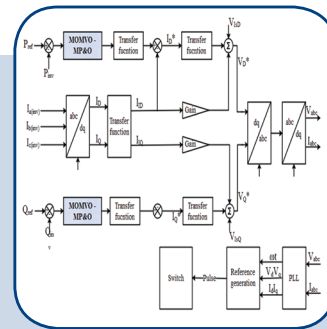
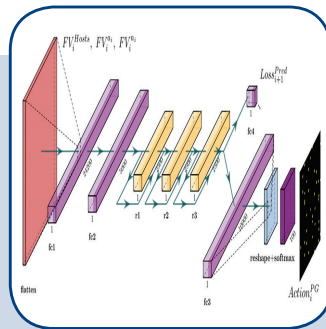
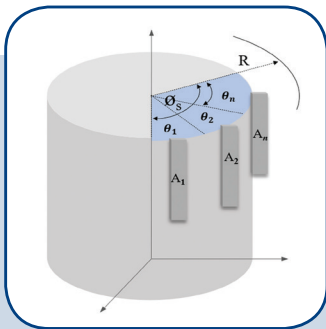
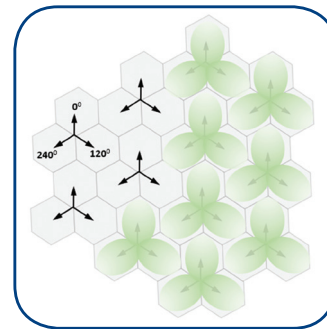
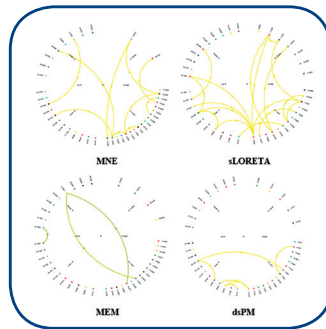


International Journal of Electrical and Computer Engineering Systems



INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia

Osijek, Croatia | Volume 15, Number 10, 2024 | Pages 819 - 904

The International Journal of Electrical and Computer Engineering Systems is published with the financial support
of the Ministry of Science and Education of the Republic of Croatia

CONTACT

**International Journal of Electrical
and Computer Engineering Systems
(IJECES)**

Faculty of Electrical Engineering, Computer
Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b, 31000 Osijek, Croatia
Phone: +38531224600, Fax: +38531224605
e-mail: ijeces@ferit.hr

Subscription Information

The annual subscription rate is 50€ for individuals,
25€ for students and 150€ for libraries.
Giro account: 2390001 - 1100016777,
Croatian Postal Bank

EDITOR-IN-CHIEF

Tomislav Matić
J.J. Strossmayer University of Osijek,
Croatia

Goran Martinović
J.J. Strossmayer University of Osijek,
Croatia

EXECUTIVE EDITOR

Mario Vranješ
J.J. Strossmayer University of Osijek, Croatia

ASSOCIATE EDITORS

Krešimir Fekete
J.J. Strossmayer University of Osijek, Croatia

Damir Filko
J.J. Strossmayer University of Osijek, Croatia

Davor Vinko
J.J. Strossmayer University of Osijek, Croatia

EDITORIAL BOARD

Marinko Barukčić
J.J. Strossmayer University of Osijek, Croatia

Tin Benšić
J.J. Strossmayer University of Osijek, Croatia

Matjaz Colnarič
University of Maribor, Slovenia

Aura Conci
Fluminense Federal University, Brazil

Bojan Čukić
University of North Carolina at Charlotte, USA

Radu Dobrin
Mälardalen University, Sweden

Irena Galić
J.J. Strossmayer University of Osijek, Croatia

Ratko Grbić
J.J. Strossmayer University of Osijek, Croatia

Krešimir Grgić
J.J. Strossmayer University of Osijek, Croatia

Marijan Herceg
J.J. Strossmayer University of Osijek, Croatia

Darko Huljenić
Ericsson Nikola Tesla, Croatia

Željko Hocenski
J.J. Strossmayer University of Osijek, Croatia

Gordan Ježić
University of Zagreb, Croatia

Ivan Kaštelan
University of Novi Sad, Serbia

Ivan Maršić
Rutgers, The State University of New Jersey, USA

Kruno Miličević
J.J. Strossmayer University of Osijek, Croatia

Gaurav Morghare
Oriental Institute of Science and Technology,
Bhopal, India

Srete Nikolovski
J.J. Strossmayer University of Osijek, Croatia

Davor Pavuna
Swiss Federal Institute of Technology Lausanne,
Switzerland

Marjan Popov
Delft University, Nizozemska

Sasikumar Punnekkat
Mälardalen University, Sweden

Chiara Ravasio
University of Bergamo, Italija

Snježana Rimac-Drlje
J.J. Strossmayer University of Osijek, Croatia

Krešimir Romić
J.J. Strossmayer University of Osijek, Croatia

Gregor Rozinaj
Slovak University of Technology, Slovakia

Imre Rudas
Budapest Tech, Hungary

Dragan Samardžija
Nokia Bell Labs, USA

Cristina Seceleanu
Mälardalen University, Sweden

Wei Siang Hoh
Universiti Malaysia Pahang, Malaysia

Marinko Stojkov
University of Slavonski Brod, Croatia

Kannadhasan Suriyan
Cheran College of Engineering, India

Zdenko Šimić
The Paul Scherrer Institute, Switzerland

Nikola Teslić
University of Novi Sad, Serbia

Jami Venkata Suman
GMR Institute of Technology, India

Domen Verber
University of Maribor, Slovenia

Denis Vranješ
J.J. Strossmayer University of Osijek, Croatia

Bruno Zorić
J.J. Strossmayer University of Osijek, Croatia

Drago Žagar
J.J. Strossmayer University of Osijek, Croatia

Matej Žnidarec
J.J. Strossmayer University of Osijek, Croatia

Proofreader

Ivanka Ferčec
J.J. Strossmayer University of Osijek, Croatia

Editing and technical assistance

Davor Vrandečić
J.J. Strossmayer University of Osijek, Croatia

Stephen Ward
J.J. Strossmayer University of Osijek, Croatia

Dražen Bajer
J.J. Strossmayer University of Osijek, Croatia

Journal is referred in:

- Scopus
- Web of Science Core Collection
(Emerging Sources Citation Index - ESCI)
- Google Scholar
- CiteFactor
- Genamics
- Hrčak
- Ulrichweb
- Reaxys
- Embase
- Engineering Village

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003
Published: quarterly
Circulation: 300

IJECES online
<https://ijeces.ferit.hr>

Copyright

Authors of the International Journal of Electrical
and Computer Engineering Systems must transfer
copyright to the publisher in written form.

TABLE OF CONTENTS

Implementation of Cyber Network’s Attacks Detection System with Deep Learning Designing Algorithms	819
<i>Original Scientific Paper</i>	
Lubna Emad Kadhim Saif Aamer Fadhil Sumaia M. Al-Ghuribi Amjed Abbas Ahmed Mohammad Kamrul Hasan Shahrul A. Mohd Noah Fatima N. AL-Aswadi	
Mask FORD-NET: Efficient Detection of Digital Image Forgery using Hybrid REG-NET based Mask-RCNN	829
<i>Original Scientific Paper</i>	
Priscilla Whitin S. Sivakumar M. Geetha M. Devaki A. Bhuvanesh Kiruthiga Balasubramaniyan A. Ahilan	
Deep Reinforcement Learning for Dynamic Task Scheduling in Edge-Cloud Environments	837
<i>Original Scientific Paper</i>	
D. Mamatha Rani Supreethi K.P. Bipin Bihari Jayasingh	
Scaling and Dynamic Resource Reallocation in NFV: Challenges and Research Perspectives	851
<i>Review Paper</i>	
Tung Thanh Hoang Linh Manh Pham Hoai Son Nguyen	
EMF Exposure Reduction Using Weighted Angle Model for Multi-Technology Sectorized BS	865
<i>Original Scientific Paper</i>	
Mohammed S. Elbasheir Rashid A. Saeed Salaheldin Edam	
Comparison Between Different Source Localization and Connectivity Metrics of Spiky and Oscillatory MEG Activities.....	875
<i>Original Scientific Paper</i>	
Ichrak ELBehy Abir Hadriche Rahma Maalej Nawel Jmail	
An Enhancement of Grid Integration in Renewable Energy Systems Using Multi-Objective Multi-Verse Optimization	885
<i>Original Scientific Paper</i>	
Bharathi R.B. Vijaya Laxmi Shashank Bhat	
A New Proposed Triple Active Bridge Converter for Fuel Cell Applications: Study, Control and Energy Management	897
<i>Case Study</i>	
Abdelkarim Aouiti Mokhtar Abbassi Faouzi Bacha	
About this Journal	
IJECEs Copyright Transfer Form	

Implementation of Cyber Network's Attacks Detection System with Deep Learning Designing Algorithms

Original Scientific Paper

Lubna Emad Kadhim

University of Imam Al-Kadhum,
College of Imam Al-Kadhum (IKC), Department of
Computer Techniques Engineering
10011, Baghdad, Iraq
lubnaemad@alkadhum-col.edu.iq

Saif Amer Fadhil

University of Imam Al-Kadhum,
College of Imam Al-Kadhum (IKC), Department of
Computer Techniques Engineering
10011, Baghdad, Iraq
saifaamer@alkadhum-col.edu.iq

Sumaia M. Al-Ghuribi

Prince Sattam bin Abdulaziz University,
Faculty of Computer Engineering & Sciences,
Department of Software Engineering
Alkharj 11942, Riyadh, Saudi Arabia
Taiz University, Faculty of Applied Sciences,
Department of Computer Science, Taiz, Yemen
s.alghuribi@psau.edu.sa

Amjed Abbas Ahmed*

Universiti Kebangsaan Malaysia (UKM),
Faculty of Information Science and Technology,
Center for Cyber Security, 43600, Bangi, Malaysia
University of Imam Al-Kadhum,
College of Imam Al-Kadhum (IKC),
Department of Computer Techniques Engineering
10011, Baghdad, Iraq
amjedabbas@alkadhum-col.edu.iq

*Corresponding author

Mohammad Kamrul Hasan

Universiti Kebangsaan Malaysia (UKM),
Faculty of Information Science and Technology,
Center for Cyber Security, 43600, Bangi, Malaysia
mkh@ukm.edu.my

Shahrul A. Mohd Noah

Universiti Kebangsaan Malaysia (UKM),
Faculty of Information Science and Technology,
Centre for Artificial Intelligence Technology (CAIT)
43600, Bangi, Malaysia
shahrul@ukm.edu.my

Fatima N. AL-Aswadi

UCSI University,
Institute of Computer Science and Digital Innovation
56000, Kuala Lumpur, Malaysia
Hodeidah University,
Faculty of Computer Science and Engineering
Al Hudaydah, Yemen
Fatima.Nadeem@ucsiuniversity.edu.my

Abstract – The internet has become indispensable for modern communication, playing a vital role in the development of smart cities and communities. However, its effectiveness is contingent upon its security and resilience against interruptions. Intrusions, defined as unauthorized activities that compromise system integrity, pose a significant threat. These intrusions can be broadly categorized into host intrusions, which involve unauthorized access and manipulation of data within a system, and network intrusions, which target vulnerabilities within the network infrastructure. To mitigate these threats, system administrators rely on Network Intrusion Detection Systems (NIDS) to identify and respond to security breaches. However, designing an effective and adaptable NIDS capable of handling novel and evolving attack strategies presents a significant challenge. This paper proposes a deep learning-based approach for NIDS development, leveraging Self-Taught Learning (STL) and the NSL-KDD benchmark dataset for network intrusion detection. The proposed approach is evaluated using established metrics, including accuracy, F-measure, recall, and precision. Experimental results demonstrate the effectiveness of STL in the 5-class categorization, achieving an accuracy of 79.10% and an F-measure of 75.76%. This performance surpasses that of Softmax Regression (SMR), which attained 75.23% accuracy and a 72.14% F-measure. The paper concludes by comparing the proposed approach's performance with existing state-of-the-art methods.

Keywords: cyber network, deep learning, intrusion detection system, network intrusion

Received: May 18, 2024; Received in revised form: July 23, 2024; Accepted: July 24, 2024

1. INTRODUCTION

An intrusion detection system (IDS) [1] analyzes and monitors an organization's network devices [2]. If an invasion is identified, this system notifies users and stops additional harm. The two types of intrusion sensing systems are anomaly-dependent network intrusion detection systems (ANIDS) [3] and signature-dependent network intrusion detection systems. SNIDS [4], such as Snort, come pre-configured with attack signatures. To detect a compromise in network security, traffic is compared to the signatures that have been applied. If an ADNIDS identifies a deviation from normal traffic patterns, the observed network traffic is labeled as an intrusion. SNIDS are useful for detecting known threats since they have a high detection accuracy and a low false alarm rate. Due to the limited number of attack signatures that can be pre-installed, it is difficult for an IDS to recognize new or inventive attacks [5-9].

ADNIDS, on the other hand, are very successful in locating previously undiscovered, unique threats. Even though they have a higher rate of false positives, ADNIDS have gained considerable recognition in the research community due to their potential for identifying new attacks. Two impediments hinder the development of effective and adaptable NIDS capable of protecting against unknown future threats. Firstly, the sheer volume of information available makes it challenging to select suitable criteria for identifying anomalies within network data. Since attack methods constantly change and evolve, traits effective for one type of attack may be ineffective for another. Currently, insufficient labeled traffic data from real networks is available for developing effective NIDS. Creating such a labeled dataset from raw, real-time network traffic traces requires significant effort and time. Network managers are very cautious about disclosing security breaches within their networks, as they strive to protect both individual user privacy and the organization's trade secrets related to internal network structure [5]. NIDS are designed to distinguish between normal and anomalous traffic patterns.

Many NIDS employ feature selection to achieve more accurate classifications, which involves selecting a subset of meaningful features from recorded traffic. Feature selection can help reduce the risk of overfitting during training by removing irrelevant features and noise [6]. Comprehending sounds, images, and speech using deep learning algorithms is a relatively recent development. These approaches enable the construction of effective feature representations from large amounts of unlabeled data, which can then be applied to smaller, labeled datasets for supervised classification. Data from both labeled and unlabeled distributions may originate from various sources, but they should ideally be related [7].

Deep learning approaches were expected to address the challenges of building effective NIDS [8]. It is fea-

sible to gather unlabeled network traffic data from various sources across the network and then apply deep learning methods to extract useful feature representations. These features can then be used for supervised classification on a smaller, labeled dataset containing both normal and anomalous traffic. Such a dataset could be used to analyze traffic trends. Collecting labeled traffic data is possible within a controlled, secure, and isolated network environment [10].

Self-taught learning (STL) is a powerful approach for Network Intrusion Detection Systems (NIDS), providing a robust mechanism for detecting anomalous patterns in network traffic. STL typically utilizes large amounts of unlabeled network data to train deep learning models without requiring annotations. This data, often unstructured, undergoes feature extraction, where important characteristics such as packet size, protocol type, and source/destination IP addresses are extracted. A deep learning model can then leverage unsupervised learning, often through an autoencoder architecture, for pretraining. During pretraining, the model aims to minimize the difference between the original input and its reconstruction, effectively learning to represent normal network behavior. This process allows the model to encode normal network behavior. During deployment, deviations from this learned behavior can indicate anomalies, such as intrusions or attacks.

Once deployed, the trained model continuously monitors all incoming network traffic. Leveraging the learned representations, the model can effectively identify anomalies based on emerging patterns in real-time. This proactive approach enables network administrators to preempt or promptly detect threats before they escalate into more serious security breaches. Moreover, the model can adapt to new threats as it continuously learns from new data, refining its understanding of normal network behavior. Therefore, STL, particularly when applied to deep learning models for NIDS, offers a promising approach to strengthening cybersecurity and defending against evolving network threats.

The following are our contribution towards this research work:

- To achieve this goal, we propose a novel deep learning approach for NIDS based on self-taught learning, utilizing sparse autoencoders and softmax regression.
- This approach enabled the development of our proposed NIDS. We evaluate its performance on the NSL-KDD intrusion dataset, a widely used benchmark derived from the original KDD Cup 1999 dataset.
- We evaluate the performance of our STL-based NIDS on the NSL-KDD dataset and compare its effectiveness with existing methods.

2. RELATED WORK

It is important to note that this discussion focuses solely on studies that utilized the NSL-KDD dataset to evaluate effectiveness. Henceforth, any mention of a dataset refers to NSL-KDD. One of the earliest studies [11] employed Artificial Neural Networks (ANNs) with enhanced backpropagation to develop an intrusion detection system. This research utilized the entire training dataset, allocating 70% for training, 15% for validation, and 15% for testing. As anticipated, performance degraded when evaluated on unlabeled data. The training dataset will be analyzed. Another study [12] employed the J48 decision tree classifier with 10-fold cross-validation. This experiment utilized a reduced feature set of 22 attributes instead of the complete set of 41. Similar research [13] demonstrated that the Random Forest model achieved the lowest false alarm rate, surpassing other supervised tree-based classifiers.

Numerous two-level classification schemes have also been proposed. One study [14] used a Discriminative Multinomial Naive Bayes (DMNB) model as the base classifier, with nominal features converted to binary using a controlled filtering approach at the second level and 10-fold cross-validation. This concept was further developed using Random Forest and Ensembles of Balanced Nested Dichotomies (END) [15]. END is an abbreviation for ensembles of balanced nested dichotomies. As expected, this approach resulted in an increased detection rate and a reduced false positive rate.

Another study [16] proposed a novel two-level technique that first applies Principal Component Analysis (PCA) for feature reduction and then utilizes a Support Vector Machine (SVM) for classification, achieving high detection accuracy. While this approach, using the full training dataset with 41 features, demonstrated promising results, reducing the feature set to 23 improved the detection accuracy for specific attack types, albeit with a slight decrease in overall performance. Building upon their previous work, the authors of [17] first ranked features based on information gain and then applied behavior-based feature selection to reduce the feature set to 20. This approach led to an improvement in the reported accuracy on the training dataset. The secondary group of experiments utilized both training and testing datasets. An earlier study [18] combined fuzzy classification with a genetic algorithm, achieving a detection accuracy of at least 80% with a 1% false positive rate.

One notable study [19] revealed a significant performance degradation when training and testing data were combined. This research employed unsupervised clustering techniques to address its research questions. Similarly, another study [20] utilizing both training and testing datasets employed a k-nearest neighbors approach, achieving slightly higher detection accuracy and a lower false positive rate. Compared to the SVM-RBF approach, the Optimum-Path Forest (OPF) strat-

egy, which utilizes graph partitioning for feature classification, has been shown to achieve a higher detection rate, although it is not as widely adopted as other techniques. The study [21] employed a deep learning approach, utilizing a Deep Belief Network (DBN) for feature selection and a Support Vector Machine (SVM) for classification. This approach, when trained on the training data, achieved an accuracy of approximately 92.84%.

Our research, which also utilizes both training and testing datasets, builds upon these earlier works by exploring the application of deep learning for NIDS. The authors of [22] adopt a semi-supervised learning scheme, similar to the one used in [23]. Their technique was validated using real-world data from the KDD Cup 1999 dataset, which was also used for training. Our approach differs in that we specifically focus on the NSL-KDD dataset to evaluate the feasibility of using deep learning for NIDS. Furthermore, we employ a sparse autoencoder for completely unsupervised feature learning. The authors of [24] propose a deep learning approach for network traffic analysis based on sparse autoencoders. However, instead of focusing on intrusion detection, their research concentrates on identifying anomalous protocols within TCP traffic.

3. METHODOLOGY

3.1. SELF-TAUGHT LEARNING (STL)

Self-Taught Learning (STL) is a deep learning-based classification technique that operates in two stages. The first stage, known as Unsupervised Feature Learning (UFL), focuses on constructing robust feature representations from a large volume of unlabeled data. This type of learning, free from human supervision, relies solely on the inherent structure within the data. The learned representation is then utilized in the subsequent stage to categorize labeled data. A key assumption in STL is that even if the unlabeled and labeled data originate from different distributions, there should be some underlying relationship or shared features between them. Fig. 1 presents a diagrammatic representation of the STL architecture.

Two common approaches for UFL are Gaussian Mixtures and Sparse Autoencoders. This study employed a Sparse Autoencoder for feature learning due to its simplicity and efficiency. In a Sparse Autoencoder, the roles of the traditional neural network layers (input, hidden, output) are reinterpreted. The input layer of the neural network corresponds to the output layer of the Sparse Autoencoder, while the hidden layer remains the same. Finally, the output layer of the neural network aligns with the input layer of the Sparse Autoencoder. Both the output and input layers consist of " N " nodes, while the hidden layer is composed of " K " nodes. The Sparse Autoencoder aims to reconstruct the original input data at its output layer, thereby learning a compressed and meaningful representation in the hidden layer.

The sigmoid function, $g(z)=1/(1+exp(-z))$, is used to activate the nodes in the hidden and output layers, indicated by hW, b , respectively:

$$hw, b(x) = g(Wx + b) \quad (1)$$

$$J = \frac{1}{2m} \sum_{i=1}^m |x_i - \hat{x}_i|^2 + \frac{\lambda}{2} \left(\sum_{k,n} W^2 + \sum_{n,k} V^2 + \sum_k b_1^2 + \sum_n b_2^2 \right) + \beta \sum_{j=1}^K KL(\rho || \rho_j^{\wedge}) \quad (2)$$

The cost function minimized during backpropagation in a sparse autoencoder is represented by Equation (2). This function consists of three key components:

- **Reconstruction Error:** The primary term represents the average sum-of-squared errors between the input and reconstructed output over all "m" training data points. This term encourages the autoencoder to learn a faithful representation of the input data.
- **Weight Decay:** The second term incorporates weight decay, controlled by a weight decay parameter. This term helps prevent overfitting by penalizing large weights, thus promoting a smoother and more generalizable model.
- **Sparsity Penalty:** The first term in the equation is the sparsity penalty factor. This term, crucial for enforcing sparsity in the hidden layer, encourages most hidden units to have low average activation levels.

Equation (3) defines the sparsity penalty using the Kullback-Leibler (KL) divergence:

$$KL(\rho || \rho_j^{\wedge}) = \rho \log \frac{\rho}{\rho_j^{\wedge}} + (1 - \rho) \log \frac{1 - \rho}{1 - \rho_j^{\wedge}} \quad (3)$$

Where the sparsity penalty term β is governed by a sparsity limited parameter " ρ " having a measure ranging 0 to 1, where parameter's value can be any number between 0, 1. The $KL(\rho || \rho_j^{\wedge})$ reaches its minimum value when $\rho = \rho_j$, in which j denotes the average activation of hidden unit "j" over all training inputs "x". In other words, the penalty is minimized when the average activation of each hidden unit closely matches the desired sparsity level (ρ).

Once the optimal parameters $W, b1$ are learned from the unlabeled data (xu) using the sparse autoencoder, we can generate feature representations for the labeled data (xl). This is achieved by calculating:

$a = hW, b1(xl)$. The modified representation of the attributes is used in the next step.

The sparsity penalty factor, appearing as the first term in the equation, ensures that the hidden layer maintains relatively low average activation levels. This factor, formally known as the Kullback-Leibler (KL) divergence, is defined in Equation (3).

The sparsity penalty term, β , is governed by the sparsity parameter " ρ ," which ranges from 0 to 1. The $KL(\rho || \rho_j^{\wedge})$ reaches its minimum value when $\rho = \rho_j$, where

"j" represents the average activation of hidden unit "j" over all training inputs "x".

We first determine the optimal parameters $W, b1$ using the sparse autoencoder on the unlabeled data (xu). Subsequently, we assess the feature representation for the labeled data (xl) using these learned parameters. This representation, denoted as a , is calculated as: $a = W, b1(xl)$.

3.2. WORKING OF STL

STL has been successfully incorporated into deep learning models for Network Intrusion Detection Systems (NIDS). Below is a simplified explanation of the STL workflow:

Unlabeled Data Collection: Initially, a large dataset of raw network traffic data is collected. This data encompasses various network packets and flows but lacks any labels indicating whether the traffic is benign or malicious.

Feature Extraction: This stage involves extracting relevant features from the raw network data. These features may include packet size, protocol type, source and destination IP addresses, port numbers, and other characteristics. Feature extraction enables the model to identify patterns and predict future network behavior.

Pre-training on Unlabeled Data: A deep learning model, such as an autoencoder or another type of artificial neural network, is trained on the unlabeled data. The model's objective during pre-training is to reconstruct the input data accurately, thereby learning meaningful representations by minimizing the difference between the input and its reconstruction.

Fine-tuning on Labeled Data: The pre-trained model is then fine-tuned using a smaller labeled dataset. This step may not always be necessary, depending on the size of the labeled data and the pre-trained model's performance.

Intrusion Detection: Once trained, the model functions as an intrusion detection system. During inference, the model receives new network traffic data and classifies it as either benign or malicious based on learned patterns. If the model detects any anomalies or deviations from normal behavior, it triggers an alert.

Feedback Loop: The NIDS can incorporate a feedback mechanism where the generated alerts are used to further improve the model's performance over time. This feedback loop allows for continuous learning and adaptation to new threats.

3.3. DATASET

This study utilizes the NSL-KDD dataset, a modified and refined version of the original KDD Cup 99 dataset. The KDD Cup 99 dataset, based on network traffic collected during the 1998 DARPA IDS assessment program, has undergone significant changes. The original data-

set consists of raw network data gathered over seven weeks of training and two weeks of testing. The testing data includes several attack types absent from the training data. This deliberate omission aims to simulate real-world scenarios where novel attacks, often inspired by previous ones, emerge. This characteristic enhances the dataset's ability to evaluate the accuracy of intrusion detection systems in identifying unknown threats. The NSL-KDD dataset comprises five million TCP/IP connection records for training and two million for testing.

For many years, the KDD Cup dataset served as a standard benchmark for evaluating NIDS. However, a significant drawback is the high redundancy within the dataset. A substantial portion of the records in both the training (78%) and testing (75%) sets are duplicates. This redundancy biases learning algorithms towards the most frequent attack types, resulting in poorer performance on less common but potentially more dangerous attacks. For example, a simple machine learning model achieved a minimum accuracy of 98% on the training data but only 86% on the testing data. This discrepancy makes it challenging to compare different

IDSs and training methods fairly. The NSL-KDD dataset addresses these limitations. It improves upon the KDD Cup dataset in two key ways:

1. Redundancy Removal: Duplicate records are removed from both the training and testing sets.
2. Difficulty-Based Sampling: Records are categorized based on their difficulty level for learning algorithms. The NSL-KDD dataset then samples records randomly from various difficulty levels, ensuring a more balanced and representative distribution of attack types.

The training data comprises 23 traffic classes, consisting of 22 attack classes and one normal class. The test data set is more diverse, containing 39 traffic classes. These include 21 attack classes present in the training data, 16 novel attack classes, and one normal class. Each attack class falls into one of four categories based on its intended impact: Probing, Denial of Service (DoS), Remote to Local (R2L), and User to Root (U2R). Table 1 presents the distribution of normal and attack traffic within the training and testing sets.

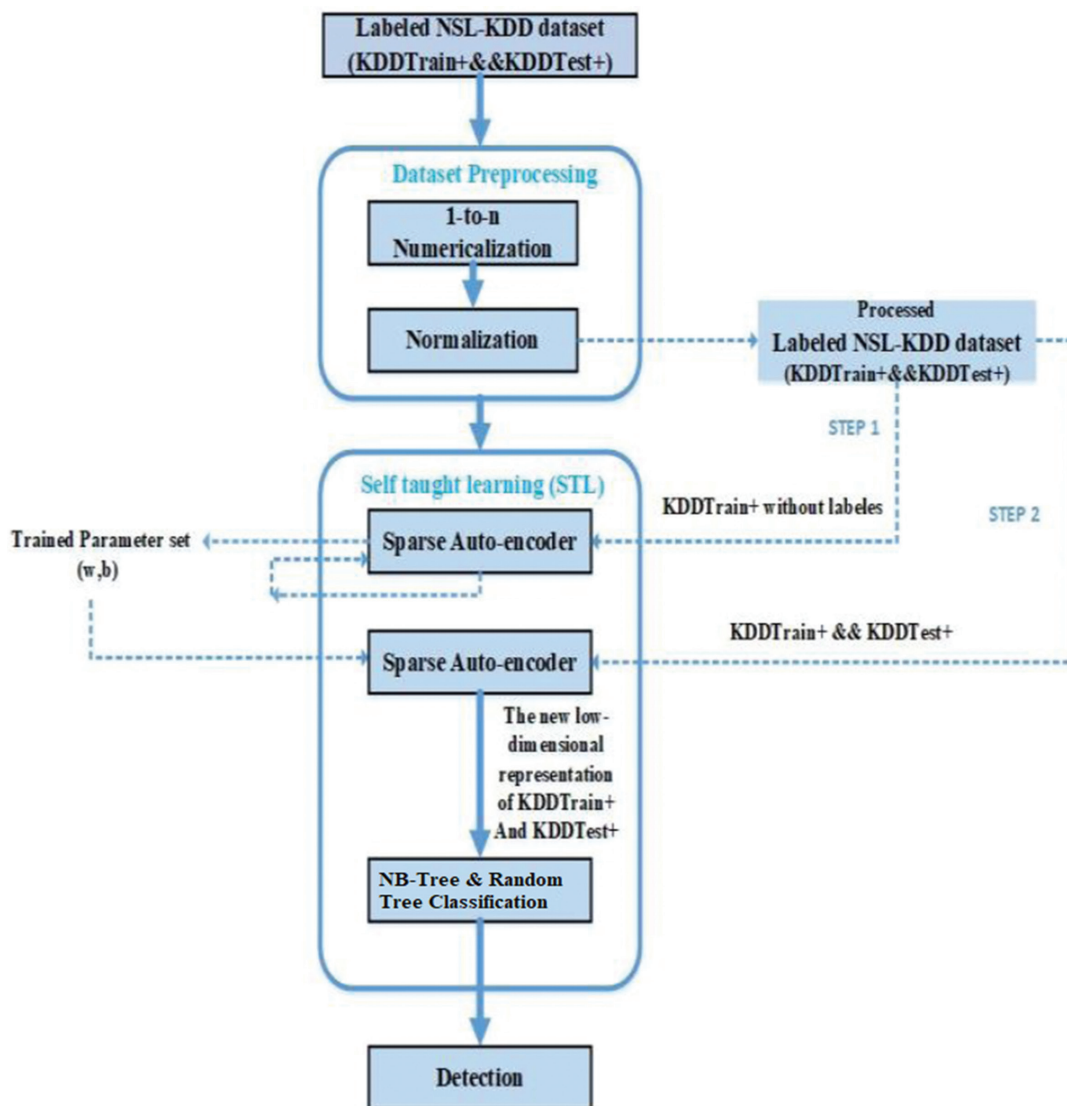


Fig. 1. STL NIDS architecture

Table 1. Distribution of Normal and Attack Traffic in the NSL-KDD Dataset

Traffic	Training	Test	
Normal	67343	9711	
Attack	DoS	45927	7458
	U2R	52	67
	R2L	995	2887
	probe	11656	2421

4. RESULTS AND DISCUSSION

The following describes the two methods employed to evaluate the NIDS performance. The first method utilizes the entire dataset for both training and testing, resulting in a high accuracy rate and a low false positive rate. However, this approach lacks independent evaluation. The second method addresses this limitation by splitting the dataset into separate training and testing sets. This independent evaluation, while more realistic, yields lower accuracy due to the differing conditions under which the training and testing data were collected. To ensure a comprehensive assessment, we prioritize the results obtained using the second, more realistic, method. However, for completeness, we also present the results from the first method.

4.1. PERFORMANCE ASSESSMENT, NSL IMPLEMENTATION

The dataset, as described in the previous section, contains numerous attributes, each of which can take on a range of values. Before employing self-taught learning, the dataset requires preparation. This involves converting nominal features into discrete attributes using '1-to-n encoding'. Additionally, the value of the 'num outbound cmds' feature is set to 0 for all entries in both the training and testing sets, as this feature is currently absent from the database. Following these steps, the dataset yields 121 features. Fig. 2 illustrates how the sigmoid function generates values in the output layer during the feature learning phase. As shown in the diagram, this function produces values ranging from 0 to 1.

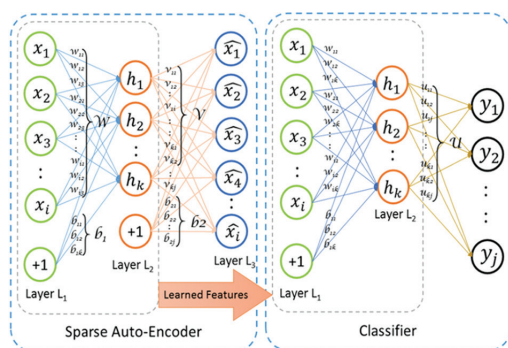


Fig. 2. STL based learning stages

4.1.1 Training phase

The classification accuracy of self-taught learning (STL) was evaluated on the training data using 10-fold

cross-validation. Performance was assessed for two, five, and twenty-three class scenarios and compared against a baseline of soft-max regression (SMR) trained on the same data without prior knowledge. As illustrated in Fig 3, STL significantly outperforms SMR in the two-class classification task. However, for the five-class and twenty-three class scenarios, the performance of both methods is comparable. Furthermore, our analysis determined that STL achieves a consistently high classification accuracy exceeding 98% across all tested categorization scenarios.

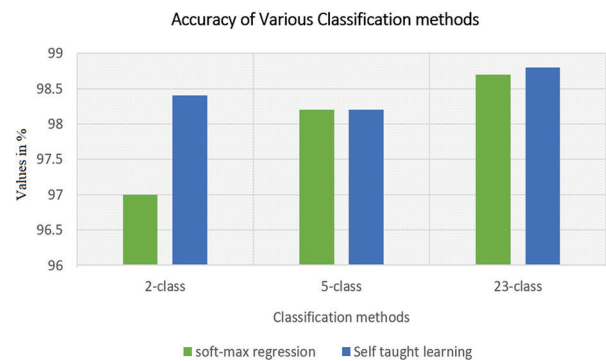


Fig. 3. Classification accuracy

Certain data categories were excluded from the 10-fold cross-validation evaluation of the five-class and twenty-three-class scenarios. Consequently, these metrics were only assessed for the two-class classification task. Our analysis revealed that STL consistently outperformed SMR across all evaluated metrics. Specifically, as depicted in Fig 4, STL achieved an F-measure of 98.84%, while SMR attained 96.79%. Notably, STL's performance on the training data approaches the highest accuracy levels reported in the literature for similar tasks.

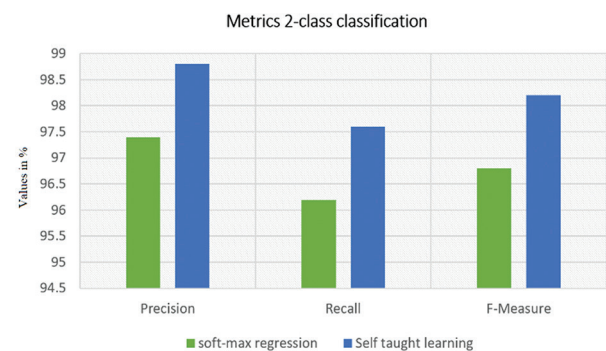


Fig. 4. Precision, recall, F-measure data

4.1.2. Testing phase

We evaluated the performance of STL and SMR on both two-class and five-class classification tasks using the held-out testing data. As illustrated in Fig. 5, STL consistently outperforms SMR in terms of accuracy. For the two-class scenario, STL achieved an accuracy of 88.39%, surpassing the 78.06% accuracy obtained by SMR. This result also compares favorably to previous studies, with the highest reported accuracy for a similar task using NB-

Tree being 82% [24]. In the five-class scenario, STL maintained its advantage with an accuracy of 79.10%, compared to 75.23% for SMR. Fig. 6 and 7 provide a detailed breakdown of the performance metrics for the five-class and two-class tasks, respectively, including F-measure, accuracy, and recall. Interestingly, while STL demonstrates superior overall accuracy in the two-class case, its accuracy (85.44%) is notably lower than that of SMR (96.56%) when considering only Figure 6. However, STL exhibits a significantly higher recall rate (95.95%) compared to SMR (63.73%), ultimately leading to a higher F-measure (90.4% for STL vs. 76.8% for SMR). This discrepancy highlights the importance of considering multiple evaluation metrics. For the five-class classification (Figure 7), the results follow a similar trend, with STL achieving a higher F-measure (75.76%) than SMR (72.14%).

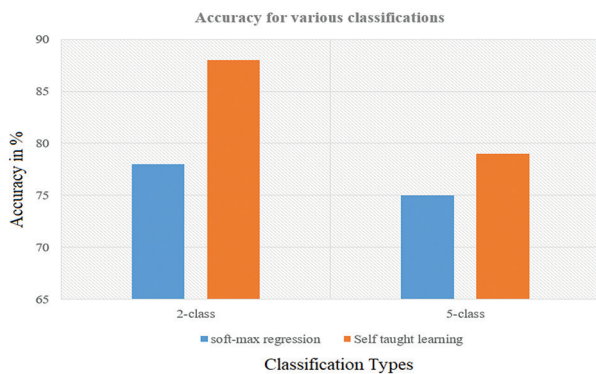


Fig. 5. Classification accuracy



Fig. 6. An evaluation of accuracy metrics for two-class classification

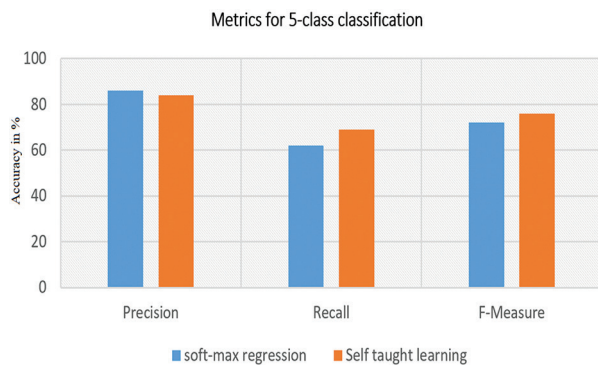


Fig. 7. An evaluation of accuracy metrics for 5-class classification

Comparing Self-Taught Learning (STL) with Softmax Regression (SMR) provides researchers and practitioners valuable insights into the trade-offs between unsupervised and supervised learning for Network Intrusion Detection Systems (NIDS). STL excels in exploratory data analysis, uncovering patterns and anomalies within unlabeled data. Conversely, SMR offers a robust framework for classification when labeled data is available, enabling effective cross-situational categorization. Understanding these distinct strengths and weaknesses is crucial for selecting the most appropriate algorithm based on the specific requirements and constraints of the NIDS application.

Self-Taught Learning (STL), an unsupervised learning approach, demonstrates particular efficacy in scenarios where labeled data is limited, especially for binary anomaly detection in Network Intrusion Detection Systems (NIDS). STL's ability to independently learn meaningful representations from unlabeled data is crucial for identifying previously unseen threats, which often evade detection by supervised methods reliant on predefined labels. However, STL may exhibit limitations in multi-class classification tasks, potentially leading to lower recall rates compared to supervised learning. This stems from STL's lack of predefined intrusion classes, making it challenging to differentiate between various attack types during training. Additionally, interpreting the learned representations within STL models can be less intuitive than those in supervised learning models, potentially hindering explainability. Therefore, a thorough understanding of STL's operational characteristics and limitations, as documented in the literature, is essential for justifying its suitability and effectiveness for specific NIDS applications.

Empirical studies consistently demonstrate that applying STL-based deep learning models to NIDS leads to significant improvements in intrusion detection and prevention rates. These enhancements are evident across several key aspects:

Improved Detection Accuracy: A primary evaluation metric for NIDS is the model's ability to accurately distinguish between benign and malicious network traffic. STL-based approaches consistently outperform conventional methods in this regard. This superior performance stems from their ability to leverage latent representations learned from unlabeled data, enabling the detection of novel attack patterns not encountered during training. Consequently, STL-based NIDS exhibit enhanced detection capabilities, strengthening overall network security.

Reduced False Positives: Minimizing false positives is crucial in NIDS, as excessive alerts can overwhelm security teams, leading to alert fatigue and potentially missed threats. STL-based models excel in this area due to their capacity to discern subtle anomalies within complex traffic patterns. This ability to effectively differentiate between benign and malicious events sig-

nificantly reduces false alarms, optimizing resource allocation for security teams.

Adaptability to New Threats: The dynamic nature of cybersecurity threats necessitates adaptable defense mechanisms. STL-based NIDS models possess inherent flexibility, continuously learning from incoming network data to refine their detection patterns. This adaptability enables them to effectively identify and respond to emerging attack types and evolving malicious tactics, ensuring the NIDS remains a relevant and effective security measure.

Scalability and Efficiency: Effective NIDS solutions must handle the demands of large-scale networks without compromising performance. STL-based models are well-suited for such environments, often designed with computational efficiency in mind. This allows them to analyze vast volumes of network traffic in real-time without imposing excessive overhead on system resources.

Overall, these findings highlight the substantial benefits of incorporating STL-based deep learning models into NIDS, paving the way for more robust and resilient network security solutions.

5. CONCLUSION

This paper presented an effective and adaptable deep learning-based approach for enhancing Network Intrusion Detection Systems (NIDS). The proposed NIDS leverages a sparse autoencoder for unsupervised feature learning, followed by a soft-max regression classifier for anomaly detection. The system's performance was rigorously evaluated using the NSL-KDD benchmark dataset, demonstrating its effectiveness in identifying network intrusions. Comparative analysis revealed that our NIDS outperforms existing methods for both normal and anomaly detection on the test data. While alternative approaches like Stacked Autoencoders and Deep Belief Networks, also derived from sparse autoencoders, show promise for unsupervised feature learning when combined with classifiers such as J48, NB-Tree, or Random Forest, these methods achieved superior results when applied directly to the dataset. Our experiments with Self-Taught Learning (STL) based deep learning models for NIDS highlight the significant advantages of incorporating STL. The results indicate that STL enhances network intrusion security by enabling: higher accurate detection rates, minimized false alarms, adaptive learning of new threats over time, and scalability to large networks. Future research directions include exploring techniques for effectively training STL-based NIDS while preserving data privacy. This could involve investigating privacy-preserving machine learning methods such as federated learning, differential privacy, and homomorphic encryption. These technologies can facilitate collaborative model training across distributed networks without requiring the sharing of sensitive raw data.

6. ACKNOWLEDGMENT

This work has been supported by the Universiti Kebangsaan Malaysia, Under the research grant scheme DIP 2022-021.

7. REFERENCES

- [1] L. X. Ying, M. Aman, A. Hafizah, M. S. Jalil, T. M. Omar, Z. S. Attarbashi, M. A. Abuzaraida, "Malaysia Cyber Fraud Prevention Application: Features and Functions", *Asia-Pacific Journal of Information Technology and Multimedia*, Vol. 12, No. 2, 2023, p. 312.
- [2] A. A. Ahmed, M. K. Hasan, I. Memon, A. H. M. Aman, S. Islam, T. R. Gadekallu, S. A. Memon, "Secure AI for 6G Mobile Devices: Deep Learning Optimization Against Side-Channel Attacks", *IEEE Transactions on Consumer Electronics*, Vol. 70, No. 1, 2024, pp. 3951-3959.
- [3] F. Dehkordi, K. Manochehri, V. Aghazarian, "Internet of Things (IoT) Intrusion Detection by Machine Learning (ML): A Review", *Asia-Pacific Journal of Information Technology and Multimedia*, Vol. 12, No. 1, 2023, pp. 13-38.
- [4] A. A. Ahmed, M. K. Hasan, N. S. Nafi, A. H. Aman, S. Islam, S. A. Fadhil, "Design of Lightweight Cryptography based Deep Learning Model for Side Channel Attacks", *Proceedings of the 33rd International Telecommunication Networks and Applications Conference*, Melbourne, Australia, 29 November - 1 December 2023, pp. 325-328.
- [5] N. Jafri, M. M. Yusof, "Managing Data Security Risk in Model Software as a Service (SAAS)", *Asia-Pacific Journal of Information Technology and Multimedia*, Vol. 7, No. 1, 2018, pp. 99-117.
- [6] A. K. Jakkani et al. "Design of a Novel Deep Learning Methodology for IOT Botnet based Attack Detection", *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 11, No. 9, 2023, pp. 4922-4927.
- [7] A. O. Alzahrani, M. J. Alenazi, "Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks", *Future Internet*, Vol. 13, No. 5, 2021, p. 111.
- [8] P. Reddy, Y. Adetuwo, A. K. Jakkani, "Implementation of Machine Learning Techniques for Cloud Security in Detection of DDOS Attacks", *Internation*

tional Journal of Computer Engineering and Technology, Vol. 15, No. 2, 2024, pp. 25-34.

- [9] R. Kolandaisamy, K. Subaramaniam, A. B. Jalil, "A Study on Comprehensive Risk Level Analysis of IoT Attacks", Proceedings of the International Conference on Artificial Intelligence and Smart Systems, Coimbatore, India, 25-27 March 2021, pp. 1391-1396.
- [10] A. Agbonyin, P. Reddy, A. K. Jakkani, "Utilizing Internet of Things (IOT), Artificial Intelligence, and Vehicle Telematics for Sustainable Growth in Small, and Medium Firms (SMES)", International Journal of Computer Engineering and Technology, Vol. 15, No. 2, 2024, pp. 182-191.
- [11] Y.-W. Chen, J.-P. Sheu, Y.-C. Kuo, N. Van Cuong, "Design and Implementation of IoT DDoS Attacks Detection System Based on Machine Learning", Proceedings of the European Conference on Networks and Communications, Dubrovnik, Croatia, 15-18 June 2020, pp. 122-127.
- [12] Q. Abu Al-Haija, S. Zein-Sabatto, "An Efficient Deep-Learning-Based Detection and Classification System for Cyber-Attacks in IoT Communication Networks", Electronics, Vol. 9, No. 12, 2020, p. 2152.
- [13] J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, Y. Xiang, "Deep Learning Based Attack Detection for Cyber-Physical System Cybersecurity: A Survey", IEEE/CAA Journal of Automatica Sinica, Vol. 9, No. 3, 2021, pp. 377-391.
- [14] B. Susilo, R. F. Sari, "Intrusion Detection in IoT Networks Using Deep Learning Algorithm", Information, Vol. 11, No. 5, 2020, p. 279.
- [15] T. H. Aldhyani, H. Alkahtani, "Attacks to Autonomous Vehicles: A Deep Learning Algorithm for Cybersecurity", Sensors, Vol. 22, No. 1, 2022, p. 360.
- [16] I. Ullah, Q. H. Mahmoud, "Design and Development of a Deep Learning-Based Model for Anomaly Detection in IoT Networks", IEEE Access, Vol. 9, 2021, pp. 103906-103926.
- [17] A. A. Ahmed, W. A. Jabbar, A. S. Sadiq, H. Patel, "Deep Learning-Based Classification Model for Botnet Attack Detection", Journal of Ambient Intelligence and Humanized Computing, Vol. 13, 2022, pp. 3457-3466.
- [18] V. Dutta, M. Choraś, M. Pawlicki, R. Kozik, "A Deep Learning Ensemble for Network Anomaly and Cyber-Attack Detection", Sensors, Vol. 20, No. 16, 2020, p. 4583.
- [19] N. Elmrabit, F. Zhou, F. Li, H. Zhou, "Evaluation of Machine Learning Algorithms for Anomaly Detection", Proceedings of the International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Dublin, Ireland, 15-19 June 2020, pp. 1-8.
- [20] A. Khaleghi, M. S. Ghazizadeh, M. R. Aghamohammadi, "A Deep Learning-Based Attack Detection Mechanism Against Potential Cascading Failure Induced by Load Redistribution Attacks", IEEE Transactions on Smart Grid, Vol. 14, No. 6, 2023, pp. 4772-4783.
- [21] J. Shareena, A. Ramdas, H. AP, "Intrusion Detection System for IoT Botnet Attacks Using Deep Learning", SN Computer Science, Vol. 2, No. 205, 2021, pp. 1-8.
- [22] L. Liu, P. Wang, J. Lin, L. Liu, "Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning", IEEE Access, Vol. 9, 2020, pp. 7550-7563.
- [23] J. Lansky, S. Ali, M. Mohammadi, M. K. Majeed, S. T. Karim, S. Rashidi, M. Hosseinzadeh, A. M. Rahmani, "Deep Learning-Based Intrusion Detection Systems: A Systematic Review", IEEE Access, Vol. 9, 2021, pp. 101574-101599.
- [24] G. Kocher, G. Kumar, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: Recent Developments and Challenges", Soft Computing, Vol. 25, 2021, pp. 9731-9763.

Mask FORD-NET: Efficient Detection of Digital Image Forgery using Hybrid REG-NET based Mask-RCNN

Original Scientific Paper

Priscilla Whitin*

Department of Electrical and Electronics Engineering,
VelTech Rangarajan Dr. Sagunthala R&D Institute of
Science and Technology,
Avadi, Chennai, Tamil Nadu, India.
priscillawhitin@veltech.edu.in

S. Sivakumar

Department of Electrical and Electronics Engineering,
VelTech Rangarajan Dr. Sagunthala R&D Institute of
Science and Technology,
Avadi, Chennai, Tamil Nadu, India.
ssivakumar@veltech.edu.in

M. Geetha

Department of Electrical and Electronics Engineering,
Sri Eshwar College of Engineering,
Coimbatore, Tamil Nadu, India.
geetha.m@sece.ac.in

M. Devaki

Department of Electrical and Electronics Engineering,
Velammal College of Engineering and Technology,
Madurai, Tamilnadu, India.
devaki852m@outlook.com

*Corresponding author

A. Bhuvanesh

Department of Electrical and Electronics Engineering,
PSN College of Engineering and Technology,
Tirunelveli, Tamilnadu, India.
bhuvanesh.ananthan@gmail.com

Kiruthiga Balasubramaniyan

Department of Electronics and Communication
Engineering,
K. Ramakrishnan College of Technology,
Trichy, Tamilnadu, India.
balasubramaniyankiruthiga44@gmail.com

A. Ahilan

Department of Electronics and Communication
Engineering,
PSN College of Engineering and Technology,
Tirunelveli, Tamilnadu, India.
listentoahil@gmail.com

Abstract – Digital image is a binary representation of visual data which provides a rapid method for analyzing large quantities of data. Furthermore, digital images are more vulnerable to fraud when distributed over an open channel via information and communication technology. However, the image data can be modified fraudulently by intruders using vulnerabilities in telecommunications infrastructure. To overcome these issues, this paper proposes a novel Mask-RCNN based Image FORgery Detection (Mask FORD-NET) which is developed for digital image forgery detection. Initially, the input image is passed beyond the recompression module to reduce the insignificance and complexity of the image to preserve or transfer the data efficiently. After image recompression, the recompressed image is transferred to the feature extraction phase which is done by using REG-NET. The extracted features are received to the noise cancellation and ELA converter module to analyze and reduce the ambient noise. After noise cancellation, the data are passed to the MASK-RCNN module, to detect and classify the forged images and finally provide the segmented output. The Mask FORD-NET framework is simulated by using MATLAB. The efficiency of the proposed Mask FORD-NET framework is assessed by using accuracy, precision, recall, and F1-measure. The experimental results show that the accuracy of the Mask FORD-NET framework has increased to up to 98.72% for digital image forgery detection. The accuracy of the proposed Mask FORD-NET framework is 80.72%, 86.32%, and 95.00% better than existing ASCA, VixNet, and MiniNet techniques respectively.

Keywords: Digital Image Forgery, Deep Learning, REG-NET, Mask-RCNN

Received: January 26, 2024; Received in revised form: July 29, 2024; Accepted: August 2, 2024

1. INTRODUCTION

The widespread availability of digital images due to the advancement of imaging technology and the profusion of image manipulation applications that do not require specialized knowledge has led to a significant rise in the number of forged digital images on social media [1-3]. Digital images are used in many industries, including social networking, e-government, military information, and meteorological research [4, 5]. By 2022, 72.6% of the world's population, according to the International Telecommunication Union (ITU), will have access to the Internet. This implies that about 4.1 billion people will have access to these technologies as well as other services [6].

Active and passive methods are the two categories that are utilized for digital image forgery techniques [7, 8]. The passive adaptive detection technique analyses the original image and identifies regions in which the image has been altered using several statistical and semantic criteria linked to the content of the image. The main aim of the detection method for image forgery, that address the increasing issue of image forgery [9, 10]. Digital image forensics experts can employ Deep Learning (DL) and Machine Learning (ML) to appear for forged images. It has been demonstrated that these methods offer high accuracy rates and provide protection against a wide variety of image forgery [11, 12].

Due to their reliance on JPEG compression artifacts, digital images are useful but limited in their ability to reliably detect extremely intricate counterfeits [13, 14]. Furthermore, it is challenging to understand the results and make conclusions due to the transparent nature of the digital image forgery detection process [15]. To resolve these shortcomings, a novel Mask FORD-NET framework is proposed for digital image forgery detection. The main contributions of the proposed Mask FORD-NET framework are presented as follows.

- Initially, the input image is passed beyond to the recompression module to reduce the insignificance and complexity of the image in an efficient manner.
- After image recompression, the recompressed image is transferred to the feature extraction phase which is done by using REG-NET.
- The extracted features are received to the noise cancellation and ELA converter module to analyze and reduce the ambient noise.
- After noise cancellation, the data are passed to the MASK-RCNN module, to detect and classify the forged images and finally provide the segmented output.
- The effectiveness of the proposed Mask FORD-NET framework is evaluated by accuracy, specificity, precision, F₁ measure, and recall respectively.

The work described in this paper is organized as follows: Section 2 presents a summary of related work.

Section 3 presents the deep learning-based Mask FORD-NET framework for image forgery detection. Section 4 provides a thorough explanation of the Framework's Outputs and performance assessment. Conclusions and future work are presented in Section 5.

2. LITERATURE SURVEY

Advanced techniques for identifying changes done on virtual images have emerged as a result of recent trends in image tampering. Earlier studies have been put forth based on various methods such as ML and DL that are mostly predicated on observations made during the entire image history. The following is a brief definition of several related works.

In 2022, Koul et al. [16] proposed a method using convolutional neural networks to detect clone-based image manipulation. The MICC-F2000 dataset is used to assess the suggested method, which detects fake copies with 97.52% accuracy. However, the suggested approach consists of an increased false positive rate. In 2022 Wu et al. [17] suggested a reliable image tampering detection for social media streaming. The suggested method reduces IoU by 2.6%, 2.9%, and 4.5%, on the dataset that OSN transmitted. However, the suggested approach fails to perform robustly in complex degradation scenarios.

In 2022 Kumar et al. [18] suggested using a variation in non-overlapping blocks, the detection, and location of image manipulation. With 98% accuracy for improved detection and classification, the suggested technique is assessed using SSIM parameters. However, the suggested approach consists of high computational complexity. In 2022 Ganguly et al. [19] suggested a Vision Transformer that uses the Xception Network to detect video and image forgeries based on deepfakes. The ViXNet approach was evaluated on the DFDC dataset, generating an F1 score of 79.06% and an AUC score of 86.32% for identifying hypothetical fraudsters. However, the advantages of the deepfake detection techniques still need improvement.

In 2023 Nirmalpriya et al. [20] suggested a Hybrid deep learning network can identify digital image manipulation via Aquila's sin-cosine algorithm. A replicated fraudulent detection dataset is used to assess the suggested approach, which yields TNR and TPR values of 1.003% and 0.991%, respectively. However, the suggested ASCA method is not reversible. In 2023 Sushir et al. [21] suggested enhanced detection of random image manipulation based on accurate deep learning utilizing a combination of DCCAE and ADFC. The accuracy of the suggested hybrid DCCAE approach is 98.07% for the GRIP dataset and 99.23% for the CASIA V1 dataset. However, the noise estimation of the suggested approach is not robust.

In 2023, Tyagi and Yadav [22] suggested an immediate CNN for spotting forged images. For 140,000 real and fake faces, the suggested MiniNet version obtains

an accuracy of above 95%, and for the CASIA dataset, it achieves 93%. However, the suggested MiniNet model consists of low precision. In 2023 Vijayalakshmi et al. [23] suggested utilizing deep learning and error-level analysis, to detect counterfeits through copy-and-paste methods. Using the MICC-F220 dataset, the suggested approach was assessed and yielded an overall 99.2% accuracy, 96.5% specificity, 95.79% recall, and 96.09% F_measure for improved detection. However, the suggested approach is not highly efficient.

The aforementioned findings indicate that the majority of manual feature extraction techniques for counterfeit detection mainly rely on the operator. Mask FORD-NET framework, a deep learning framework that is designed to identify the manipulation of digital images. In this framework, the time complexity is decreased, efficiency is increased, and potential human mistake is eliminated when using the Mask FORD-NET framework.

2.1. RESEARCH GAP

Following a thorough examination of the literature, the following research gaps on the suggested research challenge were identified. Although much progress has been made, there are still several barriers that prevent useful techniques from being used. The deep learning-based development of the Mask-FORD-NET framework for digital image forgery is still occurring continuously. Recent image processing techniques depend on numerous attributes. Most approaches employ DL or ML techniques to identify image modification. Reviewing the literature, however, reveals that there is a significant improvement in deep learning-based image intrusion detection.

3. THE MASK FORD-NET METHODOLOGY

In this section, a novel deep-learning based Mask FORD-NET framework is proposed for the digital image forgery detection. Fig. 1 depicts the block diagram of the Mask FORD-NET architecture.

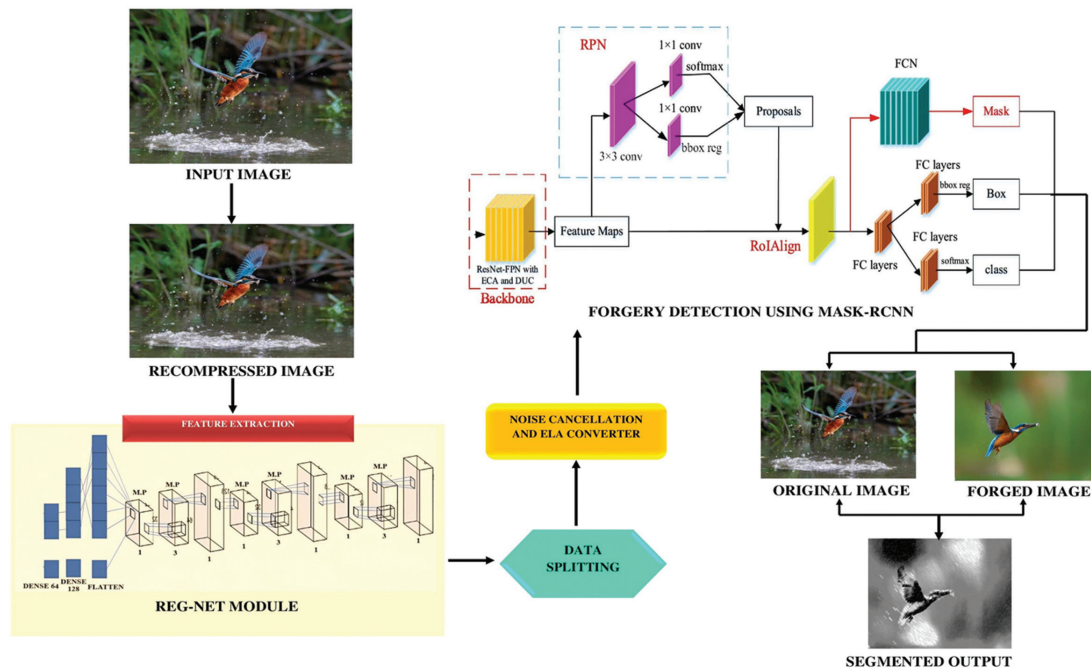


Fig. 1. The block diagram of the Mask FORD-NET framework

3.1. DATA RECOMPRESSION

The proposed method presents a framework for confirming the authenticity of fake images. This technique locates tampered locations and detects image tampering by using data compression features. In step size 2, this was completed in the interval [30, 100]. YCbCr is used in place of the RGB color system. The image's brightness information is stored in the Y channel, while the Cb and Cr channels hold the color information. The difference image is transformed to binary, where black and white regions indicate the original and modified portions of the image, to visualize the altered regions.

3.2. FEATURE EXTRACTION USING REG-NET

To extract features from compressed digital images, two types of ResNet building blocks were proposed such as bottleneck building blocks and non-bottleneck building blocks. Based on this, two different RNN-modified ResNet structural modules were obtained, one using ConvRNN as the regulator and the other using an RNN-modified ResNet structural module.

3.2.1. RegNet Module

ConvLSTM's output from the first module, H^{t1} , is represented by that from the second module, H^t . The mod-

ule's feature map is represented by X_i^t . The t -th RegNet (ConvLSTM) module can be expressed as

$$X_2^t = ReLU(BN(W_{12}^t * X_1^t + b_{12}^t)), \quad (1)$$

$$[H^t, C^t] = ReLU(BN(ConvLSTM(X_2^t, [H^{t-1}, C^{t-1}]))), \quad (2)$$

$$X_3^t = ReLU(BN(W_{23}^t * Concat[X_2^t, H^t])), \quad (3)$$

$$X_4^t = BN(W_{34}^t * X_3^t + b_{34}^t), \quad (4)$$

$$X_1^{t+1} = ReLU(X_1^t + X_4^t) \quad (5)$$

where b_{ji}^t stands for the correlation distance and W_{ij}^n stands for the convolution kernel that translates the features X_i^t to X_j^t . They are 3x3 convolution particles, W_{12}^t , W_{34}^t and W_{23}^t consists of 1x1 kernels. The batch normalization procedure is represented by BN (). Concat [] is a shorthand for the concatenation operation. The enter entity X_2^t and the previous output of *ConvLSTM* H^t in equation (1) are the input of *ConvLSTM* inside the module. *ConvLSTM* automatically determines whether the data inside the memory cell should be given to the H^t output hidden characteristic map based on the inputs.

3.2.2. Bottleneck RegNet Module

The fundamental component of the RegNet bottleneck module is the ResNet bottleneck building block. For large image processing, the bottleneck construction block was first presented. This makes it possible to represent the RegNet module bottleneck as,

$$X_2^t = ReLU(BN(W_{12}^t * X_1^t + b_{12}^t)), \quad (6)$$

$$[H^t, C^t] = ReLU(BN(ConvLSTM(X_2^t, [H^{t-1}, C^{t-1}]))), \quad (7)$$

$$X_3^t = ReLU(BN(W_{23}^t * X_2^t + b_{23}^t)), \quad (8)$$

$$X_4^t = ReLU(BN(W_{34}^t * Concat[X_3^t, H^t])), \quad (9)$$

$$X_5^t = BN(W_{45}^t * X_4^t + b_{45}^t), \quad (10)$$

$$X_1^{t+1} = ReLU(X_1^t + X_5^t), \quad (11)$$

where W_{12}^t and W_{45}^t are the two 1x1 kernels, and W_{23}^t is the 3 x 3 bottleneck kernel. The W_{34}^t is a 1 x 1 kernel for fusing features in our model.

3.3. NOISE-CANCELLATION AND ELA CONVERSION

Two important techniques were used in image forgery detection which are noise cancellation and ELA (Error Level Analysis). An image can be made noise-free by using a technique called noise removal. Noise elimination can be applied to detect altered images by eliminating artificial noise or artifacts that may have been introduced during the tampering process. The process involves using a DL algorithm to extract a variety of features from an image to identify tampering and filtering techniques are used to detect the noise in the images. Once the noise is identified, it can be removed using various filtering techniques, such as a median filter or a Gaussian filter. By removing

the noise, the algorithm can focus on the underlying structure of the image, which may provide clues about the forgery. However, ELA can be used to identify regions of an image that have been altered or compressed. ELA operates by interpreting differences in the degree of inaccuracy in various areas of an image. When an image is recompressed or modified, the error levels in the affected regions will be higher than in other areas of the image. Combining noise cancellation and ELA methods makes it possible to generate algorithms for identifying fake images that are more reliable and accurate.

3.4. DIGITAL IMAGE FORGERY DETECTION USING MASK R-CNN

Mask R-CNN enables the proper labeling of object regions and removals of those object regions from the background of each pixel level. Furthermore, Mask R-CNN may be utilized to detect the forged parts of the digital images by examining the form and edge properties of its mask images.

3.4.1. Feature pyramid network

In DL, feature extraction is a vital phase employed for extracting the relevant features present in the digital images. Especially, for feature extraction ResNet-101 is used as a feature pyramid network (FPN) model over an entire digital image. The ResNet-101 model uses the suggested rectangular zones to extract features, which are convolved into Mask R-CNN. The input data is fed into the convolutional CNN to generate the feature map. Convolutional layers are stacked, pooling layers are added, and ResNet-101 retains the residual connections. Five blocks comprise a convolutional network such as a 7x7 convolutional layer used in the first block, then 1x1, 3x3, and 1x1 convolutional layers are used in consecutive blocks. FPN strengthens the network backbone so that semantic and spatial information can be extracted from different-sized digital images.

3.4.2. Region proposal network

A Region Proposal Network (RPN) was utilized to create regions of interest with fixed points for every feature map. Multiple propositions are generated on rectangular objects with objective scores by superimposing convolutional feature maps across a small network. The digital image's foreground and background values were ascertained in this manner. The server adjusted the size and position of the digital images and chose the best limits using RPN prediction. Finally, using the regions of interest generated by the RPN layer, the FC layer builds bounding boxes and segmentation masks for particular areas of the digital image.

3.4.3. Fully convolutional network

ROI Align is utilized to modify each ROI's size to satisfy the FC input requirements before utilizing the full convolutional input. To extract pertinent attributes

from each RoI on the feature map, RoI Align employed bilinear interpolation as an alternative to the Mask R-CNN approach of RoI pooling equalization. Three prediction branches, a fully convolutional network (FCN), which is used for both classification and prediction segmentation. A regression layer, which modifies bounding box coordinates, and an FC layer, which is used for classification combined to generate the target mask in the multi-branch prediction stage. For both segmentation, bounding box classification, and analysis, the ROI alignment characteristics are fed into the head mask and bounding box head simultaneously. Through the use of all the characteristics in the soft-max layer, the results are fed to the FCN layer for classification.

Table 1. Hyperparameter settings of the proposed REG-NET technique

Parameter	Value
Training Data Ratio	60%
Validation Data Ratio	20%
Testing Data Ratio	20%
Training Time	10 hours
Optimizer	Adam
Cost Function	Binary Cross-Entropy
Batch size	64
Learning Rate	0.0001
Activation Function	Leaky ReLU
Number of Epochs	100

The hyperparameters of the proposed Mask FORD-NET method are covered in Table 1. Experimenting with different combinations of these hyperparameters can help in optimizing the performance of REG-NET for image forgery detection.

4. PERFORMANCE VALIDATION RESULT ANALYSIS

The experimental results of existing detection models and the proposed Mask FORD-NET framework are compared and analyzed in this section. The proposed work's performance is assessed using accuracy. As seen in the attached figures, it offer a thorough evaluation of the model's overall performance at a deep level of comprehension of the input data.

Dataset Description

To assess the efficacy of the proposed Mask FORD-NET framework, experiments are conducted on the widely used image tampering database, CASIA 2.0. The total number of images such as 12,614 images in BMP, JPG, and TIF formats, 5,123 are fictitious images and 7,491 are real shots presented in the dataset. Images from many genres are included in CASIA 2.0, such as those featuring people, animals, plants, architecture, objects, landscapes, textures, and interior shots. The collection contains images in a range of sizes and resolutions, from 800 x 600 pixels to 384 x 256 pixels.

4.1. PERFORMANCE METRICS

The accuracy, precision, recall and $F_{measure}$ is calculated and compared to the proposed strategy with existing approaches. They are computed as follows.

$$Accuracy = \frac{T_P + T_N}{T_{Total_Images}} \times 100 \quad (12)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (13)$$

$$Precision = \frac{T_P}{T_P + F_P} \quad (14)$$

$$F_{measure} = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100 \quad (15)$$

4.2. PERFORMANCE ANALYSIS

Fig. 2 and 3 demonstrate the proposed Mask FORD-NET framework with great accuracy throughout both training and testing.

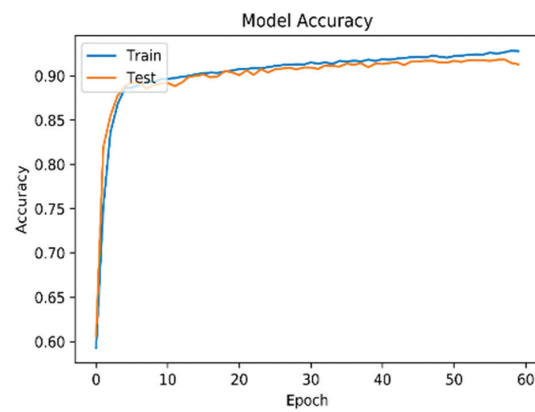


Fig. 2. Accuracy Calculation of existing Algorithm

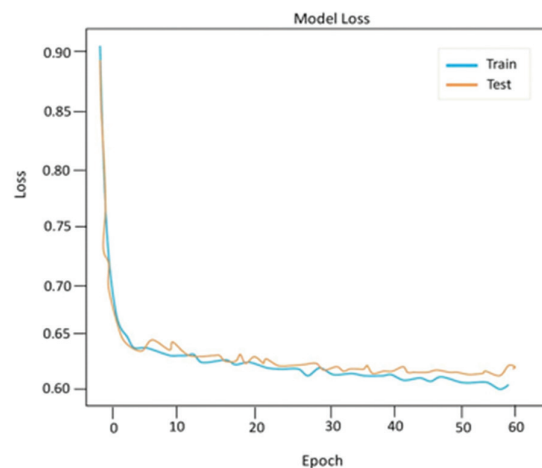


Fig. 3. Loss Calculation of Existing Algorithm

4.3. COMPARATIVE ANALYSIS

In Fig. 4, the image (A) of the table shows the input image and the image (B) of the figure shows the forged image. The image (C) of the figure insists the recompressed image and the extracted features are depicted in image (D). Finally, the segmented output is shown in image (E).

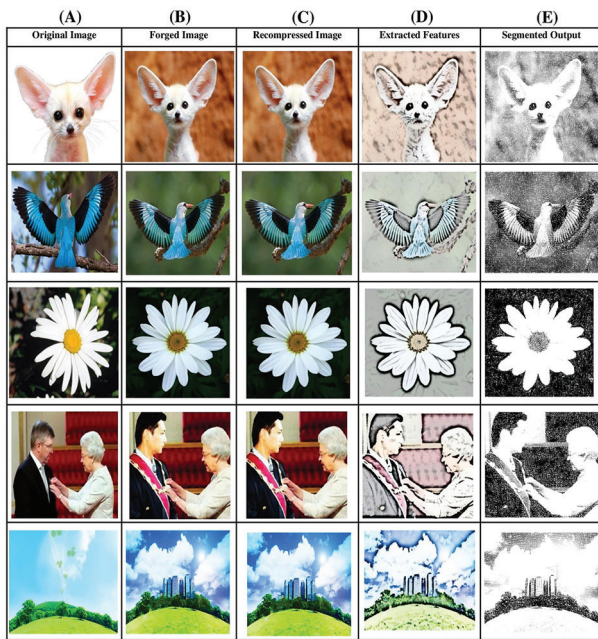


Fig. 4. Result of the proposed methodology

In comparison to alternative ways, Fig. 5 provides a better depiction of the accuracy of the proposed approach. The experimental results show that the Mask FORD-NET framework achieves 98.72% of accuracy, 90.36% of specificity, 92.25% of precision, 93.53% of F1-score, and 94.99% of recall for digital image forgery detection. The accuracy of the proposed Mask FORD-NET framework is 80.72%, 86.32%, and 95.00% better than existing ASCA, VixNet, and MiniNet techniques respectively.

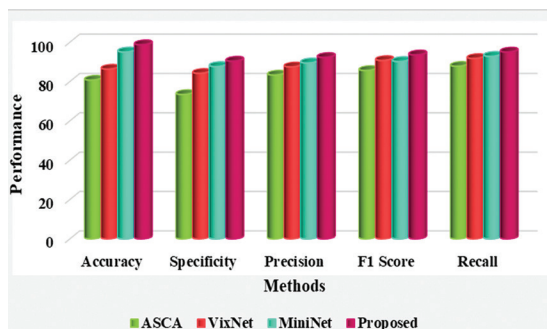


Fig. 5. Performance evaluation of existing approach

From the result analysis, we can infer that the suggested work's accuracy is 95.56 and its processing time is 31 milliseconds, compared to the existing system's accuracy of 92.23 and processing time of 30 milliseconds. For the suggested task to be accomplished more quickly and with greater precision. The existing and proposed DL approaches are graphically compared in Fig. 6. using several criteria, including accuracy, sensitivity, recall, and specificity.

The accuracy of the Bi-LSTM approach is 93.81%, the RCNN approach is 94.74%, and the corresponding accuracy of the REG-NET method is 95.47% both lower than the accuracy of the proposed Mask-RCNN methodology of 97.84% respectively.

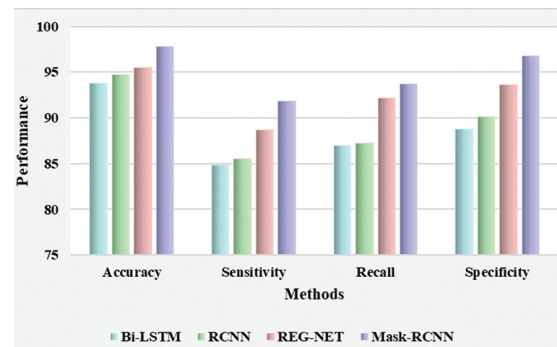


Fig. 6. Successive Rate of Mask-RCNN

5. CONCLUSION

This paper proposed a novel Mask FORD-NET framework which is developed for digital image forgery detection. Initially, the input image is passed through the recompression module to efficiently reduce insignificance and complexity. The recompressed image is then sent to the feature extraction phase using REG-NET. The extracted features are processed by the noise cancellation and ELA converter module to reduce ambient noise. Subsequently, the data are passed to the MASK-RCNN module for detecting and classifying forged images, ultimately providing the segmented output. The proposed Mask FORD-NET framework is validated by using the CASIA 2.0 image forgery dataset. The Mask FORD-NET framework is simulated by using MATLAB. According to the simulation results, a comparison is made between the proposed Mask FORD-NET framework and the existing approaches such as ASCA, VixNet, and MiniNet in terms of accuracy, precision, recall, sensitivity, and $F_{measure}$. The experimental results show that the accuracy of the Mask FORD-NET framework has increased to up to 98.72% for digital image forgery detection. The accuracy of the proposed Mask FORD-NET framework is 80.72%, 86.32%, and 95.00% better than existing ASCA, VixNet, and MiniNet techniques respectively. In the future, integrating multiple modalities such as text and audio, along with image content analysis, to detect sophisticated multimedia forgeries and deepfake content. Additionally, to pave the path for further study on identifying various forms of image forgery, the proposed Mask FORD-NET framework will aid in the field of image forgery detection.

6. REFERENCES:

- [1] N. Kaur, N. Jindal, K. Singh, "A deep learning framework for copy-move forgery detection in digital images", *Multimedia Tools and Applications*, Vol. 82, No. 12, 2023, pp. 17741-17768.
- [2] A. K. Jaiswal, R. Srivastava, "Detection of copy-move forgery in digital image using multi-scale, multi-stage deep learning model", *Neural Processing Letters*, Vol. 54, No. 1, 2022, pp. 75-100.

- [3] S. Walia, K. Kumar, M. Kumar, "Unveiling digital image forgeries using Markov based quaternions in frequency domain and fusion of machine learning algorithms", *Multimedia Tools and Applications*, Vol. 82, No. 3, 2023, pp. 4517-4532.
- [4] M. A. Anwar, S. F. Tahir, L. G. Fahad, K. Kifayat, "Image forgery detection by transforming local descriptors into deep-derived features", *Applied Soft Computing*, Vol. 147, 2023, p. 110730.
- [5] W. Lu, W. Xu, Z. Sheng, "An interpretable image tampering detection approach based on cooperative game", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 2, 2022, pp. 952-962.
- [6] G. Zhao, C. Qin, H. Yao, Y. Han, "DNN self-embedding watermarking: Towards tampering detection and parameter recovery for deep neural network", *Pattern Recognition Letters*, Vol. 164, 2022, pp. 16-22.
- [7] T. Nazir, M. Nawaz, M. Masood, A. Javed, "Copy move forgery detection and segmentation using improved mask region-based convolution network (RCNN)", *Applied Soft Computing*, Vol. 131, 2022, p. 109778.
- [8] C. You, H. Zheng, Z. Guo, T. Wang, X. Wu, "Tampering detection and localization base on sample guidance and individual camera device convolutional neural network features", *Expert Systems*, Vol. 40, No. 1, 2023, p. e13102.
- [9] G. Mariappan, A. R. Satish, P. B. Reddy, B. Maram, "Adaptive partitioning-based copy-move image forgery detection using optimal enabled deep neuro-fuzzy network", *Computational Intelligence*, Vol. 38, No. 2, 2022, pp. 586-609.
- [10] Y. Rao, J. Ni, W. Zhang, J. Huang, "Towards jpeg-resistant image forgery detection and localization via self-supervised domain adaptation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, pp. 1-12. (in press)
- [11] D. R. Ramji, C. A. Palagan, A. Nithya, A. Appathurai, E. J. Alex, "Soft computing-based color image demosaicing for medical Image processing", *Multimedia Tools Applications*, Vol. 79, 2020, pp. 10047-10063.
- [12] P. G. Sreelekshmi, M. Bhagavathi Priya, V. Vishu, "Deep forgery detect: enhancing social media security through deep learning-based forgery detection", *International Journal of Data Science and Artificial Intelligence*. Vol. 1, No. 1, 2023, pp. 9-19.
- [13] R. Ganeshan, S. Muppidi, D. R. Thirupurasundari, B. S. Kumar, "Autoregressive-Elephant Herding Optimization based Generative Adversarial Network for copy-move forgery detection with Interval type-2 fuzzy clustering", *Signal Processing: Image Communication*, Vol. 108, 2022, p. 116756.
- [14] S. Kumar, S. K. Gupta, M. Kaur, U. Gupta, "VI-NET: A hybrid deep convolutional neural network using VGG and inception V3 model for copy-move forgery classification", *Journal of Visual Communication and Image Representation*, Vol. 89, 2022, p. 103644.
- [15] C. You, H. Zheng, Z. Guo, T. Wang, X. Wu, "Tampering detection and localization base on sample guidance and individual camera device convolutional neural network features", *Expert Systems*, Vol. 40, No. 1, 2023, p. e13102.
- [16] S. Koul, M. Kumar, S. S. Khurana, F. Mushtaq, K. Kumar, "An efficient approach for copy-move image forgery detection using convolution neural network", *Multimedia Tools and Applications*, Vol. 81, No. 8, 2022, pp. 11259-11277.
- [17] H. Wu, J. Zhou, J. Tian, J. Liu, Y. Qiao, "Robust image forgery detection against transmission over online social networks", *IEEE Transactions on Information Forensics and Security*, Vol. 17, 2022, pp. 443-456.
- [18] S. Kumar, S. K. Gupta, U. Gupta, M. Agarwal, "Non-overlapping block-level difference-based image forgery detection and localization (NB-localization)", *The Visual Computer*, Vol. 39, No. 12, 2023, pp. 6029-6040.
- [19] S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, R. Sarkar, "ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection", *Expert Systems with Applications*, Vol. 210, 2022, p. 118423.
- [20] G. Nirmalapriya, B. Maram, R. Lakshmanan, M. Navaneethakrishnan, "ASCA-squeeze net: Aquila sine cosine algorithm enabled hybrid deep learning networks for digital image forgery detection", *Computers & Security*, Vol. 128, 2023, p. 103155.
- [21] R. D. Sushir, D. G. Wakde, S. S. Bhutada, "Enhanced blind image forgery detection using an accurate deep learning-based hybrid DCCA and ADFC", *Multimedia Tools and Applications*, Vol. 83, 2024, pp. 1725-1752.
- [22] S. Tyagi, D. Yadav, "MiniNet: a concise CNN for image forgery detection", *Evolving Systems*, Vol. 14, No. 3, 2023, pp. 545-556.
- [23] K. N. V. S. K. Vijayalakshmi, J. Sasikala, C. Shanmuganathan, "Copy-paste forgery detection using deep learning with error level analysis", *Multimedia Tools and Applications*, Vol. 83, No. 2, 2024, pp. 3425-3449.

Deep Reinforcement Learning for Dynamic Task Scheduling in Edge-Cloud Environments

Original Scientific Paper

D. Mamatha Rani*

TGSWRAFPDC(W), Bhongir
Department of Computer Science, Bhongir, Telangana- 508126, India
mamatha3004@gmail.com

Supreethi K P

Jawaharlal Nehru Technological University
Department of Computer Science and Engineering, Hyderabad, Telangana- 500085, India
supreethi.pujari@jntuh.ac.in

Bipin Bihari Jayasingh

CVR College of Engineering/IT Department
Hyderabad, Telangana- 501510, India
bipinbjayasingh@cvr.ac.in

*Corresponding author

Abstract – With The advent of the Internet of Things (IoT) and its use cases there is a necessity for improved latency which has led to edgecomputing technologies. IoT applications need a cloud environment and appropriate scheduling based on the underlying requirements of a given workload. Due to the mobility nature of IoT devices and resource constraints and resource heterogeneity, IoT application tasks need more efficient scheduling which is a challenging problem. The existing conventional and deep learning scheduling techniques have limitations such as lack of adaptability, issues with synchronous nature and inability to deal with temporal patterns in the workloads. To address these issues, we proposed a learning-based framework known as the Deep Reinforcement Learning Framework (DRLF). This is designed in such a way that it exploits Deep Reinforcement Learning (DRL) with underlying mechanisms and enhanced deep network architecture based on Recurrent Neural Network (RNN). We also proposed an algorithm named Reinforcement Learning Dynamic Scheduling (RLbDS) which exploits different hyperparameters and DRL-based decision-making for efficient scheduling. Real-time traces of edge-cloud infrastructure are used for empirical study. We implemented our framework by defining new classes for CloudSim and iFogSim simulation frameworks. Our empirical study has revealed that RLbDS out performs many existing scheduling methods.

Keywords: Task Scheduling, Edge-Cloud Environment, Recurrent Neural Network, Edge Computing, Cloud Computing, Deep Reinforcement Learning

Received: November 17, 2023; Received in revised form: June 12, 2024; Accepted: June 17, 2024

1. INTRODUCTION

Unprecedented growth of cloud-assisted use cases has led to compelling Cloud Service Providers (CSPs) to optimize resource usage in the presence of Service Level Agreements (SLAs). Ubiquitous adoption of technological innovation such as the Internet of Things (IoT) has led to the emergence of fog and edge computing phenomena which leverage latency. In the presence of IoT applications, the scheduling of tasks is challenging for many reasons such as network hierarchy, heterogeneity of resources, mobility of devices, resource-constrained devices and stochastic behaviour of nodes [1]. Tradition-

al cloud scheduling algorithms are not sufficient to harness the power of the dynamic computing environment made up of cloud, fog and edge resources. To overcome this problem, different scheduling algorithms came into existence. Reinforcement learning is one such technique used with the machine learning approach [2]. Many learning-based task scheduling approaches came into existence. Their merits and demerits are summarized in Table 1 and Table 2 in Section 2. The advantages of the research in [1] include consideration of dynamic environments and heterogeneous cores. However, it does not consider adaptive QoS, edge cloud, decentralized environment and presence of stochastic workloads.

The work in [3] considered edge cloud and also heterogeneous cores for their task scheduling research.

However, it does not support adaptive QoS, dynamic and decentralized environments, edge cloud and stochastic workloads. The merits of [4] include the consideration of dynamic environment, stochastic workloads and heterogeneous cores. But it lacks adaptive QoS, support for edge cloud and decentralized environments. The research in [5] and [6] has similar findings. Their method has provision for considering dynamic environments, heterogeneous cores, adaptive QoS and stochastic workloads. But is not designed for edge cloud and decentralized environments. In [7], there is consideration of dynamic environment, stochastic workloads, adaptive QoS and heterogeneous cores but does not support decentralized and edge-cloud environments. The work in [8] supports dynamic environments and stochastic workloads. However, it has limitations to deal with heterogeneous cores, adaptive QoS, edge cloud and decentralized environments. There is a similarity in the task scheduling methods proposed in [9] and [10].

Their methods are dynamic supporting adaptive QoS and stochastic workloads besides dealing with heterogeneous cores. However, they do not support decentralized and edge cloud environments. The scheduling research in [11] supports dynamic environments along with stochastic workloads. They also deal with heterogeneous cores and adaptive QoS. However, the drawback is that those methods do not consider decentralized and edge-cloud environments.

Concerning optimization parameters, Table 2 provides research gaps in existing solutions. Research in [1] is based on a heuristics approach and considers energy and SLA violation parameters. Their research lacks in the study of response time and cost of scheduling which are crucial for task scheduling. The work in [3] is also based on the heuristics method but considers cost and energy parameters. It does not throw light on response time and SLA violations. In [4], their method is based on Gaussian process regression and considers two parameters such as energy and SLAs. It has no support for optimization of cost and response time.

The task scheduling research in [5] and [6] is based on the Deep Queue Learning Network (DQN) method and supports cost and energy parameters for optimization. However, they have no optimization of SLAs and response time. In [7] Q-learning-based phenomenon is used considering energy and cost dynamics for optimization. However, it lacks optimization of response time and SLAs. Deep Neural Network (DNN) is the scheduling method used in [8] and it has support for optimization of cost and SLA parameters. It lacks support for energy and response time optimizations. The work in [9] and [10] is based on the Double DQN (DDQN) method and it supports only energy parameters for optimization. It lacks support for response time, cost and SLA optimizations. In [11] DRL method is used for task scheduling by

considering response time for optimization. However, it does not support the optimization of SLAs, cost and energy. From the literature, it is observed that there is a need for a more comprehensive methodology in edge-cloud environments for task scheduling. Our contributions to this paper are as follows.

1. We proposed a learning-based framework known as the Deep Reinforcement Learning Framework (DRLF). This is designed in such a way that it exploits Deep Reinforcement Learning (DRL) with underlying mechanisms and enhanced deep network architecture based on Recurrent Neural Network (RNN).
2. We proposed an algorithm named Reinforcement Learning Dynamic Scheduling (RLbDS) which exploits different hyperparameters and DRL-based decision-making for efficient scheduling.
3. Our simulation study has revealed that the proposed RLbDS outperforms many existing scheduling methods.

The remainder of the paper is structured as follows. Section 2 reviews prior works on existing task scheduling methods for cloud and edge-cloud environments. Section 3 presents details of the proposed system including the system model, DRL mechanisms and the underlying algorithm. Section 4 presents the results of the empirical study while Section 5 concludes our work and provides directions for the future scope of the research.

2. RELATED WORK

This section reviews prior works on existing task scheduling methods for cloud and edge-cloud environments. VM plays a vital role in cloud infrastructure for resource provisioning. Beloglazov and Bu proposed a method for improving resource utilization in the cloud through VM migration and consolidation. They found that VM live migration has the potential to exploit idle nodes in cloud data centres to optimize resource utilization and reduce energy consumption. They considered the dynamic environment and presence of heterogeneous cores for their task scheduling study. Their method is based on a heuristics approach. It considers SLA negotiations and algorithms designed to support optimizations such as energy efficiency and SLAs. Their algorithm monitors VMs and their resource usage. By considering VM consolidation and VM live migration, their method is aimed at reducing energy consumption and adherence to SLAs. This method lacks adaptive QoS and support for dynamic workloads. Pham and Huh [3] proposed a task scheduling method based on a heuristics approach for such an environment.

It is designed to work for heterogeneous cores in fog-cloud. They considered optimizations such as energy efficiency and cost reduction by scheduling tasks in an edge-cloud environment. Their algorithm is based on heuristics towards reducing cost and energy consump-

tion. It is based on graph representation. Towards this, their method exploits the task graph and processor graph. Given the two graphs representing tasks and resources, their method finds appropriate resource allocation for given tasks. It has a provision for determining task priority and then choosing the most suitable node for the execution of the task.

Bui et al. [4] proposed an optimization framework for the cloud with a predictive approach. They could predict the dynamics of resource utilization for scheduling by employing a method named Gaussian process regression. The prediction result helped them to minimize the number of servers to be used to process the requests leading to a reduction of energy usage. Their method is, however, based on heuristics and is not suitable for dynamic workloads and edge-cloud environments. Cheng et al. [2] explored DRL based approach towards task scheduling and resource provisioning in the cloud. They further optimized the Q-learning method to reduce the task rejection rate and improve energy efficiency. Huang et al. [5] and Mao et al. [6] followed the DRL approach for improving task scheduling performance in a cloud computing environment.

In [5] DRL based online offloading method is proposed based on deep neural networks. It is a scalable solution since it is a learning-based approach. In [6] DeepRM is the framework proposed for task scheduling considering efficient resource management. Both

methods are based on the DQN approach rather than heuristics. Both methods considered optimization parameters such as energy and cost. In other words, they are designed to reduce energy consumption and also the cost incurred for task execution in cloud environments. They support stochastic workloads and adaptive QoS. However, they do not support edge-cloud environments and do not optimize SLA and response time parameters.

Basu et al. [7] focused on the problem of live migration of VMs based on the RL-based Q-learning process. Their methodology improves live migration and heuristics-based existing approaches. Towards this end, their method exploits the Megh and RL-based model to have continuous adaptation to the runtime situations towards leveraging energy efficiency. Xu et al. [8] defined a DNN approach named LASER to support deadline-critical jobs with replication and speculative execution. Their implementation of the framework is designed for the Hadoop framework. Zhang et al. [9] defined a DDQN method towards energy efficiency in edge computing. It is based on the Q-learning process and also the dynamic voltage frequency scaling (DVFS) method that has the potential to reduce energy usage. As Q-learning is not able to recognize continuous system states, they extended it to have double-deep Q-learning. Table 1 shows provides a summary of findings among existing scheduling methods.

Table 1. Merits and demerits of existing scheduling methods compared with the proposed method

Reference	Dynamic	Stochastic Workload	Decentralized	Edge Cloud	Adaptive QoS	Heterogeneous
[1]	Yes	No	No	No	No	Yes
[3]	No	No	No	Yes	No	Yes
[4]	Yes	Yes	No	No	No	Yes
[5], [6]	Yes	Yes	No	No	Yes	Yes
[7]	Yes	Yes	No	No	Yes	Yes
[8]	Yes	Yes	No	No	No	No
[9], [10]	Yes	Yes	No	No	Yes	Yes
[11]	Yes	Yes	No	No	Yes	Yes
[18]	Yes	No	No	No	Yes	Yes
[19]	Yes	No	No	Yes	Yes	No
[20]	Yes	No	No	No	Yes	Yes
[21]	Yes	No	No	No	Yes	Yes
[22]	Yes	No	No	No	Yes	Yes
[23]	Yes	Yes	No	No	Yes	Yes
[25]	Yes	No	No	No	No	No
[26]	Yes	No	No	No	Yes	Yes
[27]	Yes	No	Yes	No	Yes	Yes
Proposed (RLbDS)	Yes	Yes	Yes	Yes	Yes	Yes

Similar to the work of [2], Mao et al. [6] employed DDQN for efficient resource management. This kind of work is also found in Li et al. [10]. Both have employed the DRL technique towards job scheduling over diversified resources. However, these learning-based methods are not able to withstand stochastic environments. Mao et al. [6] and Rjoubet al. [11] investigated DRL based approach for task scheduling in edge-cloud. However, they considered only response time in their research. Its drawback is that

they could not exploit asynchronous methods for optimization of their methods towards robustness and adaptability. There is a need to improve it by considering the dynamic optimization of parameters in the presence of stochastic workloads. Skarlat et al. [12] explored IoT service placement dynamics in fog computing resources while Pham et al. [13] focused on cost and performance towards proposing a novel method for task scheduling. Brogi and Forti [14] investigated on deployment of QoS-aware IoT

tasks in fog infrastructure. Task prioritization [15], DRL for resource provisioning [4, 7], energy-efficient scheduling using Q-learning [16] and DRL usage in 5G networks [17] are other important contributions.

As presented in Table 1, we summarize our findings leading to important research gaps. The summary is made in terms of different parameters such as dynamic environment, presence of stochastic workload, decentralized environment, usage of edge cloud, consideration for adaptive QoS and presence of heterogeneous cores for task scheduling. Table 1 also provides the proposed method and its merits over existing methods.

Almutairi and Aldossary [18] proposed a novel method for IoT tasks to offload in the edge-cloud ecosystem. It is designed to serve latency-sensitive applications in a better way. It has a fuzzy logic-based approach for inferring knowledge towards decision-making in the presence of resource utilization and dynamic resource utilization. Ding et al. [19] considered an edge-cloud environment to investigate stateful data stream applications. They proposed a method to judge state migration overhead and make partitioning decisions based on the dynamically changing network bandwidth availability. Murad et al.

[20] proposed an improved version of the min-min task scheduling method to deal with scientific workflows in cloud computing. It could reduce the minimum completion time besides optimizing resource utilization. Bulej et al. [21] did their research on the management of latency in the edge-cloud ecosystem towards better performance in task scheduling in the presence of dynamic workloads. It is designed to explore the upper bound of response time and optimize the performance further. Almutairi and Aldossary [22] proposed an edge-cloud system architecture to investigate modelling methodology on task offloading. It has offloading latency models along with various

offloading schemes. Their simulations are made using Edge CloudSim. They intend to improve it in future with fuzzy logic.

Zhang and Shi [23] explored workflow scheduling in an edge-cloud environment. They analyzed different possibilities in workflow scheduling in such an ecosystem. They opined that workflow applications need novel approaches in the scheduling process. Zhao et al. [24] focused on task scheduling along with security to prevent intrusions in edge computing environments. They considered low-rate intrusions and focused on preventing them along with task scheduling. It is a Q-learning-based approach designed to meet runtime requirements based on the learning process. Zhang et al. [25] proposed a time-sensitive algorithm that dynamically caters to the needs of deadline-aware tasks in edge-cloud environments. It considers job size and server capability in a given dynamic and hierarchical scenario. It is a multi-objective task considering execution time, cost and reduction of SLAs. Lakhan et al. [26] proposed a task scheduling approach for IoT tasks considering a hybrid mechanism consisting of task scheduling and task offloading. Singh and Bhushan [27] proposed a method for task scheduling based on Cuckoo Search Optimization (CSO). It has an integrated local search strategy. From these recent works, it is found that they targeted IoT kind of workflows in edge-cloud environments. There is Q-Learning used in one of the papers. However, deep reinforcement learning is not found in the latest works. Service placement in edge resources using DRL [28], dynamic scheduling [29] and task offloading [30] are other important contributions. Table 2 provides a summary of findings among existing scheduling methods in terms of optimization parameters. Magotra [41] focused on energy-efficient approaches in cloud infrastructures by developing adaptive solutions that could help the system towards proper VM consolidation, leading to better performance.

Table 2. Optimization parameters considered by existing scheduling methods

Reference	Method	Optimization Parameters			
		SLA Violations	Cost	Response Time	Energy
[1]	Heuristics	Yes	No	No	Yes
[3]	Heuristics	No	Yes	No	Yes
[4]	Gaussian Process Regression	Yes	No	No	Yes
[5], [6]	DQN	No	Yes	No	Yes
[7]	Q Learning	No	Yes	No	Yes
[8]	DNN	Yes	Yes	No	No
[9], [10]	DDQN	No	No	No	Yes
[11]	DRL (REINFORCE)	No	No	Yes	No
[18]	SJF	No	No	Yes	Yes
[19]	Cloud Computing	No	No	Yes	No
[21]	Cloud computing	No	Yes	Yes	Yes
[23]	CSA	No	Yes	No	No
[24]	Cloud computing	No	Yes	Yes	Yes
[25]	Cloud computing	No	Yes	No	No
[27]	CSP	No	Yes	Yes	No

As presented in Table 2, we summarized the existing methods in terms of optimization parameters and the approach considered in the task scheduling research. The optimization parameters considered for the comparative study of existing methods are SLA violations, cost, response time and energy.

Table 2 also provides the proposed method and its merits over existing methods. Table 1 and Table 2 provide very useful insights reflecting gaps in the research. Our work in this paper is based on such research gaps as those tables reveal the merits of the proposed system.

3. PROPOSED SYSTEM

We proposed a DRL-based framework for dynamic task scheduling in an edge-cloud environment. This section presents the framework and proposed algorithm besides DRL mechanisms.

3.1. PROBLEM DEFINITION

Considering an edge-cloud environment, let H be a collection of hosts denoted as $\{H_1, H_2, H_3, \dots, H_n\}$ where n indicates a maximum number of hosts. A task T can be assigned to host H . Scheduling is considered as the assignment of T to H . However, in terms of RL , the system state is mapped to an action. Here action does mean allocation of T to H . T may be an active task that could be migrated to a new H or a newly arrived task. At the beginning of an interval, denoted as SI_i , the system state initially is denoted as $state_i$ which reflects the hosts and their parameters, tasks yet to be allocated in the prior interval, denoted as $(a_{i-1} \setminus I_i)$ beside newly arrived tasks denoted as n_i . For each task, denoted as $a_i (= a_{i-1} \cup n_i \setminus I_i)$, the scheduler needs to take an action, denoted as $Action_i$, for the system interval SI_i in terms of either allocating it to a host or migrating to a new host. A task is satisfying Let $m_i \subseteq a_{i-1} \setminus I_i$ is considered a migratable task. A scheduler can be understood as a model which reflects a decision-making function $State_i \rightarrow Action_i$. Here loss function associated with the model for a given interval denoted as $Loss_i$ is computed based on task allocations. Therefore, the problem of realizing an optimal model is expressed in Eq. 1.

$$\begin{aligned} & \text{Minimize}_{Model} \sum_i Loss_i \\ & \text{Subject to } \forall i, Action_i = Model(State_i) \\ & \forall i \forall T \in m_i \cup n_i, \{T\} \leftarrow Action_i(T) \end{aligned} \quad (1)$$

Different notations used in our work are presented in Table 3.

3.2. OUR SYSTEM MODEL

We considered infrastructure or resources for scheduling in an edge-cloud environment. The resources are heterogeneous. Edge resources are nearby while cloud resources reside in a remote data centre. Therefore, each host in the infrastructure is different in response time and computational power. Edge resources are closer and exhibit low response times but they do have limited resources and computational power. Cloud resources

take more response time but they do have high computational power. Our system model is presented in Fig. 1.

The edge and cloud nodes are part of computing resources. These resources are managed by the resource management module. This module has several components or sub-modules to deal with resource management either directly or indirectly. The scheduler module is responsible for either scheduling a task T to a host H or migrating a task from one host to another host based on runtime dynamics. The dynamic workload is generated by IoT devices being used by different users. The workload contains several tasks with varied requirements. Resource management module takes the workload and follows DRL based (learning-based) approach in task allocation or task migration. These decisions are based on the ideal objective functions and the requirements associated with tasks. The requirements may include deadline, bandwidth, RAM and CPU.

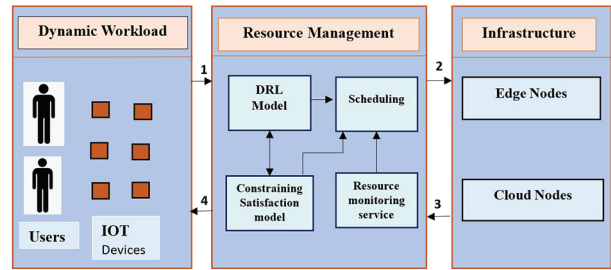


Fig. 1. Our system model

The workload is generated automatically to evaluate the functionality of the proposed system. Our system has a DRL model which influences the scheduler module in decision-making. There are multiple schedulers to be used at runtime to serve dynamically generated workloads. In the process, there is the distribution of workload among hosts leading to faster convergence. Each resource in edge-cloud accumulates local gradients associated with corresponding schedulers besides synchronizing them to update models. The DRL module follows asynchronous updates. The constraint satisfaction module takes suggestions as input from DRL and finds whether it is valid. Here valid does mean a task is in migration or the host's capacity is optimally being used.

3.3. WORKLOAD GENERATION

We generate workload programmatically to evaluate the proposed system. Since IoT devices and user's demands are dynamic, there is a change in the bandwidth and computational requirements of tasks. The whole execution time in our system is divided into several scheduling intervals. Each interval is assumed to have the same duration. SI_i denotes the i^{th} scheduling interval. This interval has a start time and end time denoted as t_i and t_{i+1} respectively. Each interval has active tasks associated with it. They are the tasks being executed and denoted as a_i . The tasks that have been completed at the beginning of the interval are denoted as I_i while

newly arrived tasks that are dynamically generated by the workload generator are denoted as n_i .

3.4. OUR LEARNING-BASED APPROACH FOR SCHEDULING

We proposed a framework known as the Deep Reinforcement Learning Framework (DRLF), as shown in Fig. 2, which exploits a learning-based approach using the DRL model for dynamic task scheduling in an edge-cloud environment. The framework supports several scheduling intervals. The framework has a workload generator which generates tasks (n_i) and gives them to the scheduling and migration module. The tasks given to the scheduler are in turn given to the resource monitoring module which schedules new tasks and migrates existing tasks if required to ensure optimal resource utilization, load balancing and latency in task completion. The scheduler activity changes the state of the edge-cloud environment.

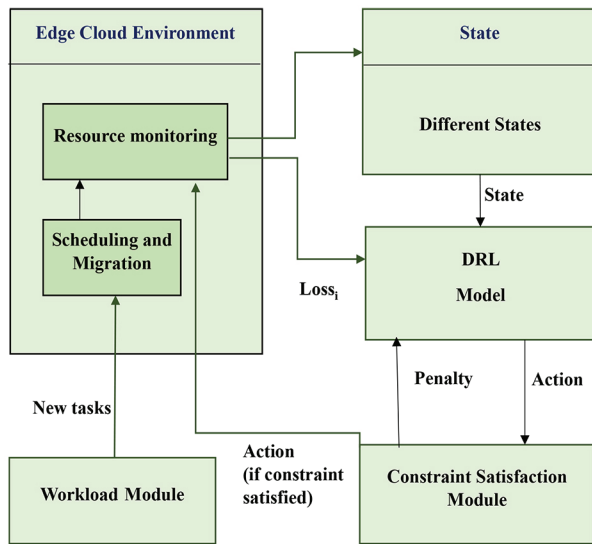


Fig. 2. Proposed Deep Reinforcement Learning Framework (DRLF) for task scheduling in edge-cloud environment

Every time $State_i$ is updated by the resource monitoring module it is given to the DRL model. The state information consists of hosts' feature vectors, new tasks n_i and the rest of the tasks associated with the previous interval and denoted by $(a_{i-1} \setminus V_i)$. The resource monitoring module also gives $Loss_i$ data to the DRL model. The DRL model suggests an action, denoted as $Action_{i-1}^{PG}$, based on the state information to the constraint satisfaction module and updates parameters as expressed in Eq. 2. This module then determines $Penalty_i$ to the DRL model.

$$Loss_i^{PG} = Loss_i + Penalty_i \quad (2)$$

This process continues iteratively. Once the constraint is satisfied, the constraint satisfaction module gives the suggested action (Action) by the DRL model to the resource management module. It then computes $Penalty_{i+1}$ about SI_{i+1} the next scheduling interval.

Table 3. Notations used in our work

Notation	Description
a_i	Indicates a set of active tasks linked to SI_i
H_i	Indicates i^{th} host in a given set of hosts
I_i	Indicates the initial set of tasks of SI_i
m_i	Indicates a decision for task migration
n_i	It indicates a task allocation decision
$Action_i^{PG}$	Scheduling actions at the beginning of SI_i
$Loss_i^{PG}$	Loss function at the beginning of SI_i
SI_i	Denotes i^{th} scheduling interval
T_i^S	It indicates i^{th} in a given set of tasks
$\{T\}$	Indicates the host to which task T has been assigned
AEC	Average Energy Consumption
AMT	Average Migration Time
ART	Average Response Time
Hosts	Indicates a collection of hosts in the edge-cloud environment
N	Indicates the maximum number of hosts
T	Denotes a task to be executed

Based on the action received from the constraint satisfaction module, the resource management module either allocates a new task to a specific host or migrates tasks, denoted as $(a_{i-1} \setminus V_i)$, of the preceding interval. This will result in an update from a_{i-1} to a_i . Then the tasks associated with a_i are executed for SI_i and the cycle continues for SI_{i+1} .

3.5. DEEP LEARNING ARCHITECTURE

The DRL model is built based on an enhanced Recurrent Neural Network (RNN) architecture. It has the functionality to achieve reinforcement learning. In the process, it approximates $State_i$ towards $Action_i^{PG}$ which is an action bestowed from the DRL model to the constraint satisfaction module for a given scheduling interval. The enhanced RNN can ascertain temporal relationships between input space and output space. This deep learning architecture is shown in Fig. 3. After each interval, cumulative loss and policy are predicted by a single network

The network has two fully connected layers, denoted as $fc1$ and $fc2$, configured. These are followed by three recurrent layers, denoted as $r1$, $r2$ and $r3$, with skip connections. The given 2D input is flattened and sent to dense layers. The output of $r3$ is given to two fully connected layers denoted as $fc3$ and $fc4$. The $fc4$ outputs a 2D vector of 100×100 . It does mean that the model can deal with 100 tasks allocated to 100 hosts in cloud infrastructure. Eventually, a softmax function is employed to the second dimension to have values $[0,1]$ and the resultant value in a row becomes 1. For interpretation O_{jk} denoting a probability map, indicates that there is a probability of a task T_j^{ai} being assigned to H_k . At the $fc4$, a cumulative loss function $Loss_{i+1}^{PG}$ is computed. The layers in the network are made up of a Gated Recurrent Unit that have the capacity to model the temporal dimension of a given task and also the

characteristics of the host comprising of bandwidth, RAM and CPU. The Gated Recurrent Unit (GRU) layers tend to have increased network parameters leading to complexity. This problem is addressed by exploiting skip connections towards gradient propagation faster.

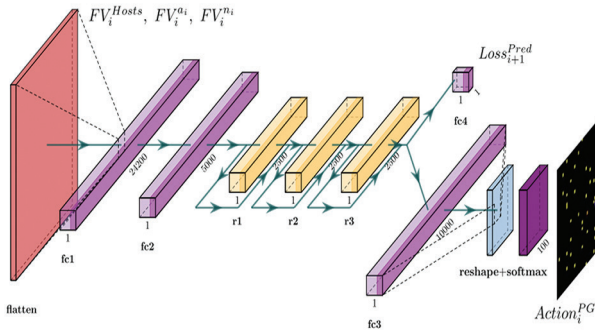


Fig. 3. Architecture of an RNN variant used to realize the DRL model

This model takes $State_i$ as input which is represented in the form of a 2D vector. This vector contains a continuous element FV_i^{Hosts} , and another continuous element $FV_i^{n_i}$ and $FV_i^{a_{i-1} \setminus V_i}$ has categorical host indices. Therefore, pre-processing is required to transform host indices into one hot vector with a maximum size of n . Then there is a need for the concatenation of all feature vectors. Afterwards, each element in the resultant vector is subjected to normalization based on a range of values $[0, 1]$. Each element has a feature denoted as f_e while min_{f_e} and max_{f_e} denote their minimum and maximum values respectively. These values are computed relying on the dataset with the help of two heuristics namely local regression and maximum migration time. Afterwards, standardization is carried out feature-wise using the expression in Eq. 3.

$$e = \begin{cases} 0 & \text{if } max_{f_e} = min_{f_e} \\ \min \left(1, \max \left(0, \frac{e - min_{f_e}}{max_{f_e} - min_{f_e}} \right) \right) & \text{otherwise} \end{cases} \quad (3)$$

Once pre-processing of the given input is carried out, it is fed to the network (Fig. 3) where it first flattens the pre-processed input before sending it through dense layers. The output of these layers is transformed into $Action_i^{PG}$. We employed a backpropagation algorithm to ascertain the biases and weights of the network. The learning rate is kept adaptive from 10 to 2 and later on, $1/10^{th}$ value based on reward change associated with the preceding 10 iterations is not greater than 0.1. Automatic differentiation is exploited to modify the parameters of the network using $Loss_i^{PG}$ as a reward. Gradients of local networks are accumulated across the edge nodes periodically in an asynchronous fashion towards the update of global network parameters. Towards this end, a gradient accumulation rule expressed in Eq. 4 is followed.

$$d\theta \leftarrow d\theta - \alpha \nabla_{\theta} \log[\pi(State_i; \theta')(Loss_i^{PG} + CLoss_{i+1}^{pred})] + \alpha \nabla_{\theta} (Loss_i^{PG} + CLoss_{i+1}^{pred} - CLoss_{i+1}^{pred})^2 \quad (4)$$

Where local and global network parameters are denoted as θ' and θ respectively, it has a log term to indicate a change direction in parameters and the $(Loss_i^{PG} + CLoss_{i+1}^{pred})$ term denotes cumulative loss predicted in a given episode that begins with $State_i$. Mean Square Error (MSE) is a gradient associated with the cumulative loss predicted. Finally, there is the transformation of output from $Action_i^{PG}$ to $Action_i$ by the constraint satisfaction module and the same is given to the resource management module.

3.6. Algorithm Design

We proposed an algorithm to realize the optimal scheduling of given tasks in the edge-cloud ecosystem. It is presented in Algorithm 1.

Algorithm: Reinforcement Learning based Dynamic Scheduling (RLbDS)

Inputs:

Size of batch **B**

Maximum intervals for scheduling N

1. Begin
2. For each interval n in N
3. IF $n \% B == 0$ and $n > 1$ Then
4. Compute loss function
5. $Loss_i^{PG} = Loss_i + Penalty_i$
6. Use $Loss_i^{PG}$ in the network (Fig. 3) for backpropagation
7. End If
8. $State_i \leftarrow PreProcess(State_i)$
9. Feed $State_i$ to the network (Fig. 3)
10. $pMap \leftarrow Output$ of RL model (network as in Fig. 3)
11. $(Action_i, Penalty_{i+1}) \leftarrow ConSatMod(pMap)$
12. Resource monitoring module takes $action$
13. DRL model takes $Penalty_{i+1}$
14. ResourceMonitoring($Action_i$) migrates active task
15. Execution of all tasks in interval n in edge-cloud
16. End For
17. End

Algorithm 1. Reinforcement Learning based Dynamic Scheduling (RLbDS)

The algorithm takes the size of batch B and maximum intervals for scheduling N and performs optimal scheduling of given tasks of every interval in edge-cloud resources. The algorithm exploits the enhanced RNN network (Fig. 3) to update the model from time to time towards making DRL-based decisions for scheduling. At each interval of scheduling, there is an iterative process for taking care of pre-processing and feeding the state to the DRL model. Based on the action suggested by DRL, the constraint satisfaction module specifies a penalty when there is an ideal scheduling decision, that is notified to resource monitoring which schedules new tasks and also performance migration of active tasks based on the decisions rendered.

3.7. LOSS FUNCTION COMPUTATION

In the proposed learning model we want to optimize, in each interval, with minimal $Loss_i$. The model is also designed to adapt to the state that dynamically changes while mapping $State_i$ to $Action_i$. Towards this end, $Loss_i$ is a metric defined to update model parameters. Besides different metrics that result in normalized value 0 or 1 are defined. Average energy consumption is a metric defined as the edge cloud resources have different sources of energy as discussed in [32]. The consumed energy by host $h \in Hosts$ is multiplied by a factor $\alpha_h \in [0, 1]$ that is associated edge-cloud deployment strategy. The normalized AEC is computed as in Eq. 5.

$$AEC_i^{Hosts} = \frac{\sum_{h \in Hosts} \alpha_h \int_{t=t_i}^{t_{i+1}} p_h(t) dt}{\sum_{h \in Hosts} \alpha_h p_h^{max}(t_{i+1}-t_i)} \quad (5)$$

Where the power function of host h is denoted by $p_h(t)$ linked to time and its maximum possible power is denoted as p_h^{max} .

Average response time is another metric defined to be used for interval SI_i . ART for all tasks is normalized by maximum response time. ART is computed as in Eq. 6.

$$ART_i = \frac{\sum_{t \in I_{i+1}} \text{Response Time}(t)}{|I_{i+1}| \max_i \max_{t \in I_i} \text{Response Time}(t)} \quad (6)$$

The average migration time metric is defined for a given SI_i . It reflects all tasks' average migration time in the interval normalized by maximum migration time. AMT is computed as in Eq. 7.

$$AMT_i = \frac{\sum_{t \in a_i} \text{Migration Time}(t)}{|a_i| \max_i \max_{t \in I_i} \text{Response Time}(t)} \quad (7)$$

Cost (C) is yet another metric defined for SI_i . It indicates the total incurred cost in the interval and is computed as in Eq. 8.

$$Cost_i = \frac{\sum_{h \in Hosts} \int_{t=t_i}^{t_{i+1}} C_h t(t) dt}{\sum_{h \in Hosts} C_h^{max}(t_{i+1}-t_i)} \quad (8)$$

Average SLA violation is another metric for SI_i . It reflects SLA violation dynamics as expressed in Eq. 9.

$$SLVA_i = \frac{\sum_{t \in I_{i+1}} SLA(t)}{|I_{i+1}|} \quad (9)$$

To minimize the resultant value for all the aforementioned metrics, as used in [16] and [33], the $Loss_i$ metric is defined as expressed in Eq. 10.

$$Loss_i = \alpha. AEC_{i-1} + \beta. ART_{i-1} + \gamma. AMT_{i-1} + \delta. Cost_{i-1} + \epsilon. SLVA_{i-1} \quad (10)$$

such that $\alpha, \beta, \gamma, \delta, \epsilon \geq 0 \wedge \alpha + \beta + \gamma + \delta + \epsilon = 1$.

Different users can have varied QoS needs and hyper-parameters ($\alpha, \beta, \gamma, \delta, \epsilon$) need to be set with different values. As discussed in [33], [34] and [35] it is important to optimize energy consumption in cloud infrastructure. Therefore, it is essential to optimize loss. Even when other metrics are compromised, it is possible to optimize loss. In such a case, the loss can have $\alpha = 1$ while

the other metrics can have 0. As discussed in [36] traffic management and healthcare monitoring are sensitive to response time. In such cases, loss can have $\beta = 1$ while other measures can have 0. In the same fashion, setting hyper-parameters is application-specific.

$$Loss_i^{PG} = Loss_i + Penalty_i \quad (11)$$

As specified in the works such as [37] and [38], the penalty is to be included in neural network modes. With the penalty, the model can update parameters towards minimizing $Loss_i$ and ensure constrained satisfaction. Therefore, for neural network loss function is defined as in Eq. 11.

4. RESULTS AND DISCUSSION

This section presents our simulation environment, the dataset used and the results of experiments.

4.1. SIMULATION SETUP

We built a simulation application using Java language. The IDE used for development is the IntelliJ Idea 2022 version. CloudSim [39] and iFogSim [40] libraries are used to have a simulation environment. Scheduling intervals are considered equal to be compatible with other existing works [4, 7, 41]. Cloudlets or tasks are generated programmatically from the Bitbrain dataset collected from [42].

The two simulation tools such as iFogSim and CloudSim are extended with required classes to facilitate the usage of cost, response time and power parameters associated with edge nodes. New modules are created to incorporate simulation of IoT devices with mobility with delayed task execution, variations in bandwidth and communication with deep learning model. Additional classes are defined to have constraint satisfaction modules and also take care of input formats, output formats and pre-processing. Based on the provision in CloudSim, a loss function is implemented. The dataset collected from [43] has traces of real workload run on Bitbrain infrastructure. This dataset contains logs of workloads of more than 1000 VMs associated with host machines. The workload information contains time-stamp, RAM usage, CPU usage, CPU cores requested, disk, network and bandwidth details. This dataset is available at [44] to reproduce our experiments. The dataset is divided into 75% and 25% VM workloads for training and testing respectively. Training deep learning model is done with the former while the latter is used to test the network and analyse results.

4.2. ANALYSIS OF RESULTS

We evaluated the performance of the proposed algorithm named RLbDS by comparing it with state-of-the-art methods such as Local Regression and Minimum Migration Time (LR-MMT) [41], Median Absolute Deviation and Maximum Correlation Policy (MAD-MC) [41], DDQN [44] and REINFORCE [9]. LL-MMT works for dynamic workloads

considering minimum migration time and local regression. It has heuristics to have task selection and overhead detection. MAD-MC is also a dynamic scheduler which is based on maximum correlation and median absolute deviation heuristics. DDQN is a deep learning-based approach that exploits RL to schedule tasks. DRL method is also based on RL which is based on policy gradient. The results reveal the sensitivity dynamics hyperparameters, such as $(\alpha, \beta, \gamma, \delta, \epsilon)$, of the proposed RLbDS about model learning and its impact on different performance metrics.

Model training is given with 10 days of simulations while testing is carried out with 1-day simulation time.

4.2.1. Impact of Hyperparameters on RLbDS

The performance of the proposed algorithm named RLbDS is analysed with loss function associated with many hyperparameters such as $(\alpha, \beta, \gamma, \delta, \epsilon)$. Experiments are made with value 1 set to each of the hyperparameters. The rationale behind this is that when the value is set to 1, it could provide optimal performance.

Table 4. Performance of RLbDS with different hyper parameters

Model Parameters	Total Energy (Watts)	Time (milliseconds)	Fraction of SLA Violations	Total Cost (USD)	Time (seconds)	Number of completed tasks
$\alpha=1$	1.37	8.5	0.17	6305.5	4.45	815
$\beta=1$	1.43	8.18	0.17	6306.5	4.3	830
$\gamma=1$	1.51	8.8	0.148	6307.5	3.65	845
$\delta=1$	1.38	8.78	0.178	6304.5	4.15	810
$\epsilon=1$	1.44	8.22	0.134	6307.8	3.75	850

As presented in Table 5, the performance of RLbDS is provided in terms of the number of performance metrics.

Table 5. Performance of RLbDS compared against existing algorithms

Models	Total Energy (Watts)	Time (milliseconds)	Fraction of SLA Violations	Total Cost (US Dollar)	Time (seconds)	Number of completed tasks
LR-MMT	0.959	8.58	0.06	6325	4.5	700
MAD-MC	0.95	8.4	0.13	6325	4.3	800
DDQN	0.85	8.8	0.07	6325	4	850
REINFORCE	0.82	8.35	0.06	6300	3.8	850
RLbDS	0.73	7.7	0.04	6000	3.3	1000

Loss function with different hyperparameters has its influence on the performance of the RLbDS algorithm as presented in Fig. 4. The network learning process differs with changes in hyperparameters. Energy consumption differed when the loss function used different hyperparameters. With $\alpha=1$ RLbDS consumed 1.37 watts, with $\beta=1$ it needed 1.43 watts, with $\gamma=1$ the algorithm consumed 1.51 watts, with $\delta=1$ it required 1.38 watts and with $\epsilon=1$ RLbDS consumed 1.44 watts. The least energy is consumed when $\alpha=1$ (all energy consumption values are given in $1*10^8$ format). The average response time of the algorithm RLbDS is influenced by each hyperparameter. With $\alpha=1$ RLbDS required 8.5 milliseconds, with $\beta=1$ it needed 8.18 milliseconds, with $\gamma=1$ the algorithm needed 8.8 milliseconds, with $\delta=1$ it required 8.78 milliseconds and with $\epsilon=1$ RLbDS required 8.22 milliseconds. The least response time is recorded when $\beta=1$.

SLA violations are also studied with these hyperparameters. It is observed that they influence a fraction of SLA violations. With $\alpha=1$ the fraction of SLA violations caused by RLbDS is 0.17, with $\beta=1$ also it is 0.17, with $\gamma=1$ the algorithm showing 0.148, with $\delta=1$ it is 0.178, and with $\epsilon=1$ RLbDS caused by 0.134. The last fraction of SLA violations is recorded when $\epsilon=1$. The total cost is

also analysed in terms of USD (as per the pricing calculator of Microsoft Azure [45]).

It was observed earlier that hyperparameters have an impact on energy consumption. Since energy consumption attracts the cost of execution in the cloud, obviously these parameters have an impact on the cost incurred. With $\alpha=1$ the total cost exhibited by RLbDS is 6305.5, with $\beta=1$ it is 6306.5, $\gamma=1$ the algorithm showed 6307.5, with $\delta=1$ it is 6304.5, and with $\epsilon=1$ RLbDS caused 6307.8. The least cost is recorded when $\delta=1$.

Average task completion time is also analysed with different hyperparameters. With $\alpha=1$ the average task completion time exhibited by RLbDS is 4.45 seconds, with $\beta=1$ it is 4.3, with $\gamma=1$ the algorithm showed 3.65, with $\delta=1$ it is 4.15, and with $\epsilon=1$ RLbDS caused 3.75. The least average task completion time is recorded when $\gamma=1$ (all average task completion values are given in $1*10^6$ format). The total number of tasks completed with scheduling done by RLbDS is also influenced by hyperparameters. With $\alpha=1$ the number of completed tasks achieved by RLbDS is 815, $\beta=1$ it is 830, $\gamma=1$ the algorithm showed 845, with $\delta=1$ it is 810, and with $\epsilon=1$ RLbDS showed 850 tasks to be completed. The least number of completed tasks is recorded when $\delta=1$.

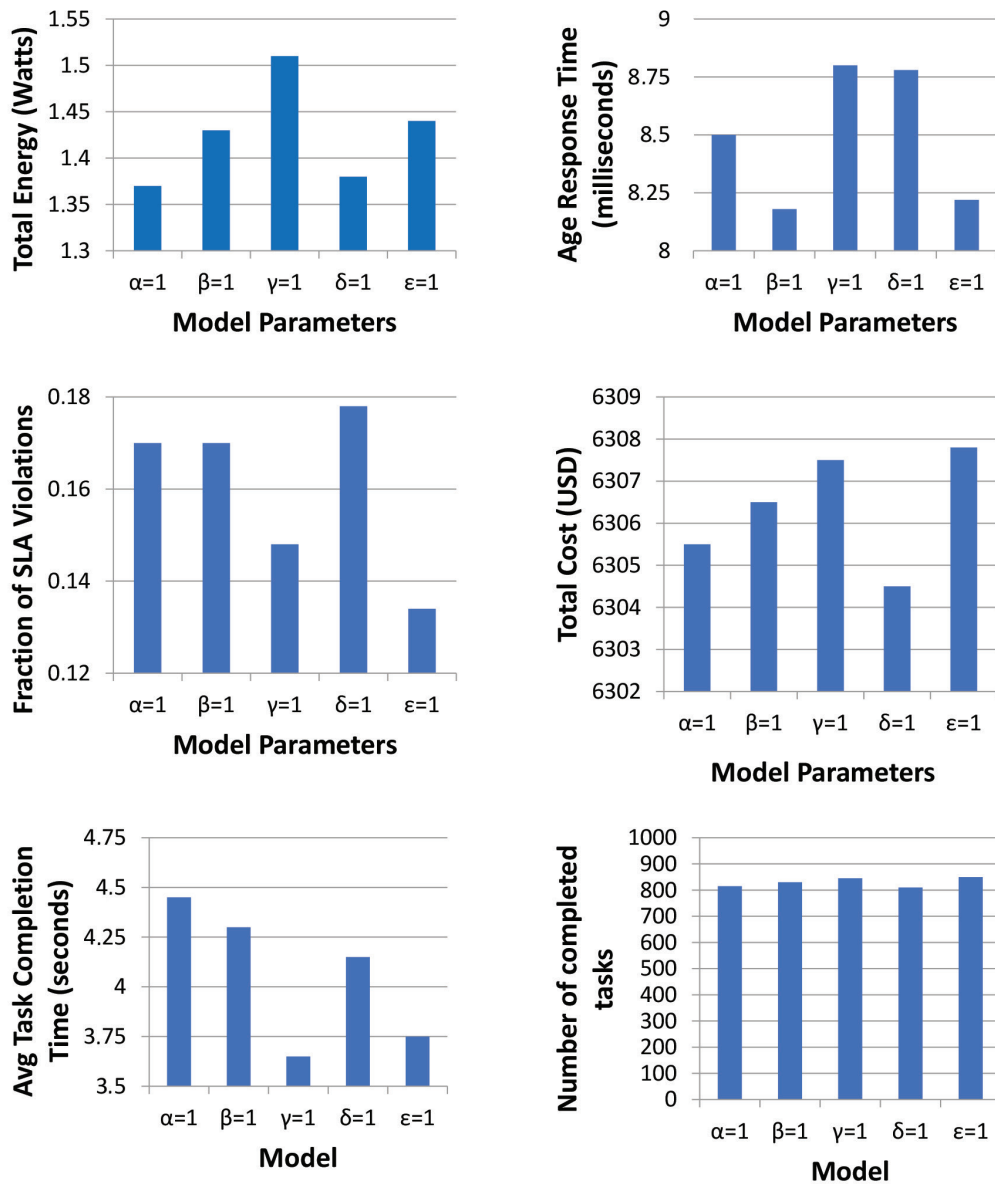


Fig. 4. Performance dynamics of proposed RLbDS algorithm with different model parameters associated with loss function

4.2.2. Performance Comparison with State of the Art

Our algorithm RLbDS is compared against several existing algorithms as presented in Fig. 5. Total energy consumption values are provided in 1×10^8 watts format. LR-MMT algorithm consumed 0.959, MAD-MC 0.95, DDQN 0.85, REINFORCE 0.82 and the proposed RLbDS consumed 0.73. The energy consumption of RLbDS is found to be the least among the scheduling algorithms. Average response time is another metric used for comparison. LR-MMT algorithm exhibited an average response time of 8.58 milliseconds, MAD-MC 8.4, DDQN 8.8, REINFORCE 8.35 and the proposed RLbDS required 7.7 milliseconds. The average response time of RLbDS is found to be the least among the scheduling algorithms. SLA violations are another important metric used for comparison. LR-MMT algorithm exhibited a fraction of SLA violations as 0.06, MAD-MC 0.13,

DDQN 0.07, REINFORCE 0.06 and the proposed RLbDS exhibited 0.04. The fraction of SLA violations of RLbDS is found least among the scheduling algorithms.

Algorithm compared with the state-of-the-art

Total cost in terms of USD is another metric used for comparison. This metric is influenced by energy consumption. LR-MMT algorithm needs 6325 USD, MAD-MC 6325, DDQN 6325, REINFORCE 6300 and the proposed RLbDS needed 6000 USD. The total cost of RLbDS is found least among the scheduling algorithms. Concerning average task completion time, the LR-MMT algorithm needs 4.5 seconds, MAD-MC 4.3, DDQN 4, REINFORCE 3.8 and the proposed RLbDS requires 3.3 seconds. The average task completion time of RLbDS is found to be the least among the scheduling algorithms (average task completion time is given in 1×10^6 seconds format). The number of completed tasks is another observation made in our empirical study.

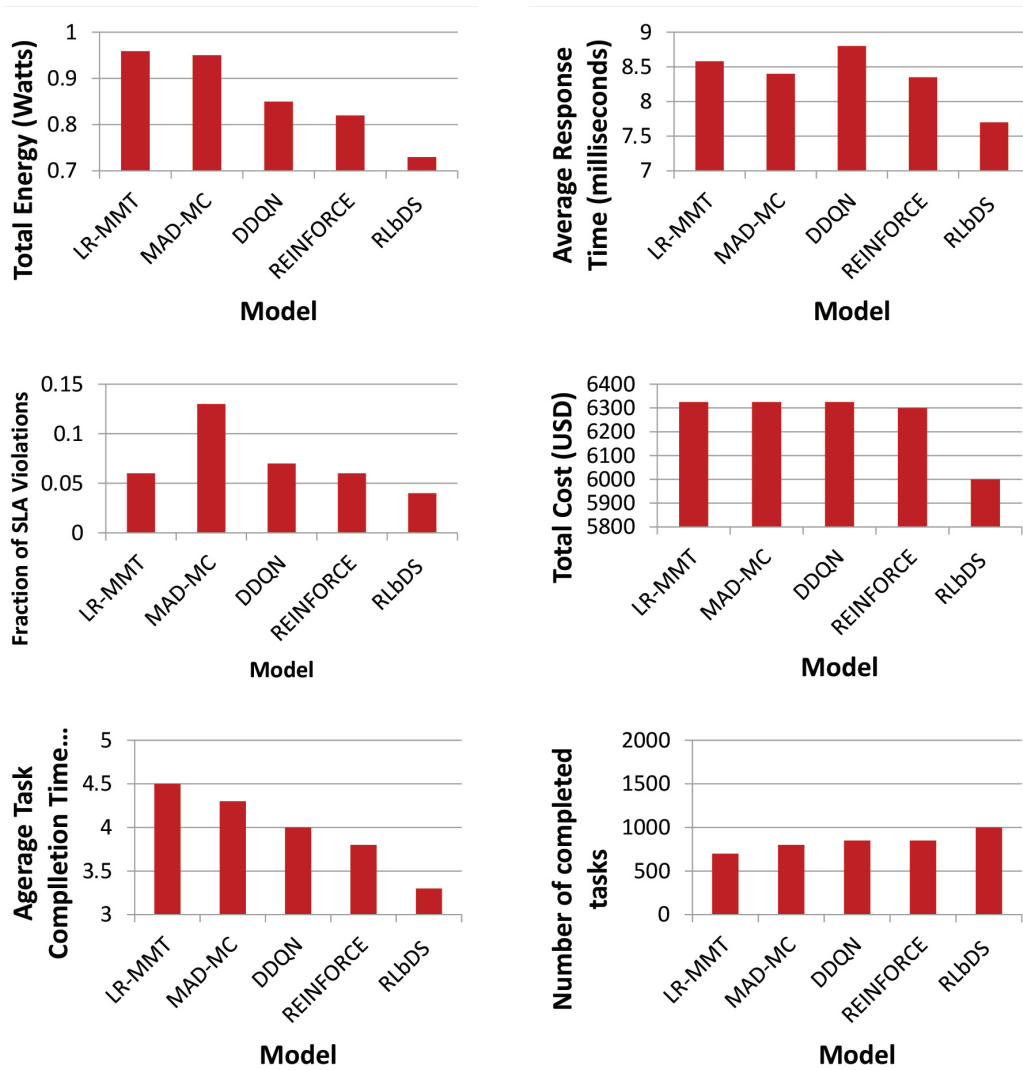


Fig. 5. Performance of proposed RLbDS algorithm compared with the state of the art

LR-MMT completed 700 tasks, MAD-MC 800, DDQN 850, REINFORCE 850 and the proposed completed 1000 tasks. The average task completion time of RLbDS is found to be the least among the scheduling algorithms.

4.2.3. Performance with Number of Recurrent Layers

Considering optimal values for hyperparameters scheduling overhead and loss dynamics against the number of

recurrent layers are analysed. Overhead is computed as the ratio between the total duration of execution and the time taken for scheduling. Empirical study has revealed that the number of recurrent layers in the proposed architecture (Fig. 3) influences the loss and overhead.

As presented in Table 6, loss value and scheduling overhead against several recurrent layers are observed. Loss value and scheduling overhead are analysed against several recurrent layers as presented in Fig. 6.

Table 6. Performance against the number of recurrent layers

Number of recurrent layers	Performance	
	Loss value	Scheduling overhead (%)
0	3.69	0.009
1	3.4	0.010
2	2.9	0.010
3	2.6	0.010
4	2.5	0.019
5	2.4	0.029

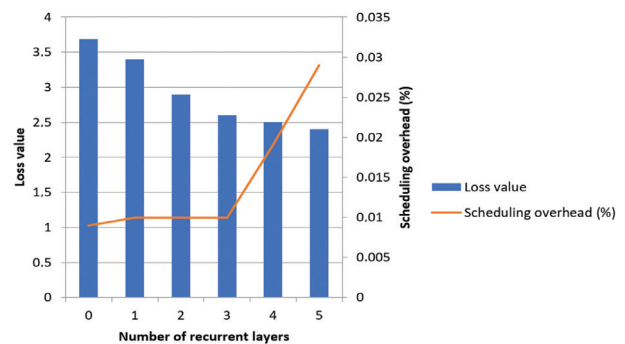


Fig. 6. Performance analysis with the number of recurrent layers

Several layers influence the loss value. Loss value decreases (performance increases) as the number of layers is increased. However, the scheduling overhead is increased with the number of recurrent layers.

4.2.4. Scalability Analysis

The scalability of the proposed algorithm is analysed in terms of speedup and efficiency. The analysis is made against the number of hosts. As presented in Table 7, the performance of the proposed algorithm in terms of its scalability is provided.

Table 7. Scalability analysis

Number of recurrent layers	Performance	
	Speed-up	Efficiency
1	1	1
5	5	0.8
10	9	0.785
15	13	0.775
20	17	0.765
25	19	0.725
30	21	0.7
35	23	0.650
40	25	0.630
45	26	0.570
50	27	0.525

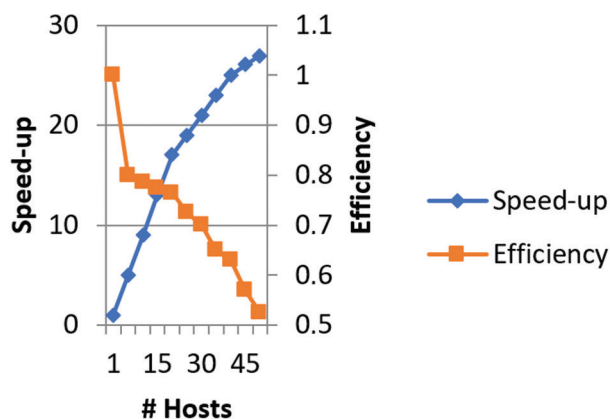


Fig.7. Scalability analysis in terms of speedup and efficiency

There is a trade-off observed between scalability and efficiency as presented in Figure 7. When the number of hosts is increased, there is a gradual decrease in efficiency while there is a gradual increase in speedup. From the experimental results, it is observed that the proposed RLbDS is found to be dynamic and can adapt to runtime situations as it is a learning-based approach.

Its asynchronous approach helps it in faster convergence. In the presence of dynamic workloads and device characteristics, RLbDS adapts to changes with ease.

5. CONCLUSION AND FUTURE WORK

We proposed a learning-based framework known as the Deep Reinforcement Learning Framework (DRLF).

This is designed in such a way that it exploits Deep Reinforcement Learning (DRL) with underlying mechanisms and enhanced deep network architecture based on Recurrent Neural Network (RNN). We also proposed an algorithm named Reinforcement Learning Dynamic Scheduling (RLbDS) which exploits different hyperparameters and DRL-based decision-making for efficient scheduling. Real-time traces of edge-cloud infrastructure are used for empirical study. We implemented our framework by defining new classes for CloudSim and iFogSim simulation frameworks. We evaluated the performance of the proposed algorithm named RLbDS by comparing it with state-of-the-art methods such as LR-MMT, MAD-MC, DDQN and REINFORCE. The results reveal the sensitivity dynamics hyperparameters, such as $(\alpha, \beta, \gamma, \delta, \epsilon)$, of the proposed RLbDS about model learning and its impact on different performance metrics. Our empirical study has revealed that RLbDS outperforms many existing scheduling methods. In future, we intend to improve our framework for container scheduling and load balancing.

6. REFERENCES

- [1] A. Beloglazov, R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centres", *Concurrency and Computation: Practice and Experience*, Vol. 24, No. 13, 2012, pp. 1397–1420.
- [2] M. Cheng, J. Li, S. Nazarian, "DRL-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers", *Proceedings of the 23rd Asia and South Pacific Design Automation Conference*, Jeju, Korea, 22-25 January 2018, pp. 129-134.
- [3] X.-Q. Pham, E.-N. Huh, "Towards task scheduling in a cloud-fog computing system", *Proceedings of the 18th Asia-Pacific Network Operations and Management Symposium*, Kanazawa, Japan, 5-7 October 2016, pp. 1–4.
- [4] D.-M. Bui, Y. Yoon, E.-N. Huh, S. Jun, S. Lee, "Energy efficiency for a cloud computing system based on predictive optimization", *Journal of Parallel and Distributed Computing*, Vol. 102, 2017, pp. 103-114.
- [5] L. Huang, S. Bi, Y. J. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks", *IEEE Transactions on Mobile Computing*, Vol. 19, No. 11, 2020, pp. 2581-2593.
- [6] H. Mao, M. Alizadeh, I. Menache, S. Kandula, "Resource management with deep reinforcement

learning", Proceedings of the 15th ACM Workshop on Hot Topics in Networks, Atlanta, GA, USA, 9-10 November 2016, pp. 50-56.

- [7] D. Basu, X. Wang, Y. Hong, H. Chen, S. Bressan, "Learn-as-you-go with Megh: Efficient live migration of virtual machines", IEEE Transactions on Parallel and Distributed Systems, Vol. 30, No. 8, 2019, pp. 1786-1801.
- [8] M. Xu, S. Alamro, T. Lan, S. Subramaniam, "Laser: A deep learning approach for speculative execution and replication of deadline-critical jobs in the cloud", Proceedings of the 26th International Conference on Computer Communication and Networks, Vancouver, BC, Canada, 31 July - 3 August 2017, pp. 1-8.
- [9] Q. Zhang, M. Lin, L. T. Yang, Z. Chen, S. U. Khan, P. Li, "A double deep Q-learning model for energy-efficient edge scheduling", IEEE Transactions on Services Computing, Vol. 12, No. 5, 2019, pp. 739-749.
- [10] F. Li, B. Hu, "Deepjs: Job scheduling based on deep reinforcement learning in the cloud data centre", Proceedings of the 4th International Conference on Big Data and Computing, Guangzhou, China, 10-12 May 2019, pp. 48-53.
- [11] G. Rjoub, J. Bentahar, O. A. Wahab, A. S. Bataineh, "Deep and reinforcement learning for automated task scheduling in large-scale cloud computing systems", Concurrency and Computation: Practice and Experience, Vol. 33, No. 23, 2020, pp.1-14.
- [12] O. Skarlat, M. Nardelli, S. Schulte, M. Borkowski, P. Leitner, "Optimized IoT service placement in the fog", Service Oriented Computing and Applications, Vol. 11, No. 4, 2017, pp. 427-443.
- [13] X.-Q. Pham, N. D. Man, N. D. T. Tri, N. Q. Thai, E.-N. Huh, "A cost-and performance-effective approach for task scheduling based on collaboration between cloud and fog computing", International Journal of Distributed Sensor Networks, Vol. 13, No. 11, 2017, pp. 1-16.
- [14] A. Brogi, S. Forti, "QoS-aware deployment of IoT applications through the fog", IEEE Internet of Things Journal, Vol. 4, No. 5, 2017, pp. 1185-1192.
- [15] T. Choudhari, M. Moh, T.-S. Moh, "Prioritized task scheduling in fog computing", Proceedings of the ACMSE Conference, New York, NY, USA, March 2018, pp. 22:1-22:8.
- [16] Q. Zhang, M. Lin, L. T. Yang, Z. Chen, P. Li, "Energy-efficient scheduling for real-time systems based on deep learning model", IEEE Transactions on Sustainable Computing, Vol. 4, No. 1, 2017, pp. 132-141.
- [17] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, L.-C. Wang, "Deep reinforcement learning for mobile 5g and beyond Fundamentals, applications, and challenges", IEEE Vehicular Technology Magazine, Vol. 14, No. 2, 2019, pp. 44-52.
- [18] J. Almutairi, M. Aldossary, "A novel approach for IoT tasks offloading in edge-cloud environments. Journal of Cloud Computing", Journal of Cloud Computing, Vol. 10, 2021, p. 28.
- [19] S. Ding, L. Yang, J. Cao, W. Cai, M. Tan, Z. Wang, "Partitioning Stateful Data Stream Applications in Dynamic Edge Cloud Environments", IEEE Transactions on Services Computing, Vol. 15, No. 4, 2021, pp. 2368-2381.
- [20] S. S. Murad, R. Badeel, N. S. A. Alsandi, Rafi, "Optimized Min-Min Task Scheduling Algorithm For Scientific Workflows In A Cloud Environment", Journal of Theoretical and Applied Information Technology, Vol. 100, No. 2, 2022, pp. 480-506.
- [21] L. Bulej et al. "Managing latency in edge cloud environment", Journal of Systems and Software, Vol. 172, 2021, pp. 1-15.
- [22] J. Almutairi, M. Aldossary, "Investigating and Modeling of Task Offloading Latency in Edge-Cloud Environment. Computers", Materials & Continua, Vol. 68, No. 3, 2021, pp. 1-18.
- [23] R. Zhang, W. Shi, "Research on Workflow Task Scheduling Strategy in Edge Computer Environment", Journal of Physics: Conference Series, Vol. 1744, 2021, pp. 1-6.
- [24] X. Zhao, G. Huang, L. Gao, M. Li, Q. Gao, "Low load DIDS task scheduling based on Q-learning in an edge computing environment", Journal of Network and Computer Applications, Vol. 188, 2021, pp. 1-12.
- [25] Y. Zhang, B. Tang, J. Luo, J. Zhang, "Deadline-Aware Dynamic Task Scheduling in Edge-Cloud Collaborative Computing", Electronics, Vol. 11, 2022, pp. 1-24.
- [26] A. Lakhani et al. "Delay Optimal Schemes for Internet of Things Applications in Heterogeneous Edge Cloud Computing Networks", Sensors, Vol. 22, pp. 1-30.
- [27] M. Singh, S. Bhushan, "CS Optimized Task Scheduling for Cloud Data Management", International Journal of Engineering Trends and Technology, Vol. 70, No. 6, 2022, pp. 114-121.

- [28] Y. Hao, M. Chen, H. Gharavi, Y. Zhang, K. Hwang, "Deep Reinforcement Learning for Edge Service Placement in Softwarized Industrial Cyber-Physical System", *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 8, 2021, pp. 5552-5561.
- [29] S. Tuli et al. "Dynamic Scheduling for Stochastic Edge-Cloud Computing Environments using A3C learning and Residual Recurrent Neural Networks", *IEEE Transactions on Mobile Computing*, Vol. 21, No. 3, 2022, pp. 1-15.
- [30] Q. Zhang, L. Gui, S. Zhu, X. Lang, "Task Offloading and Resource Scheduling in Hybrid Edge-Cloud Networks", *IEEE Access*, Vol. 9, 2021, pp. 940-954.
- [31] L. Roselli, C. Mariotti, P. Mezzanotte, F. Alimenti, G. Orecchini, M. Virili, N. Carvalho, "Review of the present technologies concurrently contributing to the implementation of the Internet of things (IoT) paradigm: RFID, green electronics, WPT and energy harvesting", *Proceedings of the Topical Conference on Wireless Sensors and Sensor Networks*, San Diego, CA, USA, 25-28 January 2015, pp. 1-3.
- [32] S. Tuli, N. Basumatary, S. S. Gill, M. Kahani, R. C. Arya, G. S. Wander, R. Buyya, "Healthfog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments", *Future Generation Computer Systems*, Vol. 104, 2020, pp. 187-200.
- [33] S. Sarkar, S. Misra, "Theoretical modelling of fog computing: a green computing paradigm to support IoT applications", *IET Networks*, Vol. 5, No. 2, 2016, pp. 23-29.
- [34] Z. Abbas, W. Yoon, "A survey on energy conserving mechanisms for the Internet of things: Wireless networking aspects", *Sensors*, Vol. 15, No. 10, 2015, pp. 24818-24847.
- [35] P. Kamalinejad, C. Mahapatra, Z. Sheng, S. Mirabbasi, V. C. Leung, Y. L. Guan, "Wireless energy harvesting for the Internet of things", *IEEE Communications Magazine*, Vol. 53, No. 6, 2015, pp. 102-108.
- [36] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, P. Liljeberg, "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach", *Future Generation Computer Systems*, Vol. 78, 2018, pp. 641-658.
- [37] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization", *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, 6-11 August 2017, pp. 22-31.
- [38] R. Doshi, K.-W. Hung, L. Liang, K.-H. Chiu, "Deep learning neural networks optimization using hardware cost penalty", *Proceedings of the IEEE International Symposium on Circuits and Systems*, Montreal, QC, Canada, 22-25 May 2016, pp. 1954-1957
- [39] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, R. Buyya, "Cloudsim: a toolkit for modelling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", *Software: Practice and Experience*, Vol. 41, No. 1, 2011, pp. 23-50.
- [40] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh, R. Buyya, "ifogsim: A toolkit for modelling and simulation of resource management techniques in the internet of things, edge and fog computing environments", *Software: Practice and Experience*, Vol. 47, No. 9, 2017, pp. 1275-1296.
- [41] Bhagyalakshmi Magotra. (2023). "Adaptive Computational Solutions to Energy Efficiency in Cloud Computing Environment Using VM Consolidation". *Archives of Computational Methods in Engineering*. (2022), pp.1790-1818
- [42] S. Shen, V. van Beek, A. Iosup, "Statistical characterization of business-critical workloads hosted in cloud datacenters", *Proceedings of the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, Shenzhen, China, 4-7 May 2015, pp. 465-474.
- [43] Bitbrain Dataset, <http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains> (accessed: 2024)
- [44] Q. Zhang, M. Lin, L. T. Yang, Z. Chen, P. Li, "Energy efficient scheduling for real-time systems based on the deep q-learning model", *IEEE Transactions on Sustainable Computing*, Vol. 4, No. 1, 2017, pp. 132-141.
- [45] Microsoft Azure Pricing Calculator, <https://azure.microsoft.com/en-au/pricing/calculator/> (accessed: 2024)

Scaling and Dynamic Resource Reallocation in NFV: Challenges and Research Perspectives

Review Paper

Tung Thanh Hoang

Electric Power University, Faculty of Information Technology
Hanoi, Vietnam
tunght@epu.edu.vn

Linh Manh Pham*

VNU University of Engineering and Technology,
Faculty of Information Technology, Department of Network and Computer Communications
Cau Giay, Hanoi, Vietnam
linhmp@vnu.edu.vn

Hoai Son Nguyen

VNU University of Engineering and Technology,
Faculty of Information Technology, Department of Network and Computer Communications
Cau Giay, Hanoi, Vietnam
sonnh@vnu.edu.vn

*Corresponding author

Abstract – Network Function Virtualization (NFV) has brought incredible experiences for Internet users and network operators. NFV enables the implementation of Virtualized Network Functions (VNFs) as software running in High Volume Servers (HVSs) to execute a Service Function Chain (SFC) to satisfy service demands of Internet users. During the execution of SFCs, VNFs and Virtual Links (VLs) tend to change their resource requirements due to the dynamic nature of the end user's demands. In this paper, we focus on dynamic resource allocation to the elements of SFC throughout the SFC process to adapt to the elasticity in demand from users by providing an overall picture of NFV and the scaling problem of SFC. We then review and analyze related studies on dynamic resource allocation of NFV systems during SFC operation and analyze the results of these projects. The most recent works are also classified based on several criteria to highlight their approaches, achievements, and also shortcomings. Finally, we introduce some research directions to deal with the scaling problem during SFC operation that needs more attention from researchers to inspire future work in the elastic operation of NFV-enabled systems.

Keywords: elasticity, network function virtualization, optimization, resource reallocation, scaling

Received: May 25, 2024; Received in revised form: September 11, 2024; Accepted: September 11, 2024

1. INTRODUCTION

1.1. MOTIVATION

The Internet has achieved incredible development in recent years, the number of Internet users is constantly increasing to reach 5.3 billion users (approximately 66% of the global population) by 2023 [1]. Therefore, the network infrastructures are continuously improved to meet the increasing needs of users.

Traditional network systems are mainly built from dedicated hardware devices such as routers, load balancers, firewalls, etc. The need to continuously upgrade infrastructure to satisfy the demand of Internet users

creates pressure on Capital Expenditure (CAPEX), such as buying equipment, space for placing devices, and Operation Expenditure (OPEX), such as electricity bills or labor expenses for Network Service Providers. Additionally, the operating of physical appliances is also a waste of physical resources, while this approach only enables a single user to use a device at a time instead of sharing resources to leverage idle resources for others. Network Function Virtualization was developed to overcome the limitations of traditional networks.

NFV was first introduced by the European Telecommunications Standards Institute (ETSI) [2], decoupling network functions (e.g., firewall, Intrusion Detection System (IDS), Network Address Translation (NAT), load

balancer, etc.) from their dedicated hardware by deploying these network functions as software on High Volume Servers (Fig. 1) and then providing them to tenants. These network functions are now called Virtual Network Functions.

The release of NFV comes with several benefits, including: *i) Reducing resource wastage*: The nature of NFV is virtualization. By virtualizing physical resources (i.e., compute, storage, network), controllers in NFV systems can flexibly allocate and reallocate resources provided to VNFs. As a result, idle physical resources should always be utilized; *ii) Elasticity*: User demands are dynamic and may cause variations in system resource consumption. Because the resource allocation

in NFV is flexible, as mentioned above, changes in service requests from users can be easily satisfied by granting more or releasing resources to these service requests; *iii) Minimizing CAPEX and OPEX*: Using NFV, service providers can reduce investments when buying Commercial Off-the-Shelf (COST) servers instead of spending money on high-cost dedicated hardware to deploy their system. Next, most NFV platforms are open-source, which means they are free. Additionally, automation mechanisms in NFV also reduce human operational activities; *iv) Fast error remediation*: Because elements in NFV systems are 'soft', administrators can quickly fix errors when unexpected things occur. Furthermore, the system can be easily re-implemented in the worst cases.

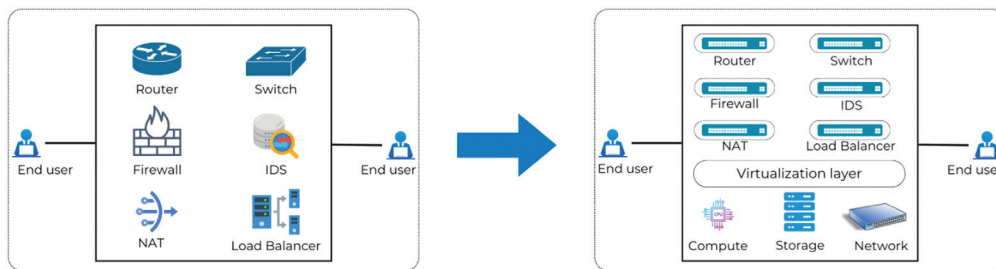


Fig. 1. The concept of virtualizing dedicated network devices to be software

1.2. RELATED WORK

As part of activities on the Internet, Internet users send data from and to the equipment. Traffic flows between these devices may need to pass through several network functions. In NFV-enabled systems, the service provider assigns VNFs to traffic flows to complete network services as expected by users. A combination of VNFs and possibly Physical Network Functions (PNFs) in a specific order can create a so-called Service Function Chain [3] or Network Service (NS) to satisfy the user's demand, as illustrated in Fig. 2. In this article, 'service function chain' and 'network service' are used equivalently.

Recent surveys focused on providing an overall picture of NFV. In the survey of Bo et al., the authors explained NFV concepts, terminologies, and architecture of NFV and introduced some projects that address hot topics in NFV such as VNF placement, scheduling, migration, chaining, and multicast [4]. Herrera et al. introduced a complete survey of the resource allocation in NFV within three stages: VNF Chain Composition, Embedding and Scheduling [5]. Yang et al. explored challenges and opportunities and offered some potential research directions in security issues for NFV [6]. The paper of Yanghao et al. [7] presented the variants of the resource allocation problem in NFV and provided a basic and standard mathematical model for the resource allocation problem for SFCs. The authors also offered some prominent research trends.

Other surveys [8], [9], [10], [11] placed emphasis on the placement of constituents (e.g., VNFs, CNFs (Con-

tainer Network Functions), VMs (Virtual Machines), etc.) of NSs at the initialization of SFC. In which the authors attempted to explore solutions to answer the question: "What is the best strategy for the placement problem to get the highest system performance with the lowest cost (in terms of minimizing the volume of resources that servers and links provide to SFC's elements)?".

1.3. OUR CONTRIBUTIONS AND PAPER ORGANIZATION

In recent years, along with the elasticity in cloud computing, the issue of flexibility in resource allocation for the operation of SFCs has also been the subject of concern. Especially, in the context of SFCs, it is always necessary to adjust to changes in tenant requirements. However, whereas most surveys are paying attention to resource allocation at the SFC's initialization, there is no overall picture of resource reallocation during the SFC's operation due to the dynamicity of requests from users, although this problem has become a hot trend in the last several years, as we will point out in sub-section 2.3.

In this paper, we focus on studying aspects related to dynamic resource reallocation to elements of SFCs during the operation of SFCs to meet the flexibility of end-user needs.

The main contributions of this survey are summarized as follows:

- An overview of NFV-enabled systems is presented, as well as a clarification of the scale issue encountered when operating virtual functions.

- We review and analyze the most recent work in dealing with the SFC scaling problem based on four aspects: The problem that the project tries to solve, the proposed solutions of the authors, the measured parameters in the research, and the experimental results. We also point out several deficiencies of existing research.
- Finally, according to the analysis results, we summarize existing solutions using a comparison table.

Since this comparison, we offer and explain in detail some potential research directions in this field.

The rest of this paper is organized as follows: We analyze the background of NFV in Section 2, including NFV architecture and the scaling problem in the operation of SFCs, and then we briefly describe most recent works in the field. Previous efforts are summarized based on specific criteria and recommend several promising avenues in the field in Section 3. Section 4 concludes our work.

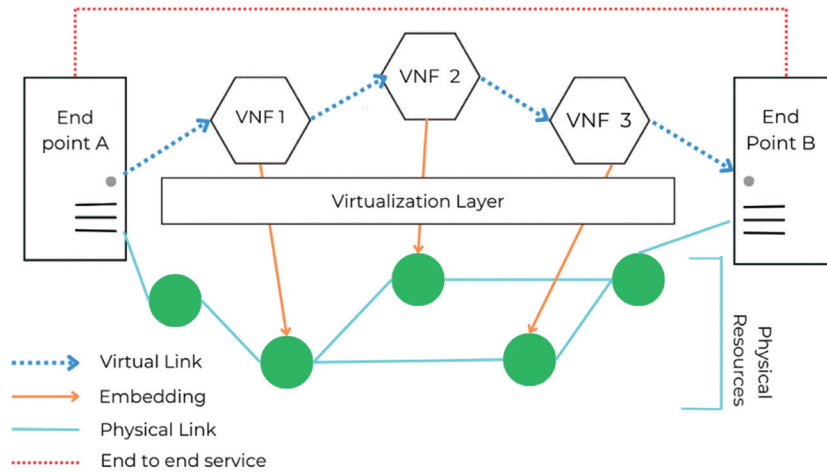


Fig. 2. An example of an SFC [12]

2. BACKGROUND AND RELATED WORK

2.1. NFV ARCHITECTURE

Network Function Virtualization is a network architecture where network functions are managed and deployed on hosts (physical computers, virtual machines, or containers) instead of traditional dedicated physical devices. Virtualization technology plays a key role in NFV, in which physical resources (e.g., compute, storage, network) are abstracted by hypervisors before allocated to upper layers. Figure 3 depicts the NFV architecture model consisting of the components as follows:

The NFV Infrastructure (NFVI) block consists of:

The *hardware and software infrastructure* that provides the platform for deploying VNFs. They are: i) servers that provide compute or storage capabilities [13], these servers can be either physical or virtual; ii) network facilities, including connection devices, transmission media, and network cards.

The *virtualization layer* is right above the hardware resources for the purpose of abstracting the underlying physical resources to create virtual resources. Therefore, this layer is also called a hypervisor. Currently, there are several well-known hypervisors on the market, such as KVM, Microsoft Hyper-V, ESXi, Xen, etc.

Virtualization infrastructure is virtualized resources that are abstracted by the hypervisors. It includes virtual compute (i.e., CPU), virtual storage (i.e., RAM), and virtual network (i.e., bandwidth). These resources

mainly constitute a virtualization environment to implement VNFs.

VNF layer: This layer plays a vital role in the NFV system. In this layer, network functions are deployed on virtualized resource platforms as software. VNFs perform network functions such as NAT, firewall, load balancing, etc., replacing traditional physical devices in the network. Each VNF may consist of one or several VNF Components (VNFC), which are orchestrated by the corresponding Element Management (EM). EM collects information about the operation of VNFs to provide to the VNF manager (VNFM). The set of EMs will make up an Element Management System (EMS). The market for VNFs is highly tremendous, including some notable names such as Suricata for IDS, HAproxy for load balancers, Open vSwitch for switches, etc.

The Management and Orchestration (MANO) block is a constituent of three sub-blocks:

Virtualization Infrastructure Manager (VIM): VIM manages and coordinates the virtualized resources of the system.

VNF Manager (VNFM) is responsible for managing VNFs, including: i) VNF Lifecycle Management (LCM); ii) VNF configuration management of the configuration parameters of a VNF/VNFC; iii) VNF information management for the value changes of VNF-related indicators; iv) VNF Performance Management (PM); v) VNF Fault Management (FM). VNFM can be deployed to manage a single VNF or a group of VNFs.

NFV Orchestrator (NFVO) in-charges of: i) handling the lifecycle management of NSs and their constituents; ii) NS performance measurements and NS fault management; iii) onboarding and management of Network Service Descriptors (NSDs) (detailed in Section 2.3); iv) onboarding and management of PNF Descriptor archives; v) onboarding and management of VNF Packages; vi) management of software images.

Operation Support System/Business Support System (OSS/BSS) is a system that supports the operation of the NFV system by interacting with operators and customers.

2.2. SCALING IN NFV

To complete a service request from a client, VNFs are logically connected to form an SFC. For example, traffic in a video conference session may need several network functions such as load balancing, video encoding, HTTP services, etc. Fig. 2 depicts an example of an SFC. In this example, traffic from end point A to end point

B must be handled by three network functions, VNF 1, VNF 2, and VNF 3. A combination of three VNFs in this order forms an SFC.

To implement an SFC, the controller needs to determine a forwarding graph, which is called VNF Forwarding Graph (VNFFG) based on the physical forwarding graph between physical nodes. The mapping of physical network forwarding graph and SFC is depicted in Fig. 4. In which, to form an SFC from user U1 to user U2 consisting of 3 VNFs in sequence: $E \rightarrow B \rightarrow A$, a virtual path will be constructed, starting from U1, pass through server 1, server 3, and server 4 in the cloud environment, before reaching U2 at the end of the path. Each server serves as a physical node for VNFs, and we must be aware that they can reside in different data centers and belong to multiple service providers. The path between the physical nodes is called a physical link, whereas the connection of ordered VNFs: E, B, and A is called a virtual link. An SFC is the constitution of VNFs and VLs.

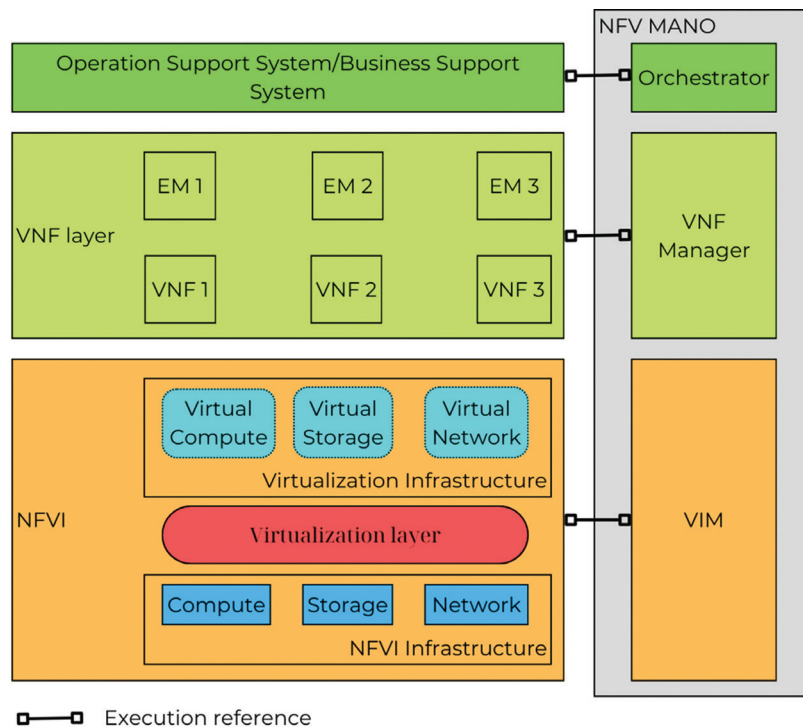


Fig. 3. NFV reference architecture [14]

Because user demands are constantly changing, tenants may have to increase or decrease their service requests by changing the volume of traffic or the quality of service while using the network service. Due to this variability, during the operation of the SFC, VNFs and VLs can be overloaded and need more resources or underloaded and need to be revoked resources to adapt to the change and to use efficiently the resources of servers and links. This leads to a phenomenon called 'scaling'.

Scaling in SFC is the term for VNFs and VLs that need to add/release resources (scaling up/down) [15] or need to add/delete the instances of VNFs (scaling

out/in) [16], [17] as shown in Fig. 5. These concepts of scaling up/down (Fig. 5b) or scaling in/out (Fig. 5c) are also known as vertical scaling and horizontal scaling, respectively. For more detail, in Fig. 5b, VNF B requires more virtual resources (i.e., vCPU, vRAM) to handle larger volumes of traffic flows and the node hosting this VNF will grant more after checking its remaining capacity. Similarly, when the volume of data flow decreases, VNF B will return redundant resources to the physical node to enhance resource utilization. Horizontal scaling is depicted in Fig. 5c, which means that to deal with the increase in ingress data flow, other replicas of VNF B are deployed on other nodes and the traffic will be split

in a certain ratio to be transmitted to all instances of VNF B. Figure 5d illustrates migration. It is a term referring to the movement of VNFs from one server to another.

In this situation, the current instance of VNF B is terminated and the virtual link from VNF A to VNF B will be rerouted to other instances of VNF B on other nodes.

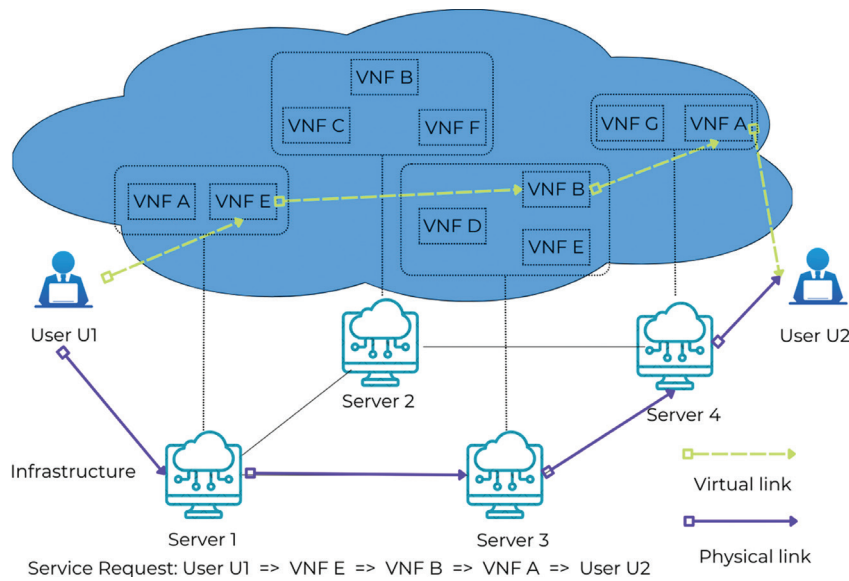


Fig. 4. A mapping of physical network and Service Function Chain [4]

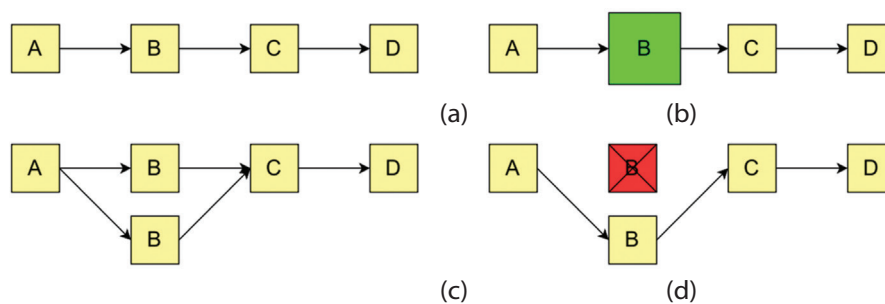


Fig. 5. VNF scaling models. (a) Service Function Chain, (b) Vertical Scaling, (c) Horizontal scaling, (d) Migration

To conclude, to guarantee the normal operation of NSs, controllers may need to: i) allocate more resources to VNFs participating in SFCs in case of scaling-up; ii) release resources granted to VNFs when scaling-down; iii) deploy more VNF instances on other servers and then split traffic flow with a determined proportion to pass through these instances in case of scaling-out; iv) delete VNF instances on idle servers to save resources in case of scaling-in; v) delete VNFs from resource-starving servers and deploy them on other servers that have enough capability, then re-direct the traffic flow to a new server in case of migration; and vi) no scale. Together with VNFs, VLs are also impacted. VLs that connect to and from the scaled VNF also need to be adjusted bandwidth to adapt with an increment or decrement of the VNF. Additionally, horizontal scaling and migration also occur when VLs between VNFs starve resources. In that case, VNFs related to those VLs also need to be deployed in other servers to reduce the amount of traffic across overloaded VLs.

In NFV systems, management operations take place in the MANO block. VIM is responsible for receiving

virtual resources from hypervisors and granting them to VNFs, whereas VNFM performs activities to manage VNFs, including initializing and terminating VNFs. During the process of creating VNFs and constructing SFCs, the initial resource requirements of VNFs are declared in the VNF Descriptor (VNFD), as described in subsection 2.3. These initial resource indices, along with operational system metrics such as transmission latency, data traffic, etc., serve as input factors for scaling-related decisions such as triggering scaling events, requesting additional resources, etc. For example, on the OpenStack open-source platform, Tacker runs as MANO and is executed on controller nodes to deploy and coordinate NFV-related tasks such as VIM registration, VNF creation, SFC construction, and resource allocation/reallocation.

2.3. SFC SCALING SOLUTION

One of the objectives of organizations when implementing NFV-enabled networks is to optimize the resource utilization of the system. The resource allocation problem in NFV is divided into three stages during the cre-

ation of the network system [5], including: i) considering the constraint between VNFs in the SFC chain (VNF chain composition problem); ii) determining the best place for the deployment of VNFs on physical servers (VNF Embedding problem); and iii) scheduling VNFs operations (VNF Scheduling). Previous works emphasized these stages [18], [19], [20], [21], [22]. Some papers took into account the flexibility factor when allocating resources for VNFs during deployment to ensure a minimum amount of resources for the operation of VNFs [23], [24], [25].

The appropriate resource allocation solution must be determined during the system deployment and service chain initialization stages. However, resource reallocation during system operation and SFC implementation should also be considered. SFCs must be added/released resources for a variety of reasons, including failures, security concerns, and changes in user needs during operation. In case of force majeure, it may be necessary to move one or some VNFs to another server to avoid service disruptions. In recent years, many researchers have begun to pay attention to the problem of scalability of VNFs (as well as VFs) and flexible resource reallocation for the operation of the SFC.

Hui et al. proposed a model to increase the success rate of scaling by developing an algorithm called ElasticNFV based on two main ideas: i) allocating more vCPU and vMemory to VNFs when needed; ii) in case there are not enough resources to grant more, move the VNFs to other servers [26]. ElasticNFV provides a Two-Phase Minimal Migration (TPMM) algorithm to minimize the migration time and embedding cost of VNF replicas. The experimental results showed that the TPMM algorithm outperforms two previous solutions, Sandpiper [27] and Oktopus [28], in terms of migration time and cost. At the same time, in a small test bed, ElasticNFV also achieved better resource utilization than FreeFlow [17]. The proposed algorithm is fine-grained when combining vertical scaling and migration to have efficient utilization of resources, a short scaling period, and a fast response time. However, the work can be improved by leveraging horizontal scaling. In which the controller can replicate more VNF instances and then tear the traffic to pass through both the current VNF instance and its new replicas.

The article [29] offered a mechanism called ENVI (Elastic resource Flexing for Network function Virtualization) that uses both features at the VNF level and infrastructure level to construct a machine-learning-based decision engine that can detect VNFs that need to be scaled. This mechanism continuously collects information about VNFs and their resource utilization, then it is fed to train a neural network. The evaluation shows that the neural network model outperforms other classification models such as decision tree, random forest, and logistic regression, with measures of accuracy, precision, recall, Receiver Operating Characteristic (ROC), and Area Under ROC curve (AUROC) and therefore can be a promising approach for scaling detection.

Zhao et al. [30] presented a model that considers the resource utilization of physical machines and the transmission delay of SFCs in response to the scaling out of SFCs. This model begins by continuously collecting information about the resource usage of the SFC, comparing it to a predetermined threshold, and deciding whether to scale in or out. VNFs in the scale-in list will be turned off to save resources, and the algorithm will prepare specifications for VNFs in the scale-out list to deploy them on other servers. The proposed algorithm proves that it brings about better physical machine resource utilization and transmission delay in comparison with traditional greedy algorithms. However, whereas the main idea of the algorithm is to migrate resource-starving VNFs to other nodes, scaling problems are affected by many complicated factors and can be sophisticatedly treated with other scaling models such as vertical scaling and horizontal scaling before migration.

The paper of Toosi et al. [31] dealt with techniques to solve resource utilization problems of SFCs using a resource threshold and the algorithm based on this threshold to decide whether the resources of the chain need to be augmented or reduced or not. To evaluate the performance of the solution, the authors defined two baselines: NoScale-Min represents the performance parameters of the system when the amount of resources is set as the initial value of the system, and NoScale-Max represents system parameters when providing maximum resources and assuming that Service-Level Agreement (SLA) violations never occur. What both baselines have in common is that they represent the performance statement of the system if the system is not scaled. Experimental results show that the ElasticSFC algorithm brings the SLA violation rate and the average response time of the requests close to that of NoScale-Max while saving resources by dynamically allocating resources based on workload. The migration phase is simply performed by finding the closest node to the traffic flow to deploy new VNF instances. This can be improved by implementing VNF replacement with an optimized VNF placement mechanism.

Dong et al. proposed a hybrid solution called HSM (Hybrid Scaling Method) [32]. The method allows to increase the success scaling rate of SFC by applying the IVS (Improve Vertical Scaling) algorithm to the server hosting that VNF and the virtual links connecting to that VNF when the required resource at a time exceeds the resource provided according to SLA. If this allocation fails, the IHS (Improve Horizontal Scaling) technique will be used to generate a new instance of that VNF and deploy it to another server. To lower scaling failure ratios, the IVS algorithm combines vertical scaling with traffic splitting, which supplies bandwidth for the increased bandwidth demand of virtual links by utilizing additional physical links. Experimental results showed that, compared to ElasticNFV [26] and ElasticSFC [31], the HSM method brings about superior success scaling rates with lower resources. However, in this article, the authors assumed

that the substrate system is secure without any attacks or failures of the network's elements. This may not reflect the real-world system. Additionally, the proposed horizontal scaling mechanism is an approach that is close to migration when creating new VNF instances and re-directing the traffic flow to those instances instead of replicating instances and splitting traffic to pass through both the current VNF and its new replicas. The HSM algorithm can be refined by tearing data flow to pass to new replicas of VNF before creating new VNF and rerouting the traffic flow as IHS does.

Cao et al. proposed a dynamic resource allocation mechanism that allows adjusting the operation of SFCs to guarantee the NS provision for end users in case the hardware infrastructure (node or link) fails [33]. This minimizes service interruptions in NFV-enabled vehicular networks. Nevertheless, although the project focuses on the flexible resource allocation problem, the authors did not mention the changes in resource demand for services. Therefore, they ignored vertical scaling and horizontal scaling.

In practice, network topology may change continuously over time because of adding or removing VNF instances. Eliminating changes in the network topology is a good way to reduce costs. Yifu et al. [34] proved that VNF scaling is an NP-hard problem. Then, the authors proposed an online algorithm to assist the VNF horizontal scaling problem, which includes two parts: The first part is a forecasting model based on Fourier series to mitigate frequent updates to the network topology, and the second is an algorithm to place the right VNF instance. The experimental results show that the approach can save 20% of costs while retaining performance parameters.

Rankothge et al. presented a framework based on metaheuristic Iterated Local Search (ILS) for autonomously reallocating resources to three scaling models (vertical scaling, horizontal scaling, and migration) [35]. The results of the experiment show that the proposed framework can return an optimal solution in several milliseconds, whereas Integer Linear Programming (ILP) might take some minutes to converge. The authors also explored how optimization is affected by the different scaling models and the optimization goals, then proved out that adopting only vertical scaling should be avoided and horizontal scaling is a method that trade-offs between CPU resources, system instability, and accepting more scaling requests.

The paper of Houidi et al. [36] focused on solving the VNFFG extension problem during SFC's operation. That is, tenants are likely to add more network functions or new forwarding paths into their services as demands arise and as their consumer base and profiles evolve. To maximize the number of extended requests while maintaining the stability of the original system and to avoid service disruption with a minimum execution time, the authors first addressed the problem through an ILP model, which can bring good performance in-

dexes in reasonable problem sizes, as a baseline to evaluate proposed heuristic algorithms (e.g., a Steiner Tree-based algorithm and an Eigendecomposition based algorithm). The Steiner Tree is proven to be the best solution for the VNFFG extension problem as it archives high successful extension ratios with an acceptable execution time for large scales. The Eigendecomposition algorithm can bring smaller execution time in high connection environments.

To get close to the real world, the project [37] takes into account the concurrent operation of multiple SFCs. In which, a VNF instance can join more than one SFC simultaneously. When migration occurs, SFCs constituted by these VNF instances may be affected. With the objective of reducing end-to-end delays for all affected SFCs while guaranteeing network load balancing after migration happens, Li et al. first formalized the VNF instance migration and SFC reconfiguration problem using a mathematical model. Finally, the authors proposed a multi-stage heuristic algorithm based on optimal order to solve the problem. The heuristic algorithm has three stages: i) determining the order of VNF instances to migrate. In which, the VNF instances that less affect SFCs can be prioritized to migrate; ii) determining the candidate nodes to migrate to; iii) calculating the minimum influence requirement and making decisions on migration. The results show that the proposed algorithm can reduce the average delay of 16% to 25% for various scale networks while maintaining the balance of network load. In the same vein, the study [38] focuses on utilizing multipath routing to distribute network traffic more efficiently, thereby improving network performance and reliability. By implementing multipath routing in NFV, the authors aim to address congestion issues and optimize resource utilization across the network. The proposed solution demonstrates significant improvements in balancing the load across multiple network paths, leading to enhanced overall system performance.

While most studies cannot achieve optimization in both efficiency and scalability, Yu et al. [39] developed a hybrid technique to address vertical and horizontal scaling, with the goal of providing an optimum solution in large-scale systems. By determining use cases for a specific scaling approach, the study pointed out that the priority rules for scaling method selection can be based on the comparison from six aspects, the results are: Vertical scaling can bring more efficient *resource utilization* than horizontal scaling; the *scaling period* of vertical scaling is smaller than horizontal scaling; vertical scaling has faster *response time* than horizontal scaling; for *compatibility*, horizontal scaling has more advantages than vertical scaling because some VNFs cannot improve their performance by granting more resources; horizontal scaling has better *scalability* than vertical scaling because vertical scaling is limited to physical machine capacity; two scaling methods have similar performance in *robustness*. According to

the above comparison, the authors conclude that vertical scaling has a higher priority than horizontal scaling. The experimental results showed that the proposed approach has acceptance ratios and resource utilization better than FreeFlow [17] and ElasticNFV [26].

In large network systems, the paper [40] considered the flexibility of VNF deployment and SFC orchestration based on network conditions. Besides the dynamicity of user requests, VNFs themselves can also modify the traffic amount during their execution. To minimize resource costs while satisfying VNF dependency and traffic volume scaling, Zeng et al. proposed a heuristic approach named TAIVP (Traffic Aware and Interdependent VNF Placement) consisting of three components: i) the SFC construction component is used to construct SFC with the lowest network resource cost while ensuring the constraint of the VNF dependency; ii) the path planning function determines a shortest path from source to destination based on the A-star algorithm; iii) and the SFC embedding function places VNFs on nodes based on the order of VNFs in the SFC and the discovered shortest path. The results reveal that the TAIVP algorithm can reduce network costs by 10.2% and increase the acceptance ratio of service requests by 7.6% on average. However, there are still some limitations to the project. The authors did not consider the delay of the service requests, which is an important factor in NFV. Additionally, the heuristic algorithms cannot provide a solution that is close to the optimal one.

ETSI defined a framework, namely NSD (NS Descriptor) [41], that is integrated inside the NFV MANO block for automatic detection of resource requirement changes. The key concept is that developers will define a discrete set of Instantiation Levels (ILs) for NS (NS-ILs), which NSs can be resized to during their lifecycle. The similarity for VNF-ILs and VL-ILs are found in VNF Descriptor (VNFD) and VL Descriptor (VLD), respectively. This framework can reduce the work for scaling research when they do not need to care about how to detect scaling events but only need to focus on developing solutions to deal with them. Adamuz-Hinojosa et al. analyzed how ILs are designed in NSD [42]. The authors also figured out how the scaling requirement of NSs, VNFs, and VLs can be triggered automatically by using NS-ILs, VNF-ILs, and VL-ILs, respectively.

In QoS enhancement, guaranteeing end-to-end reliability is a crucial factor. NFV-enabled networks are vulnerable due to frequent hardware and software errors. These hazards can come from many reasons, such as server failures, broken links, software errors, cyberattacks, etc. There are a number of projects that pay attention to this issue.

In order to ameliorate system reliability when failures happen, the paper [43] introduced a novel redundancy scheme while considering the VNFFG structure to avoid over backup and the utilization reduction of the underlying resource. The key concept of the solution is to place backup VNFs on high-reliability nodes. From

the simulations, the proposed mechanism can cut down the backup cost by up to 46% and keep high acceptance ratios with respect to the existing algorithms.

Liu et al. in the paper [44] proposed a Mixed Integer Linear Programming (MILP) to address the reliability-aware service chaining mapping problem and an on-line algorithm based on the joint protection redundancy model and backup selection scheme to improve the acceptance ratio of service requests while minimizing the consumption of physical resources. The main concept is to find an efficient mapping strategy for each SFC while maintaining constraints with two main steps for two mapping schemes: The primary scheme is the mapping of VNFs along the shortest path from ingress to egress nodes, the backup scheme is the mapping of redundant VNFs that can be used when any element in the SFC fails. The proposed novel online learning algorithm optimizes the management cost and service reliability while maintaining capacity and reliability constraints with the acceptability of delay.

Additionally, the Q-learning is adjusted to select backup VNFs in the chain. The results show that the proposed approach can significantly enhance the service request acceptance ratio while reducing resource consumption in comparison to two other backup algorithms.

To detect SFC failure in real-time, the paper [45] proposed a mechanism to jointly recover failures, prevent faults, and manage resources efficiently. In the article, the authors attempt to optimize the probability of failure in networking equipment in the case of changes in network topology. The issue is mathematically formulated as an optimization problem called the Optimal Fog-Supported Energy-Aware SFC rerouting algorithm (OFES). The proposed mechanism called Heuristic OFES (HFES) includes a near-optimal heuristic to solve the OFES problem in polynomial time by guaranteeing that the probability of fault is always less than a pre-defined threshold. The simulation results point out that the average failure probability of HFES is up to 40% higher than OFES.

In recent years, Artificial Intelligence (AI) has attracted a lot of attention from the public. Researchers have started to use machine learning algorithms to solve SFC scaling issues. Jing et al. [46] designed a Long Short-Term Memory (LSTM)-based algorithm for predicting user demands. Based on predicted results, the authors proposed a proactive method to deal with the vertical and horizontal scaling problems of VNFs. The project [15] also used an online machine learning algorithm to predict upcoming user traffic, then proactively assign a new instance of VNF and reroute the data flow with fewer resources. The research of Namjin et al. improved the Graph Neural Network (GNN) architecture and utilized a few techniques from other domains, such as image processing and natural language processing, to efficiently obtain a node representation of networking information for the VNF placement problem [47]. Therefore, the proposed method can be more effective in solving the VNF deployment problem for the scaling-in.

Table 1. Approaches and methods of existing works

Research	Approach	Scaling model	Migration	Failure
[15]	To use an online learning to proactively predict upcoming traffic demands. Then efficiently create new instances of VNFs and provide optimal route for service chain.	None	✓	✗
[17]	Splitting data flow to perform load sharing between VNF instances.	Horizontal scaling	✗	✗
[26]	To use existing Kernel-based Virtualization Machine (KVM) techniques to perform dynamic resource allocation and a TPMM algorithm for optimizing migration cost.	Vertical scaling	✓	✗
[30]	Turning off VNFs that do not use up resources and deploy VNFs that need to be scaled out on another server.	None	✓	✗
[31]	To release/grant more computing resources for VNFs, bandwidth resources for virtual links.	Both	✓	✗
[32]	Considering vertical scaling and horizontal scaling to achieve a higher success scaling rate. Split the data stream to share the load among VNF instances.	Vertical scaling	✓	✗
[33]	Reallocating the deployment location of the element (VNF or virtual link) on the faulty device.	None	✗	✓
[34]	Forecasting service request changes based on the Fourier Series to reduce the frequency of network topology changes.	Horizontal scaling	✗	✗
[35]	To use a framework based on metaheuristic Iterated Local Search (ILS) to automatically reallocate resources to three scaling models.	Both	✓	✗
[36]	Optimizing the number of extended requests with an acceptable execution time when there are changes in constituents of SFC by ILP model and two heuristic algorithms.	None	✗	✗
[37]	To use a multi-stage heuristic algorithm based on optimal order to handle migration problem with participating of a VNF in multiple SFCs simultaneously.	None	✓	✗
[39]	To determine the priority of scaling method in deadling with scaling events to have optimal solution in large scale networks.	Both	✓	✗
[40]	To use a heuristic approach to construct SFCs and place VNFs with considering the VNF can change volume of traffic flow itself.	None	✗	✗
[43]	Placing backup VNFs on high-reliability nodes.	None	✗	✓
[44]	Determining two VNFs mapping schemes for normal operation and for failure use cases.	None	✗	✓
[45]	To ensure that the fault probability is always less than a threshold.	None	✗	✓
[46]	Proposing an algorithm base on LSTM to predict user demands. Then solve VNFs vertical and horizontal scaling problem base on predicted results.	Both	✗	✗
[47]	To adapt the GNN architecture and use a few techniques to obtain a better node representation for the VNF deployment task. Therefore, proposed approach can help to solve scaling-in and out of VNFs.	Horizontal scaling	✗	✗
[48]	Defining a MILP model for the problem of resilient SFC to be able to recover from a failure.	None	✗	✓

3. DISCUSSION

3.1. SUMMARY

To provide an overview picture of dealing with the SFC scaling problem, in this section, we briefly summarize the existing studies based on the following criteria: Approach and scaling model. From the reviews in sub-section 2.3, we realize that most projects tend to ignore migration scaling, while this method has its own advantages. Therefore, we involve the problem of migrating SFC elements in this comparison for an adequate view. That is, we will examine whether the study considers the migration phase of the VNFs or not. We are also interested in failure situations. Did the research take into account the possibility of system failure? Because reliability plays a vital role in satisfying QoS, especially in the current context, where network compromises cannot be ignored [51].

In Table 1, we are involved in many articles coming from various purposes, although we want to focus on the scaling problem. This is because while reviewing current work, we realize that, aside from solving the SFC scaling issue, there are a number of studies that involve aspects that are close to the scaling. For example, the paper [36] considers the addition of VNFs into SFC instead of granting more resources or changing the

embedding of VNFs. To have a deep view of guaranteeing the reliability of the NFV-enabled system, we examined the roles of the failure factor in the SFC operation, then we found that most studies treated failure-related factors at the initialization of SFC [44], [45]. The authors tried to enhance reliability at the VNF placement stage and ignored failures during the execution of SFCs. That means failures are underestimated in the SFC scaling problem. We can also see that, in terms of scaling strategies, the majority of the listed research only tackles the problem using a single scaling model.

3.2. EMERGING RESEARCH CHALLENGES

Thanks to the interest of researchers, in recent years, the issue of flexible resource reallocation for the SFC has achieved significant improvements. According to our investigation, various features of dynamic resource reallocation in NFV may need to be further exploited. This section covers potential future research directions that need to be explored in the development of NFV.

Taking advantage of the elasticity of the Cloud

Since virtualization technology, servers can be deployed as containers (e.g., Docker). In this case, applications can be deployed and destroyed in milliseconds [49].

In horizontal scaling or migration scaling, the controller needs to deploy new VNFs on other servers in the system to satisfy scaling demands. This action is close to the VNF placement problem [30], [47], while most of the existing works are trying to place VNFs on running servers, then chaining these VNFs to initialize SFC without considering the flexibility of the cloud environment. Because the time to deploy and destroy VNFs on servers can be very short, when there is an SFC that needs to be deployed or scaled, the controller will select the appropriate nodes, and then deploy the VNFs to those hosts instead of selecting existing VNFs on nodes. Simultaneously, every node that does not participate in any SFC will be turned into idle mode to save energy and maximize the remaining volume of resources. Nevertheless, in the case of multiple types of servers collaborating with each other in the NFV system, this approach has drawbacks. In this case, the physical servers may take time to boot up and initialize the VNF instances. Consequently, the total convergence time of the SFC construction process may become longer and may even create bottlenecks in the system.

Considering all scaling models simultaneously

As demonstrated in Table 1, previous studies have evaluated one or more models when scaling events occur. Nonetheless, the authors tend to deal with the problem by a single scaling method. For example, articles [17], [34], [47] only leverage horizontal scaling to deal with the increase in traffic volume. This is a coarse-grained approach that may lead to a decrease in resource utilization and successful scaling ratio. In the paper [39], Yu et al. examined three scaling models and concluded that to deal with scaling events, the priority of vertical scaling is highest, followed by migration and horizontal scaling. Therefore, involving various scaling models can bring about more fine-grained solutions and could make important contributions to resource optimization for NFV operations. However, incorporating many constraints and input factors into a single problem may cause an increase in the complexity of the algorithm, while the nature of the optimization problem is trading off objectives and dependencies such as maximizing system performances while minimizing resource consumption.

Using machine learning to solve SFC scaling

As mentioned in Section 2.3, in recent years, many researchers have paid attention to using machine learning algorithms to solve the SFC scaling problem. AI has gained many achievements in image processing, text processing, etc. In NFV, it has been adapted to solve the VNF placement problem [50], VNF forwarding graph embedding problem [22]. Deep learning and reinforcement learning algorithms also joined to handle scaling problems in SFC [15], [47], [46], [48] and got highlight results. Therefore, adjusting more advanced machine learning algorithms will be a future approach. However, the accuracy of machine learning models is determined by various factors, including the quality of the training data, the effectiveness of the algorithm uti-

lized, and the dataset size. Whereas data flow and on-line behaviors are complex, diverse and unpredictable. As a result, these approaches may have a certain ratio of wrong decisions such as triggering scaling events at the inappropriate time.

Flexible resource allocation in the event of failures

As figured out in Table 1, most scaling-related studies did not take into account system failures, including software errors and infrastructure crashes, while the availability of NSs must be continuously guaranteed. Therefore, when errors occur, the system needs a mechanism to react to these failures. According to reviews in sub-section 2.3, there are a number of projects involving system failures in order to maintain reliability. However, most authors mitigate faults by implementing strategies at the SFC initialization. That means they ignore fault-related factors during SFC execution. Note that these failures can come from errors at many levels (e.g., in physical servers, in virtual servers, VNF errors, etc.) and for many reasons. Therefore, besides the approach of mitigating failures at the initializing SFC stage, considering failures when measuring SFC scaling problems can improve system reliability.

4. CONCLUSIONS

Network Function Virtualization is a promising field. NFV research has grown exponentially in recent years. In which, the problem of optimizing resources for NFV operation is the focus of attention. In this paper, we provide a picture of a narrow field in resource optimization. We cover the basics of NFV and scaling in the operation of SFCs. We also present a taxonomy of recent studies in the field of solving SFC scaling during its operation. Comparisons of existing works show that there were several inadequate aspects to consider in researching the scaling of the SFC. Finally, we offer some bright directions for the future to deal with resource reallocation in the operation of network services to adapt to the dynamic demands of users.

5. ACKNOWLEDGEMENT

This project is partially supported by the European Union's Horizon 2020 Research and Innovation Programme RISE under grant agreement no. 823759 (REMESH).

6. REFERENCES

- [1] Cisco, "Cisco Annual Internet Report (2018–2023) White Paper", <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html> (accessed: 2023)
- [2] ETSI, "Network functions virtualisation: An introduction, benefits, enablers, challenges and call

- for action”, SDN and OpenFlow World Congress, Darmstadt-Germany, 2012.
- [3] J. Halpern, C. Pignataro, “Service function chaining (SFC) architecture”, Technical Report RFC, 2015.
- [4] Y. Bo, W. Xingwei, L. Keqin, D. K. Sajal, H. Min, “A comprehensive survey of Network Function Virtualization”, *Computer Networks*, Vol. 133, 2018, pp. 212-262.
- [5] J. G. Herrera, J. F. Botero, “Resource Allocation in NFV: A Comprehensive Survey”, *IEEE Transactions on Network and Service Management*, Vol. 13, No. 3, 2016, pp. 518-532.
- [6] W. Yang, C. Fung, “A survey on security in network functions virtualization”, *Proceedings of the IEEE NetSoft Conference and Workshops*, Seoul, Korea, 6-10 June 2016, pp. 15-19.
- [7] X. Yanghao, L. Zhixiang, W. Sheng, W. Yuxiu, “Service Function Chaining Resource Allocation: A Survey”, arXiv:1608.00095, 2016.
- [8] W. Attaoui, E. Sabir, H. Elbiaze, M. Guizani, “VNF and CNF Placement in 5G: Recent Advances and Future Trends”, *IEEE Transactions on Network and Service Management*, Vol. 20, No. 4, 2023, pp. 4698-4733.
- [9] A. Laghrissi, T. Taleb. “A Survey on the Placement of Virtual Resources and Virtual Network Functions”, *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 2, 2019, pp. 1409-1434.
- [10] J. Sun, Y. Zhang, F. Liu, H. Wang, X. Xu, Y. Li, “A survey on the placement of virtual network functions”, *Journal of Network and Computer Applications*, Vol. 202, 2022.
- [11] X. Li, C. Qian, “A survey of network function placement”, *Proceedings of the 13th IEEE Annual Consumer Communications & Networking Conference*, Las Vegas, NV, USA, 9-12 January 2016, pp. 948-953.
- [12] ETSI GS NFV-IFA 006, “Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Vi-Vnfm Reference Point — Interface and Information Model Specification”, 2018.
- [13] ETSI GS NFV-IFA 001, “Network Functions Virtualisation (NFV); Infrastructure Overview”, 2015.
- [14] ETSI GS NFV-IFA 002, “Network Functions Virtualisation (NFV); Architectural Framework”, 2014.
- [15] X. Fei, F. Liu, H. Xu, H. Jin, “Adaptive VNF Scaling and Flow Routing with Proactive Demand Prediction”, *Proceedings of IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, USA, 16-19 April 2018, pp. 486-494.
- [16] P. Shoumik, L. Chang, H. Sangjin, J. Keon, P. Aurojit, R. Sylvia, R. Luigi, S. Scott, “E2: A framework for nfv applications”, *Proceedings of the 25th Symposium on Operating Systems Principles*, Monterey, CA, USA, 4-7 October 2015, pp. 121-136.
- [17] R. Shriram, W. Dan, J. Hani, W. Andrew, “Split/merge: System support for elastic execution in virtual middleboxes”, *Proceedings of the 10th USENIX conference on Networked Systems Design and Implementation*, Lombard, IL, USA, 2-5 April 2013, pp. 227-240.
- [18] M. T. Beck, J. F. Botero, K. Samelin, “Resilient allocation of service Function chains”, *Proceedings of the IEEE Conference on Network Function Virtualization and Software Defined Networks*, Palo Alto, CA, USA, 7-10 November 2016, pp. 128-133.
- [19] S. Mehraghdam, M. Keller, H. Karl, “Specifying and placing chains of virtual network functions”, *Proceedings of the IEEE 3rd International Conference on Cloud Networking*, Luxembourg, Luxembourg, 8-10 October 2014, pp. 7-13.
- [20] M. Wang, B. Cheng, B. Li, J. Chen, “Service Function Chain Composition and Mapping in NFV-Enabled Networks”, *Proceedings of the IEEE World Congress on Services*, Milan, Italy, 8-13 July 2019, pp. 331-334.
- [21] M. T. Beck, J. F. Botero, “Coordinated Allocation of Service Function Chains”, *Proceedings of the IEEE Global Communications Conference*, San Diego, CA, USA, 6-10 December 2015, pp. 1-6.
- [22] P. T. A. Quang, Y. Hadjadj-Aoul, A. Outtagarts, “A Deep Reinforcement Learning Approach for VNF Forwarding Graph Embedding”, *IEEE Transactions on Network and Service Management*, Vol. 16, No. 4, pp. 1318-1331.
- [23] M. Karimzadeh-Farshbafan, V. Shah-Mansouri, D. Niyato, “A Dynamic Reliability-Aware Service

- Placement for Network Function Virtualization (NFV)", *IEEE Journal on Selected Areas in Communications*, Vol. 38, No. 2, 2020, pp. 318-333.
- [24] M. Mechtri, C. Ghribi, D. Zeglache, "A Scalable Algorithm for the Placement of Service Function Chains", *IEEE Transactions on Network and Service Management*, Vol. 13, No. 3, 2016, pp. 533-546.
- [25] Y. T. Woldeyohannes, A. Mohammadkhan, K. K. Ramakrishnan, Y. Jiang, "A scalable resource allocation scheme for NFV: Balancing utilization and path stretch", *Proceedings of the 21st Conference on Innovation in Clouds, Internet and Networks and Workshops*, Paris, France, 19-22 February 2018, pp. 1-8.
- [26] H. Yu, J. Yang, C. Fung, "Fine-Grained Cloud Resource Provisioning for Virtual Network Function", *IEEE Transactions on Network and Service Management*, Vol. 17, No. 3, 2020, pp. 1363-1376.
- [27] W. Timothy, S. Prashant, V. Arun, Y. Mazin, "Black-box and gray-box strategies for virtual machine migration", *Proceedings of the 4th USENIX Symposium on Networked Systems Design & Implementation*, Cambridge, MA, USA, 11-13 April 2007.
- [28] B. Hitesh, C. Paolo, K. Thomas, R. Ant, "Towards predictable datacenter networks", *ACM SIGCOMM*, Vol. 41, No. 4, 2011.
- [29] C. Lianjie, S. Puneet, F. Sonia, S. Vinay, "ENVI: Elastic resource flexing for Network function Virtualization", *HotCloud'17: Proceedings of the 9th USENIX Conference on Hot Topics in Cloud Computing*, Santa Clara, CA, USA, 10-11 July 2017.
- [30] X. Zhao, X. Jia, Y. Hua, "An Efficient VNF Deployment Algorithm for SFC Scaling-out Based on the Proposed Scaling Management Mechanism", *Proceedings of the Information Communication Technologies Conference*, Nanjing, China, 29-31 May 2020, pp. 166-170.
- [31] T. N. Jadel, S. Jungmin, C. Qinghua, B. Rajkumar, "ElasticSFC: Auto-Scaling Techniques for Elastic Service Function Chaining in Network Functions Virtualization-based Clouds", *Journal of Systems and Software*, Vol. 152, 2019, pp. 108-119.
- [32] D. Zhai, X. Meng, Z. Yu, H. Hu, X. Han, "A fine-grained and dynamic scaling method for service function chains", *Knowledge-Based Systems*, Vol. 228, 2021.
- [33] H. Cao, H. Zhao, D. X. Luo, N. Kumar, L. Yang, "Dynamic Virtual Resource Allocation Mechanism for Survivable Services in Emerging NFV-Enabled Vehicular Networks", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 11, 2022, pp. 22492-22504.
- [34] Y. Yifu, G. Songtao, L. Pan, L. Guiyan, Z. Yue, "Forecasting Assisted VNF Scaling in NFV-Enabled Networks", *Computer Networks*, Vol. 168, 2019.
- [35] W. Rankothge, H. Ramalhinho, J. Lobo, "On the Scaling of Virtualized Network Functions", *Proceedings of the IFIP/IEEE Symposium on Integrated Network and Service Management*, Arlington, VA, USA, 8-12 April 2019, pp. 125-133.
- [36] O. Houdi, O. Soualah, W. Louati, D. Zeglache, "Dynamic VNF Forwarding Graph Extension Algorithms", *IEEE Transactions on Network and Service Management*, Vol. 17, No. 3, 2020, pp. 1389-1402.
- [37] B. Li, B. Cheng, J. Chen, "A Multi-Stage Approach for Virtual Network Function Migration and Service Function Chain Reconfiguration in NFV-enabled Networks", *Proceedings of the IEEE International Conference on Web Services*, Beijing, China, 19-23 October 2020, pp. 207-215.
- [38] T.-M. Pham, L. M. Pham, "Load balancing using multipath routing in network functions virtualization", *Proceedings of the IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future*, Hanoi, Vietnam, 7-9 November 2016, pp. 85-90.
- [39] H. Yu, J. Yang, C. Fung, R. Boutaba, Y. Zhuang, "ENSC: Multi-Resource Hybrid Scaling for Elastic Network Service Chain in Clouds", *Proceedings of the IEEE 24th International Conference on Parallel and Distributed Systems*, Singapore, 11-13 December 2018, pp. 34-41.
- [40] Z. Zeng, Z. Xia, X. Zhang, Y. He, "SFC Design and VNF Placement Based on Traffic Volume Scaling and VNF Dependency in 5G Networks", *Computer Modeling in Engineering & Sciences*, Vol. 134, No. 3, 2023, pp. 1791-1814.

- [41] ETSI GS NFV-IFA 014, "Network Functions Virtualisation (NFV) Release 4; Management and Orchestration; Network Service Templates Specification", 2019.
- [42] O. Adamuz-Hinojosa, J. Ordonez-Lucena, P. Ameigeiras, J. J. Ramos-Munoz, D. Lopez, J. Folgueira, "Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow", *IEEE Communications Magazine*, Vol. 56, No. 7, 2018, pp. 162-169.
- [43] W. Ding, H. Yu, S. Luo, "Enhancing the reliability of services in NFV with the cost-efficient redundancy scheme", *Proceedings of the IEEE International Conference on Communications*, Paris, France, 21-25 May 2017, pp. 1-6
- [44] Y. Liu, Y. Lu, W. Qiao, X. Chen, "Reliability-aware service chaining mapping in NFV-enabled networks", *ETRI Journal*, Vol. 41, No. 2, 2019, pp. 207-223.
- [45] M. M. Tajiki, M. Shojafar, B. Akbari, S. Salsano, M. Conti, M. Singhal, "Joint failure recovery, fault prevention, and energy-efficient resource management for real-time SFC in fog-supported SDN", *Computer Networks*, Vol. 162, 2019, p. 106850.
- [46] T. Jing, L. Z. Jia, C. Yan, W. J. Wei, Y. Peng, L. C. Hao, "Adaptive VNF Scaling Approach with Proactive Traffic Prediction in NFV-enabled Clouds", *ACM TURC '21: Proceedings of the ACM Turing Award Celebration Conference*, Hefei, China, 30 July - 1 August 2021, pp. 166-172.
- [47] S. Namjin, H. D. Nyeong, C. Heeyoul, "Advanced Scaling Methods for VNF deployment with Reinforcement Learning", arXiv:2301.08325, 2023.
- [48] P. Tuan-Minh, N. Thi-Minh, N. Xuan-Tuan-Trung, C. Hoai-Nam, S. H. Ngo, "Fast Resource Allocation for Resilient Service Coordination in an NFV-Enabled Internet-of-Things System", *REV Journal on Electronics and Communications*, Vol. 12, 2022.
- [49] M. P. Amit, G. D. Narayan, K. Shivaraj, M. M. Mohammed, "Performance Evaluation of Docker Container and Virtual Machine", *Procedia Computer Science*, Vol. 171, 2020, pp. 1419-1428.
- [50] O. Houidi, O. Soualah, W. Louati, D. Zeghlache, "An Enhanced Reinforcement Learning Approach for Dynamic Placement of Virtual Network Functions", *Proceedings of the IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, London, UK, 31 August - 3 September 2020, pp. 1-7.
- [51] T. -T. -L. Nguyen, T. -M. Pham, L. M. Pham, "Efficient Redundancy Allocation for Reliable Service Function Chains in Edge Computing", *Journal of Network and Systems Management*, Vol. 31, No. 1, 2023.

EMF Exposure Reduction Using Weighted Angle Model for Multi-Technology Sectorized BS

Original Scientific Paper

Mohammed S. Elbasheir*

School of Electronic Engineering, College of Engineering, Sudan University of Science and Technology, Khartoum 14413, Sudan
mohd.suleiman@gmail.com

Rashid A. Saeed

Department of Computer Engineering, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
abdulhaleem@tu.edu.sa

Salaheldin Edam

School of Electronic Engineering, College of Engineering, Sudan University of Science and Technology, Khartoum 14413, Sudan
salah_edam@hotmail.com

*Corresponding author

Abstract – Mobile networks are now growing quickly due to major advancements in wireless technology especially with the introduction of the Fifth Generation New Radio (5G NR). A greater risk of exposure to electromagnetic field radiation (EMF) is being raised by the widespread deployment of base stations (BSs). Standard guidelines are set to control the amount of EMF radiation. This paper proposed a design model to de-concentrate and reduce the total exposure of the multi-technology BS with no drawbacks on network coverage level and key performance indicators (KPIs). The proposed solution applies the concept of weighted antenna's azimuth to spread the total exposure by separating the antennas in the same sector. A set of simulations is carried out to calculate the reduction in total exposure ratio (TER) and the Compliance Distance (CD). Also, A field measurement test was done in a life network to validate and evaluate the proposed model under real conditions. Furthermore, the network operation support system (OSS) records were analyzed to evaluate the impact on the network coverage and capacity behavior. The pre- and post-results demonstrate that using the proposed model enhanced the CD and TER., the results show using two azimuths reduces the CD by 23% and by 43.4% when using six antennas. Also, the field test result demonstrated a 19.23% reduction in the Total Exposure Ratio. Overall, the system records show no significant impacts were registered on network coverage level and capacity performance for the sites involved in the test.

Keywords: mobile wireless network, antenna sectorization, mobile coverage planning, EMF compliance distance

Received: May 31, 2024; Received in revised form: August 26, 2024; Accepted: August 27, 2024

1. INTRODUCTION

Rapid development and extensive installation of mobile wireless networks increase the number of BSs that transmit data. enormous numbers of site BSs have been connected to provide the network with enough coverage and boost resource capacity. This has led to extremely enormous data transfers via mobile wireless networks, which will expand significantly by 2030 [1]. 2G Global System for Mobile Communication (GSM), 3G Universal Mobile Telecommunications System (UMTS), 4G Long-Term Evolution (LTE), and most recently, 5G NR, are the most widely used technologies in mobile wireless networks. Concerns regarding the negative ef-

fects on human health are addressed due to the rise in electromagnetic fields as a result of the development of new technologies [2].

Predictions state that 5G technology will evolve into an all-purpose system [3] because it provides high capacity and enables rich features and services that expand the number of business prospects and strengthen the global economy. For 5G development, NR BSs must be installed in higher frequency bands [4], mostly by collocating them with current 2G, 3G, and 4G technologies.

A group of technologies and solutions in one place needs an examination of the overall accumulated ra-

diation and the exposure level in relation to standard limitations. Many studies indicate this evaluation even should take place during the design stage of the network before deployment as EMF is a constraint in network planning especially for 5G [5, 6].

Nearly all cellular network operators use sectorized base stations with directional antennas. The 3-sector model is particularly useful for optimizing load balancing, capacity resource management, inside and outdoor coverage, and interference reduction. Furthermore, various technologies spreading in the same directions as co-located are carried by the same sectors. Depending on the operator's design strategy [7], the group of technologies either be connected to one multi-band antenna, or might be installed into separate antennas, but all antennas of one sector are directed in the same azimuth.

The investigation of EMF for multiple technologies is a crucial topic that aims to enhance the assessment of the compliance distance and to introduce models to minimize it, especially since it requires the operators to identify the compliance distances and mark them as exclusive zones and should be not accessible for the general public.

The contribution of this study lies in providing a simple design solution to reduce the total exposure emitting from sectorized antennas of the multi-technology BS by applying horizontal separation angles between the antennas in the sector to de-concentrate the total accumulated exposure in the same direction, while maintaining the network performance with almost no impact on the coverage signal levels and performance. This model can be practically applied to the widely commonly used antennas that are currently installed for multi-technology BSs.

This manuscript is structured into nine sections including the introduction in Section 1. In Section 2, the total exposure ratio related to standard limits is briefed. The literature review and related work are discussed in Section 3. In Section 4, the proposed model is explained in detail. In Section 5, the simulation setup and results are presented and discussed. In Section 6, the in-situ assessment is explained and the results are discussed. Section 8 gives recommendations for the potential application of the proposed model. At last, Section 9 summarizes the paper's conclusion.

2. TOTAL EXPOSURE IN STANDARDS

The Federal Communication Commission (FCC), which establishes regulatory criteria in the USA [8], and the International Commission on Non-Ionizing Radiation Protection (ICNIRP) which is established in Europe [9] are two well-known organizations that have established and published standard recommendations. Governmental and national authorities in many nations have utilized these guidelines to manage the installation of EMF transmitters and the activities associated with them

[10, 11]. The FCC and ICNIRP standards make a distinction between the technical occupational workers (OW) and the general public (GP). The OW refers to the staff members who are well-trained to be aware of potential EMF hazards and are exposed to certain related scenarios, and the GP are characterized as being normal people who are exposed to electromagnetic fields and are not aware of the dangers associated with them.

The whole-body radiation reference levels have been set by ICNIRP for both the occupational workers and the normal general public under the transmitting frequency, as listed in Table 1.

Table 1. ICNIRP Reference Limits for OW and GP

Exposure Boundary	Frequency Range	E - field (V/m)	H - field (A/m)	PD (W/m ²)
OW	0.1 - 30 MHz	$660/f_M^{0.7}$	$4.9/f_M$	NA
	>30 - 400 MHz	61.0	0.16	10.00
	>400 - 2,000 MHz	$3 f_M^{0.5}$	$0.008 f_M^{0.5}$	$f_M/40$
	>2.0 - 300 GHz	N/A	N/A	50.00
GP	0.1 - 30 MHz	$300/f_M^{0.7}$	$2.2/f_M$	NA
	>30 - 400 MHz	27.70	0.073	2.00
	>400 - 2,000 MHz	$1.375 f_M^{0.5}$	$0.0037 f_M^{0.5}$	$f_M/200$
	>2.0 - 300 GHz	N/A	N/A	10.00

Also, the FCC standard has defined the maximum limits of exposure as the maximum permitted exposure (MPE) levels for the GP and OW according to the transmitting frequency band as listed in Table 2.

Table 2. The FCC exposure limits for 0.3 MHz to 100 GHz, for OW and GP

Exposure Boundary	Frequency Range	E - field (V/m)	H - field (A/m)	PD (W/m ²)
OW	0.3-3.0 MHz	614.0	1.630	100.0
	3.0-30 MHz	$1842/f$	$4.89/f$	$900/f^2$
	30-300 MHz	61.40	0.163	1.00
	0.3-1.5 GHz	-	-	$f/300$
	1.5-100 GHz	-	-	5.00
GP	0.3-1.34 MHz	614.0	1.630	100
	1.34-30 MHz	$824/f$	$2.19/f$	$180/f^2$
	30-300 MHz	27.50	0.0730	0.20
	0.3-1.5 GHz	-	-	$f/1500$
	1.5-100 GHz	-	-	1.00

3. LITERATURE REVIEW

Owing to the significance of this subject, a great deal of research and studies have been done, and more is still being done. Aiming to examine the EMF exposure inquiry and evaluation, the recently published results in worldwide organizations concerning this topic are examined and presented from a number of viewpoints and aspects. This section evaluated a few examples that were recently published, and explored their work methods, conclusions, and outcomes. The following summaries are for those works that focused on determining the total exposure and compliance boundaries:

- The authors of [12] suggested conservative formulae to calculate the whole-body and localized SAR for the main beam exposure from the BS. The heuristic nature of the proposed formulas, their applicability to a class of typical base station antennas, their creation from multiple physical observations, and the results of a comprehensive literature review, measurements, and numerical simulations of typical exposure scenarios all lend support to their creation.
- The compliance distance for 2G GSM operating at 1800 MHz was calculated by the authors in [13] based on field measurements they conducted in various locations within the university (Symbiosis International University campus Pune, India). The calculated compliance distance is 8.4 meters.
- A novel technique for measuring 5G NR exposure based on user actions, including the evaluation of auto-included exposure of base stations and user phones, was suggested by the authors in [14]. Their study is based on information from earlier RF-EMF exposure research as well as certain studies that simulate NR base stations and readings close to test sites.
- In [15], the authors measured the 4G LTE TDD mMIMO in situ while accounting for 100% traffic load and maximum system utilization. The findings indicate that the EMF level was between 7.3 and 16.1% of the ICNIRP occupational reference level, as opposed to 79.3% based on traditional conservative calculations, and that the actual compliance boundaries were reduced by 2.2–3.3 times the conservatively calculated boundaries. The authors explain this drop by pointing to the irregular and unique conduct of mMIMO beamforming. They also point out that a further fall in the compliance barrier is anticipated because actual RBS traffic loading is typically substantially lower than 100%. Pinchera et al. analyzed the power levels surrounding the 5G antenna array in [16] to determine the compliance boundary and appropriately evaluate the exposure level. Using a statistical method, the authors presented a measure called normalized average power pattern (NAPP) for determining the average power density surrounding the antenna. Their findings illustrate the compliance distances that were computed using various power reduction factor values.
- In [17], Thielens et al. (at Ghent University) used finite-difference time-domain (FDTD) simulations for a 4G LTE base station antenna at 2,600 MHz to establish the EMF exposure compliance bounds. Their findings demonstrate that when the antenna is only partially radiating, the reference levels are not conservative for the different fundamental limitations and reference levels. Furthermore, their findings demonstrate that the compliance boundaries for fundamental restrictions at lower antenna powers are provided by the 10g averaged SAR in the head and trunk of the body.

- In [18], Heliot et al. used a commercial 5G BS operating at 3.6 GHz and a mMIMO customizable test-bed operating at 2.6 GHz to examine the nature of mMIMO exposure and its effects on the compliance boundary. Their statistical exposure-based exclusive zone definition results are intriguing. According to their investigation, there are considerable changes in exposure depending on the direction of the beams. Additionally, assuming a fixed traffic load, the variance of exposure tends to decrease as the number of users increases. Conversely, regardless of the user numbers, the exposure rises sub-linearly with traffic load.

4. TER DE-CONCENTRATION FOR THE BS

More effective coverage planning is possible when the cell site coverage is divided into sectors, which means dividing the coverage area into smaller sectors served by individual antennas. RF engineers adjust the signal propagation to reflect the geographic distribution of mobile users by concentrating the coverage in particular directions as shown in Fig. 1. (A) which represents (as an example) the three-sector model with 0/120/240 degrees as antennas' azimuth for the horizontal directions. Within the same cell site, the available frequency spectrum can be utilized numerous times with sectors remaining unaffected. This expands the cell site's total capacity and enables the simultaneous service of additional customers. Also, improved load distribution throughout the cell site is made possible by antenna sectorization. Traffic can be dynamically forwarded across sectors during high usage periods to reduce congestion and guarantee that all users receive sufficient service quality. The RF engineers may simplify their planning and deployment process according to the area's nature, and they focus on improving every region separately, accounting for variables like topography, population density, and anticipated traffic patterns. Generally, sectorization is a commonly used technique in the construction and optimization of mobile cellular networks because it offers a balance between coverage, capacity, and interference control overall [19].

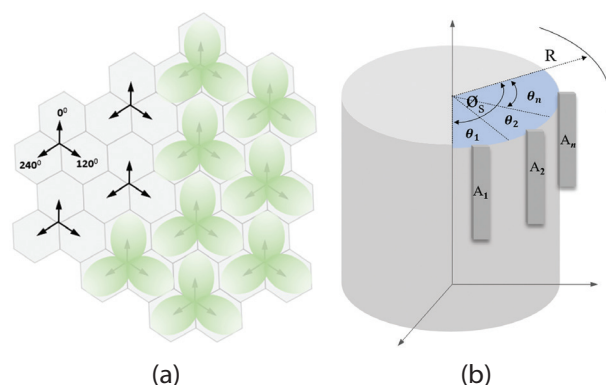


Fig. 1. (a) The three-sector model with 0/120/240 degrees is commonly used in mobile networks. (b) One sector in sectorized cells with i number of antennas.

While sectorization can enhance network performance and capacity, it increases the EMF radiation by concentrating the radiated signals from all technologies in the antenna's directions. Upon the guiltiness of ICNIRP and FCC, the total exposure should be calculated considering the accumulated power density from the transmitting sources.

Fig. 1. (b) shows a sectorized cell with n number of antennas in one sector. From [8, 9] the exposure ratio for each transmitter can be calculated using Equation (1).

$$ER_f = \frac{S_{inc,f}}{S_{inc,RL,f}} \quad (1)$$

Where ER_f is the Exposure Ratio at distance R from the antenna transmitting at frequency f . $S_{inc,f}$ and $S_{inc,RL,f}$ are the incident local power densities and their reference level at frequency f listed in Tables 1 and 2.

The $S_{inc,f}$ can be calculated using Equation (2)

$$S_{inc,f} = \frac{P_{T,f} \cdot G_{A,f}}{4 \cdot \pi \cdot R^2} \quad (2)$$

Where P_T is the transmitted power in watts, and G_A is the antenna's gain of the transmitter at frequency f .

By substituting Equation (2) in the nominator of Equation (1) it gives Equation (3) as:

$$ER_f = \frac{P_{T,f} \cdot G_{A,f}}{4 \cdot \pi \cdot R^2 \cdot S_{inc,RL,f}} \quad (3)$$

Reference to [8, 9] The total exposure ratio of n number of transmitters in one antenna TER_n can be calculated using Equation (4) as:

$$TER_n = \sum_{n,f} ER_f \quad (4)$$

Furthermore, the total exposure ratio of N sectors in one site TER_N can be calculated using Equation (5) as:

$$TER_N = \sum_{n=1}^N TER_n \quad (5)$$

Thus, by substituting Equation (4) in Equation (5), the TER_N can be calculated using Equation (6) as:

$$TER_N = \sum_{n=1}^N \sum_{n,f} ER_f \quad (6)$$

Although the three-sector model is commonly used for cell sectorization, this study considered the general case of N_s number of sectors of one BS in the proposed model. Assuming each sector contributes equally in the TER of one BS, thus, the total exposure of one sector comes through angle width ϕ_s which can be calculated simply using Equation (7):

$$\phi_s = \frac{360}{N_s} \quad (7)$$

Within one sector, the total exposure TER_N can be distributed among the whole ϕ_s by azimuth shifting the direction of each antenna towered separate sub-angle, each sub-angle has an angle width θ_n where

n is the number of antennas in one sector. Each θ_n to have an angle width based on the TER_n weight out of the TER_N , which means the θ_n is the weighted angle proportional to TER_n . Of course, this manner is to be applied for other sectors in the same site to have repeated antenna's azimuth arrangement in the way for one technology to have the same antenna angle separation between all sectors. Thus, this approach gives a model to deconcentrate the total exposure (as spreading) in sub-directions rather than having all antennas transmitting toward one direction. So, the θ_n can be calculated using Equation (8).

$$\theta_n = \phi_s * \frac{TER_n}{TER} = \frac{360}{N_s} * \frac{TER_n}{TER_N} \quad (8)$$

Finally, by substituting Equations (3, 4, 6) into Equation (8), it gives Equation (9) to calculate the θ_n

$$\theta_n = \frac{360}{N_s} * \sum_{n,f} \frac{P_{T,f} \cdot G_{A,f}}{S_{inc,RL,f}} / \left(\sum_{n=1}^N \sum_{n,f} \frac{P_{T,f} \cdot G_{A,f}}{S_{inc,RL,f}} \right) \quad (9)$$

In the assessment section (section 7), will discuss in detail the results that show this model doesn't affect the base station performance in terms of coverage level and capacity. However, it is important to mention this model is applicable for macro sites that serve in areas where continuous coverage is required around the whole site area, and it's not applicable for below such cases:

- For sites that have sector's azimuths intentionally are directed toward certain locations for special coverage and capacity requirements such as highway road sites.
- For sites that use one antenna for all technologies such as penta-band and hexa-band antennas.

5. CD FOR SECTORIZED BASE STATION

IEC62232 has stated in their guidelines the most precise compliance border possible as an iso surface pattern that may be contained in a simpler shaped volume to create more restricting parameters, such as the box-shape (horizontal, vertical, and side) that is appropriate for the sectorized site with the vertical and horizontal boundaries of the coverage antenna. The box-shape compliance range is taken into consideration in this work to assess the exposure in the two primary directions facing the horizontal and vertical beams of the antenna. Similar to related studies found in the literature [12, 13], as shown in Fig. 2 the R_{CD} is used as the distance from the transmitter at which the entire TER equals one as per ICNIRP and FCC guidelines.

The R_{CD} , where $TER=1$, can be calculated using Equation (10) as:

$$R_{CD} = \left(\frac{1}{4\pi} \sum_{n=1}^N \sum_{f > 30 \text{ MHz}}^{300 \text{ GHz}} \frac{P_{T,f} \cdot G_{A,f}}{S_{inc,RL,f}} \right)^{1/2} \quad (10)$$

As 5G uses highly massive Multi Input Multi Output (mMIMO) systems that reduce interference and boost

the cell capacity, more factors and variables were taken into account in several recent investigations of EMF exposure [12, 13, 20-22], such as the system load, actually emitted power, duty cycle, and spatio temporal. Additionally, the EMF evaluation might be carried out for actual circumstances. In [23], their results found that the actual exposure level is quite lower compared to the theoretical exposure for 5G mMIMO. In this investigation, the power weight ρ_w is added as a reduction in the entire used power P_T which is used to calculate the power density [24, 25].

Thus, Equation. (10) changed to Equation. (11) which presents the compliance distances RCD.

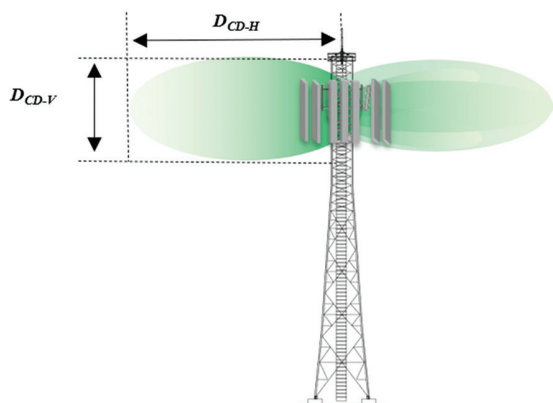
$$R_{CD} = \left(\frac{1}{4\pi} \sum_{n=1}^N \sum_{f > 30 \text{ MHz}}^{300 \text{ GHz}} \frac{\rho_{w,f} \cdot P_{T,f} \cdot G_{A,f}}{S_{inc,RL,f}} \right)^{1/2}$$


Fig. 2. Illustration of the Horizontal RCD-H and Vertical RCD-V compliance distances for macro site installed on the rooftop tower.

6. TER AND CD SIMULATION FOR 3-SECTORS BS

In the three-sector model, the sector's directions are horizontally separated by 120 degrees, so in this study, the model uses 0/120/240 degrees for sectors 1/2/3. This section discusses the theoretical TER and CD figures calculated for different scenarios of BSs that use a three-sector and examines the de-concentration options by using different azimuths for antenna directions. Table 3 lists a typical configuration for a three-sector site used for the TER and CD calculations. The BS is equipped with 6 technologies that transmit at the same time (GSM at 900 MHz, UMTS at 900 MHz, LTE at 800/1800/2100 MHz, and NR at 2600 MHz). In many countries, some operators run the 5G NR at higher frequencies such as 3.5GHz, or millimeter waves (mM) 28/39 GHz, but here the 5G NR is taken at 2.6 GHz because the calculations are validated in real-life sites operating with the frequencies and configurations listed in Table 3, and this is described in detail in next section.

Mathematical simulations are carried out to determine the TER and CD results for the proposed model using Equation (9) for the antenna's azimuths compared to the normal default azimuths where all antennas have the same direction. In this simulation, the TER

& CD are calculated for different scenarios as described in the below paragraphs.

Scenario A: Applying the default azimuth, where all the antennas (of one sector) are directed in the same azimuth angle, this includes all technologies that transmit toward one direction. This design is commonly deployed in life networks where all technologies in one sector are connected to one multi-band antenna, or separate antennas but directed in one direction. For both, all transmitters radiate in the same sector direction as illustrated in Fig. 3.

Table 3. The configurations of 3-Sector BS site with 6 Technologies

Site Setting	2G 900	3G 900	4G 800	4G 1800	4G 2100	5G 2600
Freq. Band (MHz)	900	900	800	1800	2100	2600
Freq. BW (MHz)	4	4.2	10	10	20	60
Number of Tx	2T	1T	2T	2T	4T	64T
Number of Rx	2R	2R	2R	4R	4R	64R
Tx Power (Watt)	40	40	40	60	60	200
System Load	95%	95%	95%	95%	95%	95%
Ant. Gain (dBi)	16.6	16.6	16.2	16.5	17	24.8
Horizontal BW	60o	60o	65o	65o	65o	65o
Vertical BW	7.5o	7.5o	7.8o	6o	6o	6.5o
Ant. Tilt Angle	-6o	-6o	-6o	-6o	-6o	-6o
Ant. Height	35 m	35 m	35 m	35 m	35 m	35 m

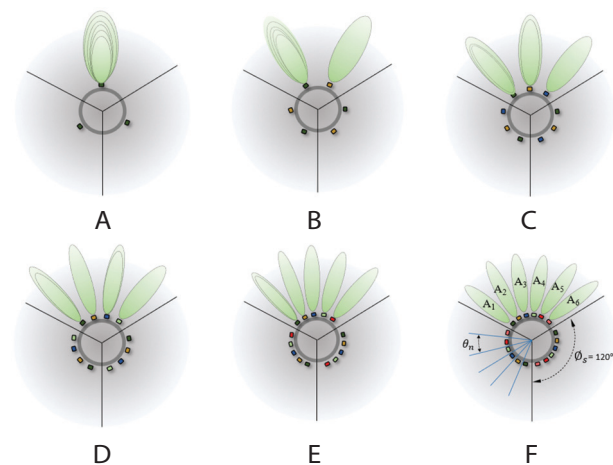


Fig. 3. Illustration of using different number of antennas in one sector in the 3-sectors model. A: 1 azimuth, B: 2 azimuths, C: 3 azimuths, D: 4 azimuths, E: 5 azimuths, F: 6 azimuths

Scenarios B, C, D, E, F: Applying the proposed model by using different azimuths, the angle separation for each technology is calculated using Equation (9). In B, two antennas are used per sector, the first antenna transmits the 900/800/1800/2100 MHz, and the second antenna transmits the 2600 MHz. In C, three antennas are used per sector (1: G900/U900/L800, 2: L1800/L2100, 3: N2600). In D, four antennas are used (1: G900/U900, 2: L800, 3: L1800/L2100, 4: N2600). In E, five an-

tennas are used (1: G900/U900, 2: L800, 3: L1800, 4: L2100, 5: N2600). In F, six antennas are used (1: G900, 2: U900, 3: L800, 4: L1800, 5: L2100, 6: N2600). Practically, the RF engineers design the antennas according to the site requirements and of course consider the company's strategy for deploying the technologies as most of the networks start with the classic system (2G, and 3G), then grow to advanced solutions including 4G and 5G.

The simulation results are concluded in Table 4 which lists the θ_n values for each technology (antenna) calculated for the corresponding scenario. Also, it shows the CD distances (in meters) reference to ICNIRP for the general public, also it lists the reduction percent compared to the default setup using one direction for all technologies of scenario A. The results show that using 2 azimuths in scenario B gives less compliance distance by 23.3% compared to scenario A which has one direction. The 3 azimuths in scenario C give 35.9% less CD, the 4 azimuths in scenario D gives 39.8% less CD, the 5 azimuths in scenario E gives 41.3% less CD, and the 6 azimuths in scenario F gives 43.4% less CD compared to scenario A. Furthermore, Table 5 lists more detailed results of the horizontal and vertical compliance distances for the mentioned six scenarios references to both standards ICNIRP and FCC limits.

Table 4. The configurations of the 3-Sector BS site that is equipped with 6 Technologies

Antenna's Azimuths	Separation Angles θ_n (degrees)						CD (m)	CD Reduction (%)
	G9	U9	L8	L18	L21	N26		
1 Azimuth				120.0°			15.58	0.00%
2 Azimuths			70.3°			49.7°	12.63	-23.3%
3 Azimuths		47.9°		22.4°		49.7°	11.45	-35.9%
4 Azimuths	30.4°	17.5°		22.4°		49.7°	11.13	-39.8%
5 Azimuths	30.4°	17.5°	11.2°	11.3°		49.7°	11.02	-41.3%
6 Azimuths	15.2°	15.2°	17.5°	11.2°	11.3°	49.7°	10.85	-43.4%

Table 5. The horizontal and vertical compliance distances for the general public and occupational workers referenced to ICNIP and FCC limits

Antenna's Azimuths	ICNIRP GP (m)		ICNIRP OW (m)	
	CDH	CDV	CDH	CDV
1 Azimuth	15.58	2.51	6.97	1.12
2 Azimuths	12.63	1.92	5.34	0.86
3 Azimuths	11.45	1.60	4.46	0.72
4 Azimuths	11.13	1.51	4.19	0.67
5 Azimuths	11.02	1.47	4.09	0.66
6 Azimuths	10.85	1.42	3.94	0.63

7. IN SITU ASSESSMENT FOR TER DECONCENTRATION

A field experiment is done in a life network to examine the results of the proposed model by applying the antenna's azimuths with angles calculated using Equation (9). To have an accurate result, the experiment is done for 4 Macro-BS sites (one cluster) that service a

residential populated district at Khubar city in the Kingdom of Saudi Arabia as shown in Fig. 4. The 4 sites belong to one public land mobile network operator (PLMN), all sites have the same configuration of three-sector, and each is equipped with 6 systems that transit as co-located in a multi-technology site.

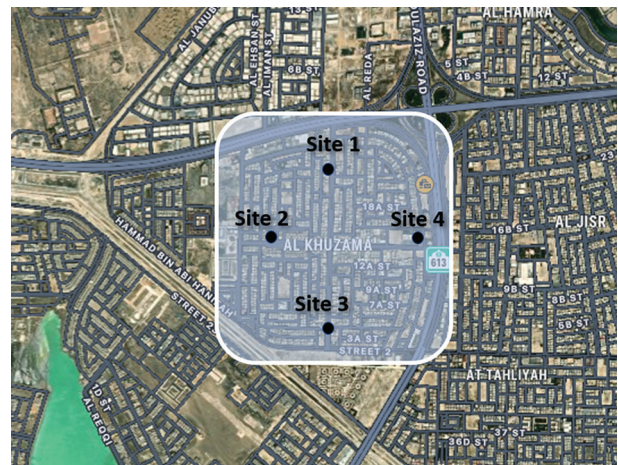


Fig 4. Google Earth map for the 4 sites where the field experiment is done

The 4 sites use 0/120/240 degrees for the sector's azimuths, where the 0 degrees starts from the north geographical direction and increases clockwise. All the sites were deployed with scenario B using two antennas, one for G9/U9/L8/L18/L21, and the second antenna for N26. Initially, all antennas were directed toward the same azimuth and were transmitted in the same direction. Then, the 2nd antennas of all sectors are redirected (rotated) to new directions with azimuths 60/180/300 degrees which gives 60 degrees as horizontal separation angles between the two antennas in each sector.

Two types of data are collected to assess the total exposure before and after applying the antenna azimuth changes, and also to evaluate the effects (impact) on the coverage signal level and capacity, as follows:

- Power density field measurement (radiation meter).
- TER from Geo-location data (system records).
- Signal level field measurements (drive test).
- Network's OSS KPIs data (system records)

7.1. POWER DENSITY FIELD MEASUREMENT

Field measurements are conducted to measure the power densities at two points in Site 1 (P1 and P2) as shown in Fig. 5. (A) and (B). Location P1 is intentionally selected facing the initial direction of the antennas at 0/120/240 degrees, and location P2 is selected facing 60/180/300 degrees.

The SRM-3006 radiation meter was configured to scan the downlink (DL) frequency ranges, the team collected approximately 7,200 measurement samples for each system technology at each point with a scan rate of 0.67 sample/second. All measurements were

conducted at the same time during the highest hours of traffic (from 07:00 to 10:00 pm). For each point, the measurements are done twice, before the antenna's azimuth changes (Pre), and after the azimuth change (Post). The results show that TER decreased at P1 by -5.41% while it increased at P2 by 5.95% referenced to the ICNIRP limit.

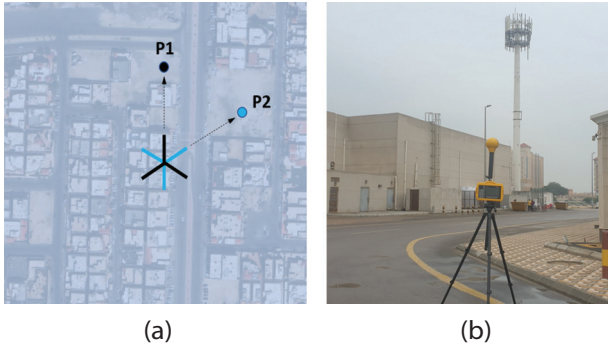


Fig. 5. (a) The measurement locations P1 and P2 at site 1, and (b) the radiation meter position which is in-front of the antenna's main direction.

7.2. TER EVALUATION FORM SYSTEM RECORDS

The network Operation Support System (OSS) continuously records and archives the statistics and information about network traffic and performance. The TER is evaluated for the whole cluster before and after the antenna changes (Pre-and-Post) using the geo-location data recorded in the OSS system.

The received levels of all technologies from all devices in the cluster are recorded before and after the azimuth changes, and it's used to calculate the TER. The geo-location system gives the data as average for pixels of 50x50 meters each, the area under this test consists of 954 pixels within the cluster polygon of 2.3 Km². The geolocation data is collected and calculated for two weeks period, one week before and one week after the antenna azimuth change. Fig. 6. summarizes the results which show the average Pre TER is 23.4x10⁻⁶, and it decreased to 18.91x10⁻⁶ after antenna azimuth is changed, this reduction is an improvement of -19.23% in average TER.

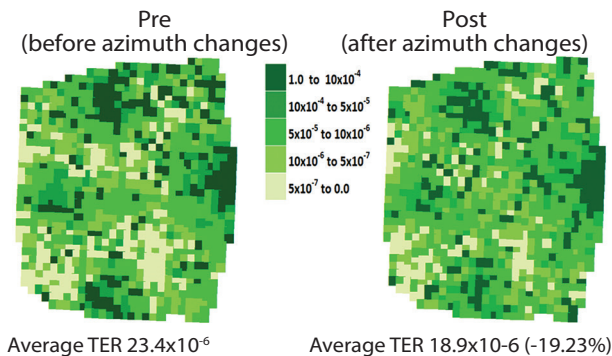


Fig. 6. The Pre and Post average TER from the geo-location records

Also, the TER distribution is evaluated and the results show that the higher range of TER (1.0 to 10x10⁻⁴) was counted for 17.0% on total pixels before the antenna changes, and it decreased to 2.7% after antenna azimuth was changed.

The above results are compared with similar related work found in [26], where a research group from Ericsson proposed an average power feedback controller to reduce the total transmitted over a specified time for the 5G MIMO system which reduces the total exposure ratio, their results show the power density becomes 25 % of the peak power after applying their power control solution. Such a solution gives 15.1% reduction in overall TER assuming the 5G contributes 20% from the total TER.

7.3. SITE'S COVERAGE PERFORMANCE

Drive test field measurements were conducted to evaluate the effect of azimuth rotation on the network received signal levels of all technologies. The team used TEMS Investigation v20 drive test tool which was installed on a laptop PC and connected to GPS and mobile user equipment (UE), as seen in Fig. 7. The measured samples were taken every 0.5 seconds, with over 3,290 measurement samples for each technology.



Fig. 7. The TEMS tool setup for field drive test measurements

The collected measurements include the Rx signal level in dBm of the Broadcast Common Control Channel (BCCCH) for 2G, the Received Signal Code Power (RSCP) for 3G, the Reference Signal Received Power (RSRP) for 4G, and the Secondary Synchronization Reference Signal Received Power (SS-RSRP) for 5G. Table 6 summarizes the results that show the average coverage level in dBm and the delta of Pre vs Post, it indicates that there is no major drawback in signal level in the whole cluster for all technologies.

Table 6. The Pre and Post Rx Levels from Drive Test measurements

Measurement Layer	G 900	U 900	L 800	L 1800	L 2100	N 2600	
	BCCCH	RSCP	RSRP	RSRP	RSRP	SS-RSRP	
Average Rx Level (dBm)	Pre	-68.9	-72.1	-77.6	-82.9	-83.3	-79.2
	Post	-69.3	-71.3	-77.0	-82.0	-82.6	-79.6
	Delta (dB)	0.4	-0.8	-0.6	-0.9	-0.7	0.4

For example, in Fig. 9 and 10, the Rx levels are plotted in the cluster map to display the details of the coverage levels and distribution for L800 and N2600 technologies. The results show the RSRP for L800 almost remains at the same average levels with -77.6 dBm at Pre and -77.0 dBm at Post with -0.6 dB delta, and have the same RSRP accumulated distribution among the cluster area. Also, the same results are obtained for N2600, the SS-RSRP almost remains at the same levels with -79.2 dBm at Pre and -79.6 dBm at Post with 0.4 dB delta.

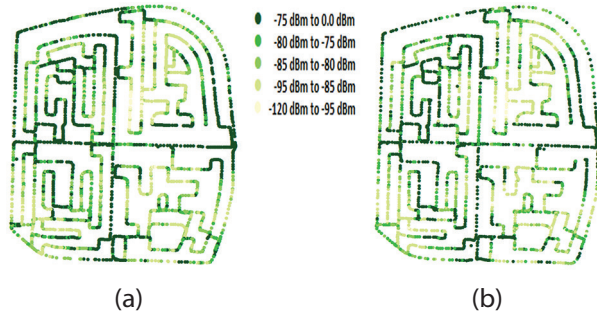


Fig. 8. 4G L800 Pre and Post RSRP levels;
a) Pre: Average RSRP = -77.6 dBm,
b) Post: Average RSRP = -77.0 dBm

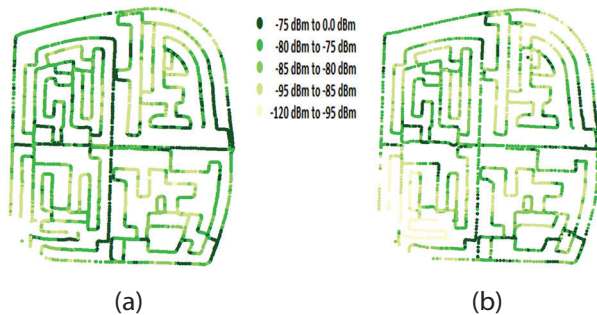


Fig. 9. 5G N2600 Pre and Post SS-RSRP levels;
a) Pre: Average SS-RSRP = -79.2 dBm
b) Post: Average SS-RSRP = -79.6 dBm

7.4. OSS PERFORMANCE RECORDS

Operation support system for mobile networks often use system row counters and statistics typically refer to a variety of metrics and data records that aid in network performance management and monitoring. In this work, some of OSS key performance indicators are employed to evaluate the behavior of the cluster under this study before and after applying the antenna azimuth changes which were implemented on the 09th of January 2024. The daily data were recorded for a continuous six weeks, three weeks pre, and three weeks post to the date of the change.

Fig. 10. shows the total daily carried traffic by the 5G N2600 for the whole cluster (4 sites), and also the system load percentage recorded based on the physical resource block (PRB) utilization. The traffic trend shows no significant change after the implementation date where the average traffic was 1.16 TB and became

1.18 TB, and the PRB utilization was 14.4% and became 14.5% with 0.7% increment.

Also, for N2600, Fig. 11. shows the daily total number of active connected users (simultaneous connection) and the user's throughput, the trend shows a very slight increase in connected users from an average of 295 to 305 users per day. And, there was almost no significant change in the user's throughput which was 78.7 Mbps and became 78.3 Mbps with -0.6% reduction.

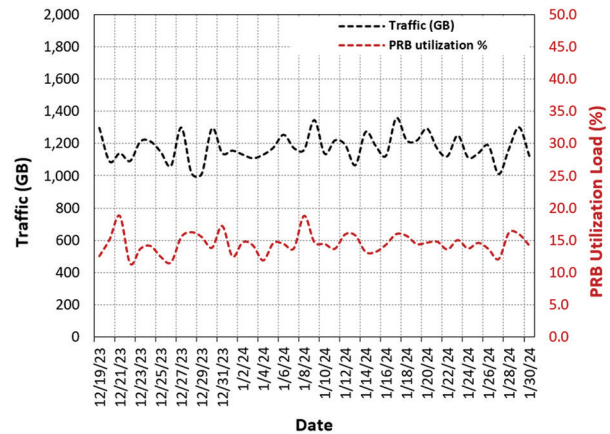


Fig 10. The daily total carried traffic and PRB utilization of the 5G N2600 for the whole cluster

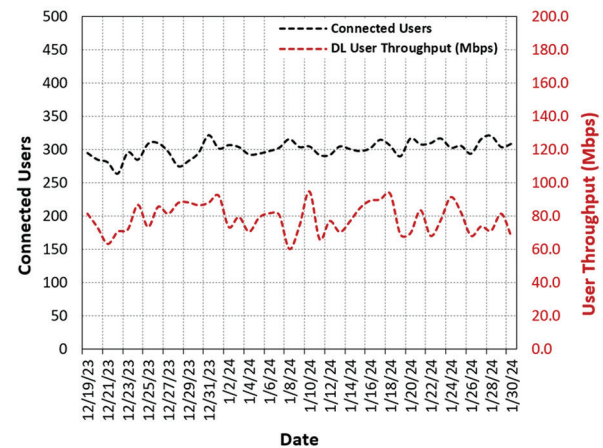


Fig.11. The daily total active connected users and throughput of the 5G N2600 for the whole cluster

8. RECOMMENDED APPLICATIONS OF THE PROPOSED MODEL

Reducing the TRE is the goal of the suggested method, and doing so will inevitably shorten compliance distances. On the other hand, neither the site performance nor the site coverage levels are impacted by the suggested approach. Nearly all national regulators require that the compliance borders be exclusive, inaccessible areas for the general public and shall be marked with warning signs or other obstacles to keep people out. To implement the compliance requirement, it is necessary to determine the compliance distances. The easier it is for mobile operators to meet the

standards, the lower the compliance limits. The authors see these requirements can be assessed in the design phase before implementation, and small adjustments to the antenna orientations can be made using the proposed model to shorten the compliance distances for the necessary cases, particularly for wall-mounted and rooftop sites where the antennas usually will be placed in close proximity to accessible areas.

Also, the mobile operators continue expanding their networks by adding and deploying new technologies such as the 5G NR into the existing on-air sites following the current sector's direction, this approach increases the total exposure ratio and consequently extends the compliance boundaries. For some sites, extending the compliance boundaries to bigger ranges might reach accessible areas specifically for some rooftop and wall-mounted sites. In these kinds of situations, the authors believe that the proposed solution is advantageous and can assist in reducing the compliance distances without affecting network performance or coverage.

9. CONCLUSION

The growing deployment of mobile base stations raises concerns about electromagnetic field radiation. This study proposes a model to reduce total exposure by adjusting antenna azimuths to spread exposure horizontally within sectors. Simulations for a 3-sector base station with six technologies showed that using two azimuths cut compliance distances by 23%, while six azimuths reduced them by 43.4%. A field test in a cluster of four live sites demonstrated a 19.23% reduction in the Total Exposure Ratio (TER) after modifying antenna azimuths using the proposed model. Performance analysis revealed no significant impact on network coverage or capacity across all technologies tested.

10. REFERENCES

- [1] A. Bajpai, A. Balodi, "Role of 6G Networks: Use Cases and Research Directions", Proceedings of the IEEE Bangalore Humanitarian Technology Conference, Vijayapur, India, 8-10 October 2020, pp. 1-5.
- [2] W. J. Koh, S. M. Mochhala, "Non-ionizing EMF hazard in the 21st century", Proceedings of the IEEE International Symposium on Electromagnetic Compatibility and IEEE Asia-Pacific Symposium on Electromagnetic Compatibility, Suntec, Singapore, 14-18 May 2018, pp. 518-522.
- [3] W. El-Beaino, A. M. El-Hajj, Z. Dawy, "On Radio network planning for next generation 5G networks: A case study", Proceedings of the International Conference on Communications, Signal Processing, and their Applications, Sharjah, United Arab Emirates, 17-19 February 2015, pp. 1-6.
- [4] C. Blackman, S. Forge, "5G deployment: State of play in Europe, USA and Asia", Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Brussels, Belgium, Technical Report PE 631.060, 2019.
- [5] L. Chiaraviglio, A. S. Cacciapuoti, G. Martinp, M. Fiore, D. Trucchi, N. B. Melazzi, "Planning 5G Networks Under EMF Constraints: State of the Art and Vision", IEEE Access, Vol. 6, 2018, pp. 51021-51037.
- [6] L. Chiaraviglio, C. Di Paolo, N. B. Melazzi, "5G Network Planning Under Service and EMF Constraints: Formulation and Solutions", IEEE Transactions on Mobile Computing, Vol. 21, No. 9, 2022, pp. 3053-3070.
- [7] C. Weng, H. Wang, K. Li, M. N. S. Swamy, "Azimuth Estimation for Sectorized Base Station With Improved Soft-Margin Classification", IEEE Access, Vol. 8, 2020, pp. 96649-96660.
- [8] FCC USA, "Evaluating Compliance with FCC Guidelines for Human Exposure to Radiofrequency Electromagnetic Fields", FCC OET Bulletin 65, FCC, 1997.
- [9] International Commission on Non-Ionizing Radiation Protection, "Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz)", Health Physics, Vol. 118, No. 5, 2020, p. 483-524.
- [10] GSMA, Public Policy 2019, <https://www.gsma.com/publicpolicy/emf-and-health/emf-policy> (accessed: 2024)
- [11] IEC, "Determination of RF Field Strength and SAR in the Vicinity of Radio Communication Base Stations for the Purpose of Evaluating Human Exposure", IEC Standard 62232:2022, 2022.
- [12] B. Thors, M. L. Strydom, B. Hasson, F. J. C. Meyer, K. Karkkainen, P. Zollman, S. Ilvonen, C. Tornevik, "On the Estimation of SAR and Compliance Distance Related to RF Exposure From Mobile Communication Base Station Antennas", IEEE Transactions on Electromagnetic Compatibility, Vol. 50, No. 4, 2008, pp. 837-848.
- [13] A. Jain, P. Tupe-Waghmare, "Radiation measurements at repeated intervals for various locations of SIU campus and calculation of compliance distance from cell tower", Proceedings of the Interna-

- tional Conference on Automatic Control and Dynamic Optimization Techniques, Pune, India, 9-10 September 2016, pp. 804-808.
- [14] M. Velghe, S. Aerts, L. Martens, W. Joseph, A. Thielens, "Protocol for personal RF-EMF exposure measurement studies in 5th generation telecommunication networks", *Environmental Health*, No. 20, 2021, pp. 1-10.
- [15] R. Werner, P. Knipe, S. Iskra, "A Comparison Between Measured and Computed Assessments of the RF Exposure Compliance Boundary of an In-Situ Radio Base Station Massive MIMO Antenna", *IEEE Access*, Vol. 7, 2019, pp. 170682-170689.
- [16] D. Pinchera, M. Migliore, F. Schettino, "Compliance Boundaries of 5G Massive MIMO Radio Base Stations: A Statistical Approach", *IEEE Access*, Vol. 8, 2020, pp. 182787-182800.
- [17] A. Thielens, G. Vermeeren, D. Kurup, W. Joseph, L. Martens, "Compliance boundaries for LTE base station antennas at 2600 MHz", *Proceedings of the 6th European Conference on Antennas and Propagation*, Prague, Czech Republic, 26-30 March 2012, pp. 889-892.
- [18] F. Hélot, T. H. Loh, D. Cheadle, Y. Gui, M. Dieudonne, "An Empirical Study of the Stochastic Nature of Electromagnetic Field Exposure in Massive MIMO Systems", *IEEE Access*, Vol. 10, 2022, pp. 63100-63112.
- [19] R. Joyce, D. Morris, S. Brown, D. Vyas, L. Zhang, "Higher Order Horizontal Sectorization Gains for 6, 9, 12 and 15 Sectorized Cell Sites in a 3GPP/HSPA+ Network", *IEEE Transactions on Vehicular Technology*, Vol. 65, No. 5, 2016, pp. 3440-3449.
- [20] H. Tataria, K. Haneda, A. F. Molisch, M. Shafi, F. Tufvesson, "Standardization of propagation models for terrestrial cellular systems: A historical perspective", *International Journal of Wireless Information Networks*, No. 28, 2021, pp. 20-44.
- [21] B. Thors, D. Colombi, Z. Ying, T. Bolin, C. Törnevik, "Exposure to RF EMF From Array Antennas in 5G Mobile Communication Equipment", *IEEE Access*, Vol. 4, 2026, pp. 7469-7478.
- [22] S. Aerts, J. Wiart, L. Martens, W. Joseph, "Assessment of long-term spatio-temporal radiofrequency electromagnetic field exposure", *Environmental Research*, No. 161, 2018, pp.136-143.
- [23] S. Aerts, L. Verloock, M. V. D. Bossche, D. Colombi, L. Martens, C. Törnevik, W. Joseph, "In-situ Measurement Methodology for the Assessment of 5G NR Massive MIMO Base Station Exposure at Sub-6 GHz Frequencies", *IEEE Access*, Vol. 7, 2029, pp. 184658-184667.
- [24] IEC, TR 62669 ed2, "Case Studies Supporting, International Electromechanical Commission", IEC Standard 62232, 2018.
- [25] B. Thors, A. Furuskär, D. Colombi, C. Törnevik, "Time-Averaged Realistic Maximum Power Levels for the Assessment of Radio Frequency Exposure for 5G Radio Base Stations Using Massive MIMO", *IEEE Access*, Vol. 5, 2017, pp. 19711-19719.
- [26] C. Törnevik, T. Wigren, S. Guo, K. Huisman, "Time Averaged Power Control of a 4G or a 5G Radio Base Station for RF EMF Compliance", *IEEE Access*, Vol. 8, 2020, pp. 211937-211950.

Comparison Between Different Source Localization and Connectivity Metrics of Spiky and Oscillatory MEG Activities

Original Scientific Paper

Ichrak ELBehy*

University of Sfax,
Faculty of Electronics and Telecommunications,
Department of STIC
Digital Research Center of Sfax, CRNS, Tunisia
ichrakchouda@gmail.com

Abir Hadriche

REsearch Groups in Intelligent Machines, Regim Lab,
Enis, Sfax University
High institute of Music, Sfax, Sfax University
Digital Research Center of Sfax, CRNS, Tunisia
Abir.hadriche.tn@ieee.org

*Corresponding author

Rahma Maalej

University of Sfax,
Faculty of Electronics and Telecommunications,
Department of STIC
Tunis km 10, Cité el Ons, Sfax Technopole,
Sakiet Ezzit, Tunisia
rahmamaalej1234@gmail.com

Nawel Jmail

Miracl Lab, Sfax University
Higher Business School, Sfax University
Digital Research Center of Sfax, CRNS, Tunisia
naweljmail@yahoo.fr

Abstract – Epilepsy is considered the second neurological disease in a coma after stroke. Famous markers of epilepsy are repetitive seizures, their origin is stroma and cortical deformation. A neurologist would be assisted by identifying Epileptogenic Zones EZ when diagnosing epilepsy. Source localization is utilized to identify regions known as EZ, which are of excessive discharges. It consists of both forward and inverse problems. The forward problem models the head through analytical and numerical methods. The inverse problem can be resolved using several techniques to locate the cerebral abnormal sources, via the electrophysiological recording biomarkers. In our study, we will investigate four distributed inverse problem methods: minimum norm estimation MNE, standardized low-resolution brain electromagnetic tomography sLORETA, maximum entropy on the mean MEM, Dynamic statistical parametric maps dSPM, to define epileptic networks connectivity of spiky and oscillatory events. We will examine the epileptic network connectivity using Phase Locking Value (PLV), Phase Transfert Entropy (PTE) for oscillatory events, cross-correlation (CC), and Granger Causality (GC) for spiky events applied on 5 pharmaco resistant subjects. We suggest rating the effectiveness of these networks in locating EZ through a phase of confrontation within iEEG transitory and oscillatory networks connectivity by exploring concordant nodes, their distance, propagation delays connection strength, and their cooperation in recognition of seizure onset zone. All studied techniques of the inverse problem, connection metrics, for both biomarkers of the 5 patients succeed in detecting at least one part of SOZ. sLORETA provides the highest concordant nodes and the closed one for spiky events using CC and GC. sLORETA also depicts the lowest propagation delay for oscillatory events using PTE. Through the 5 patients, MEM, dSPM, and MNE using CC, CG for spiky events, and PTE, PLV for oscillatory activities provide about 72 % of concordant nodes between MEG and iEEG.

Keywords: MEG, Connectivity, Epilepsy, Spike, Oscillation

Received: May 10, 2024; Received in revised form: August 30, 2024; Accepted: August 30, 2024

1. INTRODUCTION

Neurologic illness diagnosis is increasingly focusing on noninvasive modalities such as electroencephalography (EEG) and magnetoencephalography (MEG) approaches [1, 2]. EEG and MEG recordings provide a high temporal and spatial precision in highlighting brain activity and malfunction, particularly in epilepsy diagnosis. MEG requires less knowledge regarding cerebral tissue

to distinguish the origins of epileptic seizures. This could be a major cause to predispose the benefits of MEG on EEG [3, 4]. As a result, despite its cost, neurologists and biomedical researchers are exploring MEG as a supplementary method for epilepsy diagnosis. Alternatively, numerous brain regions might be involved, either as propagation zones or as epileptic discharge generators [5-7]. To identify accurate EZ, neurologists depend on network connections of MEG characteristic signals [8, 9].

Examining and assessing the network connection of MEG biomarkers (spikes and oscillatory events) [10-12] is required, beginning with source localization (forward and inverse issue) [13-15] and progressing to calculating connectivity measures [16]. Four distributed inverse methods are proposed to be investigated: minimum norm estimation MNE [17], dynamic statistical parametric maps dsPM [18], standardized low-resolution brain electromagnetic tomography sLORETA [19], and Maximum Entropy on the Mean MEM. To compare connectivity measures of epileptic spiky and gamma oscillatory events, two connectivity metrics of each event are used. Functional connectivity can be computed using several approaches, like phase synchronization measures [14], amplitude envelope correlation [20], information theoretical approach [21], and other methods.

In this study, the effectiveness of the inverse problem approaches is evaluated by exploring different connectivity metrics of two types of biomarkers to define seizure onset zones and epileptic network complexity. These inverse approaches (MNE, dsPM, sLoreta, MEM) are distributed methods that use the same initial assumptions to construct active zones with alternative hypotheses. MNE normalizes the current density map and Minimum norm estimation (MNE). It has the advantage of not requiring a specific number of sources in advance. Whereas dsPM uses noise covariance for normalization, and substitutes noise covariance with data covariance. sLORETA supposes that the entire brain areas are active within smoother maps. Finally, MEM is a technique for locating dispersed sources originally proposed that cortical parcels would be used to organize brain activity, with each active parcel. Hence, MEM can estimate a contrast of current density within each active parcel.

Connectivity brain measures are intended to look at how cortical networks interact with each other. There are three types of connection between regions: structural ("directed functional connectivity"), functional ("non-directed statistical associations"), and effective ("causal interactions"). Using Brainstorm, multiple connection measures are computed for directed and non-directed functional connectivity investigations.

Functional connectivity is estimated using Phase Locking Value (PLV) which is an alternative class of measures that considers only the relative phase of two signals by computing a phase locking value between them [22, 23]. The concept of phase locking is fundamental in dynamical systems and has been used in control systems (the phase-locked loop) as well as in the analysis of nonlinear, chaotic, and non-stationary systems. Since the brain is a nonlinear dynamical system, phase locking is a suitable method for quantifying the interaction of oscillatory gamma events. Phase Transfer Entropy (PTE) is an instantaneous phase time series, quantified by phase transfer entropy (PTE) [24]. PTE estimates whether the past of both source and target time series influences the ability to predict the target time series' future which is also suited for studying gamma networks. Correlation

is a non-directed connectivity metric that measures the relationship of two time series. Without further preprocessing of the input time series. Correlation is sensitive to volume conduction and is not frequency specific chosen to determine spiky networks. Finally, GC is a functional connectivity [25], developed in economics but recently piqued the interest of the neuroscience community since it enabled statistical influence to be estimated without the need for direct intervention [26] and also chosen to compute spiky networks.

This preprocessing chain was applied on 5 pharmaco-resistant epileptic subjects, where neurologists examined and proved the efficacy of studied biomarkers (spiky and gamma oscillatory events).

As a result, MEG concordant nodes were determined, and connection strength propagation delays and their cooperation in recognition of seizure onset zone SOZ were computed. Through 5 patients, sLORETA exhibits the highest concordant nodes and the lowest propagation delays, for both biomarkers. All proposed inverse problems within connectivity measures provide at least one part of SOZ and about 72% of nodes are detected by MEG and seen in iEEG. CG and PTE enhance the connection strength for spiky events and oscillatory activities respectively.

2. MATERIALS AND METHODS

2.1. MATERIALS

EEGLAB and Brainstorm Toolbox (a freely available collaborative tool for cerebral signal processing) was used for all analysis phases on "MATLAB" Mathwork, Natick, MA [27].

The explored signals were both MEG and iEEG for five pharmaco-resistant subjects [28]. This research involved a magnetoencephalography MEG registration for 5 patients with drug-resistance epilepsy. Acquisition and preprocessing steps were used in the clinical Neurophysiology Department of Marseille's "La Timone" hospital. An experienced neurologist (M.G.) validated patients with constant and frequent epileptic spiky and gamma activity. Registration was made with closed eyes and no activation method or movement, a 151-gradiometer device (CTF Systems Inc., Port Coquitlam, Canada) was used to capture the MEG signal. 20 epochs of 5 s each of sampling at 1025 Hz were recorded.

Intracerebral EEG signals were gathered as the Talairach stereoscopic method [22], sampled at 512Hz. Clinical, neurophysiological, and anatomical features of each patient as in [5] were taken into consideration to designate cerebral marks. CT scan and MRI examinations were detailed in [5].

In this study, for each subject, an average of 30 epileptic spikes and gamma oscillations were investigated. In total, about 300 spiky events and oscillatory ones were studied [5].

Additionally, the Institutional Review Committee of the French Institution of Health INSERM gave its approval to our experiment. The clinical data for our patients is shown in Table 1 [5].

al to our experiment. The clinical data for our patients is shown in Table 1 [5].

Table 1. The clinical data for our patients

Patients	Gender/age	Structural MRI	Histological diagnosis	MEG spike occurrence	MEG: preop versus post-op	Epilepsy surgery	Surgical outcome. Engel class (follow up)
1-BeA	F17	R lateral occipitotemporal FCD	FCD	Subcontinuous	Preoperative	R occipitotemporal cortectomy	Class 3 (8 years)
2-ZC	F26	Normal	Gliososis	Abundant	Preoperative	L occipitotemporal cortectomy	Class 3 (5 years)
3-BC	F25	L premotor FCD	FCD	Abundant	Preoperative	L premotor cortectomy	Class 3 (6 years)
4-DT	M25	R basal occipitotemporal FCD	FCD	Abundant	Preoperative	R anterior temporal lobectomy	Class 2 (2 years)
5-BoA	F31	R parietal ischemia	Gliososis	Abundant	postoperative	R parietal cortectomy	Class 4 (7 years)

2.2. METHODS

Brainstorm, EEGlab [27], Fieldtrip toolbox, and MATLAB (MathWorks, Inc.) tools were used for all signal pre-processing.

As in [5], spiky and oscillatory events were selected visually by an expert, and then a filtering step was applied to eliminate artifacts and overlap between activities. Time windows of joined spikes and oscillations independently are made. FIR filtering is applied on each window: a band-pass filter [10 45] Hz was used to eliminate slow component of spiky windows and [29] Hz for oscillatory windows. For both filters, the ripple amplitude is equal to $R_p = 3\%$, and the attenuation in the stop band is $R_s = 30$ dB.

2.2.1. Source Localization of MEG Signal

Forward Problem is a way to describe the head using analytical and numerical approaches such as boundary element method (BEM), finite element method (FEM), and finite difference method (FDM). Since the thickness of our skull is not uniform across the head, MRI determines local conductivity characteristics. Furthermore, the forward problem is solved as described in [13], by creating a multiple spheres head model for each patient. BrainVisa software was used to segment and mesh the cortex and scalp surfaces. Finally, Matlab's Brainstorm toolbox is used to register the MRI and sensors of each analyzed patient [5-24].

Inverse problem is explored to define sources that generate scalp measurements (MEG in our case) to understand cerebral function and dysfunction [30, 13, 14]. For epilepsy, the inverse problem of source localization is solved to recognize relative regions of excessive discharges and seizure buildup (damaged cerebral tissue [31]). An inverse problem is an underdetermined problem (multiple sources can yield the same potential field) so there is no unique solution. Therefore, to identify an effective solution, different hypotheses (neurophysiological, biophysical, and anatomical) as well as regularization approaches are tested and applied. Di-

polar source localization was investigated as a solution, however assumption about employed dipole number leads researchers to adopt scattered approaches.

The four proposed inverse problem methods are based on a 3D current source solution grid with fixed positions configuration that necessitates only regularization parameters to reduce the noise effect and ensure a stable source configuration. These techniques didn't require a prior source number constraint as the dipolar solution did. They provide different aspects of source localization: from simplicity and computational efficiency of MNE to statistical robustness of dSPM, precision localization capabilities of sLORETA, and incorporation of prior information in MEM. Moreover, these methods are suitable for analyzing both spiky and oscillatory events, which are crucial for understanding the dynamics of epileptic seizures" [32].

In the next section, the explored four distributed inverse problem approaches: MNE, dSPM, sLORETA, and MEM are briefly described.

Minimum norm estimation (MNE)

Minimum norm estimation (MNE) has the advantage of not requiring a specific number of sources in advance. On the other hand, it necessitates a regularization that may affect on chronological series estimation: cross-talk between sources. As a result, imposing a parsimony constraint on sources may be beneficial. An original solution of 3D current configuration that matches the analyzed signal within a minimum intensity (smallest L2-norm) is offered by the minimum norm estimate (MNE) described by [33]. MNE was proposed by [33]. It achieved an exceptional 3D current configuration solution that fits the signal under within the lowest intensity (smallest L2- 277 norm). This hypothesis can drown deeper sources since MNE focuses on superficial sources. The MNE formula is shown in Equation 1.

$$SS_{MNE} = G^T(GG^T + \lambda C)^{-1}G \quad (1)$$

λ Indicates the regularization parameter, while C represents the noise covariance matrix. Weighted solutions of MNE may be found in dSPM, eLORETA, and cMEM (their formula is based on MNE sources, S_{MNE}).

Dynamical Statistical Parametric Mapping (dSPM)

Dale et al. Suggest Dynamical Statistical Parametric Mapping (dSPM) as a different inverse problem solution. Dale et al. recommend a normalization based on a minimum norm estimate of each source noise (obtained from the MNE noise covariance matrix) as an inspiration from MNE [34-36].

Equation 2 describes dSPM as a least-squares or weighted minimal norm solution.

$$S_{dSPM} = W_{dSPM} S_{MNE} \quad (2)$$

$$W_{dSPM}^2 = \text{diag}(S_{MNE} C S_{MNE}^T) \quad (3)$$

Standardized LOw Resolution brain electromagnetic tomography (sLORETA)

According to Pascual-Marqui RD [29], the entire brain regions are activated in sLORETA's smoother maps. sLORETA swaps the noise covariance with data covariance, which accounts for uncertain number of simultaneous source activations. The ratio of the covariance matrix of sources to the gain matrix is the inverse operator L for the sLORETA technique.

$$L = R/J \quad (4)$$

R represents the source covariance matrix, assumed to be the identity gains matrix.

Maximum Entropy on the Mean (MEM)

MEM is based on a probabilistic (Bayesian) technique to estimate current source intensities from the data's informative content. MEM explores cortical parcels to organize brain activity, with each parcel being active or inactive. MEM estimates the contrast of current density inside each active parcel. The MEM's primary premise is that brain activity is segmented into discrete units. As a result, a source's activity inside a patch is correlated with its neighboring sources. An essential notion that allows MEM to be sensitive to the geographic extent of sources on cortical surfaces is using a spatial model in the MEM framework [37]. Recently, MEM has been expanded to temporal frequency to locate oscillatory and synchronous generators.

2.2.2. Functional connectivity metrics

Measures of brain connectivity enable to define interaction of cortical network. There are three types of connection between regions: structural ("directed functional connectivity"), functional ("non-directed statistical associations"), and effective ("causal interactions").

Multiple connection measures (for directed and non-directed functional connectivity investigations) were determined using Brainstorm. Computing a bivariate measure between two interested geographic time series pairs is a standard method of performing connectivity analysis.

Each brain region can be seen as a node on a connectivity graph representing resulting the connec-

tome, with connectivity metrics displayed above each graph edge. Functional connectivity was estimated using Phase Locking Value (PLV), Phase Transfer Entropy (PTE) for oscillatory gamma activities, and Correlation and Granger causality for spiky events. This choice was justified by the importance and the effect of the frequency factor and the directionality (directed versus non directed). In Table 2, a summary of these functional connectivity metrics was gathered.

Table 2. Functional connectivity metrics

Metrics	Domain	Directionality	Static(s) Dynamic(s)
Correlation	Time	Non directed	S
Granger Causality	Time	Directed	S
Phase Locking Value (PLV)	Phase	Non directed	S
Phase Transfer Entropy(PTE)	Phase	Directed	D

Phase Locking Value (PLV)

An alternative class of measures considers only the relative phase of two signals by computing a phase locking value between them [38, 22]. The concept of phase locking is fundamental in dynamical systems and has been used in control systems (the phase-locked loop) as well as in the analysis of nonlinear, chaotic, and non-stationary systems. Since our brain is a nonlinear dynamical system, phase locking is a suitable method for quantifying cortical interactions.

A more pragmatic reason for using PLV in studies of LFPs, EEG, and MEG is its resistance to amplitude fluctuations (which may contain less information about interactions) [39, 40]. PLV is an absolute value of the mean phase difference between two signals expressed as a complex unit-length vector [37-40]. If marginal distributions of two signals are uniform and signals are independent, the relative phase will be uniform and equal to zero, otherwise, (for strongly coupled signals), PLV approaches unity. PLV is frequently used to describe phase synchronization between two narrow-band signals. Consider a pair of real signals, $S_{1(t)}$ and $S_{2(t)}$, which have been band-pass filtered to a desired frequency range. The Hilbert transform can be used to obtain analytical signals from $S_{1(t)}$ and $S_{2(t)}$:

$$z_{i(t)} = S_{i(t)} + \text{HT}(S_{i(t)}) \quad (5)$$

Using analytical signals, relative phase between $z_{1(t)}$ and $z_{2(t)}$ can be computed as,

$$\Delta\phi(t) = \arg\left(\frac{z_1(t) z_2^*(t)}{|z_1(t)||z_2(t)|}\right) \quad (6)$$

The instantaneous PLV is $\text{PLV}(t) = \left| E \left[e^{j\Delta\phi(t)} \right] \right|$.

Phase Transfer Entropy (PTE)

PTE is a directed connectivity measure that evaluates transfer entropy (TE) between two instantaneous

phase time series [24]. TE calculates whether the history of both the source and target time series can affect the ability to forecast the target time series' future.

In PTE, if a phase signal $\phi\tilde{x}(t)$ causes the signal $\phi\tilde{y}(t)$, the mutual information between $\phi\tilde{y}(t)$ and the past of $\phi\tilde{x}(t)$ is computed.

Cross- Correlation (CC)

Correlation is a non-directed connectivity metric that measures the relationship of two time series, without further preprocessing of input time series. Correlation is sensitive to volume conduction and has no frequency specification. CC of signals gathered from active regions (showing a local energy peak) consists of estimating the degree of similarity between these locations using the correlation coefficient r presented in equation 6.

$$r(\tau) = \frac{cov(s1(t), s2(t - \tau))}{\sqrt{var(s1) \cdot var(s2)}} \quad (7)$$

Where r is the correlation coefficient derived between two signals (two cortical areas), cov is the covariance, var is the variance, and τ is the offset considered between studied signals.

Granger causality (GC)

Granger causality (GC) is a technique of functional connectivity created by Clive Granger in the 1960s [25] and improved by John Geweke into its current form [41]. GC specializes in economics, but has lately attracted the interest of the neuroscience community.

Previously, neuroscience depended on lesions and stimuli applied to the nervous system portion to investigate their influence on others. However, GC offered statistical measures without requiring direct intervention [26]. Even though GC has been extended to non-linear, multivariate, and time-varying conditions.

In the time domain, this may be shown as follows: if X represents a signal, it may be represented using a linear autoregressive model estimate (AR model) in two ways:

$$X(t) = \sum_{k=1}^p [A_k x(t - k)] + e_1 \quad (7)$$

$$X(t) = \sum_{k=1}^p [A_k x(t - k) + B_k y(t - k)] + e_2 \quad (8)$$

p is the quantity of previous knowledge that will be used to forecast future samples, also known as model order. In both expressions, the first model X uses only its history (and present), but the second includes the past (and present) of a second signal y . The model considers just past signals ($k \geq 1$) and ignores the current connection, making it less vulnerable to volume conditions.

The measure of GC is defined as follows:

$$F_{y \rightarrow x} = \ln \left(\frac{var(e_1)}{var(e_2)} \right) \quad (9)$$

0 if $var(e_1) = var(e_2)$ and a non negative $F_{y \rightarrow x} = \ln \left(\frac{var(e_1)}{var(e_2)} \right) > 0$.

$var(e_1) \geq var(e_2)$ Always holds, as the model can only improve when adding new.

3. RESULTS

In Fig. 1, active areas of selected epileptic spiky MEG data using MNE, dSPM, sLORETA, and MEM are depicted. MNE, dSPM, and sLORETA methods produced numerous active regions ROI (Region Of Interest), while MEM produced noticeably fewer active regions.

The four distributed inverse problem methods are explored to evaluate the coupling rate between active regions of the subject using 20 scouts in each hemisphere. For each Region of Interest (ROI), time series for both spiky and gamma oscillatory activities have been reconstructed.

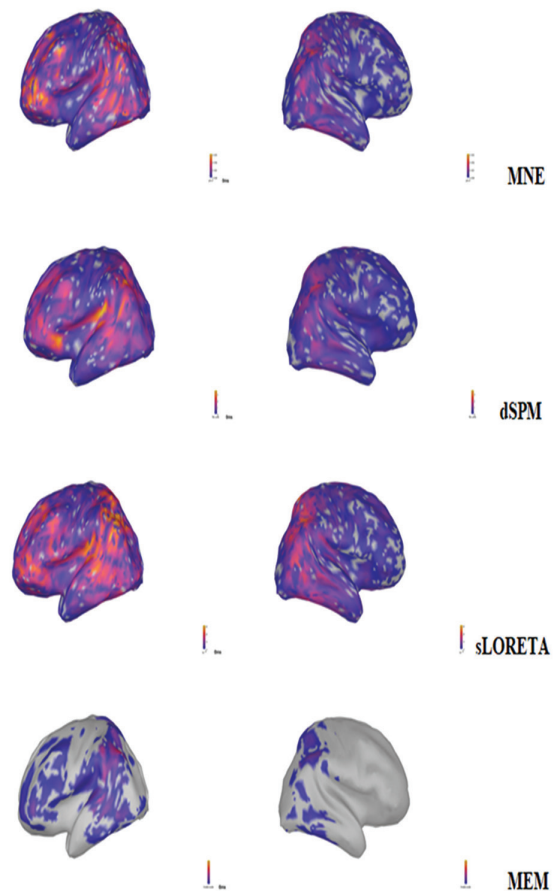


Fig.1. Active regions using 4 inverse problem methods (MNE,dSPM,sLORETA, and MEM) of spiky events

The proposed connectivity metrics are computed: Phase Transfer Entropy (PTE), Phase Locking Value (PLV) for oscillatory events, cross- correlation (CC), and Granger Causality (GC) for spiky events, as non-directed and directed functional connectivity analyses using Brainstorm.

Connectivity scores are shown as links drawn between regions of interest. These ROI are displayed as nodes labeled along graph circumference with Intensity threshold, (minimum or maximum connectivity).

In Fig. 2, GC is presented as a timing directed static measure of patient 1 spiky connectivity networks obtained by MNE, sLORETA, MEM, and dSPM.

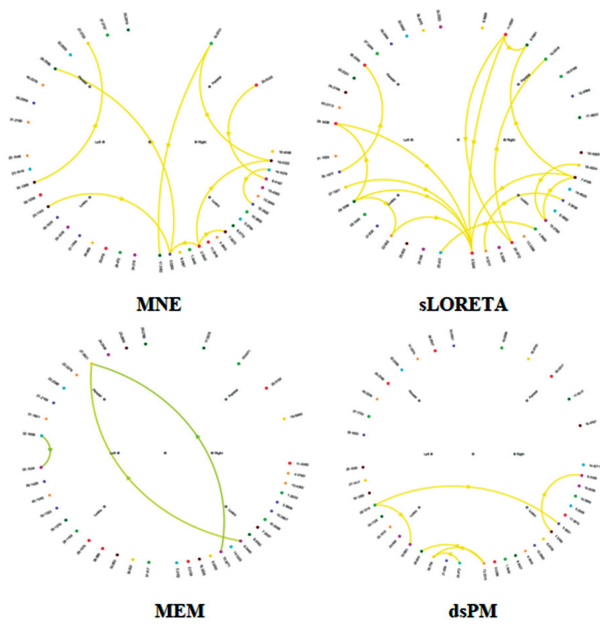


Fig. 2. Epileptic spiky network connectivity for patient 1 using Granger Causality obtained by: MNE, sLORETA, MEM, and dsPM

For the following figure, lags are imposed in the range [-149, 150] ms, the maximum threshold in each method is set to 2.00, and distance filtering to 0 mm. Connectivity is depicted by a link with direction “in” or “out” thanks to this metric specification, a measure of directed functional connectivity is obtained. dsPM shows a strong connection between ROI and MEM represents a weak connection.

Fig. 3 depicts the gamma oscillatory connectivity networks of patient 1 obtained by MNE, sLORETA, MEM, and dsPM using PLV as a phase non directed, static connectivity metric.

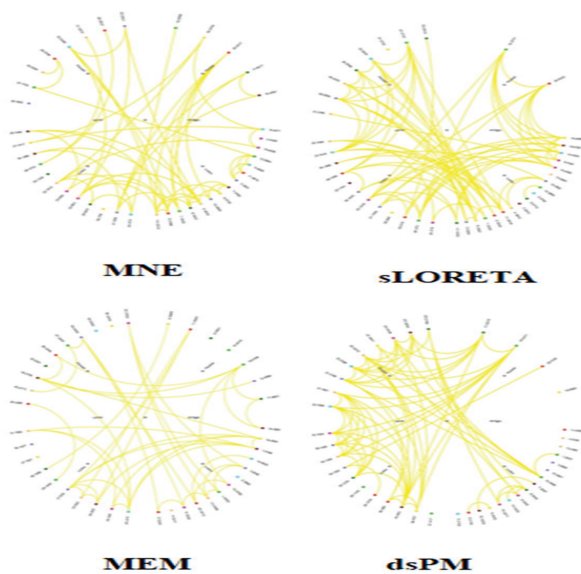


Fig. 3. Patient 1 gamma network connectivity obtained by: MNE, sLORETA, MEM, and dsPM using PLV

In Figure 3 lags in the range [-149, 150] ms are imposed. The maximum threshold in each method was set to 0.99 and distance filtering to 0 mm. A frequency band between 15 and 45 Hz (that admits the gamma band) was chosen. MNE shows a strong connection between ROI and dsPM represents a weak connection.

In Figure 4, we gathered nodes in common and a number of links between active regions for patient 1 spiky events using 4 inverse problem techniques and two metrics (CC and GC) for each one.

For both CC and GC, MEM depicts the lowest number of connections and nodes in common, hence MEM presents the lowest complexity for epileptic spiky networks.

In Fig. 5, nodes in common and several links between active regions for patient 1 gamma oscillatory events using 4 inverse problem techniques and two metrics (PLV and PTE) for each one are depicted.

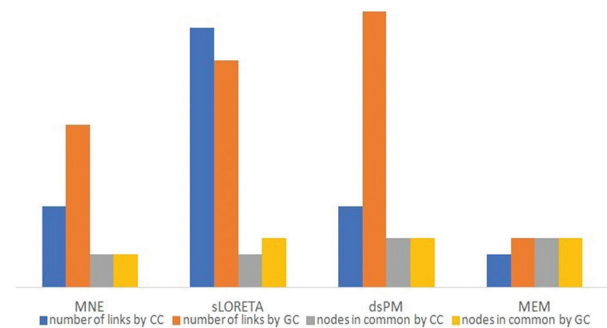


Fig.4. Patient 1 nodes in common and several links for spiky events using 4 inverse problem techniques and two metrics (CC and GC)

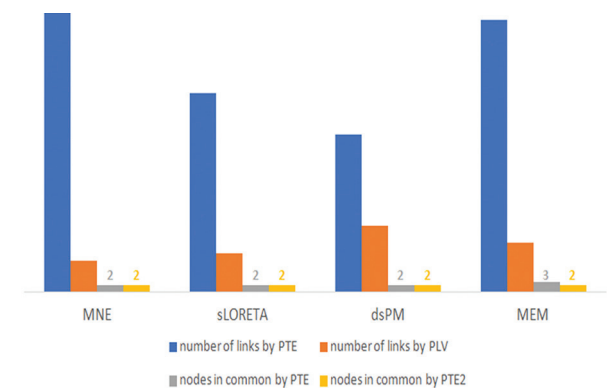


Fig. 5. Patient 1 nodes in common and several links for gamma events using 4 inverse problem techniques and two metrics (CC and GC)

For both PLV and PTE, MEM depicts the lowest number of connections and nodes in common, hence the lowest complexity of epileptic oscillatory networks. Nevertheless, MEM was able to detect parts of SOZ.

The connectivity strength of each metric of studied methods is computed, in which maximum, mean, and minimum values are applied, and depicted in Fig. 6.

GC as a functional connectivity metric provides high-

er connection strength for entire investigated inverse problem techniques with a slightly important value for sLORETA

In Fig. 7, the maximum, median and, minimum distance between common MEG nodes for the distributed methods applied to epileptic oscillatory events of patient 1 are depicted.

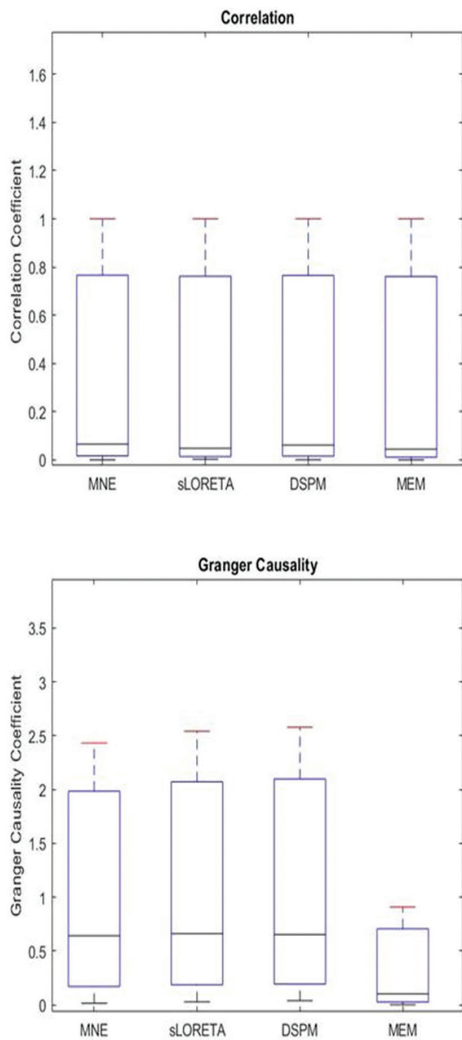


Fig. 6. Connection strength, by: MNE, sLORETA, MEM and, dsPM using CC, and GC for epileptic spiky events

For both connectivity measures, MNE depicts the closest nodes of interest within an average distance of 0.6 mm.

In Table 3, the study conducted on spiky and gamma events (using 4 distributed inverse problem techniques and 2 connectivity metrics) applied on patients in recognition of SOZ and propagation delays is gathered.

For the entire sets of patients that were investigated in this work, we noticed that the obtained network connectivity for both biomarkers and different inverse problems and metrics provide the same highlighted results (detection of SOZ, nodes in common between depth and surface , propagation delays and distance), within slight differences for patient 4.

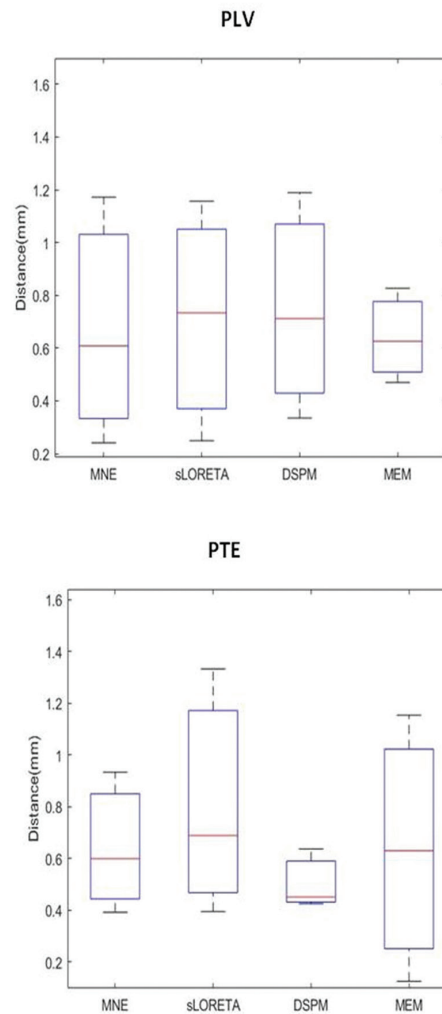


Fig. 7. Maximum, median, and minimum distance between nodes of interest obtained for gamma oscillations networks by PTE and PLV

Table 3. Recognition of SOZ per patient and Propagation delays

Methods	Spiky events		Oscillatory events	
	SOZ recognition	Average propagation Delays in ms	SOZ recognition	Average propagation Delays in ms
Non-directed MNE	yes	22	yes	24
directed MNE	yes	21	No	23
Non-directed sLORETA	yes	18	yes	20
directed sLORETA	yes	19	yes	19
Non-directed dSPM	yes	22	yes	23
directed dSPM	yes	22	yes	24
Non-directed MEM	No	24	yes	25
directed MEM	yes	23	no	23

4. DISSCUSION AND CONCLUSION

In this study, the relationship between epileptic spiky, and oscillatory events for 5 pharmaco-resistant patients [5] was established. Rating the effectiveness of a given inverse problem combining two connectivity metrics (directed and non-directed one) in locating epileptic zones was suggested. To explain cortical regions and neuronal generators of excessive discharges, we first applied four inverse problem approaches: MNE, sLORETA, dSPM, and MEM. Each technique's network connectivity using two types of connectivity metrics was computed. We investigated CC and GC for spiky activity to explore the causality effects on epileptic spiky events. Then, PLV & PTE for gamma oscillatory events were proposed to evaluate the phase synchronization effects on epileptic oscillatory networks. For both biomarker sLORETA depicts the highest number of nodes in common for directed and non directed functional connectivity; also it presents the closet nodes and the lowest delay propagation. The entire investigated inverse problem techniques and for directed and non directed functional connectivity were able to recognize parts of SOZ. CG enhances the strength of connectivity for spiky networks and PTE for oscillatory events.

This proves the effect of causality on networks' spiky topology. In this work, the distributed inverse techniques depict different topologies of networks. However, their results concerning nodes in the common delay of propagation and mutual between invasive and non invasive networks are quite close. Also, we noticed that directed and non directed metrics of connectivity did impact the complexity of networks for both biomarkers; but it doesn't change radically the studied networks. The techniques investigated could be considered as a prognosis tool for studying epileptic network connectivity.

In this study, four distributed inverse methods were investigated in the context of defining epileptic network connectivity tested on 2 types of biomarkers. A robust comparative analysis that enhances the robustness of each method in real clinical scenarios was proved. Moreover, advanced connectivity metrics directed and non-directed are explored to test the topology of network connectivity, then confronted with IEEG network connectivity. This dual approach gave a further level of analysis of epileptic network connectivity, addressing a gap in previous research which often focused on a single type of event, method, or metric. Our findings have significant implications for the pre-surgical evaluation of epilepsy patients, by demonstrating that all studied techniques successfully detect at least one part of the seizure onset zone (SOZ)

Hence to further examine and assess tools for the definition of exact cerebral generators responsible for excessive discharges and build-up of seizure using MEG signal, testing additional sets of patients as a future work is suggested. A second track that could be

also interesting is to compare outcomes of other distributed techniques, including ST-MAP (SpatioTemporal-Maximum A Posteriori), MCE (minimum current estimates), and Eloreta (exact low-resolution brain electromagnetic tomography) applied on a combination of several registration techniques.

5. REFERENCE

- [1] D. Cohen, "Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents", *Science*, Vol. 161, No. 3843, 1968, pp. 784-786.
- [2] O. Hauk, M. Stenroos, Matthias S. Treder, "Towards an objective evaluation of EEG/MEG source estimation methods: the linear approach", *NeuroImage*, Vol. 255, 2022, pp. 1053-1119.
- [3] R. Srinivasan, "Anatomical constraints on source models for high-resolution EEG and MEG derived from MR", *Technology in Cancer Research & Treatment*, Vol. 5, No. 4, 2006, pp. 389-399.
- [4] A. Hadriche, N. Jmail, J. L. Blanc, L. Pezard, "Using centrality measures to extract core pattern of brain dynamic during the resting state", *Computer Methods and Programs in Biomedicine*, Vol. 179, 2019, pp. 104-985.
- [5] N. Jmail, M. Gavaret, F. Bartolomei, P. Chauvel, J. M. Badier, C. G. Bénar, "Comparison of brain networks during interictal oscillations and spikes on magnetoencephalography and intracerebral EEG", *Brain Topography*, Vol. 29, No. 5, 2016, pp. 752-765.
- [6] A. Necibi, A. Hadriche, N. Jmail, "Assessment of Epileptic Gamma Oscillations' Networks Connectivity", *Proceedings of the International Conference on Intelligent Systems Design and Applications*, 12-14 December 2022, pp. 91-99.
- [7] R. Jarray, N. Jmail, A. Hadriche, T. Frikha, "A Comparison between modeling a normal and an epileptic state using the FHN and the epileptor model", *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 8th International Conference on Innovations in Bio-Inspired Computing and Applications*, Marrakech, Morocco, 11-13 December 2017, pp. 245-254.
- [8] C. G. Bénar, L. Chauvière, F. Bartolomei, F. Wendling, "Pitfalls of high-pass filtering for detecting epilep-

tic oscillations: a technical note on "false" ripples", *Clinical Neurophysiology*, Vol. 121, No. 3, 2010, pp. 301-310.

- [9] N. Jmail, R. Jarray, A. Hadrich, T. Frikha, C. Benar, "Separation between spikes and oscillation by stationary wavelet transform implemented on an embedded architecture", *Journal of the Neurological Sciences*, Vol. 381, 2017, p. 542.
- [10] F. Darvas, D. Pantazis, E. Kucukaltun-Yildirim, R. M. Leahy, "Mapping human brain function with MEG and EEG: methods and validation", *NeuroImage*, Vol. 23, 2004, pp. S289-S299.
- [11] F. Wendling, K. Ansari-Asl, F. Bartolomei, L. Senhadji, "From EEG signals to brain connectivity: a model-based evaluation of interdependence measures", *Journal of Neuroscience Methods*, Vol. 183, No. 1, 2009, pp. 9-18.
- [12] A. Palmi et al. "Intrinsic epileptogenicity of human dysplastic cortex as suggested by corticography and surgical results", *Annals of Neurology*, Vol. 37, No. 4, 1995, pp. 476-487.
- [13] M. Darbas, S. Lohrengel, B. Sulis, "Forward and inverse source problems for time-dependent electroencephalography", *Inverse Problems in Science and Engineering*, Vol. 4, 2022.
- [14] A. Hadriche, I. Behy, A. Necibi, A. Kachouri, C. B. Amar, N. Jmail, "Assessment of effective network connectivity among MEG none contaminated epileptic transitory events", *Computational and Mathematical Methods in Medicine*, Vol. 1, 2021, p. 6406362.
- [15] Y. Dai, W. Zhang, D. L. Dickens, B. He, "Source Connectivity Analysis from MEG and its Application to Epilepsy Source Localization", *Brain Topography*, Vol. 25, No. 2, 2012, pp. 157-166.
- [16] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods", *The Econometric Society*, Vol. 37, 1969, pp. 424-438.
- [17] F. Bartolomei, P. Chauvel, F. Wendling, "Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG", *Brain*, Vol. 131, No. 7, 2008, pp. 1818-1830.
- [18] U. Malinowska, J. M. Badier, M. Gavaret, F. Bartolomei, P. Chauvel, C. G. Bénar, "Interictal networks in magnetoencephalography", *Human Brain Mapping*, Vol. 35, No. 6, 2014, pp. 2789-2805.
- [19] C. G. Bénar, T. Papadopoulou, B. Torrèsani, M. Clerc, "Consensus matching pursuit for multi-trial EEG signals", *Journal of Neuroscience Methods*, Vol. 180, No. 1, 2009, pp. 161-170.
- [20] A. Bruns, R. Eckhorn, H. Jokeit, A. Ebner, "Amplitude envelope correlation detects coupling among incoherent brain signals", *Neuroreport*, Vol. 11, 2000, pp. 1509-1514.
- [21] M. S. Roulston, L. A. Smith, "Evaluating Probabilistic Forecasts Using Information Theory", *Monthly Weather Review*, Vol. 130, 2002, pp. 1653-1660.
- [22] R. D. Pascual-Marqui, M. Esslen, K. Kochi, D. Lehmann, "Functional imaging with low-resolution brain electromagnetic tomography (LORETA): a review", *Methods and Findings in Experimental and Clinical Pharmacology*, Vol. 24, 2002, pp. 91-95.
- [23] L. Marzetti, A. Basti, F. Chella, A. D'Andrea, J. Syrjala, V. Pizzella, "Brain Functional Connectivity Through Phase Coupling of Neuronal Oscillations: A Perspective From Magnetoencephalography", *Frontiers in Neuroscience*, Vol. 13, 2019.
- [24] M. Lobier, F. Siebenhühner, S. Palva, J. Matias Palva, "Phase transfer entropy: A novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions", *NeuroImage*, Vol. 35, 2014, pp. 853-872.
- [25] A. M. Dale et al. "Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity", *Neuron*, Vol. 26, No. 1, 2000, pp. 55-67.
- [26] J. Philippe, Lachaux, Eugenio, J. Martinerie, Francisco J. Varela, "Measuring phase synchrony in brain signals", *Human Brain Mapping*, Vol. 8, 1999, pp. 194-208.
- [27] N. Jmail et al. "A comparison of methods for separation of transient and oscillatory signals in EEG", *Journal of Neuroscience Methods*, Vol. 199, No. 2, 2011, pp. 273-289.
- [28] N. Jmail, M. Gavaret, F. Bartolomei, C.-G. Benar, "Despikifying SEEG signals using a temporal basis

- set", Proceedings of the 15th International Conference on Intelligent Systems Design and Applications, Marrakech, Morocco, 14-16 December 2015, pp. 580-584.
- [29] A. Delorme, S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis", *Journal of Neuroscience Methods*, Vol. 134, No. 1, 2004, pp. 9-21.
- [30] J. F. Geweke, "Measures of Conditional Linear Dependence and Feedback between Time Series", *Journal of the American Statistical Association*, Vol. 79, 1984, p. 388.
- [31] A. Hadriche, I. ElBehy, A. Hajje, N. Jmail, "Evaluation of techniques for predicting a build up of seizure", Proceedings of the International Conference on Intelligent Systems Design and Applications, 13-15 December 2021, pp. 816-827.
- [32] S. L. Bressler, A. K. Seth, "Wiener-Granger Causality: A well established methodology", *NeuroImage*, Vol. 58, 2011, pp. 323-329.
- [33] M. S. Hämäläinen, R. J. Ilmoniemi, "Interpreting magnetic fields of the brain: minimum norm estimates", *Medical & Biological Engineering & Computing*, Vol. 32, No. 1, 1994, pp. 35-42.
- [34] L. Kossler, T. Cecchin, O. Casparay, A. Benhadid, "EEG-MRI Coregistration and Sensor Labeling Using a 3D Laser Scanner", *Annals of Biomedical Engineering*, Vol. 39, No. 3, 2011, pp. 983-995.
- [35] D. van't Ent et al. "Spike cluster analysis in neocortical localization related epilepsy yields clinically significant equivalent source localization results in magnetoencephalogram (MEG)", *Clinical Neurophysiology*, Vol. 114, No. 10, 2003, pp. 1948-1962.
- [36] J. C. Mosher, R. M. Leahy, "Recursive MUSIC A framework for EEG and MEG source localization", *IEEE Transactions on Biomedical Engineering*, Vol. 45, 1998, pp. 1342-1354.
- [37] G. Pellegrino et al. "Accuracy and spatial properties of distributed magnetic source imaging techniques in the investigation of focal epilepsy patients", *Human Brain Mapping*, Vol. 41, No. 11, 2020, pp. 3019-3033.
- [38] M. Florian, K. Lehnertz, D. Peter, E. Christian, "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients *Physica D: Nonlinear Phenomena*", *East European Journal of Psycholinguistics*, Vol. 144, 2000, pp. 358-369.
- [39] M. Caparos, V. Louis-Dorr, F. Wendling, L. Maillard, D. Wolf, "Automatic lateralization of temporal lobe epilepsy based on scalp EEG", *Clinical Neurophysiology*, Vol. 117, No. 11, 2006, pp. 2414-2423.
- [40] C. Grova, J. Daunizea, J. M. Lina, C. G. Bénar, H. Bernali, J. B. Gotman, "Evaluation of EEG localization methods using realistic simulations of interictal spikes", *NeuroImage*, Vol. 29, No. 3, 2016, pp. 734-753.
- [41] P. Tass, M. G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkman, A. Schnitzler, H.-J. Freund, "Detection of n:m Phase Locking from Noisy Data: Application to Magnetoencephalography", *Physical Review Letters*, Vol. 81, 1998, p. 3291.

An Enhancement of Grid Integration in Renewable Energy Systems Using Multi-Objective Multi-Verse Optimization

Original Scientific Paper

Mullan Abdul Nabi*

Electrical & Electronics Engineering
JNTUA Anantapur, Anantapur, India
G Pullaiah College Of Engineering & Technology
Kurnool, India
abdulnabieeee@gpcet.ac.in

J. Surya Kumari

Electrical & Electronics Engineering
RGM College of Engineering and Technology
Nandyal, India
jdsk.23@gmail.com

*Corresponding author

Abstract – Going by the recent trends, the application of Renewable Energy Sources (RESs) has grown significantly across the world. Still, the integration of grid with photovoltaic (PV), wind and battery, remains a critical challenge as it results into power quality issues. To address the increasing need for electricity caused by industrialization and growing population, hybrid PV, wind, and battery combinations are used in this study. To accomplish an optimal energy management, this research proposes the multi-objective multi-verse optimization (MOMVO) approach, along with the modified perturb and observe (MP&O) technique. The proposed MOMVO-MP&O controller operates between the wind turbine and the battery storage system, for providing optimal power distribution and stability. The suggested model is evaluated alongside three other popular controller combinations that are, multi-verse optimization-perturb and observe (MVO-P&O), MVO-MP&O, and MOMVO-P&O. A comparative analysis is conducted, with existing methods namely, Modified-Fuzzy Direct Power Control (MF-DPC) and Adaptive Neuro-Fuzzy Inference System (ANFIS) also. From the analyzed results of this comparison, the proposed MOMVO-MP&O achieves lesser Total Harmonic Distortion (THD) of 1.86%, which demonstrates its efficiency in addressing power quality issues using hybrid RES systems.

Keywords: energy management strategy, photo voltaic, multi-objective multi-variable optimization, renewable energy sources, total harmonic distortion

Received: December 12, 2023; Received in revised form: July 17, 2024; Accepted: July 19, 2024

1. INTRODUCTION

The global search for sustainable and efficient energy solutions has been growing in the recent years, with a major focus on integrating renewable energy sources into different parts of the energy infrastructure. This innovative technology is focused on PV systems, wind turbines, and an enhanced battery storage [1, 2]. When combined effectively, these renewable energy sources have the revolutionary potential to significantly improve tiny grid systems. Grids as localized and decentralized energy distribution networks are an example of modern energy innovation allowing communities, institutions, and businesses achieve higher resilience

and energy stability [3, 4]. PV and wind energy transmission is constrained by location and intermittency. An indeterminate source is the consequence of both sources irregular power output, which is reliant on weather conditions. Moreover, RES are frequently situated distant from cities, necessitating energy-lossy transmission across long distances. Due to wind energy and solar are irregular and vulnerable for environmental fluctuations. Thus, additional intrusion into the power systems can produce significant technological issues especially in weak grids. Moreover, difficulties involving infrastructure regulations restrict the development of effective transmission. The renewable energy's potential offers better storage facilities, grid management, as

well as policy assistance to deal with these problems [5, 6]. The progress and availability of PV technology have propelled, leading the field of renewable energy sources. Solar energy, particularly in areas where sunshine is common are easily integrated into grids to provide a continuous and sustainable power source. Because solar PV installations are scattered, they enable localized generation, thereby reducing the pressure on centralized power networks and enhancing the energy self-sufficiency within grids [7, 8]. Wind turbines, another major component of the renewable energy environment, generate electricity by harnessing wind kinetic energy. Wind energy is reliable and abundant, making it an ideal resource for powering microgrids [9,10]. Wind turbines are strategically placed within or near grid areas to make use of natural wind patterns, and thus contribute to diversifying the mix of energy sources [11]. Wind's nature is reduced by combining it with other renewable energy sources and deploying energy storage devices, resulting in a continuous and stable power supply that is able to satisfy the grid's needs. In the field of energy storage, advanced battery technologies play a crucial role for optimizing renewable energy integration into the grid systems. The battery serves an important role by storing excess energy generated from PV and wind sources throughout peak production for later use during low power periods [12, 13]. Energy storage systems improve grid stability, regulate load variations, and help to properly balance supply and demand. Batteries enable microgrids to run independently by storing excess renewable energy, increasing resilience and lowering dependency on external power sources [14]. When the renewable energy sources: PV, wind, and battery storage are coupled in a micro-grid system, a synergistic energy infrastructure is created. Solar PV arrays and wind turbines generate electricity which is then stored in batteries or distributed directly to the microgrid. Their ability to seamlessly integrate different sources assures a consistent and stable energy supply, addressing the microgrid's specific energy requirements while minimizing their reliance on traditional fossil fuel-based generators [15, 16]. The incorporation of renewable energy sources into micro-grids represents an evolution in the approach to energy generation, distribution, and consumption [17]. It is consistent with global efforts to decrease carbon emissions, mitigate climate change, and accomplish long-term development ideas. Recent technology advancements utilize micro-grids, offering resilient energy solutions for global communities, and promoting sustainability on a broad scale [18]. Challenges in integrating PV, battery, and wind for grid improvement include intermittent energy supply, storage optimization, output variability, and grid stability.

Sahri et al. [19] implemented MF-DPC to improve system performance and current quality. An enhanced Maximum Power Point Tracking (MPPT) algorithm was also developed using a Fuzzy Controller (FC) for optimizing solar power. This method demonstrated nota-

ble performance improvements, especially in adverse weather and variable load conditions. However, NF-DPC depended on pre-trained models, making it challenging to adapt to real time to dynamic grid changes, hindering its responsiveness and grid stability during rapid fluctuations. Gulzar et al. [20] introduced a Battery Energy Storage System (BESS) to manage modeling, control, energy management, and operations in a hybrid grid-connected setup. A hybrid system combining PV, wind, and fuel cell (PV-Wind-FC) with an electrolyzer, and BESS was proposed to efficiently minimize the control loops and converters. This design eliminated the need for a dedicated PV converter, thus enhancing cost-effectiveness and preventing BESS overcharging. Nonetheless, the BESS method needed to enhance the system by integrating nonlinear controllers in wind, fuel cell, and the electrolyzer components. This is because nonlinear controllers improved the BESS by addressing complex dynamics and optimizing renewable energy integration and performance. Maaruf et al. [21] developed Hybrid RES (HRES) to improve reliability and efficiency using renewable technologies. A robust control strategy validated across diverse HRES conditions, affirming its effectiveness and adaptability was established. Nonetheless, to accomplish system optimization, it needed to integrate hybrid energy storage systems for supplementary support so as to upgrade the overall performance in HRES. This is because the integrating hybrid energy storage enhanced the system's stability, reliability, and efficiency by combining diverse storage technologies to improve performance in HRES. Chalamuthu et al. [22] created a grid-integrated renewable energy system featuring a hybrid series active power filter system controlled by an ANFIS. This setup was designed to cater to nonlinear or sensitive loads. The strategy involved sharing renewable energy with the grid, minimizing reliance on conventional electricity, and showcasing the potential for reduced grid consumption. However, additional information about the ANFIS approach was required to provide a comprehensive explanation for ensuring an in-depth knowledge of its operation, capabilities, and prospective applications. Ibáñez-Rioja et al. [23] introduced a Levelized Cost of Hydrogen (LCOH) calculation considering the capital and operating expenses, along with the learning curves for system components. This approach aimed at enhancing the electrolyzer full-load hours and minimizing electricity wastage in off-grid plants. Nonetheless, further improvement was required in the suggested method for future generation through amplifying efficiency and sustainability. This was because it was possible to enhance sustainability and efficiency, so as to enable surplus electricity sales that ensured system optimization and revenue generation in LCOH.

Hence, MOMVO-MP&O is used to address problems with the existing methods which include inadequate renewable energy integration, grid instability, and inefficient power management, ultimately enhancing the grid performance by optimizing resource utilization

and control strategies. The proposed MOMVO counters these challenges through optimal resource allocation and management, thereby enhancing grid performance and reliability. The contributions of the research are as follows;

- A combination of MOMVO and MP&O controllers is designed to develop an effective EMS to satisfy the load demand. The MOMVO is preferred due to its higher convergence rate to effectively switch between wind turbine and battery.
- The MOMVO-MP&O is used in the RES system to activate DC-DC converters, which results consistent power supply at different temperatures and irradiance levels
- MOMVO-MP&O enhances the grid optimization by balancing multiple objectives simultaneously, then optimizing efficiency, reliability, and sustainability for more resilient energy management.

The rest of the paper is organized as follows: Section 2 describes the proposed methodology of this research. Section 3 and Section 4 describe the results and conclusion of this overall research.

2. COMPONENT MODELING OF HYBRID ENERGY SYSTEMS

This research employs three distinct energy systems: PV modules, batteries, and wind turbines, each characterized by mathematical models as follows.

2.1. PHOTOVOLTAIC MODULES

Photovoltaic modules are used in the grid to generate sustainable energy, while the PV modules in MOMVO-MP&O contribute to grid power using MPPT control, hence amplifying grid sustainability. The system voltage and capacity are determined by connecting the PV panel either in parallel or in series connection [24]. The MPPT method is considered in the PV system to increase PV output power. The following equation (1) is used to identify the output power that depends on the I_M and V_M .

$$P_{MPPT}(t) = I_{MPPT}(t) \times V_{MPPT}(t) \quad (1)$$

Eqs. (2) and (3) provide the MPPT current and voltage, respectively.

$$I_{MPPT}(t) = I_{SC} \left\{ 1 - C_1 \left[\exp \left(\frac{V_M}{C_2 \times V_{OC}} \right) \right] \right\} + \Delta I(t) \quad (2)$$

$$V_{MPPT}(t) = V_M + \mu V_{OC} \cdot \Delta T(t) \quad (3)$$

In this context, I_{SC} signifies the short circuit current, C_1 and C_2 represent capacitance, V_M stands for the maximum voltage and V_{OC} denotes the open circuit voltage. The C_1 , C_2 , $\Delta I(t)$ and $\Delta T(t)$ of Eqs. (2) and (3) are respectively expressed in the Eqs. (4-7).

$$C_1 = \left(1 - \frac{I_M}{I_{SC}} \right) \times \exp \left(-\frac{V_M}{C_2 \times V_{OC}} \right) \quad (4)$$

$$C_2 = \left(\frac{V_M}{V_{OC}} - 1 \right) \times \left[\ln \left(1 - \frac{I_M}{I_{SC}} \right) \right]^{-1} \quad (5)$$

$$\Delta I(t) = I_{SC} \left(\frac{GT(t)}{G_{ref}} - 1 \right) + \alpha_{1,sc} \times \Delta T(t) \quad (6)$$

$$\Delta T(t) = T_c(t) - T_{c,ref} \quad (7)$$

Where, I_M signifies the maximum current, $GT(t)$ denotes the radiation of the incident on the PV surface, T_c represents the cell temperature, and $T_{c,ref}$ denotes the PV temperature in normal situations.

2.2. WIND TURBINE MODEL

When PV is idle, wind turbines provide electricity, assuring continuous power output for system stability. Wind turbines in MOMVO-MP&O maximize their ability to produce electricity through system controls, thus improving the grid sustainability and performance. The output power of a wind turbine is essentially determined by elements such as rated capacity, hub height, and specific wind speed characteristics, as shown in reference [24]. For variable speed operation in this research, a wind turbine model based on Doubly Fed Induction Generator (DFIG) is used. These wind speed characteristics contain the cut-off voltage (V_{cf}), rated speed (V_{rs}) and cut-in voltage (V_{ci}). The output of the wind turbine power is calculated using Eq. (8).

$$P_o = P_{rs} \begin{cases} 0, & V < V_{ci} \\ a \times V^3 - b \times P_{rs}, & V_{ci} < V < V_{rs} \\ 1, & V_{rs} < V < V_{cf} \\ 0, & V > V_{cf} \end{cases} \quad (8)$$

Here, rated power and wind turbine output are correspondingly denoted by P_o and P_{rs} , where $(a=P_{rs}) / (V_{rs}^3 - V_{ci}^3)$ and $b=(V_{ci}^3)/(V_{rs}^3 - V_{ci}^3)$. At a specific height level using the supplied Eq. (9), the speeds are analyzed and then transformed based on the actual turbine speed.

$$V_{hub} = V_0 \times \left(\frac{Z_{hub}}{Z_0} \right) \times \alpha \quad (9)$$

In this context, V_{hub} represents the wind turbine speed, Z_{hub} stands for the actual height of the wind turbine and V_0 represents the wind speed at the reference height. Additionally, α represents the power law exponent utilized in the evaluation and Z_0 denotes the height of the wind speed measurement.

2.3. BATTERY MODEL

Batteries supply power when wind turbines are inactive, ensuring consistent energy and maintaining grid stability during fluctuations. Batteries store surplus energy, guaranteeing grid stability and reliability. In MOMVO-MP&O, batteries are managed for peak load support and power quality optimization. RES exhibits variable output power owned to fluctuating environmental conditions. Eqs. (10) and (11) detail the charging and discharging processes of the battery, with a focus on State of Charge (SOC) conditions. Specifically, battery charging occurs when the combined power

output from PV and wind turbines surpasses the load requirements. Equally, discharge operations are initiated when the load demand exceeds the available power from renewable energy sources, as referenced in [25].

$$SOC = SOC(t-1) \times (1 - \sigma) + \left[\frac{P_{RES}(t) - P_L(t)}{\eta_{inv}} \right] \times \eta_{ch} \quad (10)$$

$$SOC = SOC(t-1) \times (1 - \sigma) + \left[\frac{P_L(t)}{\eta_{inv}} - P_{RES}(t) \right] \times \eta_{disch} \quad (11)$$

Where, the battery's charging and discharging states are correspondingly denoted by SOC and $SOC(t-1)$, while the discharge rate over one hour is denoted as σ . The power from renewable energy sources and load power are denoted by P_{RES} and P_L , respectively. Additionally, the efficiency parameters include η_{inv} for the inverter, η_{ch} for charging, and η_{disch} for discharging. Hence, the proposed MOMVO-MP&O system receives electrical output from PV modules, wind turbines, and batteries. It optimizes their utilization, guaranteeing efficient energy generation, storage and distribution for enhanced grid performance and sustainability, as described briefly in the suggested system.

2.4. PROPOSED SYSTEM

The electrical output from PV modules, wind turbines, and batteries are fed into the suggested system

to enhance the grid performance. The system optimizes energy utilization and distribution for an improved grid efficiency. This research develops an effective energy management method that makes use of two unique RES and a storage device. PV modules and wind turbines are the primary RES units evaluated for EMS. In addition, it includes a battery as a storage medium for excess electricity generated by both the PV modules and the wind turbines. The PV modules serve as the primary energy source, while the wind turbines serve as a supplementary supply, further allowing to efficiently balance load demand. When the PV modules and wind turbines are unable to generate enough energy to satisfy the load requirement, the battery serves as a backup power source. It uses the Perturb and Observe with Modified Incremental Conductance MP&O algorithm to optimize the performance of the PV modules. This algorithm activates the DC-DC converter linked with the PV modules, establishing peak power even when the irradiance and temperature are changing. Furthermore, MOMVO system is employed to control the battery and wind turbine's ON and OFF states. This improves the overall system control and administration. Fig. 1 displays the flow diagram and operation of a grid-connected RES that is integrated with the MOMVO-MP&O controller, providing clarity and comprehension of how the system works.

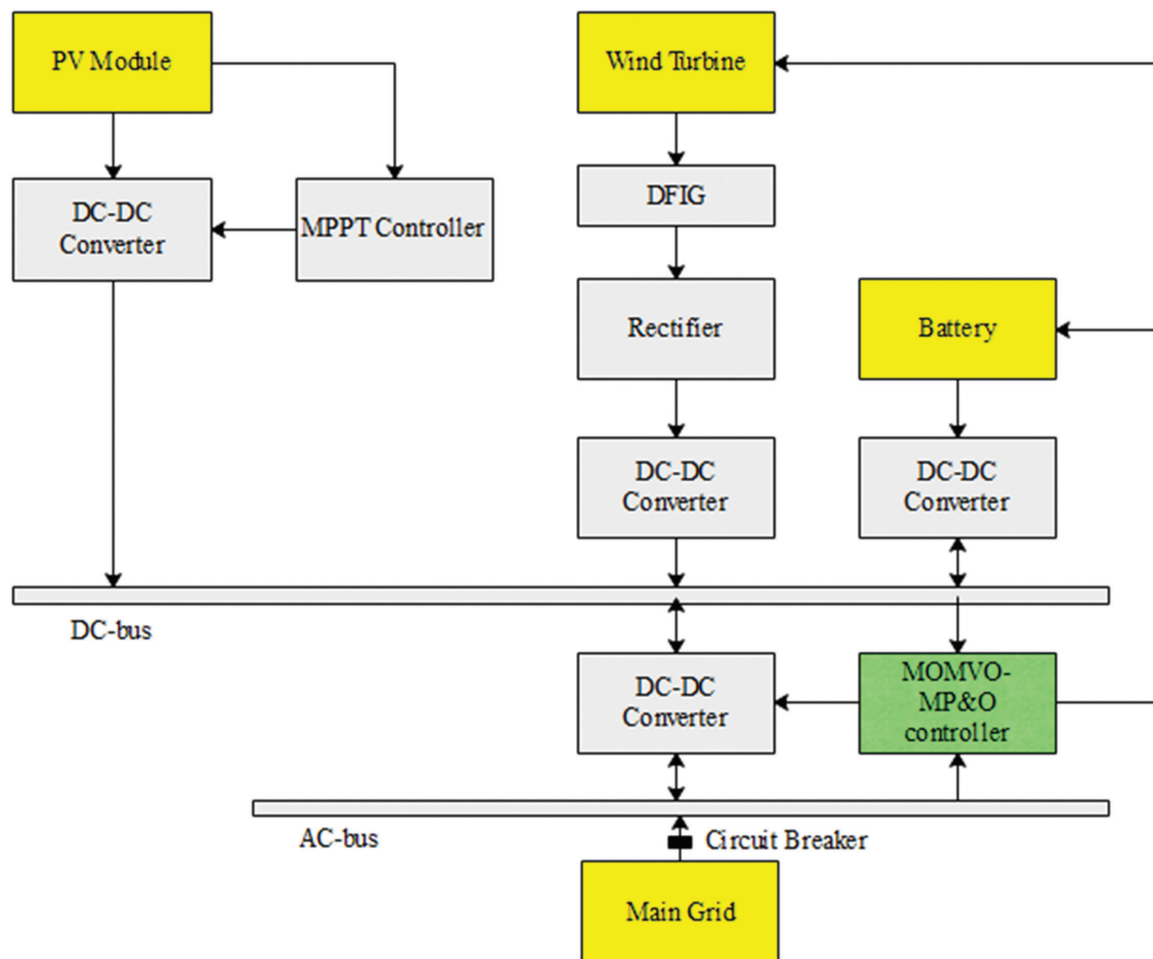


Fig. 1. Flow diagram of RES with the proposed MOMVO-MP&O controller

2.4.1. Multi Verse Optimization (MVO)

In this research, an innovative approach is introduced to improve the grid performance through the utilization of the MOMVO. MVO is used to improve grid performance using PV solar panels, wind turbines, and energy storage technologies such as batteries. The fundamental goal is to create a more efficient, stable, and long-lasting grid infrastructure. The problem with these diverse energy sources is optimizing their complicated interactions. During optimization, the following Eqs. (12-13) are applied to the universes of MVO.

$$U = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^d \\ x_2^1 & x_2^2 & \dots & x_2^d \\ \dots & \dots & \dots & \dots \\ x_n^1 & x_n^2 & \dots & x_n^d \end{bmatrix} \quad (12)$$

Where, n is the number of universes (candidate solutions) and d is the number of parameters (variables).

$$x_i^j = \begin{cases} x_k^j & r^1 < N1(U_i) \\ x_i^j & r^1 < N1(U_i) \end{cases} \quad (13)$$

The j_{th} universe $N1(U_i)$ is normalized inflation rate is connected with parameters, which are represented by x_i^j . Furthermore, k_{th} universe's parameters, x_k^j , are selected through a roulette wheel method that uses a random integer r^1 . Assume that wormhole tunnels continuously connect each universe with the best universe developed at this point to boost local alterations and raise the possibility of improving inflation rates. The mechanism is organized in the Equation (14).

$$x_i^j = \begin{cases} X_j + TDR \times ((ub_j - lb_j) \times r4 + lb_j) & r3 < 0.5 \quad r2 < WEP \\ X_j - TDR \times ((ub_j - lb_j) \times r4 + lb_j) & r3 \geq 0.5 \quad r2 \geq WEP \end{cases} \quad (14)$$

The mechanism is performed by the best universe j_{th} parameter, represented by X_j . The j_{th} variable's lower and upper limits are represented by lb_j and ub_j , respectively, whereas TDR and WEP are used as coefficients. Additional random numbers that contribute to the process are $r2$, $r3$, and $r4$, all of which fall within the interval $[0, 1]$. Furthermore, WEP, TDR is gradually increased between iterations to provide a more precise local search around the optimal universe. Eqs. (15–16) illustrates the formula for updating both coefficients.

$$WEP = min + l \times \left(\frac{max - min}{L} \right) \quad (15)$$

Where, min is the minimum max , l indicates the current iteration, and L denotes the maximum iterations.

$$TDR = 1 - \frac{1}{L^p} \quad (16)$$

The roulette wheel selection that is carried out for every factor in each universe all throughout iterations have a computational cost of either $O(n)$ or $O(\log n)$, depending upon the way it is implemented. The Eqs. (17-18) mathematically demonstrate the entire computational load.

$$O(MVO) = O(l(o(Quick\ sort) + n \times d \times (o(roulette\ wheel)))) \quad (17)$$

$$O(MVO) = O(l(n^2 + n \times d \times \log_n)) \quad (18)$$

Where, n is the number of universes, l is the maximum number of iterations, and d is the number of objects. Hence, MVO faces challenges in accurately modeling complex, variable renewable energy inputs, alongside dealing with infrastructure constraints, high computational demands, sensitivity to input parameters, and real-time adaptability. Optimizing the integration and accommodating evolving technologies within its framework remains essential for effective grid improvement using PV, wind, and battery systems.

2.4.2. Multi-Objective Multi-Versé Optimization (MOMVO)

Integrating PV, wind, and battery systems presents challenges in optimizing the grid performance using MVO. Single-objective MVO struggles to balance the competing objectives like maximizing renewable energy integration while maintaining grid stability. MOMVO addresses this by considering multiple goals, offering a more comprehensive approach. MOMVO optimizes various objectives such as maximizing renewable energy utilization and minimizing grid fluctuations, navigating the complex decisions inherent in energy grid optimization. It identifies a range of Pareto-optimal solutions by simultaneously exploring diverse solution spaces, enabling decision-making based on preferences and priorities. MOMVO allows stakeholders to select from a variety of different solutions, helping grid operators and algorithms in achieving a balance between renewable energy integration and grid stability, ultimately contributing to a better sustainable energy grid.

2.4.2.1. Cost of Electricity

It is characterized as the unit of cost per unit of delivered energy produced by hybrid micro-grid, as illustrated in the Equation (19).

$$COE = \frac{Total\ Net\ Present\ Cost\ (NPC)}{\sum_{h=1}^{h=8760} p_1(h)} \quad (19)$$

Where, $\sum_{h=1}^{h=8760} p_1(h)$ represents the summation of the annual energy output over the entire year.

2.4.2.2. Loss of Power Supply Possibility

Minimizing the LPSP which results from inadequate energy to provide the load is crucial to boosting system dependability. LPSP is used to minimize disconnection probabilities, focusing on resilience and autonomy in localized systems, reflecting a strategic shift towards robust, islanded configurations in goal-oriented frameworks. Equation (20) represents the mathematical expression of LPSP.

$$LPSP = \frac{\sum P_L(t) - (P_W(t) + P_{PV}(t) + (E_b(t-1) - E_{bmin}) + P_{diesel})}{\sum P_L(t)} \quad (20)$$

Where, $\sum P_L(t)$ represents the summation of the total power demand, $P_W(t)$ represents power generated by wind turbines at time t . Then, E_{bmin} represents the Minimum allowable energy level in the batteries. Finally, P_{diesel} represents the Power generated by diesel generators.

2.4.2.3. Renewable Factor (RF)

The RF evaluates the output rate of traditional diesel generation in the renewable energy source. Equation (21) demonstrates how the RF is expressed mathematically.

$$RF = \left(1 - \frac{\sum P_{diesel}}{\sum P_{pv} + P_W}\right) \times 100 \quad (21)$$

Ultimately, MOMVO enhances grid performance through sustainable, efficient, and balanced renewable energy integration. MOMVO's improvement in grid efficiency with PV, wind, and solar lies in generating a wide array of solutions representing potential differences between objectives. It explores the problem space, addressing goals like increasing renewable energy generation and reducing grid instability. The process involves evolving a population of solutions through different "verses," similar to distinct solution spaces, utilizing evolutionary algorithms or simulated annealing. Iterations refine outcomes, leading to a collection of Pareto-optimal solutions, where enhancing one objective may compromise another. MOMVO empowers stakeholders to choose from a variety of different solutions, helping grid operators and strategists achieve a balance between renewable energy integration and grid stability, ultimately contributing to a better, sustainable energy grid.

The two key improvements made in MOMVO to enhance grids are:

- **Optimized Renewable Energy Integration**

MOMVO improves grid performance by optimizing the integration of renewable energy sources like PV, wind, and solar. It focuses on maximizing the utilization of these sources while considering various objectives such as energy generation and grid stability. Through multi-objective optimization, MOMVO finds the best compromise solutions that efficiently integrate renewable energy into the grid, guaranteeing a sustainable, low-carbon energy mix.

- **Enhanced Grid Stability and Reliability**

MOMVO demonstrates further upgradation by effectively addressing grid stability and reliability concerns, enhancing the overall performance in power distribution systems. It optimizes the allocation and management of renewable energy resources to minimize grid fluctuations and enhance stability. By considering multiple conflicting objectives including voltage regulation, frequency control, and power quality, the MOMVO ensures a more stable and reliable grid. This results in a resilient energy infrastructure capable of accommodating intermittent renewable energy sources and meet-

ing varying demand patterns while maintaining grid integrity. In addition, the MOMVO is deployed to solve the MPPT problem which is the process of solving MPPT using MOMVO, as explained in the following section.

2.4.2. MOMVO control of PV, battery, and wind turbine

In the research, MOMVO controls the ON/OFF states of the wind turbine and battery by using power from the PV module, wind turbine, battery, and load. According to MOMVO duty cycle (modulation index), the wind turbine and battery's ON/OFF states are determined. Until the system uses MOMVO in MPPT to reach MPP, the duty cycle is updated continually. Controlling the power generation and utilization of PV modules, wind turbines, and batteries in an energy system involves smart coordination to ensure efficient and continuous power supply. Here is a brief explanation of the control strategy:

- **PV Module Control**

When the solar irradiance is insufficient or during nighttime, the PV modules are turned OFF to conserve energy. This prevents the system from attempting to generate power when sunlight is unavailable. When solar irradiance is significant, the PV modules are activated to connect solar energy and convert it into electricity.

- **Wind Turbine Control**

Wind turbines are turned ON when wind speeds are within the optimal range for power generation. They capture wind energy and convert it into electrical power. Conversely, during high wind speeds or maintenance, the wind turbines are turned OFF to prevent potential damage and establish safe operation.

- **Battery Control**

Batteries play a crucial role in storing excess energy generated by PV modules and wind turbines. When both PV modules and wind turbines are generating extra electricity, the excess power is stored in the battery for later use. The stored energy in the battery is utilized during periods of low renewable energy resources, ensuring a stable power supply by meeting load requirements effectively. This control strategy aims to optimize the utilization of renewable energy sources while maintaining a reliable power supply. It ensures that power generation order with the availability of solar irradiance and wind, and excess energy is efficiently stored and utilized, amplifying the overall efficiency and sustainability of the energy system.

2.5. P&O MODIFICATION FOR TRIGGERING PV MODULE

The results generated by the MOMVO are fed to the MP&O method for further processing and control. The modified P&O method is used in PV/Wind/Battery systems, addressing drawbacks of the P&O approach for enhanced performance. Unlike the conventional methods

prone to nonstop oscillations around MPP and deviation, the modified version utilizes both PV module voltage and current to accurately determine power and trigger the DC-DC converter switches, strengthening MPPT efficiency. The conventional P&O method considers only change in voltage (ΔV) and change in power (Δp), but the modified P&O method considers change in current (ΔI) as a third parameter to enhance the performance of MPPT. Through the modified P&O method, a total of 8 operating point perturbation cases are identified. Four cases mirror the conventional P&O method while the remaining four are tailored to sudden irradiation level changes. Detection of opposite signs in ΔI and ΔV signifies a PV module in fixed illumination, contrasting with fluctuating conditions. The modified P&O behaves like the conventional method under consistent solar irradiation.

In cases of poor tracking within the PV/Wind/Battery system, the tracking speed is augmented by doubling the step size, ensuring a prompt response to dynamic conditions. This modified P&O effectively controls the PV module when there is a change in power due to the perturbation of reference voltage and variation in sunlight. Fig. 2 depicts the controller architecture used in this PV/wind/battery system.

The effectiveness and resilience of the suggested method become apparent when observing its adept handling of power generation amidst fluctuating loads. The EMS control model ensures optimal power generation by managing system operations in response to varying loads. Furthermore, it distinctly showcases the system's efficiency in charging and discharging processes.

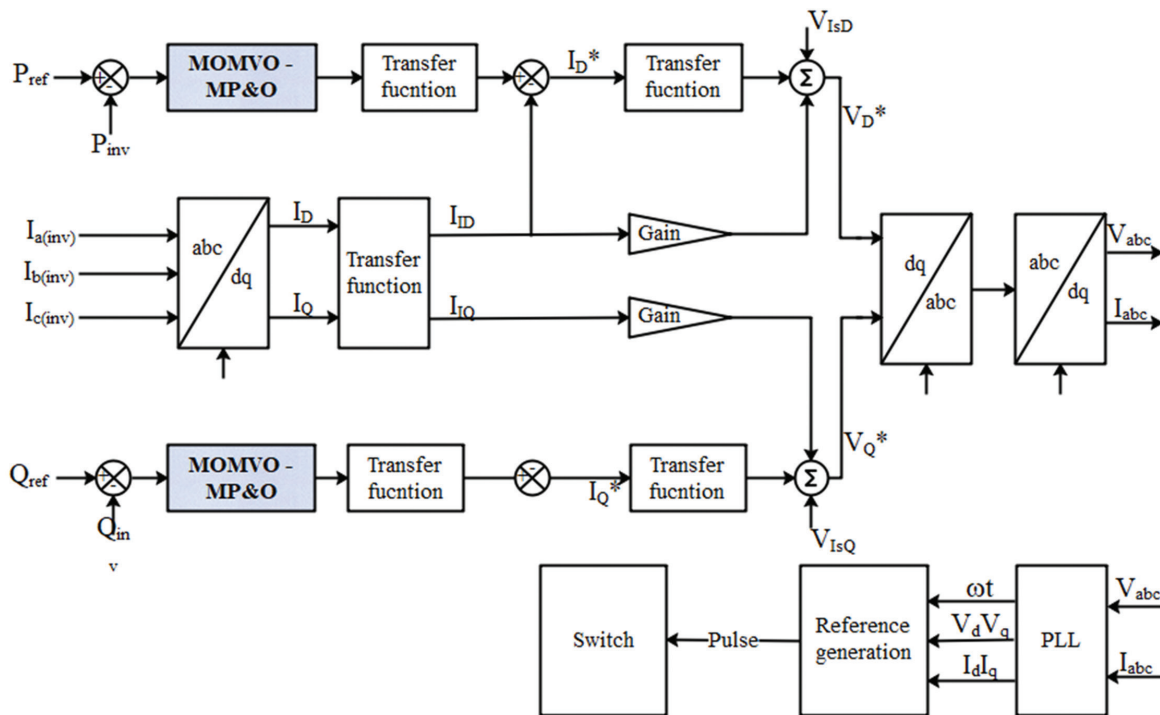


Fig. 2. Architecture of MOMVA-MP&O controller design

3. EXPERIMENTAL SETUP AND RESULTS

Many experiments are conducted to validate the effectiveness of the proposed Multiple-Objective Multi-Variable Optimization (MOMVO) for the 3S configuration, encompassing the patterns of 1, 2, and 3. This evaluation is performed under dynamic insolation levels as depicted in Tables 1, 2 and 3. Hence, the existing methods such as MultiVerse Optimization Perturb Observe (MVO-P&O), Multi-Verse Optimization – Modified Perturb Observe (MVO-MP&O) and Multi-Objective Multi-Verse Optimization – Perturb Observe (MVO-MP&O) are analyzed based on the performance analysis of power rating, extracted power, time tracking, number of iterations and maximum efficiency. Additionally, the specifications of PV, wind and battery systems are illustrated in Table 1. The simulation diagram of the proposed method is illustrated in Fig. 3.

Table 1. Specifications of PV, Wind and solar

Power systems	Specifications	Values
PV	STC power rating	150 W
	Open circuit voltage (V_{oc})	23.3V
	Minimum power circuit (I_{mp})	7.87 A
	Minimum power Voltage (V_{mp})	17.8V
Wind systems	Type	Espanda
	Cut off wind speed	3 m/s
	Cut in wind speed	25 m/s
Battery	No. of blader	2
	Type	GFM-100
	Rated voltage	2V
	Rated capacity	100Ah
	Hourly self-discharge rate	0.0001

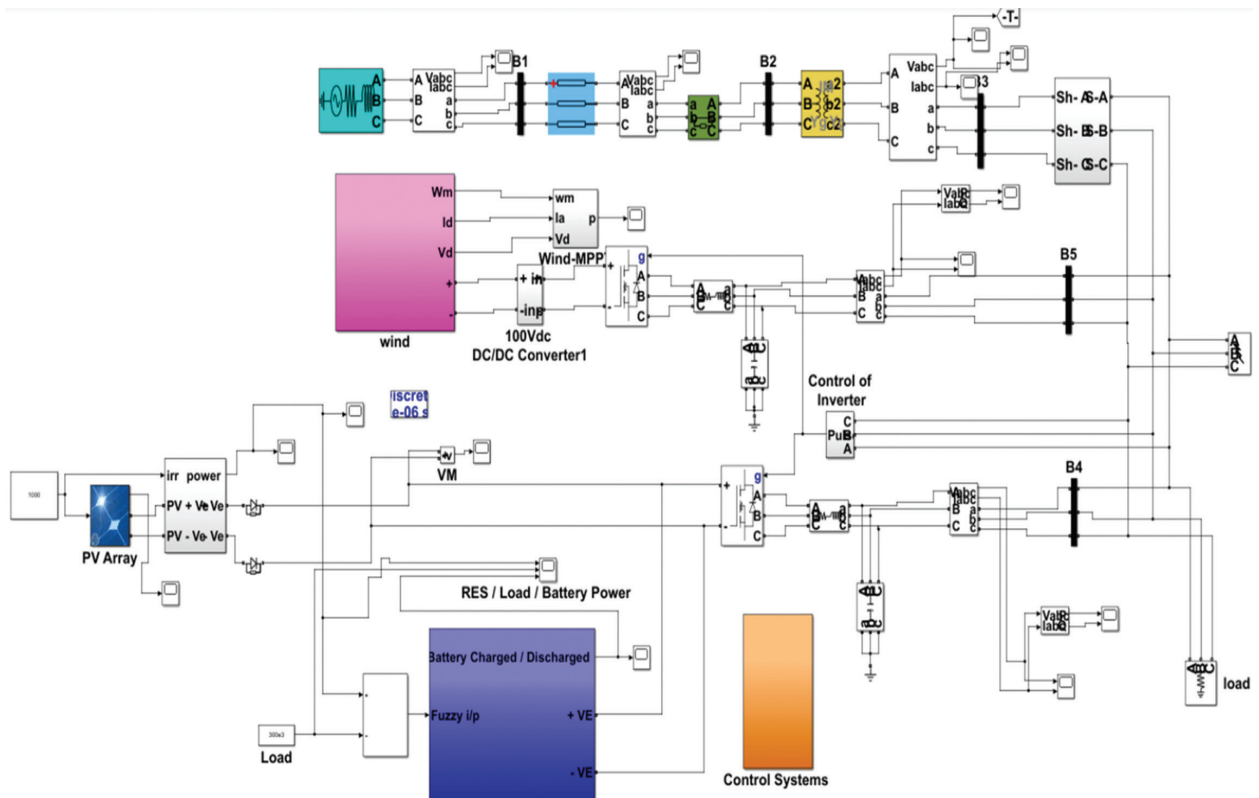


Fig. 3. Simulation diagram of the proposed method

3.1. PATTERN- 1

In this research, a 250 W PV array is subjected to varying irradiance levels in a partial shading scenario, while maintaining a relatively consistent temperature range of $T = 25$ to 25.5 °C. At $t = 0$ s, the MPPT algorithms are sequentially initiated, each corresponding to irradiance levels of $G = 1000$, 300 , and 600 W/m² on the respective PV. The MPPT algorithms' experimental waveforms are shown in the figure and tables below. The outcomes proved that the suggested method achieves superior performances in terms of power rating, extracted power, time tracking, number of iterations and maximum efficiency. The existing methods such as MVO-P&O, MVO-MP&O, and MOMVO-P&O respectively achieve 91.56%, 93.41%, and 78.74%, in terms of maximum efficiency, whereas the suggested MOMVO-MP&O method achieves a superior maximum efficiency of 98.51%, as illustrated in Table 2 and Fig. 4.

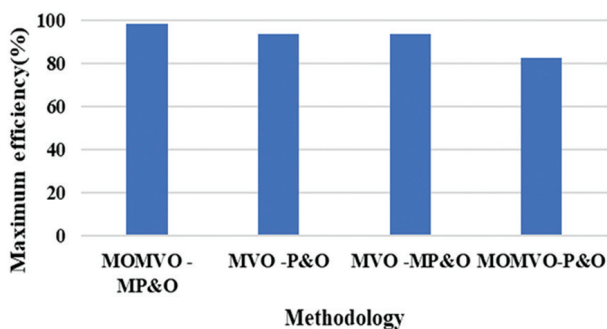


Fig. 4. Graphical representation of the proposed method for Pattern-1

Table 2. Evaluation of the proposed MOMVO with metaheuristic algorithms for pattern-1 performance analysis

Methodology	Power rating (W)	Extracted power (w)	Time Tracking (s)	Number of Iterations	Maximum Efficiency (%)
MOMVO -MP&O		104.21	0.35	13	98.51
MVO -P&O	104.50	103.45	1.50	13	91.56
MVO -MP&O		103.36	1.48	15	93.41
MOMVO-P&O		85.65	0.78	10	78.74

3.2. PATTERN- 2

The MPPT algorithms are sequentially initiated at different irradiance levels: 450, 750, and 650 W/m² for the first, second, and third PV modules, as described in this research. The experimental waveforms are illustrated in Fig. 4. Table 3 presents the response times of the MPPT algorithms to reach stability near the power point for these conditions with a 123.88 W of global power. By employing the proposed MOMVO process in a PV system, a peak power of 122.88 W is attained within 0.54 seconds and fifteen iterations, showcasing the effectiveness of the MOMVO algorithm in achieving optimal power tracking. The results show that the suggested method attains greater performances in terms of power rating, extracted power, time tracking, number of iterations, and maximum efficiency. The existing methods namely, MVO-P&O, MVO-MP&O and MOMVO-P&O correspondingly achieve 93.52%, 93.74% and 82.80% in terms of maximum efficiency, whereas that of the suggested MOMVO-MP&O method is 98.50%, as illustrated in table 3 and Fig. 5.

Table 3. Evaluation of the proposed MOMVO with metaheuristic techniques for pattern-2 performance analysis

Methodology	Power rating (W)	Extracted power (w)	Time Tracking (s)	Amount of Iterations	Maximum Efficiency (%)
MOMVO-MP&O	123.88	123.12	0.45	13	98.50
MVO-P&O		122.14	1.30	20	93.52
MVO-MP&O		120.48	0.60	16	93.74
MOMVO-P&O		116.12	0.78	15	82.80

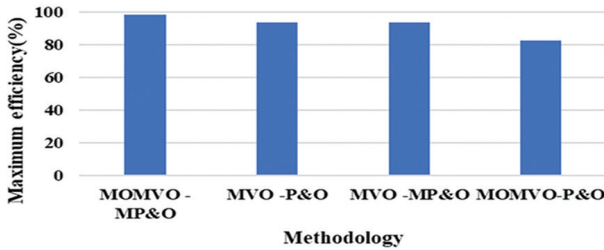


Fig. 5. Graphical representation of the proposed method for Pattern-2

3.3. PATTERN- 3

For pattern 3, the first, second, and third PV modules experience irradiances of 1000, 600, and 600 W/m², respectively. Analysis of the P-V curve reveals two peaks with the global maximum at the second peak and a local maximum at the leftmost peak. The maximum power output for pattern-3 is measured at 157.95 W. The typical MVO algorithm with a tracking duration of 1.31 s and an MPPT value of 156.67 W produce power waveforms with steady-state oscillations identical to patterns 1 and 2. In addition, the PSO algorithm is used in pattern-3 which achieves MPPT in 15 iterations with a global peak power of 138.66 W, and a tracking time of 1.02 seconds. PSO dominates those obtained by MOMVO in average. By making similar observations for the remaining rows, it attains better performance of MOMVO solutions. The results prove that the suggested method achieves higher performances in terms of Power rating, extracted power, Time tracking, number of iterations, and Maximum Efficiency. The existing methods namely, MVO-P&O, MVO-MP&O, and MOMVO-P&O correspondingly accomplish 90.54%, 94.34% and 93.80% in terms of maximum efficiency, whereas the suggested MOMVO-MP&O method accomplishes a higher performance with 98.80%, as outlined in Table 4 and Fig. 6.

Table 4. Evaluation of the proposed MOMVO with metaheuristic algorithms for pattern-3 performance analysis

Methodology	Power rating (W)	Extracted power (w)	Time Tracking (s)	Number of Iterations	Maximum Efficiency (%)
MOMVO-MP&O	157.95	158.48	0.35	12	98.80
MVO-P&O		157.69	1.25	20	90.54
MVO-MP&O		157.75	0.10	17	94.34
MOMVO-P&O		157.65	1.40	14	93.80

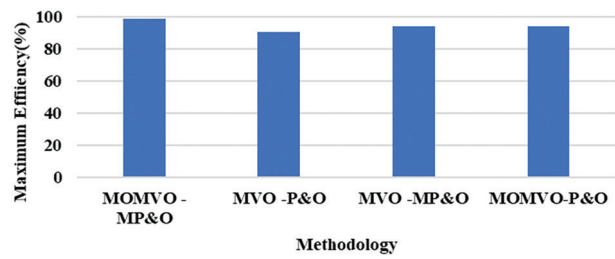


Fig. 6. Graphical representation of the proposed method for pattern-3

3.4. COMPARATIVE ANALYSIS

To test the efficiency of grid-connected RES using the MOMVO controller, a comparative analysis is conducted against the pre-established RES system designs, MF-DPC [19] and ANFIS [22]. This comparison primarily focuses on evaluating the MOMVO-MP&O controller's performance in terms of THD. From the results, it is evident that the existing method MF-DFC achieves 3.9% and ANFIS achieves 3.85% for THD performance matrices. When contrasted against the existing methods, the introduced method accomplishes a superior THD value of 1.86%. The novel MOMVO-MP&O strategically addresses the existing issues by minimizing the harmonic distortion more effectively than the existing approaches. Its meticulous parameter tuning and innovative algorithmic design synergistically mitigate distortions, showcasing commendable performance in harmonic reduction across diverse scenarios. Table 5 and Fig. 7 exhibit the contrast between the proposed method and previous methods.

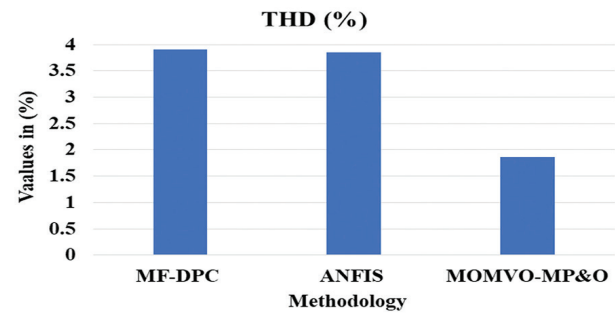


Fig. 7. Comparison of THD for MOMVO and existing methods

Table 5. Comparison of THD for MOMVO and existing methods

Method	THD (%)
MF-DPC [19]	3.9
ANFIS [22]	3.85
MOMVO-MP&O	1.86

4. CONCLUSION

This research proposed the combination of MP&O with MOMVO to design an efficient EMS that involves PV, wind, and battery system. The system's total energy density is increased when a battery is used together with

the RES. Applying the modulation index from MOMVO, the battery and wind turbine can be switched on and off efficiently. Additionally, the MP&O technique is used to obtain steady power from the grid-connected RES by enabling the DC-DC converter. Furthermore, capacitor banks are used to prevent power spikes in the given power. The grid-connected RES system MOMVO-MP&O controller is examined at different temperatures and irradiance levels. With the inclusion of the MP&O's current change and the increased MOMVO convergence rate, an efficient EMS operated by a PV, wind, and battery system can be produced. Additionally, when compared to existing methods namely MF-DPC and ANFIS the proposed MOMVO-MP&O have achieved better performances in terms of THD. The results demonstrated that a significantly higher THD value of 1.86% which is relatively higher than existing models. In future, the proposed model can be enhanced by considering more random and uncertain factors in the generation, load demand and smart battery charging technology.

5. REFERENCES

- [1] A. Naderipour, H. Kamyab, J. J. Klemeš, R. Ebrahimi, S. Chelliapan, S. A. Nowdeh, A. Abdullah, M. H. Marzbali, "Optimal design of hybrid grid-connected photovoltaic/wind/battery sustainable energy system improving reliability, cost and emission", *Energy*, Vol. 257, 2022, p. 124679.
- [2] A. L. Konde, M. Kusaf, M. Dagbasi, "An effective design method for grid-connected solar PV power plants for power supply reliability", *Energy for Sustainable Development*, Vol. 70, 2022, pp. 301-313.
- [3] N. Niveditha, M. M. R. Singaravel, "Optimal sizing of hybrid PV-Wind-Battery storage system for Net Zero Energy Buildings to reduce grid burden", *Applied Energy*, Vol. 324, 2022, p. 119713.
- [4] S. Bhattacharyya, S. Puchalapalli, B. Singh, "Operation of Grid-Connected PV-Battery-Wind Driven DFIG Based System", *IEEE Transactions on Industry Applications*, Vol. 58, No. 5, 2022, pp. 6448-6458.
- [5] S. K. Thirumalai, A. Karthick, P. K. Dhal, S. Pundir, "Photovoltaic-wind-battery and diesel generator-based hybrid energy system for residential buildings in smart city Coimbatore", *Environmental Science and Pollution Research*, Vol. 31, 2024, pp. 14229-14238.
- [6] S. Boualem, O. Kraa, M. Benmeddour, M. Kermadi, M. Maamir, H. Cherif, "Power management strategy based on Elman neural network for grid-connected photovoltaic-wind-battery hybrid system", *Computers and Electrical Engineering*, Vol. 99, 2022, p. 107823.
- [7] M. A. Judge, A. Khan, A. Manzoor, H. A. Khattak, "Overview of smart grid implementation: Frameworks, impact, performance and challenges", *Journal of Energy Storage*, Vol. 49, 2022, p.104056.
- [8] P. A. Gbadega, X. Y. Sun, "JAYA algorithm-based energy management for a grid-connected micro-grid with PV-wind-microturbine-storage energy system", *International Journal of Engineering Research in Africa*, Vol. 63, 2023, pp. 159-184.
- [9] P. Rajesh, F. H. Shajin, B. Rajani, D. Sharma, "An optimal hybrid control scheme to achieve power quality enhancement in micro grid connected system", *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, Vol. 35, No. 6, 2022, p. e3019.
- [10] R. Z. Falama, V. Dumbrava, A. S. Saidi, E. T. Houdji, C. B. Salah, S. Y. Doka, "A Comparative-Analysis-Based Multi-Criteria Assessment of On/Off-Grid-Connected Renewable Energy Systems: A Case Study", *Energies*, Vol. 16, No. 3, 2023, p. 1540.
- [11] L. Garg, S. Prasad, "A review on controller scalability for VSC-MTDC grids: challenges and applications", *Smart Science*, Vol. 11, No. 1, 2023, pp. 102-119.
- [12] R. Fathi, B. Tousei, S. Galvani, "Allocation of renewable resources with radial distribution network reconfiguration using improved salp swarm algorithm", *Applied Soft Computing*, Vol. 132, 2023, p. 109828.
- [13] R. Bisht, A. Sikander, "A novel hybrid architecture for MPPT of PV array under partial shading conditions", *Soft Computing*, Vol. 28, No. 2, 2024, pp. 1351-1365.
- [14] A. Khosravani, E. Safaei, M. Reynolds, K. E. Kelly, K. M. Powell, "Challenges of reaching high renewable fractions in hybrid renewable energy systems", *Energy Reports*, Vol. 9, 2023, pp. 1000-1017.
- [15] A. M. Jasim, B. H. Jasim, A. Flah, V. Bolshev, L. Mihet-Popa, "A new optimized demand management system for smart grid-based residential buildings adopting renewable and storage energies", *Energy Reports*, Vol. 9, 2023, pp. 4018-4035.

- [16] M. Mishra, P. Mahajan, R. Garg, "Implementation and comparison of metaheuristically modified ANN MPPT controllers under varying solar irradiance conditions", *Electrical Engineering*, Vol. 106, 2023, pp. 3427-3443.
- [17] T. Zhang, B. Sobhani, "Optimal economic programming of an energy hub in the power system while taking into account the uncertainty of renewable resources, risk-taking and electric vehicles using a developed routing method", *Energy*, Vol. 271, 2023, p. 126938.
- [18] B. Lin, C. Huang, "Promoting variable renewable energy integration: The moderating effect of digitalization", *Applied Energy*, Vol. 337, 2023, p. 120891.
- [19] Y. Sahri, S. Tamalouzt, S. L. Belaid, M. Bajaj, S. S. M. Ghoneim, H. M. Zawbaa, S. Kamel, "Performance improvement of hybrid system based DFIG-wind/PV/batteries connected to DC and AC grid by applying intelligent control", *Energy Reports*, Vol. 9, 2023, pp. 2027-2043.
- [20] M. M. Gulzar, A. Iqbal, D. Sibtain, M. Khalid, "An Innovative Converterless Solar PV Control Strategy for a Grid Connected Hybrid PV/Wind/Fuel-Cell System Coupled with Battery Energy Storage", *IEEE Access*, Vol. 11, 2023, pp. 23245-23259.
- [21] M. Maaruf, K. Khan, M. Khalid, "Robust Control for Optimized Islanded and Grid-Connected Operation of Solar/Wind/Battery Hybrid Energy", *Sustainability*, Vol. 14, No. 9, 2022, p. 5673.
- [22] P. Cholamuthu, B. Irusappan, S. K. Paramasivam, S. K. Ramu, S. Muthusamy, H. Panchal, R. S. S. Nuvvula, P. P. Kumar, B. Khan, "A Grid-Connected Solar PV/Wind Turbine Based Hybrid Energy System Using ANFIS Controller for Hybrid Series Active Power Filter to Improve the Power Quality", *International Transactions on Electrical Energy Systems*, Vol. 2022, 2022, p. 9374638.
- [23] A. Ibáñez-Rioja, L. Järvinen, P. Puranen, A. Kosonen, V. Ruuskanen, K. Hynynen, J. Ahola, P. Kauranen, "Off-grid solar PV-wind power-battery-water electrolyzer plant: Simultaneous optimization of component capacities and system control", *Applied Energy*, Vol. 345, 2023, p. 121277.
- [24] A. O. Amole, S. Oladipo, O. E. Olabode, K. A. Makinde, P. Gbadega, "Analysis of grid/solar photovoltaic power generation for improved village energy supply: A case of Ikose in Oyo State Nigeria", *Renewable Energy Focus*, Vol. 44, 2023, pp. 186-211.
- [25] N. J. Bodele, P. S. Kulkarni, "Modular battery-integrated bidirectional single-stage DC-DC converter for solar PV based DC Nano-grid application", *Solar Energy*, Vol. 259, 2023, pp. 1-14.

A New Proposed Triple Active Bridge Converter for Fuel Cell Applications: Study, Control and Energy Management

Case Study

Abdelkarim Aouiti

Computer laboratory for electrical systems,
LR11ES26, INSAT, University of Carthage,
Centre Urbain Nord BP 676 - 1080 Tunis Cedex, TUNISIA
abdelkarim.aouiti.ensit@gmail.com

Mokhtar Abbassi

Computer laboratory for electrical systems,
LR11ES26, INSAT, University of Carthage,
Centre Urbain Nord BP 676 - 1080 Tunis Cedex, TUNISIA
mok98474304@gmail.com

Faouzi Bacha*

University of Tunis, Tunisia
ENSIT, Ave Taha Hussine, 1008, Tunis
faouzi.bacha@esstt.rnu.tn

*Corresponding author

Abstract – This paper deals with a new proposed three port converter structure dedicated for two-input source hybrid systems especially for fuel cell applications. This converter is made up of three-phase triple active bridges which are galvanically isolated by means of three single phase high frequency transformers. The present converter integrates a fuel cell as the primary power source with a battery that stores energy, harnessing the unique benefits of both sources to deliver reliable power to a DC load through a single power conversion stage. In order to control the power flow between the ports, a phase shift control technique has been carried out to generate the control signals of the load and battery side bridges in reference with those of the fuel cell bridge. A detailed analysis of the proposed converter has been presented in this paper. A novel proposed energy management algorithm has been developed. This algorithm provides a robust solution for managing and distributing power flow between the converter's ports, ensuring an optimal balance of power delivery. The algorithm has been rigorously validated through simulations and experimental test, using Dspace 1104 board.

Keywords: Power flow control, three port converter, high frequency transformer, phase shift control technique

Received: February 13, 2024; Received in revised form: August 10, 2024; Accepted: August 13, 2024

1. INTRODUCTION

Today, global greenhouse gas emissions, primarily from fuel activities, have led to a 1°C rise in the global average temperature since the pre-industrial era. Projections indicate that without intervention, the global average temperature will surpass 1.5° C between 2030 and 2052. The heavy reliance on fossil fuels is unsustainable due to their significant CO₂ emissions and the impending surge in operating costs as reserves dwindle post-2030, [1]. Moreover, global oil reserves, estimated at around 1,732 billion barrels—equivalent to about 52 years of production at current rates—stood

at the end of 2020. This is a theoretical duration, considering that production from existing fields declines over time, amounting to approximately 236,295 million tonnes, [2]. To combat emissions, system designers are exploring cleaner, more efficient energy alternatives, with renewable sources expected to replace fossil fuels over time. Challenges include improving fuel cell efficiency and internal structures. Hydrogen is identified as an ideal energy source, and multi-source hybrid systems, integrating fuel cells with storage batteries and supercapacitors, are proposed to enhance dynamic response to load variations.

Recent renewable energy developments in hybrid systems, aerospace, smart grids, and portable devices have spurred challenges in designing new energy conversion systems. These systems use multi-input power electronic converters to integrate various power sources and provide a well-controlled output for diverse applications. So, converters have evolved from single input/single output to multi-input/multi-output structures, with two reported topologies in the literature [3]. The first topology adopts a conventional structure by consolidating various sources on a common bus. Each source undergoes separate power conversion stages with independent control for each converter. A communication bus facilitates information exchange between sources, leading to increased complexity and higher converter costs due to the involvement of multi-stage power conversion and necessary communication devices. The second multiport topology treats the entire structure as a single power converter, combining multiple sources with power regulation by a centralized controller. These converters, with their simple structure, minimum converter stages, and fewer devices, find applications in various fields, including hybrid power systems, hybrid vehicles, aerospace, satellite applications, portable electronic devices, and uninterruptible power supplies.

Multiport converters are categorized into two types: isolated converters and non-isolated converters. Isolated converters are utilized to separate the low voltage side from the high voltage side, mitigating the risk of electric shock and ensuring voltage compatibility. Additionally, achieving a high voltage output necessitates the inclusion of an isolation stage in converters for upcoming power conversion systems. Galvanic isolation using high-frequency (HF) transformers is imperative for these systems due to several advantages. They offer compact size, lightweight, and cost-effectiveness. Moreover, they significantly reduce noise and minimize the size of passive elements due to their high switching frequency. Furthermore, HF transformers prevent harmonic current and voltage distortions induced by the saturation of low-frequency transformer cores.

In this context, this article presents a recent trend in the development of a converter dedicated for a multi-source hybrid system. Through this work, we propose a new multiport converter topology. This structure uses three-phase three port active bridge DC-DC converters in order to benefit from the advantage of the elementary DAB-3ph structure and the three-phase configuration. In fact, the three-phase double active bridge reveals the following benefits: low electromagnetic interference due to high frequency operation, the compactness of the transformer size, the minimized effective currents through the switches and diodes, and a greatly reduced conduction losses with an inherent smooth switching of power switches, [4, 5].

The single-phase and the three phase DAB have both been firstly proposed in [6]. The characteristics of the materials and components, particularly the significant

losses brought on by the low frequency transformers, severely constrained the benefits of these converters.

After the development of the power converter components and the improvement of the core materials, the literature has shown a significant interest in the single-phase [7-14] and three-phase double active bridge [3, 4] configurations, and several works have been presented. Also, the elementary dual active bridge structure has been extended to accommodate three ports. A triple active bridge converter is proposed in [15-17]. While [15] focuses on electric vehicle charging, utilizing a generalized-harmonic-approximation method for accurate waveform prediction and achieving high efficiency (97.6%) with zero-voltage switching, [16] emphasizes ZVS across a wide load range and input voltage. In [17], a decoupling control method for a triple active bridge has been introduced, enabling simultaneous charging of two EV battery stacks while maintaining isolation.

In [18], a 3p-3ph converter for high-power applications integrates a slow transient main energy source and a fast storage device to supply a DC load. Using three bidirectional full bridge inverters and a 3p-3ph high-frequency transformer, power flow is controlled by adjusting phase shift angles among the inverter stages. While this structure offers the advantage of handling three ports, it faces limitations related to voltage level capability. In [19], a fuel cell-powered uninterruptible power supply system integrates a three-port bidirectional converter, linking the fuel cell and supercapacitor to a grid-connected inverter. It operates in stand-alone or grid-connected modes, with two bridge phase shifts controlling the DC/DC stage for simultaneous adjustment of fuel cell power and DC-link voltage. However, this structure is noted for its high current distortion compared to the three-phase structure and its low output voltage level. As to [20], a novel three-port bidirectional converter with a three-winding coupled inductor is proposed for PV systems, enabling high step-up/step-down conversion. In [21], a novel approach to integrate hybrid energy storage systems (HESSs) with renewable energy sources (RES) and loads using a current-fed dual active bridge (DAB) converter has been proposed. This method regulates load voltage, tracks RES maximum power point, and safeguards the battery from transients, enhancing system efficiency and reliability.

In this work, a novel three-port converter topology is proposed. By integrating the three-phase triple active bridge structure with the buck-boost configuration, this converter offers a solution that combines the advantages of both topology and overcomes the lack of the voltage level for the three-port configuration. This converter accommodates a fuel cell which represents the main source in the system; and a battery which allows energy storage. The advantages of the two sources are combined in order to supply a DC load while using a single conversion stage. The instantaneous power in the system can be distributed in a controlled way, which improves its dynamics and it increases its reli-

ability. This multiport converter is intended for medium to high power applications. Its main characteristics is the combination of three galvanically isolated ports with bidirectional power flow capabilities; where the voltage of the main source could be raised thanks to the employment of a boost structure, so a wide input voltage range could be applied and the gain could be adjusted by the duty cycle of switches of the principal bridge. The power flow in this system is simple, it may be achieved by introducing two phase shift angles. Moreover, the converter uses tiny passive elements, and it has very low current ripples. So, the system exhibits low power losses and important conversion efficiency. The principle of operation, as well as the control strategies adopted for the proposed converter, have been studied in details. A novel energy management algorithm has been developed to facilitate efficient and balanced power flow among the ports of the three-port converter. The algorithm's effectiveness in improving converter efficiency and power management was demonstrated through comprehensive simulations and experimental tests utilizing Dspace 1104 board.

2. SYSTEM DESCRIPTION

Fig.1 shows the proposed three port converter topology. This circuit is an extension of the three-phase DAB topology and thereby it is called a three-phase triple active bridge (TAB) converter. It is composed of three single-phase transformers with two secondaries. On the low voltage side, an interleaved boost structure is used to raise the voltage of the fuel cell. The boost mode is carried out by the inductors L_{dc1} , L_{dc2} and L_{dc3} and the three half-bridges of the low voltage side; and this to keep constant the voltage of the DC bus of port 1 and in order to allow a wide range for the variation of the fuel cell voltage. For the high voltage side, a load is connected to the first output port. As for the second output port, a battery, which acts as an energy storage device, is connected. The converter can operate in step-up mode when power is flowing from the low voltage side (LVS)

to the high voltage side (HVS) in order to supply the load and / or charge the battery. A transfer of power is also permitted from the battery to the load.

The switches of the three bridges are switched at the same frequency and these bridges generate the voltages v_{an} , v_{rM1} and v_{oM2} having a controlled phase shift with respect to the primary side (v_{an}). The maximum allowed power flow between the ports is directly related to the external inductances.

The three-phase triple active bridge (TAB) converter proposed in this study offers several significant advantages, making it an ideal solution for high power applications. In fact, it enables the combinations of three ports, including a high output voltage for the load through boosting the voltage of the main source. Moreover, the converter features a simple power flow control owing to the application of the phase shift control technique, effectively reduces current ripples.

The use of small-sized passive elements is made possible by the application of a high frequency galvanic isolation between the three ports.

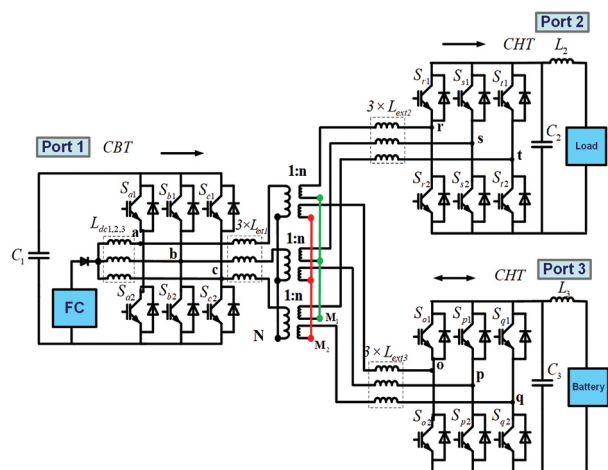


Fig. 1. Three-port three-phase active triple bridge converter diagram

Table 1. Comparison Table between the proposed converter and the existing topologies in the literature

Topology	Dual Active Bridge Multiport converter [21]	Three-port Bidirectional DC-DC converter [17]	Three-Port Three-Phase DC-DC Converter [18]	Proposed converter
Criteria				
Bidirectional power flow	✓	✓	✓	✓
Power (W)	135	5x103 (x 2 output)	103	10x103
Fully Isolated ports	x	✓	✓	✓
Output voltage (V)	100	400 (x 2 output)	800	200
Switching frequency (kHz)	50	20	100	20
Number of bridges / legs	2 briges and one leg	3	3	3
Number of switches	10	12	18	18
Power flow control	Phase shift control / Duty cycle	Phase shift control	Phase shift control	Phase shift control
Focus	Operation and control	Control	Control and transformer design	Energy management
Application	Renewable energy application	Electrical vehicle battery charging	Grid connected	Electrical Vehicle

The proposed converter exhibits exceptional bidirectional power flow capabilities and high-power density. With its compact size and fully isolated ports, it guarantees superior safety and reliability in operation.

Additionally, the converter provides a high output voltage at its output, significantly expanding its applicability across diverse tasks

Table 1 highlights the unique features and advantages of the proposed three-port converter in contrast to the existing topologies in the literature. It emphasizes the converter's versatility, efficiency, and reliability across various applications, particularly its suitability for electric vehicles and hybrid systems.

3. CONTROL STRATEGY AND DIFFERENT OPERATING MODES OF THE PROPOSED CONVERTER

A full wave control is chosen for the transformer primary-side inverter. As to the other bridges, a phase shift control has been adopted. So, the control signals could be expressed as:

$$\begin{cases} S_{j1}(\theta) = S_{r1}(\theta - \varphi_{12}), (i, j) = \{(a, r), (b, s), (c, t)\} \\ S_{k1}(\theta) = S_{r1}(\theta - \varphi_{13}), (i, k) = \{(a, o), (b, p), (c, q)\} \\ S_{m2} = \overline{S_{m1}}, m = \{a, b, c, r, s, t, o, p, q\} \end{cases} \quad (1)$$

Fig. 2 shows the Δ -model of the system where the three bridges are replaced by three voltage sources v_{ant} , v_{rM1} and v_{oM2} . The sources exchange energy through a network of inductors. The voltages v_{rM1} and v_{oM2} are respectively shifted by φ_{12} and φ_{13} with respect to the reference voltage van. The phase shifts are considered positive when the corresponding voltage is lagging in phase with respect to the reference van and they are negative when the considered voltage is leading the reference voltage van.

The inductors of Δ -model are derived from the transformer leakage inductors and the inductors L_{ext1} , L_{ext2} and L_{ext3} . Using this representation, the three-port system is broken down into three two-port subsystems which facilitates its analysis and its modelling. The magnetizing inductance, which does not contribute to the power flow, is neglected to simplify the analysis and therefore it is not shown in Fig. 2. According to the definitions of the Δ -model of Fig. 2 where all the quantities are brought back to the transformer primary-side, the relation between the phase shift angles and the power flow in the system turns out to be as follows:

$$P_{ij}(\varphi_{ij}) = \begin{cases} \frac{V_i V_j}{n L_{ij} \omega} \varphi_{ij} \left(\frac{2}{3} - \frac{\varphi_{ij}}{2\pi} \right) \text{ for } 0 \leq \varphi_{ij} \leq \frac{\pi}{3} \\ \frac{V_i V_j}{n L_{ij} \omega} \left(\varphi_{ij} - \frac{\varphi_{ij}^2}{\pi} - \frac{\pi}{18} \right) \text{ for } \frac{\pi}{3} \leq \varphi_{ij} \leq \frac{2\pi}{3} \end{cases} \quad (2)$$

Where, $V1 = V_{FC}$, $V2 = V_{load}$ et $V3 = VBT$ are the voltages at the ports; and φ_{ij} presents the phase shifts in radians and n denotes the transformation ratio of the transformer.

The phase shifts φ_{ij} are selected according to the powers that should be transferred between the ports. Taking into account the expressions of the powers P_{ij} between a port "i" and a port "j", the maximum power flow through L_{ij} occurs for the value $\pi/2$ of the phase shift angle. Once the inductors L_{ij} are fixed, the inductors L_{ext1} , L_{ext2} and L_{ext3} are determined by the Δ -T transformation and using equations presented in appendices 1.

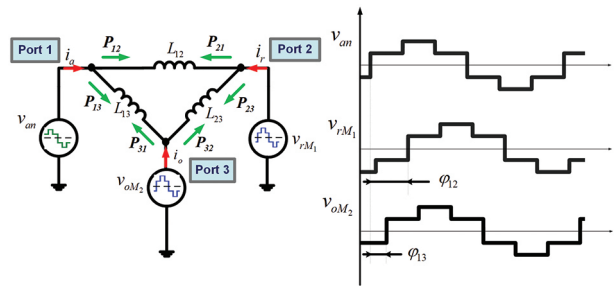


Fig. 2. Three-port three-phase active triple bridge converter diagram

Using the Δ -model in Fig. 2, the power flow at each port is a combination of the power flows through two associated branches. For a lossless system, they are:

$$\begin{cases} P_1 = P_{12} + P_{13} \\ P_2 = -P_{12} - P_{32} \\ P_3 = P_{32} - P_{13} \\ P_1 + P_2 + P_3 = 0 \end{cases} \quad (3)$$

Where $P_1 = P_{FC}$ is the power supplied by port 1 (the main source P_1) to port 2 and port 3, $P_2 = -P_{Load}$ is the power supplied to the port 2 (the charge port) (a negative sign of P_2 means that the load consumes energy) and $P_3 = P_{BAT}$ is the power drawn from port 3 (the storage port) (a negative sign of P_3 implies that the energy is stored in the battery).

The parameters of the proposed converter are given by the following table:

Table 2. Parameters of the three-port three-phase active triple bridge converter

Parameters	Designation	Value
$L_{dc1}-L_{dc2}-L_{dc3}$	Boost Inductance	1.2 μ H
L_{ext1}	Primary side external inductance	1.35 μ H
L_{ext2}	1st Secondary side external inductance	12.9 μ H
L_{ext3}	2nd Secondary side external inductance	26.3 μ H
n	Transformation ratio of the transformer	7
f_s	Switching frequency	20 kHz

By controlling the phase-shift angles between the three ports, the three-port converter operating modes can be distinguished by the combination of power flows. Since the fuel cell cannot absorb power, there are only six modes as follows:

Mode 1: the fuel cell supplies the DC load and simultaneously charges the battery.

Mode 2: the load is supplied by both sources at the same time: the fuel cell and the battery.

Modes 3: the load is only supplied by the fuel cell.

Modes 4: the load is only supplied by the battery.

Mode 5: only concerns battery charging via the fuel cell.

Mode 6: No power transit between the ports.

Table 3. The possible operating modes according to the power flow combinations

Load power	Battery power	Fuel cell power	
		$P_{FC} > 0$	$P_{FC} = 0$
$P_{Load} < 0$	$P_{BAT} > 0$	Mode 2	Mode 4
	$P_{BAT} < 0$	Mode 1	Doesn't exist
	$P_{BAT} = 0$	Mode 3	Doesn't exist
$P_{Load} = 0$	$P_{BAT} > 0$	Doesn't exist	Doesn't exist
	$P_{BAT} < 0$	Mode 5	Doesn't exist
	$P_{BAT} = 0$	Doesn't exist	Mode 6

The selection of the convenient mode is according to the load demand and the state of the charge of the battery. Once the mode is selected, the reference powers are provided and after that the phase shift are determined, as illustrated by Fig. 3 and table 4.

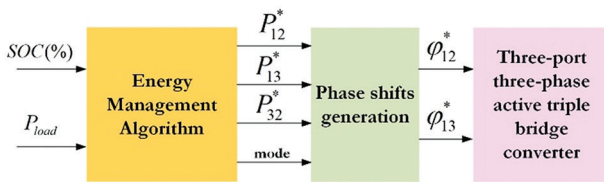


Fig. 3. Energy management and phase shift angles generation for the three-port three-phase active triple bridge converter

Table 4. The different operating modes as function of the phase shift angles φ_{12} and φ_{13}

Mode	φ_{12} (rad)	φ_{13} (rad)
Mode 1	$0 < \varphi_{12} \leq \pi/2$	$\varphi_{13} = \varphi_{12}$
Mode 2	$0 < \varphi_{12} \leq \pi/2$	$\varphi_{12} < \varphi_{13} \leq 2\pi/3$
Mode 3	$0 < \varphi_{12} \leq \pi/2$	$\varphi_{12}/2$
Mode 4	0	$(-\pi)/2 \leq \varphi_{13} < 0$
Mode 5	0	$0 < \varphi_{13} \leq \pi/2$
Mode 6	0	0

4. ENERGY MANAGEMENT

An energy management algorithm for the three-port converter has been proposed to manage the power flow in the system. This energy management algorithm, which is shown by Fig. 4, aims to optimize the power flow between the fuel cell, battery, and load in order to maximize the efficiency and longevity of the system. By monitoring the state of charge of the battery, the power demand of the load, and the output power of the fuel cell, the algorithm can dynamically adjust the power flow to maintain a balance between the different components and ensure that the load is always receiving the required amount of power.

To assess the power distribution strategy's effectiveness and observe system behavior across various modes of operation, we employed a load profile and battery state of charge as depicted in Fig. 5. The load

power, which is depicted in blue, varies between 0 and 1200 W, reflecting typical real-world operating conditions that test the system's ability to adapt dynamically to fluctuating power demands. Meanwhile, the state of charge, shown in green, ranges from 20 to 100%. The fuel cell provides 1200 W and the battery offers 1000 W. Figs. 6 and 7 illustrate the transferred powers and power distribution across the ports for both experimental and simulation tests.

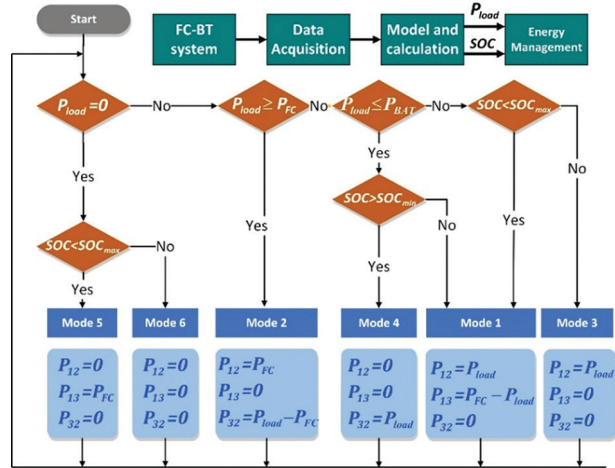


Fig. 4. The proposed energy management algorithm

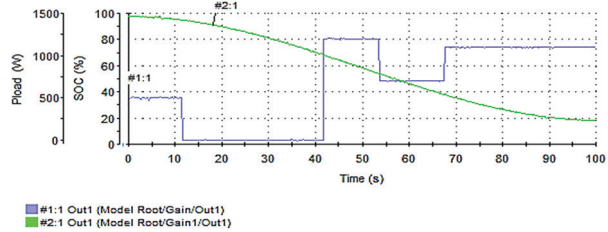


Fig. 5. Load power and state of the charge profiles

The comparison between these figures demonstrates the algorithm's accuracy in predicting power distribution under varying loads.

Between 0 and 10 seconds, the load power remains steady at 500 W, demonstrating the system's ability to maintain stable operation under moderate demand conditions. During this period, the battery exceeds its maximum state of charge, and the algorithm effectively directs the battery to supply the required power to the load (mode 4). Between 10 and 30 seconds, with no load demand and the battery is fully charged, the system correctly identifies that no power transfer is needed (mode 6), thereby conserving energy. From 30 to 40 seconds, as the battery's charge level is below its maximum state of charge, the system shifts to charging the battery from the fuel cell, efficiently managing power flow to prepare for future demand. Between 40 and 55 seconds, the fuel cell is delivering the 500W power required by the load (mode 2), highlighting the fuel cell's role in sustaining load demands during periods when the battery is inactive or not available. From 55 to 68 seconds, once again, the battery takes over by supplying the necessary power

to the load (mode 4), showcasing the system's flexibility and its ability to seamlessly switch between power sources depending on the operational needs.

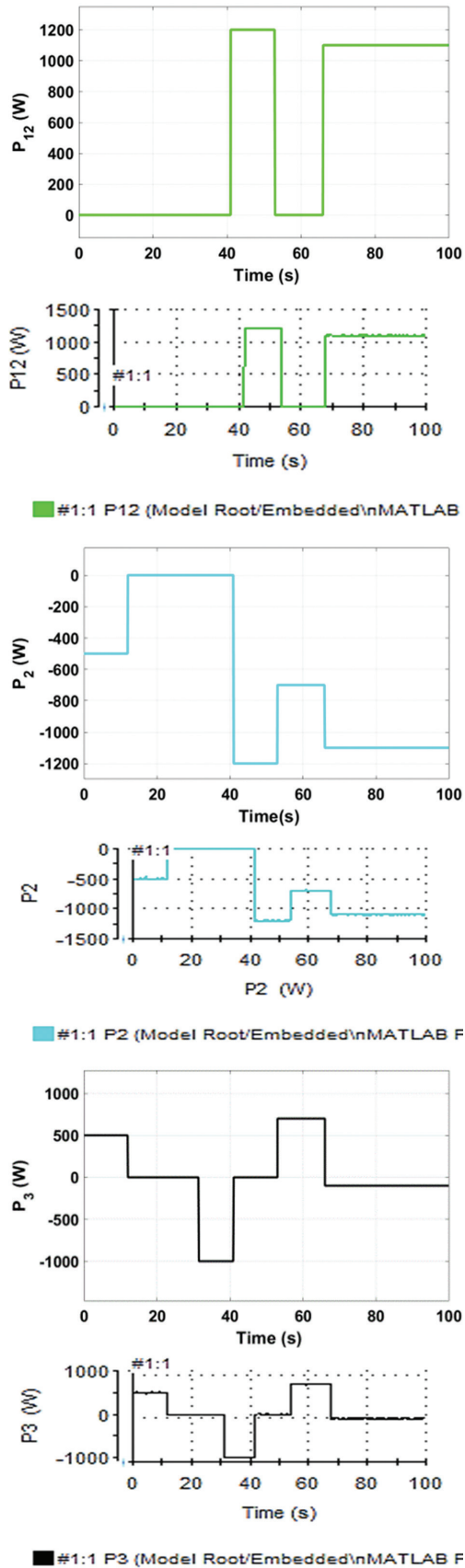


Fig. 6. Transferred power between the ports. (a to c) simulation (A to C) Experimental result

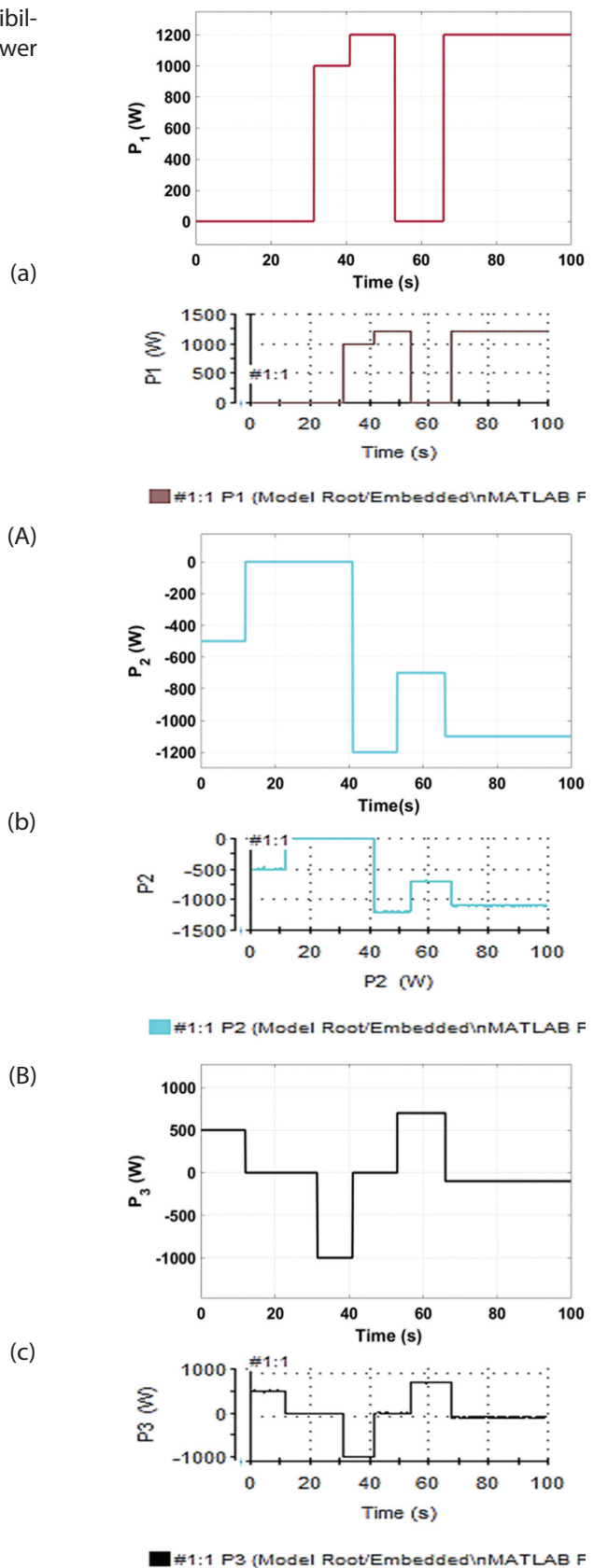


Fig. 7. Powers at the ports. (a to c) simulation (A to C) Experimental result

During the period from 68 to 100 seconds, with the battery's charge level dropping below the minimum threshold, the algorithm prioritizes recharging the battery from the fuel cell, while simultaneously ensuring the

load receiving continuous power (mode 1). This operation illustrates the system's capacity for long-term power management, maintaining the battery's health while meeting load requirements.

The close alignment between experimental and simulation results in Figs. 6 and 7 validates the algorithm's robustness and the capability to efficiently supply the required power under various load and state of charge conditions. The proposed algorithm continuously monitors the load power demand, adapting the power flow accordingly. It draws from the battery when the load requires more power than the fuel cell provides and charges the battery with excess power when the load demands less than the fuel cell generates. This strategy ensures optimal power sharing, allowing the fuel cell to operate only when necessary. Consequently, the proposed algorithm proves successful in optimizing power distribution and demonstrating its practicality.

The complexity of the proposed energy management algorithm is notably reduced due to its rule-based optimization approach, which simplifies decision-making processes compared to more complex algorithms such as dynamic programming. This simplicity is advantageous when implementing the algorithm in real-time applications. During experimental tests using the dSPACE 1104 board, the algorithm demonstrated exceptional execution speed, facilitated by the board's rapid sampling capabilities. The dSPACE 1104's high sampling rate, a key characteristic of its design, ensures that the algorithm can continuously monitor and adjust power flow with minimal latency. This rapid processing is crucial for maintaining efficient power management and system stability, particularly in dynamic scenarios where load demands and battery states fluctuate. The rule-based nature of the algorithm further contributes to its efficient execution, as it reduces computational overhead by relying on predefined rules rather than iterative calculations. Overall, the combination of the dSPACE 1104's fast sampling and the algorithm's straightforward structure supports effective real-time performance and reliable operation.

5. CONCLUSION

In this paper, we proposed a three-port three-phase DC-DC converter structure tailored for various applications. This structure integrates a fuel cell as the main power source, providing the average power required by the load. Additionally, it includes a battery that functions as a storage device, catering to peak and fluctuating load power demands. An energy management algorithm was also introduced to regulate power flow within the system, ensuring the battery supplies power efficiently to the load.

This strategy not only optimizes power distribution but also highlights the pros of our approach, including improved power sharing and system reliability. However, potential challenges include the complexity of

managing multiple power sources and ensuring seamless transitions between them. Future work could focus on refining the control algorithm and exploring alternative energy storage options to further enhance the system's efficiency and applicability.

APPENDICES 1: TRANSFORMATION FROM Δ STRUCTURE TO T

$$M_1 = L_m$$

$$M_{12} = \left(\frac{1}{L_m} + \frac{1}{L_{12}} + \frac{1}{L_{31} + L_{23}} \right)^{-1}, M_{13} = \left(\frac{1}{L_m} + \frac{1}{L_{31}} + \frac{1}{L_{12} + L_{23}} \right)^{-1}$$

$$M_2 = n^2 \left(L_m + \left(\frac{1}{L_{12}} + \frac{1}{L_{31} + L_{23}} \right)^{-1} \right), M_{21} = n^2 \left(\frac{1}{L_{21}} + \frac{1}{L_{31} + L_{23}} \right)^{-1} \quad (4)$$

$$M_{23} = n^2 \left(\frac{1}{L_{23}} + \frac{1}{L_{12} + \left(\frac{1}{L_m} + \frac{1}{L_{31}} \right)^{-1}} \right)^{-1}$$

$$M_3 = n^2 \left(L_m + \left(\frac{1}{L_{31}} + \frac{1}{L_{12} + L_{23}} \right)^{-1} \right), M_{31} = n^2 \left(\frac{1}{L_{31}} + \frac{1}{L_{12} + L_{23}} \right)^{-1}$$

$$M_{32} = n^2 \left(\frac{1}{L_{23}} + \frac{1}{L_{31} + \left(\frac{1}{L_m} + \frac{1}{L_{12}} \right)^{-1}} \right)^{-1} \quad (5)$$

$$\Delta M_{12} = M_1 - M_{12}, \quad \Delta M_{13} = M_1 - M_{13}$$

$$L_M = \sqrt{\frac{\Delta M_{12} \Delta M_{13}}{1 - \frac{1}{2} \left(\frac{M_{32}}{M_3} + \frac{M_{23}}{M_2} \right)}}, L_{ext1} = M_1 - L_M$$

$$L_{ext2} = M_2 \left(1 - \frac{\Delta M_{12}}{L_M} \right), L_{ext3} = M_3 \left(1 - \frac{\Delta M_{13}}{L_M} \right) \quad (6)$$

Where,

$L_{12}=L_{21}$: inductance between port 1 and 2 in delta equivalent model.

$L_{13}=L_{31}$: inductance between port 1 and 3 in delta equivalent model.

$L_{23}=L_{32}$: inductance between port 2 and 3 in delta equivalent model of the transformer.

$L_{ext1}, L_{ext2}, L_{ext3}$: inductances of the T model of the transformer.

M_{ij}, M_i : quantities as functions of L_{ij} inductances.

n : the turn ratio of the transformer.

L_m : magnetizing inductance in Delta model.

6. REFERENCES:

- [1] M. R. Allen, "IPCC Framing and Context", Global Warming of 1.5 °C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty", Cambridge University Press, 2022, pp. 49-92.

- [2] BP, "BP Statistical Review of World Energy", <https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2021-full-report.pdf> (accessed: 2024)
- [3] S. Farajdadian, A. Hajizadeh, M. Soltani, "Recent developments of multiport DC/DC converter topologies, control strategies, and applications: A comparative review and analysis", *Energy Reports*, Vol. 11, 2024, pp. 1019-1052.
- [4] A. Aouiti, A. Soyed, F. Bacha, "Control and Study of the Bidirectional Three Phase DAB Converter", *Proceedings of the 8th International Conference on Control, Decision and Information Technologies*, Istanbul, Turkey, 17-20 May 2022, pp. 1008-1013.
- [5] A. Aouiti, A. Soyed, F. Bacha, "Study and Control of the 3ϕ -DAB DC-DC Converter with a Boost Structure", *Proceedings of the 9th International Conference on Control, Decision and Information Technologies*, Rome, Italy, 2023, pp. 2621-2626.
- [6] R. W. A. A. De Doncker, D. M. Divan, M. H. Kheraluwala, "A three-phase soft-switched high-power-density DC/DC converter for high-power applications", *IEEE Transactions on Industry Applications*, Vol. 27, No. 1, 1991, pp. 63-73.
- [7] S. Yalcin et al. "Experimental analysis of phase shift modulation methods effects on EMI in dual active bridge DC-DC converter", *Engineering Science and Technology*, Vol. 43, 2023, p. 101435.
- [8] J. Yin, X. He, J. Lu, H. S.-H. Chung, "Phase-Shift Control with Unified PWM/PFM for Improved Transient Response in a Bidirectional Dual-Active-Bridge DC/DC Converter", *IEEE Transactions on Industrial Electronics*, Vol. 70, No. 9, 2023, pp. 8862-8872.
- [9] T.-Q. Duong, S.-J. Choi, "Deadbeat Control with Bivariate Online Parameter Identification for SPS-Modulated DAB Converters", *IEEE Access*, Vol. 10, 2022, pp. 54079-54090.
- [10] J. Guacaneme, G. Garcerá, E. Figueres, I. Patrao, R. González-Medina, "Dynamic modeling of a dual active bridge DC to DC converter with average current control and load-current feed-forward", *International Journal of Circuit Theory and Applications*, Vol. 43, 2015, pp. 1311-1332.
- [11] A. Mansour, B. Faouzi, G. Jamel, E. Ismahen, "Design and analysis of a high frequency DC-DC converters for fuel cell and super-capacitor used in electrical vehicle", *International Journal of Hydrogen Energy*, Vol. 39, No. 3, 2014, pp. 1580-1592.
- [12] H. Atallah et al. "Analysis of the Dual Active Bridge-Based DC-DC Converter Topologies, High-Frequency Transformer, and Control Techniques", *Energies*, Vol. 15, No. 23, 2022, p. 8944.
- [13] F. Slah, A. Mansour, A. Abdelkarim, F. Bacha, "Analysis and design of an LC parallel-resonant DC-DC converter for a fuel cell used in an electrical vehicle", *Journal of Circuits, Systems and Computers*, Vol. 27, No. 8, 2018, p. 1850119.
- [14] A. Aouiti, A. Soyed, F. Bacha, "Study and Analysis of 2-Phase DAB DC-DC Converter for Fuel Cell Applications", *Proceedings of the IEEE International Conference on Advanced Systems and Emergent Technologies*, Hammamet, Tunisia, 29 April - 1 May 2023, pp. 1-6.
- [15] L. Jiang, D. Costinett, "A triple active bridge DC-DC converter capable of achieving full-range ZVS", *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, Long Beach, CA, USA, 20-24 March 2016, pp. 872-879.
- [16] S. Zou, J. Lu, A. Khaligh, "Modelling and control of a triple-active-bridge converter", *IET Power Electronics*, Vol. 13, No. 5, 2020, pp. 961-969.
- [17] A. Panchbhai, A. Kumar, "Simplified control of TAB converter for scalable multibattery charging system", *International Journal of Circuit Theory and Applications*, Vol. 52, No. 8, 2024, pp. 3797-3816.
- [18] H. Tao, J. L. Duarte, M. A. M. Hendrix, "High-Power Three-Port Three-Phase Bidirectional DC-DC Converter", *Proceedings of the IEEE Industry Applications Annual Meeting*, New Orleans, LA, USA, 23-27 September 2007, pp. 2022-2029.
- [19] H. Tao, J. L. Duarte, M. A. M. Hendrix, "Line-Interactive UPS Using a Fuel Cell as the Primary Source", *IEEE Transactions on Industrial Electronics*, Vol. 55, No. 8, 2008, pp. 3012-3021.
- [20] X. Jun, Z. Xing, Z. Chong-Wei, L. Sheng-Yong, "A novel three-port bi-directional DC-DC converter", *Proceedings of the 2nd International Symposium on Power Electronics for Distributed Generation Systems*, Hefei, China, 16-18 June 2010, pp. 717-720.
- [21] S. Kurm, V. Agarwal, "Interfacing Standalone Loads With Renewable Energy Source and Hybrid Energy Storage System Using a Dual Active Bridge Based Multi-Port Converter", *IEEE Journal of Emerging and Selected Topics in Power Electronics*, Vol. 10, No. 4, 2022, pp. 4738-4748.

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia.

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to <https://ijeces.ferit.hr>. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper;
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-

in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Authors of the International Journal of Electrical and Computer Engineering Systems must transfer copyright to the publisher in written form.

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as "the Work"):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author's personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

Author/Authorized Agent

Date

CONTACT

International Journal of Electrical and Computer Engineering Systems (IJECES)
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Josip Juraj Strossmayer University of Osijek
Kneza Trpimira 2b
31000 Osijek, Croatia
Phone: +38531224600,
Fax: +38531224605,
e-mail: ijeces@ferit.hr