) [12].

In the last two decades, and in order to solve the OPF and EI problems, several methods of optimization are formulated such as Artificial neutral networks (ANN) [13], Artificial bee colony (ABC) and Incremental artificial bee colony (IABC) [14-15], Bacterial foraging algorithms (BFA) [16], Cuckoo search algorithm (CSA) [17], Harmony search (HS) [18], Evolution programming (EP) [19], Differential evaluation (DE) [20], Modified differential evaluation (MDE) [21], Tabu search (TS) [22], Simulated annealing (SA) [23], Gravitational search algorithms (GSA) [24], Evolutionary algorithm [25], Genetic algorithms (GA) [26], Particle swarm optimization (PSO) [27], Modified Particle swarm optimization (MPSO) [28], Ant colony optimization (ACO) [29], Tree-seed algorithm (TSA) [30], Moth Swarm Algorithm (MSA) [31], Sine-cosine algorithm (SCA) [32], Firefly Algorithm [33], Modified imperialist competitive algorithm (MICA) [34], Shuffled frog leaping algorithm (SFLA) [35], Electromagnetism-like mechanism method (ELM) [36], Ant-lion optimizer [37], Interior search algorithm [38], Wind driven optimization (WDO) method [39], Machine learning and modified grasshopper optimization algorithms [40], Rao algorithm [41], Artificial Eco-system optimization [42], Hamiltonian technique [43], Teaching-learning-studying-based optimization [44] and Grey wolf optimizer (GWO) [3, 45]. Variants of these algorithms were proposed to handle multi-objective functions in electric power systems.

The proposed GSA approach is tested and illustrated by numerical examples based on IEEE 30-bus test system.

With comparison, the obtained results validate the advantage of the proposed approach over many other methods in terms of solution quality.

2. PROBLEM FORMULATION

The OPF and EI are nonlinear optimization problems, represented by a predefined objective function f , subject to a set of equality and inequality constraints [46]. Generally, these problems can be expressed as follows.

$$\text{Min } f(x, u) \quad (1)$$

Subject to

$$h(x, u) = 0 \quad (2)$$

$$g(x, u) \leq 0 \quad (3)$$

$$x_{\min} \leq x \leq x_{\max} \quad \& \quad u_{\min} \leq u \leq u_{\max} \quad (4)$$

where $f(x, u)$ is a scalar objective function to be optimised, $h(x, u)$ and $g(x, u)$ are, respectively, the set of nonlinear equality constraints represented by the load flow equations and inequality constraints consists of state variable limits and functional operating constraints. x and u are the state and control variables vectors respectively. Hence, x and u can be expressed as given

$$x' = \{P_{G_1}, |V_{L_1}|, \dots, |V_{L_{n_L}}|, Q_{G_1}, \dots, Q_{G_{n_g}}, S_1, \dots, S_{n_{br}}\} \quad (5)$$

where, P_G , Q_G , V_L and S_k are the generating active power at slack bus, reactive power generated by all generators, magnitude voltage of all load buses and apparent power flow in all branches, respectively. n_g , n_L and n_{br} are, respectively, the total number of generators, the total number of load buses and the total number of branches.

The set control parameters are represented in terms of the decision vector as follows:

$$u' = \{P_{G_2}, \dots, P_{G_{n_g}}, |V_{G_1}|, \dots, |V_{G_{n_g}}|, Q_{1_{com}}, \dots, Q_{n_{com}}, T_1, \dots, T_{n_T}\} \quad (6)$$

where, P_G are the active power generation excluding the slack generator, V_G are the generators magnitude voltage, T is tap settings transformers, and Q_{com} are the reactive power compensation by shunt compensator, n_T and n_{com} are the total number of transformers and the total number of compensators units, respectively.

2.1. Single-Objective Function

In general, the single-objective function is a nonlinear programming problem. In this paper, four single objectives commonly found in OPF and EI have been considered.

2.1.1. Cost without valve-point optimization

The objective function of cost optimization f_1 of quadratic cost equation for all generators as given below

$$f_1 = \min \sum_{k=1}^{n_g} C(P_{gk}) = \min \sum_{k=1}^{n_g} a_k + b_k P_{gk} + c_k P_{gk}^2 \quad (7)$$

where f_1 is the total generation cost in (\$/h). P_{gk} and n_g are the active power output generated by the i^{th} generator and the total number of generators. a_k , b_k and c_k are the cost coefficients of the generator k .

2.1.2. Cost with valve-point optimization

When the valve point loading is taken into account, this model can be used as is, except for the shape of the objective function instead of being a quadratic function it is now a non-convex and smooth function as shown in Fig. 1 [14-15].

Generally, when every steam valves begins to open, the valve-point shows rippling. However, the characteristics of input-output of generation units make nonlinear and non-smooth of the fuel costs function. To consider the valve-point

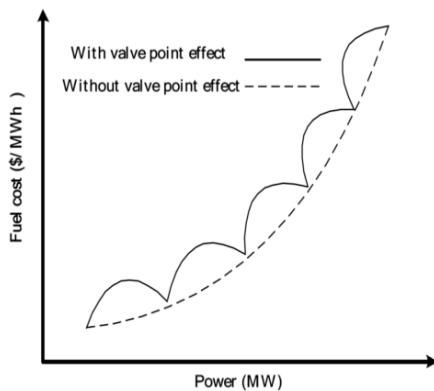


Fig. 1: Fuel cost curve of units with valve-point effects.

effect, the sinusoidal function is incorporated into the quadratic function. Typically, this function is represented as follows [14-15, 26].

$$f_2 = \min \sum_{k=1}^{n_g} \left[a_k + b_k P_{gk} + c_k P_{gk}^2 + d_k \sin(e_k (P_{gk}^{\min} - P_{gk})) \right] \quad (8)$$

where d_k and e_k are the cost coefficients of unit with valve-point effect.

2.1.3. Active power loss optimization

The active power loss function f_3 in (MW) to be minimized can be expressed as follows

$$f_3 = \sum_{k=1}^{n_b} G_{kj} \left[V_k^2 + V_j^2 - 2V_k V_j \cos \theta_{kj} \right] \quad (9)$$

where, V_k and V_j are the magnitude voltage at buses k and j , respectively, G_{kj} is the conductance of line kj , θ_{kj} is the voltage angle between buses k and j , and n_b is total number of buses.

2.1.4. Emission optimization

The emission function is the sum of exponential and quadratic functions of real power generating. Using a quadratic equation, emission of harmful gases is calculated in (ton/h) as given below [34, 46-47].

$$f_4 = \min \sum_{k=1}^{n_g} 10^{-2} (\alpha_k + \beta_k P_{gk} + \gamma_k P_{gk}^2) + \zeta_k \exp(\lambda_k P_{gk}) \quad (10)$$

where f_4 is the emission function in (ton/h), $\alpha_k, \beta_k, \gamma_k, \zeta_k$ and λ_k are the emission coefficients of the generator k .

2.2. Multi-Objective Optimization

In all multi-objective functions, we use the weighted aggregation function. The function used in the case of weighted aggregation is given by (11).

$$\text{Min} F = \sum_{i=1}^{n_f} \omega_i f_i \text{ with } \omega_i \geq 0 \text{ and } \sum_{i=1}^{n_f} \omega_i = 1 \quad (11)$$

where $\sum_{i=1}^{n_f} \omega_i = 1$ & $i = 1 : n_f$, ω_i is the weighting factor and n_f is the number of objective function considered.

2.3. Equality Constraints

These equality constraints are the sets of nonlinear load flow equations that govern the power system, i.e.:

$$\begin{cases} P_{gk} = P_k + P_{dk} \\ Q_{gk} - Q_{ComK} = Q_k + Q_{dk} \end{cases} \quad (12)$$

where P_{gk} and Q_{gk} are, respectively, the scheduled active and reactive power generations at bus k . P_k, Q_k are the active and reactive power injections at bus k . P_{dk}, Q_{dk} and Q_{comk} are the active and reactive power loads at bus k and the reactive power compensation at bus k .

2.4. Inequality Constraints

The inequality constraints $g(x, u)$ are represented by the system operational and security limits, listed below

$$P_{gk}^{\min} \leq P_{gk} \leq P_{gk}^{\max} \text{ where } k = 1, \dots, n_g \quad (13)$$

$$Q_{gk}^{\min} \leq Q_{gk} \leq Q_{gk}^{\max} \text{ where } k = 1, \dots, n_g \quad (14)$$

$$V_k^{\min} \leq V_k \leq V_k^{\max} \text{ where } k = 1, \dots, n_b \quad (15)$$

$$\theta_k^{\min} \leq \theta_k \leq \theta_k^{\max} \text{ where } k = 1, \dots, n_b \quad (16)$$

$$T_k^{\min} \leq T_k \leq T_k^{\max} \text{ where } k = 1, \dots, n_T \quad (17)$$

$$Q_{Comk}^{\min} \leq Q_{Comk} \leq Q_{Comk}^{\max} \text{ where } k = 1, \dots, n_{Com} \quad (18)$$

$$S_{kj} \leq S_{kj}^{\max} \text{ where } k = j = 1, \dots, n_b \quad (19)$$

where, n_T , n_{Com} , T and Q_{Com} are the total number of transformers, total number of compensators, transformers tap settings, the reactive power compensation and S_{kj}^{\max} is the maximum apparent power between buses k and j .

2.5. Gravitation Search Algorithm

Gravity Search Algorithm (GSA) is one of the recent algorithms developed by Rashidi et al. [48]. GSA is also a meta-heuristic method inspired by Newtonian laws of gravitation and mass interactions [24, 47-48]. The agents in the GSA method are the targets whose performance is measured by their masses. Each agent attracts another agent by a force of gravity which is inversely proportional to the square of the distance between the agents and directly proportional to the product of their mass. By means of the Newtonian law of motion this force creates a global movement of all agents towards the heavier masses.

Compared to lighter agents, heavier agents move very slowly which correspond to good solutions to the problem [24, 49].

In the GSA method [48-50], the agents/vectors of the solution are considered as objects and their performance is measured by their masses. Each mass (agent) has specified by four specifications: position of the mass, inertial mass, active gravitational mass and passive gravitational mass. The position of the mass corresponds to the solution of the

problem, and its gravitational and inertial masses are computed using a fitness function. The algorithm is navigated by properly adjusting the gravitational and inertial masses. By lapse of iteration cycles, it is expected that masses be attracted by the heaviest mass. This heaviest mass will present an optimum solution in the search space [24].

The GSA could be considered as an isolated masses system. It is like a small artificial world of masses obeying the Newtonian laws of gravitation and motion. More precisely, masses obey the following two laws [50].

Now, let us consider a system with N_a agents (masses). The position of the i^{th} agent is defined by

$$x_i = (x_i^1, \dots, x_i^d, \dots, x_i^D) \quad \text{where } i = 1, \dots, N_a \quad (20)$$

where x_i^d represents the positions of the i^{th} agent in the d^{th} dimension, which is a candidate solution to the problem, D is the space dimension of the problem and N_a is total number of agents in the swarm [48].

Initially, the agents of the solution are defined according to Newton gravitation theory. At a specific iteration t , the force acting on i^{th} mass from j^{th} mass according to Newton gravitation theory is defined randomly as follows

$$F_{ij}^d(t) = G(t) \frac{M_{pi}(t) \times M_{aj}(t)}{R_{ij}(t) + \varepsilon} (x_j^d(t) - x_i^d(t)) \quad (21)$$

where M_{pi} is the mass of the object i , M_{aj} the mass of the object j , $G(t)$ is the gravitational constant at time t , ε is a small constant and $R_{ij}(t)$ is the Euclidean distance between two agents i and j given as follows.

$$R_{ij}(t) = \|x_i(t), x_j(t)\|_2 \quad (22)$$

To give a stochastic characteristic to the algorithm, it is expected that the total force that acts on i^{th} agent in d^{th} dimension be a randomly weighted sum of d^{th} components of the forces exerted from other agents given by the following equation

$$F_i^d(t) = \sum_{\substack{j=1 \\ j \neq i}}^{N_a} rand_j F_{ij}^d(t) \quad (23)$$

where $rand_j$ is uniform random variable in the interval $[0, 1]$, this random is used to give a randomized characteristic to the search [48, 50].

The law of motion is used directly to calculate the acceleration of i^{th} agent, at time t in the d^{th} dimension. This acceleration is proportional to the force acting on that agent, and inversely proportional to the mass agent. a_i^d is given as

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \quad (24)$$

where $M_{ii}(t)$ is the inertial mass of the i^{th} agent and $a_i^d(t)$ is the acceleration of i^{th} agent in the d^{th} dimension at iteration t .

Moreover, a search strategy can be defined on this idea to find the next velocity and position of the agent. Further, the next velocity of any agent is considered a fraction of its current velocity and current acceleration. Therefore, the next velocity and the next position of an agent can be calculated as [50-51].

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \quad (25)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (26)$$

v_i^d and x_i^d are, respectively, the velocity and the position of an agent. The gravitational constant, G , which is initialized randomly at the starting, and given in terms of the initial gravitational constant (G_0) and iteration (t) expressed by (27).

$$G = G_0 \exp\left(-\alpha \frac{t}{t_{max}}\right) \quad (27)$$

where α is a user specified constant, t and t_{max} are the current and the total numbers of iterations, respectively. G_0 is set to 100, α is set to 20 [48].

The masses of agents are computed using fitness evaluation. The heavier mass of an agent, the more influential is that agent concerning the solution it represents. The masses are updated as follows:

$$M_{ai} = M_{pi} = M_{ii} \quad \text{for } i = 1, 2, \dots, N_a \quad (28)$$

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \quad (29)$$

where $fit_i(t)$ represents the fitness of the j^{th} agent at iteration t , $best(t)$ and $worst(t)$ represents the best and worst fitness value of all agents at generation t .

$$M_i(t) = \frac{m_i(t)}{\sum_1^{N_a} m_i(t)} \quad (30)$$

where $M_i(t)$ is the agent mass of i at iteration t . For a minimization problem

$$best(t) = \min_{j \in \{1, \dots, N_a\}} fit_j(t) \quad (31)$$

$$worst(t) = \max_{j \in \{1, \dots, N_a\}} fit_j(t) \quad (32)$$

The total force acting on the i^{th} agent is computed as follows:

$$F_i^d(t) = \sum_{\substack{j \in K_{best} \\ j \neq i}}^{N_a} rand_j F_{ij}^d(t) \quad (33)$$

K_{best} is the set of first K agents with the best fitness value and the biggest mass, which is a function of time with the initial value, K_0 and it decreases with time. In such a way, all agents apply the forces at the beginning, and as time passes, K_{best} is linearly decreased to 1. At the end, there will be only one agent applying force to the others.

2.5.1. Implementation GSA in OPF problem

At the beginning of the GSA algorithm, in the search space each agent is placed at a certain point, which defines a solution to the problem. Then, the customers are retrieved and their next locations are calculated according to (18) and (19). Other parameters of the algorithm such as masses M , gravitational constant G , and acceleration a are calculated using equations (27)-(30), and (24), respectively, and updated each iteration. The Flowchart of GSA used in this works is shown in Fig. 2.

Below we will present the steps of the GSA method to solve the problem of OPF.

Step 1: Initialization the population size of agent vectors, $N_a = 25$, $t_{\max} = 250$, $G_0 = 100$, $\alpha = 10$

Step 2: Generation of the initial vectors of the agents N_a having that $(n + 1)$

Step 3: Calculate the values of fitness error of total population, N_a , as shown by Equation (31).

Step 4: Calculation of population best solution (hgbest).

Step 5: Update $G(t)$, $best(t)$, $worst(t)$ and $M_i(t)$ for $i = 1, \dots, N_a$.

Step 6: Calculate the sum of forces in different directions.

Step 7: Calculate the factor velocities and accelerations.

Step 8: Update the position of agent's.

Step 9: Repeat steps 3 through 8 until the stopping criterion is met (either the maximum number of iterations or near global optimal solution, hgbest) is met.

3. SIMULATION & RESULTS

The five generators system, IEEE 30-bus system is used throughout this work to test the proposed algorithm. This system consist, 30 buses, 6 generators units and 41 branches, 37 of them are the transmissions lines and 4 are the tap changing transformers. One of these buses is chosen like as a reference bus (slack bus), the buses containing generators are taken the PV buses, the remaining buses are the PQ buses or loads buses. It is assumed that 9 capacitors compensation is available at buses 10, 12, 15, 17, 20, 21, 23, 24 and 29. The network data, the cost and emission coefficients of the five generators are referred in [52]. The one-line diagram IEEE 30-bus system is shown in Fig. 3.

The total loads of active and reactive powers are 283.4 (MW) and 126.2 (MVar), respectively, with 24 control variables. The basis apparent power used in this paper is 100 (MVA). The simulation results of load flow problem of test system are summarized in Table 1.

3.1.1. Case 1: Cost without valve point effect

In this case, the cost has resulted in 801.7517 (\$/h), which is considered 8.3608 % lower than the initial case (load flow). Fig. 4 shows the convergence characteristic of cost using GSA. Table 1 summarizes the optimal control variables setting in this case.

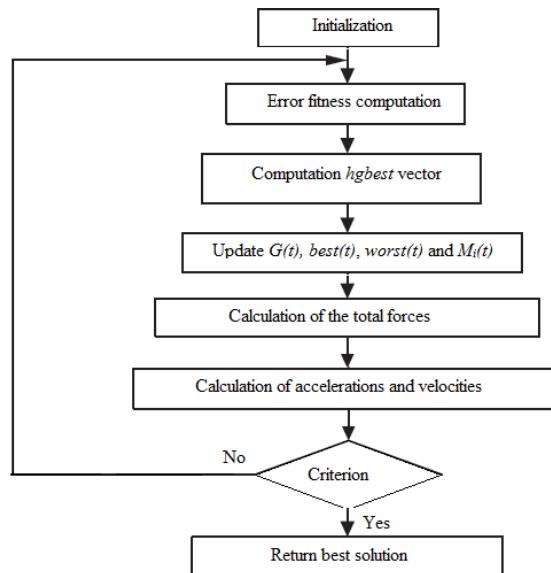


Fig. 2: Flowchart of GSA.

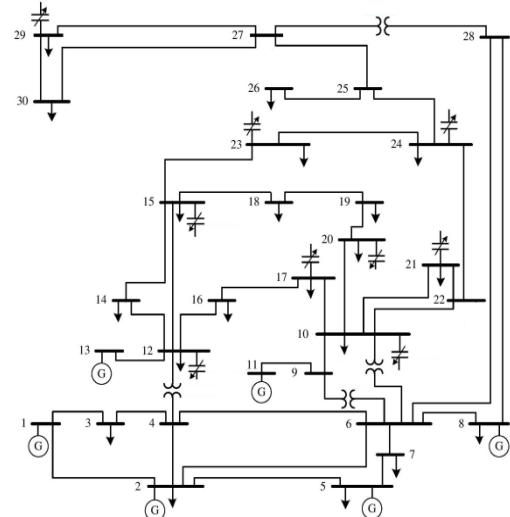


Fig. 3: Single-line diagram of IEEE 30-bus system.

3.1.2. Case 2: Cost optimization with valve point effect

In this case, the cost has resulted in 834.85 (\$/h), which is considered 3.837% lower than the initial case. Table 1 summarizes the optimal control variables of this case. Fig. 4 illustrates the convergence algorithms for case 2.

3.1.3. Case 3: Active power loss optimization

The optimal control variables of this case are introduced in Table 1. Fig. 5 shows the convergence characteristics of active power losses using GSA algorithm. The active power loss has dramatically decreased to 5.4074 (MW) which is considered 81.5905% lower than the basic case.

3.1.4. Case 4: Gas emission optimization

In this case, the emission reduction yielded 0.2162 (ton/h), which is considered 97.7962% lower than initial case.

Table 1: Results of cases 1, 2 and 3 for test system.

| Control variables | Optimal values | | |
|---------------------------|----------------|---------------|---------|
| | Case 1 | Case 2 | Case 3 |
| | Cost w/o valve | Cost w/ valve | Losses |
| P _{G2} (MW) | 49.3452 | 36.9639 | 42.1640 |
| P _{G5} (MW) | 21.1403 | 16.3928 | 49.9853 |
| P _{G8} (MW) | 21.2490 | 10.3052 | 34.9261 |
| P _{G11} (MW) | 11.9704 | 11.1055 | 29.8481 |
| P _{G13} (MW) | 12.0421 | 12.0007 | 37.5303 |
| V ₁ (pu) | 1.0860 | 1.0603 | 1.0696 |
| V ₂ (pu) | 1.0673 | 1.0330 | 1.0559 |
| V ₅ (pu) | 1.0380 | 1.0062 | 1.0326 |
| V ₈ (pu) | 1.0391 | 0.9967 | 1.0396 |
| V ₁₁ (pu) | 1.0926 | 1.0056 | 1.0781 |
| V ₁₃ (pu) | 1.0444 | 1.0668 | 1.0255 |
| Q _{com10} (MVAr) | 1.3214 | 1.0262 | 3.2206 |
| Q _{com12} (MVAr) | 1.2352 | 2.0189 | 0.6561 |
| Q _{com15} (MVAr) | 1.9913 | 2.5747 | 4.0699 |
| Q _{com17} (MVAr) | 3.1741 | 2.0459 | 2.8506 |
| Q _{com20} (MVAr) | 0.9824 | 2.5540 | 2.3771 |
| Q _{com21} (MVAr) | 3.8632 | 3.7772 | 3.8207 |
| Q _{com23} (MVAr) | 3.8792 | 1.4481 | 4.3263 |
| Q _{com24} (MVAr) | 2.4345 | 2.1937 | 2.3369 |
| Q _{com29} (MVAr) | 2.6679 | 2.4615 | 1.5492 |
| T ₆₋₉ | 1.0042 | 1.0037 | 1.0329 |
| T ₆₋₁₀ | 1.0021 | 0.9824 | 0.9452 |
| T ₄₋₁₂ | 0.9574 | 0.9404 | 0.9911 |
| T ₂₈₋₂₇ | 0.9762 | 0.9630 | 0.9873 |
| Cost in (\$/h) | 800.751 | 834.85 | 912.19 |
| loss in (MW) | 9.0937 | 12.264 | 3.837 |
| Emission (ton/h) | 0.3117 | 0.3211 | 0.2161 |
| Slack in (MW) | 176.7467 | 208.899 | 92.7820 |
| CPU time (s) | 87.2648 | 86.756 | 77.595 |

The optimal settings of control variables of this case are detailed in [Table 2](#). The convergence characteristic of emission using GSA method is shown in [Fig. 6](#).

3.1.5. Case 5: Cost and active loss optimization

The multi-objective control variables considering cost and active loss are tabulated in [Table 2](#). [Fig. 7](#) shows the trend of optimization for this case using GSA method.

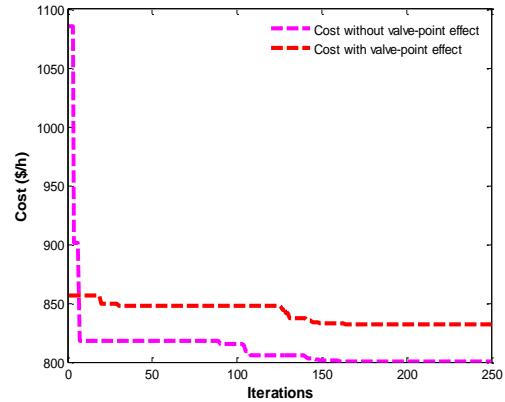
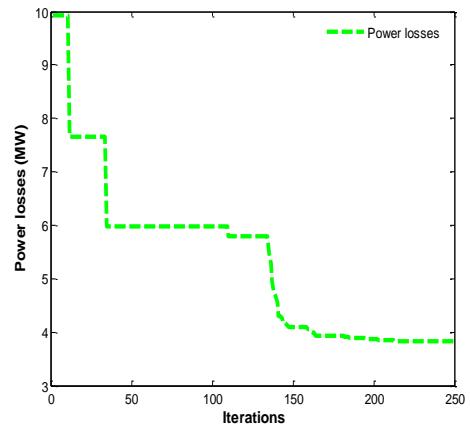
3.1.6. Case 6: Cost and gas emission optimization

[Fig. 8](#) shows the convergence characteristics obtained in case 6. The results of this case are tabulated in [Table 3](#).

3.1.7. Case 7: Cost, active power loss and gas emission

The control variables setting of multi-objectives considering cost, active power loss and emission are given in [Table 3](#). The convergence characteristics of this case are shown in [Fig. 9](#).

In order to obtain the desired set of non-dominant solution points, we run the algorithm with different weight factor. Therefore, the multi-objective problem is transformed into a single objective problem using the linear summation of weight factors according to (34).

**Fig. 4:** Convergence of algorithm for cases 1 and 2.**Fig. 5:** Convergence of algorithm for case 3.

$$f_{\text{multi-objective}} = w_c \cdot f_1 + w_l \cdot f_3 + w_e \cdot f_4 \quad (34)$$

where w_c , w_l and w_e are, respectively, the weight factor for cost, losses and emission functions and $w_c + w_l + w_e = 1$. [Table 4](#) shows the obtained results using different weight factors.

From the results presented in [Table 1](#) and [Figs. 4, 5](#), and [6](#) it can appear that, the GSA method is considered to have given best results for multi-objective OPF based combined economic dispatch and emission because they obtained better results compared to those known references.

The developed GSA has been implemented and used to solve the OPF combined economic dispatch with valve-point effect and emission for IEEE 30-bus system under varying operating conditions. The cost function is considered to be quadratic function.

From [Figs. 7, 8](#), and [9](#), all cases study of multi-objective results obtained the minimum values after 120 iterations.

[Table 5](#) shows a comparison between the obtained single and multi-objective results of costs, power losses and emission with the results obtained in literature.

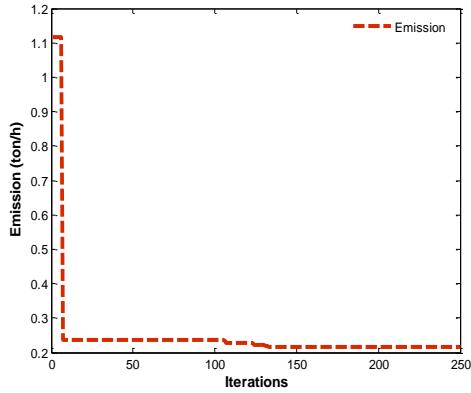


Fig. 6: Convergence of algorithm for case 4.

Table 2: Results of cases 4 and 5 for IEEE 30-bus system.

| Control variables | Optimal values | | |
|--------------------|----------------|---------------|---------------|
| | Case 4 | Case 5 | |
| P_{G2} (MW) | 55.2162 | 51.7103 | 48.8408 |
| P_{G5} (MW) | 48.4827 | 30.7894 | 30.4953 |
| P_{G8} (MW) | 22.8868 | 34.9894 | 34.9624 |
| P_{G11} (MW) | 29.9986 | 23.4313 | 19.1652 |
| P_{G13} (MW) | 31.4252 | 21.1699 | 21.3077 |
| V_1 (pu) | 1.0863 | 1.0713 | 1.0673 |
| V_2 (pu) | 1.0728 | 1.0581 | 1.0501 |
| V_5 (pu) | 1.0228 | 1.0321 | 1.0202 |
| V_8 (pu) | 1.0097 | 1.0420 | 1.0309 |
| V_{11} (pu) | 1.0493 | 1.0644 | 1.0623 |
| V_{13} (pu) | 1.0248 | 1.0602 | 1.0551 |
| Q_{com10} (MVAr) | 1.4597 | 3.7160 | 4.6011 |
| Q_{com12} (MVAr) | 4.2850 | 2.9271 | 4.2547 |
| Q_{com15} (MVAr) | 1.3286 | 3.9829 | 1.4724 |
| Q_{com17} (MVAr) | 3.0464 | 1.2142 | 4.5426 |
| Q_{com20} (MVAr) | 3.3183 | 2.6623 | 3.9967 |
| Q_{com21} (MVAr) | 3.1756 | 3.1168 | 4.9962 |
| Q_{com23} (MVAr) | 3.1649 | 1.3367 | 4.0620 |
| Q_{com24} (MVAr) | 3.0629 | 4.3422 | 3.5058 |
| Q_{com29} (MVAr) | 2.7770 | 3.3533 | 1.4227 |
| T_{6-9} | 0.9902 | 0.9953 | 0.9658 |
| T_{6-10} | 1.0132 | 0.9589 | 1.0312 |
| T_{4-12} | 0.9408 | 1.0010 | 0.9731 |
| T_{27-28} | 0.9755 | 0.9819 | 0.9557 |
| Cost in (\$/h) | 1025.9600 | 824.87 | 862.78 |
| loss in (MW) | 5.245 | 05.827 | 06.284 |
| Emission (ton/h) | 0.229 | 0.2524 | 0.2557 |
| Slack in (MW) | 100.636 | 127.1374 | 134.913 |
| CPU time (s) | 77.306 | 81.0773 | 80.756 |

The proposed method to solve the OPF combined economic dispatch with valve-point effect and emission is considered to have given the best results because the results obtained using the GSA method are better compared to those published recently in several researches papers.

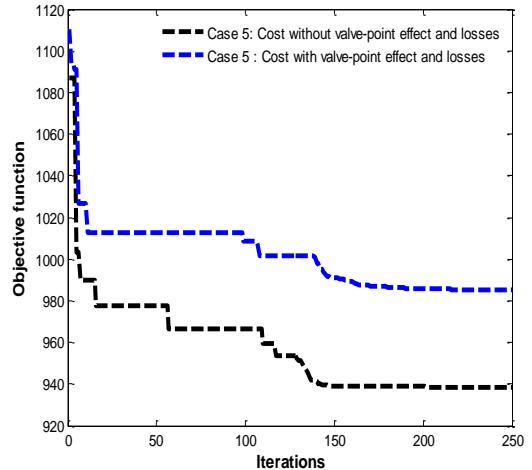


Fig. 7: Convergence of algorithm for case 5.

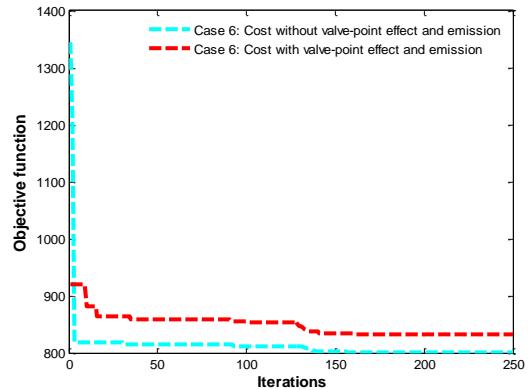


Fig. 8: Convergence characteristics for case 6.

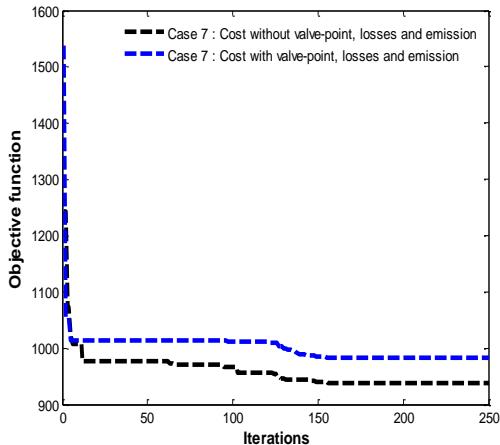


Fig. 9: Convergence characteristics for the case 7.

Through the results obtained in Table 5, we note that the optimal values of the different objective functions are affected by the change of weight factors. The larger the weight factor, the more optimal the value of the objective function.

Table 3: Results of cases 6 and 7 for IEEE 30-bus system.

| Control variables | Optimal values | | | |
|---------------------------|----------------|---------------|----------------|----------------|
| | Case 6 | | Case 7 | |
| | w/o valve | With valve | w/o valve | with valve |
| P _{G2} (MW) | 48.9786 | 42.8649 | 52.2267 | 47.7413 |
| P _{G5} (MW) | 21.1630 | 16.6036 | 30.5107 | 29.2543 |
| P _{G8} (MW) | 21.3428 | 13.9896 | 35.0000 | 34.9274 |
| P _{G11} (MW) | 11.8295 | 10.0373 | 24.4152 | 25.4630 |
| P _{G13} (MW) | 12.0012 | 12.0919 | 20.6417 | 17.2960 |
| V ₁ (pu) | 1.0809 | 1.0859 | 1.0717 | 1.0722 |
| V ₂ (pu) | 1.0624 | 1.0660 | 1.0577 | 1.0568 |
| V ₅ (pu) | 1.0325 | 1.0365 | 1.0327 | 1.0291 |
| V ₈ (pu) | 1.0375 | 1.0217 | 1.0424 | 1.0366 |
| V ₁₁ (pu) | 1.0459 | 1.0812 | 1.0821 | 1.0777 |
| V ₁₃ (pu) | 1.0180 | 1.0277 | 1.0597 | 1.0507 |
| Q _{com10} (MVAr) | 3.9395 | 2.2226 | 4.1934 | 3.0310 |
| Q _{com12} (MVAr) | 2.7670 | 1.6456 | 1.9471 | 3.1556 |
| Q _{com15} (MVAr) | 3.1526 | 1.4549 | 2.1550 | 2.0520 |
| Q _{com17} (MVAr) | 2.4079 | 3.7126 | 2.0085 | 3.9209 |
| Q _{com20} (MVAr) | 3.5253 | 2.0909 | 3.0977 | 1.8054 |
| Q _{com21} (MVAr) | 4.3362 | 0.7274 | 4.8694 | 4.6344 |
| Q _{com23} (MVAr) | 3.7058 | 1.3824 | 2.3529 | 3.5211 |
| Q _{com24} (MVAr) | 4.9791 | 4.1553 | 4.6928 | 3.2256 |
| Q _{com29} (MVAr) | 1.0175 | 1.8858 | 3.1813 | 1.9244 |
| T ₆₋₉ | 1.0588 | 0.9958 | 1.0198 | 0.9929 |
| T ₆₋₁₀ | 0.9963 | 1.0163 | 0.9503 | 1.0325 |
| T ₄₋₁₂ | 1.0190 | 0.9525 | 0.9884 | 1.0253 |
| T ₂₇₋₂₈ | 1.0084 | 0.9633 | 0.9882 | 0.9817 |
| Cost in (\$/h) | 800.89 | 834.69 | 825.170 | 862.90 |
| loss in (MW) | 09.157 | 10.951 | 5.775 | 6.1693 |
| Emission (ton/h) | 0.3203 | 0.3203 | 0.3203 | 0.3203 |
| Slack in (MW) | 177.246 | 198.764 | 126.381 | 134.887 |
| CPU time (s) | 74.072 | 84.152 | 83.059 | 79.020 |

Table 4: Results of case 7 with different weight factors.

| | <i>w_c</i> | <i>w_l</i> | <i>w_e</i> |
|-------------------------|----------------------|----------------------|----------------------|
| Cost (\$/h) | 0.9 | 0.08 | 0.02 |
| Cost (\$/h) | 862.90 | 864.71 | 865.10 |
| Losses (MW) | 6.1693 | 6.2012 | 6.1921 |
| Losses (MW) | 0.3203 | 0.3101 | 0.3100 |
| Emission (ton/h) | 0.8 | 0.15 | 0.15 |
| Cost (\$/h) | 863.10 | 864.71 | 865.10 |
| Losses (MW) | 6.9731 | 6.9958 | 6.9547 |
| Losses (MW) | 0.3199 | 0.3158 | 0.30258 |
| Emission (ton/h) | 0.5 | 0.25 | 0.25 |
| Cost (\$/h) | 863.80 | 864.71 | 865.10 |
| Losses (MW) | 6.3257 | 6.2012 | 6.1921 |
| Losses (MW) | 0.3302 | 0.3354 | 0.3434 |
| Emission (ton/h) | 0.338 | 0.335 | 0.327 |
| Cost (\$/h) | 866.65 | 862.90 | 862.90 |
| Losses (MW) | 6.7000 | 7.2121 | 7.123 |
| Losses (MW) | 0.24156 | 0.2315 | 0.3058 |
| Emission (ton/h) | 0.24156 | 0.2315 | 0.3058 |

Table 5: Comparison of obtained and literature results.

| Methods | Ref. | Cost (\$/h) | Losses (MW) | Emission (\$/ton) |
|-----------------|------|---------------|--------------|-------------------|
| <i>Case 5</i> | | | | |
| Proposed | - | 824.87 | 5.827 | 0.2524 |
| MSA | [36] | 859.191 | 4.540 | - |
| IABC | [19] | 854.913 | 4.982 | - |
| PSO | [31] | 878.873 | 7.810 | - |
| MDE | [25] | 820.880 | 5.594 | - |
| <i>Case 6</i> | | | | |
| Proposed | - | 800.89 | 9.157 | 0.267 |
| GA | [30] | 820.166 | - | 0.271 |
| MICA | [39] | 865.066 | - | 0.222 |
| <i>Case 7</i> | | | | |
| Proposed | - | 825.17 | 5.775 | 0.227 |
| GA | [30] | 793.605 | 8.450 | 0.187 |
| IABC | [19] | 851.611 | 4.873 | 0.223 |
| ABC | [18] | 854.916 | 4.982 | 0.228 |

4. CONCLUSION

The GSA method was successfully implemented in this paper to find the optimum OPF control variables for single objective and multi-objective optimization. The versatility of the multi-objective OPF optimization is illustrated by different tests systems by changing the parameters of GSA method such as population size N_a and control parameters, α and G_0 . The analysis performance of used methodology is illustrated by the numerical and graphical results as shown in all tables and figures. The proposed method has fast convergence time in all cases test due of obtained performance. Through the obtained results, the power generation cost, active losses and emission index were significantly reduced to 5.85%, 61.61% and 44.63% %, respectively, from the base case and these results obtained are considered good results compared to some references. The effectiveness and robustness of used method are demonstrated by the obtained results. Therefore, it can be recommended to future researchers as a promising algorithm for solving some more complex engineering optimization problems. However, we have to mention that it becomes slow if the numbers of system variables are increased. It is found that the average CPU time increases rapidly as system size increases and convergence slows down.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Nabil Mezhoud: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Mohamed Amarouayache:** Conceptualization, Formal analysis, Investigation, Project administration, Resources Software, Supervision, Visualization, Writing - original draft.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double

publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] J. H. Talaq, F. El-Hawary, and M. E. El-Hawary, "A summary of environmental/economic dispatch algorithms," *IEEE Transactions on Power Systems*, vol. 9, no. 3, pp. 1508-1516, 1994.
- [2] C. Kumar, and Ch. P. Raju, "Constrained OPF using Particle Swarm Optimization," *Int. J. of Em. Tech. and Adv. Eng.*, vol. 2, no 2, pp. 235-241, 2012.
- [3] A. A. Mohamed, A. A. M. El-Gaafary, Y. S. Mohamed, and A. M. Hemeida, "Multi-Objective Modified Grey Wolf Optimizer for OPF," in *IEEE Eighteenth International Middle East Power Systems Conference (MEPCON)*, 2016.
- [4] H. W. Dommel, and W. F. Tinney, "OPF Solutions", *IEEE Transactions on Power Apparatus and Systems*, vol. 87, no. 10, pp. 1866-1876, 1968.
- [5] S. A. H. Soliman, and A. A. H. Mantawi, *Modern Optimization Techniques with Applications in Electric Power System*. Springer, New York, USA, 2012, pp. 281-346.
- [6] L. L. Lai, and J. T. Maimply, "Improved genetic algorithms for optimal power flow under both normal contingent operation states," *International Journal of Electrical Power & Energy Systems*, vol. 19, no. 5, pp. 287-292, 1997.
- [7] H. Nicholson, and M. J. H. Sterling, "Optimum dispatch of active and reactive generation by quadratic programming," *IEEE Transactions on Power Apparatus and Systems*, vol. 72, pp. 644-654, 1973.
- [8] Costa, G. R. M. da., Costa, and C. E. U., "Improved newton method for optimal power flow problem," *International Journal of Electrical Power and Energy Systems*, vol. 22, no. 7, pp. 459-462, 2000.
- [9] T. C. Giras, and S. N. Talukdar, "Quasi-newton method for optimal power flows," *International Journal of Electrical Power and Energy Systems*, vol. 3, no. 2, pp. 59-64, 1981.
- [10] B. Stott, and E. Hobson, "Power system Security control calculations using linear programming, Part-2," *IEEE Transactions on Power Apparatus and Systems*, vol. 97, no. 5, pp. 1721-1731, 1978.
- [11] A. M. Sasson, "Nonlinear programming solutions for load-flow, minimum-loss, and economic dispatching problems," *IEEE Transactions on Power Apparatus and Systems*, vol. 88, no. 4, pp. 399-409, 1969.
- [12] F. Capitanescu, M. Glavic, D. Ernst, and L. Wehenkel, "Interior-point based algorithms for the solution of optimal power flow problems," *Electric Power Systems Research*, vol. 77, pp. 508-517, 2007.
- [13] L. L. Lai, *Intelligent System Applications in Power Engineering: Evolutionary Programming and Neural Networks*. Wiley, New York, USA, 1998.
- [14] S. Mouassa, and T. Bouktir, "Artificial bee colony algorithm for solving OPF problem considering the valve point effect," *International Journal of Computer Applications*, vol. 112, no. 1, pp. 45-53, 2015.
- [15] D. Aydin, and S. Özyön, "Solution to non-convex economic dispatch problem with valve point effects by incremental artificial bee colony with local search," *Applied Soft Computing*, vol. 13, no. 5, pp. 2456-2466, 2013.
- [16] Z. Zakaria, T. K. A. Rahman, and E. E. Hassan, Economic Load Dispatch via an Improved Bacterial Foraging Optimization, in *IEEE 8th International Power Engineering and Optimization Conference*, Langkawi, The Jewel of Kedah, Malaysia, 2014.
- [17] S. M. Abd-Elazim, and E. S. Ali, "Optimal power system stabilizers design via Cuckoo search algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 75, no. 1, pp. 99-107, 2016.
- [18] A. H. Khazali, and M. Kalantar, "Optimal reactive power dispatch based on harmony search algorithm," *International Journal of Electrical Power and Energy Systems*, vol. 33, no. 3, pp. 684-692, 2011.
- [19] J. Yuryevich, and K. P. Wong, "Evolutionary programming based optimal power flow algorithm," *IEEE Transactions on Power Systems*, vol. 14, no. 4, pp. 1245 - 1250, 1999.
- [20] A. A. Abou El Ela, M. A. Abido, and S. R. Spea, "Optimal power flow using differential evolution algorithm," *Electric Power Systems Research*, vol. 91, no. 7, pp. 878-885, 2010.
- [21] S. Sayah, and K. Zehar, "Modified differential evolution algorithm for optimal power flow with non-smooth cost functions," *Energy Conversion and Management*, vol. 4, pp. 3036-3042, 2008.
- [22] M. A. Abido, "Optimal power flow using Tabu search algorithm," *Electric Power Components and Systems*, vol. 30, N°. 5, pp. 469-483, 2010.
- [23] Y.J. Jeon, and J. C. Kim, "Application of simulated annealing and Tabu search for loss minimization in distribution systems," *International Journal of Electrical Power and Energy Systems*, vol. 26, no. 1, pp. 9-18, 2004.
- [24] A. R. Bhowmik, and A. K. Chakraborty, "Solution of optimal power flow using non-dominated sorting multi-objective gravitational search algorithm," *International Journal of Electrical Power and Energy Systems*, vol. 62, no. 4, pp. 323-334, 2014.
- [25] S. S. Reddy, P. R. Bijwe, and A. R. Abhyankar, "Faster evolutionary algorithm based optimal power flow using incremental variables," *International Journal of Electrical Power & Energy Systems*, vol. 54, no. 1, pp. 198-210, 2014.
- [26] C. Yasar, and S. Özyön, "A new hybrid approach for nonconvex economic dispatch problem with valve-point effect," *Energy*, vol. 36, no. 10, pp. 5838-5845, 2011.

- [27] M. A. Abido, "OPF using particle swarm optimization," *Electrical Power and Energy System*, vol. 24, no. 1, pp. 563-571, 2002.
- [28] J. Y. Kim, H. S. Lee, and J. H. Park, "A modified particle swarm optimization for OPF," *Journal of Electrical Engineering and Technology*, vol. 2, no. 4, pp. 413-419, 2007.
- [29] A. Ketabi, and A. A. R. Feuillet, "Application of the ant colony search algorithm to reactive power pricing in an open electricity market," *International Journal of Electrical Power & Energy Systems*, vol. 32, no. 6, pp. 622-628, 2010.
- [30] A. El-Fergany, and H. M. Hasanien, "Tree-seed algorithm for solving optimal power flow problem in large-scale power," *Syst. Inc. Val. and Comp.*, vol. 64, no. 1, pp. 307-316, 2018.
- [31] A. A. A. Mohamed, Y. S. Mohamed, A. A. M. El-Gaafary, and A. M. Hemeida, "Optimal power flow using moth swarm algorithm," *Electric Power Systems Research*, vol. 142, pp. 190-206, 2017.
- [32] A. F. Attia, R. A. El-Sehiemy, and H. M. Hasanien, "OPF solution in power systems using a novel sine-cosine algorithm," *Electrical Power and Energy Systems*, vol. 99, no. 3, pp. 331-343, 2018.
- [33] S. S. Padaiyatchi, "Hybrid DE/FFA algorithm applied for different optimal reactive power dispatch problems," *Australian Journal of Electrical and Electronics Engineering*, vol. 17, no. 3, pp. 203-210, 2020.
- [34] M. Ghasemi, S. Ghavidel, M. M. Ghanbarian, M. Gharibzadeh, and A. A. Vahed, "Multi-objective optimal power flow considering the cost, emission, voltage deviation and power losses using multi-objective modified imperialist competitive algorithm," *Energy*, vol. 2014, pp. 1-14, 2014.
- [35] A. K. Khamees, A. El-Rafei, N. M. Badra, and A. Y. Abdelaziz, "Shuffled frog leaping algorithm," *International Journal of Engineering, Science and Technology*, vol. 9, no. 1, pp. 55-68, 2017.
- [36] H. R. El-Hana Bouchekara, M. A. Abido, and A. E. Chaib, "Optimal power flow using an improved electromagnetism-like mechanism method," *Electric Power Components and Systems*, vol. 44, no. 4, pp. 434-449, 2016.
- [37] I. N. Trivedi, P. Jangir, and S. A. Parmar, "Optimal power flow with enhancement of voltage stability and reduction of power loss using ant-lion optimizer," *Cogent Engineering*, 2016.
- [38] B. Bentouati, S. Chettih, and L. Chaib, "Interior search algorithm for optimal power flow with non-smooth cost functions", *Cogent Engineering*, vol. 2017, pp. 1-17, 2017.
- [39] R. Senthilkumar, P. Sk. Karimulla, K. B. V. S. R. Subrahmanyam, and R. Deshmukh, "Solution for optimal power flow problem using WDO algorithm," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 2, pp. 889-895, 2021.
- [40] F. Hasan, A. Kargarian, and A. Mohammadi, "A survey on applications of machine learning for optimal power flow," in *IEEE Texas Power and Energy Conference (TPEC)*, 2020.
- [41] S. Gupta, N. Kumar, L. Srivastava, H. Malik, A. Anvari-Moghaddam, and F. P. A. García Márquez, "Robust optimization approach for optimal power flow solutions using Rao algorithms," *Energies*, vol. 14, pp. 1-28, 2021.
- [42] K. Bhattacharjee, K. Shah, and J. Soni, "Solving economic dispatch using artificial EcoSystem-based optimization," *Electric Power Components and Systems*, vol. 50, no. 1, 2022.
- [43] H. T. Ul-Hassan, M. F. Tahir, K. Mehmood, K. M. Cheema, A. H. Milyani, and Q. Rasool, "Optimization of power flow by using Hamiltonian technique," *Energy Reports*, vol. 6, no. 11, pp. 2267-2275, 2020.
- [44] E. Akbari, M. Ghasemi, M. Gil, A. Rahimnejad, and S. A. Gadsden, "Optimal power flow via teaching-learning-studying-based optimization algorithm," *Electric Power Components and Systems*, vol. 49, no. 6-7, pp. 584-601, 2021.
- [45] L. Dilip, R. Bhesdadiya, R. I. Trivedi, and P. Jangir, "OPF Problem Solution using Multi-objective Grey Wolf Optimizer Algorithm" in *Intelligent Communication and Computational Technologies, Networks and Systems*, 2018, pp. 191-201.
- [46] N. Mezhoud, B. Ayachi and B. Ahmed. Wind Driven Optimization Approach based Multi-objective OPF and Emission Index Optimization. *International Research Journal of Multidisciplinary Technovation*, vol. 4, no. 2, pp. 21-41, 2022.
- [47] S. Jiang, Z. Ji, and Y. Shen, "A novel hybrid particle swarm optimization and gravitational search algorithm for solving economic emission load dispatch problems with various practical constraints," *Electrical Power and Energy Systems*, vol. 55, no. 11, pp. 628-644, 2014.
- [48] Rashedi E, Rashedi E, Nezamabadi-pour, and H. Saryazdi, "GSA: a gravitational search algorithm," *Information Sciences*, vol. 179, no. 13, pp. 2232-2248, 2009.
- [49] S. Duman, U. Güvenç, Y. Sönmez, and N. Yörükeren, "Optimal power flow using gravitational search algorithm," *Energy Conversion and Management*, vol. 59, pp. 86-95, 2012.
- [50] S. Jiang, C. Zhang, and S. Chen, "Sequential hybrid particle swarm optimization and gravitational search algorithm with dependent random coefficients," *Hindawi Mathematical Problems in Engineering*, vol. 2020.
- [51] S. Mirjalili, S. Zaiton, and M. Hashim, "A new hybrid PSOGSA algorithm for function optimization," *International in 2010 Conference on Computer and Information Application*, pp. 374-377, 2010.

[52] K. Y. Lee, and M. El-Sharkawi, *Modern heuristic optimization techniques: Theory and applications in power systems*, New York, USA, 2008.



BIOGRAPHY

Nabil Mezhoud is a Lecturer Professor at the University of Skikda, Algeria. He received the Ph.D. degree in Electrical Engineering in 2017. He published several research papers in conferences, journals and reviews. His areas of interest are: modeling, simulation and application of FACTS and HVDC systems, application of intelligent techniques to optimal power flow (OPF) problem, hybrid and multi-objective OPF, power system stability and control, integrating of renewable energy into electrical

networks and smart grid systems. Email: mezhounab@yahoo.fr.



Mohamed Amarouayache is a Lecturer Professor at the University of Skikda, Algeria. He received the Ph.D. degree in Electrical Engineering in 2014. He published several research papers in conferences, journals and reviews. Currently, he is member of LES laboratory. His areas of interest are: renewable energy modelling and simulation, power system control and intelligent techniques, integrating renewable energy into electrical networks and smart grid systems. Email: amarouayachemohamed@yahoo.fr.

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Smart AI-based Video Encoding for Fixed Background Video Streaming Applications

Mohammadreza Ghafari¹ , Abdollah Amirkhani^{2,*} , Elyas Rashno³ , and Shirin Ghanbari⁴

¹Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

²School of Automotive Engineering, Iran University of Science and Technology, Tehran, Iran

³Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

⁴Department of Computer Science and Electronic Engineering, University of Essex, Essex, UK

* Corresponding Author: amirkhani@ieee.org

Abstract: This paper is an extension of our previous research on presenting a novel Gaussian Mixture-based (MOG2) Video Coding for CCTVs. The aim of this paper is to optimize the MOG2 algorithm used for foreground-background separation in video streaming. In fact, our previous study showed that traditional video encoding with the help of MOG2 has a negative effect on visual quality. Therefore, this study is our main motivation for improving visual quality by combining the previously proposed algorithm and color optimization method to achieve better visual quality. In this regard, we introduce Artificial Intelligence (AI) video encoding using Color Clustering (CC), which is used before the MOG2 process to optimize color and make a less noisy mask. The results of our experiments show that with this method the visual quality is significantly increased, while the latency remains almost the same. Consequently, instead of using morphological transformation which has been used in our past study, CC achieves better results such that PSNR and SSIM values have been shown to rise by approximately 1dB and 1 unit respectively.

Keywords: Artificial Intelligence, video coding, background subtraction, color clustering, mixture of gaussian model.

Article history

Received 14 March 2022; Revised 28 June 2022; Accepted 06 August 2022; Published online 28 February 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

M. Ghafari, A. Amirkhani, E. Rashno, and S. Ghanbari, "Smart AI-based video encoding for fixed background video streaming applications," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 37-44, 2023. DOI: [10.22055/jaree.2022.40295.1051](https://doi.org/10.22055/jaree.2022.40295.1051)



1. INTRODUCTION

Artificial Intelligence (AI), as a branch of computer science, has been able to affect people's lives. More specifically, AI has provided algorithms for computer professionals that were previously difficult to implement. Nowadays, AI is widely used within image processing applications including face recognition or object detection. However, due to the complexities and challenges of video compression, the feasibility of using these algorithms in video compression has been less used. One of the main applications of video compression is in traffic cameras and video surveillance (CCTV) that are continuously streaming data to their users. Since these CCTV cameras are streaming video to their destinations, day and night, it seems necessary to use a

method that can take up less bandwidth. Practically, CCTV is one of the most important technologies in the security field.

Today, due to network limitations, CCTVs are applied at variant private and public places [1]. CCTV is used within real applications including identity offenders and prevent crime purposes [2]. An efficient CCTV technology was proposed by Harikrishnan [3] for business activities. In this paper, the author studied the impact of area condition on the CCTV technology and its utilization, and by using data analysis statistics, they proposed practical ideas in improving CCTV cameras. To address its challenges, Carli [4] tried to introduce CCTVs as a tool for crime prevention. Also, Goold [5] tried to address challenges in regards to CCTV responsibilities. On the other hand, Welsh and Farrington [6]

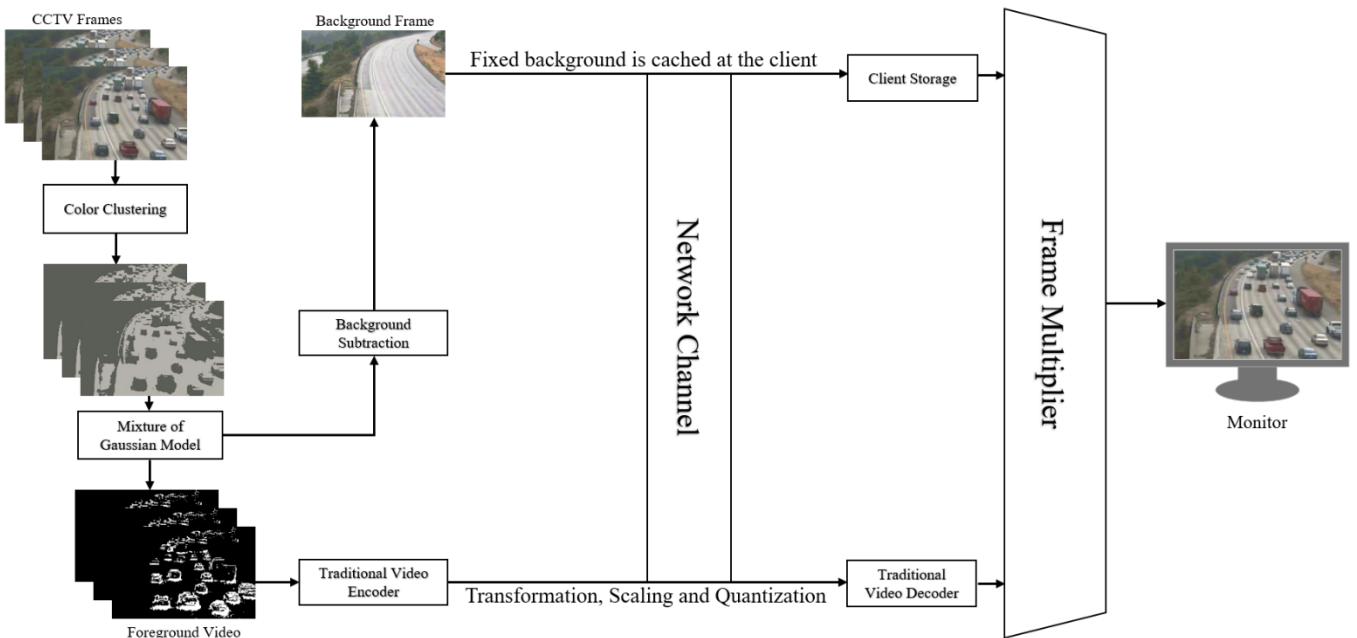


Fig. 1: Our proposed structure for AI-based video coding.

suggested a new method to improve the quality of CCTV and its efficiency. In order to use the CCTV camera as an analytical tool with image processing, the important features of 3D surveillance are extracted by the OpenCV library to introduce a new method of video analytics [7]. Another use of CCTV is as an analytical tool as proposed by Kumar et al. [8], whereby they define a new model for shot detection in MATLAB. The use of Digital Video Recorder (DVR) and Network Video Recorder (NVR), both are major components of CCTV cameras that perform compression and video capturing. In this way, a video compression method specifically for CCTVs can reduce the bandwidth but has an impact on image quality [9]. Another challenging issue with CCTV cameras is the limitation of its storage, which has been the subject of several studies. The main reason of this challenge is due to the large video files which are constantly recorded all the time. To overcome this challenge, cloud-based services are now available that can provide unlimited storage space [10]. Also, there are methods relating to storage optimization as to optimize CCTV storage and remove similar frame footage. This process may lead to the reduction of video size [11].

Today, video image enrichment in CCTV cameras is performed by artificial intelligence algorithms. This enrichment is achieved through visual models, the most well-known being object tracking, pattern recognizing and face detection [12]-[14]. CCTV image transfer optimization depends on video compression algorithms. The main reason for this claim is that by optimizing the video compression, the volume of the transmitted video is reduced and in addition to requiring less storage space, it also occupies less bandwidth.

Among conventional video coding techniques, block-based estimation is a practical method to reduce temporal redundancy in video streaming. In this way, a background-base model overlaying on High Efficiency Video Coding (HEVC) in a surveillance video streaming can reduce the network bandwidth occupation [15]. In a similar work, Zhang et al. [16] proposed a background-based method that tries to

separate the background from the foreground for achieving better accuracy in surveillance video coding. In this paper, the authors have used two methods namely, Background Reference Prediction (BRP) and Background Differences Prediction (BDP) to encode the video frames. In line with image transfer problems, Guo et al. [17] proposed a novel algorithm to overcome the high-quality of large videos files during the transformation process by background removal. Predicting adaptive motion units, namely, PA-search is a novel predictive method that has been proposed for motion estimation [18]. In this approach the author defined a smart algorithm for searching the motions of video frames while keeping algorithm compression complexities at a minimum. In another similar work, Kim and Lee [19] have proposed an efficient method to separate the background from foreground based on the integer motion estimation method for video compression. This video coding system helps to reduce the external power consumption to the encoder side and optimize its efficiency. Double background-based coding, as proposed by Li et al. [20] is defined by two different background frames based on reconstructed frames and the original frames, simultaneously. This method, tested by HM14.0, helped to improve the compression performance by saving the average of 17.32 bit rate. The distance parameter is one of the challenges of CCTVs, which makes increases the complexity of face identification. To overcome this challenge, a two phased object recognition is defined by Celine and Agustin [21] that enhances the facial quality. At the first phase the dataset including faces are created, then at the second phase, good quality images are produced. Moreover, automatically detecting illegal activities such as weapon detection can be critical for security applications. However, as the detection of such objects is time consuming, it practically affects the efficiency of such systems. In this regard, the weapon detection system has been optimized using Yolov4 with an accuracy of 90% in detection [22]. Another detection system has been investigated by Powale et al. [23] that identifies people even in low resolution images. The results of the study show that the proposed detection system based on

Convolutional Neural Networks (CNNs), has been able to reach an accuracy of 94.55% for the TinyFace [24] dataset. For object detection, between Deep Learning models, YOLO has shown to outperform. Pillai et al. [25] have proposed a Mini-YOLO method that has comparable accuracy with YOLO, however, the model size and computational cost were reduced significantly. In another application, YOLO has been used for real-time applications due to its high speed in object detection. In this regard, a novel real-time approach based on machine learning and deep learning methods to detect human faces in CCTV images has been investigated by Rehmat Ullah et al. M. S. Pillai, G. Chaudhary, M. Khari, and R. G. Crespo, "Real-time image enhancement for an automatic automobile accident detection through CCTV using deep learning", *Soft Computing*, vol. 25, no. 18, pp. 11929-11940, 2021.

[26]. The novel part of this research refers to the pre-processing part that enhances the image quality for better detection. In a similar work, Pan et al. [27], use YOLOv3 with the help of the COCO dataset to achieve better detection from CCTV video streaming. The experimental results of this method showed that YOLOv3 has performed 44% accuracy during the day and 41% accuracy during the night.

Reviewing these works, have motivated us to define a novel video coding system, which can efficiently reduce the network bandwidth. The fundamental concept of this coding is referred to the background subtraction process, whereby the fixed background is sent through the channel only once. In this regard, it is cached at the client side and is multiplied to the moving objects.

The rest of the paper is arranged as follows. **Section 2** describes the proposed scientific methodology and its experimental results are presented in **Section 3**. Finally, **Section 4** concludes this paper and discusses future work.

2. PROPOSED METHOD

The proposed compression technique extends the authors previous work [28] based upon an improved Mixture of Gaussian Model, MOG2 algorithm to separate the background image from moving objects in the video, while for getting better visual quality, Color Clustering (CC) method is added before the MOG2 process. In fact, instead of compressing the entire video frame, this technique causes the video frames to be separated into moving objects and the background image and whereby the background image occupies the network bandwidth only once. In other words, the moving objects frames create the foreground video while the background image creates the background video. Our previous results showed that for using background subtraction, the produced mask from MOG2 is very noisy and it should be denoised using Morphological Transformation (MT). As the visual quality of this transformation is not satisfactory, the first step of the proposed architecture is to use CC as a clean mask.

Fig. 1 shows the overall process of compressing and retrieving video images on a Network Video Recorder (NVR). In this process, video frames are captured by the CC block for clustering the colors. Then it passes to MOG2 block to estimate the fixed background and separate foreground

from background parts. At the end of this block, the background subtraction operation is started to detect and extract moving objects. It is necessary to pass one full cycle for achieving background and after the CC and MOG2 process, the fixed background image is cached at the client-side. At the server's side, the foreground video is going to be compressed by conventional video codecs. In our approach, we have used H.265, namely the HEVC video codec to compressed moving objects. At the end, to the decoder side, the stored background image and encoded foreground video are multiplied to achieve the full video frames including foreground and background. With these explanations, our proposed architecture consists of three main parts, each of which is discussed below.

2.1. Color Clustering

The idea of using CC is due to its capability to better separate the components of objects. In our study, as the foreground objects includes the moving objects, it would be more efficient to separate colors then applying the background subtraction method. In this regard, as we wanted to separate colors in order for having better masks, and we have just two kinds of objects, foreground and background objects, we have set CC=2 in order for getting a denoised mask. Clustering method has been done by k-means color quantization which in RGB channel we have used Euclidean distance for comparing and clustering the colors. For every colorful image we have three channels including red, green and blue which creates the color-based image. Since we are looking for the distance of each color, the distance between two colors calculates by (1):

$$d = \sqrt{(R_2 - R_1)^2 + (G_2 - G_1)^2 + (B_2 - B_1)^2} \quad (1)$$

where R, G, B are the representatives of the red, green and blue channels.

2.2. Gaussian Mixture Model

The basis of the MOG2 method is a pixel-by-pixel review in a video sequence in which motion information is obtained from differences between the frames. Using this method, moving objects are detected that can be a tool for creating a mask. Multiplying this mask throughout the video will separate the background image from moving objects. The process of creating the mask is relevant to Gaussian distribution.

For every Gaussian distribution η we have

$$\eta(t) = \frac{1}{\sqrt{2\pi}\omega} \exp(-(t - \mu)^T \omega^{-1} (t - \mu)) \quad (2)$$

where μ is the mean vector and ω is the covariance matrix. If we assume that there are k Gaussian distributions according to each t variable, which refers to the frame sequences in the time domain, the Gaussian Mixture Model (GMM) is a mixture of k Gaussians that describes the random variable t .

This assumption leads to (3):

$$Pr(t) = \sum_{k=1}^{k=K} W_k * \eta(t) s.t: \sum_{k=1}^{k=K} W_k = 1 \quad (3)$$

where W_k is the k^{th} weight of the Gaussian model.

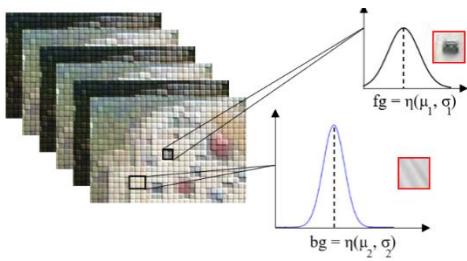


Fig. 2: Foreground and background Gaussian modelling.

Since the background frames appear frequently during video streaming, the ideal background Gaussian model should have a high weight with low variance. To select the first n Gaussians as Background Models (BM) we have:

$$BM = \operatorname{argmin}_n \left(\sum_{k=1}^n W_k > T_n \right) \quad (4)$$

where T_n is the model selection threshold. The range of this threshold for MOG2 model is from 1 to 255. Based on Monte Carlo analysis as performed by Matczak et al. M. Ghafari, A. Amirkhani, E. Rashno, S. Ghanbariet, "Novel gaussian mixture-based video coding for fixed background video streaming," in 2022 12th Iranian/Second International Conference on Machine Vision and Image Processing (MVIP), Ahvaz, Iran, 2022.

[29] we set this variable threshold to 16, which exactly fits our requirements. During the subtracting process, if block B is matched with BM, it would be labelled as background block, otherwise it is selected as the foreground block. Fig. 2 depicts the subtraction processing phase for creating the foreground and background model.

2.3. Video Encoder

The video compression technique has been used for many years. Due to the H.264 codec, an evolution was made in video compression techniques. Recently, with the standardization of the H.265 codec, many devices will support this codec, which can significantly decrease the bandwidth being used. Due to the popularity and advantages of this codec, this research has also used this codec.

2.4. Frame Multiplier

By separating the background image from the original video frames, on the decoder side, the image must be multiplied by the foreground frames that have been compressed by the video encoder to restore the original frames. In fact, at this point, it is enough to multiply the moving objects on the pre-cached background image to restore the same original frames. Since the detection of moving objects by the MOG2 method is faced with a small error, in some parts of the frame noise is visible and whereby causes this method not to be an ideal filter to separate the background image from moving objects. In this regard, the proposed methodology used MT to remove noises and to create an ideal mask.

3. EXPERIMENTAL RESULTS

In applying the proposed methodology, a highway video streaming dataset [30] that completely covers the mentioned

requirement is used. The camera position is fixed such that the background can be extracted from the main frames. By applying this method, the experimental results of this article can be examined in four areas.

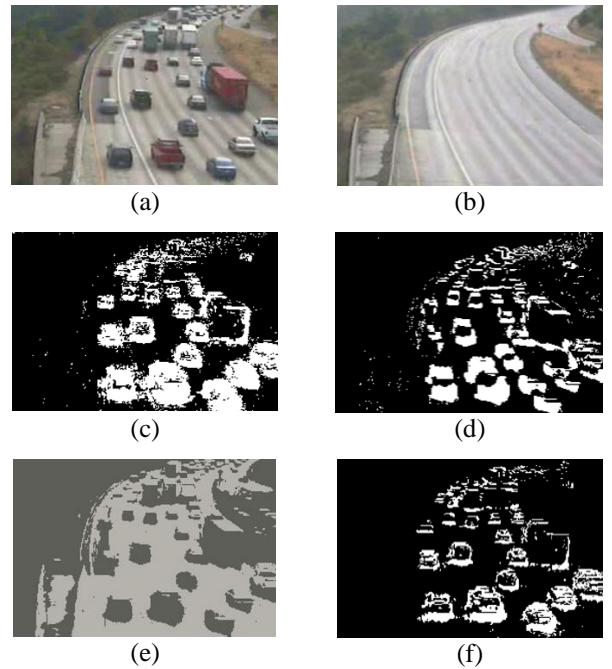


Fig. 3: Procedure of creating an ideal mask from the original frame, (a) Original frame, (b) Subtracted background, (c) MOG2 mask, (d) MOG2 mask after MT, (e) CC frame, (f) Ideal mask (CC+MOG2).

3.1. Numerical Calculations

Due to the removal of the background image from the main video frames and the compression of only foreground frames, pixels per frame will be significantly reduced. In fact, by reducing the number of pixels per frame, the number of numerical calculations for the compression operation at the encoder side will be reduced by the same amount. Fig. 3 illustrates the number of pixels reduced by comparing the 100 original frames for both MT+MOG2 and CC+MOG2 methods. Analyzing this figure shows that in the initial 100 frames, the total number of pixels per frame will be reduced by an average of 35000 pixels for MT+MOG2 method while the CC+MOG2 gives a better reduction, 38000 pixels reduction per frame. In other words, 45.57% of the processing load is reduced by MT+MOG2 method and 49.47% by CC+MOG2 method. Henceforth, the CC method will result in better reductions as its associated mask is cleaner as compared to the MT method.

3.2. Bandwidth Transmission

The occupied network bandwidth will decrease dramatically for two reasons. The first reason is that due to the subtraction process, the background image is occupied the bandwidth only once. The second reason is that the bandwidth occupation rate is directly dependent on the content within the frames and since the content of each frame is reduced, the transmitted bandwidth will be saved.

Fig. 6 illustrate the I/P/B frames bandwidth occupation rate for three methods. Since the I-frame occupies a very large amount of bandwidth, it will be very important to check this

frame rate. This research illustrates that due to the removal of large quantities of pixels, which is referred to the separation of foreground and background frames, it has been able to

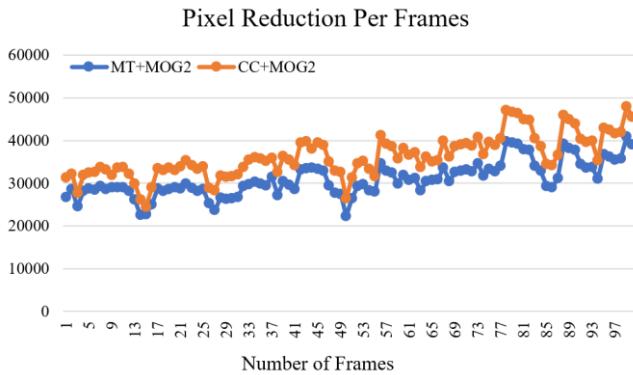


Fig. 4: Pixel reduction per frames to reduce the total calculation.

reduce the bandwidth occupation by 48% in the I-frame using MT+MOG2 and 67.4% through CC+MOG2. In return, for this significant reduction in the reference frames, P/B-frames have been increased to a very small extent, which will take up very little bandwidth compared to reference frames.

3.3. Visual Quality

Due to the rapid changes of moving objects in the real-time video streaming, noises will be present in the Gaussian filter. To optimize this noisy mask to an ideal mask, our proposed last methodology used MT to reduce noises. Due to applying several filters during the AI process, there will have a slight reduction in visual quality compared to the original video. However, in our newer method, instead of using MT, we have used CC to achieve better results.

Structural Similarity Index Measurement (SSIM) and Peak Signal to Noise Ratio (PSNR) algorithms are useful visual quality assessment algorithms in which the aim is to examine the qualitative changes compared to the original CCTV video.

Fig. 8 and **Fig. 9** compare the visual quality of three compressed videos using the CC+MOG2, MT+MOG2 and, traditional solutions. The investigation of the above compressed videos for PSNR and SSIM values illustrates an approximate 1 dB and 1 unit improvement by using CC+MOG2 in comparison to only using MT+MOG2 solution.

3.4. Latency

As the proposed coding system uses a background subtraction process, it is expected that latency is slightly increased. The main part of the process which pushes delays in an end-to-end cycle refers to MT process, which needs to scan whole the frame then creates a noise free mask. In our newer strategy, as we have used CC instead of MT, it is expected to perform better in reducing the total latency. The FFMPEG encoding process depicts that approximately twice the time is required for MT+MOG2 encoding, and the use of the CC+MOG2 method has resulted in very little reduction in latency. This result shows that the majority of latency is due to using MOG2 and not MT or CC processes. **Fig. 10**

illustrates the encoding latency for the three above methodologies.

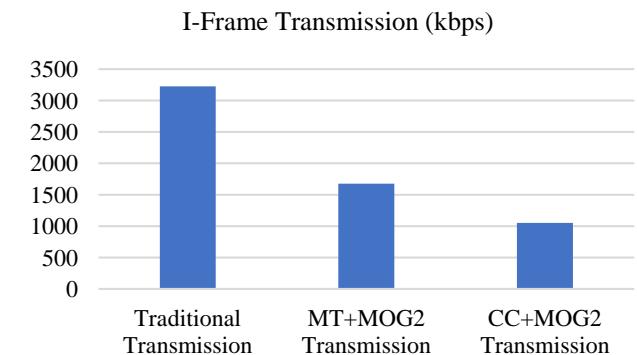


Fig. 5: Comparison of I-Frame transmission reductions.

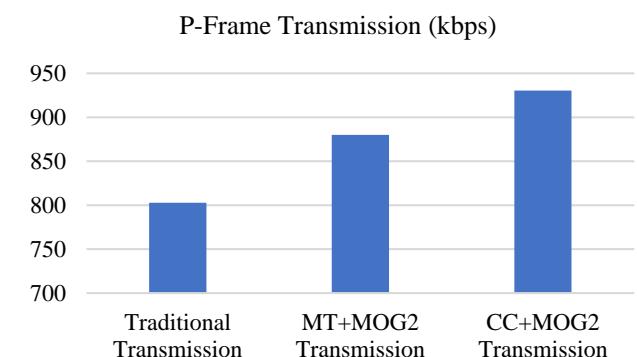


Fig. 6: Comparison of P-Frame transmission reductions.

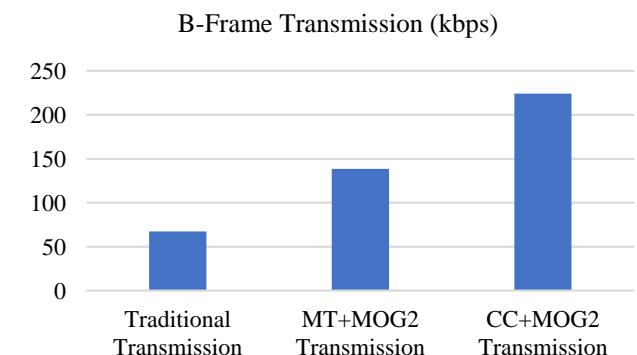


Fig. 7: Comparison of B-Frame transmission reductions.

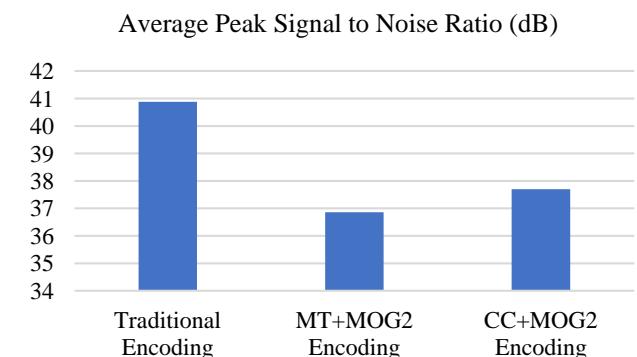


Fig. 8: The PSNR comparison of traditional, MT+MOG2 and CC+MOG2 encoding.

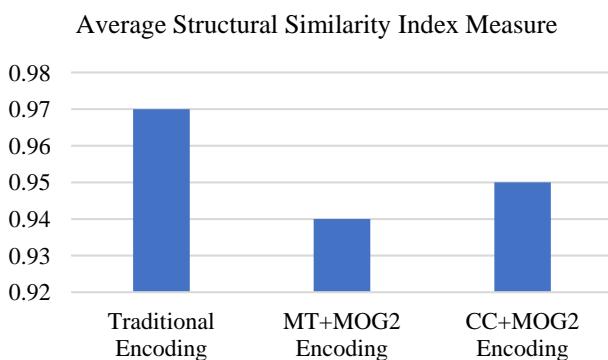


Fig. 9: The SSIM comparison of traditional, MT+MOG2 and CC+MOG2 encoding.

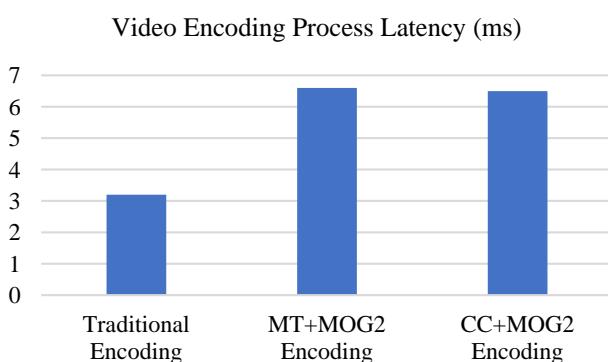


Fig. 10: Comparison of video encoding processes.

4. CONCLUSION

In this paper, an end-to-end AI-based video coding for separating the foreground and background frames with the help of using CC method has been improved. Through this procedure the total number of encoder computations was reduced while the visual quality reduction was improved by the CC process. Also, as the background is sent just one at a time, the network bandwidth occupation is reduced by 48% by MT+MOG2 and 67.4% by CC+MOG2. In visual quality assessment, 1dB and 1 unit improvement in PSNR and SSIM were achieved through the proposed method. For future works, the CC process can be changed by investigation on adaptive thresholding method to check if the optimization and improvement would be increased.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Mohammadreza Ghafari: Conceptualization, Data curation, Formal analysis, Methodology, Project administration. **Abdollah Amirkhani:** Project administration, Supervision, Validation. **Elyas Rashno:** Investigation, Methodology. **Shirin Ghanbari:** Validation, Writing - original draft, Writing - review & editing.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] S. Arora, K. Bhatia, and V. Amit, "Storage optimization of video surveillance from CCTV camera," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 710-713.
- [2] S. Vítek, M. Klíma, and L. Krasula, "Video compression technique impact on efficiency of person identification in CCTV systems," in *2014 International Carnahan Conference on Security Technology (ICCST)*, 2014, pp. 1-5.
- [3] G. Harikrishnan, "Role of CCTV in business organization: A case study," *International Journal of Commerce, Business and Management*, vol. 3, no. 3, pp. 466-470, 2014.
- [4] V. Carli, "Assessing CCTV as an effective safety and management tool for crime-solving, prevention and reduction," Montreal, 2008. [Online]. Available: <https://policycommons.net/artifacts/1217708/assessing-cctv-as-an-effective-safety-and-mangement-tool-for-crime-solving-prevention-and-reduction/1770796>
- [5] B. Goold, "Public area surveillance and police work: The impact of CCTV on police behaviour and autonomy," *Journal of Surveillance and Society*, vol. 1, no. 2, pp. 191-203, 2003.
- [6] B. C. Welsh, and D.P. Farrington, "Effects of closed circuit television surveillance in reducing crime," *Campbell Systematic Reviews*, vol. 4, no. 1, pp. 1-73, 2008.
- [7] B. Srilaya, LV. Kumar, and LNP. Boggavarapu, "Surveillance using video analytics," in *Proc. of Int. Conf. on Computational Intelligence and Information Technology*, 2013.
- [8] Kumar, M. B. Punith, and P. S. Puttaswamy, "Video to frame conversion of TV news video by using MATLAB," *International Journal of Advance Research in Science and Engineering*, vol. 3, no.3, pp. 95-101, 2014.
- [9] D. R. Zaghar, and TE. Abdulabbas, "A Compression Algorithm for Video Surveillance System," *International Journal of Computer Applications*, vol. 118, no.24, pp. 19-22, 2015.
- [10] CCTV Services (Jan. 6, 2023). Unlimited Video Storage. [Online]. Available: <http://www.cctvservices.net/unlimited-video-storage/>
- [11] S. Arora, K. Bhatia, and V. Amit, "Storage optimization of video surveillance from CCTV camera," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 710-713.
- [12] C. Held, J. Krumm, P. Markel, and R. P. Schenke, "Intelligent Video Surveillance," *Computer*, vol. 45, no. 3, pp. 83-84, 2012.

- [13] F. Lv, J. Kang, R. Nevatia, I. Cohen, and G. Medioni, "Automatic tracking and labeling of human activities in a video sequence," in *Proc. the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS04)*, 2004.
- [14] P. C. Ribeiro, J. Santos-Victor, and P. Lisboa, "Human activity recognition from video: Modeling feature selection and classification architecture," in *International Workshop on Human Activity Recognition and Modeling*, pp. 61-70, 2005.
- [15] L. Ma, H. Qi, S. Zhu, and S. Ma, "A fast background model based surveillance video coding in HEVC," in *IEEE Visual Communications and Image Processing Conference*, Valletta, pp. 237-240, 2014.
- [16] X. Zhang, T. Huang, Y. Tian, and W. Gao, "Background-modeling-based adaptive prediction for surveillance video coding," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 769-784, 2014.
- [17] S. Guo, Y. Wang, Y. Tian, P. Xing, and W. Gao, "Quality-progressive coding for high bit-rate background frames on surveillance videos," in *2015 IEEE International Symposium on Circuits and Systems*, 2015, pp. 2764-2767.
- [18] Y. Tian, J. Yan, S. Dong, and T. Huang, "PA-Search: predicting units adaptive motion search for surveillance video coding," *Computer Vision and Image Understanding*, vol. 170, pp. 14-27, 2018.
- [19] H. Kim, and H. Lee, "A low-power surveillance video coding system with early background subtraction and adaptive frame memory compression," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 4, pp. 359-367, 2017.
- [20] H. Li, W. Ding, Y. Shi, and W. Yin, "A double background based coding scheme for surveillance videos," in *Data Compression Conference*, Snowbird, UT, 2018, pp. 420-420.
- [21] J. Celine and S. A. A, "Face Recognition in CCTV Systems," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2019, pp. 111-116.
- [22] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," *IEEE Access*, vol. 9, pp. 34366-34382, 2021.
- [23] S. Powale, A. Dhanawade, S. Bagwe, S. Kawale, N. L. Chutke, and S. Chavan, "Person identification in low resolution CCTV footage using deep learning," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 236-240.
- [24] Computer Vision Group, School of Electronic Engineering and Computer Science, Queen Mary University of London. (Jan. 6, 2023). TinyFace: Face Recognition in Native Low-resolution Imagery. [Online]. Available: <https://qmul-tinyface.github.io/>
- [25] M. S. Pillai, G. Chaudhary, M. Khari, and R. G. Crespo, "Real-time image enhancement for an automatic automobile accident detection through CCTV using deep learning", *Soft Computing*, vol. 25, no. 18, pp. 11929-11940, 2021.
- [26] R. Ullah et al., "A real-time framework for human face detection and recognition in CCTV images," *Mathematical Problems in Engineering*, article 3276704, 2022.
- [27] Pan, S.-H., S.-C.J.S. Wang, and Materials, "Identifying vehicles dynamically on freeway CCTV images through the Yolo deep learning model," *Sensors and Materials*, vol. 33, no. 5, pp. 1517-1530, 2021.
- [28] M. Ghafari, A. Amirkhani, E. Rashno, S. Ghanbariet, "Novel gaussian mixture-based video coding for fixed background video streaming," in *2022 12th Iranian/Second International Conference on Machine Vision and Image Processing (MVIP)*, Ahvaz, Iran, 2022.
- [29] G. Matczak, and P. Mazurek, "Comparative Monte Carlo Analysis of Background Estimation Algorithms for Unmanned Aerial Vehicle Detection," *Remote Sensing*, vol. 13, no. 5, 2021.
- [30] Kaggle. (Jan. 6, 2023). *Highway Traffic Videos Dataset*. [Online]. Available: <https://www.kaggle.com/aryashah2k/highway-traffic-videos-dataset>

BIOGRAPHY



Mohammadreza Ghafari received his Telecommunication master's degree from Amirkabir University of Technology. He has graduated with his B.Sc. degree, Electrical Engineering and was honoured as the best student of the year at IRIB university. His background is mainly covering computer networks and video processing which he also participated in many national and academic projects in these fields. He has been working on Software Defined Network, Cloud Computing and Video Compression methods for many years, while his main interest is Cloud Gaming.



Abdollah Amirkhani received the M.Sc. and Ph.D. degrees (with honors) in electrical engineering from Iran University of Science and Technology (IUST), Tehran, in 2012 and 2017, respectively. He earned the Outstanding Student Award (2015) from the First Vice President of Iran. In 2016, he was awarded by the Ministry of Science, Research and Technology. He is an Assistant Professor in the school of automotive engineering at IUST. He is the Associate Editor of the "Engineering Science and Technology, an International Journal". He has been actively involved in several National R&D projects, related to the development of new methodologies and learning algorithms based on AI techniques. His research interests are in machine vision, fuzzy cognitive maps, data mining and machine learning.



Elyas Rashno received the B.Sc. degree in computer engineering from the Isfahan University of Technology, Isfahan, Iran, in 2015, the M.Sc. degree in artificial intelligence from the Iran University of Science and Technology, Tehran, Iran, in 2018. Since 2018, he has been an R&D member of a company that works on deep learning models. His research interests include machine learning, deep learning, image processing, biometrics, content-based image retrieval, and evolutionary computing.



files.

Shirin Ghanbari received her M.Sc. in E-Commerce and Ph.D. in object segmentation and video tracking from the Computer Science and Electrical Engineering department from the University of Essex, United Kingdom, in 2005 and 2010. Recently, she leads a team for data analysis using the latest AI technologies for both text and video

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Improving the Quality of ECG Signal Using Wavelet Transform and Adaptive Filters

Amir Hatamian^{1,*} , Farzad Farshidi² , Changiz Ghobadi¹ , Javad Nourinia¹ , and Ehsan Mostafapour¹

¹ Department of Electrical Engineering, Urmia University, Urmia, Iran

² Department of Biomedical Engineering, Islamic Azad University, Science and Research Branch, Tehran

* Corresponding Author: a.hatamian@urmia.ac.ir

Abstract: The increasing risk of cardiovascular diseases, stress, high blood pressure, obesity, sleep disorders, and depression causes electrocardiogram (ECG) monitors to be used for diagnosing health. The main objective of this research is to enhance the quality of the ECG signal using wavelet transform and adaptive filters. This research has been made as descriptive-analytic and the method is used in the signal processing stages to calculate the ECG modulation spectrum, the spectral-modulation filtering scheme, and the ECG database from the standard algorithm and performance criteria. The results of the simulation indicate that the conversion of Sym4 and the adaptive filter with the size of 0.0005 and the length of the filter of 25 signals to the noise will be greatly improved to reveal the main features of the ECG signal.

Keywords: ECG signal quality, wavelet, adaptive filter.

Article history

Received 27 May 2021; Revised 03 October 2022; Accepted 21 October 2022; Published online 28 February 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

A. Hatamian, F. Farshidi, C. Ghobadi, J. Nourinia, and E. Mostafapour "Improving the quality of ECG signal using wavelet transform and adaptive filters," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 45-53, 2023.

DOI: [10.22055/jaree.2022.37567.1030](https://doi.org/10.22055/jaree.2022.37567.1030)



1. INTRODUCTION

Different electronic health devices are designed for the elderly in the community and the growing population of these elderly citizens who live alone. This approach arises from the increased interest in personal health care and the growth of cardiovascular diseases. In monitor of a 24hours electrocardiogram (ECG), a data compression method is needed to efficiently use resources and to reduce the time of data transmission. In the telecommunication industry, a compression algorithm for compressing ECG signals has been provided based on selecting the important sub-bands of the wavelet packet transform with the aim of minimizing the data while the quality of the reconstruction signal is desirable. The proposed algorithm consists of four stages for compression and four stages for reconstruction [1]. Wavelet transform, wave compression techniques are suitable for displaying transients, such as sounds recorded in sound or high-frequency components in two-dimensional images (for example, an image of stars in the night sky). This means that the transient elements of a data signal can be represented by a smaller amount of information, unless in case of some other

changes including the discrete deformation is more used. The discrete wavelet transform has been successfully used to compress electrocardiographic signals (ECG). In this work, the high correlation between the wavelet coefficients of the signals of successive heart cycles is used by linear prediction [2]. Improvement methods of signal quality of wavelet transform are generally based on a threshold which is rooted in the assumption of energy concentration in a small number of discrete wavelet transform coefficients. The major disadvantage of threshold methods in the ECG signal is to create the distortion in the target signal. For this reason, in recent years, a variety of filters, and especially the adaptive filters, are used to correct wavelet transform coefficients. In the telecommunications industry, a new structure for improving the quality of the ECG signal with the help of using an adaptive and threshold filter on wavelet coefficients has been provided that in this method, instead of threshold, the wavelet coefficients are improved by passing through an adaptive filter. This causes the estimated signal to be largely close to the major signal and the estimate error to be reduced [3]. One of the problems of artisans is the processing of biomedical data (such as electrocardiography) that separates

the desirable signal from noises caused by interference of power lines, external electromagnetic fields, interference of high frequency and random and breathing movements. Different types of digital filters are used to remove signal components from undesired frequencies. It is difficult to use constant coefficient filters to reduce random noises. Because the behaviour of the system is not dependent on the specific time. To overcome this problem, an adaptive filter technique is required. Electrocardiogram (ECG) is the most important parameter for controlling cardiac activity. By fully analyzing the shape of the ECG signal, the doctor can detect various types of deviation. In some medical signal applications, useful signals are excelled by various components [4]. To avoid interference of ECG signals, designing a system is needed. To design and achieve a reliable system, there are two issues to consider. The accuracy and processing time should be appropriate. Basically, these methods should be categorized as adaptive and non-adaptive filter. There are some cavities in non-flexible filter or fixed filter. Adaptive filters are used for cancelling noise in different areas. Adaptive filters are capable of responding to their blow, and as a result, there is little information of the signal or information which can be planned for extracting the signal from unwanted information. This will help improve the signal-to-noise ratio. Adaptive filters are for those plans that some of the parameters of the processing operations required are not initially known or cannot be modified [5]. In order to overcome this limitation, algorithms of increasing ECG quality are strongly needed that can act under the broad spectrum of the levels of noise. Typically, two methods were previously investigated to enhance quality; filtering, wavelet contraction. For example, improving the quality of the ECG was performed using multiple repetitions of a moving filter that could have been about 10 dB for noisy signals with a signal-to-noise ratio [5]. A lot of studies have been done on cardiac signals that the Esmaili (2018) analyzed ECG signal in a research by using the features of ECG cardiac signal (morphological and characteristics derived from wavelet transform) and neural network [6] the results show that (25 normal files and 20 non-normal files) of these signals, 64 characteristics are obtained (48 characteristics based on wavelet transform and 16 morphological characteristics) that as inputs in neural network. The result indicated the effectiveness of the algorithm used. Rezai and Khodadadi (2016) investigated the use of an adaptive filter to remove ECG noise from the surface EMG signal [7], and the results show that the fast convergence of the least squared algorithm has made it possible for the proposed filter to effectively remove electrocardiogram noise remove from the surface electromyogram signal. Mohseni et al. (2015) investigated the new method based on the usage of a variety of comparative filters to remove artifacts from the ECG signal in a research [8], and the results of calculations of the signal-to-noise ratio for LMS, EBLMS, ENLMS and ELMS filters respectively 31.557, 3.516, 3.516, 6.830 and 3.038 were obtained. Belgorzi and Elshafiey (2017) investigated fixed-wavelet transforms and adaptive filter for improving the quality of the ECG signal in a research [9], and the results of fixed wavelet showed that an increase in the signal of ECG quality is successfully performed in terms of signal-to-noise ratio. In a study, Tobon and Falk (2016) examined modulation filtering to improve the quality of ECG in a study [10] and the findings obtained showed that the proposed

algorithm can be used to improve the quality of wearable ECG monitors even in severe conditions, so it can play a key role in training and monitoring the performance of peak sport. Sehamby and Singh (2016) investigated the elimination of noise by using an adaptive filter in the ECG signal in a study [11], and the ECG signals are weak and easily sensitive to noise and interference. In this research, the implementation of scale of the least squares was presented. Sharma et al. (2015) investigated and designed an adaptive filter for reducing the noise in the ECG in a research [12], and experimental results have shown that rate of convergence increases for small amounts of step. Increasing the risk of cardiovascular disease, stress, high blood pressure, obesity, sleep disorders and depression makes use of portable electrocardiogram monitors (ECG) for diagnosing health, but other parts of the market regarding medical programs are emerging. However, low-cost electromagnetic devices have shown that they are susceptible to numerous artifacts, including muscle contractions, base noises and movement, so the quality of the signal decreases and ultimately prevents heartbeat to change and to analyze it. To overcome this limitation, the algorithms of increasing ECG quality are highly needed which can operate under a broad spectrum of noise levels [13]-[21]. ECG is substantially based on the electrical conductivity of the heart. Normal conduction has started and distributes in a predictable pattern. Deviations from this pattern can be a natural or pathological change. An ECG is not equivalent to the activity of mechanical pump of the heart. For example, electrical activity produces an ECG which needs to pump the blood, but no pulse is felt. Ventricular fibrillation produces an electromagnetic wave, but it is too inefficient to produce the sustained cardiac output. Some rhythms have a good cardiac output, and some have bad heart output. Finally, an echocardiogram model or other anatomical imaging methods is useful in evaluating the mechanical function of the heart. Therefore, in this study, the increase of the quality of the ECG signal will be reviewed by using wavelet transform and adaptive filters.

2. MATERIALS AND METHOD

In the signal processing stages for calculating ECG modulation spectrum, modulation filtering scheme and ECG data base, the standard algorithm and performance criteria are used. The wavelet-based algorithms are the most popular, so here it is used to evaluate the proposed algorithm. A subset of 100 signals, each with 5 levels was used to optimize the parameters of the standard algorithm. It was found that the universal contraction method with soft threshold and a mother wavelet with 8 decomposition levels resulted in the best performance in this subset. Based on above information, the equations for the variables are set.

3. FORMULATION

In standard algorithm we have the followings.

$$\dot{V} = q_{in} - q_{out} \quad (1)$$

$$\dot{B} = \left(\mu - \frac{q_{in}}{V} \right) B \quad (2)$$

where V is volume, B is biomass, μ is vacuum permeation, and q is signal.

$$\dot{S} = q_{in}(c_{in} - S) - r_1 M_w B \quad (3)$$

where M_w is molar mass, r_1 is rate of absorbance, S is layer concentration, and c_{in} is time of the involved sample (per hour).

4. WAVELET TRANSFORM ALGORITHM

Wavelet transform is an efficient tool for signal processing and is used in many areas such as elimination of images noise, audio and video processing, pattern recognition, image encoding and compression. For this reason, it is important to provide solutions to increase the speed of implementation of the wavelet transform. One of the best solutions is the use of parallel processing. In telecommunications, a new architecture for implementation of a two-dimensional discrete wavelet transform has been provided to be used in image compression. The structure in this research is aimed at reducing the complexity of hardware and software, as well as optimizing the number of consumables and increasing the frequency of work. This implementation includes a processor unit for calculating discrete wavelet transform coefficients and a control unit for controlling data flow in the processor and generating memory address lines and an external memory unit for storing wavelet transform coefficients. The wavelet has some minor benefits over the Fourier transform in reducing the calculations when examining certain frequencies. However, they are rarely more sensitive and in fact, the conventional wavelet is mathematically the same. A short-term Fourier transform using a Gaussian window is called Morelet. The exception, when searching for signals of known form is non-sinusoidal (e.g. heartbeats); in this case, using convergent wavelets can analyze the standards. The function $\Psi \in L^2(\mathbb{R})$, wavelet transform is called. This transformation can also be expressed as Hilbert. Hilbert is a complete and comprehensive transformation. Hilbert's base is defined as the function (4):

$$\Psi_{jk}: j, k \in \mathbb{Z} \quad (4)$$

$$\Psi_{jk}(x): s^{\frac{j}{2}}\Psi(2^j x - k), \quad j, k \in \mathbb{Z} \quad (5)$$

If it is under the internal standard of $L^2(\mathbb{R})$, as bellow:

$$(f \cdot g) = \int_{-\infty}^{+\infty} f(x)g(x)dx \quad (6)$$

$$(\Psi_{jk}, \Psi_{lm}) = \int_{-\infty}^{+\infty} \Psi_{jk}(x)\overline{\Psi_{lm}(x)}dx = \delta_{j1} \delta_{km} \quad (7)$$

Where δ_{j1} : Kronecker delta.

$$f(x) = \sum_{j,k=-\infty}^{\infty} c_{jk} \Psi_{jk}(x) \quad (8)$$

With the convergence of the above set that seems to be convergence is in it, we reach the objective. Such a representation of "f" is known as a wavelet series. This means that it is a two-dimensional wavelet. The transform of the consistent wavelet of integral transformation is defined in (9):

$$[W_{\Psi}f](a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} \Psi\left(\frac{x-b}{a}\right) f(x) dx \quad (9)$$

Wavelet coefficient of c_{jk} is as (10):

$$c_{jk} = [W_{\Psi}f](2^{-j}, k2^{-j}) \quad (10)$$

The idea of wavelet transforms is that transformation should only allow to growth for the changes at expansion time, not to the shape. This case affected by selecting the appropriate basis functions that are possible for this task. It is expected that changes in the expansion of time correspond to corresponding frequency analysis of basis function. Based on the uncertainty principle of signal processing, we have:

$$\Delta t \Delta \omega \geq \frac{1}{2} \quad (11)$$

where ω is angular frequency, and t is time. In Figs. 1 and 2 we present the basis functions.

A: when the Δt is big:

1. Inappropriate time resolution.

2. Appropriate frequency resolution.

3. Low frequency.

B: when the Δt is small:

1. Appropriate time resolution.

2. Inappropriate frequency resolution.

3. High frequency.

In other words, the basis function of Ψ can be considered as the promissory respond of a system that performance of $x(t)$ is filtered in it. Provides a signal for converting time and frequency related information. Therefore, the wavelet transform contains information that is similar to the short-time Fourier transform but differs from the special features of the wavelet, which are shown at the time of correction in the basis function of analysis frequencies. The difference in time separation of the ascending frequency for Fourier transform and wavelet transform is shown in Fig. 3.

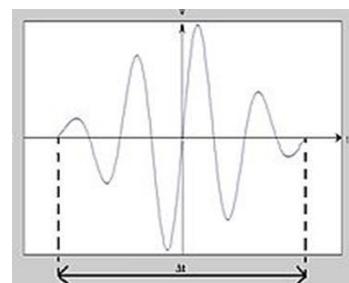


Fig. 1: $\Psi - \Delta t$ curve.

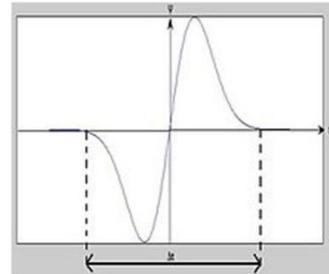


Fig. 2: $\Psi - \Delta t$ curve.

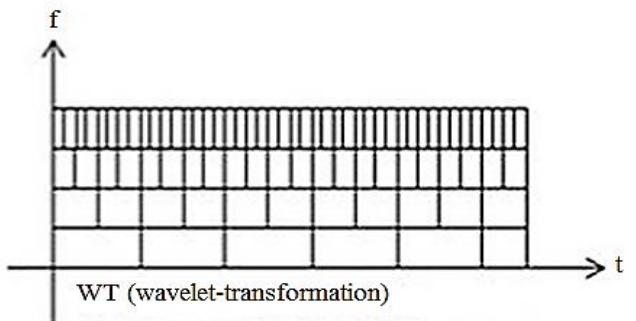


Fig. 3: Wavelet transform.

5. SIMULATION

5.1. Receiving Data

First, we receive the data from the following address, which contains two data sets of 100 and 200. Serial data 100 is randomly selected from 4000 pieces, serial data 200 includes rare, and important arrhythmias that are not well represented by random selection. Each of the data includes three files: Reference, Signals, Header, and Annotations. In Fig. 4, we present the set of ECG data that we used in this paper.

5.2. Using The Database of MIT-BIH in MATLAB

To use the data in the MATLAB software, we downloaded the 100m.mat matrix from the database. The sample frequency is considered 360 Hz, so any beat may range from 441 to 234 samples. The ECG signal is visible after the initial pre-processing and sampling in Fig. 5.

5.3. Noise Signal of ECG

Medical monitoring devices are highly sensitive to biomedical signal recording and require more accurate results for each proper diagnosis. The ECG signal of a healthy person repeats once every 0.8 seconds, which means a very low frequency. The low frequency signal is destroyed by the interference of the 50 Hz sound voltage line; this noise is also the source of interference with the biomedical signal recording. The 50 Hz power line interference frequency is approximately equal to the ECG frequency, so this 50 Hz noise can eliminate the ECG signal output. To simulate an ECG signal contaminated with this noise, we added a 50 Hz noise to the simulated ECG signal, and thus the ECG noise signal is visible in Fig. 6.

| Reference annotations | Signals | Header |
|-------------------------|-------------------------|-------------------------|
| 100.atr | 100.dat | 100.hea |
| 101.atr | 101.dat | 101.hea |
| 102.atr | 102.dat | 102.hea |
| 103.atr | 103.dat | 103.hea |
| 104.atr | 104.dat | 104.hea |
| 105.atr | 105.dat | 105.hea |
| 106.atr | 106.dat | 106.hea |
| 107.atr | 107.dat | 107.hea |
| 108.atr | 108.dat | 108.hea |
| 109.atr | 109.dat | 109.hea |
| 111.atr | 111.dat | 111.hea |
| 112.atr | 112.dat | 112.hea |
| 113.atr | 113.dat | 113.hea |
| 114.atr | 114.dat | 114.hea |
| 115.atr | 115.dat | 115.hea |
| 116.atr | 116.dat | 116.hea |

Fig. 4: Data formats in the MIT-BIH database.

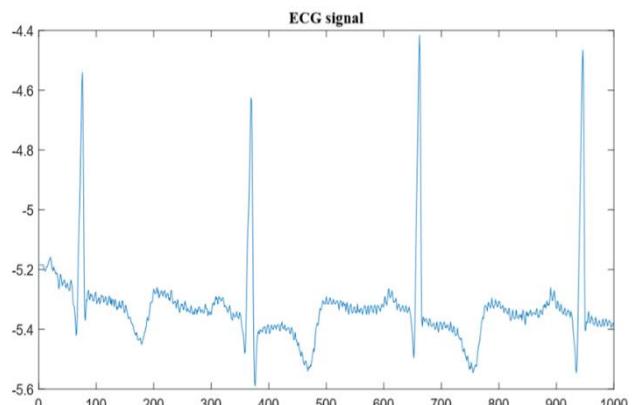


Fig. 5: ECG signal.

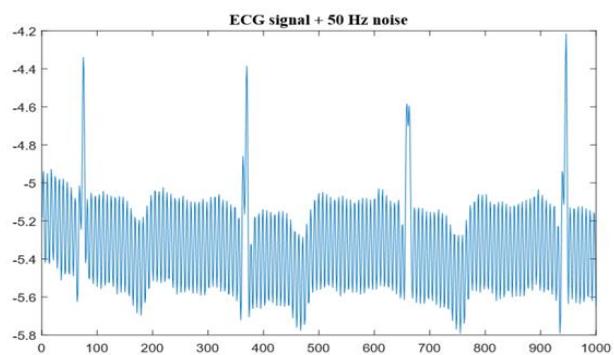


Fig. 6: ECG signal contaminated with 50 Hz noise.

5.4. Wavelet Transform

The common problem in recording heart signal is noise and even body and eye movement, which causes error in the recording of signal or its analysis. Noise or artifact can limit the use of ECG signal, and it is necessary to eliminate its effect. In this section, wavelet transform is used to eliminate or reduce of ECG signal noise. Wavelet transform is very practical as a method of time-frequency analysis. Wavelet transform to increase the signal to noise ratio is an effective and efficient method. In this study, wavelet operator has been used at several different levels of analysis to achieve optimal transform for improve the ECG signal noise and extract the feature from this signal. The applied wavelet function is considered on the EGC signal for this research as the following Fig. 7.

ECG signal after applying transform on the wavelet (mentioned above) are visible in the several different levels as the following figures:

Regarding the observed results, after the simulation shown in Figs. 8, 9 and 10, the transformation of the level 3 wavelet from the other surfaces eliminates the noise from the ECG signal by maintaining the feature of the signal well, and desirable wavelet is addressed. In this case, the signal to noise will be equal to 14.4642.

5.5. Design of Adaptive Filter

In this research, least average squares algorithm has been used which has been simulated by taking into account different values for μ and filter length, which are further discussed in more detail.

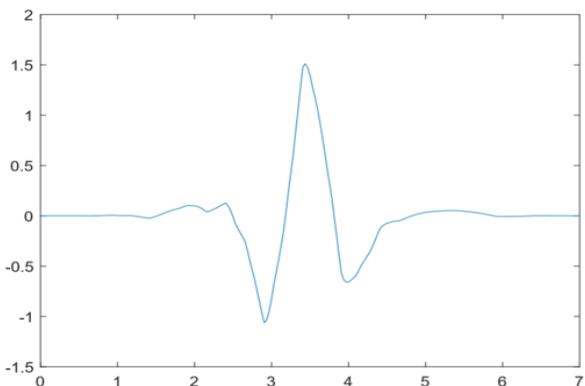


Fig. 7: The applied wavelet function on the ECG signal.

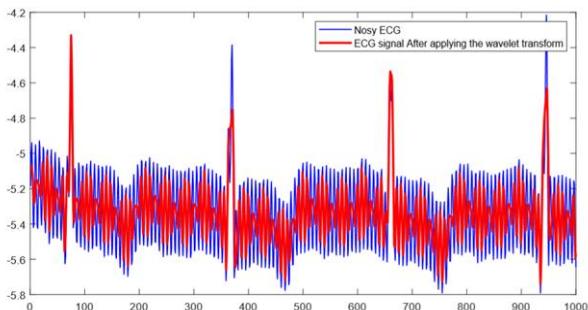


Fig. 8: ECG signal before and after applying transform on wavelet level 2.

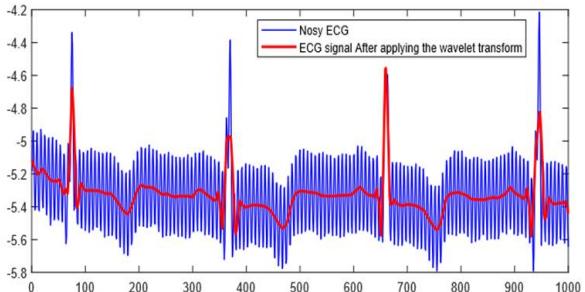


Fig. 9: ECG signal before and after applying transform on wavelet level 3.

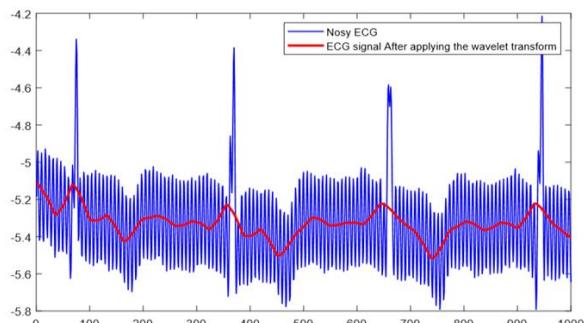


Fig. 10: ECG signal before and after applying transform on wavelet level 5.

- 1) Results for adaptive filter with $\mu=0.0005$ and filter length of 15 are presented in [Fig. 11](#).
- 2) Results for adaptive filter with $\mu=0.0005$ and filter length of 20 are presented in [Fig. 12](#).

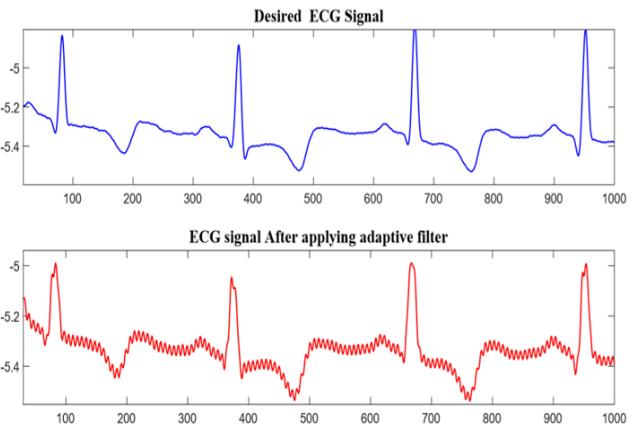


Fig. 11: desirable ECG signal and ECG signal obtained from applying adaptive filter with $\mu=0.0005$ and filter length of 15.

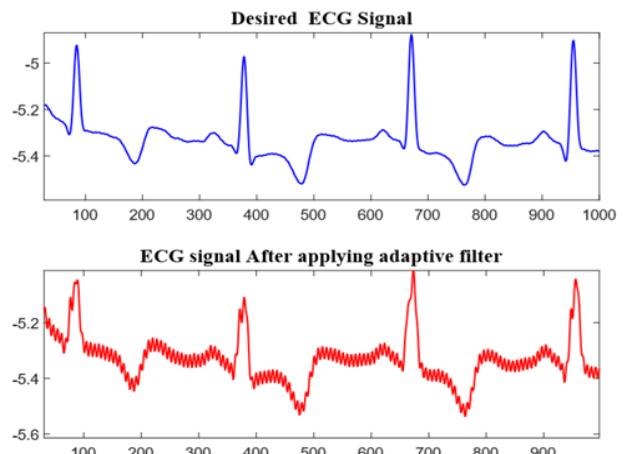


Fig. 12: desirable ECG signal and ECG signal obtained from applying adaptive filter with $\mu=0.0005$ and filter length of 20.

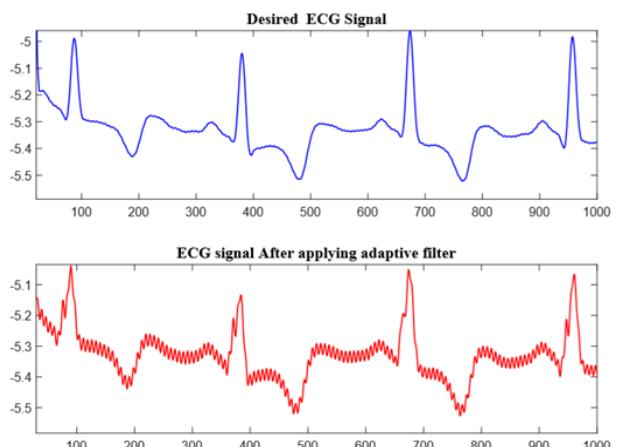


Fig. 13: desirable ECG signal and ECG signal obtained from applying adaptive filter with $\mu=0.0005$ and filter length of 25.

- 3) Result of adaptive filter with $\mu=0.0005$ and filter length of 25 are presented in [Fig. 13](#).
- 4) Result of adaptive filter with $\mu=0.0009$ and filter length of 15 are presented in [Fig. 14](#).
- 5) Result of adaptive filter with $\mu=0.0009$ and filter length of 20 are presented in [Fig. 15](#).

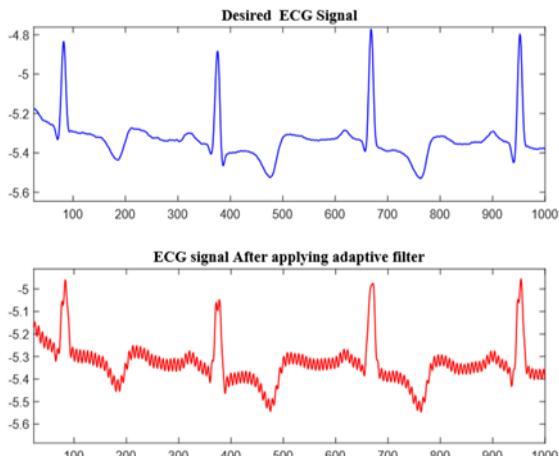


Fig. 14: desirable ECG signal and ECG signal obtained from applying adaptive filter with $\mu=0.0009$ and filter length of 15.

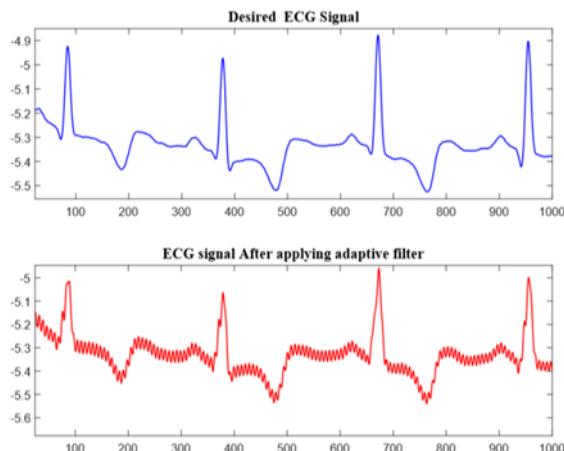


Fig. 15: desirable ECG signal and ECG signal obtained from applying adaptive filter with $\mu=0.0009$ and filter length of 20.

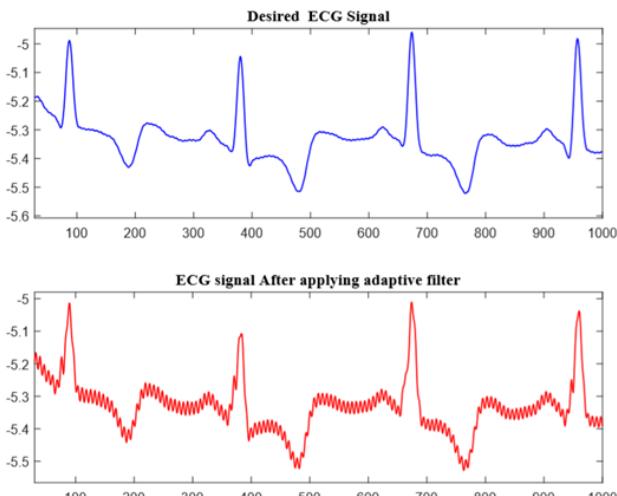


Fig. 16: desirable ECG signal and ECG signal obtained from applying adaptive filter with $\mu=0.0009$ and filter length of 25.

6) Result of adaptive filter with $\mu=0.0009$ and filter length of 25 are presented in Fig. 16.

Table 1: Adaptive filter modes.

| Adaptive filter (LMS) | Filtered signal with adaptive filter |
|--------------------------------------|--------------------------------------|
| $\mu=0.0005$ and filter length of 15 | 1.2111 |
| $\mu=0.0005$ and filter length of 20 | 3.5586 |
| $\mu=0.0005$ and filter length of 25 | 4.5103 |
| $\mu=0.0005$ and filter length of 15 | 0.377 |
| $\mu=0.0005$ and filter length of 20 | 3.1193 |
| $\mu=0.0005$ and filter length of 25 | 4.1828 |

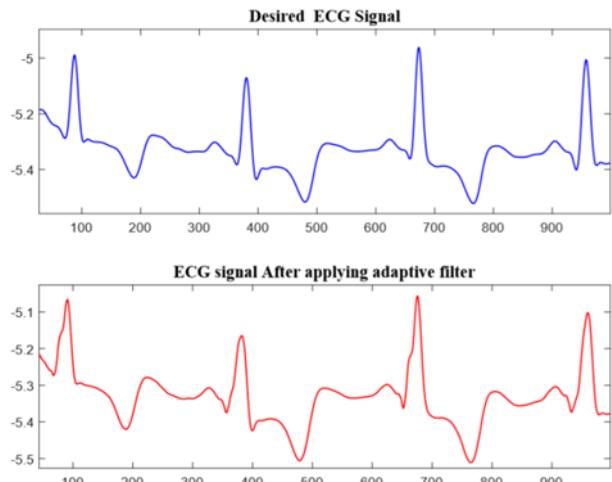


Fig. 17: desirable ECG signal and ECG signal obtained from applying transform on wavelet and adaptive filter with $\mu=0.0005$ and filter length of 25.

5.6. Signal-to-Noise Values for Different Adaptive Filter Modes Materials and Methods

Table 1 presents the various LMS filter modes that we used in this paper.

The results indicate that by increasing the filter length in the adaptive filter, signal-to-noise ratio will be higher. The simulation results after the transformation of wavelet and adaptive filter of $\mu=0.0005$ size of the window 25 will be as follows (see Fig. 17).

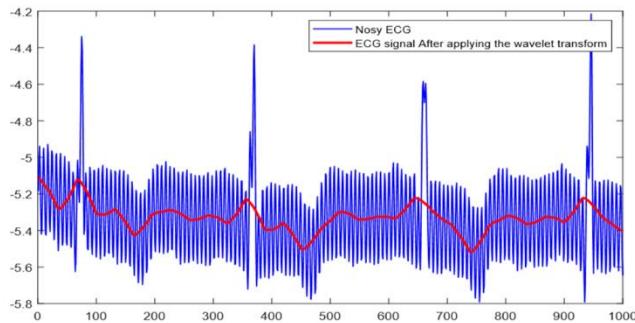
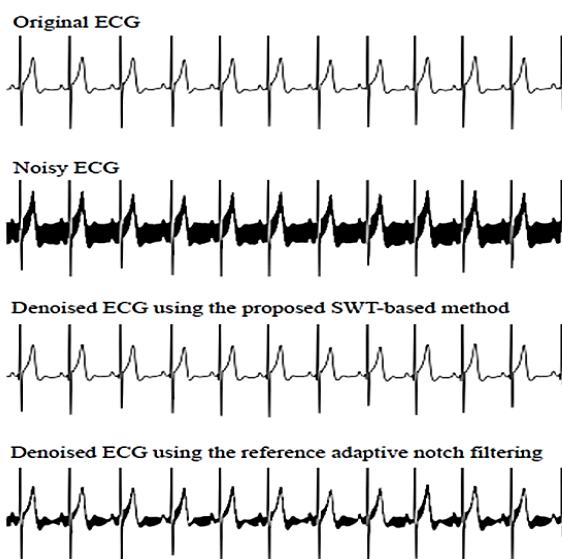
5.7. Comparing Results

The results of the simulations indicate that the transformation of the sym4 level 3 wavelet eliminates the noise significantly from the ECG signal and increases the signal to noise. Also, by checking the LMS adaptive filter in different modes with step size (μ) and different filter lengths determined with trial and error that for the size of the step 0.0005 and the filter length of the 25, the signal to noise and signal amplitude would be desirable. In Table 2 we present an explanatory comparison between our findings and the findings of references [22] and [23].

As it is observed in Figs. 18 and 19, the result obtained from stimulation by using mentioned method is more efficient compared to [22] and reduced the noise.

Table 2: Comparison table.

| Reference no. | Article title | Release year | Result |
|---------------|---|--------------|---|
| [22] | An efficient wavelet transform-based algorithm for reducing power noise from electrocardiogram | 2018 | It introduces easy and efficient algorithm for suppressing the PLI from the ECG. In summary, the input signal is divided into four wavelet levels, and the resulting coefficients are used to eliminate the estimated PLI from the TQ intervals of the threshold. The ECG denoised signal is then reconstructed by calculating the inverse wavelet transform. |
| [23] | ECG signal filtering based on CEEMDAN with a minimum time interval range and more accurate statistics for selecting related modes | 2018 | Enhancement of ECG signal based on collective experimental state of branching with compatibility noise (CEEMDAN) and Statistics Order Order (HOS) |

**Fig. 18:** ECG signal before and after applying the wavelet transform by simulation software.**Fig. 19:** ECG signal before and after applying noise elimination in [22].

6. CONCLUSION

The results of the simulations indicate that the transformation of the sym4, level 3 transforms well the noise from the ECG signal and increases the signal to noise. In addition, by checking the LMS adaptive filter in different modes with step size (μ) and different filter lengths was determined by trial and error that step size of 0.0005 and filter length of 25, signal to noise, and signal amplitude will be desirable. In this study, we investigated the performance of

digital filters on the ECG signal, and in particular, we applied the adaptive filter and the wavelet transform on an ECG signal that received from the MIT-BIH database. The results of the simulation show that the Sym4 wavelet transform and adaptive filter with step size of 0.0005 and filter length of 25, well improves the signal to noise to the high level and will be able to detect the main features of the ECG signal. In this study, in particular, the wavelet transform and adaptive filter were used to eliminate the noise and extraction of signal features. In this regard, in order to continue the pathway of the researcher, we study more about digital filters such as Kalman filter on the ECG signal.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Amir Hatamian: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Project administration. **Farzad Farshidi:** Formal Analysis, Project administration, Software, Validation. **Changiz Ghobadi:** Project administration, Supervision, Writing - review & editing. **Javad Nourinia:** Project administration, Supervision, Writing - review & editing. **Ehsan Mostafapour:** Methodology, Writing - original draft, Writing - review & editing.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] F. Faraji-Kheirabadi, "An effective method for compression of ECG signals," M.Sc. Thesis, Razi University, Faculty of Technology and Engineering, 2015 (in Persian).
- [2] J. Liu, "Shannon wavelet spectrum analysis on truncated vibration signals for machine incipient fault detection," *Measurement Science and Technology*, vol. 23, no. 5, article 055604, 2012.
- [3] E. Esmaili, and A. Rafiee, "Simultaneous use of the adaptive filter and wavelet transform to improve the

- quality of the ECG signal," in *The 5th International Conference on Electrical and Computer Engineering with an emphasis on indigenous knowledge*, Tehran, 2017 (in Persian).
- [4] S. Hussian, and S. M. Babitha, "Noise removal from cardiac signals using various adaptive algorithms," in *Proceedings of International Academic Conference on Electrical, Electronics and Computer Engineering*, 2013.
- [5] S. K. Salih, S. A. Aljunid, S. M. Aljunid, and O. Maskon, "Adaptive filtering approach for denoising electrocardiogram signal using moving average filter," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 5, pp. 1065-1069, 2015.
- [6] A. Esmaili, "Arrhythmia modeling of ECG signals by using artificial neural networks," in *the 2nd National Conference on advanced developments in the field of energy and oil and gas industries*, Saveh, 2018 (in Persian).
- [7] A. Rezaei, and S. Khodadadi, "Use of adaptive filter to eliminate electrocardiogram noise from superficial electromyogram," in *the 2nd International Conference on new findings of science and technology*, 2016 (in Persian).
- [8] E. Mohseni, F. Goldoost, and N. Safdarian, "New method based on the use of a variety of adaptive filters for the elimination of artifacts from the ECG signal," in *the international conference of nonlinear systems and the optimization of electrical and computer engineering*, Iran, 2015 (in Persian).
- [9] A. N. S. Belgurzi, and I. Elshafiey, "A power line interference canceler using wavelet transform and adaptive filter for ECG signal," in *2017 International Conference on Computer and Applications (ICCA)*, IEEE, 2017, pp. 206-210.
- [10] D. P. Tobón, and T. H. Falk, "Adaptive modulation spectral filtering for improved electrocardiogram quality enhancement," in *Computing in Cardiology Conference (CinC)*, IEEE, 2016, pp. 441-444.
- [11] R. Sehamby, and B. Singh, "Noise Cancellation using Adaptive Filtering in ECG Signals: Application to Biotelemetry," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 2, pp. 237-244, 2016.
- [12] I. Sharma, R. Mehra, and M. Singh, "Adaptive filter design for ECG noise reduction using LMS algorithm," in *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, IEEE, 2015, pp. 1-6.
- [13] D. Zhang et al., "An ECG signal de-noising approach based on wavelet energy and sub-band smoothing filter," *Applied Sciences*, vol. 9, no. 22, 4968, 2019.
- [14] B. Yang, C. Yu, and Y. Dong, "Capacitively coupled electrocardiogram measuring system and noise reduction by singular spectrum analysis," *IEEE Sensors Journal*, vol. 16, no. 10, pp. 3802-3810, 2016.
- [15] P. S. Gokhale, "ECG signal de-noising using discrete wavelet transform for removal of 50Hz PLI noise," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 5, pp. 81-85, 2012.
- [16] L. El Bouny, M. Khalil, and A. Adib, "ECG signal denoising based on ensemble emd thresholding and higher order statistics," in *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSiP)*, IEEE, 2017, pp. 1-6.
- [17] H. T. Patil, and R. S. Holambe, "New approach of threshold estimation for denoising ECG signal using wavelet transform," in *2013 Annual IEEE India Conference (INDICON)*, IEEE, 2013, pp. 1-4.
- [18] W. Jenkal et al., "An efficient algorithm of ECG signal denoising using the adaptive dual threshold filter and the discrete wavelet transform," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 3, pp. 499-508, 2016.
- [19] S. Cuomo, G. De Pietro, R. Farina, A. Galletti, and G. Sannino, "A revised scheme for real time ecg signal denoising based on recursive filtering," *Biomedical Signal Processing and Control*, vol. 27, pp. 134-144, 2016.
- [20] O. Singh, and R. K. Sunkaria, "ECG signal denoising based on empirical mode decomposition and moving average filter," in *2013 IEEE International Conference on Signal Processing, Computing and Control (ISPCC)*, IEEE, 2013, pp. 1-6.
- [21] M. A. Kabir, and C. Shahnaz, "Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains," *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 481-489, 2012.
- [22] J. Ródenas, M. García, J. J. Rieta, and R. Alcaraz, "An Efficient Algorithm Based on Wavelet Transform to Reduce Powerline Noise From Electrocardiograms," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45, IEEE, 2018.
- [23] L. El Bouny, M. Khalil, and A. Adib, "ECG signal filtering based on CEEMDAN with hybrid interval thresholding and higher order statistics to select relevant modes," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13067-13089, 2019.

BIOGRAPHY



Amir Hatamian was born in Urmia, Iran in 1995. He received his B.Sc. M.Sc. degrees from Urmia University, in 2017 and 2019, respectively, both in telecommunication engineering. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Urmia University. His research interests include medical signal processing, pattern recognition, and antennas.



Farzad Farshidi was born in Urmia, Iran in 1990. He received his B.Sc. M.Sc. degrees from Tabriz Azad University, in 2013 and 2015, respectively, both in bioelectric medical engineering. Since 2015 he is pursuing the Ph.D. degree in the Azad University, Science and Research Branch, Tehran. His research interests include medical Image and signal processing, pattern recognition, and classifying networks.



Changiz Ghobadi received his B.Sc. in Electrical Engineering-Electronics and M.Sc. degrees in Electrical Engineering from Isfahan University of Technology, Isfahan, Iran and Ph.D. degree in Electrical-Telecommunication from University of Bath, Bath, UK in 1998. From 1998 he was an assistant professor and now he is a professor in the Department of Electrical Engineering of Urmia University, Urmia, Iran. His primary research interests are in antenna design, radar and adaptive filters.



Javad Nourinia received his B.Sc. in Electrical and Electronic Engineering from Shiraz University and M.Sc. degree in Electrical and Telecommunication Engineering from Iran University of Science and Technology, and Ph.D. degree in Electrical and Telecommunication from University of Science and Technology, Tehran Iran in 2000. From 2000 he was an assistant professor and now he is a professor in the Department of Electrical Engineering of Urmia University, Urmia, Iran. His primary research interests are in antenna design, numerical methods in electromagnetic, microwave circuits.



Ehsan Mostafapour was born in west Azarbayjan Province, Urmia, Iran, in 1988. He received his B.S. and the M.S. in 2010 and 2012, respectively, both in telecommunication engineering. He received his Ph.D. degree from the Department of Electrical Engineering, Urmia University in 2018. His research interests include stochastic and adaptive signal processing. Dr. Mostafapour has published more than 30 scientific papers in multiple journals and is a fulltime reviewer for the respected journals of Wireless personal communication, IEEE TVT, IEEE Access, ACES, Journal of communication engineering (JCE), etc.

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Effect of Changes in the Parameters of a Modular Converter in Its Controllability Range in Fuel Cell Applications

Mohammad Afkar¹ , Parham Karimi^{1,*} , Roghayeh Gavagsaz-Ghoachani¹ , Matheepot Phattanasak² , and Serge Pierfederici³

¹ Mechanical and Energy Engineering, Department of Renewable Energy, Shahid Beheshti University, Tehran, Iran

² Department of Teacher Training in Electrical Engineering, King Mongkut's University of Technology North Bangkok, Thailand

³ LEMTA Université de Lorraine, CNRS Nancy, France

* Corresponding Author: parha.karimi@gmail.com

Abstract: In fuel cell systems, voltage balancing is an important consideration. The utilization of a modular construction based on a three-level boost converter was able to balance DC voltage. This paper investigates the effect of parameter variations, such as inductors and capacitors, on the converter's steady-state controllable areas. The plot of the inductor current and the voltages of the output capacitors are illustrated for different scenarios. The system simulation results were performed using MATLAB / Simulink software.

Keywords: Fuel cell modular converter, voltage balance, parameter changes, controllability.

Article history

Received 10 June 2022; Revised 14 October 2022; Accepted 06 December 2022; Published online 28 February 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

M. Afkar, P. Karimi, R. Gavagsaz-Ghoachani, M. Phattanasak, and S. Pierfederici, "Effect of changes in the parameters of a modular converter in its controllability range in fuel cell applications," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 54-61, 2023. DOI: [10.22055/jaree.2022.41056.1061](https://doi.org/10.22055/jaree.2022.41056.1061)



1. INTRODUCTION

Nowadays, the vital role of renewable energy in human life cannot be ignored. The fuel cell (FC) is a worthwhile energy-harvesting technology. It has attracted much attention in microgrid and electric hybrid vehicle applications [1, 2]. Much progress has been made in FCs, which has caused the formation of different types of FCs. Even though FCs have a great variety, they have the same operating principles and a high power density.

The polymer fuel cell (PEMFC) is popular among fuel cells [3]. Thanks to the solid electrolyte, PEMFC is shown high resistance to gas. PEMFC takes advantage of the reaction of hydrogen and oxygen to generate DC electric power. PEMFC can be set up quickly by taking advantage of low operating temperatures. These advantages cause this kind of FC in applications like vehicles and emergency systems which need high speed to be practical [3]. The advantage and disadvantages of PEMFC are enumerated in Table 1 [4-5]. Expanding the life of this type of FC is the major challenge

of this technology. Although the oxygen and hydrogen inputs are connected to the FC stack in parallel, the electrical outlets are linked in series. The series connection is to boost the output voltage. Because of the series connection of cells, the whole system's lifespan depends on each cell's lifespan [6].

The phenomenon of the snowball effect is one of the significant challenges facing the FC. Fig. 1 shows the snowball effect in an FC. Chain reactions in this effect can lead to the destruction of the FC. One of the effective parameters in cell destruction is membrane drying.

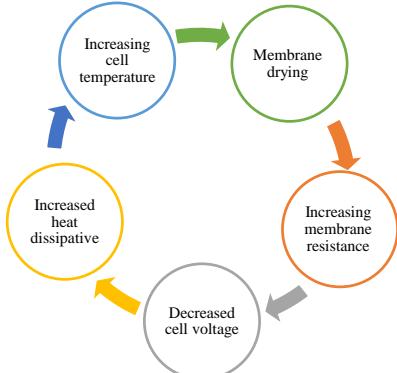
Proper energy management can increase the life of any cell. One solution to prevent this is to regulate water as a product produced in the cell by regulating the FC current. A DC-DC converter can regulate current [7]. A particular unit assures water management in the fuel cell, but flow management is done with the assistance of a DC-DC converter.

DC-DC power converters are among the most important

Table 1: Advantages and disadvantages of PEMFC [4-5].

| Advantages | | | |
|-----------------------------|---------------------------|--------------------|--------------------|
| Good efficiency | Long life | Flexibility in use | Solid electrolyte |
| Non-corrosive electrolyte | Low operating temperature | Fast start-up | High power density |
| Consumable water production | Simple design | Easy production | Reasonable price |

| Disadvantages | | | |
|--------------------------------|------------------------|-----------------------|--------------------------------|
| Sensitivity to carbon monoxide | High cost of catalysts | Limited manufacturers | Complexity of water management |

**Fig. 1:** Snowball effect in fuel cell.

and thought-provoking issues in hybrid systems [8]. These converters have various applications in different industries. The applications of these converters are listed in Fig. 2 [9-17].

A group of FC cells is connected in series to make an FC stack. Connecting each stack to its converter and the layout of output converters in series to DC-DC converters can effectively boost voltage [7, 18]. This structure is illustrated in Fig. 3. The possibility of controlling the output current of each cell individually in this method gives the freedom to manage and control the FC. This method can be used to solve the snowball effect.

Each FC stack operates independently according to its specific conditions. Therefore there is another new challenge. This problem occurs due to the voltage imbalance in output capacitors C1 to C4 (Fig. 3) caused by unequal stack power production. Risen stress on switches and power components and cell life reduction can be caused by voltage imbalance.

According to [18], a modular structure is suggested to solve the voltage imbalance problem in the photovoltaic system. This structure solves the problem of voltage imbalance by sharing a capacitor between two converters. Fig. 4 shows the considered structure consisting of two modules. Three capacitors are used in this circuit. The voltages of C1 and C2 and as well as C2 and C3, are equal. As a result, capacitor voltages C1 and C3 are equal. The equalization of the output voltages means that the voltage is balanced.

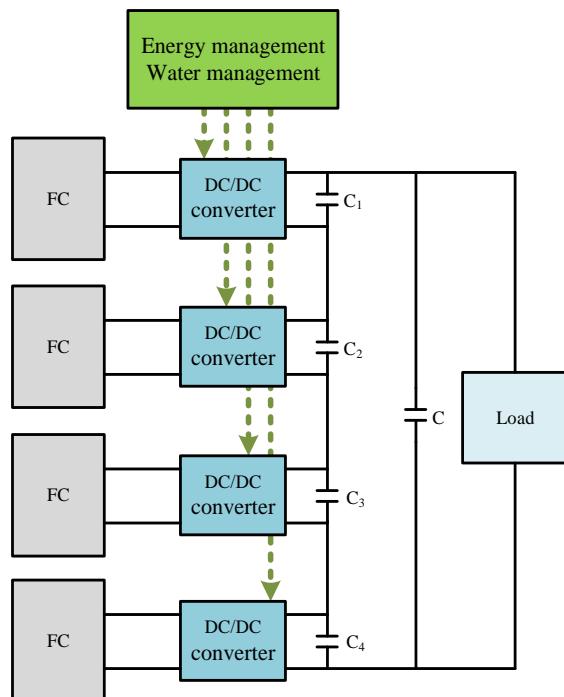
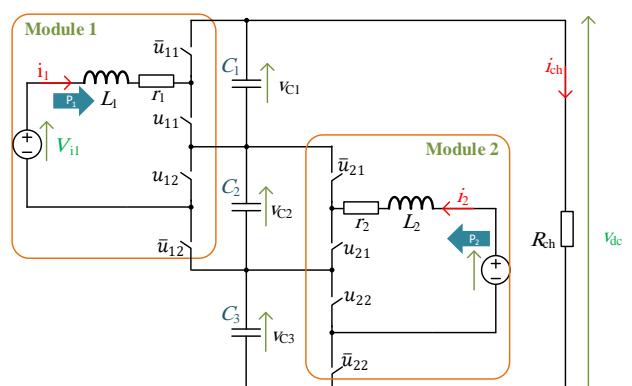
**Fig. 2:** DC-DC power converter applications [7-15].**Fig. 3:** Utilizing separate converters in the fuel cell.**Fig. 4:** Studied system: DC modular system.

Fig. 5 depicts some of the advantages of the researched structure in terms of modularity and the use of a three-level boost. The proposed system's shortcomings include a large

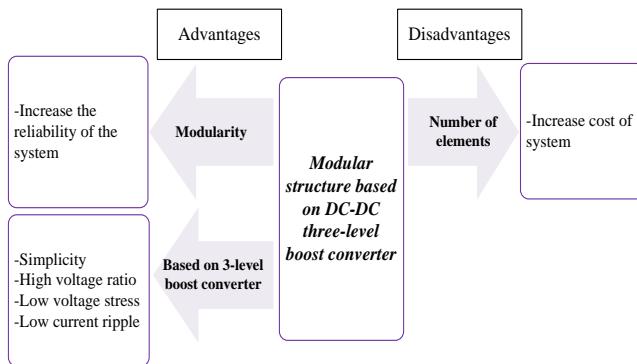


Fig. 5: Advantages and disadvantages of the studied structure.

number of switches, which raises the cost. However, despite these disadvantages, the use of this structure is justified.

The commandable areas of the considered converter in [7] are investigated. If the calculated duty cycles are between 0 and 1, the system is in the commandable areas.

The inductor and capacitor values can be planned by restricting the current ripple. Besides, they can be designed by choosing the high-frequency voltage ripple at a specific switching frequency. The inductor, capacitor, and frequency are determined using an optimization procedure that considers volume, cost, and efficiency limitations [19-20].

Because the parameters of each system will undergo possible changes in its operating point, in this paper, the change of capacitor and inductor values on the voltage balance of the system is investigated. In research [18], the capacitor and inductor values for both modules are considered equally. Moreover, in [18], the robustness of this modular converter and its sliding mode controller is investigated by changing the value of the capacitor and inductor change from 50% to 150%. Although, the effect of this change in these elements on the waveforms of inductor current or capacitor voltage and especially voltage balance is not considered. In this study, additional work was performed to further investigate and ensure that the performance of the system that we introduced in [18] does not affect by changes in capacitance and inductance, especially tracking reference currents and voltage balance.

According to Fig. 6, fuel cell systems deal with a snowball effect challenge that can damage fuel cells. To relieve this problem, separate converters might be utilized. Furthermore, boost converters are widely utilized in fuel cell systems since the output voltage is low. These converters can be coupled in series to enhance the produced voltage further. A voltage imbalance occurs in the system when a separate converter is used. Each FC may operate under different conditions and, therefore, has a different voltage at its output terminal.

The current and voltage waveform of inductors and capacitors are displayed in this paper to explore the influence of inductors and capacitor changes.

The paper structure is as follows: After reviewing the research literature in the introduction, the second part introduces the studied system. The governing equations of

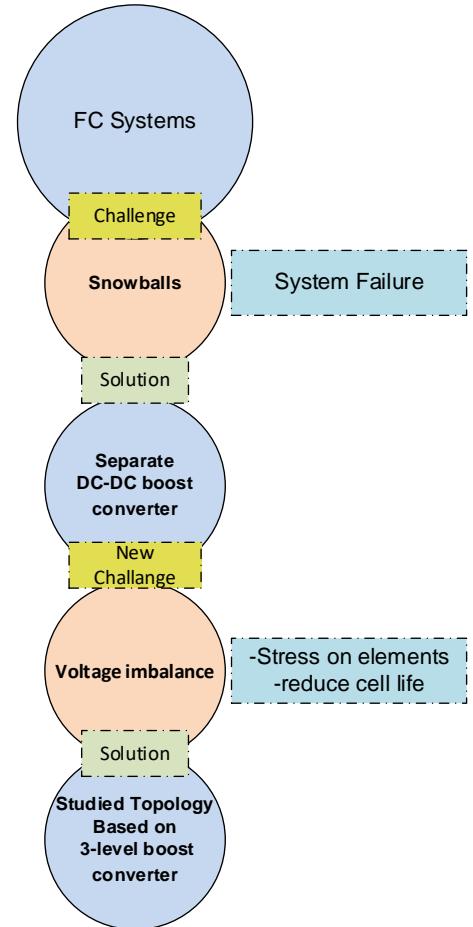


Fig. 6: Application of the studied structure in FC systems.

that system are recalled. In the third section, the results of the system simulation are shown. At the end is the conclusion.

2. STUDIED SYSTEM

In this paper, the three-level modular converter is considered. Fig. 4 shows the studied system, which contains two modules.

An input voltage source, an inductor, and its switches are important elements that build a module.

Four switches exist in each module; two are main switches, and the others are complementary switches. The two main switches take advantage of the interleaving technique for making the command signals. In modulation, the signal phase of the first control signal shifts half the period to the second control signal.

The system incorporates five state equations shown in Table 2. The state variables are input inductors' currents and the other is capacitors' voltages.

3. SIMULATION OF THE SYSTEM AND ITS RESULTS

System simulation has been done with the use of MATLAB/Simulink software. A large-signal average model for the two-module system is used for simulation [18]. The simulation parameters are given in Table 3.

Table 2: System state equation.

| | |
|-----------|---|
| | $L_1 \frac{di_1}{dt} = -r_1 i_1 + v_{i1} - (1 - d_{11})v_{c1} - (1 - d_{12})v_{c2}$ |
| Inductor | $L_2 \frac{di_2}{dt} = -r_2 i_2 + v_{i2} - (1 - d_{21})v_{c2} - (1 - d_{22})v_{c3}$ |
| | $C_1 \frac{dv_{c1}}{dt} = (1 - d_{11})i_1 - i_{ch}$ |
| Capacitor | $C_3 \frac{dv_{c3}}{dt} = (1 - d_{22})i_2 - i_{ch}$ |
| | $C_2 \frac{dv_{c2}}{dt} = (1 - d_{12})i_1 + (1 - d_{21})i_2 - i_{ch}$ |

Table 3: System parameters.

| Parameter | Value |
|------------------|--------------|
| V_{i1}, V_{i2} | 12 V |
| L_1, L_2 | 0.9 mH |
| r_1, r_2 | 0.06, 0.01 Ω |
| C_1, C_2, C_3 | 100 μF |
| f_s | 10 kHz |
| R_{ch} | 8.52 Ω |

Several scenarios are performed to change the values of inductors and capacitors. **Table 4** is considered several cases. The simulation results for each case are presented in **Figs. 10–14**. In each figure, two operating points are considered for input power reference values (P_{ref}). For all figures, the power reference $P_{ref} = 120$ W for the upper figure (a) and $P_{ref} = 500$ W for the lower figure (b). As shown in **Table 4**, there is a normal case in which the parameters are at their nominal values. Only the capacitor or inductor is changed in other cases (Case 1, Case 2, and Case 3).

3.1. Normal Case

In this scenario, the current waveforms of the inductor and the voltage of the capacitor in a steady state are conducted without changing the parameter.

Fig. 7 shows the waveforms of the currents i_{L1} and i_{L2} and the corresponding reference currents, i_{ref1} and i_{ref2} . It can be observed that in both operating points (120 W and 500 W), inductor currents are tracking their reference current. Moreover, in those operating points, voltage balance is achieved, as V_{c1} , V_{c2} and V_{c3} are equal.

The voltage waveforms of capacitors C_1 , C_2 , and C_3 as a function of time are seen in **Fig. 8**.

3.2. First Case (Case1)

In this case, the common capacitor is changed. **Figs. 9 and 10** display the voltage waveform of capacitors C_1 ,

Table 4: Different scenarios.

| Scenario | Changes | | Value | Figure Number |
|----------|----------------|----------|---|--|
| Normal | Without change | | $C_1=100 \mu F$ $C_3=100 \mu F$ $C_2=100 \mu F$ | Fig. 7 Fig. 8 |
| Case 1 | Change C_2 | Increase | $C_1=100 \mu F$ $C_3=100 \mu F$ $C_2=200 \mu F$ | Fig. 9 |
| | | Decrease | $C_1=100 \mu F$ $C_3=100 \mu F$ $C_2=50 \mu F$ | Fig. 10 |
| Case 2 | Change C_1 | Increase | $C_3=100 \mu F$ $C_2=100 \mu F$ $C_1=200 \mu F$ | Fig. 11 |
| | | Decrease | $C_3=100 \mu F$ $C_2=100 \mu F$ $C_1=50 \mu F$ | Fig. 12 |
| Case 3 | Change L_2 | Increase | $L_1=0.9 mH$ $L_2=2L_1$ | Fig. 13 |
| | | Decrease | $L_1=0.9 mH$ $L_2=0.5 L_1$ | Fig. 14 |

C_2 , and C_3 as a function of time. In these two figures, the value of the common capacitor C_2 changes. In **Fig. 9**, by increasing the value of C_2 , the ripple of its voltage decreases. In **Fig. 10**, voltage waveform increases rather than other capacitors. Nonetheless, the voltage balance is still achieved in increasing and decreasing values of capacitor and in both operating points (120 W and 500 W).

3.3. Second Case (Case 2)

In this case, the non-common capacitor is changed. **Figs. 11 and 12** present the voltage waveform of capacitors C_1 , C_2 , and C_3 as a function of time. In these two figures, the value of capacitor C_1 changes. These figures represent that similar to Case 1, by changing the value of this capacitor, voltage balance is achieved, and changing this capacitor only changes the ripple of its voltage waveform.

3.4. Third Case (Case 3)

In this case, one of the inductors is changed. **Figs. 13 and 14** illustrate the waveforms of the currents i_{L1} and i_{L2} and the corresponding reference currents, i_{ref1} and i_{ref2} . In these two figures, the value of inductor L_2 changes. In **Fig. 13**, the ripple of the current waveform is decreased due to the increase in the value of L_2 . In **Fig. 14**, L_2 is decreased, and its current waveform has larger ripples than L_1 current waveform. In both of them, tracking reference currents are done properly.

According to all the waveforms obtained from the simulation, it can be seen that changes in the value of the inductor or capacitor only change the current or voltage ripple. Changes in the parameters do not affect the voltage balance of the capacitors.

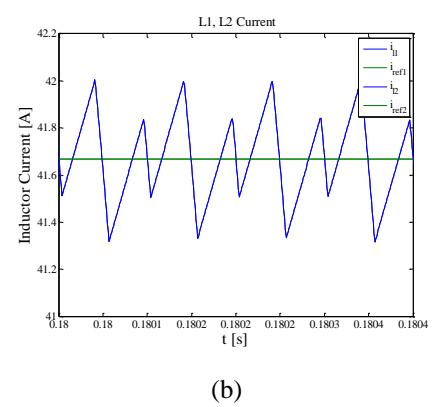
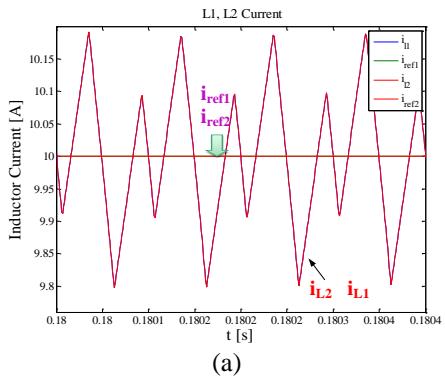


Fig. 7: Simulation waveforms: Inductor currents for L_1 and L_2 and related reference currents (normal state). Input power: (a) 120 W, (b) 500 W.

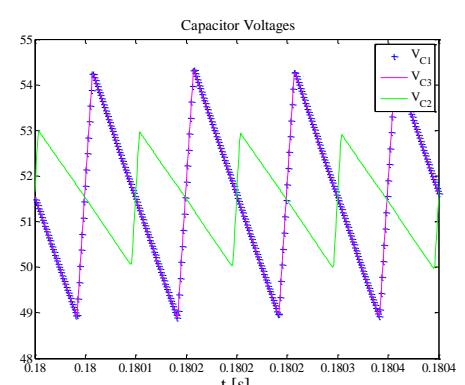
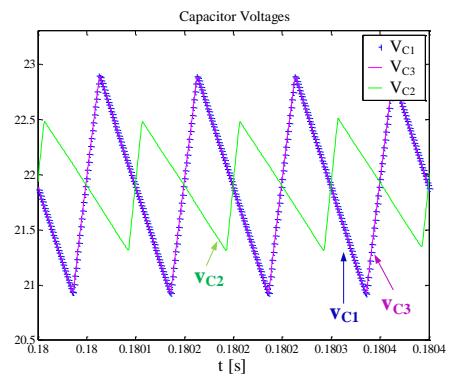


Fig. 9: Simulation waveforms: Voltage of capacitors C_1 , C_2 , C_3 first state ($C_2 = 200 \mu F$). Input power: (a) 120 W, (b) 500 W.

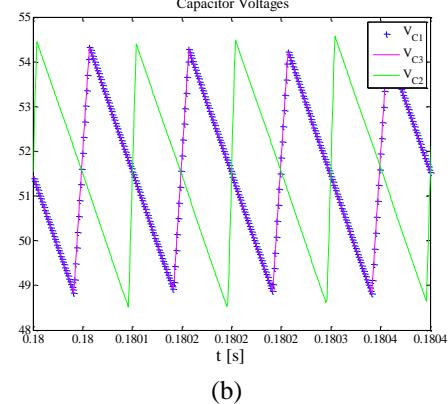
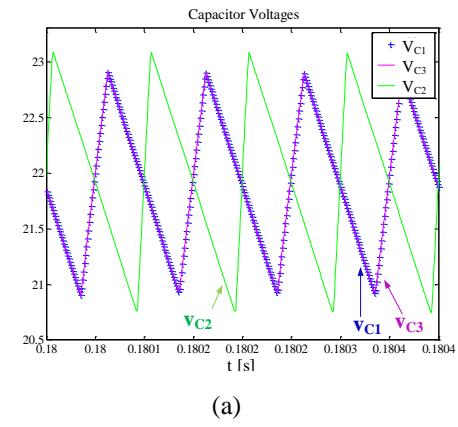


Fig. 8: Simulation waveforms: Voltage of capacitors C_1 , C_2 , C_3 (normal mode). Input power: (a) 120 W, (b) 500 W.

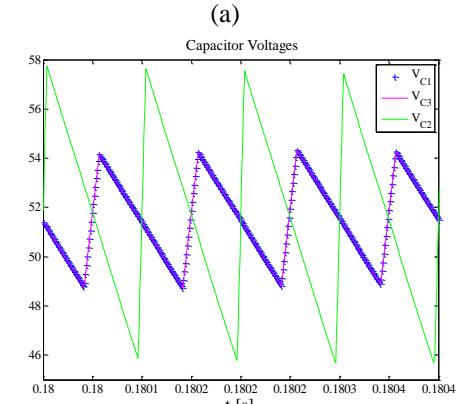
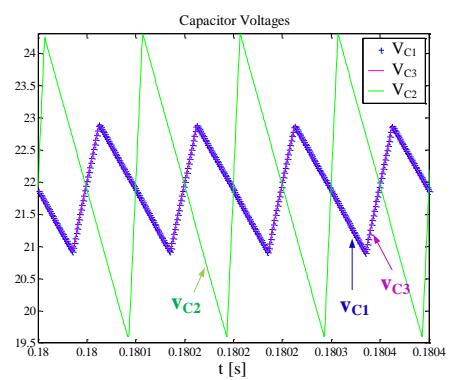
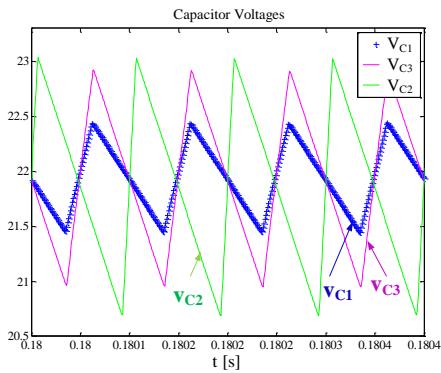
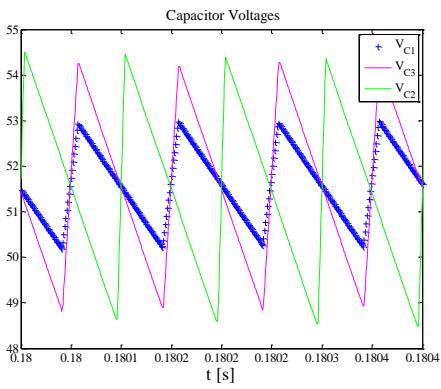


Fig. 10: Simulation waveforms: Voltage of capacitors C_1 , C_2 , C_3 first state ($C_2 = 50 \mu F$). Input power: (a) 120 W, (b) 500 W.

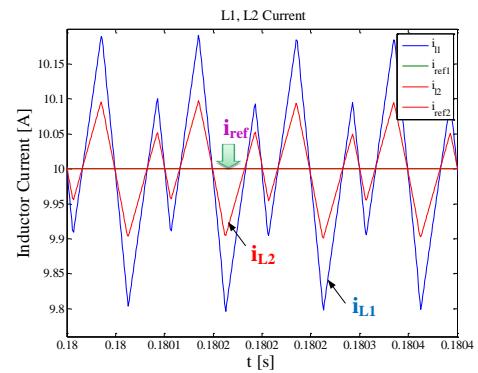


(a)

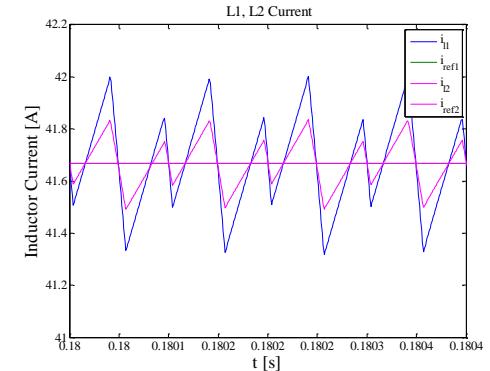


(b)

Fig. 11: Simulation waveforms: Voltage of capacitors C_1 , C_2 , C_3 second state ($C_1 = 200 \mu\text{F}$). Input power: (a) 120 W, (b) 500 W.

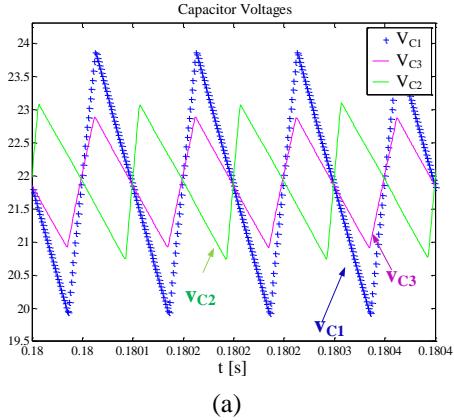


(a)

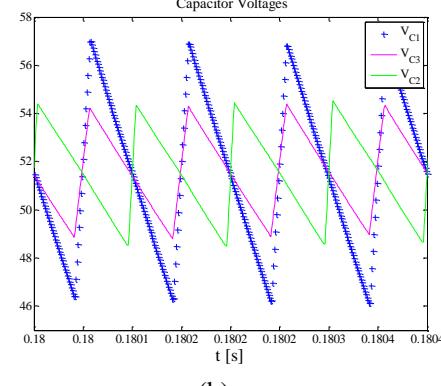


(b)

Fig. 13: Simulation waveforms: currents of L_1 and L_2 inductors and related reference currents (third case) ($L_2 = 2L_1$). Input power: (a) 120 W, (b) 500 W.

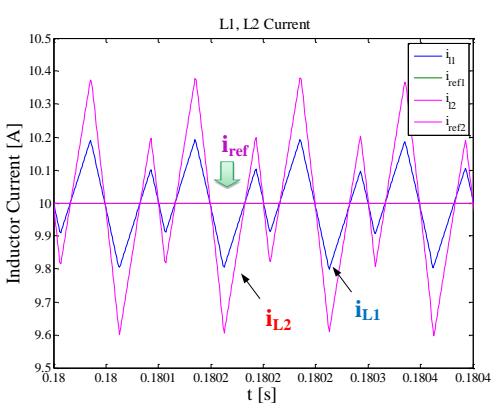


(a)

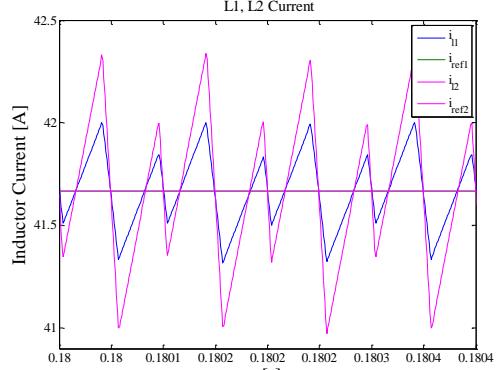


(b)

Fig. 12: Simulation waveforms: Voltage of capacitors C_1 , C_2 , C_3 second state ($C_1 = 200 \mu\text{F}$). Input power: (a) 120 W, (b) 500 W.



(a)



(b)

Fig. 14: Simulation waveforms: currents of L_1 and L_2 inductor and related reference currents (third case) ($L_2 = 0.5L_1$). Input power: (a) 120 W, (b) 500 W.

4. CONCLUSION

In this paper, a DC modular converter is considered to balance the output voltage. This converter can be used in fuel cell applications. In several scenarios, the values of the inductors and the capacitors of the modules changed. Using MATLAB/Simulink software, the waveforms of the system are plotted. The simulation results are performed on the variation in capacitor and inductance to investigate their modification effects. It was observed that changing the values of capacitors and inductors had no effect on the capacitance-voltage balance in the steady-state regime. As a result, the inductance (L) and capacitor (C) values do not affect the controlled zones.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Mohammad Afkar: Conceptualization, Data curation, Investigation, Software, Visualization, Writing - original draft, Writing - review & editing. **Parham Karimi:** Conceptualization, Data curation, Investigation, Visualization, Writing - review & editing. **Roghayeh Gavagsaz-Ghoachani:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing - review & editing. **Matheepot Phattanasak:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - review & editing. **Serge Pierfederici:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] S. Saib, Z. Hamouda, and K. Marouani, "Energy management in a fuel cell hybrid electric vehicle using a fuzzy logic approach," in *2017 5th International Conference on Electrical Engineering - Boumerdes (ICEEB)*, 2017.
- [2] S-M. Nosratabadi, R. Hemmati, M. Bornapour, and M. Abdollahpour, "Economic evaluation and energy/exergy analysis of PV/Wind/PEMFC energy resources employment based on capacity, type of source and government incentive policies: Case study in Iran," *Sustain. Energy Technol. Assessments*, vol. 43, 2021.
- [3] S. R. Paital, P. C. Pradhan, A. Mohanty, P. K. Ray and M. Viswawandya, "Power management in wind-fuel cell-ultracapacitor based autonomous hybrid power system," in *IEEMA Engineer Infinite Conference*, New Delhi, 2018.
- [4] X. Lü, Y. Qu, Y. Wang, C. Qin, and G. Liu, "A comprehensive review on hybrid power system for PEMFC-HEV: Issues and strategies," *Energy Conversion and Management*, vol. 171, pp. 1273-1291, 2018.
- [5] J. Zhao, Z. Tu, and S. H. Chan, "Carbon corrosion mechanism and mitigation strategies in a proton exchange membrane fuel cell (PEMFC): A review," *Journal of Power Sources*, vol. 488, article 229434, 2021.
- [6] M. Bahrami et al., "Design and modeling of an equalizer for fuel cell energy management systems," *IEEE Transactions on Power Electronics*, vol. 34, no. 11, pp. 10925-10935, Nov. 2019.
- [7] M. Afkar, R. Gavagsaz-Ghoachani, M. Phattanasak, and S. Pierfederici, "Commandable areas of a modular converter for DC voltage imbalance mitigation in fuel cell systems," *Sustainable Energy Technologies and Assessments*, vol. 48, Dec. 2021.
- [8] M. Z. Hossain, N. A. Rahim, and J. a/l Selvaraj, "Recent progress and development on power DC-DC converter topology, control, design and applications: A review," *Renewable & Sustainable Energy Reviews*, vol. 81, Part 1, pp. 205–230, 2018.
- [9] W. Li, and X. He, "Review of Nonisolated High-Step-Up DC/DC Converters in Photovoltaic Grid Connected Applications," *IEEE Transactions on Industrial Electronic*, vol. 58, no. 4, pp. 1239–1250, 2011.
- [10] W. Chen, Q. Huang, C. Li, G. Wang, and W. Gu, "Analysis and comparison of medium voltage high power DC/DC converters for offshore wind energy systems," *IEEE Transactions on Power Electronic*, vol. 28, no. 4, pp. 2014–2023, 2013.
- [11] S. F. Tie, and C. W. Tan, "A review of energy sources and energy management system in electric vehicles," *Renewable Sustainable Energy Review*, vol. 20, pp. 82–102, 2013.
- [12] L. M. Fernandez, P. Garcia, C. A. Garcia, and F. Jurado, "Hybrid electric system based on fuel cell and battery and integrating a single dc/dc converter for a tramway," *Energy Conversion & Management*, vol. 52, no. 5, pp. 2183–2192, 2011.
- [13] L. Costa, S. A. Mussa, and I. Barbi, "Multilevel buck/boost-type DC-DC converter for high power and high voltage application," *IEEE Transactions on Industrial Application*, vol. 50, no. 6, pp. 3931-3942, 2014.
- [14] S. Ratanapanachote, H. J. Cha, and P. N. Enjeti, "A digitally controlled switch mode power supply based on matrix converter," *IEEE Transactions on Power Electronics*, vol. 21, no. 1, pp. 124–130, 2006.
- [15] G. Sulligoi, D. Bosich, G. Giadrossi, L. Zhu, M. Cupelli, and A. Monti, "Multiconverter medium voltage dc power systems on ships: constant-power loads instability solution using linearization via state feedback control", *IEEE Transactions on Smart Grid*, vol. 5, pp. 2543–2552, 2014.
- [16] G. J. M. de Sousa and M. L. Heldwein, "Three-phase unidirectional modular multilevel converter," in *15th European Conference on Power Electronics and Applications (EPE)*, 2013.

- [17] Y. Tang, D. Fu, T. Wang and Z. Xu, "Hybrid switched-inductor converters for high step-up conversion," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 3, pp. 1480-1490, March 2015.
- [18] M. Afkar, R. Gavagsaz-Ghoachani, M. Phattanasak, J.-P. Martin, S. Pierfederici, "Proposed system based on a three-level boost converter to mitigate voltage imbalance in photovoltaic power generation systems," *IEEE Trans. Power Electron.*, vol. 37, no. 2, pp. 2264-2282, Feb. 2022.
- [19] D. Dell'Isola, "Optimization of DC/DC converters for embedded systems including dynamic constraints," Dissertation, Groupe de recherche en énergie électrique de Nancy (Vandœuvre-lès-Nancy), Université de Lorraine, 2020.
- [20] A. M. Roldan, A. Barrado, J. Pleite, J. Vazquez, and E. Olias, "Size and cost reduction of the energy storage capacitors," in *Proc. Applied Power Electronics Conference and Exposition (APEC)*, 2004, pp. 723-729.

BIOGRAPHY



Mohammad Afkar received the B.Sc. and M.Sc. degrees in electrical engineering and renewable energy engineering from the Shahid Beheshti University, Tehran, in 2014 and 2020, respectively.

He is with the laboratory of Renewable energies engineering (REDSBU), Shahid Beheshti University, Tehran, Iran. His research interests include power electronics, control and converters for fuel cell (FC), and photovoltaic (PV) systems.



Parham Karimi received the M.Sc. degree in renewable energy engineering from the Shahid Beheshti University, Tehran, in 2021. In addition, he received a bachelor's degree in biomedical engineering from the Islamic Azad University of Najafabad, Isfahan, Iran.

He is cooperated with "Laboratoire d'Energétique et de Mécanique Théorique et Appliquée" (LEMTA), King Mongkut's University of Technology North Bangkok (KMUTNB) and Shahid Beheshti University in research field. His research

interests include modular converters and engineering education.



Roghayeh Gavagsaz-Ghoachani received the M.Sc. degree from the Institut National Polytechnique de Lorraine (INPL), Nancy, France, in 2007, and the Ph.D. degree from the Université de Lorraine, France, in 2012, all in electrical engineering. She is with the Department of Renewable Energies Engineering, Shahid Beheshti University, Tehran, Iran. She is also a researcher in the "Groupe de Recherche en Energie Electrique de Nancy" (GREEN), and the "Laboratoire d'Energétique et de Mécanique Théorique et Appliquée" (LEMTA), Université de Lorraine, France. Her current research interests include the stability study, control of power electronics systems and renewable energy.



Matheepot Phattanasak received the B.Sc. and ME degrees in electrical engineering from King Mongkut's Institute of Technology North Bangkok, Thailand, in 1996 and 2004, and the Ph.D. degree in electrical engineering in 2012 from Université de Lorraine, France.

He is currently a Full Professor with the Department of Teacher Training in Electrical Engineering (TE), King Mongkut's University of Technology North Bangkok (KMUTNB). His current research interests include power electronics, and their controllers.



Serge Pierfederici received the Dipl.-Ing. from the Ecole Nationale Supérieure d'Electricité et Mécanique, Nancy, France, in 1994, and the Ph.D. degree in electrical engineering from the Institut National Polytechnique de Lorraine, Nancy, France, in 1998.

Since 2009, he has been engaged as a Full Professor at the University of Lorraine, Nancy, France. He is authored and coauthored more than 200 papers which are published in the international peer-reviewed journals. His research interests include the stability study of distributed power systems, modeling, and control of power electronic systems, and distributed control of multisources and multicarrier microgrids. Prof. Pierfederici was the recipient of several IEEE awards and he serves on the Editorial Boards of the international peer-reviewed journals.

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Robustness Analysis of Model Reference Adaptive Controller in The Presence of Input Saturation Using Describing Function Method

Fatemeh Tavakkoli¹, Alireza Khosravi^{1,*}, Pouria Sarhadi^{1,2}

¹ Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol 47148-71167, Iran

² Research Fellow in Autonomous Vehicles, Department of Mechanical Engineering Science, University of Surrey, Guildford, GU2 7XH, UK

* Corresponding Author: akhosravi@nit.ac.ir

Abstract: This work represents a new method for robustness analysis of the model reference adaptive controller (MRAC) in the presence of input saturation. Saturation is one of the nonlinear factors affecting the stability of control systems which must be considered in controller design and stability analysis experiments. Various methods are presented for the stability and robustness analysis of adaptive control systems, and employment of describing function (DF) can be attractive and practical, due to the appropriate effectiveness of DF in estimating limit cycles and also the application of quasi-linearization theory. In this work, the stability analysis and a limit cycle estimation of a saturated system in the frequency domain are performed. The controller parameters are adjusted in a way that the system achieves its stable limit cycle in the presence of the initial conditions for the states. Moreover, the efficiency of the proposed method for second-order systems is reported in the presence of symmetric saturation and uncertainty model in Rohrs's counterexample as the unmodeled dynamics. The results demonstrate the proposed method provides a proper analysis of system stability during the changes in the control parameters and the saturation amplitude.

Keywords: Input saturation, unmodeled dynamic, describing function, frequency response, model reference adaptive control.

Article history

Received 22 June 2022; Revised 30 October 2022; Accepted 06 December 2022; Published online 29 March 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

F. Tavakkoli, A. Khosravi, and P. Sarhadi, "Robustness analysis of model reference adaptive controller in the presence of input saturation using describing function method," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 62-69, 2023.

DOI: [10.22055/jaree.2022.41169.1062](https://doi.org/10.22055/jaree.2022.41169.1062)



1. INTRODUCTION

The signal level that a stimulus can deliver is usually limited by physical or safety constraints. These limits exist in all control systems including force, torque, voltage, and flow. The impact of amplitude saturation in the design of control system often depends on the control system performance. This effect is ignored in some systems and an appropriate functioning of most systems occurs when the amplitude saturation is taken into account [1-2]. Therefore, the controller design in the presence of saturation and its identification has recently been studied. One of the first works to address this issue is metaheuristic-based optimization algorithms [3]. The saturated system is identified using different optimization methods and described the differences between them. in addition, due to the destructive effects of saturation on the system, many controllers are vulnerable to the nonlinear factors due to the performance changes of the

closed-loop system that often cause an instability in the system. In order to solve this issue, anti-windup compensators are designed by investigating the effects and properties of saturation on the system and controller to guarantee stability of the system as well as preventing the occurrence of saturation in the control systems. These designs are based on different definitions of saturation effects [4]. The investigation of the stability analysis and the stabilization of linear systems is performed in the presence of saturation and different kinds of Lyapunov functions such as polyhedral, quadratic and Lure which are used to model the saturation section to analyze the behavior of the closed-loop nonlinear system [5]. Additionally, there are several methods for designing anti-windups such as LMI [6-8]. Accordingly, a coefficient can be determined based on the optimal solution of LMI, which affects the controller state equations and provides stability of the loop system [9]. Moreover, the compensation design is also proposed that the main idea is

that they combine Nussbaum gain technique into backstepping control to compensate the saturation input, mainly used in spacecraft [10]. Other types of compensators including nonlinear multi-input and multi-output (MIMO) systems [11], ship steering control [12], switched nonlinear systems [13], and hydro-turbine governing systems [14] are also introduced.

The model reference adaptive controller (MRAC) is an attractive method to design adaptive systems due to its acceptable performance. Accordingly, some compensators have been proposed for the MRAC in the presence of saturation. Modern compensators are used for the PID MRAC controller which is implemented on the practical system of the autonomous underwater vehicle (AUV) [15]. In addition to MRAC method, the positive μ method is proposed to reduce the saturation effects in the adaptive system in [16]. In the positive μ method, a coefficient known as μ is considered in the design of the control signal and affects the control signal when the system is saturated, consequently reduces the value of the signal. In addition to designing compensators in the presence of saturation, robustness analysis in the presence of unmodeled dynamics and uncertainties in a system is an important subject in the adaptive control [17]. Berk Altin and Kira Barton have shown how unmodeled dynamic (Rohrs counterexample [18]) causes instability in model reference adaptive iterative learning control (MRAILC) [19]. Eugene Lavretsky et al. have proposed a method based on the MRAC to investigate the general stability of the system with unmodeled dynamics [20]. The model reference adaptive controller was also used for the robustness of linear time-variant systems with temporal delay [21]. In these works, the stability analysis of the adaptive control system has been carried out with the Lyapunov function -albeit in the absence of saturation- and presented methods for the robustness of the controller.

One way to analyze the stability of nonlinear systems is the describing function. A frequency response method is a powerful tool for analyzing and designing linear control systems. However, frequency analysis cannot be applied directly for nonlinear systems because the frequency response function cannot describe the nonlinear system. Therefore, the describing function is used to approximate the analysis and estimate the nonlinear behaviors. The most important application of the describing function method is to estimate the limit cycle of nonlinear systems [22-24]. Additionally, the describing function has been examined in the analysis of nonlinear systems with memory [25-26]. Recently, the describing function is used for the robustness analysis of reference model adaptive systems in the presence of unmodeled dynamics and the describing function of the reference model adaptive controller [27].

In this work, the robustness of adaptive control systems to unmodeled dynamics has been upgraded to be used for second-order system with input saturation [27]. Adaptation, plant and saturation rules are converted into a lure model, which consists of a linear plant, nonlinear saturation on the forward path and an isolated nonlinear part in the feedback path. The describing function is then used to analyze the system in the frequency domain. By placing the DF of nonlinear parts and analysis via the Nyquist diagram, the prediction of the limit cycles of the system is achieved. The

main goal of this work is to estimate the limit cycle the system and determine the approximate initial conditions for the adaptive system with nonlinear factors, which leads to the robustness of the system to reach its stable limit cycle. The application of the proposed method is investigated by simulating the second-order system in the presence of symmetric saturation and Rohrs counterexample as unmodeled dynamics. In addition, the rest of the paper is organized in a way to address the issues presented as follows. In Section II, the main problem is explained in the presence of saturation and the transformed parameters of the controller and the update rules obtained by transformations. In Section III, the describing function of the nonlinear parts is calculated and the limit cycle of the closed-loop system is estimated by drawing the Nyquist diagram. In Section IV, the second-order system with unmodeled dynamics and input saturation is included to indicate the usefulness of the proposed analysis method.

2. PROBLEM DESCRIPTION

In this section, the control system is explained in the presence of amplitude saturation. To simplify the analysis, new parameters are introduced, based on which updating rules are obtained. The second-order system under investigation has the following state equations (1).

$$\begin{cases} \dot{x}_p = A_p x_p + B_p(v(t) + f(x_p)) \\ f(x_p) = k_p x_p \end{cases} \quad (1)$$

where x_p is a state variable, $B_p \in R^{2 \times 1}$ and $A_p \in R^{2 \times 2}$ are known and constant, (A_p, B_p) are controllable, $f(x_p)$ is the linear state-dependent uncertainty, and $v(t)$ is the scalar input. The unmodeled dynamic is as (2):

$$\begin{cases} x_\eta = A_\eta x_\eta(t) + B_\eta U_{sat}(t) \\ v(t) = C_\eta^T x_\eta(t) \end{cases} \quad (2)$$

where $x_\eta \in R^{m \times 1}$, $C_\eta^T \in R^{1 \times m}$, and $A_\eta \in R^{m \times m}$ is a Hurwitz matrix with $G_\eta = C_\eta^T (sI_{m \times m} - A_\eta)^{-1} B_\eta$ and $(C_\eta^T, A_\eta, B_\eta)$ are controllable and observable, U_{sat} is the saturated control input which is defined as written in (3).

$$U_{sat}(u) = \begin{cases} u_{max} & \text{if } u > u_{max} \\ u & \text{if } -u_{max} < u < u_{max} \\ -u_{max} & \text{if } u < -u_{max} \end{cases} \quad (3)$$

$$u = \theta^T(t)x_p(t)$$

where u is the control input which is calculated by the controller, u_{max} is the maximum control signal that can be created by the stimulus, and $\theta^T(t) = [\theta_0 \ \theta_1]$ are the adaptation rules (4) (refer to Fig. 1).

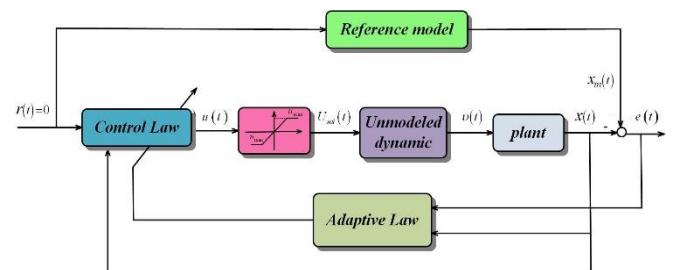


Fig. 1: The block diagram of the system in the presence of saturation.

$$\dot{\theta} = -\Gamma x_p B_p^T P e \quad (4)$$

where P is obtained from solving the Lyapunov.

In equations $A_m^T P + PA_m = -Q$, and $e = x_p - x_m$, x_m is the reference state variable and is defined as (5), and (6):

$$\begin{aligned} \dot{x}_m &= A_m x_m + B_m r(t) \\ A_m &= A_p + B_p \theta^{*T} \\ A_m &= A_p + B_p (-K_{LQR})^T \\ \theta^* &= -K_{LQR} \end{aligned} \quad (5) \quad (6)$$

where $r(t)$ is the reference input, A_m and B_m are the reference model state matrices, A_m is a Hurwitz matrix, and $\theta^{*T}(t) = [\theta_0^* \theta_1^*]$ are the ideal adaptive rules. A linear designing technique is employed to determine θ^{*T} . The linear-quadratic regulator (LQR) method is chosen as a tool for reference model control design. R_{LQR} and Q_{LQR} are its weight matrices and K_{LQR} is its vector for control parameters.

To simplify the analysis and determine the effect of saturation and unmodeled dynamic on the adaptive system, non-singular transformations are defined as equation (7):

$$\begin{aligned} \varepsilon(t) &= Ce(t) \\ v(t) &= M\theta(t) \end{aligned} \quad (7)$$

where $\varepsilon(t)$ is the transformed error, $v(t)$ is the transformed parameter, C and M are the transform matrices. The transform matrices are defined as (8). The proof of the equations are expressed in reference [21-27].

$$C = [c_0 \ c_1]^T \quad M = p_b C P^{-1} \quad (8)$$

where $p_b = \sqrt{B_m^T P B_m}$, and C is defined as (9):

$$c_0 = p_b^{-1} P B_m, \quad c_1 c_1^T = P - c_0 c_0^T \quad (9)$$

Remark 1 : From (8) and (9) the following equations can be obtained:

$$c^T_0 B_m = p_b, \quad c^T_1 B_m = 0$$

$$CP^{-1}C^T = I$$

Using (5) and (8) the $n \times n$ matrix below is determined as (10):

$$\begin{aligned} \tilde{A}_m &= CA_m P^{-1} C^T = \begin{pmatrix} \alpha_{00} & a_1 \\ a_0 & \tilde{A}_m' \end{pmatrix} \\ \alpha_{i,j} &= c_i^T A_m P^{-1} c_j \quad \forall i, j = \{0, 1\} \end{aligned} \quad (10)$$

where $\tilde{A}_m' \in R^{(n-1) \times (n-1)}$. It can be shown that \tilde{A}_m' and \tilde{A}_m are Hurwitz matrices. The error dynamic is defined as follows according to the closed loop system equation (11):

$$\dot{e} = A_m e + B_p \tilde{\theta}^T x_p + B_p \eta \quad (11)$$

where $\eta = v - U_{sat}$ is dependent on the saturated control input and the unmodeled dynamic, and $\tilde{\theta} = \theta - \theta^*$ is defined with θ^* in (6). The transformed error $\dot{\varepsilon}_i = C_i^T \dot{e}$ is obtained from (7) and (11) as equation (12).

$$\dot{\varepsilon} = \begin{bmatrix} c^T_0 A_m e + p_b \tilde{\theta}^T x_p + p_b \eta \\ c^T_1 A_m e \end{bmatrix} \quad (12)$$

Considering Remark1 and using (10) and (12) the equation (13) is concluded:

$$\dot{\varepsilon}_1 = \tilde{A}'_m \varepsilon_1 + a_0 \varepsilon_0 \quad (13)$$

$\dot{\varepsilon}_0$ can also be rewritten as (14):

$$\begin{aligned} \dot{\varepsilon}_0 &= (\alpha_{00} + \tilde{v}_0) \varepsilon_0 + (a_1 + \tilde{v}_1) \varepsilon_1 + \dots \\ &\quad + p_b \eta + \tilde{v}_0 m_0 + \tilde{v}_1 m_1 \end{aligned} \quad (14)$$

where $v^* = M\theta^*$, $\tilde{v}_i = v_i - v_i^*$, and $m_i = C_i^T x_m$. The proposed adaptive rules are the revised standard adaptive rules using projection algorithm as (15).

$$\dot{\theta} = M^{-1} \dot{v}$$

$$\dot{v}_i = proj(\{M\theta\}_i, -\{M\Gamma x_p B_m^T P e\}_i) \quad (15)$$

$$\text{where } \Gamma = \gamma P$$

$$proj(\theta_i, y_i) = f(x) = \begin{cases} \frac{\theta_{i,max}^2 - \theta_i^2}{\theta_{i,max}^2 - \theta_i^2}, & \theta_i \epsilon \Omega_i \Lambda \theta_i y_i > 0 \\ y_i, & \text{Otherwise} \end{cases}$$

$$\overline{\Omega}_i = \{\theta_i \epsilon \tilde{R} | \theta'_{i,max} \leq \theta_i \leq \theta_{i,max}\}$$

$$\underline{\Omega}_i = \{\theta_i \epsilon \tilde{R} | -\theta_{i,max} \leq \theta_i \leq -\theta'_{i,max}\}$$

$$\Omega_i = \overline{\Omega}_i \cup \underline{\Omega}_i$$

And the positive constant $\theta_{i,max} > \theta'_{i,max}$.

Remark 2 : It can be shown that if $\forall t \geq t_a \ \| \theta_i(t_a) \| \leq \theta_{i,max} \Rightarrow \| \theta_i(t) \| \leq \theta_{i,max}$ then the projection algorithm can guarantee the limitations of θ_i which is independent of the system dynamic.

Considering $\dot{v}_i = p_b C_i^T P^{-1} \theta$, and the transformed error as (16),

$$\dot{v}_i = \gamma' proj(v_i, (\varepsilon_i + m_i) \varepsilon_0) \quad (16)$$

in which $\gamma' = \gamma p_b^2$, Remark 2 guarantees that v_i^* will ultimately converge into the projection region.

3. DESCRIBING FUNCTION

The describing function is a classic tool to analyze the existence of the limit cycles in nonlinear systems based on the frequency response method [28]. The idea of the method is based on Gaussian linearization, so that the nonlinear part is considered as a single block, the describing function is a complex coefficient based on the main harmonics of the nonlinear system whose input is sinusoidal and its output is obtained through the Fourier series. The DF has many applications in nonlinear controllers and extensive researches have been performed where different methods such as Two-Sinusoid-Input Describing Function (TSIDF) and Dual-Input Describing Function (DIDF) have been proposed [26-29]. Despite the favourable characteristics of the describing functions, it is rarely used in adaptive control because of the complexity of the analysis of nonlinear systems with memory [29]. In a recent study, the stability of the MRAC is analyzed and the describing function of the controller is calculated using the DF method [25]. Although the DF is an approximate method, it is superior to other methods for nonlinear system analysis due to the desirable properties of the frequency response technique. In the following section, the method of the obtaining the function is introduced first, based on that,

the main system is divided into two linear and nonlinear parts. Then, the DF of the nonlinear sections is calculated and the stability analysis of the system will be provided via plotting the Nyquist diagram.

3.1. Calculating The Describing Function

If the nonlinear section is considered as a block with a sinusoidal input of amplitude A and the frequency ω , i.e., $x(t) = A\sin(\omega t)$, its output, $w(t)$, is often a periodic function, despite it often being a non-sinusoidal (Fig. 1). Using the Fourier series, the periodic function $w(t)$ can be extended as (17):

$$w(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(n\omega t) + b_n \sin(n\omega t)] \quad (17)$$

where the Fourier coefficients a_n and b_n , which are often a function of A and ω , are determined by the equations below:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} w(t) d(\omega t)$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} w(t) \cos(n\omega t) d(\omega t)$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} w(t) \sin(n\omega t) d(\omega t)$$

The describing function should have some conditions, one of which being $a_0 = 0$. Furthermore, the main component is considered in the Fourier series. That is equation (18):

$$\begin{aligned} w(t) &\approx w_1(t) = a_1 \cos(\omega t) + b_1 \sin(\omega t) \\ &= M \sin(\omega t + \varphi) \end{aligned} \quad (18)$$

where

$$M(A, \omega) = \sqrt{a_1^2 + b_1^2}$$

$$\varphi(A, \omega) = \arctan\left(\frac{a_1}{b_1}\right)$$

Equation (18) represents the main component corresponding to a sinusoidal input which is a sinusoidal with the same frequency as the input. This sinusoidal can be written as follows in a complex display as (19):

$$w_1 = M e^{j(\omega t + \varphi)} = (b_1 + j a_1) e^{j\omega t} \quad (19)$$

Similar to the concept of the frequency response function which was the ratio of the sinusoidal input and the sinusoidal output of a system in the frequency domain, the describing function of a nonlinear element is the complex ratio of the main component of the nonlinear element to the sinusoidal input. This means equation (20):

$$N(A, \omega) = \frac{M e^{j(\omega t + \varphi)}}{A e^{j\omega t}} = \frac{M}{A} e^{j\varphi} = \frac{1}{A} (b_1 + j a_1) \quad (20)$$

3.2. Lure Model

The proposed controller is designed based on the proven concepts in [21-22] and it will be assumed for transforming into a lure model that $x_m(t_0) = 0$. Since it is assumed for analyzing the describing function that $r = 0$, therefore $\forall t \geq t_0$ $x_m(t) = 0$ and $\forall t$ $m_i(t) = 0$. Considering (14) and (16), the equation for the adaptive controller is as (21), and (22):

$$\dot{\varepsilon}_0 = (\alpha_{00} + \tilde{\nu}_0) \varepsilon_0 + (a_1 + \tilde{\nu}_1) \varepsilon_1 + p_b \eta$$

$$\dot{\varepsilon}_i = -\gamma_i \varepsilon_0 \varepsilon_i \quad (21)$$

where:

$$\gamma_i = \begin{cases} \frac{\nu_{i,max}^2 - \varepsilon_i^2}{\nu_{i,max}^2 - \varepsilon_{i,max}^2}, & \text{if } |\varepsilon_i| \geq \nu_{i,max} \\ \gamma', & \text{Otherwise} \end{cases}$$

$$\gamma' = \gamma p_b^2$$

$$\nu_{i,max} = \nu_{i,max}' + \epsilon_i, \epsilon_i > 0 \quad (22)$$

Furthermore, from (21) we have $\varepsilon_0 = -\gamma_1 \varepsilon_0^2$. Thus ε_0 is negative for all times if $\varepsilon_0(t_0) > -\nu_{0,max}$. Hence, the parameter ε_0 will ultimately converge to $\nu_{0,max}$ so, it is assumed that $\varepsilon_0(t_0) = -\nu_{0,max}$. The control input is calculated according to Fig. 2 as follows:

$$u = p_b^{-1} \nu^T \varepsilon = p_b^{-1} (-\nu_{0,max} \varepsilon_0 + \varepsilon_1 \varepsilon_1) \quad (23)$$

where $\nu_{0,max}$ is a constant.

To separate the linear part from the nonlinear one, the plant and its unmodeled dynamic are considered as the linear block G_0 , the section for updating the adaptation rules as a nonlinear block in the feedback, and the saturation is considered as a nonlinear block in the forward path (Fig. 3).

3.3. Analyzing the Describing Function (DF)

Typically, to compute the DF of the nonlinear part, all of it is considered as a single entity. Since there are two distinct nonlinear parts in here, the DF of each part is calculated individually. In order to obtain the DF of the adaptive rules section, it is assumed that one of its inputs is sinusoidal in the form of $\varepsilon_1 = A_1 \sin(\omega_1 t)$ which is produced from the linear part according to Fig. 4. ε_0 is acquired by putting ε_1 in (13) where we will have equations (24) and (25).

$$\varepsilon_0(t) = \frac{A_1(\omega_1 \cos(\omega_1 t) - \tilde{A}'_m \sin(\omega_1 t))}{a_0} \quad (24)$$

$$\begin{aligned} u(t) &= \\ &\left(\frac{4\tilde{A}'_m A_1^3 \varphi - 8a_0 A_1 v_{1,max} \omega_1 - 3A_1^3 \gamma' \omega_1}{8a_0 \omega_1 p_b} \right) \sin(\omega_1 t) \\ &+ \left(\frac{4A_1^3 \omega_1 - 8A_1 \tilde{A}'_m v_{0,max} \omega_1}{8a_0 \omega_1 p_b} \right) \sin(\omega_1 t) \\ &- \left(\frac{A_1^3 \gamma' \tilde{A}'_m - 8A_1 v_{0,max} \omega_1^2}{8a_0 \omega_1 p_b} \right) \cos(\omega_1 t) \\ \varphi &= \tan^{-1}\left(\frac{\omega_1}{-\tilde{A}'_m}\right) \end{aligned} \quad (25)$$

The obtained $u(t)$ is approximate. Hence, according to the definition of the DF in (20), with $\varepsilon_1 = A_1 \sin(\omega_1 t)$ and the output in (25), the describing function of the nonlinear part in the feedback is acquired.

$$\begin{aligned} N_{A,L}(A_1, \omega_1) &= \left(\frac{4\tilde{A}'_m A_1^2 \varphi - 8a_0 v_{1,max} \omega_1 - 3A_1^2 \gamma' \omega_1}{8a_0 \omega_1 p_b} \right) \\ &+ \left(\frac{4A_1^2 \omega_1 - 8A_1 \tilde{A}'_m v_{0,max} \omega_1}{8a_0 \omega_1 p_b} \right) \\ &- j \left(\frac{A_1^2 \gamma' \tilde{A}'_m - 8v_{0,max} \omega_1^2}{8a_0 \omega_1 p_b} \right) \end{aligned} \quad (26)$$

To obtain the describing function of the saturation

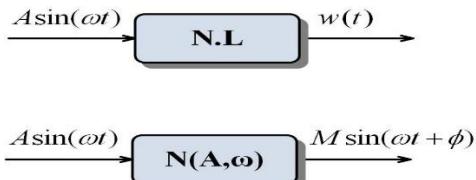


Fig. 2: The block diagram of a nonlinear element. Below: the display of its describing function [24].

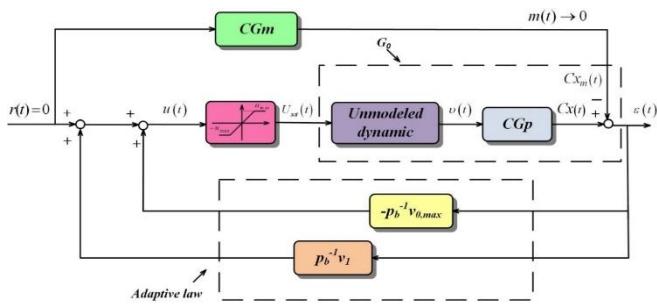


Fig. 3: The closed loop system with transformed states.

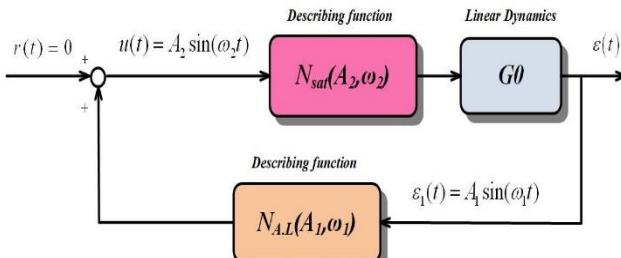


Fig. 4: The simplified block diagram of the closed loop system with the describing function.

section, it is assumed that $u(t) = A_2 \sin(\omega_2 t)$. This way if $A_2 > u_{max}$ then the describing function will be as equation (27):

$$N_{sat}(A_2) = \frac{2}{\pi} \left[\arcsin\left(\frac{u_{max}}{A_2}\right) + \frac{u_{max}}{A_2} \sqrt{1 - \left(\frac{u_{max}}{A_2}\right)^2} \right] \quad (27)$$

and if $A_2 < u_{max}$ then $N_{sat}(A_2) = 1$. The performance of the DF is that it allows the stability analysis of a nonlinear system in the frequency domain to be evaluated in the same manner as linear systems. Consider the system in Fig. 4 where the describing function of nonlinear elements is placed instead of the elements themselves. If it is assumed that $\varepsilon_1 = \text{Asin}(\omega t)$ then the frequency response of the close-loop will be:

$$\begin{aligned} G_0(j\omega)N_{A,L}(A, \omega)N_{sat}(A) + 1 &= 0 \\ \Rightarrow G_0(j\omega)N_{A,L}(A, \omega)N_{sat}(A) &= -1 \end{aligned} \quad (28)$$

Therefore, the describing function can be used to present the stability and robustness of the closed-loop nonlinear systems in a graphical form, such as the Nyquist diagram. In this way that the Nyquist diagram of the linear part of the system $G(j\omega)$ can be plotted as usual and the describing function $-1/(N_{A,L}(A, j\omega)N_{sat}(A))$ for different amplitudes on the same axes. In this way, the intersection point of the two diagrams shows the amplitude and the frequency of the limit cycle. Alternatively, the diagram of $G(j\omega)N_{A,L}(A, \omega)N_{sat}(A)$

is drawn for different amplitudes similar to the Nyquist diagram. Then, the point ‘-1’ is the point where the oscillation can occur [30]. It can be assumed stability analysis that the sinusoidal input has some phase, that is $\varepsilon_1(t) = \text{Asin}(\omega t + B)$. Accordingly, there are two equations and three variables for acquiring the limit cycle. Thus, the obtained solution will not be unique anymore. Therefore, it is assumed that the main harmonic does not have any phase.

Finally, given the initial conditions, it is observed that the obtained estimation for the limit cycle is correct and its amplitude and frequency has a very slight difference with the value acquired from the analysis. In the next section, this method is applied to a second-order system.

4. SIMULATION EXAMPLE

In this section, the intended method for stability analysis is simulated for a practical plant. By introducing the system in (1), the describing function of the adaptive rules in (26), and saturation in (27), the amplitude and frequency of the limit cycle will be achieved using (28), and after the analysis with the Nyquist diagram, the system is given the initial conditions. The analysis will be repeated for different values of the controller parameter and the saturation amplitude. The system of generic transport aircraft (DC-8) airplane is considered as the main system in the proposed method [31].

$$\begin{aligned} \begin{pmatrix} \dot{\alpha} \\ \dot{q} \end{pmatrix} &= \begin{pmatrix} \frac{Z_\alpha}{V} & 1 + \frac{Z_q}{V} \\ \frac{M_\alpha}{M_q} & M_q \end{pmatrix} \begin{pmatrix} \alpha \\ q \end{pmatrix} + \begin{pmatrix} \frac{Z_\delta}{V} \\ \frac{M_\delta}{M_\delta} \end{pmatrix} \Lambda(v(t) + f(x_p)) \\ f(x_p) &= f(\alpha, q) = K_\alpha \alpha + K_q q \end{aligned}$$

where α (deg) is the aircraft angle of attack, q (deg/s) is the pitch rate, V (ft/s) is the air speed (considered constant), $M_\delta, M_q, M_\alpha, Z_\delta, Z_q, Z_\alpha$ are the stability converters of the plane, $\Lambda > 0$ is the loss of control effectiveness, and $f(x_p)$ is the uncertainty of the dynamics of the system.

$$\begin{aligned} A_p &= \begin{pmatrix} -0.8060 & 1.0 \\ -9.1486 & -4.59 \end{pmatrix} & B_p &= \begin{pmatrix} -0.04 \\ -4.59 \end{pmatrix} \\ \Lambda &= 0.5, \quad K_\alpha = 1.5M_\alpha, \quad K_q = 0.5M_q & & \end{aligned} \quad (29)$$

highly damped second-order unmodeled dynamics [18-27], described by

$$G_\eta = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad \text{with } \zeta = 0.9912, \quad \omega_n = 15.1327 \quad (30)$$

The control signal δ_e (deg) is the elevator deflection in this system and control rules for determining the reference model are obtained using the LQR method with weight matrices of $R_{LQR} = 1$ and $Q_{LQR} = \text{diag}(0.5, 0.5)$. Other unknown parameters are selected as (31):

$$\gamma' = 1, u_{max} = 30, \quad B_m = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 0.2 & 0 \\ 0 & 40 \end{pmatrix} \quad (31)$$

It can be shown that:

$\tilde{A}_m' = 1.4995$, $a_0 = 31.22$, $\theta^* = [-0.0178 \quad 0.2185]^T$. $v_{0,max}$ and $v_{1,max}$ are selected to be 5 and 3.2 respectively. The value for the amplitude and the frequency of the limit cycle is obtained to be $A^* = 27$ and $\omega^* = 6.28$ using (28).

Moreover, if all parameters are adjusted then G_0 and $N_{A,L}$ are showed according to (32). Since the frequency of the limit cycle is determined by the plant, three different values are observed for the amplitude by drawing the Nyquist diagram for $G_0(j\omega)N_{total}(j\omega)$. As the amplitude is reduced, the diagram will circle around the point '-1' and vice versa. As explained before, the amplitude whose Nyquist diagram crosses the '-1' point is the amplitude of the limit cycle. Fig. 5 depicts the Nyquist diagram of the system for three different values of amplitudes.

$$G_0(S) = \frac{-4.58S - 546.6}{S^4 + 30.4S^3 + 282.9S^2 + 1346S + 9567}$$

$$N_{A,L}(A^* + j\omega^*) = -2.6432 - 1.7036i \quad (32)$$

Other amplitudes and frequencies may apply to (28) such as $A = 5.5$ and $\omega = 16.5$. But a limit cycle is stable if all of the paths around it will ultimately converge to it and these conditions are formulated as below:

$$\frac{\partial(\Im(-N(A, \omega)G_0(j\omega)))}{\partial\omega} \Big|_{A^*, \omega^*} > 0$$

$$\frac{\partial(\Re(-N(A, \omega)G_0(j\omega)))}{\partial A} \Big|_{A^*, \omega^*} > 0$$

Eventually, considering the system in (1), and (29), the unmodeled dynamic in (2), (30), the saturation in (3), the reference model in (5), and the adaptation rules in (31), the stable initial condition is acquired by changing the initial conditions $\varepsilon_1(0)$ and other initial conditions as follows:

$$x_m(0) = 0, x_\eta(0) = 0$$

$$x_p(0) = C^{-1}\varepsilon(0), \varepsilon_0(0) = 0.35\varepsilon_1(0)$$

Also, the controller parameters are as below:

$$v_{0,max} = 5, \varepsilon_0 = 0.1v_{0,max}, v_{1,max} = 3.2, \varepsilon_1 = 0.02v_{1,max}$$

$$v_{i,max} = v'_{i,max} + \varepsilon_i$$

In Fig. 6, the transformed error is displayed for two different initial values. Regarding the obtained values for the amplitude and frequency from the analysis, it is observed that the describing function method could successfully predict the limit cycle and according to the practical features of the system, the initial conditions of the states were less than 30 degrees.

The saturated control signal for this limit cycle is almost similar to the saturated control signal which is obtained by applying a sinusoidal input with the amplitude and frequency of the limit cycle. These two signals will approximately coincide with each other in most times because the limit cycle reaches its stability and this shows the appropriate initial conditions and the correct estimation of the limit cycle (Fig. 7).

As seen in (27), and (28), it is obvious that the DF is dependent on $v_{0,max}$, $v_{1,max}$ and u_{max} . This means that by changing these values, the limit cycle of the system will change. By changing the amplitude from 30 to 15 and the controller parameters being constant $v_{1,max} = 3.2$, and $v_{2,max} = 5$, the values $A = 45.5$ and $\omega = 6.3$ are obtained for the limit cycle from the frequency analysis. This limit cycle is shown in Fig. 8 for two different initial conditions.

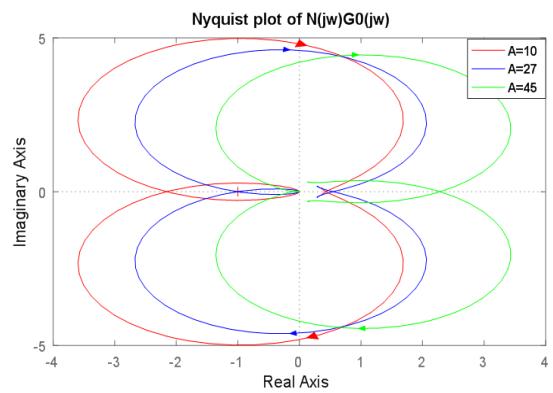


Fig. 5: The Nyquist diagram of $G(j\omega)N(A, \omega)$.

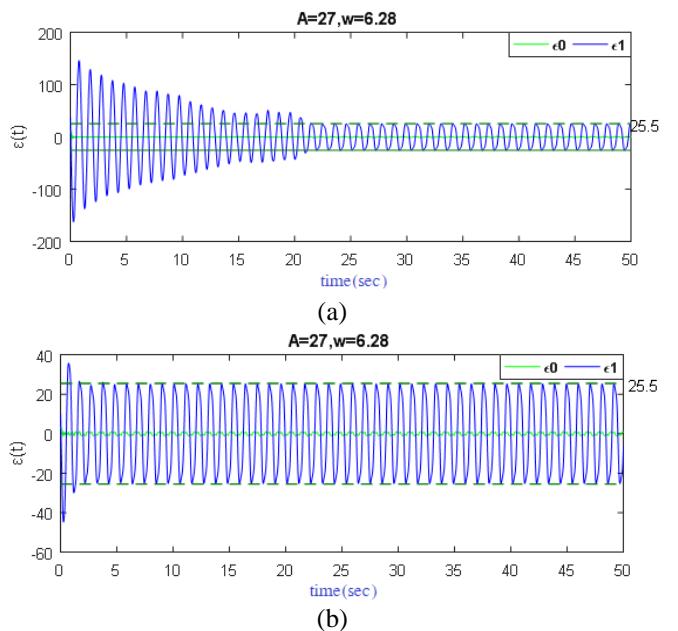


Fig. 6: The display of the stable limit cycle for two different initial conditions and the values $u_{max} = 30$, $v_{1,max} = 3.2$, and $v_{0,max} = 5$, (a) $\varepsilon_1(0) = 38$ and $x_p(0) = [15.6 \ 20.9]^T$, and (b) $\varepsilon_1(0) = 10$ and $x_p(0) = [4.14 \ 5.5]^T$.

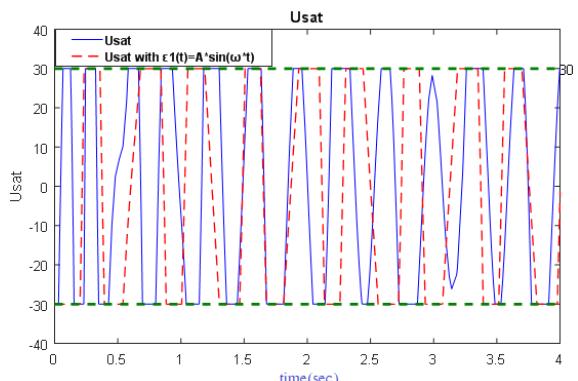


Fig. 7: The saturated control signal with $u_{max} = 30$ and the initial conditions of $\varepsilon_1(0) = 10$ and the next with the input $\varepsilon_1(t) = A * \sin(\omega * t)$.

Now, if the values are set to $v_{0,max} = 0.2$, and $v_{1,max} = 1.3$ and the saturation amplitude stays constant $u_{max} = 15$, the

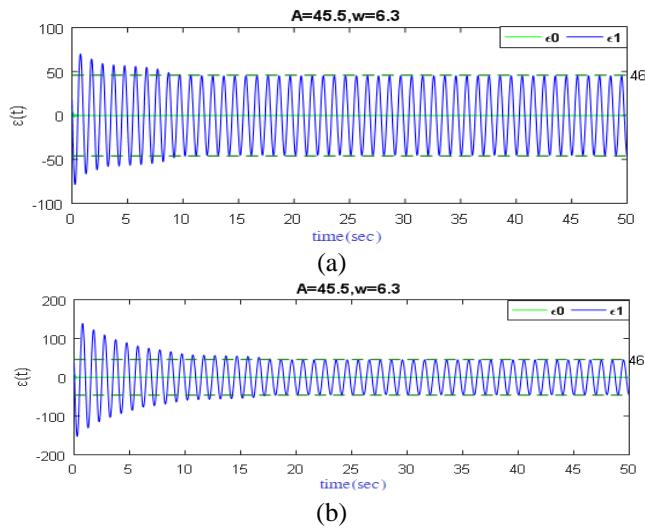


Fig. 8: The stable limit cycle for $u_{max} = 15$, $v_{0,max} = 0.2$ and $v_{1,max} = 1.3$, (a) The transformed error for initial conditions $\varepsilon_1(0) = 18$ and $x_p(0) = [7.4 \ 9.8]^T$, and (b) The transformed error for initial conditions $\varepsilon_1(0) = 35$ and $x_p(0) = [14.5 \ 19.3]^T$.

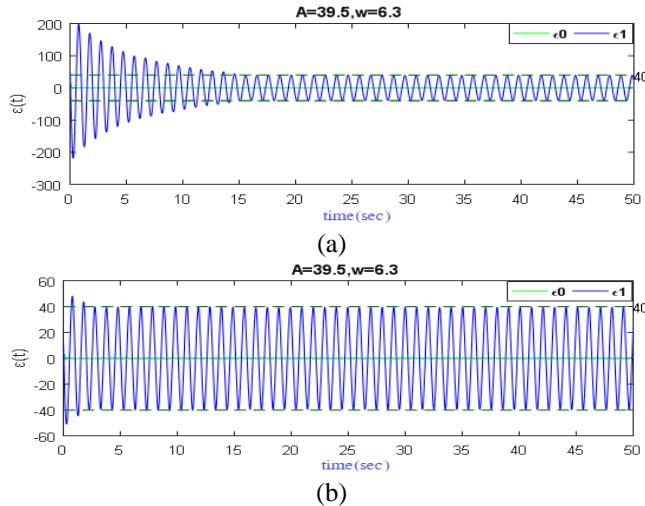


Fig. 9: The stable limit cycle for $u_{max} = 15$, $v_{0,max} = 0.2$ and $v_{1,max} = 1.3$, (a) The transformed error for initial conditions $\varepsilon_1(0) = 50$ and $x_p(0) = [20.7 \ 27.46]^T$, and (b) The transformed error for initial conditions $\varepsilon_1(0) = 12$ and $x_p(0) = [4.96 \ 6.6]^T$.

amplitude and the frequency of the limit cycle are acquired as $A = 39.5$ and $\omega = 6.3$ from the analysis. As seen in Fig. 9, the numerical simulation confirms the estimation obtained from the DF method. The limit cycle has a slight difference with the results of the analysis, the difference can be attributed to the describing function method because it is an approximate analytical method. The control signal and the saturated control signal for these updates and saturation parameters with the initial conditions $x_p(0) = [4.96 \ 6.6]^T$ are shown in Fig. 10.

5. CONCLUSION

In this work, an accurate stability analysis method for the model reference adaptive controller for second-order systems

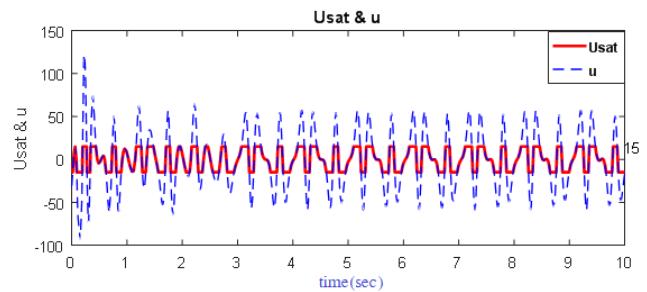


Fig. 10: The control signal and the saturated control signal for initial conditions $\varepsilon_1(0) = 12$, $u_{max} = 15$, $v_{0,max} = 0.2$, and $v_{1,max} = 1.3$.

in the presence of saturation, unmodeled dynamics and the uncertainty of the system has been proposed. The recognition of the effects of nonlinear factors such as saturation in the stability analysis of systems is undeniable. Since the describing function is one of the accurate methods for nonlinear systems analysis and in the general system, there were two nonlinear parts including saturation and the nonlinear adaptive controller, the DF method was employed for the stability analysis of the system and to predict the stable limit cycles. Using this analytical method, the accurate estimation of the limit cycle was performed and the parameters of the applied algorithm were adjusted in the MRAC and conditions were established so that the system reached its stable limit cycle given the initial conditions of variables. One of the main features of the proposed method is the use of frequency analysis to predict the limit cycles of the system and the correct approximation of their amplitude and frequency. The simulation results demonstrate the integrity of the method for second-order systems by changing the saturation amplitude and the controller adjustment parameter.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Fatemeh Tavakkoli: Formal analysis, Software, Writing - original draft, Writing - review & editing. **Alireza Khosravi:** Conceptualization, Supervision, Validation. **Pouria Sarhadi:** Validation, Supervision.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors

REFERENCES

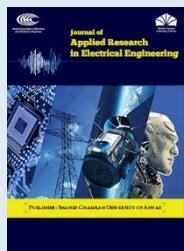
- [1] D. S. Bernstein, and A.N. Michel, "A chronological bibliography on saturating actuators," *International Journal of robust nonlinear control*, vol. 5, no. 5, pp. 375-380, 1995.
- [2] V. Kapila, and K. Grigoriadis, *Actuator Saturation Control*, ser. Control Eng. New York, NY: Marcel Dekker, 2002.
- [3] Sarhadi, P., A. Khosravi, and V. Bijani, "Identification of nonlinear actuators with time delay and rate saturation

- using meta-heuristic optimization algorithms," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems Control Engineering*, vol. 229, no. 9, pp. 808-817, 2015.
- [4] P. Hippe, *Windup in control: its effects and their prevention*. Springer Science & Business Media, 2006.
- [5] S. Tarbouriech, et al., *Stability and stabilization of linear systems with saturating actuators*. Springer Science & Business Media, 2011.
- [6] S. Galeani et al., "A magnitude and rate saturation model and its use in the solution of a static anti-windup problem," *Systems & Control Letters*, vol. 57, no. 1, pp. 1-9, 2008.
- [7] K. Kefferpütz, B. Fischer, and J. Adamy, "A nonlinear controller for input amplitude and rate constrained linear systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 10, pp. 2693-2697, 2013.
- [8] N Wada, and M. Saeki, "Anti-windup synthesis for a model predictive control system," *IEEJ Trans Elec Electron Eng.*, vol. 11, no. 6, pp. 776-785, 2016.
- [9] D. Li, N. Hovakimyan, and C. Cao, "Positive invariant set estimation of adaptive controller in the presence of input saturation," *International Journal of Adaptive Control Signal Processing*, vol. 27, no. 11, pp. 1012-1030, 2013.
- [10] Q. Hu, X. Shao, Y. Zhang, L. Guo, "Nussbaum-type function-based attitude control of spacecraft with actuator saturation," *Int J. Robust Nonlinear Control*, vol. 28, no. 8, pp. 2927-2949, 2018.
- [11] A. Gelb, and W. E. Van der Velde, *Multiple-input describing functions and non-linear system design*. 1968.
- [12] C. Y. Tzeng, and K. F. Lin, "Adaptive ship steering autopilot design with saturating and slew rate limiting actuator," *International Journal of Adaptive Control Signal Processing*, vol. 14, no. 4, pp. 411-426, 2000.
- [13] J. Zhang, and T. Raïssi, "Saturation control of switched nonlinear systems," *Nonlinear Analysis: Hybrid Systems*, vol. 32, pp. 320-336, 2019.
- [14] Z. Peng, and W. Guo, "Saturation characteristics for stability of hydro-turbine governing system with surge tank," *Renewable Energy*, vol. 131, pp. 318-332, 2019.
- [15] P. Sarhadi, A. R. Noei, and A. Khosravi, "Model reference adaptive PID control with anti-windup compensator for an autonomous underwater vehicle." *Robotics and Autonomous Systems*, vol. 83, pp. 87-93, 2016.
- [16] M. C. Turner, "Positive μ modification as an anti-windup mechanism," *Systems & Control Letters*, vol. 102, pp. 15-21, 2017.
- [17] P. A. Ioannou, and J. Sun, *Robust adaptive control*. Courier Corporation, 2012.
- [18] C. Rohrs, L. Valavani, M. Athans, G. Stein, "Robustness of continuous-time adaptive control algorithms in the presence of unmodeled dynamics," *IEEE Transactions on Automatic Control*, vol. 30, no. 9, pp. 881-889, 1985.
- [19] B. Altin, and K. Barton, "Rohrs' example revisited: On the robustness of adaptive iterative learning control," *Asian Journal of Control*, vol. 20, no. 3, pp. 993-1002, 2018.
- [20] H. S. Hussain, M. Matsutani, A. M. Annaswamy, and E. Lavretsky, "Robust adaptive control in the presence of unmodeled dynamics: A counter to Rohrs's counterexample," in *AIAA Guidance, Navigation, and Control (GNC) Conference*, American Institute of Aeronautics and Astronautics, 2013.
- [21] M. Matsutani, A. Annaswamy, and E. Lavretsky, "Guaranteed delay margins for adaptive systems with state variables accessible," in *2013 American Control Conference*, IEEE, 2013.
- [22] D. Atherton, and S. Spurgeon, *Nonlinear Control Systems, Analytical Methods*, Wiley Encyclopedia of Electrical and Electronics Engineering, 1999.
- [23] K. S. Narendra, and A. M. Annaswamy, *Stable adaptive systems*, Courier Corporation, 2012.
- [24] J. -J. E. Slotine, and W. Li, *Applied nonlinear control*. Vol. 199, Prentice Hall Englewood Cliffs, NJ, 1991.
- [25] R. Sridhar, "A general method for deriving the describing functions for a certain class of nonlinearities," *IRE Transactions on Automatic Control*, vol. 5, no. 2, pp. 135-141, 1960.
- [26] Vander Velde, W.E., *Multiple-input describing functions and nonlinear system design*. 1968: McGraw-Hill, New York.
- [27] H. S. Hussain, C. S. Subedi, A. M. Annaswamy, and E. Lavretsky, "Robustness of adaptive control systems to unmodeled dynamics: A describing function viewpoint," in *AIAA Guidance, Navigation, and Control Conference*, American Institute of Aeronautics and Astronautics, 2017.
- [28] L. T. Aguilar, I. Boiko, L. Fridman, and R. Iriarte, *Self-oscillations in dynamic systems*, Springer, 2015.
- [29] J. E. Gibson, *Nonlinear automatic control*, McGraw-Hill Book Company, New York, 1963.
- [30] C. Fielding, and P. Flux, "Non-linearities in flight control systems," *The Aeronautical Journal*, vol. 107, no. 1077, pp. 673-686, 2003.
- [31] E. Lavretsky, and K. Wise, *Robust and Adaptive Control: With Aerospace Applications*. Springer, 2012.

Copyrights

© 2023 by the author(s). Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –NonCommercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Investigating the Effect of Geometric Design Parameters on the Mutual Inductance Between Two Similar Planar Spiral Coils With Inner and Outer Diameter Limits

Ata Ollah Mirzaei ¹, Amir Musa Abazari ^{1,*} , and Hadi Tavakkoli ²

¹ Department of Mechanical Engineering, Faculty of Engineering, Urmia University, Urmia 5756151818, Iran

² Department of Mechanical and Aerospace Engineering, Hong Kong University of Science and Technology, Hong Kong

* Corresponding Author: am.abazari@urmia.ac.ir

Abstract: Nowadays, planar spiral coils are widely used in different applications. Mutual inductance of two adjacent coils, is one of the critical operating principles in near-field wireless power and data transmission systems, significantly impacting their performance. Hence, in this study, the mutual inductance between two similar concentric planar spiral coils is investigated. The effect of main parameters, including the track width, w , and the space between two consecutive turns, s , with a fixed inner and outer diameter of the coils are investigated. The Taguchi method using the L16 array in Minitab environment is used to optimize design parameters. The samples of applied Taguchi, are modeled and simulated via ANSYS Maxwell. The results show that the mutual inductance increases by reducing the two investigated parameters. Based on the Taguchi analysis, it is revealed that the effect of the response for both of the investigated parameters is very close. By applying the main effect analysis the obtained results are verified. This interesting result is important in the design of planar spiral coils while we have fabrication limitations in a real sensor design realization.

Keywords: Mutual inductance, planar coil, 3D modeling.

Article history

Received 26 March 2022; Revised 06 November 2022; Accepted 06 December 2022; Published online 29 March 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

A. O. Mirzaei, A. M. Abazari, and H. Tavakkoli, "Investigating the effect of geometric design parameters on the mutual inductance between two similar planar spiral coils with inner and outer diameter limits," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 70-74, 2023. DOI: [10.22055/jaree.2022.40372.1053](https://doi.org/10.22055/jaree.2022.40372.1053)



1. INTRODUCTION

Planar coils are widely utilized in diverse applications, from blood pressure measurement to fabrication of electronic cards. These coils are considered as a proper wireless connection choice, especially in a limited space [1]. Calculation of inductance and resistance for planar coils is critical because of their vital and direct role in the performance of coils. They are called planar or flat coils due to placement of the coil components almost on a plane. This is a huge advantage comparing with solenoid coils; because they occupy less space than solenoid coils; so, they are suitable for applications with size constraints such as Micro-electromechanical systems (MEMS) [2] or implantable medical components (e.g., heart pumps). Planar coils can be built on a rigid or non-rigid substrate; it means that they can be integrated on the Printed Circuit Boards (PCBs) as well as flexible substrates. Planar coils can be produced in the batch production systems, which leads to a cost-effective manufacturing process. According to these features, planar

coils have different applications such as remote health monitoring, wireless power transmission, induction heating, and radiofrequency detection [3]. Mutual inductance of two adjacent coils, is one of the critical operating principles in near field wireless power and data transmission systems, significantly impacting their performance. Hence, in this study, the mutual inductance between two similar concentric planar spiral coils are investigated. In particular, two different analytical methods exist to compute the mutual inductance between two coils, including loop inductance procedure and partial inductance procedure. While the loop inductance procedure calculates the inductance considering the whole structure, the partial inductance procedure deals with each part of the coil. One of these methods can be employed as an optimal method to solve the problem based on the geometric features of the structure. For example, in structures with unparalleled planar coils, the loop method makes the situation complicated. Grover [4] presented a model to calculate the mutual inductance between two wires located in the desired position in the space; this method calculates the mutual

inductance between the studied two wires well. Based on this model, Cheng and Shu [5] provided a relationship for computing the mutual inductance between two square-shaped coaxial loops using the partial inductance procedure; although the method can calculate the mutual inductance between the square-shaped loops, but it is not suitable for non-square and polygon loops. Greenhouse [6], also tried to use the partial method to calculate the inductance of a square-shaped planar spiral coil. Abbaspour et al. [7] used the partial inductance to obtain the mutual inductance between two square-shaped coils, which was limited by the shape and location of the coils. Tavakkoli et al. [8] suggested a novel model based on the partial inductance procedure that calculates the mutual inductance between two parallel coaxial planar spiral coils with arbitrary number of sides; in their study, they have compared the results of their method with practical experiments and computer simulations and show that their method works well in comparison to the others' result. However, this method requires two coaxial and paralleled coils, limiting its application. Ji et al. [9] used relationship to calculate the mutual inductance between the circular and square coils using the circular matrix. Inferring from Newman's relationship, they proposed a method to calculate the mutual inductance between two coils in the arbitrary states. This method can be considered as an essential step in the field due to removing spatial constraints from the computation. However, it can calculate the mutual inductance between two coils in the arbitrary states but it is geometrically limited to circular and square coils. Due to the importance of the planar spiral coils, there are a lot of studies around this subject. The design parameters of the planar spiral coils and their optimization are important and at the same time, interesting topic in this field. The design parameters directly affect the different properties of the coils. For example, Chen et al. [10] investigated the effect of design parameters on the resonant frequency in the double layer printed spiral coils to transfer the wireless power. In the previous studies the direct impact of geometrical parameters, such as the track thickness, track width or space between two consecutive turns of the coil, on the mutual inductance is not studied. Hence, in this paper, we investigate the effect of mentioned parameters on the mutual inductance between two similar planar spiral coils with internal and external diameter limits. Considering the whole permutations for all of design parameters, can be time-consuming and sometimes it is impossible. Therefore, in order to investigate the effect of geometric parameters on mutual inductance, we used the Taguchi design method which reduces the number of the cases and can achieve a desired result in an optimal state. The Taguchi method optimizes design parameters to minimize variation before optimizing design to hit mean target values for output parameters [11]. The extracted cases of Taguchi are modelled and simulated via ANSYS Maxwell to get the simulation results and then, the simulation results are analyzed using the Taguchi analysis method. The main effect analysis is used to verify the final results.

2. MATERIALS AND METHODS

2.1. Design of Simulations

As it stated, Taguchi method is a powerful tool for the design of high quality systems. It is a simple, efficient and systematic approach to optimize designs for performance,

quality, and cost. The methodology is valuable when the design parameters are qualitative and discrete. Parameter design via Taguchi can optimize the performance characteristics through the setting of design parameters and reduce the sensitivity of the system performance to sources of variation [11]. Fewer experiments/simulations means less time and cost. Taguchi provides an orthogonal array of variables and levels for experiments/simulations. Taguchi method proposes a minimum number of necessary experiments/simulations to reach a proper conclusion [12]. So, due to the large number of the cases for an all-inclusive analysis of geometric parameters, we tried to use the Taguchi method to design our cases for an optimal simulation process; and here, the impact of the width of the track, w , and space between two consecutive turns of the coil, s , on the mutual inductance between two concentric similar planar circular coils considering a limit for inner diameter, d_i , and also a limit for outer diameter, d_o , is investigated (see Fig. 1).

According to Table 1, two parameters, i.e., w and s are introduced to Minitab software as factors in four levels. The designs are simulated using the ANSYS Maxwell software according the Taguchi L16 orthogonal array which is applied in Minitab software. Table 2 presents the configuration of the designs.

2.2. Finite Element Analysis (FEA)

FEA methods are widely used in scientific and industrial studies. In the field of electromagnetics, also, simulation software are widely developed. ANSYS Maxwell software is one of them which is widely used in the study of various problems in the field of electromagnetics. In particular, recently, the use of this software for studying the mutual inductance in planar inductors attracted a great interest. Due to its capabilities and ease of use, in this article, ANSYS Maxwell software is used to simulate and calculate the mutual inductance between two planar coils.

The samples of Taguchi's result, are modelled and simulated via ANSYS Maxwell. The coils are designed in the circular shape with a rectangular cross-section; d_i , d_o , track thickness, and vertical space between two fixed coils are considered 5, 15, 0.02, and 10 mm, respectively. In all of the models, coils' material are considered to be "copper" and the surrounding environment is assumed to be "air". Fig. 2 shows a sample of two planar spiral coils in ANSYS Maxwell software. It is noted that in the modelling of samples and

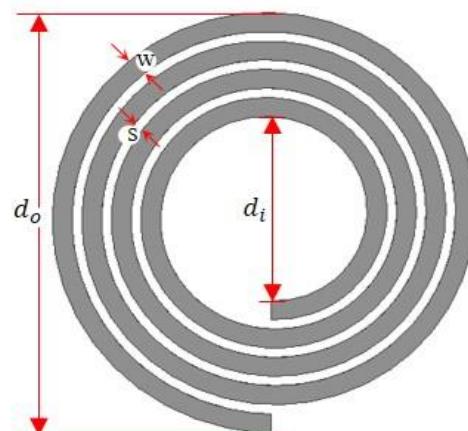
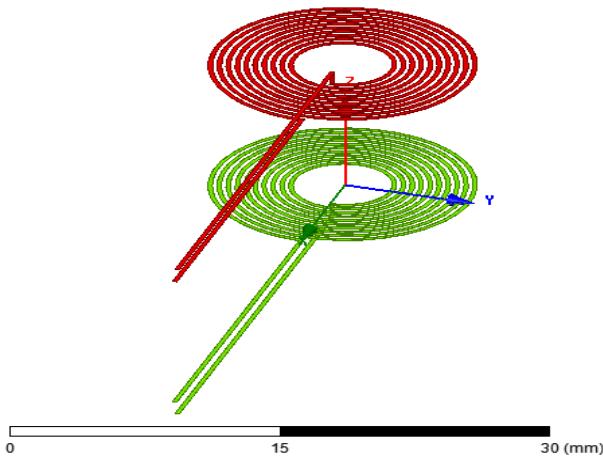


Fig. 1: The used parameters in the modeling.

Table 1: Variables and levels.

| Variables | Levels | | | |
|-----------|--------|---|-----|-----|
| | 1 | 2 | 3 | 4 |
| w [mm] | 1 | 2 | 3 | 3.5 |
| s [mm] | 0.5 | 1 | 1.5 | 2 |

**Fig. 2:** A sample of two concentric planar spiral circular coils in ANSYS Maxwell software.**Table 2:** The configuration of simulations using Taguchi's L16 orthogonal array and the obtained results.

| sample number | w | s | Mutual Inductance [nH] |
|---------------|---|---|------------------------|
| 1 | 1 | 1 | 73.87 |
| 2 | 1 | 2 | 41.46 |
| 3 | 1 | 3 | 18.83 |
| 4 | 1 | 4 | 12.41 |
| 5 | 2 | 1 | 41.52 |
| 6 | 2 | 2 | 19.90 |
| 7 | 2 | 3 | 12.89 |
| 8 | 2 | 4 | 8.74 |
| 9 | 3 | 1 | 19.13 |
| 10 | 3 | 2 | 13.21 |
| 11 | 3 | 3 | 8.27 |
| 12 | 3 | 4 | 6.32 |
| 13 | 4 | 1 | 12.61 |
| 14 | 4 | 2 | 8.79 |
| 15 | 4 | 3 | 6.43 |
| 16 | 4 | 4 | 5.49 |

based on (1), the number of coil turns is a function of w and s , while internal and external diameters of planar coils are considered the same in all samples.

$$N = \frac{d_o - d_i}{2(w + s)} \quad (1)$$

where N , d_i , d_o , s and w are the number of turns, inner diameter, outer diameter, space between two consecutive turns, and track width of the coil, respectively.

3. RESULTS

The results for the effect of the track width and spacing between two consecutive turns on the mutual inductance between two concentric planar spiral circular coils by applying a limit on inner and outer diameters is presented

here. The results of the simulation for proposed Taguchi's samples are reported in Table 2 as well as in Fig. 3.

Fig. 3 illustrates the variation of the mutual inductance value between two similar concentric planar spiral circular coils at the vertical distance of 10 mm from each other, the inner diameter of 5 mm, outer diameter of 15 mm, and track thickness of 0.02 mm with respect to the variation of space between two consecutive turns and track width of the coil. According to this graph, one can easily observe that the decrease of studied parameters, i.e. w and s , leads to an increase in mutual inductance value. Here, it should be notified that reducing w and s will increase the mutual inductance, but w and s reduction will increase DC resistance and coupling capacitance, respectively.

Fig. 4 shows the magnetic field density on the zy plane (Fig. 2) for sample 1 at Table 2.

Finally, to investigate and analyze the results of simulations, the mutual inductance values are entered into Minitab Software, and then Taguchi design analysis is performed on the data. The results of this analysis are presented in Table 3 and Fig. 5. Table 3 shows the signal-to-noise ratio for each factor at different levels. This ratio expresses the scattering around the specific value. The more the ratio is, the less the scattering is. So, the impact of the scattering variable will be more important.

In Table 3, delta which shows the difference between the highest and lowest mean response values for each factor, represents the relative impact of each factor on the response. The more the delta for each factor is, the more the impact on the response is. Regarding the delta value, the effective factors can be ranked [13]. According to the results in Table 3 and the delta values, which are 28.31 and 28.54 for w and s , respectively, it can be concluded that the variation impact of both parameters on the mutual inductance between two planar coils is almost the same.

Fig. 5 shows diagram graph which is related to the analysis of the main effect. It represents the impact of variables on the output values. According to section w , at the left side of Fig. 5, one can see that the smaller the track width is, the more the mutual inductance is. Also, the s section, at the right side, indicates that the smaller the space between two consecutive turns is, a higher value for the mutual inductance can be obtained. Therefore, the main effect analysis verifies the results of previous analyses.

4. CONCLUSION

In this article, the impact of track width, w , and spacing between two consecutive turns of the coils, s , on the mutual inductance between two concentric planar spiral circular coils was investigated for a fixed inner and outer diameters. For this purpose Taguchi method was used to design an optimal simulation table. Then, ANSYS Maxwell software was used to calculate the mutual inductance between two coils. The mutual inductance obtained from simulation results, are analyzed by Taguchi concepts and the main effect analysis is also performed to verify the results. The obtained results and performed analyses shows that the reduction of w and s lead to an increase in mutual inductance. Based on the results, it is shown that the effect of both of the investigated parameters,

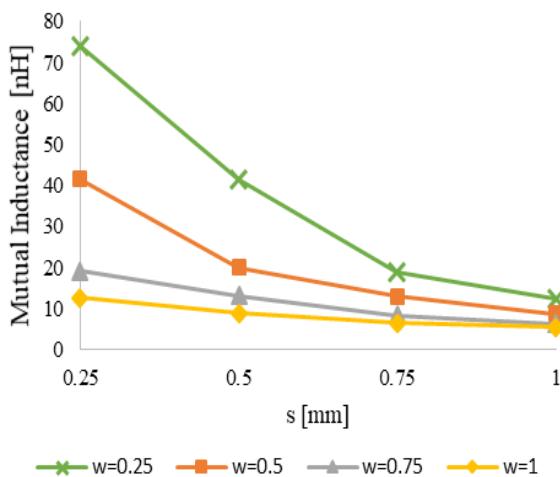


Fig. 3: The variation of mutual inductance value with respect to the variation of the space between two consecutive turns, s , and track width, w , of the coils.

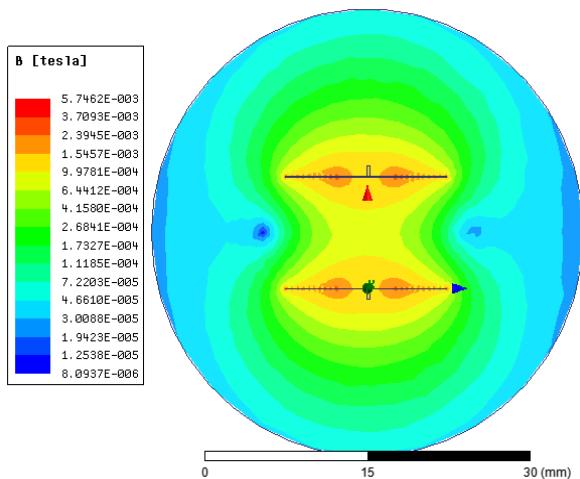


Fig. 4: The magnetic field density on the zy plane (Fig. 2) for sample 1 at Table 2.

Table 3: Signal to noise ratio (larger is better).

| Level | w | s |
|-------|-------|-------|
| 1 | 36.64 | 36.78 |
| 2 | 20.76 | 20.84 |
| 3 | 11.83 | 11.61 |
| 4 | 8.33 | 8.24 |
| Delta | 28.31 | 28.54 |
| Rank | 2 | 1 |

i.e. w and s , is very close to each other. This compelling result is important in design of planar spiral coils where fabrication issues can be a challenge in real applications. Finally, we declare that although, reducing w and s will increase the mutual inductance, it will increase DC resistance and coupling capacitance. The important point is that the mutual inductance is not the only effective factor in the systems associated with planar coils; and the coil resistance and coupling capacitance must also be considered in the design of

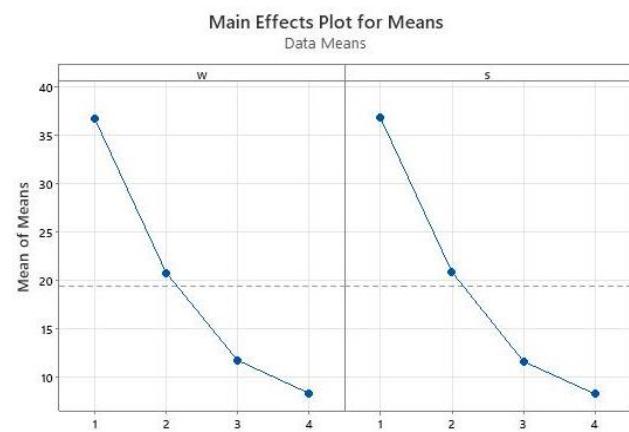


Fig. 5: The result obtained from the main effect analysis.

such a system. So, these side effects must be considered depending on any specific application which can be studied in the future works.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Ata Ollah Mirzaei: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing - original draft. **Amir Musa Abazari:** Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing - original draft, Writing - review & editing. **Hadi Tavakkoli:** Conceptualization, Investigation, Supervision,

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] L. Qian, M. Chen, K. Cui, G. Shi, J. Wang, and Y. Xia, "Modeling of mutual inductance between two misalignment planar coils in wireless power transfer," *IEEE Microwave and Wireless Components Letters*, vol. 30, no. 8, pp. 814-817, 2020.
- [2] C. H. Ahn, and M. G. Allen, "Micromachined planar inductors on silicon wafers for MEMS applications," *IEEE Transactions on Industrial Electronics*, vol. 45, no. 6, pp. 866-876, 1998.
- [3] G. M. Moreton, "The design and development of a planar coil sensor for angular displacements," Ph.D. Thesis, Cardiff University, 2018.
- [4] F. W. Grover, *Inductance calculations: working formulas and tables*. Courier Corporation, 2004.
- [5] Y. Cheng, and Y. Shu, "A new analytical calculation of the mutual inductance of the coaxial spiral rectangular coils," *IEEE Transactions on Magnetics*, vol. 50, no. 4, pp. 1-6, 2013.

- [6] H. Greenhouse, "Design of planar rectangular microelectronic inductors," *IEEE Transactions on parts, hybrids, and packaging*, vol. 10, no. 2, pp. 101-109, 1974.
- [7] E. Abbaspour-Sani, R. -S. Huang, and C. Y. Kwok, "A linear electromagnetic accelerometer," *Sensors and Actuators A: Physical*, vol. 44, no. 2, pp. 103-109, 1994.
- [8] H. Tavakkoli, E. Abbaspour-Sani, A. Khalilzadegan, A.-M. Abazari, and G. Rezazadeh, "Mutual inductance calculation between two coaxial planar spiral coils with an arbitrary number of sides," *Microelectronics Journal*, vol. 85, pp. 98-108, 2019.
- [9] Y. Ji, H. Wang, J. Lin, S. Guan, X. Feng, and S. Li, "The mutual inductance calculation between circular and quadrilateral coils at arbitrary attitudes using a rotation matrix for airborne transient electromagnetic systems," *Journal of Applied Geophysics*, vol. 111, pp. 211-219, 2014.
- [10] K. Chen, and Z. Zhao, "Analysis of the double-layer printed spiral coil for wireless power transfer," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 1, no. 2, pp. 114-121, 2013.
- [11] W. P. Yang, and Y. Tarng, "Design optimization of cutting parameters for turning operations based on the Taguchi method," *Journal of Materials Processing Technology*, vol. 84, no. 1-3, pp. 122-129, 1998.
- [12] R. K. Roy, *A primer on the Taguchi method*. Society of Manufacturing Engineers, 2010.
- [13] Minitab 18 Support. (2019). Interpret the key results for Analyze Taguchi Design. [Online]. Available: <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/doe/how-to/taguchi/analyze-taguchi-design/interpret-the-results/key-results/>

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Partial Discharge Pattern Recognition in GIS Using External UHF Sensor

Reza Rostaminia* , Mehdi Vakilian , and Keyvan Firouzi

*Electrical Engineering Department and Centre of Excellence in Power System Management and Control,
 Sharif University of Technology, Tehran 1458889694, Iran*
 * Corresponding Author: Reza_rostaminia@ee.sharif.edu

Abstract: Partial Discharge (PD) measurement is one of the best solutions for condition assessment of Gas Insulated Switchgears (GISs). For having Condition-based maintenance of GIS, online PD monitoring is of great importance. For this aim, Ultra High Frequency (UHF) PD sensors should be installed inside the GIS during the installation. However, in most installed GISs in industries, the internal UHF PD sensors are not installed. In this paper, a new method for online defect type recognition according to external UHF PD sensors and based on the time-frequency representation of PD signal is proposed. In this case, four artificial defect types named protrusion on the main conductor, protrusion on the enclosure, free moving metal particle, and metal particle on spacer are implanted inside the 132 kV L-Shaped structure of one phase in enclosure GIS. The signal energy at each level of the decomposed signal by Discrete Wavelet Transform (DWT) is applied for features of each defect type. The trends of signal energy variations at each frequency range of signal are applied for discriminating between each defect type. The Deep Feed Forward Network (DFFN) classifier is applied for PD pattern recognition. The results show the benefits and simplicity of the proposed method for PD signal classification, independent from the position of the PD sensor, especially in the case of online PD monitoring of GIS.

Keywords: Gas insulated switchgear (GIS), partial discharge (PD), ultrahigh frequency (UHF) measurements, pattern recognition, time-frequency representation.

Article history

Received 30 March 2022; Revised 31 October 2022; Accepted 06 December 2022; Published online 29 March 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

R. Rostaminia, M. Vakilian, and K. Firouzi, "Partial discharge pattern recognition in GIS using external UHF sensor," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 75-86, 2023. DOI: [10.22055/jaree.2022.40395.1054](https://doi.org/10.22055/jaree.2022.40395.1054)



1. INTRODUCTION

Partial Discharge (PD) monitoring is one of the most useful method for detecting insulation defects in the early stages of development in Gas Insulated Switchgears (GIS). PD occurrence can be monitored via different methods such as conventional electrical IEC 60270 method [1]-[2], gas analysis method [3], ultrasonic method [4], optical detection method [5] and electromagnetic method in GIS [6]. The conventional electrical IEC 60270 technique is one the most popular method for PD detection in GIS [7]. In this method, the PD apparent charge can be measured via a coupling device. The main advantage of this method is quantifying the discharge level in pico-colomb range [8]. However, the main disadvantage of using the IEC 60270 method is needing to use coupling capacitor device which imposed some restrictions, especially in at-site testing [9]. The ratio of coupling capacitor value to the test object equivalent

capacitance has great impact on measurement sensitivity. This is in great importance especially in case of at-site GIS testing with long length, since applying large value of coupling capacitor can be difficult. Furthermore, the other disadvantage of this method is its sensitivity to internal and external interferences [10].

The acoustic detection method can be employed for measuring the acoustic waves generated by PD within the GIS. The advantage of this method is its well quantification of the PD level and the risk assessment during online operation [11]. However, the main disadvantage of this method is the high uncertainty in the obtained results due to existence of noise and electric interferences [12]. The optical spectrum of measured signal via optical sensors installed in a GIS system varies due to presence of contaminants or surface roughness. This is another way for partial discharge detection [13]. However, the application of optical detection method is restricted due to limitation in accessing the GIS interior

design information. Also, some components of GIS such as solid spacers and cable or cable sealing ends can be unchecked during PD measurement via this method. Another method to detect occurrence of PD within the GIS is by analysis of by-products of SF₆ decomposition [14]. With quantifying the decomposed components of SF₆, the level of insulation degradation and the type of defects can be reached [15]. One of the dis advantage of this method is that it should be performed off-line and the generated by-products due to decomposition of SF₆ is time-consuming process. Also, some defect types, such as: occurrence of voids within the epoxy insulating material of GIS spacer or another insulator can be undetectable with this method.

Ultra High Frequency (UHF) sensors can be applied to measure the radiated electromagnetic waves due to PD occurrence within the GIS. The internal and external UHF sensors can be exploited for measuring PD in GIS [16-17].

Unlike the preferences of internal sensors compared to the external ones due to better sensitivity and noise immunity features, however in most GIS in service, these internal sensors have not been installed. Accordingly, the external sensors can be applied for PD monitoring. One of the main challenges in UHF PD monitoring systems of GIS is distinguishing the type of defects. The recorded PD wave shape can be changed with defect type within the GIS [18]. However, some parameters such as the distance between the measuring sensor and the defect [18], the propagation path due to configuration of GIS such as L-Shaped and T-Shaped structure, spacer thickness and disconnectors [19-21], can influence on PD measured wave shape. Recorded PD data can be represented via Phase-Resolved Partial Discharge (PRPD) or Time-Resolved pattern. In PRPD pattern, usually the PD magnitude in (mV), phase angle of applied voltage at the time of PD occurrence and the number of PD occurred are represented [22]. In time-resolved patterns of PD wave shape of PD signal is recorded and the type of defects can be distinguished due to the fact that each defect can result in specified PD wave shape. Although the efficiency of PRPD pattern analysis in recognition of defect type is proven but, one of its main challenge is occurring two or more defects simultaneously within the GIS. In this case, the segregation of patterns can be difficult [23]. In some works, the time-domain pattern is applied for PD defect recognition [24]. In [25], the envelope detection circuit (based on a mathematical method for waveform estimation) is applied to extract the time-domain signal. Then, the noise reduction technique based on wavelet is used to extract the de-noised PD wave shape. The pattern classification is performed using the “back propagation neural network” (BPNN). In [26], the Damper-Shafer (DS) theory is established for feature extraction from both PRPD and time-domain pattern of PD within 126 kV GIS. In [27], the Convolutional Neural Network (CNN) is used to classify the extracted features from time-frequency

representation of the signal. In [28], the gray scale image for time-frequency representation of PD signal using S-Transform is applied to extract features from five regions of Transverse Electromagnetic (TE) mode. The extracted features based on low order moments and J-criterion are feed to three classification methods, named Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and particle swarm optimized Extreme Learning Machine (ELM). Finally, it is shown that the ELM has a better performance not only in respect to the classification accuracy, but also in the learning process and the test speeds.

However, in most of the recent works for pattern recognition of PDs in GIS, the employed data are related to the recorded PD data from the UHF sensors installed inside the GIS enclosure. Also, in some works, the proposed methods are complex and in practice, their implementation is difficult. Since the internal UHF PD sensors have not been installed in most of GIS under operations in the electric power industries around the world, the only way for UHF PD online monitoring is the application of the external UHF sensor. Thus, the authors propose a method for pattern recognition of recorded PD data captured by an external UHF PD sensor installed on a sample GIS busbar. Therefore, the L-shape arrangement of GIS sample busbar with artificial embedded defects are applied for the PD measurements. Then, the time-frequency representation of PD signal from wavelet transform is used to identify the features that can discriminate among the various defects. It is shown the energy of the signals in various frequency ranges are different for different type of defects. Finally, the DFNN classifier is applied for classification of the detected PD defect types.

2. EXPERIMENTAL SETUP

The L-shape structure of gas insulated busbar is selected for implanting four artificial defects and it is shown in [Fig. 1](#). This L-shape arrangement of GIS model is constructed with one phase conductor in enclosure and the rated voltage for this model is 132 kV. The absolute SF₆ gas pressure in this L-shape model is 4 bar. This L-shape structure is connected to oil-insulated busbar which is terminated with an air-insulated bushing to apply high voltage as shown in [Fig. 1](#).

Two inspection windows are considered at the end of the straight part and the L-shape part of this model. These windows are designed in such a way that a UHF PD sensors can be placed in them for capturing the PD data. The UHF PD sensors installation location is shown in [Fig. 2](#).

Two inspection windows are considered at the end of the straight part and the L-shape part of this model. These windows are designed in such a way that a UHF PD sensors can be placed in them for capturing the PD data. The UHF PD sensors installation location is shown in [Fig. 2](#).

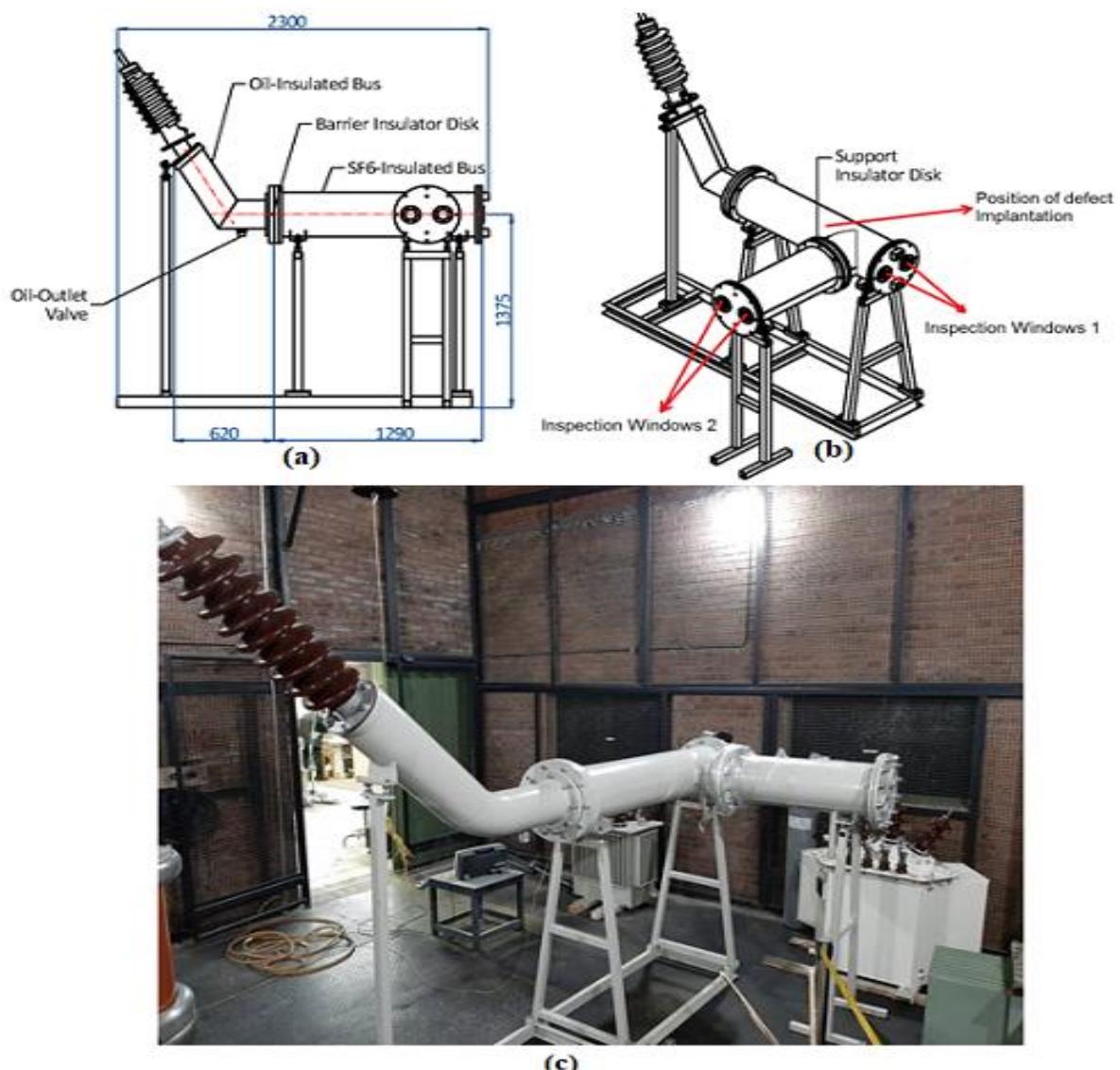


Fig. 1: The L-shape arrangements of GIS experimental busbar, (a) side view (b) 3D view (c) installed in HV laboratory of Electrical Engineering Department, Sharif University of Technology.

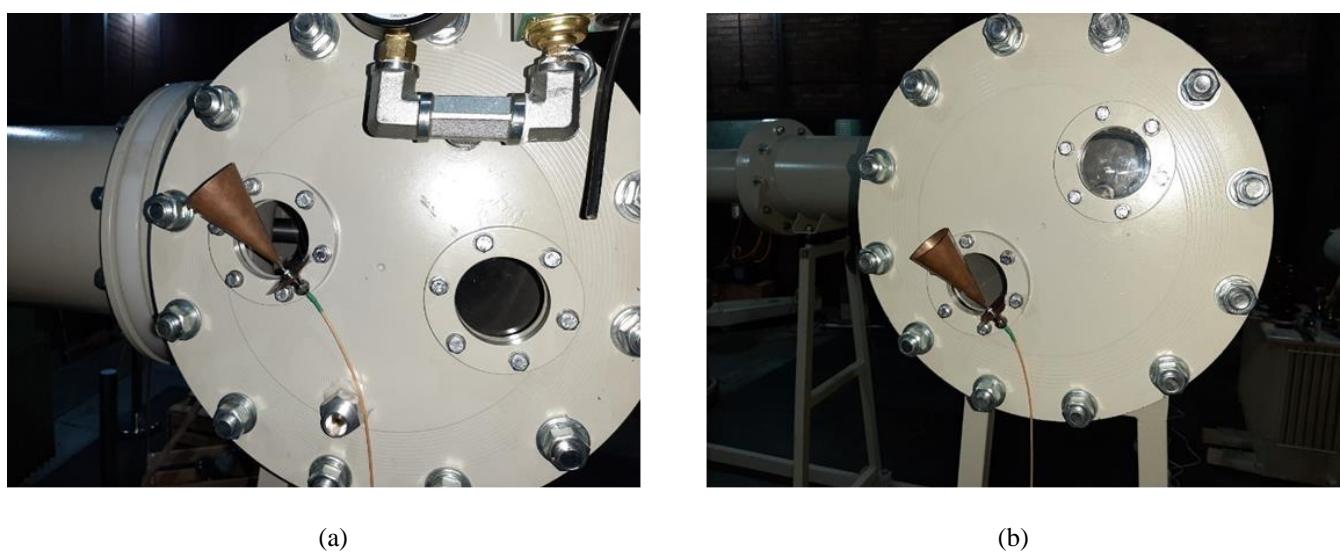


Fig. 2: External UHF PD Sensor installation on observing window of L-Shape GIS model, (a) UHF PD sensor 1, (b) UHF PD sensor 2.

Two types of UHF PD sensors, called externally and internally connected UHF sensors, can be applied for PD measurements in GIS. Although the internally connected PD sensors are more sensitive than the externally connected ones, but in most energized GIS in industries, the internally PD sensors have not been installed inside the GIS. Accordingly, the condition monitoring and PD measurements can just only be performed with externally connected PD sensors. Various types of externally UHF antenna such as bio-conical log periodic antenna, loop antenna, horn antenna, dipole antenna and planar antenna can be applied for PD measurements [29]-[33]. The external PD sensors can be applied both at the bush type spacers (spacers with no metallic coverage) and at the inspecting windows of disconnect and earthing switches. All the above mentioned antennas have features and quality that can be chosen for reaching the specified goal. As an example, horn antenna have the most sensitivity at higher frequencies which the dipole and log-periodic antenna have better frequency response at lower range of frequencies. However, there are some restrictions for using these antennas due to the position where they should be applied. In this case, the dipole antenna can have the best performance since be directly connected to the flanges.

In this study, the aim is discriminating the difference between various types of defects due to time-domain measured PD signals, the horn antenna is applied for measuring UHF PD signals. The frequency response of the horn antenna UHF PD sensor is presented in Fig. 3.

Four typical defects named: 1- metal protrusions on main HV conductor, 2- metal protrusions on grounded enclosure, 3- free moving metal particle and 4- A metal particle on epoxy resin insulating material of GIS spacer are implemented in to this GIS busbar system as artificial defects (Fig. 4). The

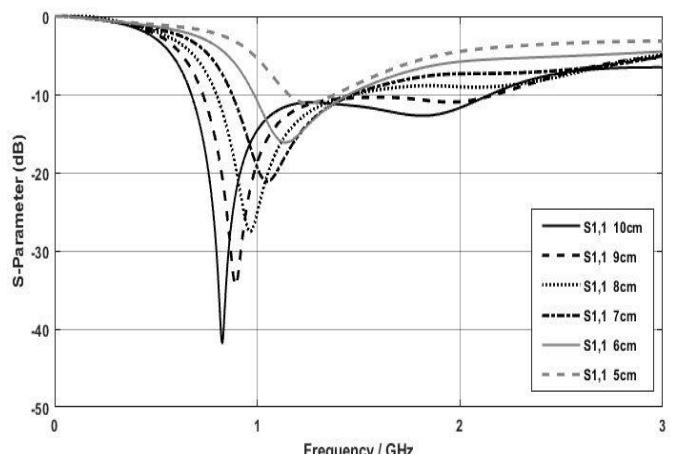


Fig. 3: Frequency Response of horn antenna.

defects number 1 and 2 are created via a 2 cm sharp point needle, soldering it to the main HV conductor and enclosure of the model, respectively (Fig. 4a, and b). While, defect No. 3 is implanted in this experimental GIS system using some little free moving metal particles; each of them 1 mm in length (Fig. 4c). A 2 cm length of metal wire is located in a fixed position on the surface of epoxy insulating material of the GIS system spacer; between the core HV conductor and the grounded enclosure, to represent the defect No. 4 (as shown in Fig. 4d).

The main circuit and the experimental test setup for measurement of electromagnetic radiation due to the PD occurrence are depicted in Fig. 5. The positions of implanted the defects are in the straight section of the GIS busbar model; as shown in Fig. 1.

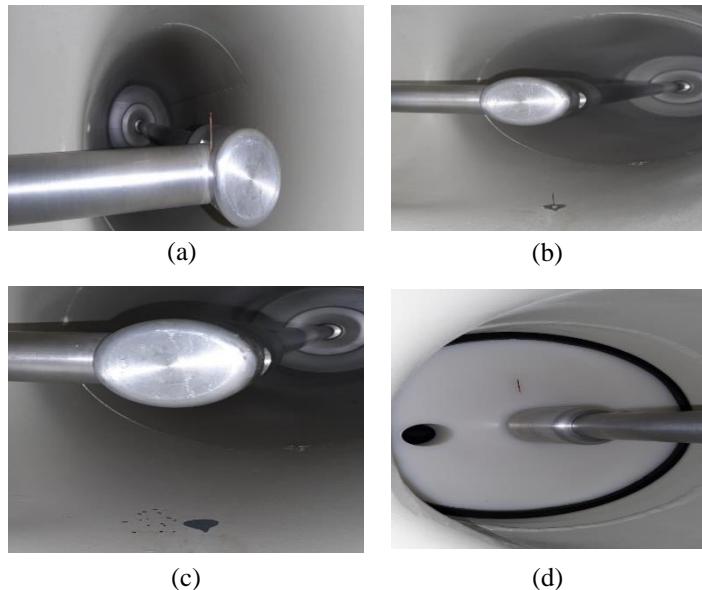


Fig. 4: Four artificial defect types, as applied to the experimental L-shaped GIS busbar, (a) protrusion on the main HV conductor, (b) protrusion on the enclosure of GIS, (c) free moving metal particle, (d) metal on epoxy resin of GIS spacer.

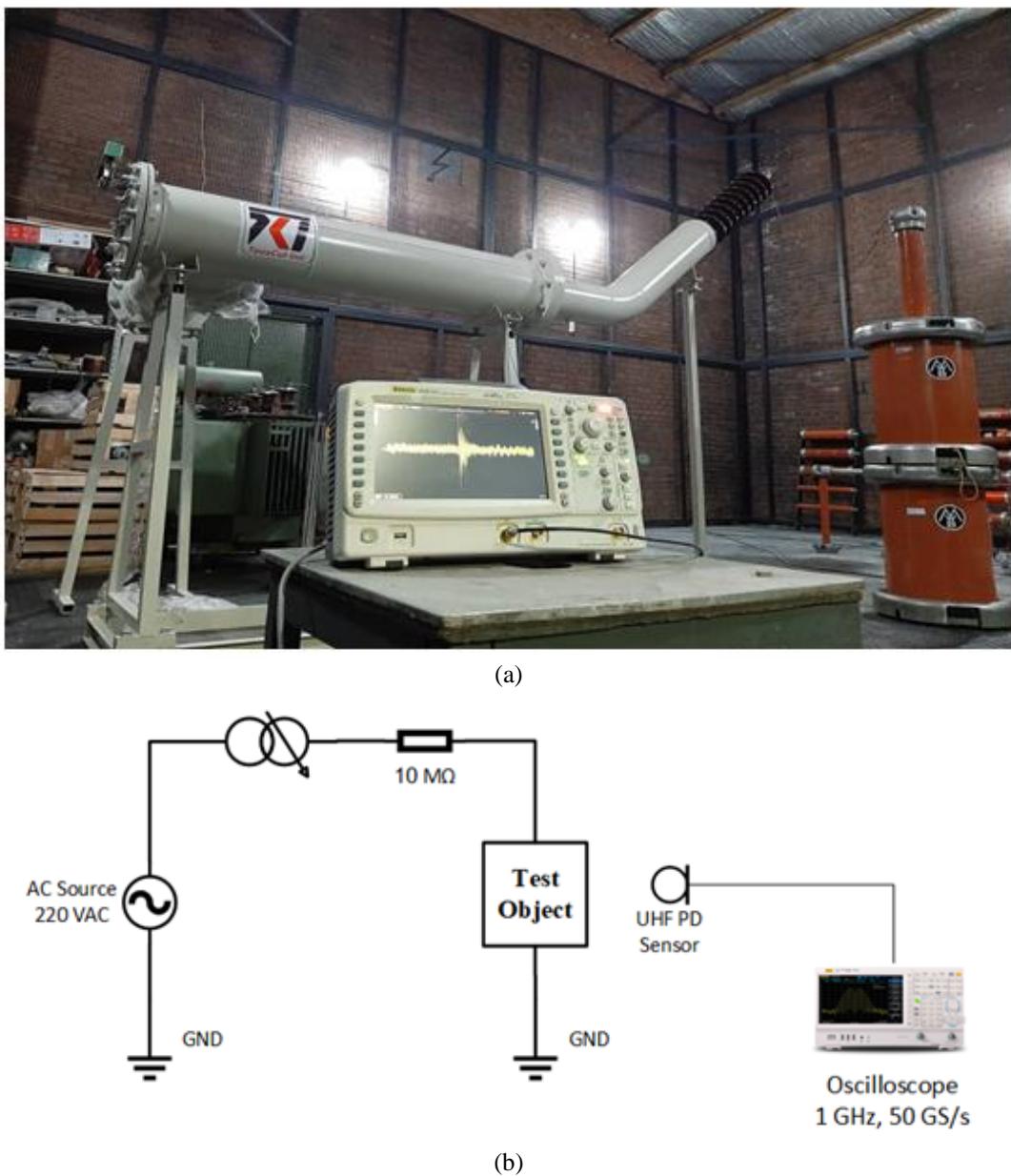


Fig. 5: (a) The test setup in the HV laboratory, and (b) the PD test circuit using UHF sensor.

3. DATA ACQUISITION

The recorded time-domain PD pulse for four defect types implanted in the L-shape GIS experimental model, using UHF sensor No. 1 are represented in Fig. 5. The magnitude of PD pulses are per-united based on the maximum measured value of PD for each defect by sensor No. 1. Each PD signal is recorded over 1100 time samples and each sample has a length of 0.02 ns.

As it can be seen the PD wave shape related to each defect have some differences compared to each other. It needs to be considered that these recorded data are based on the positioning of UHF PD sensors No. 1, which is just directly in line with defects implantation. In Fig. 6, the recorded time-domain PD wave shape are related to the four defect types, through the UHF PD sensor No. 2 measurement.

As it can be seen, there are significant differences between recorded PD signal waveform from sensor No. 1 and

sensor No. 2. These differences are strongly related to the position of each measuring UHF PD sensor. Based on their positions, the radiations and reflections of electromagnetic PD waves on their ways to reach the UHF sensors can result in different time-domain PD recorded wave shapes. Since the PD sensor No. 1 is very close to the defect's position, the radiated electromagnetic PD waves are directly reached to this sensor. Therefore, the recorded PD data have less distortions and fluctuations. However, the recorded PD waveform by sensor No.2 contains a lot of fluctuations. The other important fact is that the amplitude of the recorded PD data strongly depends on the relative location of the defect and sensor position. It is obvious, that as the defect position is more close to the UHF PD sensor results in higher amplitude of measured PD. This is very important especially in case of using external PD sensor. Accordingly, the proposed method for defect type recognition is independent from PD magnitude and the amplitude of presented PD wave shape in Fig. 6 and Fig. 7 are per-united.

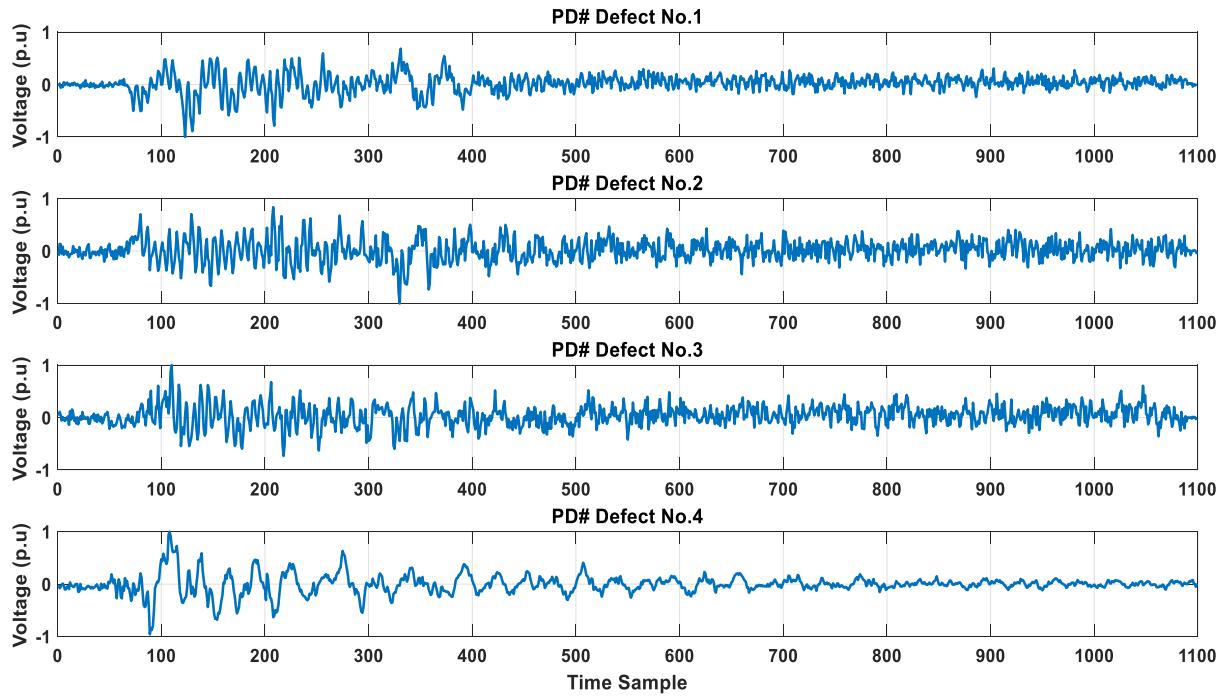


Fig. 6: PD time-domain signal measured by sensor No. 1 for the four different defect models.

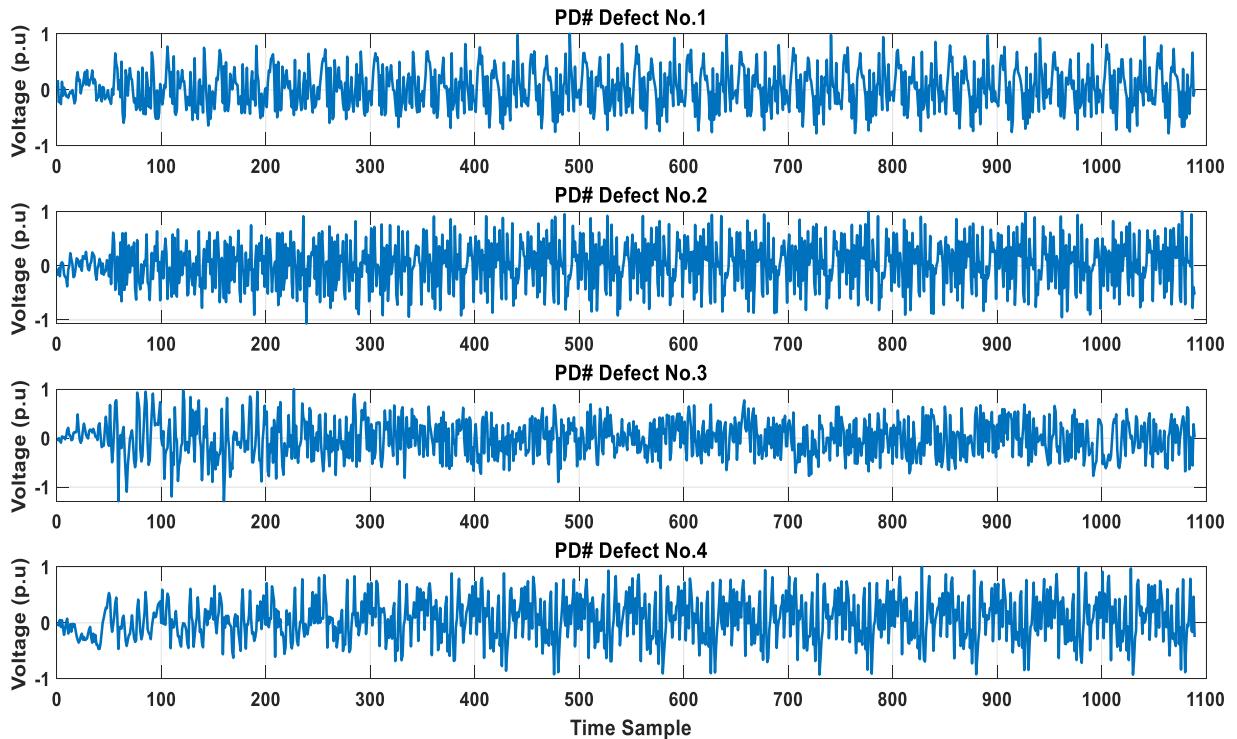


Fig. 7: PD time-domain signal measured by sensor No. 2 for four defect models.

4. FEATURE EXTRACTION AND PATTERN CLASSIFICATION

As it can be seen from **Fig. 5** and **Fig. 6**, the recorded PD data from one specific defect have considerable differences in their wave shape based on positioning of UHF PD sensor. Accordingly, some features related to their wave shape may have not enough capability for discrimination between each defect type. To solve this problem, the time-frequency

representation of PD signal based on Discrete Wavelet Transform (DWT) is proposed in this paper.

4.1. Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a well-known choice for representing time- frequency representation of signals with lots of distortions. With selecting proper mother wavelet, a signal can be decomposed in different levels to some of its different frequency ranges at specific times. Based on decomposition level, the coefficient of similarity between

original signal wave and mother wavelet is represented in time-domain. The decomposition level of original signal wave strongly depends on frequency components of original signal. At each level of decomposition, a low-pass and high-pass filter is applied to the signal. The output of low-pass and high-pass filters are called the signal approximation and signal details, respectively.

The “db-4” is selected as mother wavelet for signal decomposition. The frequency range of recorded PD wave is restricted to 1 GHz due to limitation of frequency ranges of oscilloscope. Accordingly, three decomposition levels are selected and their related frequency ranges for “signal approximation” and “signal details” at each decomposed level of wavelet transform are shown in Fig. 8. At decomposition level 1 of PD signal, the approximation and details of the signal have the components with frequency ranges 0-0.5 GHz and 0.5 to 1 GHz, respectively. Also at decomposition level 2, the coefficients of the signal in approximation and details are in frequency ranges between 0-250 MHz and 250-500 MHz, respectively. Finally, at decomposition level 3, the approximation and details of signal have frequencies 0-125 MHz and 125-250 MHz, respectively.

In Fig. 9 and Fig. 10, the main PD signal wave and the approximation at level 3 and details at level 1, 2 and 3 of discrete wavelet transform for the four defect types of UHF sensor 1 and 2 are shown, respectively.

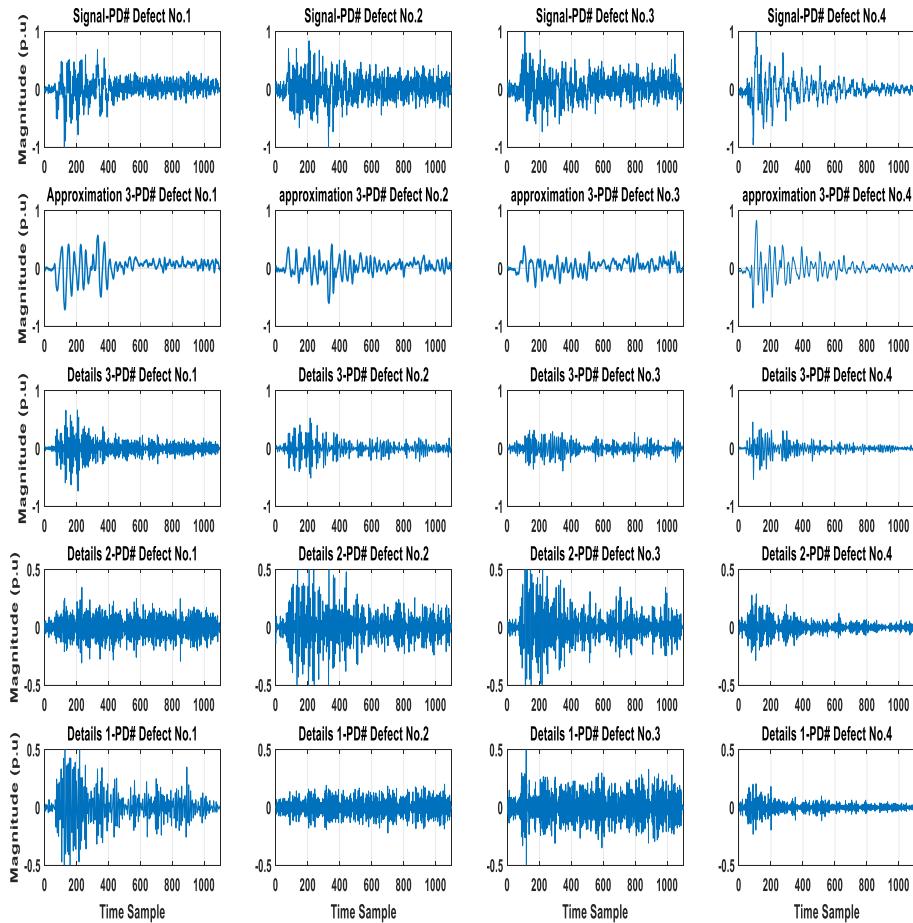


Fig. 9: The Main signal and the approximation and details of DWT for four defect types of UHF sensor 1.

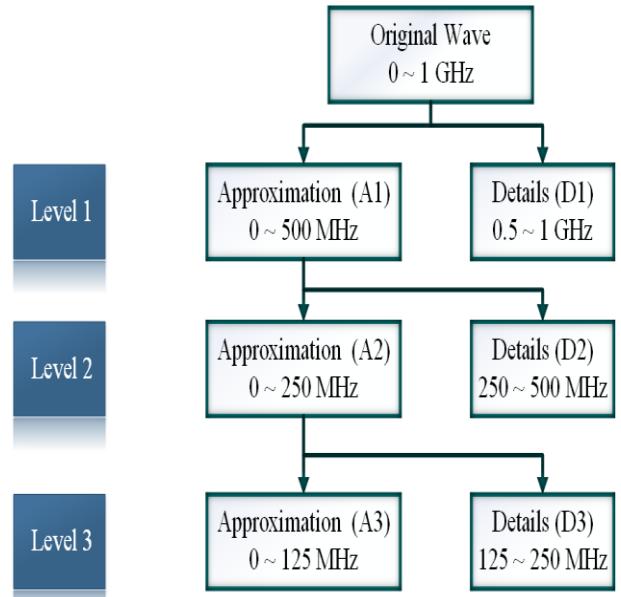


Fig. 8: The Discrete Wavelet Transform decomposition level applied to measured PD signal.

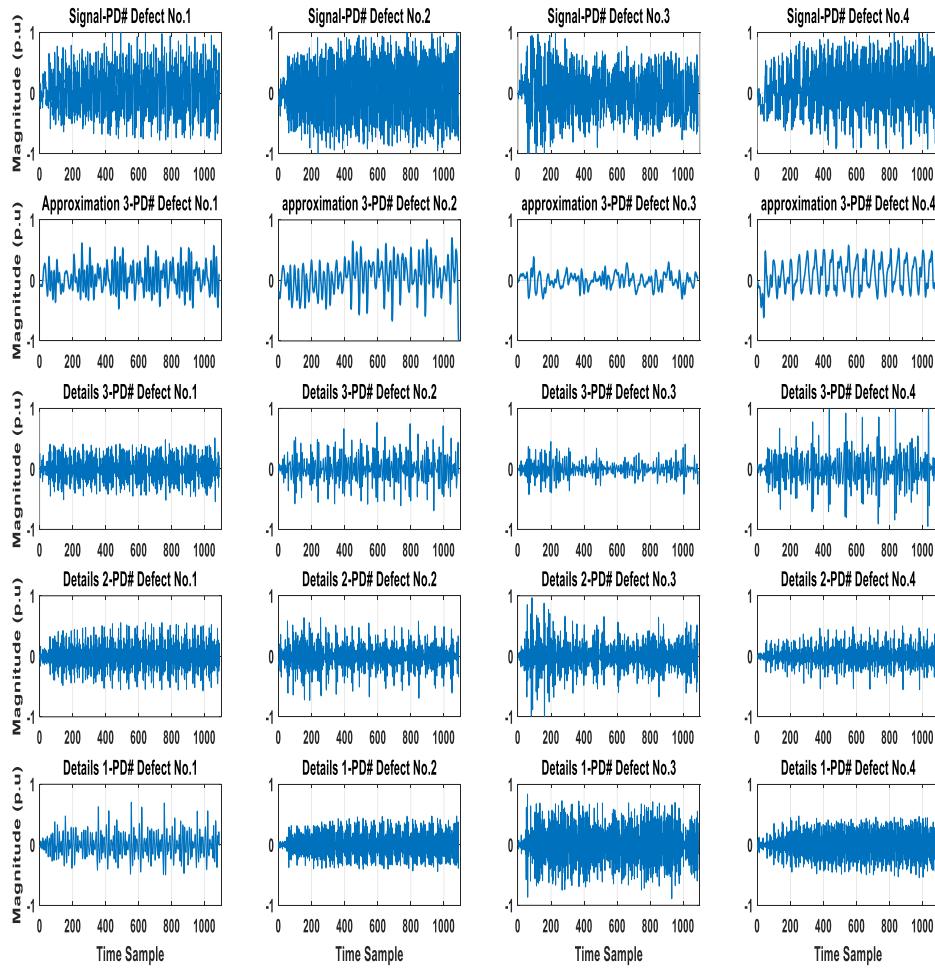


Fig. 10: The Main signal and the approximation and details of DWT for four defect types of UHF sensor 2.

Since the magnitude of PD pulse strongly depends on the distance of PD source and the UHF PD sensor, then the features extracted for pattern recognition should be independent of the PD pulse magnitude. Therefore, the magnitude of each PD pulse is presented in per unit, using its maximum as base. This is the main challenge for defect type identification based on time-domain PD pulse. In some past works time-domain PD data is employed for defect type identification, using pulse shape parameters such as the rise time, fall time, maximum amplitude, average amplitude and other parameters of PD signal [25-26]. However, due to presence of extensive distortions in each level of decomposition and also in main signal, the main wave shape features of PD signal are not proper for pattern recognition [27-33]. This is especially more critical for defects which are far from or are not in line with the PD sensor position [34]. Since, the fluctuations of PD pulse in different frequency ranges are different as it can be seen from Fig. 7 and Fig. 8, each defect type results in a specific radiated electromagnetic wave energies over some different frequency ranges. Accordingly, to solve the abovementioned problem, the signal's energy in each frequency ranges based on (1) is selected as its main feature for each defect type:

$$\text{Signal Energy} = \sqrt{\sum_{i=1}^n |a_i|^2} \quad (1)$$

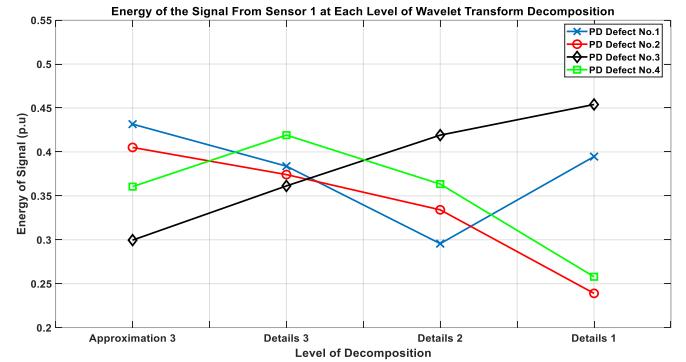


Fig. 11: Signal energy from sensor No. 1 at each level of WT decomposition (explained in Fig. 7).

where a_i is the coefficient of the signal at each level of wavelet transform decomposition and is the number of time sample or recorded PD signal.

The comparison of signal energies in different frequency ranges are depicted in Fig. 10 and Fig. 11 for the captured signals by sensors 1 and 2, respectively. As it can be seen from these figures, the trends of pulse energy variations in each frequency range differs due to the defect type of pulse origin. The pattern of these changes seems to be unique based on defect types and it is independent from positioning of UHF PD source.

In Fig. 10 and Fig. 11, for PD defect No. 1 the signal energy at approximation level 3 is maximum value among all the decomposition levels and the signal energy at details level 2 is the minimum value. However, the signal energy is increased in details 1 from details 2. For defect No. 2, the maximum and minimum amount of signal energies are related to the approximation 3 and details 1, respectively. The signal energy is decreased at each level of decomposition and the main energy of the signal is in frequency ranges between 0-125 MHz. For defect No. 3, the signal energy increased from approximation 3 to details 1. It can be seen the maximum energy of the signal corresponds to details 1 and frequency ranges between 0.5-1 GHz. However, for defect No. 4, the maximum and minimum energy of signal are in details 3 and details 1, respectively. In this type of defect, first the PD signal energy is increased from approximation 3 to details 3 and then the energy is decreased to details 1. The important point in investigation of signal energy at different range of frequency is that these trends of variations are related to the type of defects and they are independent from position of UHF PD sensor. These trends are presented in Table 1.

4.2. Deep Feed Forward Network (DFFN)

Artificial neural network is a well-known method for classification of different PD defect types [35]. In this case, the deep feed forward network is the most popular and simplest one in classification problems. In this paper, the feed forward neural network with three hidden layers are applied for PD defect type classification. As it can be seen in Fig. 12, the signal energy at approximation 3, details 3, details 2 and details 1 are selected as inputs of neural network.

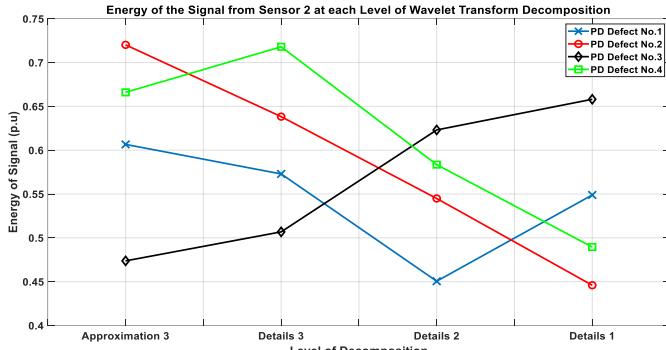


Fig. 12: Signal energy from sensor No. 2 at each level of WT decomposition (signal approximation level 3, and signal details levels 1, 2, and 3 as defined in Fig. 7).

Table 1: The signal energy variation trends at each level of decomposition.

| | Signal Energy Variations | | | | Pattern Code |
|----------------|-----------------------------------|-----------------------------|-----------------------------|---------|--------------|
| | from Approximation 3 to Details 3 | from Details 3 to Details 2 | from Details 2 to Details 1 | | |
| | ↗ | ↘ | ↗ | (0,0,1) | |
| PD Defect No.1 | ↘ | ↘ | ↗ | (0,0,1) | |
| PD Defect No.2 | ↘ | ↘ | ↘ | (0,0,0) | |
| PD Defect No.3 | ↗ | ↗ | ↗ | (1,1,1) | |
| PD Defect No.4 | ↗ | ↘ | ↘ | (1,0,0) | |

At hidden layer 1, the comparison of signal energy between two adjacent levels of decomposition is done. Then, the rectified linear unit function is applied in hidden layer 2 to make the output of the signal between 0 and 1. In hidden layer 3, the outputs of compared energy of signal, based on values 0 and 1 are compared to each other for making pattern code. The created pattern code with the one presented in Table 1 result in type of defects classification. For training the Deep feed-forward network 200 recorded PD waveform for each type of defects are selected. Then, the 300 recorded data are applied to validate the proposed method for partial discharge pattern recognition. In Table 2, the misclassification matrices for the proposed PD pattern recognition method is presented.

As it can be seen in Table 2, the maximum and minimum Identification Percentage (IP) are for data related to defect No. 3 and defect No. 2, respectively. However, the average IP for the proposed PD pattern recognition is 94.5% and this means that the proposed methods especially for on-line PD monitoring of GIS is acceptable. It should be noted in the proposed method, there is no need to employ an internal UHF PD sensor be installed inside the GIS and also it is independent from the location of PD sensors.

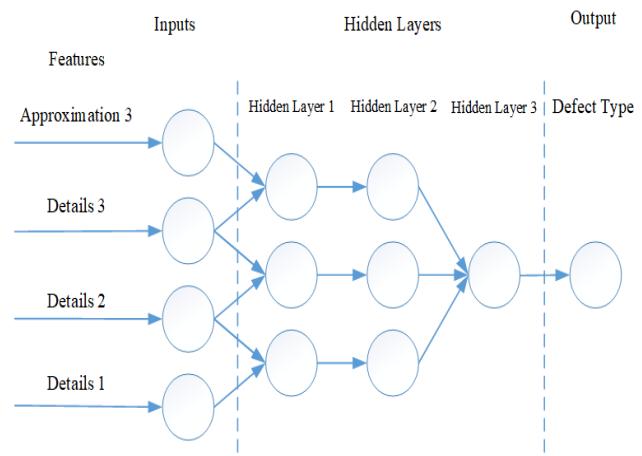


Fig. 13: The deep feed forward network classifier for PD pattern recognition.

Table 2: The misclassification matrices for the proposed PD pattern recognition method in GIS.

| | | Target Class | | | |
|---|----------|--------------|----------|----------|----------|
| | | Defect 1 | Defect 2 | Defect 3 | Defect 4 |
| Output Class | Defect 1 | 282 | 15 | 3 | 6 |
| | Defect 2 | 15 | 276 | 1 | 10 |
| | Defect 3 | 1 | 0 | 291 | 2 |
| | Defect 4 | 2 | 9 | 5 | 282 |
| Correct Identification Percentage (IP%) | | 94 % | 92 % | 98 % | 94 % |
| Total Accuracy | | 94.5 % | | | |

5. SUMMARY AND CONCLUSION

Online partial discharge monitoring is one of the most important method to assess the condition of GIs. However, in most installed Gas insulated switchgears, the internal UHF PD sensors are not installed inside the GIS. In this paper, online partial discharge pattern recognition method is presented based on measured PD data from external UHF PD sensor. The time-frequency representation of signal from discrete wavelet transform is applied for feature extractions of each PD defect model. Four artificial defect models are implanted inside the 132 kV L-Shape GIS model. The feature extracted based on signal energy at each level of DWT decomposition are independent from positioning of UHF PD sensors. This is important when the UHF PD sensors are not installed inside the GIS busbar. Then, there would be just some few locations on GIS busbar which the external UHF PD sensors can be implanted. The trends in partial discharge signal energy variations in each level of DWT decomposition, present a significant potential for pattern recognition. By using the Deep Feed-Forward Network, the classification accuracy of the proposed method for PD pattern recognition is about 94.5 %.

ACKNOWLEDGEMENT

Authors would like to thank Mr. Mohammad Bagher Souzanchi the managing director of Toskat Co., for providing an L-shaped busbar of GIS, which is specially designed and manufactured for experimental research work on partial discharge monitoring in a GIS system to the high voltage laboratory of Electrical Engineering Department of Sharif University of Technology. Also for providing the necessary support for the third author during his post-doctoral study, in which this research work is carried out.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Reza Rostaminia: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Roles/Writing - original draft, Writing - review & editing. **Mehdi Vakilian:** Investigation, Supervision, Validation, Visualization, Writing - review & editing. **Keyvan Firouzi:** Data curation, Supervision, Validation, Visualization, Writing - review & editing.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] H. Wang, H. E. Jo, and S. J. Kim, "Measurement and analysis of partial discharges in SF₆ gas under HVDC," *Meas. J. Int. Meas. Confed.*, vol. 91, pp. 351–359, 2016.
- [2] Y. Khan, A. A. Khan, and F. N. A. Budiman, "Partial discharge pattern analysis using support vector machine to estimate size and position of metallic particle adhering to spacer in GIS," *Electr. Power Syst. Res.*, vol. 116, pp. 391–398, 2014.
- [3] D. Lim, and S. Bae, "Study on oxygen / nitrogen gas mixtures for the surface insulation performance in gas insulated switchgear," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 3, pp. 1567–1576, 2015.
- [4] H. X. Ji, L. Cheng-rong, and P. Zhi-kai, "Influence of tip corona of free particle on PD patterns in GIS," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 24, no. 1, pp. 259–267, 2017.
- [5] M. Ren, M. Dong, and J. Liu, "Statistical analysis of partial discharges in SF₆ gas via optical detection in various spectral ranges," *Energies*, vol. 9, no. 3, p. 152, 2016.
- [6] D. Dai, X. Wang, J. Long, M. Tian, G. Zhu, and J. Zhang, "Feature extraction of GIS partial discharge signal based on S-transform and singular value decomposition," *IET Sci. Measur. Technol.*, vol. 11, no. 2, pp. 186–193, Mar. 2017.
- [7] IEC Standard, "High-voltage test techniques - Partial discharge measurements," IEC 60270, 2001.
- [8] A. Bargigia, W. Koltunowicz, and A. Pigin, "Detection of partial discharges in gas insulated substations," *IEEE Trans. Power Deliv.*, vol. 7, no. 3, pp. 1239–1249, 1992.
- [9] I. M. Welch, O. Farish, B. F. Hampton, and D. Templeton, "Partial discharge diagnostics for gas insulated substations," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 2, no. 5, pp. 893–905, 1995.
- [10] V. M. Ibrahim, Z. Abdul-Malek, and N. A. Muhamad, "Status review on gas insulated switchgear partial discharge diagnostic technique for preventive maintenance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 9–17, 2017.
- [11] H. Okubo, and A. Beroual, "Recent trend and future perspectives in electrical insulation techniques in relation to sulfur hexafluoride (SF₆) substitutes for high voltage electric power equipment," *IEEE Elec. Insul. Mag.*, vol. 27, no. 2, pp. 34–42, 2011.
- [12] L. E. Lundgaard, M. Runde, and B. Skyberg, "Acoustic diagnosis of gas insulated substations: a theoretical and experimental basis," *IEEE Trans. Power Deliv.*, vol. 5, no. 4, pp. 1751–1759, 1990.
- [13] S. Yoshida, H. Kojima, N. Hayakawa, F. Endo, and H. Okubo, "Light emission spectrum depending on propagation of partial discharge in SF₆," in *Conference Record of IEEE International Symposium on Electrical Insulation*, 2008, pp. 365–368.
- [14] Siying Wu, Fuping Zeng, Ju Tang, Qiang Yao, and Yulong Miao, "Triangle fault diagnosis method for SF₆ gas-insulated equipment", *IEEE Trans. Power Deliv.*, vol. 34, no. 4, 2019, pp. 1470–1477.
- [15] S. Okabe, S. Kaneko, T. Minagawa, and C. Nishida, "Detecting characteristics of SF₆ decomposed gas sensor for insulation diagnosis on gas insulated

- switchgears," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 15, no. 1, pp. 251–258, 2008.
- [16] Y. Xu, W. Liu, and W. Gao, "Investigation of disc-type sensors using the UHF method to detect partial discharge in GIS," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 5, pp. 3019–3027, 2015.
- [17] K. Khotimah, U. Khayam, and Y. Suwarno, Tai, M. Kozako, and M. Hikita, "Design of dipole antenna model for partial discharge detection in GIS," in *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2015, pp. 186–191.
- [18] W. Gao, D. Ding, W. Liu, and X. Huang, "Analysis of the intrinsic characteristics of the partial discharge induced by typical defects in GIS," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 20, no. 3, pp. 782–790, 2013.
- [19] S. Okabe, and S. Kaneko, "Electromagnetic wave propagation in a coaxial pipe GIS model," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 14, no. 5, pp. 1161–1169, 2007.
- [20] W. Gao, D. Ding, W. Liu, and X. Huang, "Propagation attenuation properties of partial discharge in typical in-field GIS structures," *IEEE Trans. Power Deliv.*, vol. 28, no. 4, pp. 2540–2549, 2013.
- [21] Q. Li, et al, "Influence of GIS structure on propagation of electromagnetic waves of partial discharge," in *PEAM 2011 - Proceedings: 2011 IEEE Power Engineering and Automation Conference*, 2011, vol. 2, pp. 128–135.
- [22] F. Álvarez, F. Garnacho, J. Ortego, and M. Á. Sánchez-Urán, "Application of HFCT and UHF sensors in online partial discharge measurements for insulation diagnosis of high voltage equipment," *Sensors (Switzerland)*, vol. 15, no. 4, pp. 7360–7387, 2015.
- [23] S. Das and P. Purkait, "Φ-q-n pattern analysis for understanding partial discharge phenomena in narrow voids," in *IEEE Power and Energy Society 2008 General Meeting: Conversion and Delivery of Electrical Energy in the 21st Century*, PES, 2008.
- [24] R. Rostaminia, M. Saniei, M. Vakilian, S. S. Mortazavi, and V. Parvin, "Accurate power transformer PD pattern recognition via its model," *IET Science, Measurement & Technology*, vol. 10, no. 7, pp. 745–753, 2016.
- [25] L. Li-Xue, H. Cheng-Jun, Z. Yi, and J. Xiu-Chen, "Partial discharge diagnosis on GIS based on envelope detection," *WSEAS Transactions on Systems*, vol. 7, no. 11, pp. 1238–1247, 2008.
- [26] L. Li, J. Tang, and Y. L. Liu, "Partial discharge recognition in gas insulated switchgear based on multi-information fusion," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 22, no. 2, pp. 1080–1087, 2015.
- [27] G. Li, M. Rong, X. Wang, X. Li, and Y. Li, "Partial discharge patterns recognition with deep convolutional neural networks," in *2016 International Conference on Condition Monitoring and Diagnosis*, Xi'an, China, 2016, pp. 324–327.
- [28] F. Bin, F. Wang, Q. Sun, S. Chen, J. Fan, and H. Ye, "Identification of ultra high-frequency PD signals in gas-insulated switchgear based on moment features considering electromagnetic mode," *High Volt.*, vol. 5, no. 6, pp. 688–696, 2020.
- [29] H. Guo, F. Lu, and K. F. Ren, "Simulation and measurement of PD induced electromagnetic wave leakage in GIS with metal belt," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 21, no. 4, pp. 1942–1949, 2014.
- [30] H. Guo, H. Qiu, L. Yao, F. Huang, and K. F. Ren, "Investigation on polarization characteristics of PD-induced electromagnetic wave leakage in GIS with metal belt," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 3, pp. 1475–1481, Jun. 2016.
- [31] S. Kaneko, S. Okabe, M. Yoshimura, H. Muto, C. Nishida, and M. Kamie, "Detecting characteristics of various type antennas on partial discharge electromagnetic wave radiating through insulating spacer in gas insulated switchgear," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 16, no. 5, pp. 1462–1472, 2009.
- [32] Y. Wang, Z. Wang, and J. Li, "UHF Moore fractal antennas for online GIS PD detection," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 852–855, 2016.
- [33] T. Ju, X. Zhongrong, Z. Xiaoxing, and S. Caixin, "GIS partial discharge quantitative measurements using UHF microstrip antenna sensors," in *Proc. 2007 Annual Report - Conference on Electrical Insulation and Dielectric Phenomena*, Vancouver, BC, Canada, Oct. 2007, pp. 116–119.
- [34] R. Rostaminia, M. Vakilian, and K. Firouzi, "Influence of gas insulated switchgear configuration components on UHF PD signals," *Journal of Applied Research in Electrical Engineering*, vol. 1, no. 2, pp. 139–148, 2022.
- [35] R. Candela, G. Mirelli, R. Schifani, "PD recognition by means of statistical and fractal parameters and a neural network," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 7, no. 1, pp. 87–94, 2000.

BIOGRAPHY



Reza Rostaminia received the B.Sc. degree in Electrical and Electronics Engineering from Babol Noshirvani University of Technology, Babol, Iran in 2007, The M.Sc. degree in Electrical Power Engineering from Khaje Nasir Toosi University of Technology (KNTU), Tehran, Iran in 2011 and the Ph.D. degree in Electric Power Engineering (High Voltage Engineering) from Shahid Chamran University of Ahwaz, Ahwaz, Iran in 2017. He was a post-doctoral researcher at the same university from 2019 to 2020. From 2015 to 2018, he was a sabbatical study in Electric Power Engineering (High Voltage Engineering) from Sharif University of Technology, Tehran, Iran in 2021. He joined the Parsian Substation Development Company, Tehran, Iran, where he is currently Technical Head of HV Substation Equipment Department. His research interest is High Voltage Engineering, Dielectrics and Insulation, Partial Discharge, Condition Monitoring and

Diagnosis of High Voltage Equipment, High Voltage Substation Design and Earthing design.



Mehdi Vakilian (M'88-SM'15) received the B.Sc. degree in electrical engineering and the M.Sc. degree in electric power engineering from the Sharif University of Technology, Tehran, Iran, in 1978 and 1986, respectively, and the Ph.D. degree in electric power engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1993. From 1981 to 1983, he was with Iran Generation and Transmission Company, and then with the Iranian Ministry of Energy from 1984 to 1985. Since 1986, he has been with the Faculty of the Department of Electrical Engineering, Sharif University of Technology. During 2001–2003, and 2014–2018 he was the Chairman of the department. During 2003 to 2004, and part of 2007, he was on leave of study at the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia. His research interests include transient modeling of power system equipment, especially power

transformers, optimum design of high-voltage equipment insulation, monitoring of power system equipment and their insulations, especially with partial discharge measurement, power system transients, and distribution system studies.



Keyvan Firuzi received the B.Sc. degree in Electrical and Electronics Engineering from University of Tabriz, Tabriz, Iran in 2012, The M.Sc. degree in Electrical Power Engineering and the Ph.D. degree in Electric Power Engineering (High Voltage Engineering) from Sharif University of Technology, Tehran, Iran in 2014 and 2019 respectively. He was a post-doctoral researcher at the same university from 2019 to 2020. From 2015 to 2018, he was a Research Scientist with Niroo Research Institution (NRI). In 2021, he joined the Electrical and Electronics Engineering Department at METU, where he is currently working as an Assistant Professor. His research interest is High Voltage Engineering, Dielectrics and Insulation, Partial Discharge, Condition Monitoring and Diagnosis of High Voltage Equipment, Signal Processing, and Machine Learning.

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

A Feedforward Active Gate Voltage Control Method for SiC MOSFET Driving

Hamidreza Ghorbani* and Jose Luis Romeral Martinez

MCIA-Motion Control and Industrial Applications, Universitat Politècnica de Catalunya, Barcelona, Spain

* Corresponding Author: hamidreza.ghorbani@upc.edu

Abstract: A new active gate drive for Silicon carbide (SiC) metal–oxide–semiconductor field-effect transistor (MOSFET) is proposed in this paper. The SiC MOSFET as an attractive replacement for insulated gate bipolar transistor (IGBT) has been regarded in many high power density converters. The proposed driver is based on a feedforward control method. This simple analog gate driver (GD) improves switching transient with minimum undesirable effect on the efficiency. This paper involves the entire switching condition (turn on/off), and the GD is applied to the SiC base technology of MOSFET. To evaluate the performance of the proposed GD, it will be compared with a conventional gate driver. The presented GD is validated by experimental tests. All the evaluations are carried out in a hard switching condition and at high-frequency operation.

Keywords: Active gate driver (AGD), SiC MOSFET, switching condition, feedforward control.

Article history

Received 08 January 2022; Revised 09 January 2023; Accepted 19 February 2023; Published online 29 March 2023.

© 2023 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this articleH. Ghorbani, and J. L. R. Martinez, "A feedforward active gate voltage control method for SiC MOSFET driving," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 87-94, 2023. DOI: [10.22055/jaree.2023.39698.1045](https://doi.org/10.22055/jaree.2023.39698.1045)**1. INTRODUCTION**

Conventional IGBTs are conventional switches in the structure of the power converters. However, because of some weak points such as operation in low-speed switching and low-temperature condition, the studies have been driven to silicon carbide (SiC) technology. SiC technology in power switches has emerged as a serious alternative to overcome the disadvantages of Si-switches. SiC device has some advantages such as higher operating frequency and temperature and lower on-resistance due to its bandgap and unipolar nature [1-2]. Moreover, due to its fast switching behaviour and shorter switching time, a better switching efficiency can be expected. In order to gain as efficiently as possible, engineers try to switch as fast as possible. However, the high speed switching in SiC MOSFETs increases the electromagnetic interference (EMI) emission. Therefore, the existing trade-off between efficiency improvement and EMI reduction through switching control brings a challenge in the gate driver designing. In addition, the SiC MOSFET normally has large input capacitance and higher threshold voltage, therefore more complex and sensitive driver is needed [3].

Several gate drivers (GDs) have been presented to improve the mentioned trade-off between fast switching and EMI [4-8]. However, most of them have been assigned to Si-

MOSFET or IGBT applications. Also, mainly they can be categorized in the closed-loop controller. Typically such controllers are effective and comprehensive for GDs, but in general, they increase the complexity of the GD's circuit. Therefore, some of the presented approaches are not attractive solutions for industrial. Moreover, in high-frequency operation rates, when SiC MOSFET is under hard switching condition; the advent of EMI problem is possible. Hence, designing proper GD for SiC MOSFETs has significant importance.

1.1. Overview of Gate Drivers for Power Devices

In order to enhance power density in switch-mode applications, operation at high switching frequency is necessary due to reducing the size of its passive component as well as it reduces the size of the heatsink. Thereby, the operation of IGBT is limited at the low switching frequencies (<20 kHz) [9]. However, in high speed switches the transition rates of current and voltage (di/dt and dv/dt respectively) get higher values. Also, parasitic inductance results some problematic oscillations and overshoots in current and voltage waveforms [10]. Changing the gate resistor R_g is known as a conventional solution [11-12]. Although the overshoot suppression can be achieved by high R_g value. However, the i_d and V_{ds} both get lower slopes which cause to increase switching times. As a result, the increased R_g

sacrifices additional switching losses. Another conventional driving technique is the use of an external gate-source capacitance (C_{gs}) in the GD circuit [13]. However, it increases the input capacitance (C_{ies}). This technique is used for IGBT's gate drivers due to roughly better efficiency (compared to the method of solely R_g increasing). Nevertheless, the gate capacitance is a parasitic element which potentially provokes transients and it can create some parasitic problems such as imposing the stress and crosstalk problem [14-15]. As a result, this may not be a favorable solution for SiC MOSFET applications which typically has large input capacitance. To improve the existing trade-off between switching loss, stress and EMI; diverse approaches have been proposed such as applying snubber circuits in Si and SiC devices [16-17] active gate voltage controlling [18] active gate current driving [19] resonant gate drivers [16, 20], etc. All mentioned techniques could be used for driving SiC MOSFETs. Although, these GDs can minimize stress from the power device, however, these deal with more complexity or more cost and more switching losses. The control of GDs is not the single possible way for EMI reduction, rather with using a better design of PCB layout the parasitic (stray) inductances can be reduced, and consequently we will have less EMI problems.

1.2. SiC MOSFET Gate Drivers

The SiC MOSFETs are widely employed in power converters due to its advantages. This switching technology inherently has lower trans-conductance compared to Si-MOSFETs or IGBTs. Thus, higher orders of gate-source voltage are required for switching-on. Also, the gate-source voltage pulse is commonly asymmetrical. Therefore, different values of R_g should be used in their GDs [21]. Conventionally, two different gate resistance is used in the drive circuit for controlling each turn path. This common driver controls both turn-on and turn-off paths separately. A gate boost circuit was introduced for SiC MOSFET driving in [22] which had reduced the switching losses, however, the transient and overshoot had not been reduced. The same technique is presented in [23] as well. Many studies for controlling di/dt and dv/dt transition by closed-loop control method have been reported [23-26]. Such controllers have been allocated to guarantee the safe operation of MOS-gate switches under different and variable loads. However, they increase the complexity of the driver's circuit.

According to the presented overview, most of the offered approached are related to efficiency improvement and for solving some other issues such as EMI reduction, overshoot suppression, stability improvement etc. mainly the presented solutions have fallen in the complex closed-loop GD controllers. This paper presents a simple control method for driving SiC MOSFETs. The control concept is based on a feed-forward controller. The effective performance of the controller beside its simple structure is the main advantage of this GD. The purpose is the switching transient improvement with a minimum undesirable effect on efficiency. In the next section, the operation of SiC MOSFET and the principles of new GD are presented.

2. ACTIVE GATE DRIVER

2.1. Principles of Proposed Controller

The test circuit is represented in Fig. 1. The controller is applied into the gate circuit. The profile of V_{gg} voltage signal is changed by the controller and it is delivered to the gate port of MOSFET. In order to test in a hard-switching condition, the load is highly inductive.

The schematic of the proposed controller is depicted in Fig. 2. Since the SiC MOSFET meets several intervals during the switching conditions, controller changes the profile of gate signal during MOSFET's active region (Fig. 3a shows these intervals). The modification process of gate signal has been demonstrated in Fig. 3b. The turn-on is initiated at t_0 , and step voltage (from $-V_{EE}$ to $+V_{CC}$) is applied to the gate. As shown in Fig. 2, each switching state is separated from the other by diodes for individual controlling. The positive side of voltage signal is driven by d_1 and the d_2 conducts its negative for turn-on and turn-off controlling respectively. Both control paths have same structure and operate based on same concept.

However, the required parameters should be defined according to each switching condition. In both cases, the controller gets a step voltage value (ΔV_g) according to (1) from the input. Depending on the suppression rate of overshoot at each swathing state, a portion of the input value is given to the corresponding control path. K_i and K_v represent these coefficients for turn-on and turn-off controlling respectively.

In each switching condition, weakened signal with a negative coefficient is summed with the same positive signal

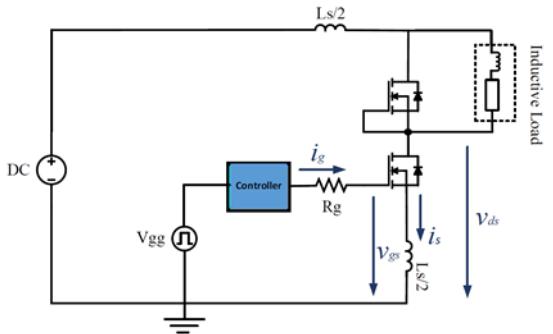


Fig. 1: Schematic of test circuit.

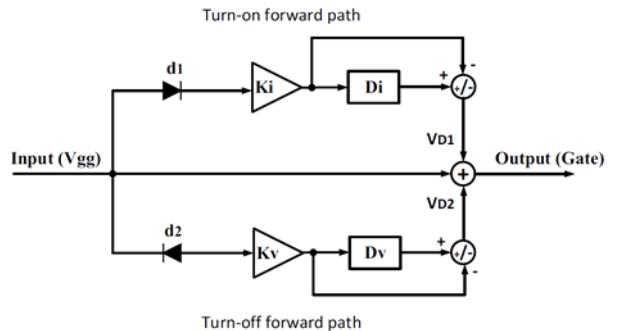
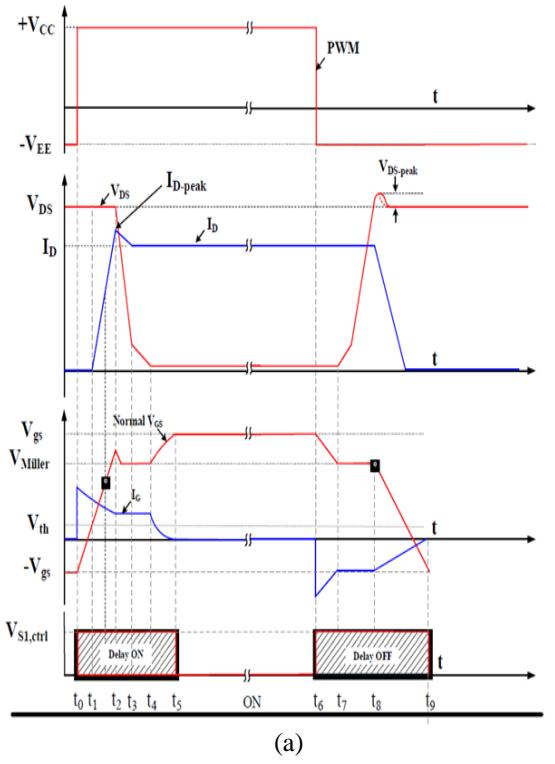
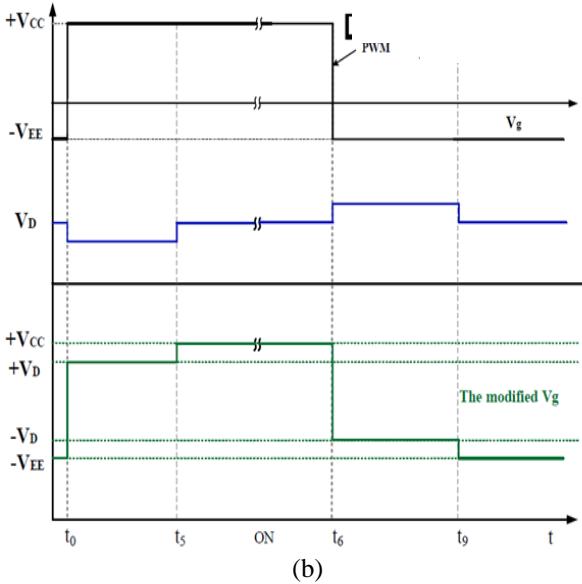


Fig. 2: Block diagram of the proposed feedforward controller.



(a)



(b)

Fig. 3: (a) The transient behavior of MOSFET switching, and (b) controller performance for V_g modification.

that has a delay. These delays are created by blocks D_i and D_v . The applied D_i delay covers whole turn-on ($t_0 < t < t_5$) and D_v covers turn-off ($t_6 < t < t_9$) intervals. The resultant voltage signal is called V_D here. Finally, the original V_{gg} signal after summing with V_{D1} and V_{D2} results modified V_g signal which is applied on gate port for driving MOSFET. The modified V_g affects to the current transient while turn-on condition and as well as to the dynamic of voltage during turn-off condition.

$$\Delta V_g = V_{cc} - V_{EE} \quad (1)$$

According to what has been presented in [27], with a little approximation the current and voltage transitions may be defined as (2) and (3), which both depend on a different

voltage value between $V_g(+/-)$ to $V_{gs}(th)$. This differential voltage value affects to the injected gate current i_g in each switching state. Thereby, the used technique can be effective to control of both current and voltage transitions.

$$\frac{di_d}{dt} = g_m \cdot \frac{V_{cc} - V_{gs}(th) - \frac{i_d}{2 \cdot g_m}}{C_{iss} \cdot R_g} \quad (2)$$

$$\frac{dv_{ds}}{dt} = \frac{V_{EE} + V_{gs(plateau)}}{C_{gd} \cdot R_g} \quad (3)$$

In (2), g_m is the trans-conductance and $C_{iss} = C_{gs} + C_{gd}$ is the input capacitance of the SiC MOSFET. All details about the switching process of the MOSFET is fairly well demonstrated in [10]. Here, we present the performance of the new gate driver for controlling di_d/dt and dv_{ds}/dt rates.

2.2. Parameters and Limitations

For tuning the controller in each switching state two parameters are necessary. K_i coefficient and the D_i delay for turn-on and K_v coefficient and the D_v delay for turn-off. In the case of delays, these parameters can be determined based on application note or experimental observations. As already mentioned, the delay time must cover whole active region times. Because of the time difference between switching on and off, two individual delays have been considered for corresponding states as shown in Fig. 2. It should be noted that the margin determination for delays is not a delicate factor. Because after finishing D_i or D_v delay, the modified V_g returns to its original value when MOSFET is in the saturation (steady state) region. For this reason, the delay time could be defined much longer than switching time. The K coefficient determines V_D voltage value (see Fig. 3b and (5)) or in other word, it determines the ΔV_g voltage value while controlling time. As a result, the injected gate current and then switching transient will get effect by that. V_D is the reduced voltage level during turn on/off. Since the SiC MOSFETs driving is asymmetric and the absolute value of V_{EE} is smaller than V_{cc} , thus the change domain of the V_{gg} for each switching state must be determined individually. For turn-on condition K_i coefficient can be obtained by the following equations.

$$V_{gs, th} < \sigma_1 < V_{cc} \quad (4)$$

$$V_D^+ = V_{cc} - \sigma_1 \quad (5)$$

$$\Delta V_{m1} = V_D^+ - V_{EE} \quad (6)$$

$$K_i = 1 - \left(\frac{\Delta V_{m1}}{\Delta V_g} \right) \quad (7)$$

As well as for turn-off condition, K_v coefficient can be defined as:

$$V_{EE} < \sigma_2 < 0 \quad (8)$$

$$V_D^- = V_{EE} - \sigma_2 \quad (9)$$

$$\Delta V_{m2} = V_{cc} - V_D^- \quad (10)$$

$$K_v = 1 - \left(\frac{\Delta V_{m2}}{\Delta V_g} \right) \quad (11)$$

In the mentioned equations σ_1 and σ_2 are variable factors which should be selected according to the desired di_d/dt and dV_{ds}/dt rates respectively with considering the limitations presented in (4) and (8). For turn-on condition the modified V_g has been limited by $V_{gs,th}$ and for turn-off condition it has been limited by zero. Accordingly, smaller ΔV_m has higher impact on the switching transient and oscillations suppression.

2.3. Controller Tuning

According to the presented method, the level of applied intermediate voltages and their time duration should be determined. The influence of each control parameter on the switching transient behaviour is explained here. Also, the optimal interval values for each switching condition should be determined.

2.3.1. Tuning for turn-on

Based on what expressed in previous section, K_i coefficient determines the level of intermediate voltage which can be reduced up to MOSFET's threshold voltage. As a result, the applied intermediate voltage affects to di/dt and current overshoot at turn-on. The time duration of intermediate voltage is another consideration that must be long enough to cover turn-on active region. In this case study, 3 μs has been considered for D_i . In order to realize which level of reduced gate voltage provides a desirable di/dt and current overshoot, the corresponding MOSFET is tested by different intermediate voltage values. Fig. 4 shows the effect of controller on MOSFET behaviour at turn-on.

2.3.2. Tuning for turn-off

Also, the voltage transition (dv/dt) and overshoot ($V_{DS-peak}$) are being affected by intermediate voltage while turn off condition (according to (3)). The resultant intermediate gate voltage through K_v and its consequence on MOSFET behaviour at turn-off has been reflected in Fig. 5. In this controlling stage, D_v is 2 μs which covers whole transient behaviour of MOSFET while turn-off with considering worst case.

3. TEST CONDITION

Experimental tests evaluate the performance of the proposed gate driver. The test circuit is a standard clamped-inductive circuit which is depicted in Fig. 1. The driving power SiC MOSFET and the clamped SiC MOSFET both are from a same type (SCT2080KE). The parasitic inductance (LS) which comes from the loop of the PCB and power devices is 120 nH. The load current is 6 A, and the value of L in load is 330 μ H. A square signal with 50% of duty cycle and frequency at 100 kHz has been applied to the input. The voltage of dc-bus is 400 V and the V_{gg} supply for original gate driver is -5/+18 V. The applied gate resistor (R_g) for turning-on is 33 ohms and for turning-off is 46 ohms. The experimental waveforms have been captured by a Tektronix MSO 4054 (500 MHz) digital oscilloscope. The insulators and the safety instruments for protection are not demonstrated here.

3.1. Optimized Tuning

The product of multiplication of the drain-source voltage $V_{ds}(t)$ to output current $I_d(t)$ during the switching time results

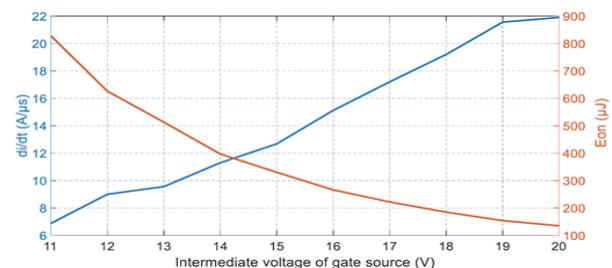


Fig. 4: The effect of intermediate gate-voltage levels on the peak value of current transient and di/dt while turn-on control domain.

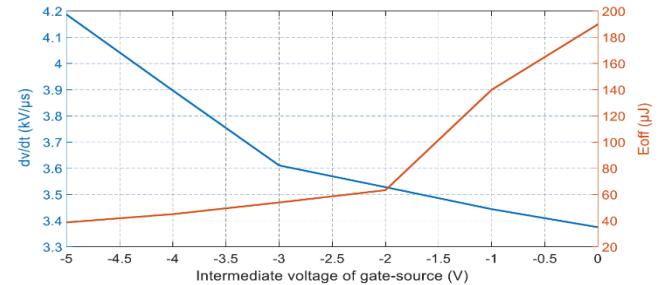


Fig. 5: The effect of intermediate gate-voltage levels on voltage transition and voltage overshoot in turn-off control domain.

corresponding switching loss. The lost energy while turn-on and turn-off can be calculated from (12) and (13), respectively. Accordingly, to reach an optimized design, the effect of K_i and K_v on switching loss and peak value of oscillations are evaluated.

$$E_{on} = \int_{t_0}^{t_5} v_{ds(t)} \times i_{d(t)} dt \quad (12)$$

$$E_{off} = \int_{t_6}^{t_9} v_{ds(t)} \times i_{d(t)} dt \quad (13)$$

First, each one of the switching losses and peak values of current transient (at turn-on) and voltage overshoot (at turn-off) must be normalized as below equations.

$$\alpha = \frac{E_{on}}{E_{min,on}} \quad (14)$$

$$\beta = \frac{E_{off}}{E_{min,off}} \quad (15)$$

In this analysis, the minimum value of switching loss (E_{min}) is assumed when the minimum possible value of R_g has been used. This value for each switching condition is 6.3 Ω . Also, in this condition the maximum peak value of current transient ($i_{d,max,peak}$) and maximum voltage overshoot ($V_{ds,max,ov}$) can be measured.

$$\gamma = \frac{di_d/dt}{d_{id}/dt - max} \quad (16)$$

$$\delta = \frac{dv_{ds}/dt}{d_{vds}/dt - max} \quad (17)$$

α and γ present the normalized values of the lost energy and peak value of current oscillations at turn-on condition respectively. Also, β and δ represent the normalized values of the lost energy and voltage overshoot at turn-off condition

respectively. With these assumptions, the optimal intermediate gate voltages for each switching state can be obtained. Fig. 6 and Fig. 7 show these optimal intermediate gate-source voltages.

To realize the effect of V_{GS} value on the transient behaviour of switch while turning-on equation (18) represents the relation of current peak normalized value (see (16)) to the normalized value of lost energy (see (14)) for a specific VGS value at turn-on condition. Also, in the same way it can be realized for turning-off condition by (19).

The test results base on (18) have been reflected in Table 1 and for turn-off condition have been reflected in Table 2. Then, the highest value of γ/α column expresses the highest impact of V_{GS} value or in other word it belongs to optimum value of V_{GS} .

$$V_{GS_{on}} = \left| \frac{\gamma_n}{\alpha_n} \right|_{max} \quad (18)$$

$$V_{GS_{off}} = \left| \frac{\delta i}{\beta i} \right|_{max} \quad (19)$$

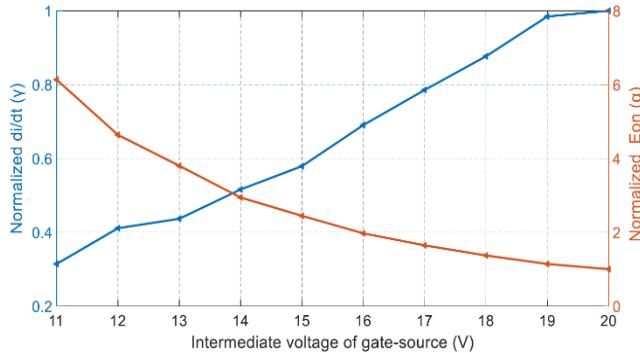


Fig. 6: Optimal intermediate voltage for gate-source (V) at turn-on.

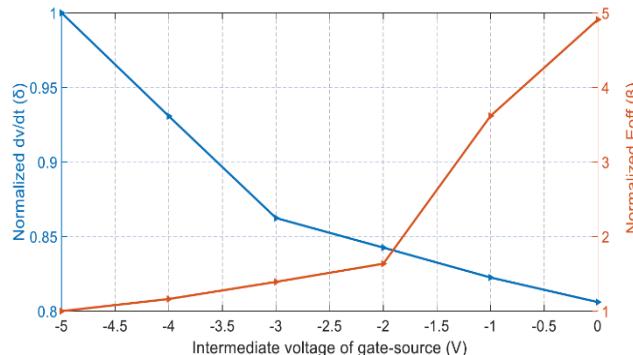


Fig. 7: Optimal intermediate voltage for gate-source (V) at turn-off.

Table 3 shows the optimal setting at both switching states and corresponding values. Through original GD, SiC MOSFET has been driving with $V_g = +20/-5$ V and the implemented external gate resistor (R_g) is valid for the new GD as well. Though defined coefficients, K_i delivers $V_{D+} = 18$ V to the gate for switching-on and by K_v it gets $V_{D-} = -4$ V while turn-off state.

4. EXPERIMENTAL VALIDATION

The proposed GD is validated by experimental tests. In order to evaluate the performance of new GD, the transient behaviour of the MOSFET in both with original GD and with proposed GD are compared with together.

4.1. The Test Results

The profile of output parameters (i_d and V_{DS}) of MOSFET when it is driven by original GD are demonstrated in Fig. 8. Then in next figure, for a closer look, the switching behaviour of MOSFET driven by new GD is zoomed in different tuning conditions.

As can be seen in Fig. 9, the overshoot value in output current and corresponding oscillations can be suppressed by applying different K_i . The biggest suppression rate belongs to which has smallest ΔV_{ml} (see (6) and Fig. 3). However, the optimized value ($V_{D+} = 18$ V and $V_{D-} = -4$ V) for driving is compared with original gate driver. Fig. 10 and Fig. 11 represent the i_d current waveform while turning on and off conditions.

Although the proposed GD may suppresses the overshoot up to 5.2 A, however, the optimized tuning condition the overshoot can be reduced up to 5.5 A. In this tuning condition,

Table 1: Optimal VGS value in turn-on condition.

| n | V_{GS} | γ | α | γ/α |
|---|----------|----------|----------|-----------------|
| 1 | 19 | 0.0155 | 0.141 | 0.1099 |
| 2 | 18 | 0.01078 | 0.229 | 0.4707 |
| 3 | 17 | 0.0913 | 0.274 | 0.3332 |
| 4 | 16 | 0.095 | 0.326 | 0.2914 |
| 5 | 15 | 0.111 | 0.474 | 0.2341 |
| 6 | 14 | 0.063 | 0.497 | 0.1267 |

Table 2: Optimal VGS value in turn-off condition.

| i | V_{GS} | γ | α | γ/α |
|---|----------|----------|----------|-----------------|
| 1 | -4 | 0.0693 | 0.163 | 0.425 |
| 2 | -3 | 0.0683 | 0.23 | 0.297 |
| 3 | -2 | 0.0198 | 0.243 | 0.0815 |

Table 3: The controller tuning parameters.

| | P | Value | P | Value | $E_{on}(\mu J)$ | $E_{off}(\mu J)$ | $E_{min}(\mu J)$ | $i_d, max(A)$ | $V_{ds, max, or}(V)$ |
|-------------------|-------|-----------|----------|-------|-----------------|------------------|------------------|---------------|----------------------|
| <i>Turn – on</i> | K_i | 0.122 | α | 1.52 | 190 | - | 125 | 6.6 | - |
| | D_i | 3 μs | γ | 0.86 | | | | | |
| <i>Turn – off</i> | K_v | 0.1 | β | 1.18 | - | 46 | 39 | - | 580 |
| | D_v | 2 μs | δ | 0.93 | | | | | |

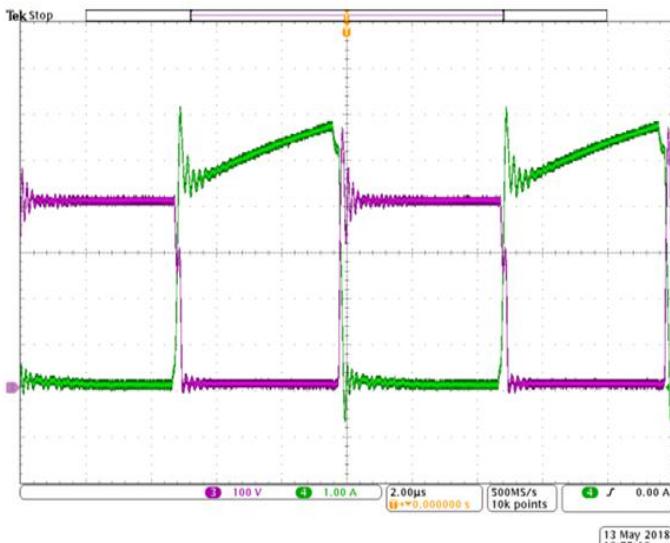


Fig. 8: Output voltage and current of MOSFET driven by original GD.

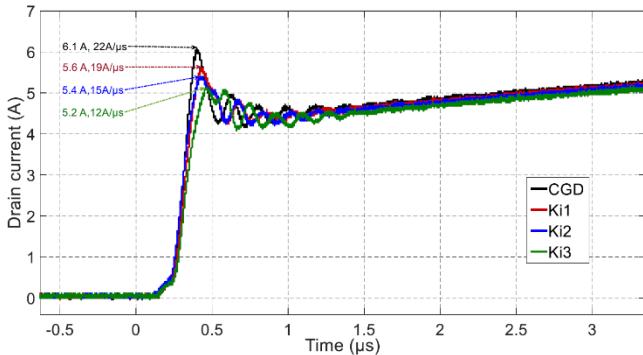


Fig. 9: Zoomed view of drain current with different K_i .

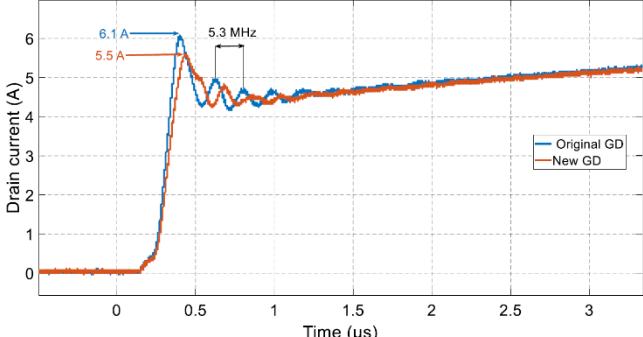


Fig. 10: Drain current with new GD (optimal tuning value) and original GD at turning-on.

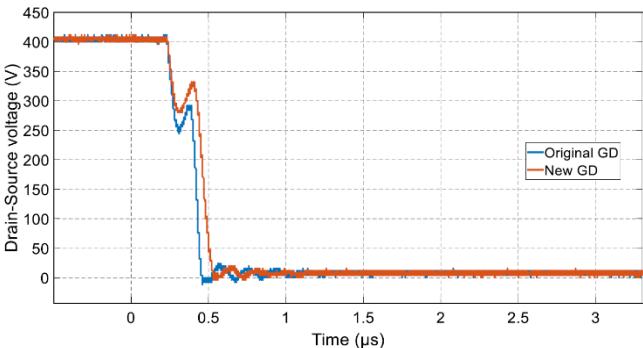


Fig. 11: Drain current with new GD (optimal tuning value) and original GD at turning-off.

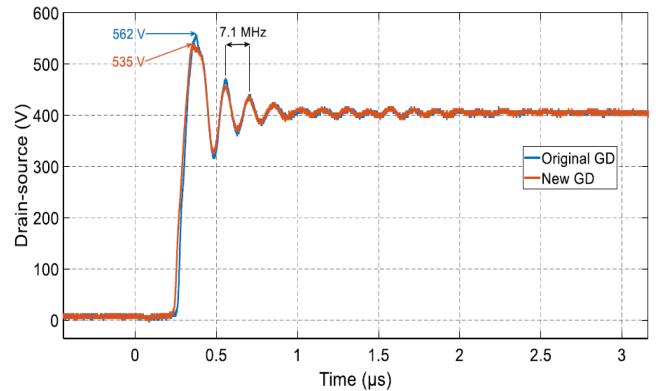


Fig. 12: Drain-Source voltage with new GD (optimal tuning value) and original GD at turning-on.

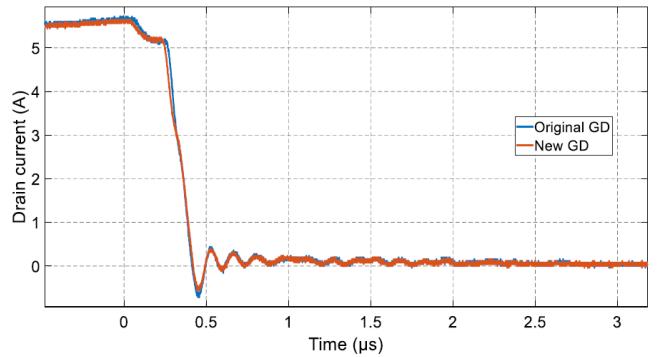


Fig. 13: Drain-Source voltage with new GD (optimal tuning value) and original GD at turning-off.

the slope of the current (di/dt) in fundamental frequency (100 KHz) ten times has been increased compare to its maximum value. Also the current fluctuation in switching-on condition with 5.3 MHz (see Fig. 10) highly has been removed which both manners help to eliminate EMI problem from switch mode power supplies.

This comparison can be carried out in the case of drain-source voltage as well. Fig. 12 and Fig. 13 demonstrate the output voltage profiles with original and new GDs in both switching condition.

The obtained results show that output voltage at turn-off condition gets minimum effect from applied control method.

4.2. Performance Index

Based on previous subsection, the experimental setup of the gate driver for both switching states has developed. In order to observe the effect of the applied GD, an analytical test between new GD and original GD (with minimum $R_{g,ext}$ values) has been carried out. The purpose of this comparison is the evaluation of transient behaviour during the operation of new GD and the CGD. Another criteria in this analysis is the comparison of the switching losses between these GDs. The change of gate resistor ($R_{g,ext}$) is known as a conventional driver for MOS-channel switches [27-28]. So, as a conventional solution, $R_{g,ext}$ is increased up to 15 Ω (for turning on) and 22 Ω (for turning off) to achieve the same level of overshoot suppression in current and voltage that new GD presents in its operation. In this condition the switching losses of both GDs can be calculated according to (12) and (13). The results have reflected in Table 4.

Table 4: The performance index.

| Gate driver | I_d (A) overshoot | Voltage overshoot (V) | E_{on} (μ J) | E_{off} (μ J) | di_d/dt (A/ μ s) | dv_{ds}/dt (KV/ μ s) |
|---|---------------------|-----------------------|---------------------|----------------------|------------------------|----------------------------|
| Original gate driver | 6.1 | 562 | 125 | 39 | 22 | 4.2 |
| New gate driver | 5.5 | 535 | 190 | 46 | 19.2 | 3.9 |
| CGD, $R_{g, on}=15 \Omega$ $R_{g, off}=22 \Omega$ | 5.5 | 535 | 212 | 50 | 18.5 | 3.6 |

5. CONCLUSION

Based on the obtained results, the new active voltage gate driver has improved transient behavior of the SiC MOSFET in both switching condition. Although this technique mostly improves the switching in turn-on condition, however it has better performance compared to CGD. In this study, we tried to use an optimized gate drive for switching in both switching states. Applying the proposed feedforward controller on SiC technology MOSFETs, optimal tuning of active voltage GD and evaluation of this GD by performance index were the important topics of this study.

ACKNOWLEDGMENTS

This work was supported by the Generalitat de Catalunya, grant number SGR 2017 SGR 967.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Hamidreza Ghorbani: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **Jose Luis Romeral Martinez:** Conceptualization, Supervision, Validation, Writing - review & editing.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] A. Niwa, T. Imazawa, T. Kimur, T. Sasaya, T. Isobe, and H. Tadano, "Novel dead time controlled gate driver using the current sensor of SiC-MOSFET," in *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 2015, pp. 001651 – 001656.
- [2] J. Millan, P. Godignon, X. Perpina, A. Perez-Tomas, and J. Rebollo, "A survey of wide-bandgap power semiconductor devices," *IEEE Trans. Power Electron.*, vol. 29, no. 5, pp. 2155-2163, 2014.
- [3] F. Mo, J. Furuta, and K. Kobayashi, "A low Surge voltage and fast speed gate driver for SiC MOSFET with switched capacitor circuit," in *2016 IEEE 4th Workshop on Wide Bandgap Power Devices and Applications (WiPDA)*, 2016, pp. 282 - 285.
- [4] Y. Lobsiger, and J. W. Kolar, "Closed-loop di/dt and dv/dt IGBT gate driver", *IEEE Trans. Power Electronic*, vol. 30, no. 6, pp. 3402 - 3417, 2015.
- [5] Z. Wang, X. Shi, M.L. Tolbert, F. Wang, and B.J. Blalock "A di/dt feedback-based active gate driver for smart switching and fast overcurrent protection of IGBT modules", *IEEE Trans. Power Electronics*, vol. 29, no. 7, pp. 3720 - 3732, 2014.
- [6] Y. Lobsiger, and J. W. Kolar, "Closed-loop di/dt and dv/dt control and dead time minimization of IGBTs in bridge leg configuration", in *2013 IEEE 14th Workshop on Control and Modeling for Power Electronics (COMPEL)*, 2013.
- [7] H. Ghorbani, V. Sala, A. Paredes, and L. Romeral, "Embedding a feedforward controller into the IGBT gate driver for turn-on transient improvement" *Microelectronics Reliability*, vol. 80, pp. 230 – 240, 2018.
- [8] L. Shu, J. Zhang, F.Z. Peng, and Z. Chen, "A voltage controlled current source gate drive method for IGBT devices," in *Proc. of IEEE Energy Conversion Congr. and Expo. (ECCE)*, 2014.
- [9] S. Yin, K. J Tseng, R. Simanjorang, and P. Tu, "Experimental comparison of high-speed gate driver design for 1.2-kV/120-A Si IGBT and SiC MOSFET modules", *IET Power Electronics*, vol. 10, no. 9, pp- 979 – 986, 2017.
- [10] J. Wang, H. S. h. Chung, and R. T. h. Li, "Characterization and experimental assessment of the effects of parasitic elements on the MOSFET switching performance," *IEEE Trans. Power Electron.*, vol. 28, no. 1, pp. 573–590, 2013.
- [11] Y. Lobsiger, and J. W. Kolar, "Closed-loop di/dt and dv/dt IGBT gate driver", *IEEE Trans. Power Electronic*, Vol. 30, no. 6, pp. 3402 - 3417, June. 2015.
- [12] P. Nayak, K. Hatua "Active gate driving technique for a 1200 V SiC MOSFET to minimize detrimental effects of parasitic inductance in the converter layout", *IEEE Trans. Industry Appl.*, vol. 54, no. 2, pp. 1622-1633, 2018.
- [13] H. G. Lee, Y. H. Lee, B. S. Suh, and D. S. Hyun, "An improved gate control scheme for snubberless operation of high power IGBTs," in *1997 Thirty-Second IAS Annual Meeting Industry Applications Conference*, 1997, pp. 975- 982.
- [14] S. Jahdi, O. Alatise, J. A. O. Gonzalez, R. Bonyadi, L. Ran, and P. Mawby, "Temperature and switching rate

- dependence of crosstalk in Si-IGBT and SiC power modules,” *IEEE Trans. Ind. Electron.*, vol. 63, no. 2, pp. 849–863, 2016.
- [15] D. W. Peters, “Turn-On Delay Time of MOS Transistors”, *Proceedings of the IEEE*, vol. 56, no. 1, pp. 89 – 90, 1968.
- [16] N. Oswald, P. Anthony, N. McNeill, and B. H. Stark, “An experimental investigation of the trade-off between switching losses and EMI generation with hard-switched all-Si, Si-SiC, and all-SiC device combinations,” *IEEE Trans. Power Electron.*, vol. 29, no. 5, pp. 2393–2407, 2014.
- [17] X. Gong, and J. A. Ferreira, “Investigation of conducted EMI in SiC JFET inverters using separated heat sinks,” *IEEE Trans. Ind. Electron.*, vol. 61, no. 1, pp. 115–125, 2014.
- [18] P. R. Palmer; J. Zhang, and X. Zhang, "SiC MOSFETs connected in series with active voltage control," in *2015 IEEE 3rd Workshop on Wide Bandgap Power Devices and Applications (WiPDA)*, Blacksburg, VA, USA, 2015, pp. 60-65.
- [19] L. Shu, J. Zhang, F. Peng, and Z. Chen, “An active current source IGBT gate drive with closed-loop di/dt and dv/dt control” *IEEE Trans. Power Electron.*, vol. 32, no. 5, pp. 3787 - 3796, 2017.
- [20] H. Fujita, “A resonant gate-drive circuit capable of high-frequency and high-efficiency operation,” *IEEE Trans. Power Electron.*, vol. 25, no. 4, pp. 962–969, 2010.
- [21] D. Peftsis, and J. Rabkowski, “Gate and base drivers for silicon carbide power transistors: An overview,” *IEEE Trans. Power Electron.*, vol. 31, no. 10, pp. 7194–7213, 2016.
- [22] K. Yamaguchi, Y. Sasaki, and T. Imakubo, “Low loss and low noise gate driver for SiC-MOSFET with gate boost circuit,” in *IECON 2014-40th Annual Conference of the IEEE Ind. Electron. Soc.*, pp. 1594-1598.
- [23] Z. Wang, X. Shi, M.L. Tolbert, F. Wang, and B.J. Blalock “A di/dt feedback-based active gate driver for smart switching and fast overcurrent protection of IGBT modules”, *IEEE Trans. Power Electronics*, vol. 29, no. 7, pp. 3720 - 3732, 2014.
- [24] H. RiaziMontazer, and S. K. Mazumder, “Optically switched-drive-based unified independent dv/dt and di/dt control for turn-off transition of power MOSFETs,” *IEEE Trans. Power Electron.*, vol. 30, no. 4, pp. 2338–2349, 2015.
- [25] H. RiaziMontazer, A. Rahnamaee, A. Mojab, S. Mehrnami, S. K. Mazumder, and M. Zefran, “Closed-loop control of switching transition of SiC MOSFETs,” in *2015 IEEE Applied Power Electronics Conference and Exposition (APEC)*, pp. 782–788, 2015.
- [26] Fink. K, and S. Bernet, “Advanced Gate Drive Unit With Closed-Loop dic/dt Control”, *IEEE Trans. Power Electronics*, vol. 28, no. 5, pp. 2587 - 2595, 2013.
- [27] A. Paredes, V. Sala, H. Ghorbani, and L .Romeral, “A novel active gate driver for improving SiC MOSFET switching trajectory”, *IEEE Trans. Ind. Electron.*, vol. 64, no. 11, pp. 9032-90422, 2017.
- [28] A. Paredes, V. Sala, H. Ghorbani, and L .Romeral, “A novel active gate driver for silicon carbide MOSFET”, in *Annual Conference of the IEEE Industrial Electronics Society IECON*, 2016, pp. 3172-3177.
- [29] J. D. Kagerbauer, and T. M. Jahns, “Development of an active dv/dt control algorithm for reducing inverter conducted EMI with minimal impact on switching losses,” in *2007 IEEE Power Electronics Specialists Conference*, Orlando, FL, USA, 2007, pp. 894-900.

Copyrights

© 2023 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).





Research Article

Design of Low-Power Approximate Logarithmic Multipliers with Improved Accuracy

Mojtaba Arab Nezhad and Ali Mahani*

Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman 7616913439, Iran

* Corresponding Author: amahani@uk.ac.ir

Abstract: Approximate computing is considered a promising way to design high-performance and low-power arithmetic units recently. This paper proposes an energy-efficient logarithmic multiplier for error-tolerant applications. The proposed multiplier uses a novel technique to calculate the powers of two products to reduce critical path complexity. Also, a correction term is provided to improve the multiplier accuracy. Additionally, the use of approximate adders in our design is investigated, and optimal truncation length is obtained through simulations. We evaluated our work both in accuracy and hardware criteria. Experiments on a 16-bit proposed multiplier with approximate adder show that power-delay product (PDP) is significantly reduced by 34.05% compared to the best logarithmic multipliers available in the literature, while the mean relative error distance (MRED) is also decreased by 21.1%. The results of embedding our multiplier in the dequantization step of the JPEG standard show that the image quality is improved in comparison with other logarithmic multipliers. In addition, a subtle drop in image quality compared to utilizing exact multipliers proves the viability of our design.

Keywords: Logarithmic multiplier, approximate computing, error-tolerant.

Article history

Received 19 December 2020; Revised 15 June 2021; Accepted 13 August 2021; Published online 20 August 2021

© 2021 Published by Shahid Chamran University of Ahvaz & Iranian Association of Electrical and Electronics Engineers (IAEEE)

How to cite this article

M. Arab Nezhad, and A. Mahani, "Design of low-power approximate logarithmic multipliers with improved accuracy," *J. Appl. Res. Electr. Eng.*, vol. 2, no. 1, pp. 95-102, 2023. DOI: [10.22055/jaree.2021.36119.1018](https://doi.org/10.22055/jaree.2021.36119.1018)

1. INTRODUCTION

Significant computational demands of large-scale applications such as scientific computing, social media, and financial analysis have exceeded available resources [1]. Machine learning algorithms are becoming more accurate every day and, in many areas, have gone beyond human accuracy, but this accuracy comes at the expense of increased computations [2]. Due to recent advances in technology and the end of Dennard scaling, it has become difficult to improve the performance of computing systems at current power levels [3]. A wide range of applications that require huge computations can maintain their output well enough despite some computational error. Some of these applications are as follows [4]:

- Applications such as machine learning and adaptive filters that are inherently error-tolerant.
- In digital signal processing, because the inputs are often noisy, accuracy is limited.

- In image processing, due to limitations in human cognition, the existence of some errors in calculations is not detectable in the output.

Approximate computing introduces some errors in the calculations but simplifies the arithmetic operations. Therefore, approximate computing can be considered as a promising way to reduce power consumption. Approximate computing techniques can be applied to various levels, such as hardware, architecture, algorithm, and software [5]. Adders and multipliers are the arithmetic units that are the main subject of hardware-level approximations [6]. In the aforementioned applications, there are an abundant number of arithmetic processing that involve addition and multiplication. To design high-performance arithmetic processors, it is necessary to optimize the performance and power consumption of its main components, namely adders and multipliers. For this reason, much attention has been paid to approximate computing techniques at the circuit level to improve these units.

Multiplication is more elaborate than the addition operation and has always been a limiting factor to improve speed and area [7]. Hence enhancing this operation can result in considerable improvement in the whole design. Also, most applications mentioned above consist of some dominant kernels that intensively rely on multiplication. So, multipliers become primary candidates for approximation computing to improve whole system performance [8]. A conventional multiplier consists of partial product generation, accumulation, and final addition [9]. Various parts of a multiplier are capable of applying approximate techniques [5]. Different approximation approaches are proposed to design highly efficient multipliers [10]: Approximate recursive multipliers are built of 2×2 approximate multiplier blocks to form a complete multiplier [11]. In [12], a dynamic truncation method based on leading-one position has been introduced, which reduces the multiplier bit-width. Paper [13] proposed using an $m \times m$ multiplier to design approximate $n \times n$ multipliers where $m < n$. Approximate radix-4 booth multiplier [14] is another multiplication technique. A different category from traditional multipliers are logarithmic multipliers that use binary logarithms to simplify multiplication operations. In the logarithm domain multiplication converts to addition. Multipliers that use logarithm transformation are inherently erroneous. Such error occurs for the following reasons: 1) a limited number of precision bits and 2) errors that happen at the time of transformation to the logarithmic system. Mitchell introduced the first logarithmic multiplier [15]. In conventional approximate multipliers, the accuracy is high, but the area and power consumption are also high. But in logarithmic multipliers reducing hardware overhead as well as reducing power, take precedence over multiplier accuracy. This property makes logarithmic multipliers suitable for large-scale applications that require high parallelism [16], [17]. In this paper, we introduce a new logarithmic multiplier to optimize power consumption and reduce hardware area and latency, while improving multiplier accuracy in terms of error amount as well as error distribution. The main contributions of this article are summarized as follows:

- A new multiplication algorithm is presented that uses less hardware resources than previous designs and is therefore more power efficient.
- We have introduced and used a correction term that improves the multiplier error characteristic.
- A new method has been proposed to calculate the product of the power of number two, which reduces the critical path delay significantly.
- The use of approximate adders in the proposed design has been investigated and the truncation length parameter for compromise between circuit complexity and multiplier approximation error has been introduced so that the proposed design can be adjusted for different applications.

The rest of the paper is organized as follows: in Section 2, we have introduced logarithmic multipliers, notably Mitchell's algorithm. The main problems of these multipliers are described, and the main approaches to alleviate them are reviewed. The first part of Section 3 is devoted to proposing the multiplication algorithm. The remainder of this section deals with the hardware architecture of the multiplier. Error

analysis and simulation results are presented in Section 4. Section 5 implements our multiplier in JPEG image compression and decompression standard and evaluates the output image's quality. Finally, Section 6 concludes the paper.

2. REVIEW AND RELATED WORKS

Due to the complexity of the multiplication operation, approximate multipliers are designed for trade-off between accuracy and design efficiency. Various approximation methods have been proposed to simplify multiplier circuit. Exploring available references shows that approximation techniques in multipliers are mainly grouped in logarithmic and non-logarithmic categories. Non-logarithmic multipliers usually use approximation techniques to simplify different parts of a typical multiplier such as partial product generation [14, 22] and partial product accumulation [23, 24]. These multipliers have relatively low approximation errors and a more complex hardware instead. Logarithmic multipliers, as their name implies, convert complex multiplication operation into simpler addition operation in the logarithm domain, which results in more compact hardware than non-logarithmic multipliers. Unlike a conventional multiplier, a logarithmic multiplier needs logarithm conversion, addition, and antilogarithm stages. Because of inherent error in logarithm transformation, they are approximate multipliers. There are different ways to convert binary numbers into the logarithmic numbers system: 1) iterative methods, which are very time-consuming and need several cycles to converge to an acceptable result. 2) look-up table-based methods that are accurate but need complex and increased hardware. 3) using a piece-wise linear approximation of the function $\log x$. The third method is high-speed, and implementing this method needs relatively fewer resources. The First logarithmic multiplier, which uses a piece-wise linear approximation, was proposed by Mitchell [15]. Here, the algorithm is briefly expressed. Assume that we want to multiply two fixed-point numbers A and B ; they can be represented in the form $2^{k_{A,B}}(1 + x_{A,B})$ and x_A, x_B are between $[0,1]$. 2^{k_A} and 2^{k_B} are the largest powers of two smaller than or equal to A and B , respectively. It means k_A and k_B represent the position of the most significant one in A and B . Taking the logarithms of A and B , we have $\log_2 A, B = k_{A,B} + \log_2(1 + x_{A,B})$. Mitchell's method to compute this term is to use the approximation $\log_2(1 + x) \approx x$. Thus, the multiplication is simply calculated with only shift and add operations. The problem with Mitchell's algorithm is that this method has a relatively large error and always underestimates the logarithms, so the product is, in any case, smaller than or equal to exact results. The Mitchell's multiplier accuracy improvement methods can be categorized into four main groups [25], as shown in Fig. 1.

Mitchell's method is based on a piece-wise linear approximation in which the lines are in intervals between powers of two. Each line has two intersections with the exact logarithm curve; intersections are at powers of two. In divided approximation methods, a range of Mitchell's algorithms is divided into some more fine-grained intervals, and in each, a more precise equation is derived to approximate the curve better. In [26], the authors proposed implementing a logarithm converter based on Mitchell's method. For

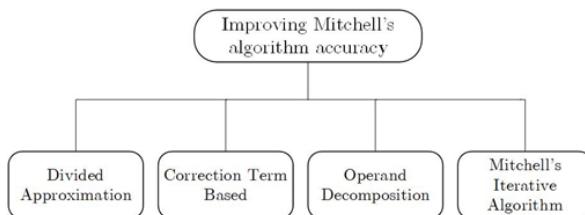


Fig. 1: Different classes of improving Mitchell's accuracy.

accuracy improvement, regions between powers of two are split into some smaller intervals, and within them, a more precise equation is placed instead of Mitchell's original equation for approximating the logarithms. Worst case relative error in original Mitchell's conversion is 5.36%, but in [26], this error is reduced to 0.93, 0.43, and 0.15 percent for 2-region, 3-region, and 6-region correcting algorithms, respectively. For calculating these equations, an error-correcting circuit must be placed in hardware, which means increased hardware. Mahalingam et al. [25] used the operand decomposition technique to reduce the error in Mitchell's multiplier. Operand decomposition was first introduced in [27] to reduce the array's switching activity and the tree multipliers. In [15], Mitchell showed that his algorithm is more accurate when there is no carryover in the mantissa part during the summation step. Operand decomposition reduces the number of "1" bits in the decomposed operands; this means less chance to produce a carryover from the mantissa part into the integer part when summing up the logarithms. It has been shown that operand decomposition reduces Mitchell's multiplication error by 44.7% on average, and to achieve accuracy further, this work can be used with other error reduction methods. The main drawback of this work is its hardware overhead: The multipliers are doubled; moreover, there is a need for a decomposition circuit and an adder to compute the final product.

Correction term-based approaches add a term to the results obtained from the original Mitchell's algorithm to reduce the error. McLaren in [28] showed that multiplication error is only related to the mantissa part (fractional parts x_A and x_B), thus repeating for every characteristic. For error correction, different correction values would be added to the final result based on various combinations of x_A and x_B ; but it is impractical. McLaren split the range of $x \in [0,1]$ into eight regions of 0.125 and made a table of correction values for each combination of these ranges. With this modification mean of the errors reduces from 3.614 to 0.0363. The paper states that their method has increased area and power consumption about 30% over the original Mitchell's algorithm. The first iterative logarithmic multiplier was presented in [15]. The magnitude of error in Mitchell's algorithm is $2^{k_1+k_2}(x_1x_2)$ when there is no carry and $2^{k_1+k_2}(x'_1x'_2)$ when we have carryover from mantissa respectively (x'_1 and x'_2 are the two's complements of x_1 and x_2). Considering x_1x_2 or $x'_1x'_2$ as a new product, if this term is computed with another Mitchell multiplier and this correction value is added with the approximate product computed before, the error reduces significantly at the cost of extra hardware.

Several articles have attempted to optimize Mitchell's multiplier hardware. In [5], three different approximate adders are exploited in the adder stage of a logarithmic

multiplier. They tried various truncation lengths for adders and reported the effects on hardware efficiency and error criteria. The logarithmic multiplier in [17] was improved in different aspects: they used efficient fully parallel leading-one detectors, exploited efficient shift amount calculation, and finally introduced parameter w (the truncation width) and designed a customizable logarithmic multiplier for compromising between hardware costs and accuracy. A modified exact adder is proposed in [16], as in the final addition of the multiplier, some states do not occur; they can use a simplified adder.

The one-sided error distribution of Mitchell's method is another problem that must be considered. In [28], correction terms changed the distribution. About 68% of errors fall in the range -1.21 and 1.29, while errors in the original algorithm are between 0.507 and 6.721. in [5], using an inexact set-one adder causes a somewhat double-sided error distribution. Authors in [29] have proposed a novel logarithm conversion algorithm that differs from Mitchell's. in this algorithm, instead of finding the most significant power of two smaller than the operands, they find the nearest power of two to the operand. This modification leads to a reduced error and a double-sided error distribution, which avoids error accumulation in many applications like matrix multiplication. However, finding the nearest ones needs more complex hardware and leads to dealing with negative numbers and subtractors. In the next section, we will present another way to improve accuracy, which at the same time reduces hardware costs. In [Section 3](#), we discuss selecting the approximation, which keeps the distribution of errors double-sided.

3. PROPOSED METHOD

In this section, our proposed approximate multiplier is introduced. At first, the multiplication algorithm is described, and then multiplier hardware is investigated.

3.1. Multiplication Algorithm

Consider operands A and B that have to be multiplied. We can represent operands as (1):

$$\begin{cases} A = h_1 + q_1 = 2^{k_1} + q_1 \text{ where } 0 \leq q_1 < 2^{k_1} \\ B = h_2 + q_2 = 2^{k_2} + q_2 \text{ where } 0 \leq q_2 < 2^{k_2} \end{cases} \quad (1)$$

Equation (1) shows the operands are decomposed into the largest power of two smaller or equal to them plus an extra term. So, the multiplication becomes from (2) and (3):

$$P_{exact} = A \times B = 2^{k_1+k_2} + 2^{k_1}q_2 + 2^{k_2}q_1 + q_1q_2 \quad (2)$$

$$P_{approx} = 2^{k_1+k_2} + 2^{k_1}q_2 + 2^{k_2}q_1 + q_1q_{2approx} \quad (3)$$

As seen in (2), the first term is 2 to the power of $k_1 + k_2$. Which can be simply computed with a shift operation. In [29], term $2^{k_1+k_2}$ was calculated by giving the summation of k_1 and k_2 to a decoder. However, here we directly shift 2^{k_1} to the left by the amount of k_2 . Two other terms, $2^{k_1}q_2$ and $2^{k_2}q_1$, are products of an arbitrary number and a power of two. To produce these terms, we shift q_1 and q_2 to the left, respectively, by k_2 and k_1 . In order to compute the last term, q_1q_2 , which itself is a product term, we have used approximation. The approximation is as follows: q_1 and q_2 are

approximated to the largest power of two smaller or equal to them, as shown in (4):

$$\begin{cases} q_1 = 2^{m_1}(1 + r_1) = 2^{m_1}x_1 & 0 \leq r_1 < 1 \\ q_2 = 2^{m_2}(1 + r_2) = 2^{m_2}x_2 & 0 \leq r_2 < 1 \end{cases} \quad (4)$$

With the approximation mentioned above, we approximate q_1 and q_2 as equation 4 with $k \in \{1, 2, 4\}$ as (5).

$$q_1 q_2 = k \times 2^{m_1+m_2} \quad (5)$$

Computing this term becomes similar to the calculation of $2^{k_1+k_2}$ which was discussed earlier, and then calculating coefficient k . So, we can obtain this term by shifting 2^{m_1} to the left by m_2 . Thus k is a power of two; the result can be obtained only by shifting $2^{m_1+m_2}$ to the left. The only approximation used in this work is the computation of the $q_1 q_2$ term. The reason why we used such approximation is discussed in the next section. The complete workflow of the proposed multiplier is described in Fig. 2.

3.2. Correction Term Selection

To select the best option for approximating $q_1 q_2$, three different k were candidates, i.e., 1, 2, and 4. This brings us three approximations $2^{m_1+m_2}$, $2^{m_1+m_2+1}$, and $2^{m_1+m_2+2}$. The absolute error for each option is calculated in (8).

$$\text{error} = |P_{\text{exact}} - P_{\text{approx}}| \quad (6)$$

Concerning (2) and (3), the equation (6) becomes:

$$\text{error} = |q_1 q_2 - q_1 q_2_{\text{approx}}| \quad (7)$$

$$\Rightarrow \begin{cases} \text{error}_1 = |2^{m_1+m_2}x_1x_2 - 2^{m_1+m_2}| & (k = 1) \\ \text{error}_2 = |2^{m_1+m_2}x_1x_2 - 2^{m_1+m_2+1}| & (k = 2) \\ \text{error}_3 = |2^{m_1+m_2}x_1x_2 - 2^{m_1+m_2+2}| & (k = 4) \end{cases} \quad (8)$$

To select the best option, we decided to pick the k , which in most cases gives us the least error. To do so, two conditions were examined, and solving these inequalities leads to the following circumscriptions (9), (10):

$$\text{error}_1 < \text{error}_2 \Rightarrow x_1 x_2 < 1.5, \quad 0 \leq x_1, x_2 < 1 \quad (9)$$

$$\text{error}_2 < \text{error}_3 \Rightarrow x_1 x_2 < 3, \quad 0 \leq x_1, x_2 < 1 \quad (10)$$

$x_1 x_2$ product is plotted in Fig. 3, and red and blue lines show the borders where $x_1 x_2$ is 1.5 and 3 respectively. This plot clearly shows that in most cases (about 68%), $k = 2$ i.e., $2^{m_1+m_2+1}$ approximation for $q_1 q_2$ has the minimum error, so we selected it for our design.

3.3. Hardware Architecture

The hardware implementation of our proposed multiplier is described. The multiplier block diagram is shown in Fig. 4. LOD units are leading-one-detectors, which their structure is taken from [30]. LOD finds the most significant 1 in its input and keeps it in output while making other bits zero. Priority encoder (PE) determines the position of the most valuable 1 in number. It also has a zero flag, which becomes high in the case of zero input. Shifter blocks are combinational barrel shifters, and their architecture is the same as shifters proposed in [31].

Algorithm 1 Proposed Approximate Multiplier

Inputs: A, B : n -bits

Output: P : $2n$ -bits

```

// Find the leading one in A and B:
1:  $h_1 \leftarrow \text{LOD}(A)$ 
2:  $h_2 \leftarrow \text{LOD}(B)$ 

// Find  $q_1$  and  $q_2$ :
3:  $q_1 \leftarrow A[n-2:0] \text{ XOR } h_1$ 
4:  $q_2 \leftarrow B[n-2:0] \text{ XOR } h_2$ 

// Find the position of the leading one:
5:  $k_1 \leftarrow \text{PE}(h_1)$ 
6:  $k_2 \leftarrow \text{PE}(h_2)$ 

// Calculate  $2^{k_1+k_2}$ 
7:  $2^{k_1+k_2} \leftarrow 2^{k_1} \ll k_2$ 
8:  $T_1 = 2^{k_1} q_2 \leftarrow q_2 \ll k_1$ 
9:  $T_2 = 2^{k_2} q_1 \leftarrow q_1 \ll k_2$ 
10:  $m_1 \leftarrow \text{PE}(q_1)$ 
11:  $2^{m_2} \leftarrow \text{LOD}(q_2)$ 
12:  $2^{m_1+m_2} \leftarrow 2^{m_2} \ll m_1$ 
13:  $2^{m_1+m_2+1} = 2^{m_1+m_2} \& '0'$ 
14:  $T_3 = 2^{k_1+k_2} + 2^{m_1+m_2+1} \leftarrow 2^{k_1+k_2} \text{ OR } 2^{m_1+m_2+1}$ 
15:  $P \leftarrow T_1 + T_2 + T_3$ 
```

Fig. 2: Proposed multiplication algorithm.

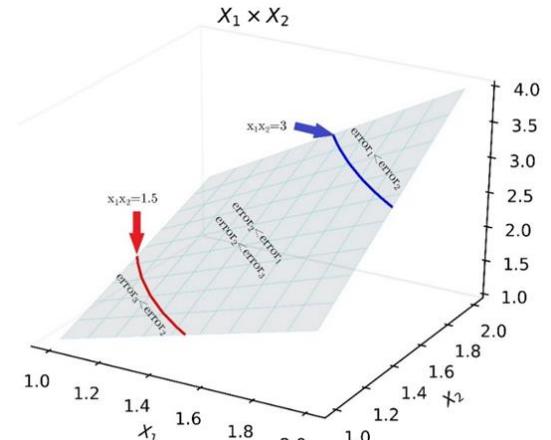


Fig. 3: x_1 and x_2 product plot separated by 1.5 and 3 lines.

Operands A and B are given to LOD1 and LOD2 as inputs. The PE1 and PE2 take the LOD1 and LOD2 outputs, which are in one hot representation format, and calculate k_1 and k_2 . q_1 and q_2 are computed by XORing h_1 and h_2 with A and B. We have used a novel approach to calculate term $2^{k_1+k_2}$. In [29], the authors have used an adder and a decoder after PEs to find this term's value. However, in this paper, the adder and the decoder are eliminated. Instead, we have placed a shifter after LOD1, and the shift amount comes from PE2, and PE1 is no longer on its path. With these modifications, it seems the level of logic and hardware area must decrease to some extent. Shifter2 and shifter3 are responsible for calculating terms $2^{k_1}q_2$ and $2^{k_2}q_1$, respectively. To approximate $q_1 q_2$, first, q_1 is given to PE3 to obtain m_1 . Note that there is no need for LOD because the PE itself finds the position of most significant '1'. On the other hand, q_2 transfers through LOD3, and 2^{m_2} is computed. With the use of shifter4, we calculate $2^{m_1+m_2}$. In this paper, to reduce the mean error, we used approximation $2^{m_1+m_2+1}$. We reach this term easily by concatenating a '0' on the right side of $2^{m_1+m_2}$.

The four terms calculated before must be added to produce the final result. To reduce the complexity of the adder

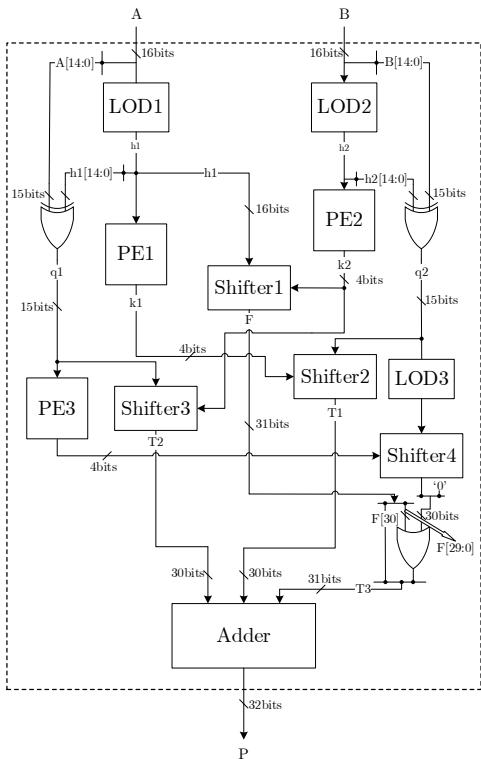


Fig. 4: Proposed approximate multiplier.

stage, we consider two terms $2^{k_1+k_2}$ and $2^{m_1+m_2+1}$. Thus $2^{k_1+k_2}$ is a power of 2 and is always greater than $2^{m_1+m_2+1}$; we can OR them to find the addition. Lastly, the three terms are summed up in an adder, and the final product is obtained.

3.4. Exploiting Approximate Adders

A large part of the logarithmic multiplier is devoted to adders. This urged us to investigate the use of approximate adders in our design. The utilization of approximate adders in logarithmic multipliers has been studied in [5]. Some types of approximate adders were exploited, and the results showed that set-one adders outperform other types. In [29], a modified version of set-one adders was introduced and employed in their logarithmic multiplier. Therefore, we bring our attention to set-one adders and explore the performance of our proposed multiplier with them. An n-bit set-one adder with m truncated bits (SOA-m) is composed of an m-bit approximate part for the least significant bits (lower part of the augend and addend) and an exact part for (n-m) most significant bits. N-bit SOA-m is depicted in Fig. 5. The expressions (11) and (12) describe lower m bit of a set-one adder:

$$\text{sum}[m-1:0] = 1 \quad (11)$$

$$c_{in} = a[m-1] \text{ AND } b[m-1] \quad (12)$$

The adder used in our design sums up three terms to build the product. It is composed of a set of 3 to 2 compressors (full adder units) followed by a ripple carry adder. As the bit width of the result is twice the inputs in multiplication, the adder is costly in terms of power and area, and the carry chain causes a relatively high delay. Set-one adder can alleviate hardware overhead because there are no logical circuits for calculating the 'm' right-hand bits of the result. So, there is a reduction of 2m full adders in our design (m full adders in compressor

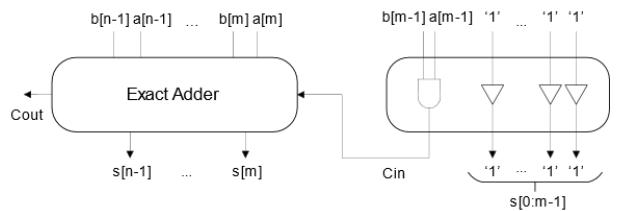


Fig. 5: SOA with m truncation bits.

stage plus m full adders in the ripple carry adder). The delay also significantly improves since SOA shrinks lengthy carry chain by m bits. The effect of approximation bit numbers in the final adder on the accuracy of the proposed multiplier is investigated to pick the best value of m. MRED criterion is chosen for this purpose. As presented in Fig. 6, for a 16-bit multiplier, selecting m to values up to 16 nearly has no impact on our multiplier accuracy. Therefore, we chose 16 for maximum hardware saving. This means our design is more robust than previous works in [5] and [29], in which the MRED started to immediately increase when m was larger than 11 and 15, respectively.

4. EXPERIMENTAL RESULTS

We evaluate our work and compare it with some similar works available in the literature in this section. Prior to experimental results, error metrics for approximate designs are introduced. These criteria are measured for our proposed algorithm. Hardware simulations are done, and hardware metrics such as area, power, and delay are assessed. For evaluation, two 16×16 multipliers with both exact and approximate adders have been considered.

4.1. Accuracy Evaluation

To assess the accuracy and error characteristics of the multiplier, the multiplication algorithm is implemented in behavioral level. Because exhaustive simulations are time-consuming, 10^7 pairs of random inputs were given to the model, and results were obtained. Error metrics, including error rate (ER), mean related error distance (MRED), and normalized error distance (NMED), are calculated. For comparison, multiplier designs available in papers [29], [15], and [5] are considered, and the results are listed in Table 1. As expected, simulation results verify that our multiplier outperforms in terms of accuracy metrics. Because of an additional correction term, the MRED and NMED measures of the proposed algorithm are lower than LM [15] and ALM-SOA [5]. Although nearest-one detectors in [29] are removed, a good selection of correction terms can compensate for the effects, and even results show a reduction in mean relative error to about 25.6% than the best available work [29]. It is evident from Table 1 that using set-one adders in our multiplier does not affect the accuracy severely, and even with a high number of truncated bits (m), accuracy metrics stay about their values with an exact adder.

4.2. Hardware Evaluation

We coded a 16-bit multiplier for hardware assessment based on the proposed algorithm. All hardware simulations were done in Synopsys Design Compiler with default settings, using TSMC 180nm technology. Table 2 presents

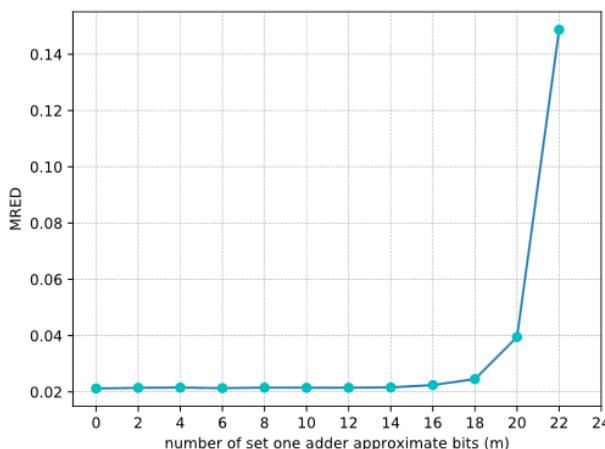


Fig. 6: effect of the number of approximate bits in adder stage on the accuracy of the 16-bit multiplier.

obtained results from simulations. Compared to the proposed multiplier in [29], we removed costly subtractors and nearest-one detectors and replaced them with an array of XORs and leading-one detectors, respectively, expecting a power and area reduction in our design. Results confirmed that our modifications caused a meaningful improvement in both area and power consumption.

The critical path in [29] is related to the path where the term $2^{k_1+k_2}$ is calculated. This path goes through an adder for computing $k_1 + k_2$ and a priority encoder for calculating $2^{k_1+k_2}$. To reduce the delay, we proposed a novel way for $2^{k_1+k_2}$ computation, in which the adder and decoder are replaced with a shifter. The results show a 24% reduction in critical path delay with respect to ILM-EA. As discussed in Section 3, our algorithm has more persistence to approximation in its final adder so that we can exploit this characteristic for further hardware improvements. Simulation results show that by setting truncation bits (m) to 16, we can achieve significant hardware savings and reduce the large carry chain of the final adder to half its length.

To decide which multiplier design is preferable overall, i.e., both accuracy and hardware metrics, we compared PDP×MRED of the multipliers. The results are presented in Table 3.

Multiplier designs with lower MRED and PDP and, as a result, with lower PDP×MRED are more favorable. As seen from Table 3, our proposed multiplier with approximate adder has the least PDP-MRED product, and therefore it is the most hardware-efficient design over others while considering accuracy.

5. JPEG APPLICATION

We employed our work in the real-world application JPEG, an image compression standard [32]. To show our proposed approximate multiplier's applicability. Image compression in jpeg is as follows: the image is first partitioned in 8×8 blocks. Then Discrete Cosine Transform

Table 1: Error metrics.

| Multiplier | MRED | ER (%) | NMED |
|----------------------------|--------|--------|--------|
| LM [15] | 0.0384 | 99.77 | 0.0092 |
| ILM-EA [29] | 0.0289 | 99.95 | 0.0069 |
| ALM-SOA-11 [5] | 0.0330 | 98.97 | 0.0080 |
| Proposed (Exact Adder) | 0.0215 | 99.95 | 0.0064 |
| Proposed (SOA-16 Adder) | 0.0228 | 99.99 | 0.0064 |

Table 2: Hardware metrics.

| Multiplier | Power (mW) | Delay (nS) | Area (μm^2) | PDP (pJ) |
|----------------------------|---------------|---------------|-----------------------------|-------------|
| LM [15] | 6.00 | 34.94 | 146338 | 209.64 |
| ILM-EA [29] | 8.85 | 34.49 | 158629 | 305.23 |
| ALM-SOA-11 [5] | 4.29 | 23.50 | 124871 | 100.815 |
| Proposed (Exact Adder) | 5.31 | 29.29 | 139595 | 155.53 |
| Proposed (SOA-16 Adder) | 4.96 | 20.68 | 122214 | 102.57 |

Table 3: PDP×MRED of approximate multipliers.

| Multiplier | PDP×MRED |
|----------------------------|----------|
| LM [15] | 8.05 |
| ILM-EA [29] | 8.82 |
| ALM-SOA-11 [5] | 3.32 |
| Proposed (Exact Adder) | 3.34 |
| Proposed (SOA-16 Adder) | 2.33 |

(DCT) is calculated for each block, and after that, the quantization step is done. This step is attained by dividing the matrix of DCT coefficients by the quantization matrix in an element-wise fashion and rounding the results. Then an entropy coding is applied to the resulted matrix to reduce the image size. Image decompression starts by decoding the data and then dequantizing the blocks. Dequantizing is done by multiplying the matrix of quantization into each block. This step is where we have exploited our multiplier. After dequantization, the inverse of DCT (IDCT) is computed, and the image is formed. For evaluating the applicability, we coded the lossy JPEG standard in MATLAB. We then implemented some approximate multipliers, including our design, in the dequantization part of the JPEG standard. Two measures Peak Signal to Noise Ratio (PSNR) and structural similarity (SSIM), are used to compare and inspect the applicability of multipliers. Both PSNR and SSIM are widely used in image processing; they assess the quality of a compressed image. The higher the PSNR, the better the quality of the compressed or reconstructed image. Simulation results in Table 4 show that using the approximate multipliers does not significantly affect the decompressed image's quality. As expected, our multiplier has the least quality reduction in output image due to its lower MRED.

Table 4: PSNR and SSIM values for decompressed images.

| Multiplier | PSNR | SSIM |
|---------------------------|---------|--------|
| Exact | 35.1281 | 0.9095 |
| LM [15] | 30.2662 | 0.9001 |
| ILM-EA [29] | 31.9424 | 0.8953 |
| ALM-SOA-11 [5] | 30.2744 | 0.8759 |
| Proposed (Exact Adder) | 32.9800 | 0.9037 |

6. CONCLUSION

In this paper, a new algorithm for logarithmic multiplication is proposed and analysed. The use of approximate adders (SOA) in the final stage of multiplication is also investigated. 16-bit multipliers were implemented using this algorithm, and the simulation results on showed that by using the appropriate correction term, the multiplier accuracy is significantly improved compared to previous similar works, so that MRED has decreased by about 25.6% compared to ILM-EA. analysing change of MRED with respect to truncation width of SOA showed that our design is more robust to adder truncation than previous designs, so that at ALM-SOA-11 and ILM-EA the best truncation width is 11 bits, but MRED in our design does not change much up to 16 bits. Which can be exploited for more hardware savings. Hardware synthesis also show an improvement of 2.12% and 12% in area, and latency respectively and a 13.5% increase in power consumption compared to best results available in the literature. The PDP×MRED criterion also shows that our multiplier shows the best performance among the existing designs by considering both error characteristics and hardware measures. Finally, we implemented our multiplier in JPEG standard, and results showed that our design is applicable in such error-tolerant applications without notable quality degradation.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Mojtaba Arab Nezhad: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Resources, Software, Visualization, Writing - original draft. **Ali Mahani:** Investigation, Project administration, Supervision, Validation, Writing - review & editing.

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. The ethical issues; including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy has been completely observed by the authors.

REFERENCES

- [1] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys (CSUR)*, vol. 48, pp. 1-33, 2016.
- [2] V. Sze, Y. -H. Chen, T. -J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295-2329, 2017.
- [3] P. Yin, C. Wang, H. Waris, W. Liu, Y. Han, and F. Lombardi, "Design and analysis of energy-efficient dynamic range approximate logarithmic multipliers for machine learning," *IEEE Transactions on Sustainable Computing*, 2020.
- [4] R. Pilipović, P. Bulić, and U. Lotrič, "A two-stage operand trimming approximate logarithmic multiplier," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2021.
- [5] W. Liu, J. Xu, D. Wang, C. Wang, P. Montuschi, and F. Lombardi, "Design and evaluation of approximate logarithmic multipliers for low power error-tolerant applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 9, pp. 2856-2868, 2018.
- [6] A. G. M. Strollo, E. Napoli, D. De Caro, N. Petra, and G. Di Meo, "Comparison and extension of approximate 4-2 compressors for low-power approximate multipliers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 9, pp. 3021-3034, 2020.
- [7] D. Nandan, J. Kanungo, and A. Mahajan, "An efficient VLSI architecture design for logarithmic multiplication by using the improved operand decomposition," *Integration*, vol. 58, pp. 134-141, 2018.
- [8] H. Saadat, H. Bokhari, and S. Parameswaran, "Minimally biased multipliers for approximate integer and floating-point multiplication," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2623-2635, 2018.
- [9] H. Jiang, L. Liu, F. Lombardi, and J. Han, "Approximate arithmetic circuits: Design and evaluation," *Approximate Circuits*, Springer, 2019, pp. 67-98.
- [10] H. Jiang, C. Liu, N. Maheshwari, F. Lombardi, and J. Han, "A comparative evaluation of approximate multipliers," in *2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2016.
- [11] P. Kulkarni, P. Gupta, and M. Ercegovac, "Trading accuracy for power with an underdesigned multiplier architecture," in *2011 24th International Conference on VLSI Design*, 2011.
- [12] S. Hashemi, and S. Reda, "Approximate multipliers and dividers using dynamic bit selection," *Approximate Circuits*, Springer, 2019, pp. 25-44.
- [13] S. Narayananamoorthy, H. A. Moghaddam, Z. Liu, T. Park, and N. S. Kim, "Energy-efficient approximate multiplication for digital signal processing and classification applications," *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 23, no. 6, pp. 1180-1184, 2015.
- [14] W. Liu, L. Qian, C. Wang, H. Jiang, J. Han, and F. Lombardi, "Design of approximate radix-4 booth multipliers for error-tolerant computing," *IEEE Transactions on Computers*, vol. 66, no. 8, pp. 1435-1441, 2017.
- [15] J. N. Mitchell, "Computer multiplication and division using binary logarithms," *IRE Transactions on Electronic Computers*, vol. 4, pp. 512-517, 1962.
- [16] M. S. Ansari, B. F. Cockburn, and J. Han, "An improved logarithmic multiplier for energy-efficient neural computing," *IEEE Transactions on Computers*, vol. 70, no. 4, pp. 614-625, 2021.
- [17] M. S. Kim, A. A. D. Barrio, L. T. Oliveira, R. Hermida, and N. Bagherzadeh, "Efficient Mitchell's approximate log multipliers for convolutional neural networks," *IEEE Transactions on Computers*, vol. 68, no. 5, pp. 660-675, 2019.

- [18] S. Mazahir, M. K. Ayub, O. Hasan, and M. Shafique, "Probabilistic error analysis of approximate adders and multipliers," *Approximate Circuits*, Springer, 2019, pp. 99-120.
- [19] J. Ma, K. Man, T. Krilavicius, S. Guan, and T. Jeong, "Implementation of high-performance multipliers based on approximate compressor design," in *Proc. Int. Conf. Electrical and Control Technol.*, 2011.
- [20] A. Momeni, J. Han, P. Montuschi, and F. Lombardi, "Design and analysis of approximate compressors for multiplication," *IEEE Transactions on Computers*, vol. 64, no. 4, pp. 984-994, 2014.
- [21] S. Hashemi, R. I. Bahar, and S. Reda, "DRUM: A dynamic range unbiased multiplier for approximate applications," in *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2015.
- [22] V. Leon, G. Zervakis, D. Soudris, and K. Pekmestzi, "Approximate hybrid high radix encoding for energy-efficient inexact multipliers," *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 26, no. 3, pp. 421-430, 2017.
- [23] M. Ha, and S. Lee, "Multipliers with approximate 4-2 compressors and error recovery modules," *IEEE Embedded Systems Letters*, vol. 10, no. 1, pp. 6-9, 2017.
- [24] D. Esposito, A. G. M. Strollo, E. Napoli, D. De Caro, and N. Petra, "Approximate multipliers based on new approximate compressors," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4169-4182, 2018.
- [25] V. Mahalingam, and N. Ranganathan, "An efficient and accurate logarithmic multiplier based on operand decomposition," in *19th International Conference on VLSI Design held jointly with 5th International Conference on Embedded Systems Design (VLSID'06)*, 2006.
- [26] K. H. Abed, and R. E. Siferd, "CMOS VLSI implementation of a low-power logarithmic converter," *IEEE Transactions on Computers*, vol. 52, no. 11, pp. 1421-1433, 2003.
- [27] M. Ito, D. Chinnery, and K. Keutzer, "Low power multiplication algorithm for switching activity reduction through operand decomposition," in *Proceedings 21st International Conference on Computer Design*, 2003.
- [28] D. J. McLaren, "Improved Mitchell-based logarithmic multiplier for low-power DSP applications," in *IEEE International [Systems-on-Chip] SOC Conference, 2003. Proceedings*, 2003.
- [29] M. S. Ansari, B. F. Cockburn, and J. Han, "A Hardware-Efficient Logarithmic Multiplier with Improved Accuracy," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2019.
- [30] D. Nandan, J. Kanungo, and A. Mahajan, "An efficient architecture of leading one detector," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 14, pp. 267-272, 2018.
- [31] J. Chen, C.-H. Chang, Y. Wang, J. Zhao and S. Rahardja, "New hardware and power efficient sporadic logarithmic shifters for DSP applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 4, pp. 896-900, 2017.
- [32] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii-xxxiv, 1992.
- [33] H. Jiang, C. Liu, L. Liu, F. Lombardi and J. Han, "A review, classification, and comparative evaluation of approximate arithmetic circuits," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 4, pp. 1-34, 2017.
- [34] T. Stouraitis and V. Palioras, "Considering the alternatives in low-power design," *IEEE Circuits and Devices Magazine*, vol. 17, no. 4, pp. 22-29, 2001.
- [35] C. Basetas, I. Kouretas and V. Palioras, "Low-power digital filtering based on the logarithmic number system," in *International Workshop on Power and Timing Modeling, Optimization and Simulation*, Springer, 2007, pp. 546-555.

BIOGRAPHY

Mojtaba Arab Nezhad received his B.S. degree in electrical engineering from Shahid Bahonar University, Kerman, Iran in 2016. He is currently working toward his M.S. degree at Shahid Bahonar University, Kerman, Iran. His research interests include computer arithmetic, approximate computing, FPGA based accelerators and accelerating deep learning algorithms.



Ali Mahani received the B. Sc. degree in electronic engineering from Shahid Bahonar University of Kerman, Iran, in 2001, The M.Sc. and Ph.D. degrees both in Electronic engineering from Iran University of Science and Technology (IUST), Tehran, Iran, in 2003 and 2009 respectively. Since then, he has been with the electrical engineering department of shahid bahonar university of kerman, where he is currently an associate professor. His research interests focus on Fault tolerant design, FPGA-based accelerators, approximate digital circuits, stochastic computing and networked systems.



Copyrights

© 2021 Licensee Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (<http://creativecommons.org/licenses/by-nc/4.0/>).



Journal of Applied Research in Electrical Engineering

Vol. 2, No. 1, Winter and Spring 2023

jaree.scu.ac.ir

| | |
|--|----|
| Analyzing the Inference Process in Deep Convolutional Neural Networks using Principal Eigenfeatures, Saturation and Logistic Regression Probes Mats Leon Richter, Leila Malihi, Anne-Kathrin Patricia Windler, and Ulf Krumnack | 1 |
| Improving Stochastic Computing Fault-Tolerance: A Case Study on Discrete Wavelet Transform Shabnam Sadeghi, and Ali Mahani | 11 |
| Investigation of the Operation of Active Superconducting Fault Current Limiters in Distribution Networks Connected to Microgrids Ahmad Ghafari, Mohsen Saniei, Morteza Razzaz, and Alireza Saffarian | 19 |
| Multi-Objective Optimal Power Flow Based Combined Non-Convex Economic Dispatch with Valve-Point Effects and Emission Using Gravitation Search Algorithm Nabil Mezhoud, and Mohamed Amarouayache | 26 |
| Smart AI-based Video Encoding for Fixed Background Video Streaming Applications Mohammadreza Ghafari, Abdollah Amirkhani, Elyas Rashno, and Shirin Ghanbari | 37 |
| Improving the Quality of ECG Signal Using Wavelet Transform and Adaptive Filters Amir Hatamian, Farzad Farshidi, Changiz Ghobadi, Javad Nourinia, and Ehsan Mostafapour | 45 |
| Effect of Changes in the Parameters of a Modular Converter in Its Controllability Range in Fuel Cell Applications Mohammad Afkar, Parham Karimi, Roghayeh Gavagsaz-Ghoachani, Matheepot Phattanasak, and Serge Pierfederici | 54 |
| Robustness Analysis of Model Reference Adaptive Controller in The Presence of Input Saturation Using Describing Function Method Fatemeh Tavakkoli, Alireza Khosravi, and Pouria Sarhadi | 62 |
| Investigating the Effect of Geometric Design Parameters on the Mutual Inductance Between Two Similar Planar Spiral Coils With Inner and Outer Diameter Limits Ata Ollah Mirzaei, Amir Musa Abazari, and Hadi Tavakkoli | 70 |
| Partial Discharge Pattern Recognition in GIS Using External UHF Sensor Reza Rostaminia, Mehdi Vakilian, and Keyvan Firouzi | 75 |
| A Feedforward Active Gate Voltage Control Method for SiC MOSFET Driving Hamidreza Ghorbani, and Jose Luis Romeral Martinez | 87 |
| Design of Low-Power Approximate Logarithmic Multipliers with Improved Accuracy Mojtaba Arab Nezhad, and Ali Mahani | 95 |



Iranian Association of Electrical
and Electronics Engineers



Shahid Chamran
University of Ahvaz