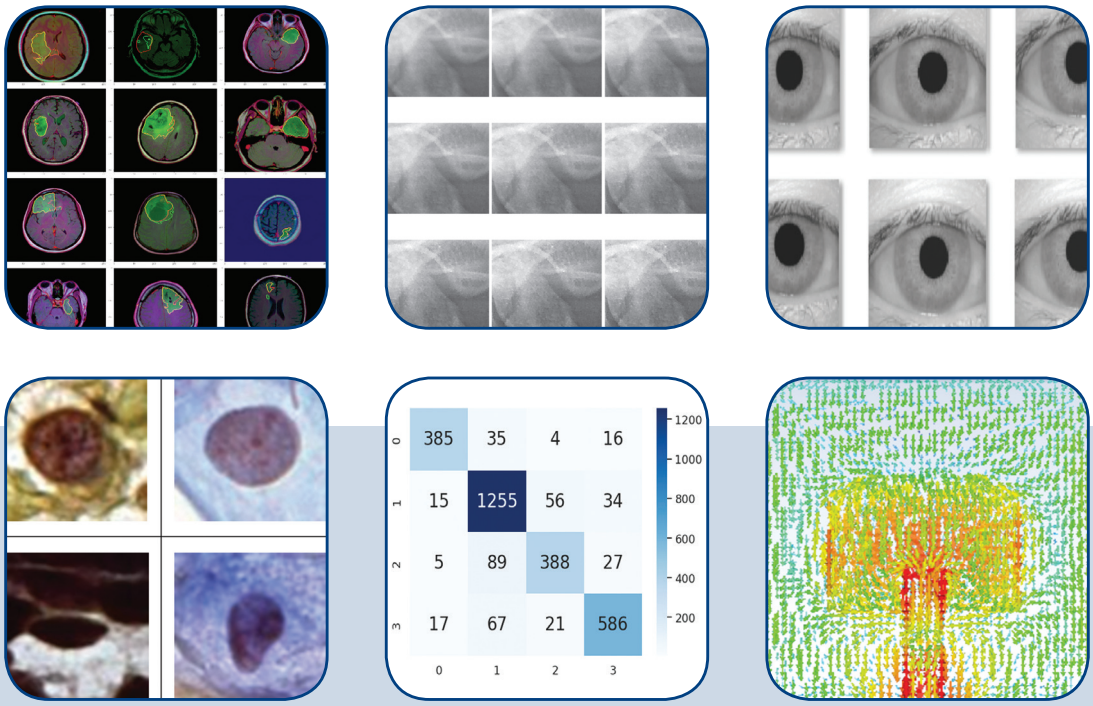
Volume 14, Number 6, 2023 ISSN 1847-6996

FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

IJECES International Journal of Electrical and Computer Engineering Systems

International Journal of Electrical and Computer Engineering Systems



ISSN: 1847-6996

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 14, Number 6, 2023 | Pages 613 - 732

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

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

TABLE OF CONTENTS

Multi-class Cervical Cancer Classification using Transfer Learning-based Optimized SE-ResNet152 model in Pap Smear Whole Slide Images

Original Scientific Paper

Krishna Prasad Battula
School of Computer Science and Engineering,
VIT-AP University, Amaravathi, Andhra Pradesh, India.
krishnaprasadbattula706@gmail.com

B. Sai Chandana
School of Computer Science and Engineering,
VIT-AP University, Amaravathi, Andhra Pradesh, India.
saichandanas869@gmail.com

Abstract *– Among the main factors contributing to death globally is cervical cancer, regardless of whether it can be avoided and treated if the afflicted tissues are removed early. Cervical screening programs must be made accessible to everyone and effectively, which is a difficult task that necessitates, among other things, identifying the population's most vulnerable members. Therefore, we present an effective deep-learning method for classifying the multi-class cervical cancer disease using Pap smear images in this research. The transfer learning-based optimized SE-ResNet152 model is used for effective multi-class Pap smear image classification. The reliable significant image features are accurately extracted by the proposed network model. The network's hyper-parameters are optimized using the Deer Hunting Optimization (DHO) algorithm. Five SIPaKMeD dataset categories and six CRIC dataset categories constitute the 11 classes for cervical cancer diseases. A Pap smear image dataset with 8838 images and various class distributions is used to evaluate the proposed method. The introduction of the cost-sensitive loss function throughout the classifier's learning process rectifies the dataset's imbalance. When compared to prior existing approaches on multi-class Pap smear image classification, 99.68% accuracy, 98.82% precision, 97.86% recall, and 98.64% F1-Score are achieved by the proposed method on the test set. For automated preliminary diagnosis of cervical cancer diseases, the proposed method produces better identification results in hospitals and cervical cancer clinics due to the positive classification results.*

Keywords*: Cervical cancer; affected tissues; image classification; Pap smear images; and loss function*

1. INTRODUCTION

The fourth most prevalent cancer in women worldwide and the seventh most common overall is cervical cancer [1]. Cervical cancer is primarily brought on by early sexual activity, poor menstrual hygiene, oral contraceptive use, smoking, a weakened immune system, early pregnancy, and sexual activity with several partners [2, 3].

The diagnosis of cervical cancers frequently uses cytopathology screening [4]. The doctor uses the brushes to extract cells from a patient's cervix during a cervical cytopathological examination, and the exfoliated cells are then placed on a glass slide [5]. If a tumor is malignant, cytopathologists examine it under a microscope to confirm this [6]. Each slide contains thousands of cells, though. As a result, inspection manually is complicated, and specialists are vulnerable to mistakes. So, a better solution is required for this issue [7, 8].

This issue is resolved by developing automatic computer-aided diagnosis (CAD) systems. Rapid and reliable pap slide analysis is possible with CAD [9]. The bedrock of early CAD was traditional machine learning and image processing techniques; the manually created features may produce poor categorization results [10, 11]. Deep learning technology has made advancements recently in several.

Sectors, including natural language processing (NLP), medical imaging, computer vision, and others [12]. Deep learning is a new type of machine learning that can automatically learn characteristics for classification. Yet, many labeled data sets are needed to train a deep learning model [13, 14]. A popular deep learning network is a Convolutional Neural Network (CNN), frequently utilized in image detection, classification, and segmentation [15, 16].

The research presented in this research is limited to various cervical cancer cases, which are challenging to categorize compared to earlier methods [17]. The highest-performing CNN model is initially chosen based on a thorough experiment [18]. Finally, it assesses the combination of performance enhancements, including optimizer selection and image enhancement [19, 20]. The generalization and high performance of DL approaches for classification are incorporated into the proposed method. It is suggested to use transfer learning to classify Pap smear images into several types of cervical cancer disorders using the deep learning-based hybrid network model.

The main contributions of the research are,

- Initially, image pre-processing is used to enhance the quality of image data by avoiding unwanted distortions. Mathematical morphology operations are used for image enhancement to increase classification performance.
- After completing a pre-processing phase, the transfer learning-based optimized SE-ResNet152 model is used for effective classification. The proposed model effectively classifies the Pap smear images into 11 classes using CRIC and SIPaKMeD dataset images.
- The reliable disease image features are successfully extracted by the proposed network model for effective classification results. The CNN hyperparameters are optimized using the DHO optimization methods, which produce results with greater accuracy in diagnosing multiple diseases.
- On the publicly available datasets, the proposed approach is assessed. The experiments are performed on the Python platform.
- The proposed approach outperforms the state efficiency concerning all other approaches, according to the experimental data.

The work is organized as follows. Part 2 of the paper discusses cutting-edge methods for classifying Pap smear images. The deep learning-based models for Pap smear image categorization that have been proposed are described in Section 3. We also cover the databases used for the experimental assessment of the effectiveness and performance of the proposed approach. Section 4 discusses the evaluation parameters and findings. Finally, the research is concluded in Section 5.

2. LITERATURE SURVEY

In this section, we review some existing deep-learning techniques for diagnoses of cervical cancer classification using Pap smear images. Desiani et al. [21] proposed the Bi-path architecture, which combined image segmentation and classification. The multiple image processing stages are proposed in this paper, including image enhancement, segmentation, and classification for recognizing and identifying cervical cancer feature on the Pap-smear images. The image quality is improved by using enhancement approaches like Normalization, CLAHE, and Adaptive Gamma Correction before segmentation. Segmentation using the CNN architecture is the first step. Using the KNN and ANN algorithms to segment data yields the second path, a classification process.

Yaman et al. [22] have suggested an approach for the early identification of cervical cancer based on an exemplar pyramid structure. This structure is mainly used for feature extraction. The research aims to categorize cervical cells in pap-smear images to detect malignancy. The model pyramid deep feature generator was created using the SIPaKMeD and Mendeley Liquid-Based Cytology (LBC) datasets. The suggested feature generator generates 21,000 features using transfer learning-based feature extraction from DarkNet19 or DarkNet53 networks in an example pyramid structure. Using Neighborhood Component Analysis (NCA), The Support Vector Machine (SVM) method classifies the 1000 features NCA chose.

A simple yet efficient ensemble method is proposed by Diniz et al. [23] for improving the classification task. To address the imbalance in the dataset, a data augmentation mechanism is applied. This research presents classification analyses of two, three, and six classes. Several pap smear image classification tasks are performed using the deep learning-based EfficientNet model.

In their proposal [24], Alquran et al. used the unique Cervical Net and Shuffle Net to develop a computer-aided cervical cancer diagnosis method. 544 resulting characteristics are retrieved automatically using pre-trained CNN fused with a new Cervical Net structure. PCA is used to reduce dimension and choose the most crucial features.

Win et al. [25] created a new computer-assisted cervical cancer screening system. The cervical cancer screening system depends on the analysis of Pap smear images. The cervical cancer screening process includes four fundamental phases. In segmenting cells, nuclei were found using an iterative shape-based technique. From the regions of segmented nuclei and cytoplasm, three significant features were extracted during the features extraction step. The approach for selecting features, RF (random forest), was employed. Bagging ensemble classifiers, such as bagged trees, boosted trees, KNN, SVM, and LD, were used to integrate the output of five classifiers in classification.

A segmentation and classification system for Pap smear images was developed by Hussain et al. [26]. The Unet model is proposed for this classification. In this model, a fully convolutional layer, densely connected blocks, and residual blocks are connected in the Unet architecture. To guarantee feature reuse-ability, the number of convolutional layers in the conventional Unet has been replaced by densely connected blocks. Residual blocks have also been included to speed up network convergence. A stacked auto-encoder-based shape representation learning model precedes the suggested approach to improve the reliability of the entire network.

The multi-class classification of cervical cells using Whole Slide Imaging (WSI) with optimal feature extraction was proposed by Bhatt et al. [27]. Conv-Net with Transfer Learning approach actualization substantiates significant Metamorphic pre- and post-neoplastic lesions diagnosis. Conv-Net with Transfer Learning approach actualization substantiates significant Metamorphic pre- and post-neoplastic lesions diagnosis. The retrieved morphological cell features can be carried over by the model recursively to future Neural Network layers for illusive learning.

3. PROPOSED METHODOLOGY

Evaluating various classification improvement strategies through an experimental process to improve the performance of automated multi-class cervical cancer classification and detection is the major concept of the proposed model. The 11 classes of cervical cancer diseases (multi-class classification) are classified in this research as ASC-H, ASC-US, SCC, HSIL, LSIL, NILM, Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate. The combination of CRIC and SIPaKMeD dataset images is used to analyze the proposed model using Pap smear images. The Data were obtained from publicly available open sources. For image pre-processing, contrast enhancement is accomplished by applying mathematical morphology processes in the Pap smear images. Then, a transfer learning-based optimized SE-ResNet152 model for classifying multiple cervical cancer diseases was proposed. The hypermeters are optimized by the DHO optimizer. The schematic diagram of the proposed method is given in Fig. 1.

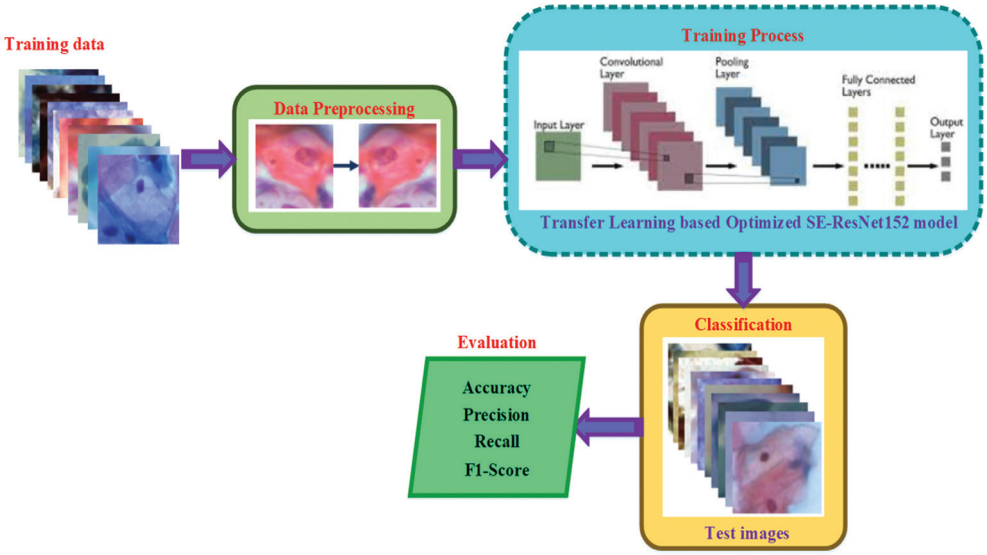


Fig. 1. Schematic diagram of the proposed methodology

3.1. IMAGE PRE-PROCESSING

Mathematical morphology has been applied to enhance contrast in the Pap smear images. The structural values of objects are the basis for how mathematical morphology methods operate. These techniques leverage interactions between classes and basic mathematical concepts to identify the components of an image and clarify spatial relationships. Two different data sets are used as input for morphological operators. In morphological operators, equations 1 and 2 define the erosion and dilation operators. This process represents the gray-level image matrix id and defines the structural element.

$$I \Theta S = \min_{m,n} \{I(x+m, y+n) - S(m,n)\} \quad (1)$$

$$I \oplus S = \max_{m,n} \{I(x-|m, y-n) + S(m,n)\} \quad (2)$$

In the erosion process, the size of the objects is reduced. Minimal information from that image is also eliminated by this erosion process. In other words, image gap size reduces, and object size rises when the dilation operator is used in reverse. As a result, the closing operator behaves similarly to applying the dilation and erosion operations to the same image in equation 3, while the opening operator acts in equation 3.

$$I \circ S = (I \Theta S) \oplus S \quad (3)$$

$$I \bullet S = (I \oplus S) \Theta S \quad (4)$$

Tiny gaps are filled in by the opening operator, who also eliminates bad relationships between artifacts and small amounts of information. Nonetheless, disk-shaped SEs are utilized more commonly than other masks for medical pictures. The size and shape of a SE are typically chosen at random.

3.2. FEATURE EXTRACTION BASED ON TRANSFER LEARNING

Extracting cervical image characteristics is a prerequisite for the multi-class classification of cervical cancer. The classifier's classification performances are directly impacted by the effectiveness of feature extraction. Traditional

deep-learning classification applications sometimes call for expensive computing resources and big datasets for training from scratch. For extracting the significant image features in this research, we utilized Modifying SE-ResNet152 models that were trained to decrease the reliance on the amount of data and accelerate model training. Textural, statistical, wavelet, and morphological features of the Pap smear images are extracted to improve the classification performance. The weights and parameters are updated by training the modified SE-ResNet152 model. For multi-class classification of cervical cancer, high-quality features of the image are extracted by fine-tuning the transfer learning model.

3.2.1. ResNet

The residual adding unit is more useful for addressing the challenge of the deeper neural network, which solves the vanishing gradients problem of the network. An entirely new trend in image recognition has been launched by ResNet. To teach desirable features, appropriate depth is still essential. The stacked layers might enhance the "levels" of features (depth). The cervical image features are effectively extracted using 152-layer ResNet (ResNet152), which increases the obtained accuracy from the increased depth and uses fine-grained and small-scale features of the Pap smear image dataset.

The residual function is represented as

$$Q(z) = P(z) - z \tag{5}$$

The parameter function is represented by P, which was directly learned through training, x is the input, and the residual function is denoted as $Q(z)$. It will be challenging to train deep networks if "simple" layers can directly fit a possible identity mapping function $P(z)=z$. The residual network is easier to optimize and converges more quickly when $P(z)=Q(z)+ z$ transformed into a residual function. Feature extraction with more extraordinary expression ability is possible by slightly increasing the network's depth.

3.2.2. SENet

The basis of SENet is its feature recalibration method, which allows it to emphasize insightful data while suppressing less helpful ones selectively. This mechanism automatically determines the value of each feature channel. Squeeze-and-excitation (SE) blocks make up the SENet architecture. The model's sensitivity to channel properties can be increased, and its performance can be significantly improved by incorporating the SE block into the current network layout. The SE-ResNet152 model's organizational structure is depicted in a schematic in Figure 5.

In ResNet152, the SE block is included for creating high-quality Pap smear image features considering the slight visual variations between various cervical cancer illnesses. Combining 1×1 and 3×3 convolutions, the residual module, as seen in Figure. 2, lowers the parameters. The dynamic relationship between channels is fitted by connecting two fully connected (FC) layers with a bottleneck structure, and the input image's global features are obtained by enabling global pooling in the network model. The first FC layer performs dimensionality reduction, followed by activation using a ReLU and restoration of the original dimension by the last FC layer. A sigmoid operation yields the normalized weights. The entire activity shown in the image can be viewed as an attention mechanism that aims to strengthen the network's capacity for representation. Based on the abovementioned processes, the proposed transfer learning-based optimized SE-ResNet152 model performance is improved for effective classification.

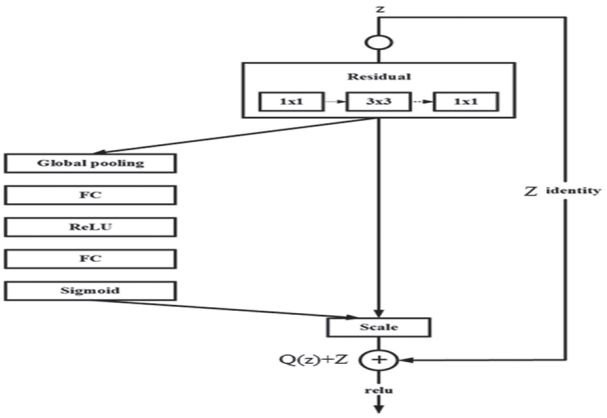


Fig. 2. SE-ResNet152 model structure example diagram

3.2.3. Loss function

Due to data imbalance, classifiers will produce poor classification results, particularly for classifications with few samples. For enhancing a loss function, the accuracy of cervical cancer identification can be significantly increased by applying a re-weighting technique.

$$CB\ (p,y) = \frac{1}{E_{n_y}} L(p,y) = \frac{1-\beta}{1-\beta^{n_y}} L(p,y) \tag{6}$$

Where n_y the ground-truth class's total sample is given by S, where x is the sample, the model's probability distribution $P=[p_1, p_2,..., P_c]$ $p_i \in [0,1]$ $L(P,y)$ is defined by the loss function. The class-balanced term is adjusted by the hyperparameter β.

The focal loss function is:

$$Focal\ Loss(z,y) = -\sum_{i=1}^{c} \left(1 - p_i^t\right)^{\gamma} \log(p_i^t) \tag{7}$$

It is possible to express the class-balanced focal loss as

$$CB\,fl\,(z,y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{c} \left(1 - p_i^t\right)^{\gamma} \log\left(p_i^t\right) \tag{8}$$

We define z_i^t as

$$z_i^t = \begin{cases} z_i, & if\ i = y \\ -z_i, & ow \end{cases} \tag{9}$$

The model's output is $z=[z_1, z_2,..., z_c]^T$ where the overall number of classes is indicated by C.

3.3. DEER HUNTING OPTIMIZATION (DHO) ALGORITHM

The way humans hunted deer served as an inspiration for the traditional DHOA. Finding the person's optimal or most productive deer hunting stance is the primary goal of traditional DHOA. Deers, also called a buck, have specific characteristics that allow them to protect themselves from predators or hunters. The distinguishing features include the capacity to detect ultrahigh-frequency noises, an excellent sense of smell, and exceptional vision.

Testing findings that demonstrate overfitting values, like in the case of the combined dataset, requires the hyperparameter tuning process. The best value is produced by the data training detection rate. Loss function values, activation function, initialization weights, learning rate value, and the number of layers and neurons must all be examined to establish the ideal model. These network model parameters are optimized for improving the accuracy and precision of the classification of Pap smear images. In each layer, the system's complexity is influenced by the number of neurons. The training and response times often rise as the number of neurons increases. Capacity in the multilayer model is controlled by the number of neurons in each layer. Automated feature engineering's complicated data representation makes adding layers beneficial to learning. Depending on how complicated the dataset is, layer addition can increase accuracy. The conventional DHOA is composed of four processes in terms of mathematics.

Phase 1: Initializing the population: Equation (10) creates the hunter population. The population of hunters and the average number of hunters are represented as H and S.

$$H = \{H_1, H_2, ...H_S\} \;\; ;1 < \mathrm{r} \leq se \tag{10}$$

Phase 2: The wind initialization and position angle: Establishing the location and wind angle is the first step in determining the ideal position for creatures to hunt deer. The wind angle is computed by Equation (11), while the position angle is computed by Equation (12). A number at random between [0, 1] is l, the present iteration is k, φ represents the wind angle, and the position angle is represented by θ notation.

$$\phi_k = 2\pi l \tag{11}$$

$$\theta_k = \phi + \pi \tag{12}$$

Phase 3: Propagation of position

The fitness function determines the optimal space, the area closest to the most optimal solution. Two roles are typically described: successor (Hsuccessor) and leader (Hlead). The propagation process makes use of the leader's position. Everyone in the community works hard to land the best positions once they are identified, and then updating the job description starts. The equation provides the hunter's encircling formula's empirical formula (13)

$$H_{k+1} = H^{lead} - M.c. \mid R \times H^{lead} - H_k \mid \tag{13}$$

In Eq (14), the current position of the hunters is represented H_k, and the upgraded position of the predators is described H_{k+1} in the following procedure. In contrast, the random number c, which ranges from 0 to 2 for considering the wind angle, is represented. This uses the random integer rd, which ranges from 0 to 1. It K_{max} gives the maximum number of iterations, and the coefficient vectors (M) are calculated below the equation,

$$M = \frac{1}{4}\log\left(K + \frac{1}{K_{\max}}\right)a \tag{14}$$

$$R=2.rd \tag{15}$$

Using position angle as a basis for propagation: During the hunter's location improvement process, this is accomplished by improving the search space by considering the position angle. The position angle is determined to increase the effectiveness of the deer hunting activity. Moreover, a parameter ds_k is derived from the difference between the visualization and wind angles for improving the position angle, as given in equation (16). Equation (17) determines the prey's vision vs_k angle.

$$ds_k = \Phi_k - vs_k \tag{16}$$

$$vs_k = \Pi/8 \; x \; rad \tag{17}$$

Equation (18) updates the position angle, and equation (19) upgrades the hunter's location by accounting for the position angle.

$$\theta_{k+1} = \theta_k + ds_k \tag{18}$$

$$H_{k+1} = H^{lead} - c. \mid \cos(\omega) \times H^{lead} - H_k \mid \tag{19}$$

The propagation process for the successor position is as follows: The encircling procedure is used to alter the vector R during the exploring stage. The vector R-value is considered smaller than one when evaluating the random search in its beginning position. Equation illustrates how updating the hunter's position is based on the successor's location (20). The search agent is chosen probabilistically if the vector R-value is less than 1; else, the best solution for improving the positions of the search agents is generated.

$$H_{k+1} = H^{successor} - M.c. \mid R \times R^{successor} - R_k \mid \tag{20}$$

Phase 4: End procedure: This is the last stage of the DHOA algorithm. The position updating procedure is repeated until the ideal location is discovered.

4. RESULT AND DISCUSSION

This section examines the many performance measurements used to assess the method, compares the proposed method to current state-of-the-art procedures, and makes comparisons. Using Python, the experimental evaluation is carried out. Each model's input requirements are fulfilled by standardizing the

size and resolution of the images. Several cervical cancer diseases are classified using the transfer learning-based optimized SE-ResNet152.

4.1. DATASETS

4.1.1. CRIC

400 conventional Pap smear images from the CRIC collection, including 11534 cells identified manually. There are six different types of images: negative for intraepithelial lesion or malignancy (NILM); low-grade squamous intraepithelial lesion (LSIL); high-grade squamous intraepithelial lesion (HSIL); squamous cell carcinoma (SCC); possibly non-neoplastic (ASC-US); atypical squamous cells of undetermined significance; cannot exclude a high-grade lesion (ASC-H); and atypical squamous cells. Including 4789 images of cervical cells that have been separated, this paper exclusively uses cropped images. Fig. 3 displays a few examples of images of the CRIC dataset.

4.1.2. SIPaKMeD

Cervical cell image classification tasks are possible using the SIPaKMeD dataset. The images collected by the camera and cropped into 4049 cervical cell images help compensate for this dataset. These images are divided into five categories: superficial intermediate, parabasal, metaplastic, koilocytotic, and dyskeratotic. Fig. 4 shows some example images of the SIPaKMeD dataset.

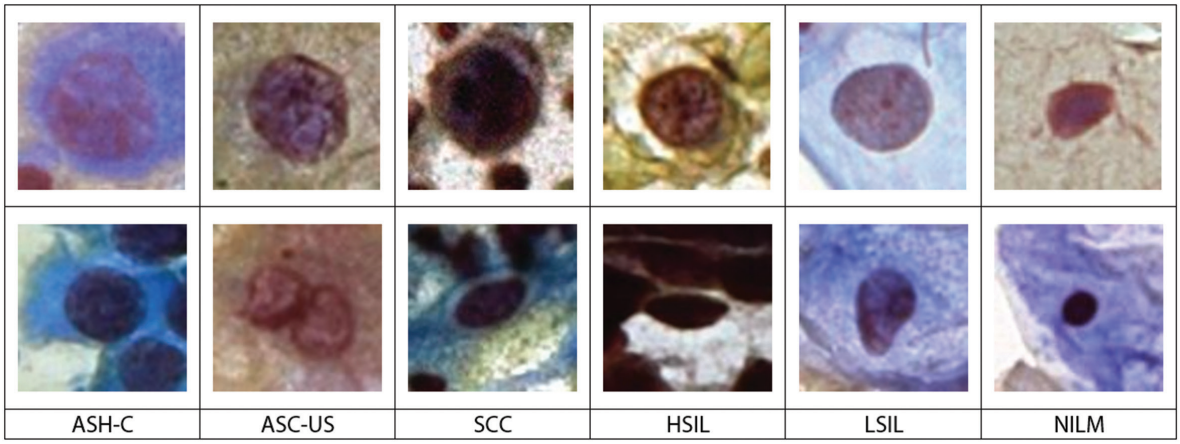


Fig. 3. Sample images of a CRIC dataset

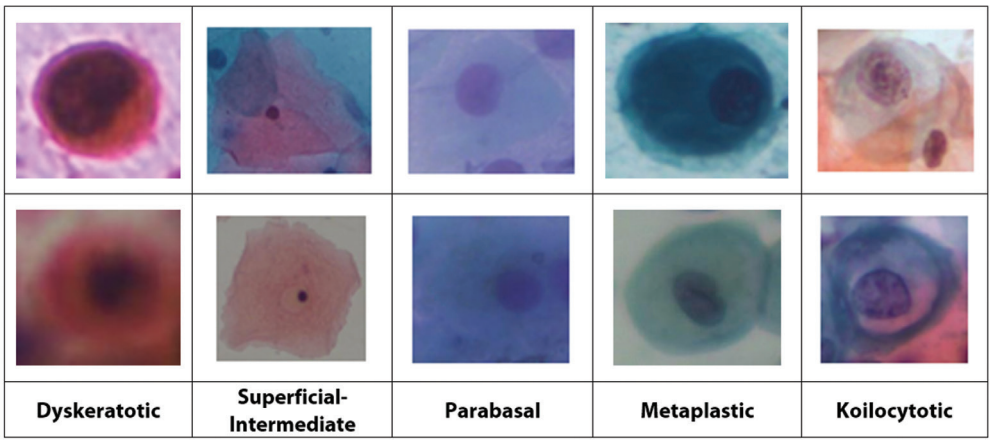


Fig. 4. SIPaKMeD dataset images

4.1.3. combination of datasets used for analysis

We integrated the CRIC and SIPaKMeD datasets for this investigation. This integration of dataset images is used for experiment analysis. As a result, we obtained 11 categories, comprising the CRIC dataset's six categories and the SIPaKMeD dataset's five categories. There are 8838 images in this combined dataset. For the test, validation, and training sets, we chose 20%, 20%, and 60%, respectively, randomly from each class of images. Table 1 displays the data configurations.

Table 1. The combined dataset's data settings

Class/Dataset	Test	Validation	Train	Total
Superficial-Intermediate	166	166	499	831
Parabasal	157	157	473	787
Metaplastic	158	159	476	793
Koilocytotic	165	165	495	825
Dyskeratotic	162	163	488	813
SLIM	518	172	172	862
LSIL	163	164	491	818

HSIL	174	175	525	874
SCC	137	137	413	687
ASC-US	148	148	446	742
ASC-H	161	161	484	806
Total	**1763**	**1767**	**5308**	**8838**

4.2. EXPERIMENTAL SETTINGS

This research experiment was performed on a local computer running Windows 10 and equipped with 32 GB of RAM. The Pytorch version we are using is 1.8.0, and we are using the Python programming language. Using the Adam optimizer's hyperparameter configuration for training the proposed network model is shown in Table 2.

Table 2. Hyperparameter settings for network training

Parameter	Value
Momentum	0.9
Decay	10-5
Learning rate	0.0002
Batch size	16
Epoch	100

4.3. EVALUATION METHODS

Calculations are made to assess the performance of the suggested model using precision, recall, F1-Score, and accuracy. The number of precisely anticipated positive samples is called True Positive (TP). There are exactly as many True Negative (TN) as predicted negative samples. False Positives (FP) are negative sample range that was expected to be positive. Moreover, False Negatives (FN) are samples that were positive but were later found to be negative. Precision is the percentage of TP in all correct predictions. The proportion of anticipated positive samples to all positive sample counts is known as a recall. A specific measure of precision and recall is the F1-Score. The proportion of correctly predicted occurrences to all occurrences is known as accuracy. The performance metrics are described as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

$$precision\ (p) = \frac{Tp}{TP + FP} \quad (22)$$

$$\text{Re}\,call\ (R) = \frac{TP}{TP + FN} \quad (23)$$

$$F1 - Score = 2 \times \frac{p \times R}{P + R} \quad (24)$$

4.4. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed model's performance is assessed throughout the test and validation phases using accuracy, F1-Score, recall, and precision calculations, as illustrated in Fig. 5. The proposed model obtains 98.95% of precision, 98.23% of recall, 98.93% of F1-Score, and 99.89% of accuracy on the validation set from the experiment results. The proposed model obtains 98.82% of precision, 97.86% of recall, 98.64% of F1-Score, and 99.68% of accuracy on the test set. The proposed model's excellent generalizability is demonstrated by the little performance difference we acquired between the validation set and the test set. The performance of the suggested model outperforms the existing approaches and yields excellent outcomes.

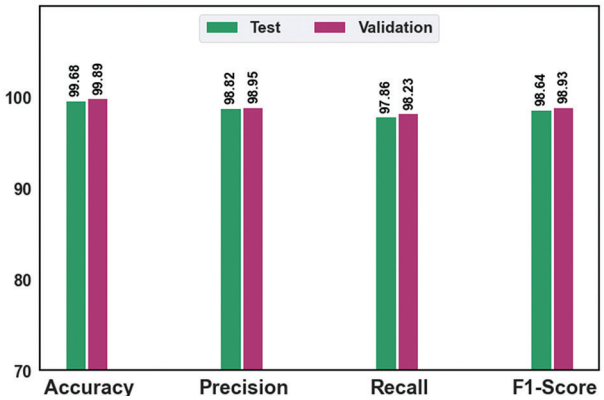
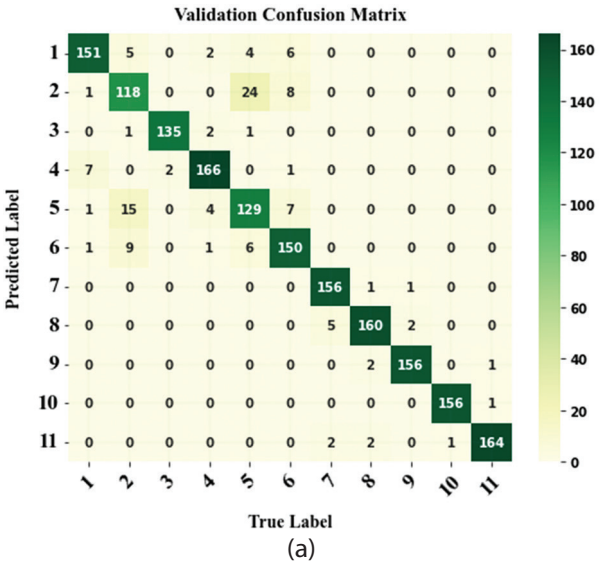
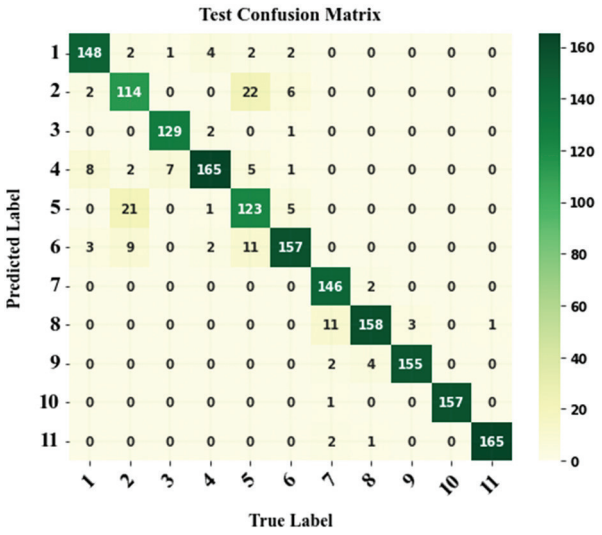


Fig. 5. The performance of the proposed transfer learning-based optimized SE-ResNet152 model

For validation and testing, Fig. 6 displays the confusion matrices using the proposed approach. The entire classification results are accurate, as seen in the validation confusion matrix. The second category's fifteen images are divided into the fifth and sixth categories, respectively, while nine are placed in the second category. The second category contains twenty-four of the five images from the fifth category. In the test confusion matrix, the result is much the same. In the sixth category, nine images from the second and twenty-one from the fifth categories were incorrectly assigned. In the fifth category, 11 images are incorrectly categorized into the sixth category, and 22 are incorrectly classified into the second category. The eighth category was incorrectly assigned to 11 images in the seventh category.



(a)



(b)

Fig. 6. The proposed model's confusion matrices for both (**a**) validation and (**b**) testing

These eleven categories are broken down into abnormal types for the first, third, fourth, seventh, and eighth categories. The proposed approach has a greater level of classification accuracy for abnormal types. Because it could mean the difference between life and death, we hope the model won't incorrectly identify cancer patients in real-world applications. Thus, the better the model's ability to accurately classify abnormal types is precisely how our suggested model performs.

4.5. COMPUTATIONAL TIME

The proposed network model is trained using 100 epochs in this experiment, and the size of the model parameters is around 120 MB. The training period is 1.3 hours long and consists of 5308 training images. Each training epoch lasts for around 82 seconds. Despite the lengthy training phase, the test lasts only 31 seconds. Each test image takes around 0.018 seconds to complete, totaling 1763 test images.

4.6. COMPARISON RESULTS

We compare the categorization outcomes of our suggested model with outcomes from previous research using Pap smear images in this section. The comprehensive analysis of the proposed transfer learning-based optimized SE-ResNet152 model in terms of accuracy is performed with the most advanced classifiers available, mentioned in Table 3.

Table 3. Comparative analysis of proposed approach and state-of-the-art approaches

References	Year	Classes	Dataset	Method	Accuracy (%)
Win et al. [25]	2020	2 classes	SIPaKMeD dataset	Bagged trees, boosted trees, KNN, SVM, and significant voting	98.27
Dhawan et al. [28]	2021	3 classes	Kaggle dataset	InceptionV3	96.1
Kanphade and Mulmule [29]	2022	-	Benchmark database	MLP with three kernels and SVM	97.14
Alquran et al. [30]	2022	7 classes	Herlev dataset	DL and cascading SVM	92
Yaman and Tuncer [22]	2022	2 classes	SIPaKMeD	SVM	98.26
			Mendeley		99.47
Proposed Model		11 classes	CRIC dataset	Transfer learning-based optimized SE-ResNet152	99.68

Yaman and Tuncer et al. [22] proposed the SVM model for pap smear image classification, it classifies two class images, and the SVM method achieves 98.26% accuracy for the SIPaKMeD dataset and 99.47% for the Mendeley dataset. Alquran et al. [30] presented a DL and cascading SVM approach to cervical cancer classification, and it classifies multiple cervical cancer diseases with an overall accuracy of 92% using the Herlev dataset. Dhawan et al. [28] used InceptionV3 to inculcate the batch normalization layer along with every activity layer and obtained an accuracy of 96.1% for cervical cancer classification. Mulmule and Kanphade [29] introduce SVM and MLP approaches, achieving 97.14% accuracy. Win et al. [25] utilize bagged trees, boosted trees, KNN, SVM, and major voting models to classify multi-class Pap smear images, achieving an accuracy of 98.27%. While compared to the prior deep learning models, the proposed transfer learning-based optimized SE-ResNet152 model achieves better results and effectively classifies the multiple cervical cancer diseases using the CRIC dataset Pap smear image dataset. The proposed transfer learning-based optimized SE-ResNet152 model has the main benefit of preventing overfitting and having no negative effects on network performance due to the classification and segmentation process.

The proposed approach provides a classification of 11 cervical cancer diseases based on Pap smear images using the largest dataset of images ever used. It is the more effective system that does so. The proposed CNN architecture has a high classification rate compared to previous methods.

The recent machine learning-based techniques like SVM, decision tree, K-means clustering, and Genetic algorithms produce better results on multi-class pap smear image classification. However, it still has a lot of limitations, including (i) Cannot separate overlapping cells, (ii) overfitting in case of many attributes, (iii) complex and time-consuming, (iv) usually detects only

round shapes, (v) the classification accuracy is low, (v1) challenging to obtain optimal parameters for nonlinear data and training is slow. These problems are solved by our proposed transfer learning-based optimized SE-ResNet152 model.

In the proposed transfer learning-based optimized SE-ResNet152 model, the effective significant related features of the image are extracted efficiently for improving the classification performance. In image classification, training deep neural networks from scratch on large datasets can be computationally expensive and time-consuming. Our transfer learning-based optimized SE-ResNet152 model helps overcome this challenge by starting with a pre-trained model that has already learned lower-level features, allowing the fine-tuning process to converge faster and with fewer training samples. There may be other options in many real-world scenarios than collecting and annotating a large dataset for a specific task. Our proposed transfer learning-based model allows the use of smaller datasets by leveraging the knowledge acquired from a larger dataset during pre-training. This leads to better generalization and improved performance, even with limited labeled data.

4.7. ADVANTAGES OF THE PROPOSED APPROACH

The suggested model successfully extrapolates characteristics that describe the inter-scale heterogeneity of the diseases, enhancing classification performance. For classifying Pap smear images, the proposed method performs better than the existing methods; the proposed method achieves a 99.68% accuracy, 98.82% precision, 97.86% recall, and 98.64% F1-Score. As a result of the proposed model's impressive classification accuracy, cervical cancer disorders are automatically detected and pre-screened.

4.8. DISADVANTAGES OF THE PROPOSED APPROACH

The experimental results indicate that the proposed approach can classify multiple cervical cancer diseases. However, this field still requires more research to be done. The proposed approach offered greater accuracy when compared to the other models. Here, a few of the limitations are highlighted,

- The proposed method's drawback is that concatenated features' dimensions are more significant. It may be possible to minimize this set of features in the future by using a feature reduction strategy.
- Transfer learning models tend to be complex, with many layers and parameters. We will use hybrid hyperparameter optimization for effective parameter tuning and reduction.

5. CONCLUSION

The transfer learning-based optimized SE-ResNet152 model is a deep learning-based framework proposed for cervical cell categorization tasks in this paper. Using a huge dataset of images, the proposed model was also tested and trained. For noise removal and contrast enhancement, these images are pre-processed initially. The transfer learning-based SE-ResNet152 model with a DHO optimizer is used to analyze this image dataset to classify and identify the multiple cervical cancer diseases (11 classes). The distinction between Pap smear images with very high accuracy, sensitivity, and sensitivity with a considerable decrease in time taken for detection and epochs according to the proposed model results. The proposed transfer learning-based optimized SE-ResNet152 model effectively classifies the Pap smear images with an overall accuracy of 99.68%. Experiment results on segmentation benchmarks demonstrated that the proposed model outperformed several existing deep-learning network models by a large margin and gained comparable performance with other typical cervical cancer classification approaches.

To enhance the efficiency of the proposed approach, we will attempt to employ different new model configurations in the future. Along with updating the module structure, improving the capacity to extract features is possible. In the future, different data pre-processing methods, including random cropping and color jittering, will be used to enhance the model's generalization abilities.

6. REFERENCE

[1] W. William, A. Ware, A. H. Basaza-Ejiri, J. Obungoloch, "A pap-smear analysis tool (PAT) for the detection of cervical cancer from pap-smear images", Biomedical engineering online, Vol. 18, 2019, pp. 1-22.

[2] P. Wang, J. Wang, Y. Li, L. Li, H. Zhang, "Adaptive pruning of transfer learned deep convolutional neural network for classification of cervical pap smear images", IEEE Access, Vol. 8, 2020, pp. 50674-50683.

[3] E. Hussain, L.B. Mahanta, C. R. Das, R. K. Talukdar, "A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network", Tissue and Cell, Vol. 65, 2020, p. 101347.

[4] D. Riana, Y. Ramdhani, T. P. Rizki, A. N. Hidayanto, "Improving hierarchical decision approach for single image classification of pap smear", Inter-

national Journal of Electrical and Computer Engineering, Vol. 8, No. 6, 2018, p. 5415.

[5] W. Liu et al. "CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron", Pattern Recognition, Vol. 130, 2022, p. 108829.

[6] M. A. Mohammed, F. Abdurahman, Y. A. Ayalew, "Single-cell conventional pap smear image classification using pre-trained deep neural network architectures", BMC Biomedical Engineering, Vol. 3, No. 1, 2021, p. 11.

[7] S. Fekri-Ershad, "Pap smear classification using combination of global significant value, texture statistical features and time series features", Multimedia Tools and Applications, Vol. 78, No. 22, 2019, pp. 31121-31136.

[8] R. Gupta, A. Sarwar, V. Sharma, "Screening of cervical cancer by artificial intelligence based analysis of digitized papanicolaou-smear images", International Journal of Contemporary Medical Research, Vol. 4, No. 5, 2017, pp. 2454-7379.

[9] D. Selvathi, W. R. Sharmila, P. S. Sankari, "Advanced computational intelligence techniques based computer aided diagnosis system for cervical cancer detection using pap smear images", Classification in BioApps: Automation of Decision Making, Lecture Notes in Computational Vision and Biomechanics, Vol. 26, Springer, Cham., 2018, pp. 295-322.

[10] V. Kumararaja, K. Deepa, "Pap smear image classification to predict urinary cancer using artificial neural networks", Annals of the Romanian Society for Cell Biology, Vo. 25, No. 2, 2021, pp. 1092-1098.

[11] C. W. Wang et al. "Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning", Scientific Reports, Vol. 11, No. 1, 2021, p. 16244.

[12] P. V. Mulmule, R. D. Kanphade, D. M. Dhane, "Artificial intelligence-assisted cervical dysplasia detection using papanicolaou smear images", The Visual Computer, Vol. 39, 2022, pp. 1-12.

[13] P. Wang, L. Wang, Y. Li, Q. Song, S. Lv, X. Hu, "Automatic cell nuclei segmentation and classification of cervical Pap smear images", Biomedical Signal Processing and Control, Vol. 48, 2019, pp. 93-103.

[14] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, K. G. Kim, "Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images", Scientific Reports, Vol. 11, No. 1, 2021, pp. 1-11.

[15] B. Nithya, V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction", SN Applied Sciences, Vol. 1, 2019, pp. 1-16.

[16] H. Chen et al. "CytoBrain: cervical cancer screening system based on deep learning technology", Journal of Computer Science and Technology, Vol. 36, 2021, pp. 347-360.

[17] V. Sellamuthu Palanisamy, R. K. Athiappan, T. Nagalingam, "Pap smear based cervical cancer detection using residual neural networks deep learning architecture", Concurrency and Computation: Practice and Experience, Vol. 34, No. 4, 2022, p. e6608.

[18] Z. K. Şenturk, U. Süleyman, "An improved deep learning based cervical cancer detection using a median filter based pre-processing", Avrupa Bilim ve Teknoloji Dergisi, Vol. 32, 2022, pp. 50-58.

[19] M. Zhao, H. Wang, Y. Han, X. Wang, H. N. Dai, X. Sun, J. Zhang, M. Pedersen, "Seens: Nuclei segmentation in pap smear images with selective edge enhancement", Future Generation Computer Systems, Vol. 114, 2021, pp. 185-194.

[20] A. Khamparia, D. Gupta, V. H. C. de Albuquerque, A. K. Sangaiah, R. H. Jhaveri, "Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning", The Journal of Supercomputing, Vol. 76, 2020, pp. 8590-8608.

[21] A. Desiani, B. Suprihatin, S. Yahdin, A. I. Putri, F. R. Husein, "Bi-path Architecture of CNN Segmentation and Classification Method for Cervical Cancer Disorders Based on Pap-smear Images", IAENG International Journal of Computer Science, Vol. 48, No. 3, 2021.

[22] O. Yaman, T. Tuncer, "Exemplar pyramid deep feature extraction based cervical cancer image classification model using pap-smear images", Biomed-

ical Signal Processing and Control, Vol. 73, 2022, p. 103428.

[23] N. D. Diniz et al. "A deep learning ensemble method to assist cytopathologists in pap test image classification", Journal of Imaging, Vol. 7, No. 7, 2021, p. 111.

[24] H. Alquran, M. Alsalatie, W. A. Mustafa, R. A. Abdi, A. R. Ismail, "Cervical Net: A Novel Cervical Cancer Classification Using Feature Fusion", Bioengineering, Vol. 9, No. 10, 2022, p. 578.

[25] K. P. Win, Y. Kitjaidure, K. Hamamoto, T. Myo Aung, "Computer-assisted screening for cervical cancer using digital image processing of pap smear images", Applied Sciences, Vol. 10, No. 5, 2020, p. 1800.

[26] E. Hussain, L. B. Mahanta, C. R. Das, M. Choudhury, M. Chowdhury, "A shape context fully convolutional neural network for segmentation and classification of cervical nuclei in Pap smear images", Artificial Intelligence in Medicine, Vol. 107, 2020, p. 101897.

[27] A. R. Bhatt, A. Ganatra, K. Kotecha, "Cervical cancer detection in pap smear whole slide images using convnet with transfer learning and progressive resizing", PeerJ Computer Science, Vol. 7, 2021, p. e348.

[28] S. Dhawan, K. Singh, M. Arora, "Cervix image classification for prognosis of cervical cancer using deep neural network with transfer learning", EAI Endorsed Transactions on Pervasive Health and Technology, Vol. 7, No. 27, 2021.

[29] P. V. Mulmule, R. D. Kanphade, "Supervised classification approach for cervical cancer detection using Pap smear images", International Journal of Medical Engineering and Informatics, Vol. 14, No. 4, 2022, pp. 358-368.

[30] H. Alquran, W. A. Mustafa, I. A. Qasmieh, Y. M. Yacob, M. Alsalatie, Y. Al-Issa, A.M. Alqudah, "Cervical cancer classification using combined machine learning and deep learning approach", Computers, Materials and Continua, Vol. 72, No. 3, 2022, pp. 5117-5134.

Effective Brain Tumor Classification Using Deep Residual Network-Based Transfer Learning

Original Scientific Paper

D. Saida
Department of Computer Science & Engineering,
University College of Engineering, Osmania University,
Hyderabad, Telangana 500007, India
saidan.dhana@gmail.com

KLSDT Keerthi Vardhan
Department of Computer Science & Engineering,
Siddhartha Institute of Engineering & Technology,
Ranga Reddy-501510, Telangana, India
keerthivardhan@siddhartha.ac.in

P. Premchand
Department of Computer Science & Engineering,
University College of Engineering, Osmania University,
Hyderabad, Telangana 500007, India
Sairadha.rajputh@gmail.com

Abstract *– Brain tumor classification is an essential task in medical image processing that provides assistance to doctors for accurate diagnoses and treatment plans. A Deep Residual Network based Transfer Learning to a fully convoluted Convolutional Neural Network (CNN) is proposed to perform brain tumor classification of Magnetic Resonance Images (MRI) from the BRATS 2020 dataset. The dataset consists of a variety of pre-operative MRI scans to segment integrally varied brain tumors in appearance, shape, and histology, namely gliomas. A Deep Residual Network (ResNet-50) to a fully convoluted CNN is proposed to perform tumor classification from MRI of the BRATS dataset. The 50-layered residual network deeply convolutes the multi-category of tumor images in classification tasks using convolution block and identity block. Limitations such as Limited accuracy and complexity of algorithms in CNN-based ME-Net, and classification issues in YOLOv2 inceptions are resolved by the proposed model in this work. The trained CNN learns boundary and region tasks and extracts successful contextual information from MRI scans with minimal computation cost. The tumor segmentation and classification are performed in one step using a U-Net architecture, which helps retain spatial features of the image. The multimodality fusion is implemented to perform classification and regression tasks by integrating dataset information. The dice scores of the proposed model for Enhanced Tumor (ET), Whole Tumor (WT), and Tumor Core (TC) are 0.88, 0.97, and 0.90 on the BRATS 2020 dataset, and also resulted in 99.94% accuracy, 98.92% sensitivity, 98.63% specificity, and 99.94% precision.*

Keywords*: Brain Tumor Segmentation, Convolutional Neural Network, Deep Residual Network, Magnetic Resonance Images, U-Net Architecture*

1. INTRODUCTION

The segmentation of brain tumors is the phenomenon of detecting the tumor area and the spread of tumor regions like active tumorous tissue, edema tissue, and necrotic tissue. The more complicated deep learning tasks are handled and levels of the tumor presented in the brain are detected by the deep residual networks [1]. A fully convoluted CNN is required to extract features of the whole image and small patches in the segmentation process. The CNN-based ResNet50 framework is used to detect the initial stage of cancer nodes in the classification block and also to identify and classify the tumor in brain images efficiently [2]. Automatic Brain Tumor Segmentation is applied to MRI Images to monitor the improving growth rate of tumors in the brain which makes it simple to improve the survival rate of the patient. Automatic brain tumor segmentation repeatedly identifies the tumor growth and gives quick results [3]. The shape, size and localization of brain tumors are evaluated by multi-modalities without any high ionization radiation effect on patients. A Computer Assisted Diagnosis (CAD) system evaluates the shape of the tumor and landmarks to predict the growth of the tumor and reduce the effect of a brain tumor [4, 5].

The differentiation of brain tumors from normal tumors causes difficulty in brain tumor segmentation and also consists of a high degree in shape, patient extension, and area. To capture the spatial information from far away at different resolutions, a multi-scale 3 Dimensional (3D) U-Nets architecture is used and the 3D depth-wise separable convolution is involved in it to reduce the cost of computation [6]. The image was captured to evaluate the information that existed in it but while capturing the data a lot of noise like pepper noise, salt & speckle noise, and Gaussian noise are reduced by using a modified iterative median filter technique and for input MRI a homomorphic wavelet filter is used [7, 8]. The noise was removed and sent to the inceptionv3 model for extracting the features, where required features are extracted using Non-Dominated Sorting Genetic Algorithm (NSGA). The brain tumor regions are accurately described and the tumor is separated from normal brain tissues by brain tumor segmentation to give the information which is used for treatment planning and diagnosis [9, 10].

The main contribution of the proposed model is evaluated as follows:

- The MRIs are preprocessed initially and performed augmentation to feed the image data to the network for segmentation and classification tasks.
- A Deep Residual Network (ResNet-50) to a fully convoluted CNN is proposed to perform tumor classification from MRI of the BRATS dataset.
- The 50-layered residual network deeply convolutes the multi-category of tumor images in classification tasks using convolution block and identity block.

This research paper includes the related works in Section 2 and Section 3 explains the proposed methodology. The results evaluated by the proposed model are given in Section 4 and Section 5 shows the comparative analysis. The conclusion of this research paper is included in Section 6.

2. RELATED WORKS

Punn and Agarwal [11] presented multi-modalities fusion, tumor extractor, and segmentation to perform tumor segmentation which are the components of a 3D deep neural network. The 3D inception U-net model was used to find the tumor patterns when the extractor component of the tumor moves through the fused images. The necrosis, edema, background, non-enhancing tumor, and enhancing tumor were the target classes used to divide each tumor region into the CT, WT, and ET. Based on the coefficient of dice and index of the Jaccard, the loss function of weighted segmentation was proposed to reduce the problem of class imbalance. The advantages held by using multi-modalities are inception convolutions, segmentation loss function, and 3D U-Net architecture which performed well on BraTS2017 and BraTS2018 datasets. The limitations of multi-modalities did not apply to real-time data fusion, and cascading. Further, it was extended to biomedical image analyses like measuring disease and image registration.

Ullah et al. [12] presented an image quality hypothesis used for the performance of classification in a statistical approach in the preprocessing stage. An updated image technique was used for better results which consist of removing noise using a median filter, contrast enhancement using the histogram equalization technique, and converting the image from grayscale to RGB (red, green, and blue). The Discrete wavelet transform was used to extract the features from MRI and reduced by color moments like mean, skewness, and standard deviation. The MRIs were classified into malignant and benign efficiently using an advanced Deep Neural Network. The drawbacks resulting in it do not apply to larger datasets, including high disease rate, enhancing the image is not performed well, and more time for computation is consumed.

Rehman et al. [13] presented a machine learning technique on Fluid-attenuated Inversion Recovery (FLAIR) scans of MRI. The Gabor filter bank was used for creating text on-map images, bilateral filtering was used for removing the noise and the features were extracted from superpixels. The main advantages of using machine learning techniques were the computational cost of image segmentation was reduced in small regions, and the performance of low-level features was increased by the segmentation of superpixel which was done on texton-map images. The limitations of this work were, stable features and small-size regions were difficult to compute and more time was consumed, difficult to build the model when samples were insufficient and the spatial data was not considered.

Amin et al. [14] presented a fusion process that was used to combine the texture and structural data of four MRI sequences T1, T1C, T2, and Flair when the brain tumor was detected. The process of fusion was done by a Discrete Wavelet Transform (DWT) with Daubechies wavelet kernel, compared to a single sequence of MRI as it provides more information on the tumor region. To remove the noise content a Partial Differential Diffusion Filter (PDDF) and segmentation tumor were detected by a global thresholding method which fed to the proposed CNN model for classifying tumor and non-tumor regions. The fused images give better results and are further extended to other modalities like Positron Emission PET and CT images were analyzed using classified results. The main drawback of using fusion images was the spatial distortion produced affects further processes like classification.

Zhang et al. [15] proposed a mechanism of using multiple encoders in segmenting brain tumors from 3D MRI to address the limitation of implementing 2D images in tumor segmentation. A new loss function has been introduced in this approach to solve the issues of voxel imbalance. The performance measures were evaluated on BRATS 2020 dataset and achieved promising dice scores with better specificity and sensitivity. However, this approach has a limitation of not

implementing better data-enhancing techniques and this can be resolved in future by enhancing generalization of the model and preventing overfitting issues.

Sharif et al. [16] proposed a four-phase brain tumor classification model that includes: 1) lesion enhancement, 2) extraction and selection of features for classification using a pre-trained inceptionV3 model and a Non-Dominated Sorted Genetic Algorithm (NSGA), 3) localization using YOLOv2-inception model, and 4) segmentation using McCulloch's Kapur entropy approach. To address the limitation of low prediction rate of brain tumors in existing CNN models that used VGG-16, SVM, and FCM approaches in tumor classification. Three validation datasets namely: BRATS 2018, BRATS 2019, and BRATS 2020 are considered in validating the performance of the proposed model in terms of accuracy, specificity, sensitivity and dice scores. However, this model is not suitable for classifying all types of tumors which is a limitation of this work and this can be resolved by implementing quantum computed algorithms for various brain tumor classification.

Sasank and Venkateswarlu [17] developed a brain tumor segmentation model to address the limitation of the existing Lattice Boltzmann Method (LBM) in tumor segmentation. It was stated that the randomly chosen parameters of LBM affect the performance of the tumor growth model. Hence to overcome this, a Modified Sunflower Optimization (MFSO) based LBM approach was proposed which chooses optimal parameters to improve the tumor growth model's performance. A Scalable Range Adaptive Bilateral Filter (SCRAB) method was used in the pre-processing stage for noise reduction and to improve edges. A fractal and multi-fractal Brownian motion (mBm) method was implemented for feature extraction. The proposed approach was validated on BRATS 2018, BRATS 2019, and BRATS 2020 benchmark datasets and achieved better results with high accuracy, specificity, sensitivity, precision, recall, and F1-score. However, this approach was only employed for segmenting tumor cells in future scan data by referring to the density information of the existing scan data. In the future, effective feature extraction techniques can be implemented in detecting tumor tissues from the first scan point.

Ullah et al. [18] proposed an automatic brain tumor segmentation model that used multiscale residual attention-UNet (MRA-UNet) to overcome the limitations of laborious and time-consuming processes in manual detection of brain tumors. In order to maintain the sequential information, MRA-UNet used three consecutive slices as its input. By using multiscale learning in a cascade way, it was able to make use of the adaptive region of interest scheme and precisely segment enhanced and core tumor regions. This suggested model achieved better dice scores when validated on BraTS2017, BraTS2019, and BraTS2020. However, the proposed model was highly sensitive to noise and also performed over segmentation which is an undesired limitation of this work.

Ullah et al. [19] proposed a Multitask semi-supervised learning framework that used auxiliary tasks for which adequate data was publicly available. To maximize the advantages of multi-task learning, the adversarial autoencoder (AAE) was used, as it has a significant capacity to acquire powerful and discriminative features. Additionally, the semi-supervised learning, which incorporated a discriminative component in an unsupervised AAE training pipeline, was made possible by the coupling of the supervised classification networks with the unsupervised AAE. This method was evaluated using the publicly available COVIDx data and achieved greater accuracy and better generalization ability through a cross dataset validation. However, the proposed model was unstable for training the process which was a limitation of the work.

Solanki et al. [20] suggested a model to detect brain tumors from 2D magnetic resonance images of the brain by utilizing a hybrid deep learning technique. This process is combined with both conventional classification methods and deep learning classification methods such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Multi-layer Perceptron (MLP), Logistic Regression (LR), and Naive Bayes (NB) for performing conventional phase of categorization. The output results demonstrated that SVM produced the most accurate results and later CNN was compared with conventional classifier techniques, which produced significant performance when evaluated on BRATS and MICCAI datasets. Standard investigative methods were not employed to look into the relationship between stroke and brain tumors. It is clear that the conventional system for the diagnosis of brain tumors and strokes lacks automated methods for the identification and segmentation of both brain tumors and strokes.

When samples were insufficient and spatial data was not considered, it was difficult to create the model and took more time. It is challenging to segment brain tumors since they differ from other tumor types to a significant degree in shape, patient extension, and area. The motivation of this work is to develop Deep Residual Network (ResNet-50) to a fully convoluted CNN to perform tumor classification from MRI of the BRATS dataset. The 50-layered residual network deeply convolutes the multi-category of tumor images in classification tasks using convolution block and identity block. A multi-scale 3D U-Nets architecture is also utilized to gather spatial data from distant objects at various resolutions, and the 3D depth-wise separable convolution is integrated into it to lower the computational cost. The summary of the related works is given in Table 1.

To overcome the limitations of existing brain tumor segmentation and classification models, a Deep residual neural network based on transfer learning is proposed in this manuscript for effective brain tumor classification. The proposed model achieves high accuracy by introducing ResNet for tumor classification and also minimizes the problem of vanishing gradient.

Table 1. Summary of Related works

Author	Methodology	Advantages	Limitations
Punn and Agarwal [11]	Multi-modality encoded fusion with 3D inception U-net and decoder model for brain tumor segmentation	The advantages of using multi-modalities are inception convolutions, segmentation loss function, and 3D U-Net architecture and performed well on BraTS2017 and BraTS2018 datasets	More computational time, with a slowed-down pace of learning in the middle of the layers of the network.
Ullah et al. [12]	A hybrid image enhancement-based brain MRI images classification technique	Accurate categorization of malignant and benign classes.	This method is not suitable for larger datasets. Poor Image enhancement and computation time.
Rehman et al. [13]	Texture-based localization of a brain tumor from MRI images by using a machine learning approach	the computational cost of image segmentation was reduced in small regions, and the performance of low-level features was increased by the segmentation of superpixel which was done on texton-map images	Stable features and small size regions were difficult to compute and more time was consumed, difficult to build the model when samples were insufficient and the spatial data was not considered
Amin et al. [14]	Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network	Spatial degradation and spatial distortions generated in the fused image were reduced.	Shift sensitivity, poor directionality, lack of phase information.
Zhang et al. [15]	ME-Net: a multi-encoder net framework for brain tumor segmentation	This method can well solve the problem of unbalanced foreground and background voxels that often occur in segmentation tasks.	The disadvantage of this method was that it did not consider spatial information and was sensitive to noise and gray-scale unevenness.
Sharif et al. [16]	Active Deep Neural Network Features Selection for Segmentation and Recognition of Brain Tumors using MRI Images	Complex operations are performed simultaneously	High computational time and presence of redundant features.
Sasank and Venkateswarlu [17]	Modified Sunflower Optimization (MFSO) based LBM approach for brain tumor segmentation.	Noise reduction and improved edges	The approach was only employed in segmenting tumor cells in future scan data by referring the density information of existing scan data
Ullah et al. [18]	automatic brain tumor segmentation model that used multiscale residual attention-UNet (MRA-UNet)	This method has an advantage of Less computational cost and simple architecture	highly sensitive to noise and also has over segmentation
Ullah et al. [19]	Multi-task semi-supervised adversarial autoencoding for COVID-19 detection based on chest X-ray images	significant capacity to acquire powerful and discriminative features	The process of training was unstable which resulted in high computational time
Solanki et al. [20]	Detection of brain tumors from 2D magnetic resonance images of the brain by utilizing a hybrid deep learning technique.	It is observed that the 5-layer CNN approach obtains the best outcome compared to the other distributions when it is trained with a learning rate of 0.001, an epoch of 10, and a training duration of 15 secs.	It is clear that the conventional system for the diagnosis of brain tumors and strokes lacks automated methods for the identification and segmentation of both brain tumors and strokes.

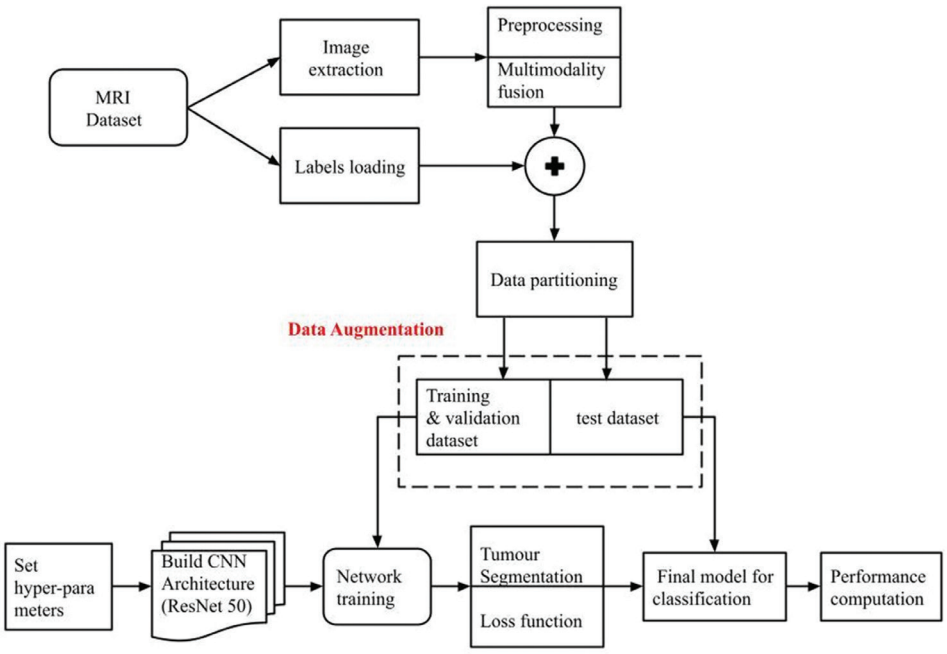


Fig. 1. Block diagram of the proposed model

3. METHODOLOGY

The proposed model for brain tumor segmentation and classification is elaborated in this section by showcasing the block diagram of the proposed architecture. The block diagram shown in Fig. 1 explains the procedure for the classification of MRIs taken from the BRAT 2020 dataset using data preprocessing, data augmentation, CNN-based residual network (ResNet-50), and segmentation.

3.1. DATASET

The BRAT 2020 dataset is considered for pre-processing which consists of MRI out of which 369 images are trained set, validation set consists of 125 images, and 166 images from the test set multimodal brain MRIs of tumor patients having high-grade (undifferentiated tumors) or low-grade (well-differentiated tumors) gliomas are included. Every patient receives T1 & T2 weighted, post-contrast T1 weighted, and FLAIR images. The MRI image has dimensions 240 x 240 x 155, where 155 represents the number of slices. The voxel spacing was the same across all MRI images, at (1 x 1 x 1) mm3. The size of each MRI input image was fixed throughout the process. The dataset includes expert annotation for each subject (i.e. ground fact). FLAIR, T1-Contrast enhanced (T1C), T1, and T2, along with ground truth are the four kinds of MRI modalities. The summary of the BRATS 2020 dataset is shown in Table 2.

Table 2. Summary of BRATS 2020 dataset

Trained set of images	369
Validation set of images	125
Test set of images	166
Types of tumors	Low-grade and high grade
Categories	T1 & T2 weighted, post contrast T1 weighted, and FLAIR images
The MRI image dimensions	240×240×155

3.2. PRE-PROCESSING STAGE

The reduction of original image size for better performance of the network is accomplished by less dimensionality and computations in the image. All of this is involved in the pre-processing stage. MRI images from the BRATS dataset are then extracted by multimodality fusion after labeling the image data as T1, T2, post-contrast T1 weighted, and FLAIR. A Gaussian filter is applied to the image and then finally image data augmentation is performed.

3.2.1. Multi-modalities Fusion

After the data was collected from BRAT 2020 using the filter, each MRI sequence is processed in the inception module (IM), and subsequently, the two sequence concatenation of T1 with T1c and T2 is accomplished with FLAIR resulting in modalities. Preprocessed modality is processed by hierarchical inception blocks instead of using batch axis for direct fusion, which is a multi-modality fusion component. The encoded fused output is then sent to the tumor extractor for deep pattern discovery in the lesion region. The proposed MRI fusion approach is claimed to be effective in improving segmentation outcomes by alternating the architectural designs after extensive tests and the input image is shown in Fig. 2.

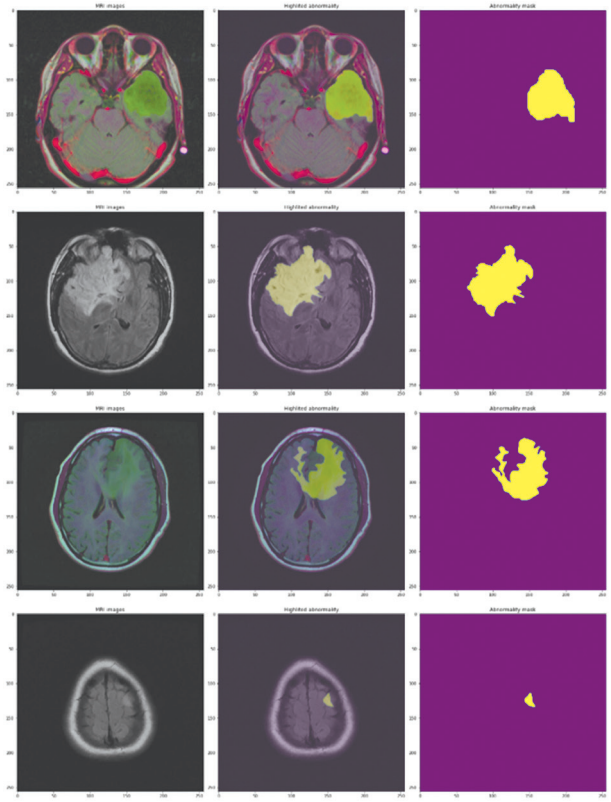


Fig. 2. Input image

3.2.2. Noise removal using a Gaussian filter

The Gaussian filter is one of the most popular filters for noise removal that is majorly used in the pre-processing stage for the detection of brain tumors. The filter removes the noisy data in the MRI scans. The majority of noise reduction techniques do not affect the image sharpness, but the fine details and smoothness across boundaries are reduced in these images by the Gaussian filter. The noise detection with noisy and noise-free images is mathematically represented in Eq. (1)

$$\mathcal{N}(x,y) = \begin{cases} 1, & if\ \mathcal{B}(x,y)\ noisy \\ 0, & if\ \mathcal{B}(x,y)\ noise\ free \end{cases} \quad (1)$$

Where, $N(x,y)$ is a noise detection image, and $B(x,y)$ is the input image.

After multimodality fusion and labeling the data, the data is filtered to remove noise by using the Gaussian function as shown in Eq. (2)

$$G\ (x,y) = \frac{1}{2\pi\delta^2} e^{-\frac{x^2+y^2}{\delta^2}} \quad (2)$$

Where x and y in the above function represent the location in the Gaussian template.

3.2.3. Data Augmentation

To train the network on unsorted data, the data is shuffled and split into training (80%), and test (20%) datasets. The augmentation technique is applied to the trained dataset to generate new data points and further enhance the model's robustness. Various data augmentation approaches, such as flipping and rotation, are utilized to enable the architecture to understand the variances during training. The degrees or angles for data augmentation flip and rotation methodology are 30, 45, 60, 90, 120, 180, 270, and 360.

3.3. TUMOR SEGMENTATION

After multi-modalities fusion, the segmentation of a brain tumor is processed for separating the tumor from healthy brain tissues. This information is helpful for diagnosis and treatment planning in typical clinical practices. U-Net, which consists of an enlarged up-sampling path for deep feature maps, significantly improves the segmentation performance of medical images. The 3D nature of multimodal MRI poses problems like memory, computation constraints and class imbalance when using the U-Net structure directly in brain tumor segmentation, which is shown in Fig. 3. A path for expanding input images and a skip connection is established to improve segmentation performance that integrates cropped features, and maps from the encoder-decoder network. The residual mapping of adding input features is given by Eq. (3)

$$y(x)=F(x)+x \tag{3}$$

Where $y(x)$ and x are input and output vectors, $F(x)$ is the residual mapping.

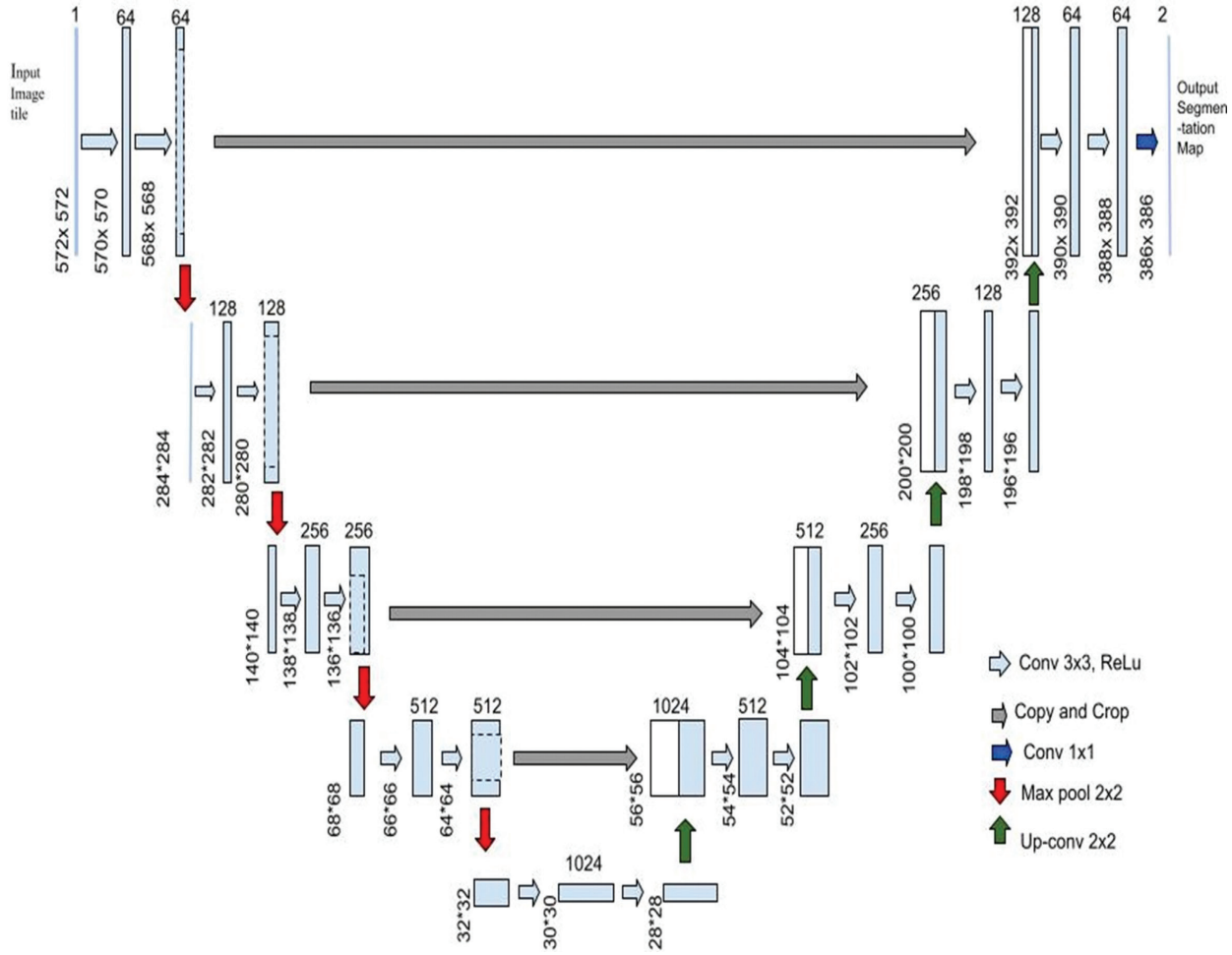


Fig. 3. U-Net architecture

The size of the tumor varies from person to person over time, hence classifying the tumor from various images is a difficult task for radiologists. To extract the features from the MRI images of the trained dataset, the U-Net-based fully convolutional networks were applied. U-Net is used for the quick segmentation process by applying convolutional blocks along with max-pooling to encode the input image into a feature representation. Segmentation using the U-Net model gives high performance in detecting the location and size of the tumor within a lesser time period. The output image of the process is shown in Fig. 4.

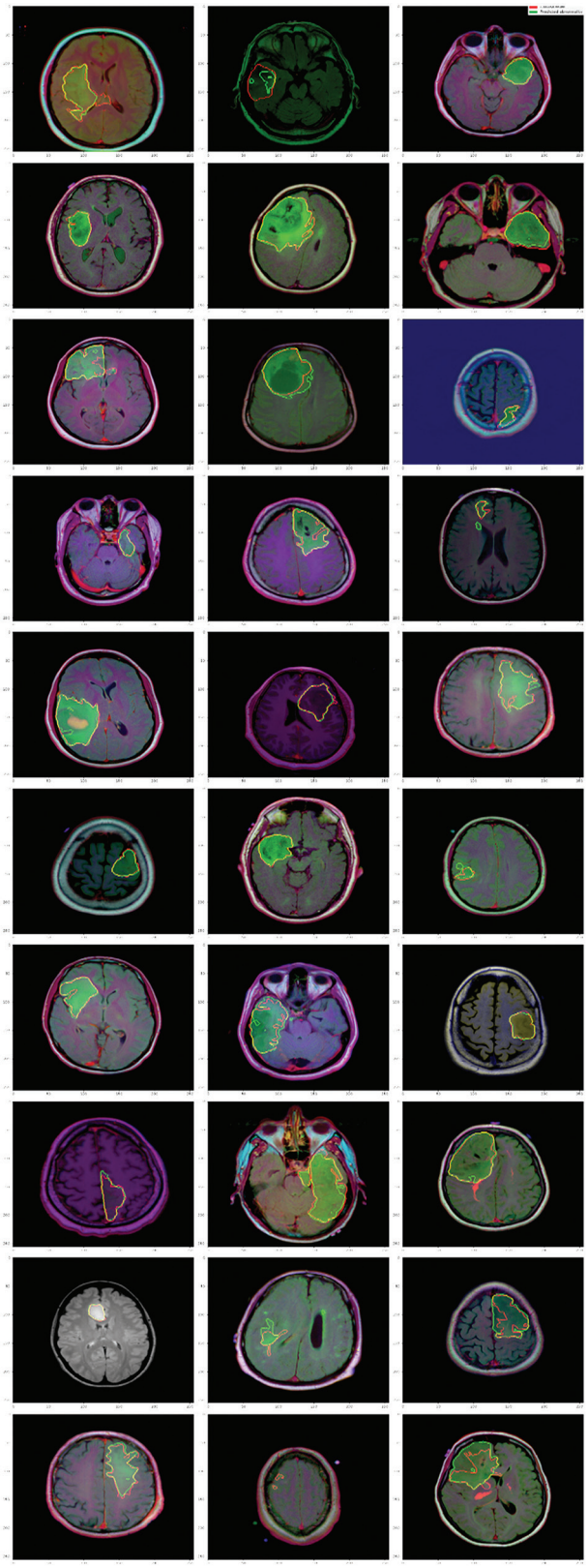


Fig. 4. Output image

3.3.1. Loss Function

The final loss function is a compound loss that is formed as the weighted sum of a few widely used segmentation loss functions. The dice loss is calculated to find the overlap between ground truth values and predicted values.

- **Dice Loss:** Dice loss is the special region-based loss that is used in error reduction that occurs due to overlap of prediction and the ground truth. To use Dice Loss, the Softmax activation function is applied to all pixel values between 0 and 1. Dice loss has the advantage of being able to deal with class imbalance issues very well.

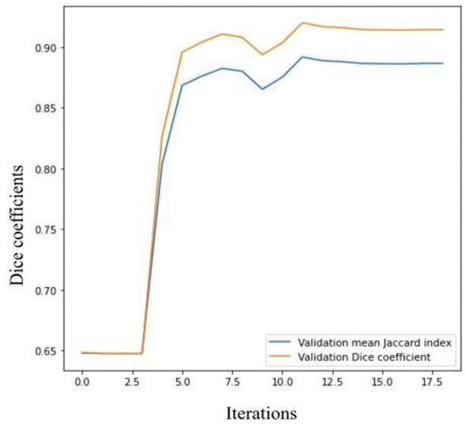


Fig. 5. Dice Loss

- **IoU Loss:** The mismatch error of segmentation is directly optimized in IoU loss. The network's output is given by the probability of a pixel falling into a specific class of the region. The pixel probabilities of output model Y0 for every pixel in the pixel vector are denoted by P and given from 1 to n. The ground truth values are given by $Y_G = [0,1]$ where 0 is the background class and 1 is the object.

To achieve an optimal solution a compound loss function that combines the IoU loss with the Dice Loss and Weighted Cross-entropy loss, is proposed at a faster convergence rate.

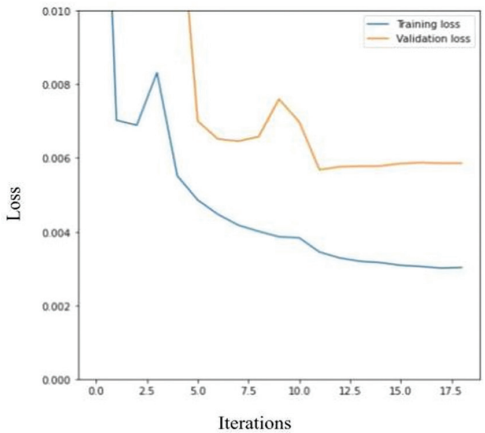


Fig. 6. IoU Loss

3.4. TUMOR CLASSIFICATION USING RESNET-50

After tumor segmentation, the classification involves ResNet with transfer learning. The Residual network (ResNet) is the popular deep network that handles the vanishing gradient problem and uses a learning approach

of deep residual to simplify the deeper neural network training and reduce the errors caused by increasing depth, as the network structure is composed of 50 layers. ResNet introduces skip connections, which identify shortcut connections to skip one or more layers. It demonstrates that the model's performance and computational time are directly proportional to the number of layers.

- In transfer learning, a model is trained on a large dataset for one task and then the pre-trained model weights are used to train for another task. Transfer learning improves performance with a simple function that involves adding new fully connected layers to pre-trained models.
- ResNet50 uses identity mapping from previous layers that help in tracking the vanishing gradient. The advantage of residual blocks is that deeper networks can be constructed without increasing the percentage of training error.
- The presence of more layers in ResNet50 help in knowing all the parameters from previous activations in the network, at the same time, solving complex problems, and also improving the overall performance. And more layers are used to retrieve more features from a brain image.

The brain images are classified efficiently to detect the brain tumor while evaluating initial cancer nodules, and this is achieved with high accuracy by using the ResNet50 model in the classification block. The tumor diagnostics were improved by using deep learning techniques without any issues wherein the deep learning increased the training speeds and accuracy. Thus, the proposed model performs brain tumor image classification using ResNet-50 through a transfer learning approach.

4. SIMULATION RESULTS

The proposed approach is implemented with Python 3.6 tensor flow and Keras deep learning platform. All experiments are conducted on a computer with the following specifications: Intel Core (TM) i7-6580 K with 16 GB RAM, frequency of CPU @ 3.360 GHz, and GPU of NVIDIA GeForce GTX 1080. The performance of the proposed model is validated on BRATS 2020 dataset in terms of Accuracy, Sensitivity, and Specificity. The training speed of the developed model is 12 minutes and 54 seconds.

Accuracy:

The test scans use the trained network for overall classification accuracy, which is the ratio of the total of true positives and negatives to the overall prediction trials.

Sensitivity:

The sensitivity metric is used to find a tumor-affected individual. It is the ratio of true positives to the total of true positives and false negatives.

Specificity:

Specificity is the reversal to the sensitivity metric where the test ability finds and designates the individual does not have a tumor. It is the measure of true negatives to the total of true negatives and false positives.

4.1. QUANTITATIVE EVALUATION

In this research, the evaluation of the proposed ResNet-50 DNN architecture is validated on the BRATS-2020 dataset, where it comprises the MRIs of brain tumor patients with different tumor types. To expose the multimodality nature of the SD scans, a U-Net architecture framework is introduced in the network to solve memory computational issues. By validating Table 3 the experimental evaluation is performed by comparing the dice scores of existing models and the proposed models in classifying WT, TC, and ET.

The proposed model achieved better dice scores with WT of 0.97, ET of 0.88, and TC of 0.90 compared to Multi Encoder net for tumor segmentation [21] dice score, WT of 0.88, ET of 0.70, and TC of 0.74. another prior CNN-based model using YOLOv2 inceptionv3 achieved dice scores, WT of 0.93, ET of 0.87, and TC of 0.79.

Table 3. Dice Scores achieved in Tumor Classification on BRATS 2020 dataset

Architecture	Dataset	Dice scores		
		WT	ET	TC
ME-Net [21]	BRATS 2020	0.88	0.70	0.74
YOLOv2 inceptionv3 [22]		0.93	0.87	0.79
Proposed model DNN-based ResNet-50		0.97	0.88	0.90

4.2. COMPARATIVE EVALUATION

The comparative analysis between the proposed model and the prior models is given in Table 4. The performance metrics accuracy, sensitivity, and specificity of prior models with the proposed model are compared in this section. A CNN-based ME-net model has achieved an accuracy of in the region of WT, sensitivity of 91.30%, and specificity of 96.73%. Another CNN-based model using YOLOv2 inceptionv3 model has achieved an accuracy of 98.52%, sensitivity of 98.78%, and specificity of 97.29%. The deep residual model achieved an accuracy of 99.94%, a sensitivity of 98.92%, specificity of 98.63% which are higher than the other two CNN-based models. The proposed Deep Residual Network overcomes the limitations of existing models such as multi-modality fusion, more vanishing gradient, and low-performance feature extraction.

The advantage of the proposed model is; it performs pre-processed modality using hierarchical inception blocks for direct fusion, and decreases vanishing gradient by using ResNet-50, because the network finds skip connections, which identify shortcut connections to skip one or more layers [23, 24]. The overfitting of the model can be overcome by using the 50-layered residual network that deeply convolutes the multi-category of tumor images in classification tasks using convolution block and identity block [25, 26].

Table 4. Performance Metrics of Proposed Model

Architecture	Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
CNN-based ME-Net model	BRATS 2020	97.46	91.30	96.73
CNN-based YOLOv2 inceptionv3		98.52	98.78	97.29
DNN-based ResNet-50 model		99.94	98.92	98.63

5. CONCLUSION

The deep residual networks improved the accuracy and training time was reduced to enhance the performance of the residual networks. The ability of deep residual networks reduced the problem of vanishing gradient. The CNN with ResNet-50 model consists of 50 layers to predict the tumor in the brain images where the input is taken from a pre-trained network that trains a million images from the ImageNet database. The 2020 BRATS dataset includes 369 images of the training set, 125 validation sets, and 166 test sets of multimodal brain MRIs of patients consisting of high-grade or low-grade gliomas. The Gaussian filter and data augmentation are used in preprocessing to remove the noise and after multi-modalities fusion, the tumor was separated from healthy brain tissue which helps with diagnosis and treatment planning. Transfer learning is used for classification with ResNet after segmentation and a popular deep network reduces the problem of vanishing gradient and the dice score resulted as 0.88 for enhanced tumor, 0.97 for detecting the whole tumor, and 0.90 for tumor core on the validation set. In future work, transfer learning can be implemented in large datasets and detect various types of tumors in the early stages.

Nomenclature

Parameters	Representation
x	Horizontal orientation in Gaussian template
y	Vertical orientation in Gaussian template
L_F	Loss function
L_{iou}	Loss of Intersection over Union
L_{Dice}	Dice loss
N	Sample set
L	Label set
y_i^l	Sampling encoder
$\hat{y}_i^{(l)}$	Predicted probability
ϵ	constant
Y_0	Output model
P	Pixel vector
Y_G^p	Pixel vector of ground truth values
Y_0^p	pixel values of output model

6. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

7. AUTHOR CONTRIBUTIONS

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, and visualization, have been done by 1st author. The supervision and project administration, have been done by the 2nd author.

8. REFERENCES:

[1] T. Zhou, S. Canu, S. Ruan, "Fusion based on attention mechanism and context constraint for multimodal brain tumor segmentation", Computerized Medical Imaging and Graphics, Vol. 86, 2020, p. 101811.

[2] S. A. A. Ismael, A. Mohammed, H. Hefny, "An enhanced deep learning approach for brain cancer MRI images classification using residual networks", Artificial Intelligence in Medicine, Vol. 102, 2020, p. 101779.

[3] S. Peng, W. Chen, J. Sun, B. Liu, "Multi-scale 3d u-nets: an approach to automatic segmentation of brain tumor", International Journal of Imaging Systems and Technology, Vol. 30, No. 1, 2020, pp. 5-17.

[4] S. Ramesh, S. Sasikala, N. Paramanandham, "Segmentation and classification of brain tumors using modified median noise filter and deep learning approaches", Multimedia Tools and Applications, Vol. 80, No. 8, 2021, pp. 11789-11813.

[5] X. Lei, X. Yu, J. Chi, Y. Wang, J. Zhang, C. Wu, "Brain tumor segmentation in MR images using a sparse constrained level set algorithm", Expert Systems with Applications, Vol. 168, 2021, p. 114262.

[6] J. Zhang, J. Zeng, P. Qin, L. Zhao, "Brain tumor segmentation of multi-modality MR images via triple intersecting U-Nets", Neurocomputing, Vol. 421, 2021, pp. 195-209.

[7] P. M. S. Raja, A. V. Rani, "Brain tumor classification using a hybrid deep autoencoder with Bayesian fuzzy clustering-based segmentation approach", Biocybernetics and Biomedical Engineering, Vol. 40, No. 1, 2020, pp. 440-453.

[8] M. A. A. Hamid, N. A. Khan, "Investigation and classification of MRI brain tumors using feature extraction technique", Journal of Medical and Biological Engineering, Vol. 40, No. 2, 2020, pp. 307-317.

[9] S. Kim, B. Kim, H. Park, "Synthesis of brain tumor multicontrast MR images for improved data augmentation", Medical Physics, Vol. 48, No. 5, 2021, pp. 2185-2198.

[10] B. Srikanth, S. V. Suryanarayana, "Multi-class classification of brain tumor images using data augmentation with deep neural network", Materials Today: Proceedings, 2021. (withdrawn)

[11] N. S. Punn, S. Agarwal, "Multi-modality encoded fusion with 3D inception U-net and decoder model for brain tumor segmentation," Multimedia Tools and Applications, Vol. 80, No. 20, 2021, pp. 30305-30320.

[12] Z. Ullah, M. U. Farooq, S. H. Lee, D. An, "A hybrid image enhancement based brain MRI images classification technique", Medical Hypotheses, Vol. 143, 2020, pp. 109922.

[13] Z. U. Rehman, M. S. Zia, G. R. Bojja, M. Yaqub, F. Jinchao, K. Arshid, "Texture based localization of a brain tumor from MR-images by using a machine learning approach", Medical Hypotheses, Vol. 141, 2020, p. 109705.

[14] J. Amin, M. Sharif, N. Gul, M. Yasmin, S. A. Shad, "Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network", Pattern Recognition Letters, Vol. 129, 2020, pp. 115-122.

[15] W. Zhang, G. Yang, H. Huang, W. Yang, X. Xu, Y. Liu, X. Lai, "ME-Net: multi-encoder net framework for brain tumor segmentation", International Journal of Imaging Systems and Technology, Vol. 31, No. 4, 2021, pp.1834-1848.

[16] M. I. Sharif, J. P. Li, M. A. Khan, M. A. Saleem, "Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images", Pattern Recognition Letters, Vol. 129, 2020, pp. 181-189.

[17] V. V. S. Sasank, S. Venkateswarlu, "An automatic tumor growth prediction based segmentation using full resolution convolutional network for brain tumor", Biomedical Signal Processing and Control, Vol. 71, No. A, 2022, p. 103090.

[18] Z. Ullah, M. Usman, M. Jeon, J. Gwak, "Cascade multiscale residual attention CNNs with adaptive ROI for automatic brain tumor segmentation", Information Sciences, Vol. 608, 2022, pp. 1541-1556.

[19] Z. Ullah, M. Usman, J. Gwak, "MTSS-AAE: Multi-task semi-supervised adversarial autoencoding for COVID-19 detection based on chest X-ray images", Expert Systems with Applications, Vol. 216, 2023, p. 119475.

[20] S. Solanki, U. P. Singh, S. S. Chouhan, S. Jain, "Brain Tumor Detection and Classification by using Deep Learning Classifier", International Journal of Intelligent Systems and Applications in Engineering, Vol. 11, No. 2s, 2023, pp. 279-292.

[21] A. Biswas, P. Bhattacharya, S. P. Maity, R. Banik, "Data augmentation for improved brain tumor segmentation", IETE Journal of Research, 2021, pp. 1-11.

[22] A. M. I. Sharif, J. P. Li, J. Amin, A. Sharif, "An improved framework for brain tumor analysis using MRI based on YOLOv2 and convolutional neural network", Complex & Intelligent Systems, Vol. 7, No. 4, 2021, pp. 2023-2036.

[23] A. R. Khan, S. Khan, M. Harouni, R. Abbasi, S. Iqbal, Z. Mehmood, "Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification", Microscopy Research and Technique, Vol. 84, No. 7, 2021, pp. 1389-1399.

[24] A. S. Akbar, C. Fatichah, N. Suciati, "Single level UNet3D with multipath residual attention block for brain tumor segmentation", Journal of King Saud University-Computer and Information Sciences, Vol. 34, No. 6B, 2022, pp. 3247-3258.

[25] S. K. Mohapatra, P. Sahu, J. Almotiri, R. Alroobaea, S. Rubaiee, A. B. Mahfouz, A. P. Senthilkumar, "Segmentation and Classification of Encephalon Tumor by Applying Improved Fast and Robust FCM Algorithm with PSO-Based ELM Technique", Computational Intelligence and Neuroscience, Vol. 2022, 2022, p. 2664901.

[26] R. Pitchai, P. Supraja, A. H. Victoria, M. Madhavi, "Brain tumor segmentation using deep learning and fuzzy k-means clustering for magnetic resonance images", Neural Processing Letters, Vol. 53, No. 4, 2021, pp. 2519-2532.

A Hybrid Metaheuristics based technique for Mutation Based Disease Classification

Original Scientific Paper

Manu Phogat
Guru Jambheshwar University of Science & Technology,
Hisar, India-125001
kunjean4181@gmail.com

Dharmender Kumar
Guru Jambheshwar University of Science & Technology,
Hisar, India-125001
dharmindia24@gmail.com

Abstract *– Due to recent advancements in computational biology, DNA microarray technology has evolved as a useful tool in the detection of mutation among various complex diseases like cancer. The availability of thousands of microarray datasets makes this field an active area of research. Early cancer detection can reduce the mortality rate and the treatment cost. Cancer classification is a process to provide a detailed overview of the disease microenvironment for better diagnosis. However, the gene microarray datasets suffer from a curse of dimensionality problems also the classification models are prone to be overfitted due to small sample size and large feature space. To address these issues, the authors have proposed an Improved Binary Competitive Swarm Optimization Whale Optimization Algorithm (IBCSOWOA) for cancer classification, in which IBCSO has been employed to reduce the informative gene subset originated from using minimum redundancy maximum relevance (mRMR) as filter method. The IBCSOWOA technique has been tested on an artificial neural network (ANN) model and the whale optimization algorithm (WOA) is used for parameter tuning of the model. The performance of the proposed IBCSOWOA is tested on six different mutation-based microarray datasets and compared with existing disease prediction methods. The experimental results indicate the superiority of the proposed technique over the existing nature-inspired methods in terms of optimal feature subset, classification accuracy, and convergence rate. The proposed technique has illustrated above 98% accuracy in all six datasets with the highest accuracy of 99.45% in the Lung cancer dataset.*

Keywords*: Feature Selection, Metaheuristic, Competitive Swarm Optimization, Whale Optimization Algorithm, Artificial Neural Network, Hybrid Techniques, Classification*

1. INTRODUCTION

Mutations are types of abnormal changes in the genetic material, especially within nucleic acid (RNA, DNA). The change in the nucleotide sequence of DNA results in an alteration in the amino acid's sequences of proteins, which leads to certain genotypic and phenotypic changes in the human body. Various diseases are caused by mutation, the most common are different types of cancers. Cancer is caused by acquired mutations, also known as somatic mutation, that occurs due to changes in nucleotide patterns such as C -> T [1]. The tumor is a very complex disease, which is driven by various factors like lifestyle, environment, and genetics. During the past few years, investigation on gene mutation at both specific loci and large scale has been carried out to increase the knowledge of molecular diversity in complex diseases like cancer. Several large-scale cancer genome projects have been carried out which have provided huge amounts of high-dimensional data. These projects include ICGC (The international cancer genome Consortium), TGCA (The Cancer Genome Atlas), and Cancer Genome Project by Trust Sanger Institute along with many other experiment-based studies that have been conducted for cancer classification [2-4]. The tremendous rise in DNA Microarray technology has given us a deep insight into various alterations and genetic variations that helps us in the early detection of complex diseases such as cancer. The genomic datasets have thousands of genes and a small number of samples, which makes the selection and classification of cancer very difficult. The high-dimensional datasets contain redundant and irrelevant genes that decrease the training strength of a classifier. So numerous techniques have been proposed in recent years for cancer classification using machine learning [5]. The researchers found machine learning, the most significant tool to perform data analysis in

biological datasets [6]. The main objective of cancer classification is to identify the biomarkers (genes) to differentiate various types of cancer. Techniques that have been proposed in the literature for gene selection methods are categorized into four methods that are as follows: wrapper, filter, embedded, and hybrid. The filter method selects the features on the basis that how to correlate with the output or based on their relationship with the output. The filter methods are fast, classifier-independent, and computationally inexpensive [7]. On the contrary wrapper methods split the data into subsets and train a model using this, the addition and deletion of features are dependent on the output of the model. So generally, wrapper methods provide better accuracy than the filter methods, but in comparison, they are computationally expensive because they tested all possible combinations of feature subsets. The hybrid methods combine the qualities of both filter and wrapper methods to create the best sub-set of features. In hybrid methods, the researchers first apply the filter method to reduce the feature size and then use the wrapper technique for the final selection of the relevant genes. The carcinoma microarray datasets are highly dimensional and suffer from a curse of dimensionality problem, which means a high number of features and a small sample size. The small p and large n state the cancer classification problem as an NP-hard problem. For the past few years, different metaheuristics techniques have been proposed in the literature to solve the different variety of problems related to real-life. The metaheuristics algorithms are easily understandable and computationally inexpensive because they provide an optimal solution in a decent amount of time, which makes them pretty useful in the problem areas of bioinformatics and computational biology. So, for the selection and classification of genes metaheuristics algorithms are used in the filter, wrapper, and hybrid techniques, and also used for parameter tuning of a classifier to improve the classification accuracy. The wrapper techniques generally used the conventional fitness function for selecting the genes, to maximize the performance of a classifier. Accordingly in this study, the authors have introduced a new fitness function to overcome the limitation of conventional fitness or basic fitness function. In the past few years, machine-learning algorithms have been used for optimal prediction in the field of computational biology. It is difficult to identify the gene signification for complex and large biological structures, so the metaheuristics techniques add new insight. Numerous nature-inspired techniques are used in cancer classification for relevant gene selection [8].

In the literature part authors have introduced a dominant metaheuristics algorithm, CSO for solving real-world problems, also due to its impressive capabilities it has also been used for gaining the nearest optimal solution in the large search space efficiently. The CSO Technique is highly effective to find the relevant gene subset, but this process needs a few improvements for slow convergence, which inspired us to search for further progress. Another popular metaheuristic technique is the whale optimization algorithm (WOA) widely used for hyperparameter tuning in various classifiers [9]. WOA leads to global optima and demonstrates rapid convergence as compared to other metaheuristic techniques. Moreover, WOA is very flexible and robust in handling a large number of decision variables.

In recent scenarios, various wrapper techniques are introduced in bioinformatics to examine the biological system for a thorough perspective. Many benchmark wrapper techniques have been explored for mutated gene selection in the tumor classification but these algorithms failed to identify the correlation between genes in their search process [10]. This leads to a rise in the computational load for identifying the optimal genes. To overcome this weakness, researchers have been exploring various hybrid evolutionary methods, such as a hybrid of PSO & GA, a hybrid of BBHA & BPSO, a hybrid of TLBO and GSA and fusion of EFS and AGOA Algorithm [11-14].

The existing hybrid methods still experience a lot of shortcomings, such as being stuck in local optima and having high execution time, so they do not accomplish adequate classification accuracy. In virtue of that, our study developed a new hybrid metaheuristics technique to conquer the shortcomings of the traditional algorithms and identify the target genes for accurate cancer prediction. The proposed technique has some impeccable advantages such as scaling down the computational complexity and boosting the accountability of the dataset; also, in addition, it can handle high dimensional data with the optimal solutions in a feasible time.

The key input of the present study is as follows:

- This article introduced a new hybrid metaheuristics technique with the combination of IMCSO and WO algorithms.
- The proposed technique introduces a new fitness function to improve classification performance
- The proposed technique optimizes the hyperparameter of the ANN classifier.

The rest of the arrangement of this article is as follows: Section 2 Comprises Related work in this the authors give a brief introduction of IBCSO, WO, and ANN techniques followed by a proposed hybrid metaheuristics technique (IBCSOWO) with an ANN model and its advantages. Section 3, focused on experimental results and discussion. Section 4 elaborates on the conclusion and future scope.

2. RELATED WORK

In the past few years, machine-learning algorithms have been used for optimal prediction in the field of computational biology. It is difficult to identify the gene signification for complex and large biological structures,

so the metaheuristics techniques add new insight. Numerous nature-inspired techniques are used in cancer classification for relevant gene selection. A hybrid gene identification technique has been proposed by, using the fusion of an Artificial bee colony (ABC) and a Genetic Algorithm (GA) [15]. The main goal of the article was to combine the advantages of both techniques to predict the relevant genes. Another gene selection technique proposed by using PSO and KNN target genes subset for tumor classification can be identified [16]. Another hybrid method using IGWO and PSO algorithms for the prediction of the most relevant genes in breast tumor classification can be acknowledged [17]. In another hybrid metaheuristics technique, Sharma et al. proposed a multi-objective framework (C-HMOSHSSA) with the fusion of unique MOSHO and SSA (Salp Swarm Algorithm), to select the optimal set of genes from high dimensional datasets [18].

Table 1. Comprises of Comparison of various tumor classification Techniques.

References	Techniques	Advantages	Disadvantages
[19]	Stacked Auto-Encoder (Deep Learning)	Formulate a clinical decision support system (DSS) to aid pharmacologists.	The performance of the model with respect to training time is poor than the existing models
[20]	Cuckoo search with crossover (Classification)	Assist both microarray and NGS-based miRNA expression data.	No Sensitivity analysis of feature subset.
[21]	Binary Bat Algorithm with SVM	Greedy Crossover proposed to rearrange the sub-optimal solutions. Overcome premature convergence	The technique is dependent on a particular dataset.
[22]	mRMR with Modified BAT Algorithm (Classification)	DNA microarray appearances empowered the simultaneous observation of expression levels of a large number of genes.	The time complexity of the proposed technique is high.
[16]	PSO and BAT (Classification)	The heuristics search technique is used to select the optimal values of K.	Cross-validation not performed, Higher complexity
[23]	Feature Score and ACO	The subset genes sampled are mapped into a dissimilarity space. Classifiers surpass the feature-based models.	The proposed technique only used one filter and wrapper technique in the study.
[24]	IG and GA (Classification)	Feed Forward neural network is used that gives good classification accuracy.	Smaller sample size with high complexity.
[25]	Gene Bank and GSA	The adaptive distance technique is used to improve the performance of the algorithm.	Slow convergence and high complexity.

The recent literature depicts the latest machine learning techniques and metaheuristics methods, which are applied, for cancer classification and shows potential results but most of them suffer from the local optima stagnation, redundancy, and slow convergence rate as depicted in Table 1. The feature space of microarray cancer datasets is large so the author used a hybrid feature selection technique to strike out the optimal feature subset.

Binary Competitive Swarm Optimization (BCSO)

Competitive Swarm Optimization is proposed by Cheng and Jin in 2015, it is a popular algorithm that uses a pair-wise competitive scenario. The CSO is considered a novel version of Particle Swarm Optimization [26]. The CSO randomly divides the population of particles into two equal-size groups. Each group is in a competitive spirit with each other and out of this competition, a particle having better fitness value is considered as a winner and directly moves to the next level. The loser particle updates its velocity and position by attaining information from the winner.

The loser velocity is updated as:

$$v_{l,d}(i+1) = r_1 v_{l,d}(i) + r_2\left(x_{w,d}(i) - x_{l,d}(i)\right) + \alpha r_3\left(\bar{x}_d(i) - x_{l,d}(i)\right) \quad (1)$$

$$x_{l,d}(i+1) = x_{l,d}(i) + v_{l,d}(i+1) \quad (2)$$

Where v_l and x_l are the velocity and position of the loser particle, x_w is the position of the winner particle, $\bar{x}$ is the mean position of the current swarm, r_1, r_2, and r_3 are three independent random vectors distributed in [0,1], α is the social factor, d is the dimension of search space and i is the iteration number.

The conventional CSO converts into binary CSO when the continuous real domain is converted into a discrete domain, so the solution can be represented in binary form. Traditionally the wrapper techniques consider Binary CSO instead of CSO. The solution represented in BCSO is either 0 or 1.

The main steps of BPSO are as follows:

- At first, randomly initialize the population of N particles. The velocity of every particle will be considered zero V=0.
- For every particle fitness is evaluated; the best fitness score particle is named gbest.
- Divide the particles into two groups on each iteration.
- The velocity of the loser particle is updated using Equation 1.
- Now the velocity is transformed into a probability value between [0,1], using a transfer function.

So, the updated position is calculated as:

$$x_{l,d}(i+1) = \begin{cases} 1, & (if\ S(v_{l,d}(i+1) > r_4) \\ 0, & (otherwise) \end{cases} \quad (3)$$

Here S is the transfer function and r_4 is a random vector dispersed in [0,1].

The BCSO is applied for the feature selection process in classification tasks. In BCSO the bit value 1 indicate feature is selected, while bit 0 is for the unselected feature [26].

Whale Optimization Algorithm (WOA)

The whale optimization algorithm introduced by Mirjalili and Lewis in 2016, is based on the social behavior of humpback whales [27]. The WOA is robust and easy to implement when compared to other nature-inspired techniques. When a whale attacks their prey, they encircle it and swim up to the surface in a shrinking circle. The WOA works in three phases: a) Shrinking encircling prey, b) Spiral shape attacking method, and c) search for prey.

The encircling behavior is expressed as:

$$\vec{D} = \left|\vec{C}\,\vec{X^*}(t) - \vec{X}(t)\right| \quad (4)$$

$$\vec{X}(t+1) = \vec{X^*}(t) - \vec{A}.\vec{D} \quad (5)$$

$$\vec{A} = 2\vec{a}.\vec{r} - \vec{a} \quad (6)$$

$$\vec{C} = 2.\vec{r} \quad (7)$$

Where $\vec{X}$ represents whale position, $\vec{X^*}$ is the general best position, a represents linearly reduced distribution within [2, 0] during iterations, t stands for present iteration, and r is random not dispersed within [0,1].

The helix shape movement of the whales inspired the spiral model, the Spiral Shape attacking method (exploita-tion phase) equation as follows:

$$\vec{X}(t+1) = \vec{D'}.e^{bl}.\cos(2\pi l) + \vec{X^*}(t) \quad (8)$$

Where $\vec{D'}$ is a vector that stores the absolute distance between $\vec{X^*}(t)$ and $\vec{X}(t)$, b is a fixed value that interprets logarithmic spiral space and l is a random numerical value ranged between [-1, 1].

If A>1 or A<-1, so for global optimizers, a search agent is revised as per a random search agent in place of the best search agent. The search for prey model equation is as follows:

$$\vec{D} = \left|\vec{C}.\overrightarrow{X_{rand}}(t) - \vec{X}\right| \quad (9)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}}(t) - \vec{X}.\vec{D} \quad (10)$$

The $\overrightarrow{X_{rand}}$ is called promptly from whales in the current iteration.

Artificial Neural Network

The ANN model proposed by McCulloch and Pitts in 1943, is based on the functionality of biological neurons [28]. The neural network consists of input neurons, which consist of input and their weight. Next is the internal neuron that provides a function, which has the summation of all weights and biases. The last is the output neuron in which the summation of weights and biases are passed through an activation function Figure 1.

The products of ANN with K elements are given as follows:

$$y(x) = \sum_{i=1}^{k} wiyi(x) \quad (11)$$

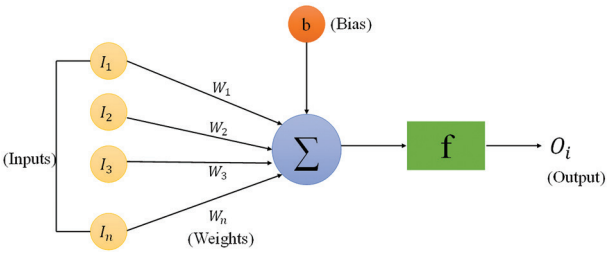


Figure 1. General Structure of ANN Network

The ANN has various architecture models such as single-layer feed-forward (perceptron), feed-forward neural network, and recurrent neural network. Several different variants are there of these architectures such as Kohonen networks, convolution neural networks, extreme learning machines, Hopfield networks, etc. In this article, a Multilayer Perceptron model is used which is part of a feed-forward neural network, the Multi-Layer Perceptron consists of an input layer, a hidden layer, and an output layer, and each node is associated with an activation function. The ANN has been widely used for tumor classification and the other different fields of bioinformatics [29].

3. PROPOSED FRAMEWORK

The proposed technique has three main aspects such as preprocessing, optimization, and classification. In the preprocessing phase, the input datasets go through a filter technique mRMR (minimum Redundancy Maximum Relevance) for gene selection. The genes are categorized according to their rank by filter method, followed by which a reduced dataset is obtained.

Pre-Processing Phase

The filter-based feature selection techniques are highly effective for filtering out the irrelevant and redundant genes in tumor classification. The efficient mRMR technique is used to generate the subset of high-quality genes. In mRMR the mutual information of variables X and c is determined based on the entropy of both X and c for each gene variable [30]. The entropy equation is followed as:

$$I(c;X) = H(c) - H\binom{c}{X} \quad (12)$$

Where $H(c)$ is entropy and $H(^c_X)$ is conditional entropy between variables and class. The main concept of minimum redundancy is to find out the genes which are

mutually maximally different from others. The average minimum redundancy is given as:

$$minZ(X,c) = \frac{1}{|s^2|}\sum_{x_j \in s} I\,(x_j ; x_k) \quad (13)$$

Where s stands for a subset of required genes and(x_j, x_k) represents mutual information between j-th and k-th genes. Again, the concept of mutual information is needed to choose a subset S with N genes having a maximum dependency on target class c. The average maximum relevance is formulated as:

$$maxV(X,c) = \frac{1}{|s|}\sum_{x_j \in s} I(\,c, x_j) \quad (14)$$

These two conditions are combined into a single criteria function $Max(V\text{-}Z)$, here mRMR is for discrete variable form. The equation for mRMR is an integration of Equations 13 and 14 and is described as:

$$j_{mRMR}(\phi) = \; I(c;X)^{-1}/_{|S^2|}\sum_{x_j \in s} I(x_j; x_k) \quad (15)$$

x_j is a selected subset of genes Sand x_k is the original genes set.

Proposed Algorithm

The mRMR filter method effectively reduces the dimension of the dataset and hands over the important genes. Now the effective classification model is applied to further scaling down the dimension of the gene subset and a hybrid algorithm is used with an ANN model to accomplish the maximum classification accuracy.

A proposed hybrid model called IBCSOWOA is used by fusion of two metaheuristics algorithms namely IBCSO and WOA, the IBCSO is used as a wrapper technique to deal with the dimension reduction and WOA is used for parameter tuning to improve the classification accuracy of the model. Algorithm 1 can show the Pseudo-code of the proposed algorithm. The IBCSO (Improved Binary Competitive Swarm Optimization) is an enhanced version of BCSO in which the swarm population is divided into three different swarms to escalate the convergence speed. The idea of tri-swarm in the proposed technique will allow two-thirds of the population to update, so more swarms will have a chance to move towards a good solution and it also balances exploration and exploitation. The mechanism of IBCSO describes in detail with the following steps:

- Initialization: At first the random particle is developed with a swarm size (m) in the multiple of three, which means the tri-division of the swarm. After initialization, all the particles are divided into three different groups.
- Tri-Competition: The particles from each group are picked up randomly and they go into a tri-competition. Out of the three only one will be declared the winner having the highest fitness value others will be named as the first and second losers.
- Updation: The winner is granted to pass immediately to the next iteration while both the loser is getting their velocity and position updated. The process of upgrading position and velocity is the same as in BPSO. The velocity and position for the first and second losers are described by the following equations: Loser1 ($l1$)

$$\upsilon_{l1,d}\,(i+1) = r_1(d,i)\upsilon_{l1,d}\,(i) + r_2\left(x_{w,d}\,(i) - x_{l1,d}\,(i)\right) + \alpha_1 r_3\left(\bar{x}_d(i) - x_{l1,d}\,(i)\right) \quad (16)$$

$$x_{l1,d}(i+1) = \; x_{l1,d}(i) \; + \upsilon_{l1,d}\,(i+1) \quad (17)$$

For Loser2 ($l2$)

$$\upsilon_{l2,d}\,(i+1) = r_4(d,i)\upsilon_{l2,d}\,(i) + r_5\left(x_{w,d}\,(i) - x_{l2,d}\,(i)\right) + \alpha_2 r_6\left(\bar{x}_d(i) - x_{l2,d}\,(i)\right) \quad (18)$$

$$x_{l2,d}(i+1) = \; x_{l2,d}(i) \; + \upsilon_{l2,d}\,(i+1) \quad (19)$$

Here $D(=m/3)$ represents the different swarms that participated in the competition. Position and velocities are $x_{w,d}\,(i)$, $x_{l1,d}(i)$, $x_{l2,d}(i)$ and $\upsilon_{w,d}(i)$, $\upsilon_{l1,d}(i)$, $\upsilon_{l2,d}(i)$ in the k-th round of competition (k=1,2,.....,k) in iteration i. r_1 to r_6 are six random numbers. α_1, α_2 are independent social factors regulating the impact of the mean position. $\bar{x}_d\,(i)$ is the mean position value of the relevant particle.

Algorithm 1: The pseudocode of Improved Binary Competitive Swarm Optimizer (IBCSO)

$P(i)$ represents the total swarm at each generation i. S represents a set of particles that do not participate in a Swarm. $X_w\,(i)$, $X_{l1}(i)$ and $X_{l2}(i)$ represent and two loser swarms respectively.

1. i=0;
2. Initialize the population p(0) randomly.
3. **While** termination criteria are not satisfied **do**
4. Search for fitness of every particle in $P(i)$;
5. $S=P(i)$, $P(i+1)=\phi$;
6. **While** $S=\phi$ **do**
7. Now arbitrarily choose three particles $X_1(i)$, X_2(i) *and* X_3(i) from S;
8. Arrange according to increasing order according to fitness function $f(X_1(i)) \le f(X_2(i)) \le f(X_3(i))$
9. Assign $X_w(i) = X_1(i)$, $X_{l1}(i) = X_2$(i) *and* X_{l2}(i) = X_3(i);
10. Add $X_w(i)$ *into* $P(i+1)$;
11. *Update* $X_{l1}(i)$, X_{l2}(i) to X_{l1} (i+1) and X_{l2} (i+1) *by equation* (16)-(19) and add to $P(i+1)$;
12. Change the velocity into probability using the S-shaped threshold function
13. Upgrade the position of the loser using equation (3)
14. If the value of the position vector is 1 then the feature is selected, for 0 feature is unselected.
15. F_s= {*All Features where position vector*=1;}
16. Remove $X_1(i)$, $X_2(i)$ and $X_3(i)$ from S;
17. **End while**

18. i=i+1;

19. **End while**

20. **Return** F_s;

So, the main idea of IBCSO is to develop a subset of highly efficient gene selection techniques with a better convergence rate and a good balance between exploration and exploitation. For the classification tasks, ANN is used with Rectified linear unit activation function with WOA used for optimizing weights.

The equation for ReLU is as follows:

$$y = \max(0, x)$$

Where y is the output function and x is the input weight.

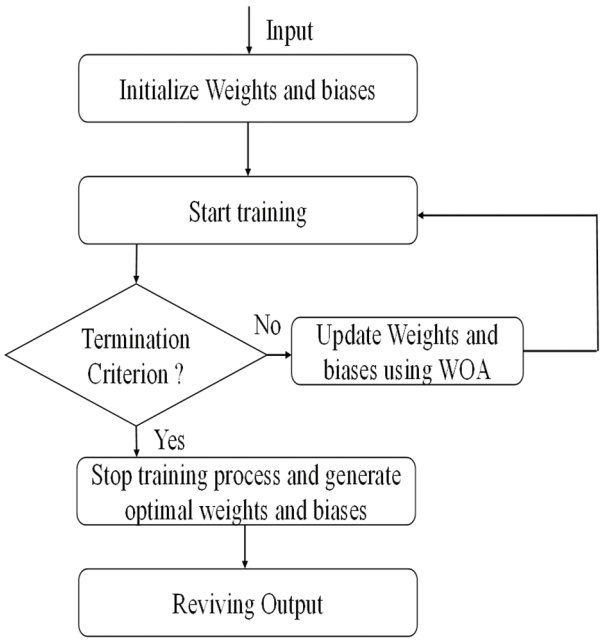


Figure 2. The Training Process of ANN

The training process of ANN is shown in Fig. 2. It takes input and achieves output based on current weights and biases. The computed output is compared with the target outcome with a loss function. After that, a backpropagation algorithm is used to update weight and bias in the next iteration. In the proposed technique WOA is used for the optimization of weights for the next iteration. The combined weights and biases in the ANN network are shown in Equation 21.

$$W = rs + 2r + 1 \tag{21}$$

Where s is the total no of input nodes and r is the total number of neurons in the hidden layer. The Mean Square Error (MSE) of ANN is the difference between predicted and actual values and it is used to alter the values of weights during backpropagation. MSE can be used by search agents (whales) as a fitness function to find out optimized weight values. The MSE can be calculated as:

$$MSE = \frac{\sum_{i=1}^{n}(O_i - O_i')^2}{n} \tag{22}$$

Where O_i is the actual output and O_i' is the predicted output for the input sample i. n is the number of samples.

Proposed fitness function

A new fitness function is proposed with respect to increasing the classification accuracy and reducing the number of genes. The fitness function is expressed as:

$$fitness(f) = \mathrm{A} * \frac{\mathrm{B}}{\Theta} + (1 - \mathrm{A}) * \Gamma \tag{23}$$

Here Γ represents the classification accuracy using ANN as a classifier. B Stands for the upper limit of selected features and A represent a constant value between 0 and 1. Θ is the measure of a chromosome length.

4. RESULT AND DISCUSSION

This section emphasizes the obtained results from the proposed technique and other methods on six mutation gene expression datasets from complex diseases such as Prostate, Lung, Breast, DLBCL, Arcene, and Dorothea. The Prostate dataset consists of 12600 genes with 136 samples out of which 77 include tumors and 56 are normal. The lung cancer dataset consists of 12533 genes with 181 tissue samples out of which 150 are of adenocarcinoma (AD-CA) and 31 are malignant pleural mesothelioma (MPM). The breast cancer dataset consists of 97 samples out of which 46 patients developed a tumor and the rest 51 are healthy samples, the number of genes in the dataset is 24481. The Diffuse Large B-cell lymphoma dataset (DLBCL) contains 11226 genes for 77 samples out of which 58 samples have large B cells and 19 are Follicular lymphoma. The Arcene dataset consists of 10000 genes and 100 samples, from which 56 are tumor samples and 44 are normal samples. Dorothea dataset consists of 800 samples and 100000 genes, from which 190 samples are positive and 610 are marked as negative. A detailed description of the datasets has been depicted in Table 2. The experimental results were obtained by using PYTHON 3.8, the observational evaluation was performed on NVIDIA Corporation TU104GL (Quadro TRX 5000) GPU and AMD 7662 (64-CORE*256) CPU along the Ubuntu 20.04.2. LTS (64-bit) operating system.

Table 2. Dataset description

No.	Datasets	Instances	Genes	Classes	Ref.
1	Prostate	136	12600	2	[31]
2	Lung	181	12533	2	[32]
3	Breast	97	24481	2	[31]
4	DLBCL	77	11226	2	[33]
5	Arcene	100	10000	2	[31]
6	Dorothea	800	100000	2	[31]

Parameter Setting

Table 3 contains the parameter values for IBCSO and WOA(ANN) techniques. The values were selected based on the results of considerable preliminary runs. The authors have tested the performance of all classifiers us-

ing 10-fold cross-validation to identify which one gives a better performance than the other methods on the top six-microarray datasets.

Table 3. Parameter Setting of the proposed technique.

S.no	Parameters	Value
1	No of generations	100
2	Population size (IBCSO, WOA)	200,100
3	Iterations	20
4	Θ Chromosome Length	50
5	Performance	Accuracy
6	$\alpha 1, \alpha 2$	0.2
7	Maximum velocity	6
8	α (WOA)	Linearly decreases from 2 to 0
9	b (WOA)	1

Evaluation Criteria

The performance of the proposed model is evaluated by the ANN classifier with the fitness function expressed in Equation 23. The proposed technique is evaluated by four different measures Specificity, Sensitivity, F-measure, Accuracy, and Matthews Correlation Coefficient (MCC). Performance measures are defined in the following equations:

$$Specificity(Se) = \frac{T_N}{T_N + F_P} \quad (24)$$

$$Sensitivity(Se) = \frac{T_P}{T_P + F_N} \quad (25)$$

$$F - score(F_s) = \frac{2*T_P}{2*T_P + F_P + F_N} \quad (26)$$

$$Accuracy(Acc) = \frac{T_P + T_N}{T_P + T_N + F_N + F_P} \quad (27)$$

$$MCC = \frac{(T_P * T_N - F_N * F_P)}{\sqrt{(T_P + F_P) + (T_P + F_N) + (T_N + F_P) + (T_N + F_N)}} \quad (28)$$

Here, T_P, T_N, F_P, and F_N are True positive, True negative, False positive, and False negative in all independent datasets. The T_p or T_N indicates exactly matched actual and predicted sample values, while F_p or F_N indicates distinct actual and predicted sample values.

Kappa Statistics

Cohen's kappa statistic is a performance model for classification, which measure the interrater reliability, which means a chance-corrected standardized measure of agreement between categorical scores produced by two raters. It measures value range from 0.0-1.0. The evaluation of Cohen's kappa is depicted as follows:

$$K = \frac{P0 - Pe}{1 - Pe} \quad (29)$$

Where $P0$ is the probability of agreement and Pe is the probability of random agreement.

Experimental Results and Analysis

The proposed technique is compared and validated with a series of algorithms frequently used on these datasets. The proposed technique is compared with different filter-based techniques such as FCBF, CMIM, mRMR, and Relief-F in Table 4. In the DNA microarray datasets, most of the genes are redundant, irrelevant, and noisy, so the top 100 genes are chosen to obtain the classification accuracy of all the methods. When IBCSOWOA is applied to the optimal gene set it increases the prediction accuracy and attains the highest accuracy 97.67% in the DLBCL dataset. Different types of nature-inspired algorithms are used in litera-ture for gene selection for complex disease classification.

In comparison with the earlier reported study, it should be found that using the mRMR filter selection method obtained 93.87% accuracy with ANN (WOA) on the DLBCL data; the proposed technique obtained the result of 97.67% accuracy with 100 selected genes on the same dataset. The highest performance is achieved by the proposed technique as 97.35% in Lung Cancer, 86.86% in Breast Cancer, 93.89% in Prostate Cancer, 94.25% in Arcene, and 91.45% in Dorothea dataset. This table states the efficiency results ANN (WOA) classifier except for the proposed technique in which a new fitness function is used; and chose the number of genes by CMIM, FCBF, mRMR, Relief-F, and the proposed technique of all datasets.

In Table 5 nature inspired algorithms such as PSO, GA, ACO, BCSO, and GWO are compared with the IBCSOWOA based on accuracy, standard deviation, and the optimal number of genes; the result demonstrates the effectiveness of the proposed technique. The optimum results between all gene selection techniques have been emphasized and marked in bold type. The most renowned metaheuristic algorithm such as GA selects the number of a gene to 18 in a DLBCL dataset with a classification performance of 97.52%. The classification performance of 99.45% in the Lung cancer dataset with 15 genes is the second maximum obtained by our proposed method. The metaheuristic algorithm PSO classification accuracy is 97.05% with 20 genes in the Lung cancer dataset.

In the ACO technique, the highest and lowest classification accuracy is 94.76% and 82.65% in DLBCL and Arcene datasets. Moreover, the BCSO algorithm obtained 96.21% classification efficiency from the DLBCL dataset with 22 genes and the GWO algorithm attain 91.25-classification accuracy from the Lung cancer dataset.

Table 4. Average classification performance with the top 100 genes from all six datasets

Techniques	DLBCL	Lung	Breast	Prostate	Arcene	Dorothea
FCBF	93.54	92.36	84.56	89.56	89.68	86.32
CMIM	92.68	92.99	83.75	87.34	90.25	85.65
mRMR	93.87	91.54	83.68	88.56	88.56	87.26
Relief-F	85.64	88.98	84.52	89.56	87.25	84.25
Proposed	**97.67**	**97.35**	**86.86**	**93.89**	**94.25**	**91.45**

Table 5. Observation of proposed technique with nature-inspired techniques with classification accuracy and STD, with the opti-mal number of selected genes.

Datasets	Performance metrics	PSO	GA	ACO	BCSO	GWO	Proposed
Prostate	#Acc ± STD # Genes	90.42± 0.53 15	91.91± 1.53 18	90.02± 1.82 21	86.35± 2.03 20	87.96± 2.34 25	**98.48± 0.68 09**
Lung	#Acc ± STD # Genes	97.05 ±0.87 20	95.61 ±0.53 16	88.46 ±1.31 19	87.96 ±2.02 24	91.25 ±0.92 25	**99.45± 0.05 15**
Breast	#Acc ± STD # Genes	93.64 ±0.87 21	92.17 ±0.96 17	90.99 ±1.52 15	91.68 ±1.63 20	88.47 ±2.13 18	**98.87 ±1.10 12**
DLBCL	#Acc ± STD # Genes	95.42 ±1.52 25	97.52 ±0.36 18	94.76 ±2.03 24	96.21 ±1.61 22	90.52 ±2.83 26	**99.62± 0.08 18**
Arcene	#Acc ± STD # Genes	90.41 ±1.03 17	91.92 ±0.86 19	82.65 ±2.31 23	87.36 ±1.53 21	89.56 ±2.52 25	**96.98± 0.08 11**
Dorothea	#Acc ± STD # Genes	87.23 ±2.89 27	94.15 ±2.51 22	86.32 ±2.74 18	87.35 ±2.57 24	84.36 ±2.63 27	**98.52 ±0.84 12**

The proposed technique is also compared with some state of arts hybrid techniques such as IWSSr+ Shuffled Frog Leaping Algorithm (SFLA) and teaching learning-based algorithm gravitational search algorithm (TLB-GSA) in Table 6. The results show the proposed method outperforms the two other hybrid techniques except in the prostate and DLBCL datasets where the accuracy of TLBOGSA and IWSSr+SFLA is comparatively higher. The highest accuracy achieved in the Lung cancer dataset with the proposed method is 99.45% with 15 optimal numbers of genes. Table 7 depicts the comparison of accuracy, sensitivity, specificity, F-measure, MCC, and kappa statistic of the optimal subset of genes obtained after applying mRMR and IBCSO algorithm is classified with four different classifiers including ANN (Optimized with WOA), the comparative evaluation states that ANN (WOA) provides more promising results. The Kappa Statistics is the highest (0.974) in the ANN classifier based on the whale optimization algorithm.

Table 6. Comparison of the proposed technique with other hybrid techniques.

Dataset	Performance	Proposed method	IWSSr+SFLA	TLBOGSA
Prostate	#Acc ± STD # Genes	97.48 ± 0.68 09	95.18 ±0.58 08	**98.42 ± 0.67 07**
Lung	#Acc ± STD # Genes	**99.45± 0.05 15**	98.16 ± 0.21 12	99.10 ± 0.02 13
Breast	#Acc ± STD # Genes	**98.87±1.10 12**	90.17 ± 0.14 11	97.87 ± 0.10 13
DLBCL	#Acc ± STD # Genes	**99.34 ± 0.08 18**	98.21 ± 0.47 15	97.26± 0.86 20
Arcene	#Acc ± STD # Genes	96.98 ± 0.08 11	**97.36 ± 0.78 9**	95.84 ± 0.84 12
Dorothea	#Acc ± STD # Genes	**98.52 ±0.84 12**	92.43 ± 0.56 21	96.87 ± 0.10 14

Table 7. Average classification performance of the proposed method using four different classifiers on six datasets.

Dataset	Measures	SVM	NB	kNN	ANN(WOA)
Prostate	Acc	94.44	94.78	90.65	97.48
	Se	96.25	93.25	89.65	96.38
	Sp	91.81	91.56	91.03	95.32
	Fmes	90.32	88.98	89.57	98.10
	MCC	0.89	0.90	0.88	0.97
	Kappa	0.88	0.89	0.86	0.95
Lung	Acc	98.16	94.97	92.87	99.45
	Se	99.10	94.11	91.35	99.21
	Sp	96.60	93.36	91.80	97.12
	Fmes	95.30	93.01	91.54	98.23
	MCC	0.97	0.94	0.95	0.98
	Kappa	0.96	0.88	0.87	0.97

Dataset	Measures	SVM	NB	kNN	ANN(WOA)
Breast	Acc	87.89	86.87	82.98	98.87
	Se	89.00	85.69	80.52	95.89
	Sp	85.55	88.07	7935	93.45
	Fmes	81.20	86.95	78.99	97.56
	MCC	0.93	0.94	0.90	0.97
	Kappa	0.84	0.83	0.80	0.96
DLBCL	Acc	98.12	96.07	94.88	99.34
	Se	98.33	95.36	95.36	99.12
	Sp	95.00	94.87	93.89	98.10
	Fmes	95.26	93.65	92.67	98.20
	MCC	0.95	0.94	0.91	0.96
	Kappa	0.96	0.92	0.90	0.96
Arcene	Acc	94.00	93.21	91.32	96.98
	Se	92.21	92.36	90.55	94.89
	Sp	95.45	90.35	89.65	92.56
	Fmes	92.46	91.58	88.39	95.32
	MCC	0.90	0.90	0.92	0.95
	Kappa	0.89	0.88	0.87	0.91
Dorothea	Acc	90.37	90.72	86.38	98.52
	Se	91.31	88.36	85.75	96.45
	Sp	89.47	89.65	84.96	93.12
	Fmes	86.97	91.56	85.32	98.13
	MCC	0.89	0.93	0.89	0.97
	Kappa	0.86	0.86	0.83	0.97

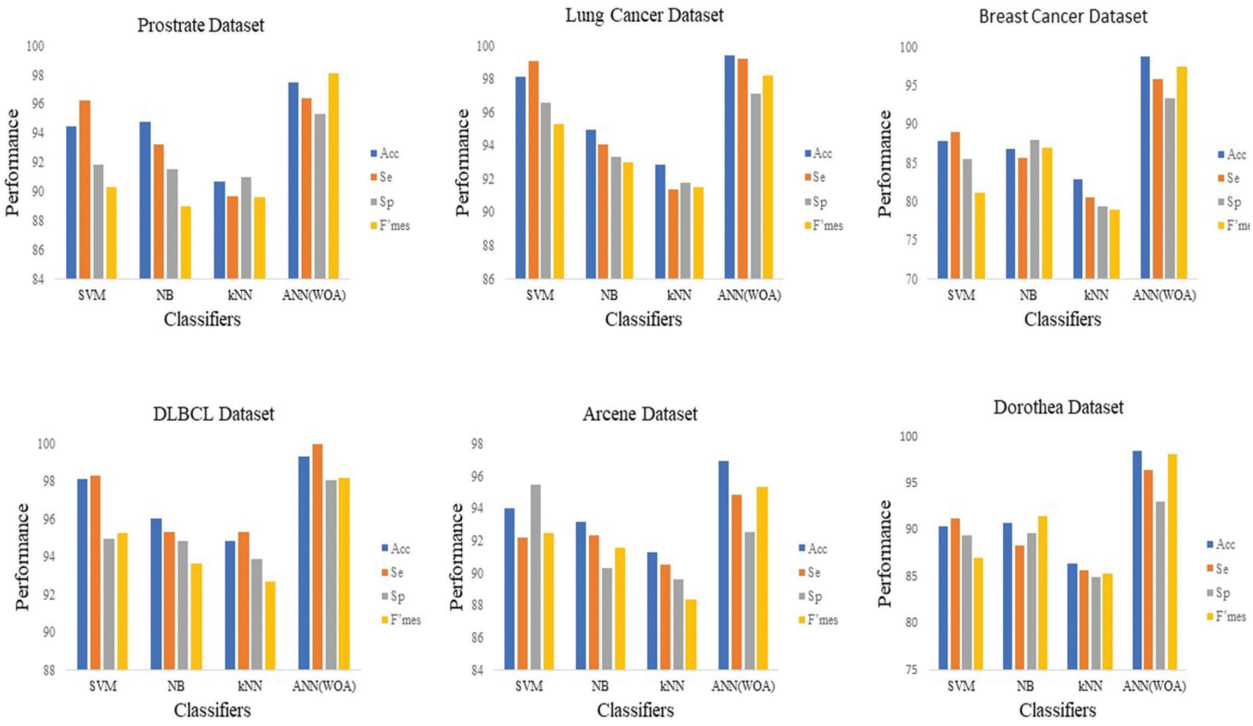


Figure 3. Performance comparison of the proposed technique using different classifiers on all the datasetsa

Fig. 3 displays the performance observation of the proposed gene selection technique with different classifiers including ANN(WOA) on all datasets. The parameters used are sensitivity (S_e), specificity (S_p), Accuracy (Acc), and F-measure (F'_{mes}).

5. CONCLUSION

Recently in the field of computational biology researchers are being attracted to the identification of marker genes related to complex diseases, especially

cancer diagnosis. However, it has been a difficult task to identify those markers due to the high dimensionality of microarray datasets. Although several existing techniques are efficient to strike out informative features from large datasets, these methods have some shortcomings such as slow convergence rate and high computational cost. To attain a good balance between exploration and exploitation and identify informative genes, a hybrid technique called IBCSOWOA is proposed to accelerate the gene selection process and improve classification accuracy. The IBCSOWOA incorporates the qualities of IBCSO and WOA techniques, the IBCSO selects the relevant feature subset and WOA optimizes the hyperparameters of ANN to improve the classification accuracy. The proposed technique is also introducing a new fitness function for identifying the informative genes. The experiments are conducted on six different biological datasets and the results illustrate that the technique outperforms other existing methods in terms of relevant gene subset selection and classification accuracy and also reduced the computational time. Out of six datasets, the proposed technique achieves more than 98% accuracy in all datasets. Therefore, it can be concluded that the proposed technique has been able to enhance the classification performance and reduce the computational time. This study also has some potential limitations such as all the datasets are microarray type and low sample size as compared to other gene datasets in the future RNA-seq datasets can be applied, as they are less noisy and more accurate. The current work can further be enhanced to incorporate deep learning techniques to improve the classification process.

Conflict of Interest

The authors of this publication declare there is no conflict of interest.

Funding Agency

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author(s) of the article.

Author Contribution

All authors have contributed equally.

6. REFERENCES

[1] M. S. Lawrence, P. Stojanov, P. Polak, C. Stewart, Y. Drier, E. Helman, J. Kim, G. Getz, "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes", Nature, Vol. 499, No. 7457, 2013, pp. 214-218.

[2] The International Cancer Genome Consortium, "International Network of Cancer Genome Projects", Nature, Vol. 464, No. 7291, 2010, pp. 993-998.

[3] L. Chin, M. L. Meyerson, "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways", Nature, Vol. 455, No. 7216, 2008, pp. 1061-1068.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J. P. Messirov, H. Coller, E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, Vol. 286, No. 5439, 1999, pp. 531-537.

[5] G. Isabelle, W. Jason, B. Stephen, V. Vladimir, "Gene Selection for Cancer Classification Using Support Vector Machines", Machine Learning, Vol. 46, No. 1, 2002, pp. 389-422.

[6] A. Sharma, R. Rani, "KSRMF: Kernelized Similarity Based Regularized Matrix Factorization Framework for Predicting Anti-Cancer Drug Responses", Journal of Intelligent & Fuzzy Systems, Vol. 35, No. 2, 2018, pp. 1779-1790.

[7] M. Phogat, D. Kumar, "Disease Single Nucleotide Polymorphism Selection Using Hybrid Feature Selection Technique", Journal of Physics: Conference Series, Vol. 1950, No. 1, 2021, p. 012079.

[8] A. K. Shukla, P. Singh, M. Vardhan, "An Adaptive Inertia Weight Teaching-Learning-Based Optimization Algorithm and Its Applications", Applied Mathematical Modelling, Vol. 77, 2020, pp. 309-326.

[9] F. S Gharehchopogh, H. Gholizadeh, "A Comprehensive Survey: Whale Optimization Algorithm and Its Applications", Swarm and Evolutionary Computation, Vol. 48, 2019, pp. 1-24.

[10] E. Zorarpaci, S. A. Ozel, "A Hybrid Approach of Differential Evolution and Artificial Bee Colony for Feature Selection", Expert Systems with Applications, Vol. 62, 2016, pp. 91-103.

[11] M. Phogat, D. Kumar, "Classification of Complex Diseases Using an Improved Binary Cuckoo Search and Conditional Mutual Information Maximization", Computación y Sistemas, Vol. 24, No. 3, 2020.

[12] E. Pashaei, N. Aydin, "Gene Selection Using Hybrid Binary Black Hole Algorithm and Modified Binary Particle Swarm Optimization", Genomics, Vol. 111, No. 4, 2019, pp. 669-686.

[13] A. K. Shukla, P. Singh, M. Vardhan, "Gene Selection for Cancer Types Classification Using Novel Hybrid Metaheuristics Approach", Swarm and Evolutionary Computation, Vol. 54, 2020, p. 100661.

[14] S. Dwivedi, M. Vardhan, S. Tripathi, A. K. Shukla "Implementation of Adaptive Scheme in Evolutionary Technique for Anomaly-Based Intrusion Detection", Evolutionary Intelligence, Vol. 13, No. 1, 2020, pp. 103-117.

[15] H. M. Alshamlan, G. H. Badr, Y. A. Alohali, "Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification", Computational Biology and Chemistry, Vol. 56, 2015, pp. 49-60.

[16] W. M. Shaban, "Insight into Breast Cancer Detection: New Hybrid Feature Selection Method", Neural Computing & Applications, Vol. 35, 2023, pp. 6831-6853.

[17] N. Kumar, D. Kumar, "An Improved Grey Wolf Optimization-Based Learning of Artificial Neural Network for Medical Data Classification", Journal of Information and Communication Technology, Vol. 20, No. 2, 2021, pp. 213-248.

[18] A. Sharma, R. Rani, "C-HMOSHSSA: Gene Selection for Cancer Classification Using Multi-Objective Meta-Heuristic and Machine Learning Methods", Computer Methods and Programs in Biomedicine, Vol. 178, 2019, pp. 219-235.

[19] K. Adem, S. Kiliçarslan, O. Comert, "Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification", Expert Systems with Applications, Vol. 115, 2019, pp. 557-564.

[20] A. Sampathkumar, R. Rastogi, S. Arukonda, A. Shankar, S. Kautish, M. Sivaram, "An efficient hybrid methodology for detection of cancer-causing gene using CSC for microarray data, Journal of Ambient Intelligence and Humanized Computing, Vol. 11, 2020, pp. 4743-4751.

[21] S. Akila, S.A. Christe, "A wrapper-based binary bat algorithm with greedy crossover for attribute selection", Expert System with Applications, Vol. 187, 2022, p. 115828.

[22] M. A. Al-Betar, O. A. Alomari, S. M. Abu-Romman, "A TRIZ-inspired bat algorithm for gene selection in cancer classification", Genomics, Vol. 112, 2020, pp. 114-126.

[23] M. Hamim, I. El Moudden, M. D. Pant, H. Moutachaouik, M. Hain, "A Hybrid Gene Selection Strategy Based on Fisher and Ant Colony Optimization Algorithm for Breast Cancer Classification", International Journal of Online and Biomedical Engineering, Vol. 17, 2021, p. 148.

[24] G. G. Afif, Adiwijaya, W. Astuti, "Cancer Detection Based on Microarray Data Classification Using FLNN and Hybrid Feature Selection", Journal RESTI, Vol. 5, 2021, pp. 794-801.

[25] A. Tahmouresi, E. Rashedi, M. M. Yaghoobi, M. Rezaei, "Gene Selection Using Pyramid Gravitational Search Algorithm", PLoS ONE, Vol. 17, 2022, p. e0265351.

[26] R. Cheng, Y. Jin, "A Competitive Swarm Optimizer for Large Scale Optimization", IEEE Transactions on Cybernetics, Vol. 45, No. 2, 2015, pp. 191-204.

[27] S. Mirjalili, A. Lewis, "The Whale Optimization Algorithm", Advances in Engineering Software, Vol. 95, 2016, pp. 51-67.

[28] W. S. McCulloch, W. Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity", The Bulletin of Mathematical Biophysics, Vol. 5, No. 4, 1943, pp. 115-33.

[29] J. Khan, J. S. Wei, L. H Saal, M. Ladanyi, F. Wastermann, P. S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks", Nature Medicine, Vol. 7, No. 6, 2001, pp. 673-679.

[30] H. C. Wu, X. G. Wei, S. C. Chan, "Novel Consensus Gene Selection Criteria for Distributed GPU Partial Least Squares-Based Gene Microarray Analysis in Diffused Large B Cell Lymphoma (DLBCL) and Related Findings", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 15, No. 6, 2018, pp. 2039-2052.

[31] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, F. Herrera, "A Review of Microarray Datasets and Applied Feature Selection Methods", Information Sciences, Vol. 282, 2014, pp. 111-135.

[32] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, R. Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma", Cancer Research, Vol. 62, No. 17, 2002, pp. 4963-4967.

[33] M. A. Shipp, K. N Ross, P. Tamayo, A. P Weng, J. L Kutok, M. Angelo, M. Reich, T. R. Golub, "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning", Nature Medicine, Vol. 8, No. 1, 2002, pp. 68-74.

Generic Paddy Plant Disease Detector (GP2D2): An Application of the Deep-CNN Model

Original Scientific Paper

Samuda Prathima
Department of ECE,
R.M.K. College of Engineering and Technology,
Chennai, India.
samudaprathimaece@rmkcet.ac.in

Praveena N G
Department of ECE,
R.M.K. College of Engineering and Technology,
Chennai, India.
drpraveenang@gmail.com

Sivachandar K
Department of ECE,
R.M.K. College of Engineering and Technology,
Chennai, India.
siva5446@gmail.com

Srigitha S Nath
Department of ECE,
Saveetha Engineering College, Chennai, India.
srigithapramodh@gmail.com

Sarala B
Department of ECE,
R.M.K Engineering College, Chennai, India.
bsa.ece@rmkec.ac.in

Abstract *– Rice is the primary food for almost half of the world's population, especially for the people of Asian countries. There is a demand to improve the quality and increase the quantity of rice production to meet the food requirements of the increasing population. Bulk cultivation and quality production of crops need appropriate technology assistance over manual traditional methods. In this work, six popular Deep-CNN architectures, namely AlexNet, VGG-19, VGG-16, InceptionV3, MobileNet, and ResNet-50, are exploited to identify the diseases in paddy plants since they outperform most of the image classification applications. These CNN models are trained and tested with Plant Village dataset for classifying the paddy plant images into one of the four classes namely, Healthy, Brown Spot, Hispa, or Leaf Blast, based on the disease condition. The performance of the chosen architectures is compared with different hyper parameter settings. AlexNet outperformed other convolutional neural networks (CNNs) in this multiclass classification task, achieving an accuracy of 89.4% at the expense of a substantial number of network parameters, indicating the large model size of AlexNet. For developing mobile applications, the ResNet-50 architecture was adopted over other CNNs, since it has a comparatively smaller number of network parameters and a comparable accuracy of 86.1%. A fine-tuned ResNet-50 architecture supported mobile app, "Generic Paddy Plant Disease Detector (GP2D2)" has been developed for the identification of most commonly occurring diseases in paddy plants. This tool will be more helpful for the new generation of farmers in bulk cultivation and increasing the productivity of paddy. This work will give insight into the performance of CNN architectures in rice plant disease detection task and can be extended to other plants too.*

Keywords*: disease identification, Convolutional Neural Networks, mobile application, paddy disease*

1. INTRODUCTION

Paddy is cultivated globally in more than hundred countries of a cultivation area of 158 million hectares [1] with an annual production of 700 million tons. Rice is one of the main staplefoods in Asian countries like India, Pakistan, China, Indonesia, Bangladesh and so on and these countries alone contribute an annual production of 640 million tons, which is almost 90% of global production. In India it is harvested in three seasons with the names: early Kharif from March to May, medium Kharif from June to October and Rabi from November to February.The possible reasons for reduction in rice yields areabnormal weather conditions, unseasonal rains and plant diseases caused by fungus, bacteria and viruses. Since we have less things to do with environmental conditions, to increase the productivity of paddy to meet the food requirements of increasing population, there is a need for a systematic identification of the disease in paddy plants at least.

Identification of diseases in paddy plants manually may result in errors and become difficult to do with large areas. For accurate detection of disease in large areas, earlier research involved various technologies such as image processing, IoT and Artificial Intelligence (AI) to address this problem. The research into applying AI for detection of diseases in paddy plants started a decade ago. From past research, it has been identified that Deep Convolutional Neural Networks (CNN) outperform other machine learning techniques for the identification of paddy plant disease. This study aims to identify the appropriate CNN architecture for Paddy plant disease identification. Hence, the performance of popular Deep-CNN architectures, including AlexNet, VGG-19, VGG-16, InceptionV3, MobileNet, and ResNet-50 for this multiclass problem was measured. These CNN architecture parameters are fine-tuned to get the maximum accuracy in classifying the plant image in one of the four classes such as healthy, Brown Spot, Hispa, or Leaf Blast. The performance of the above listed models for various hyperparameter configurations and their class predicting ability are discussed briefly in the later part of this study.

The organization of this paper as follows: Existing work is outlined in section II, pre-processing steps are presented section III, and result analysis is presented in the next section, followed by performance evaluation and the conclusions. The main contribution to this work are listed as follows:The performance of popular Deep-CNN models on Paddy plant disease identification is investigated for the Plant Village dataset. AlexNet, VGG-16, VGG-19, Inception V3, MobileNet, and ResNet-50 architectures are experimented with various batch sizes and learning rates.The high-performance CNN architecture suitable for mobile based application is developed and verified with real time paddy images.

2. RELATED WORKS

Earlier research carried out in similar applications helped to find the way for developing the web or mobile application for identifying the most common rice plant diseases. Ruoling Deng et al. [2] used an ensemble model by integrating three best performing models for this application, namely DenseNet-121, SE-ResNet-50, and ResNet-50 to develop a mobile app for rice plant disease detection. The base models are chosen based on their classification performance on a dataset of 33,026 images consisting of rice plant images affected by six types of rice diseases. In [3], Gugan Kathiresan et al. exploited transfer learning and General Adverse Networks (GAN) to develop a mobile app for the detection of disease in rice plants. Upon training and validation with three different dataset namely the rice leaf disease dataset (120 images), the rice disease image dataset (2092 images), and the leaf disease dataset (8293 images), the model achieved a cross validation accuracy of 98.38% without GAN and 98.7% with GAN. Krishnamoorthy et al. [4] explored transfer learning techniques on popular CNNs, namely VGG-16, ResNet-50, and InceptionV3,for classifying the disease-affected rice plant images into four classes. Upon hyperparameter tuning, the chosen architectures achieved the following accuracies for this multiclass classification problem: 87% for VGG-16, 93% for ResNet-50, and 95% for Inception V3. The work carried by Teja et al. [5] exploited different approaches to transfer learning techniques for feature extraction and reused the popularly trained models to minimize the prediction time. These approaches helped to achieve 99.33% testing accuracy for the InceptionV3 model.

In the study conducted by Md. Ashiqul Islam et al. [6], the performance of rice plant disease detection using the following CNN models was analysed: VGG-19, Inception ResNet-V2, ResNet-101 and Xception. The models are expected to classify the given image in any one of the following classes: Brown Spot, Leaf Blast, Bacterial Leaf Blight and Leaf Smut, which are frequent rice plant diseases in Bangladesh. Upon training Inception ResNetV2 with transfer learning, a maximum accuracy of 92.68% was achieved for five class classification, including healthy class. Swathika et al. [7] achieved an accuracy of 70% for two class classification using the ReLU activation function for Deep-CNN. Shivam et al. [8] have compared the LeNet5 model with the VGG-19 and MobileNetV2 models and concluded that MobileNetV2 is a light- weight model with a minimum number of parameters and computations as it uses the depth-wise separable Deep-CNN. They also identified that LeNet will work better for a smaller number of epochs. Shrivastava et al.[9] showed that the VGG-16 with a batch size of 8 and learning rate of 0.01works best for seven class classification tasks, including healthy class, with an accuracy of 93.11% among the compared architectures on Real Image Dataset. Divvela Srinivasa Rao et al. [10] identified the AlexNet model that gave better results for their own dataset.

Herlambang Dwi Prasetyo et al. [11] developed the web-based application using GoogLeNet architecture and Inception modules for rice plant disease detection to be used by end users. Chen et al. [12] used the transfer learning concept by combining the DenseNet pretrained on ImageNet Dataset with the Inception module and showed good accuracy for the public dataset. Dengshan Li et al. [13] have developed avideo detection architecture to detect disease in rice plants in real time using Deep-CNN for three class classification. Kalai Priya et al. [14] performed the fine tuning of the hybrid algorithms to increase the accuracy of rice plant disease detection using deep learning networks. Rishabh Sharma et al. [15] made an attempt at real-time identification of hispa rice plant disease and its severity from images taken from paddy fields in the state of Punjab, India. Shreya Ghosal et al. [16] improved the testing accuracy of pre-trained VGG-16 model using ImageNet dataset with the transfer learning technique. Hossain et al. [17] proposed a model with minimized network parameters and applied it on a novel dataset consisting

of 4199 rice plant images affected by the most common rice plant diseases of Bangladesh namely: Blast, Brown Spot, Bacterial Leaf Blight, Tungro and Sheath Blight, and achieved good accuracy.

S Ramesh et al. [18] have proposed the Jaya algorithm to detect the disease in rice plants using deep neural networks. They have generated a feedback loop as a post-processing step, to achieve the stability of their proposed algorithm. Liang et al. [19] proved that the DeepCNN outperformed the traditional handcrafted methods with the Rice Blast Disease dataset. They also stated that the performance of CNN techniques with SoftMax and CNN with Support Vector Machine (SVM) is similar. Cham et al. [20] developed a CNN model on TensorFlow for the detection of disease in rice plants. They have also used the ML kit service to store it and implement it as a mobile application. Anandhan K et al. [21] created their own dataset of 1500 images and applied R-CNN and mask R-CNN to detect the diseases in rice plants, proving mask R-CNN performs better than the other one. Zaki et al. [22] have obtained 90% accuracy for disease detection in Tomato plants using leaf images with a fine-tuned MobileNetV2 and stated that MobileNetV2 is suitable for developing a mobile application since the number of registers available on mobile are less compared to the desktop. Ahmed et al. [23] have developed a mobile application for the detection of the 38 most common diseases in 14 plant varieties. This work used an imagery dataset consisting of 96,206 images of healthy and disease affected plants. With the developed real-time application, they have obtained an overall classification accuracy of 94%. Aishwarya et al. [24] proposed a custom CNN model for the detection and identification of Tomato Plant Diseases. Kohli et al. [25] utilized Fuzzy C-means clustering for segmenting the leaf images, followed by feature extraction (color and texture) for detecting the type of disease in the leaf. The CNN also provides good support in the segmentation process, which can be used for leaf disease detection as well [26]. Verma et al. [27] used MobileNetV2 with lightweight depthwise convolutions to improve computation time and accuracy. Juyal et al. [28] adopted the R-CNN mask for correct identification and accurately mask the disease-affected region for fast identification. This work aims to utilize the power of CNN in a rice plant disease detection task.

3. PERFORMANCE OF CNN ARCHITECTURES IN RICE PLANT DETECTION

The proposed work to develop an application for identification of disease in paddy plants, namely Generic Paddy Plant Disease Detector (GP2D2), using CNN architectures involves the list of processes such as loading images, image pre-processing, dataset splitting, applying deep learning algorithms, and classification. These processes are illustrated in the work flow diagram shown in Fig. 1.

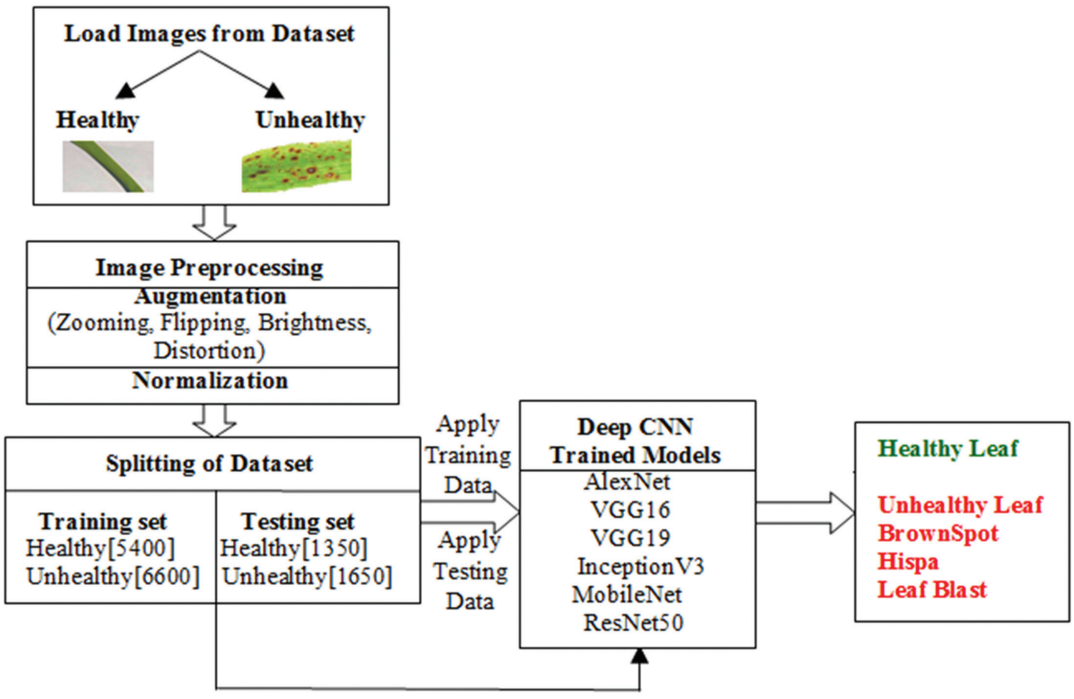


Fig. 1. Workflow of Proposed Rice Plant Disease Detection

a. Load Images from Dataset: The Plant Village dataset is chosen to train and test the CNN architecture since it consists of 3350 images belonging to three generic paddy plant disease classes and a healthy class. The sample images for each class, namely, Healthy, Brown Spot, Hispa and Leaf Blast, are shown in Fig. 2. The details about the number of images in each disease class are represented in Table 1.

b. Image Pre-processing: Augmentation helps to increase the number of images if a smaller dataset is

available in order to make the model more generalized and also to avoid the over fitting problem. Hence, the augmentation techniques such as zooming, flipping, brightness, and distortion are used in this work. The images were resized to 224x224x3. Min-max normalization was adopted since this technique preserves the relationship that exists between the original data. This technique scales the data in the range (0, 1).

c. Splitting of Dataset: After applying augmentation the dataset size increased to 15000, and the train-test split technique is used for training and testing the CNN architectures. 80% of the data were used for training the CNN model, and 20% of the data were used for testing the classification performance of the same. Table 1 gives the number of images used for training and testing in each class.

d. Training of Pre-trained CNN Models: The six pre-trained models exploited in this work are: (i) AlexNet, (ii) VGG-19, (iii) VGG-16, (iv) InceptionV3, (v) MobileNet, and (vi) ResNet-50. The above models were trained and tested on the images of the Plant Village dataset and its augmented versions. The performance of the CNN architectures is observed with different combinations of hyper parameters: (i) Learning rate = 0.01 and Batch size = 32; (ii) Learning rate = 0.01 and Batch size = 64; (iii) Learning rate = 0.001 and Batch size = 32; and (iv) Learning rate = 0.001 and Batch size = 64. The complete details of the hyper parameters used in this study are given in Table 2.

Table 1. Dataset description of Augmented Images

Class	No. of Original Images (O)	No. of Images in each Augmentation Method				Total no. of Augmentation Images (A)	Total Images (O+A)	Training Images	Testing Images
		Zooming	Flipping	Brightness	Distortion				
Healthy	1488	2160	1080	1600	422	5262	6750	5400	1350
Brown Spot	523	695	338	492	171	1696	2219	1775	444
Hispa	565	784	402	603	206	1995	2560	2048	512
LeafBlast	779	1026	544	821	301	2692	3471	2777	694

The sample images of each class, before and after augmentation techniques are shown in Fig. 2.

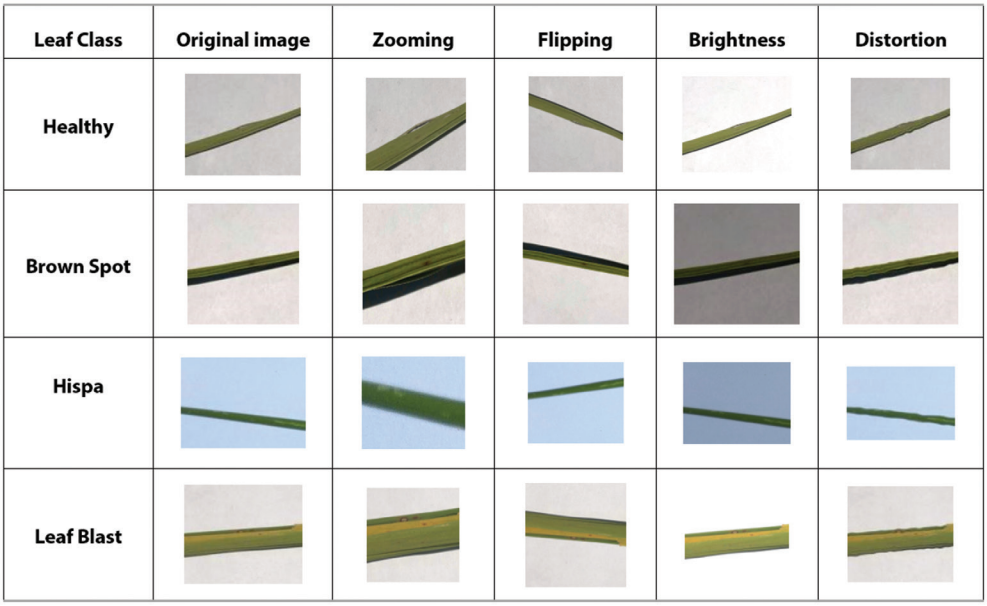


Fig. 2. Augmented images of paddy leaf diseases

Table 2. Hyper Parameters of the CNN Architecture

Hyper Parameters	Description
Batch size	32, 64
Learning rate	0.01, 0.001
Number of epochs	100
Type of optimizer	Adam
Activation function	ReLU
Loss	Categorical Cross Entropy

e. Performance Analysis

To analyse the performance of the above mentioned pre-trained architectures, the following metrics are used: Accuracy, Precision, Sensitivity, Specificity and F1 Score, which are shown in eqns.(1)-(5). The metric values are estimated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Sensitivityor Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

$$\text{F1 Score} = \frac{2TP}{2TP+FP+FN} \quad (5)$$

Where True Positive (TP) is the number of disease- affected images correctly predicted in its actual plant disease class, True Negative (TN) is the number of healthy plant images correctly predicted as healthy class, False Positive (FP) is the number of images not predicted correctly in its actual disease class or healthy class and False Negative (FN) is the number of disease-affected plant images wrongly predicted as healthy.

From Table 3, it is observed that AlexNet performed best with an average testing accuracy of 89.4% of batch size of 32 and learning rate of 0.01. Next, VGG-16 with 87.1% for 32, 0.01, followed by VGG-19 with 86.7% for 64, 0.001, ResNet-50 with 86.1% for 32, 0.001, MobileNet with 77.3% for 64, 0.01, and InceptionV3 with 42.4% for 32 batch size and 0.01 learning rate. From Table 4, it is observed that for batch size of 64, the VGG-19, MobileNet architectures performed well, and the other four architectures were performed well for batch size of 32. With a learning rate of 0.01, AlexNet, InceptionV3 performed well, while others performed well for 0.001. In Fig. 3, the computation time for the batch size of 32 and 64 with the learning rate of 0.01 and 0.001 for the pre-trained CNN architectures is shown for the different hyperparameters. The performance metrics obtained using pre-trained CNN models for four different combinations of Batch Size (BS) and Learning Rate (LR) were compared.

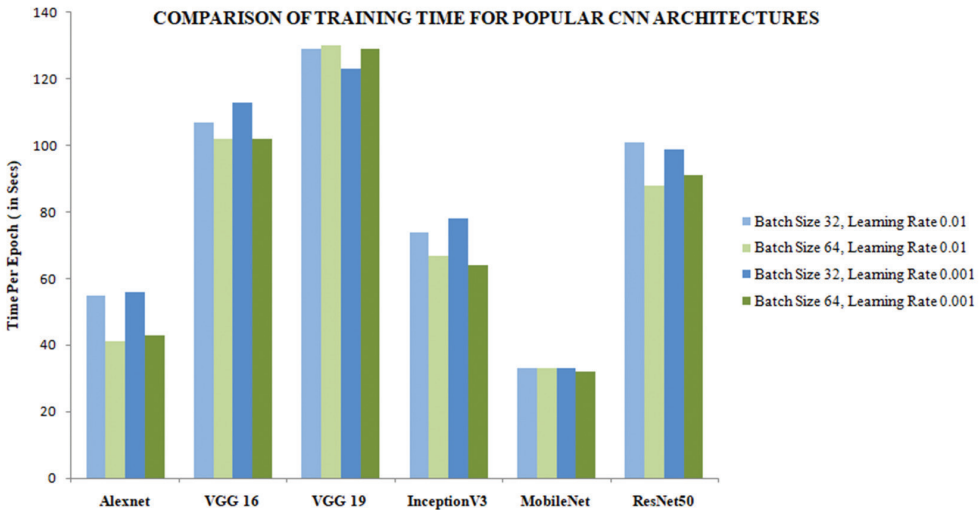


Fig. 3. Computation time of pre-trained CNN architectures for different hyperparameters

Table 3. Performance Comparison of Pre-trained CNN Architectures on Generic Paddy Plant Disease Detection

Model	Batch size	Learning Rate	Testing Accuracy	F1score	Precision	Recall	Specificity	Loss
AlexNet	32	0.01	0.8940	0.8939	0.8956	0.8923	0.8777	0.62
	64	0.01	0.8553	0.8552	0.8560	0.8544	0.8336	0.75
	32	0.001	0.8793	0.8789	0.8823	0.8757	0.8711	0.59
	64	0.001	0.8553	0.8549	0.8568	0.8531	0.8326	0.69
VGG-16	32	0.01	0.7080	0.6935	0.7396	0.6538	0.6635	1.06
	64	0.01	0.7433	0.7412	0.7615	0.7224	0.6837	1.15
	32	0.001	0.8713	0.8721	0.8731	0.8711	0.8520	1.10
	64	0.001	0.8647	0.8648	0.8651	0.8645	0.8520	1.38
VGG-19	32	0.01	0.7207	0.7128	0.7593	0.6728	0.8457	1.06
	64	0.01	0.7447	0.7427	0.7623	0.7243	0.8458	1.09
	32	0.001	0.8513	0.8510	0.8517	0.8503	0.8368	1.13
	64	0.001	0.8670	0.8675	0.8686	0.8664	0.8479	1.31
InceptionV3	32	0.01	0.4253	0.4248	0.4249	0.4246	0.3711	21.19
	64	0.01	0.4207	0.4207	0.4209	0.4205	0.3436	12.94
	32	0.001	0.2557	0.2558	0.2579	0.2538	0.2968	4.72
	64	0.001	0.2553	0.2535	0.2556	0.2515	0.2715	3.24
MobileNet	32	0.01	0.5410	0.1391	0.7498	0.0783	0.4102	1.10
	64	0.01	0.6053	0.5900	0.6205	0.5627	0.4806	1.01
	32	0.001	0.7700	0.7704	0.7753	0.7657	0.7462	1.08
	64	0.001	0.7733	0.7735	0.7670	0.7670	0.7302	1.06

ResNet-50	32	0.01	0.7620	0.7620	0.7763	0.7487	0.7011	0.97
	64	0.01	0.7907	0.7916	0.7955	0.7878	0.7468	1.21
	32	0.001	0.8610	0.8614	0.8620	0.8608	0.8402	1.58
	64	0.001	0.8583	0.8587	0.8590	0.8584	0.8406	1.26

4. RESULTS AND DISCUSSION

The summary of the performance comparison of the popular pre-trained models is presented in Table 5. AlexNet achieves highest accuracy among others, but it requires around 92 million parameters.VGG-16, VGG-19, and ResNet-50 require approximately 15 million to 24 million parameters to achieve accuracy closer to AlexNet. InceptionV3 shows very poor performance; hence, it may not be considered for this classification task. MobileNet requires a minimum number of parameters of around 4 million in order to achive accuracy. During the process of saving the adopted CNN on disk, the value of each and every single weight in the network layers is recorded. A network consisting of hundreds of millions of parameters, such as AlexNet, might occupy hundreds of megabytes, which is unacceptable on a mobile device or phone.

One solution for the mentioned problem is to reduce the resolution of the weights through quantization.

This process enables to store only the quantized values once and is used for assigning the network weights. But this process will reduce the architecture is the more preferable one for developing web app and will be a suggested one rather than implementing it as amobile application. MobileNet is an in-built mobile application that does not require accuracy, increases processing time, and also requires high computational power. Implementing the classification model in the backend of a cloud-based internet connection for rice plant disease detection has lesser classification accuracy than AlexNet and others. If accuracy is given more importance, then the targeted user needs to have good internet facilities to access the cloud-based web app through a smartphone, which has an AlexNet model as a backend model for prediction. VGG-16,VGG-19, and ResNet-50 also can be used as backend models.The Confusion matrix and the Area Under the Curve (AUC) of the best performing architectures with those particular batch sizes and learning rates are depicted in Fig. 4.

Table 4. Class Wise Performance of CNN Architecture on Generic Paddy Plant Disease Detection

Model	Batch size	Learning rate	Healthy	Brown Spot	Hispa	Leaf Blast
AlexNet	32	0.01	92.33	96.97	94.70	94.90
	64	0.01	90.14	96.57	93.13	92.73
	32	0.001	91.10	96.70	93.87	94.20
	64	0.001	88.66	96.80	93.07	92.53
VGG-16	32	0.01	76.27	93.37	85.27	86.70
	64	0.01	83.40	93.93	87.03	88.37
	32	0.001	90.13	96.93	93.27	93.93
	64	0.001	91.00	96.67	93.07	93.43
VGG-19	32	0.01	90.13	93.47	85.70	85.90
	64	0.01	89.77	94.27	93.07	93.43
	32	0.001	89.33	96.13	92.33	92.47
	64	0.001	90.57	96.33	93.10	93.40
InceptionV3	32	0.01	56.70	82.00	68.50	77.87
	64	0.01	54.47	85.37	66.57	77.73
	32	0.001	54.73	85.73	80.47	30.20
	64	0.001	55.37	85.83	83.07	26.80
MobileNet	32	0.01	59.00	91.13	83.03	75.03
	64	0.01	64.23	91.70	83.03	82.10
	32	0.001	81.70	95.00	90.33	88.23
	64	0.001	81.67	94.67	88.97	89.37
ResNet-50	32	0.01	81.37	94.37	87.13	89.63
	64	0.01	83.83	95.83	88.13	90.33
	32	0.001	89.17	96.53	93.40	94.00
	64	0.001	88.63	96.77	92.23	94.03

Table 5. Summary of Pre-trained CNN Architectures

Architecture	Accuracy	Total No of Parameters	More Compatible for Mobileapp/Webapp
AlexNet	89.40	91,784,764	Webapp
VGG-16	87.13	14,879,428	Webapp
VGG-19	86.70	20,189,124	Webapp
InceptionV3	42.53	21,810,980	Webapp
MobileNet	77.33	4,282,564	Mobileapp/Webapp
ResNet-50	86.10	24,145,668	Webapp

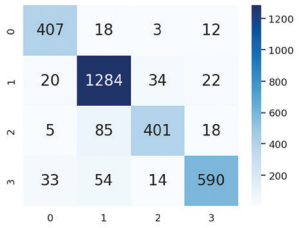
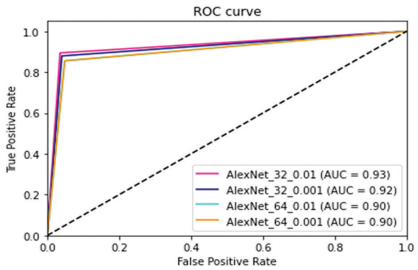
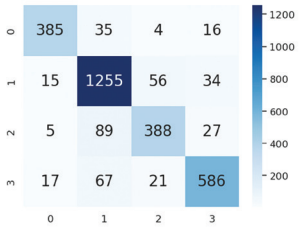
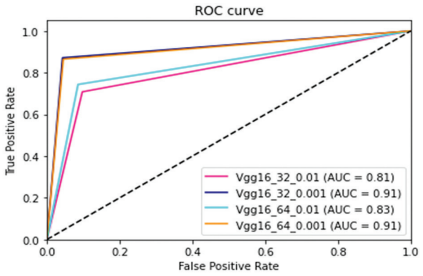
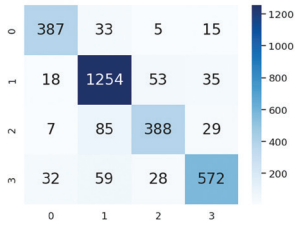
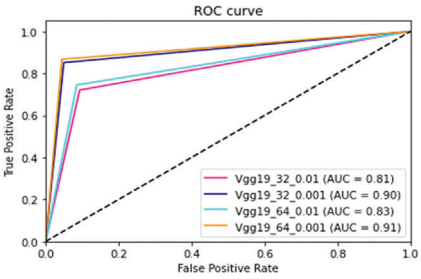
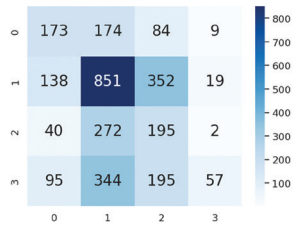
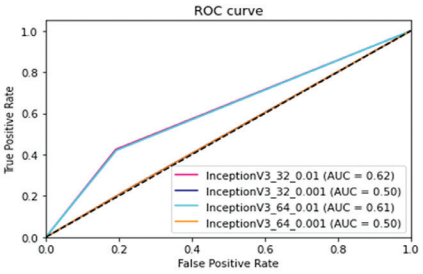
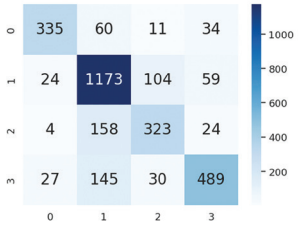
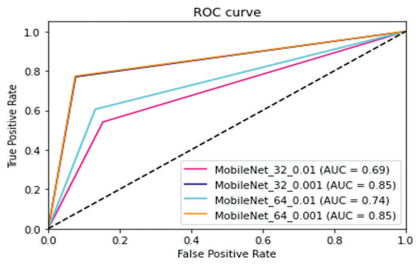
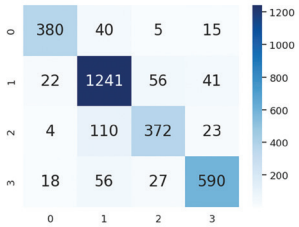
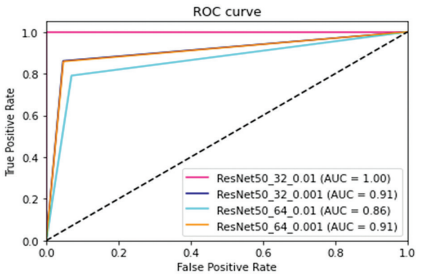
Model	Confusion Matrix	Area Under Curve (AUC)
AlexNet (Batch Size 32, Learning Rate 0.01)		
VGG-16 (Batch Size 32, Learning Rate 0.001)		
VGG-19 (Batch Size 64, Learning Rate 0.001)		
InceptionV3 (Batch Size 32, Learning Rate 0.01)		
MobileNet (Batch Size 64, Learning Rate 0.001)		
ResNet-50 (Batch Size 32, Learning Rate 0.001)		

Fig. 4. Classification performance of various pertained CNN architecures

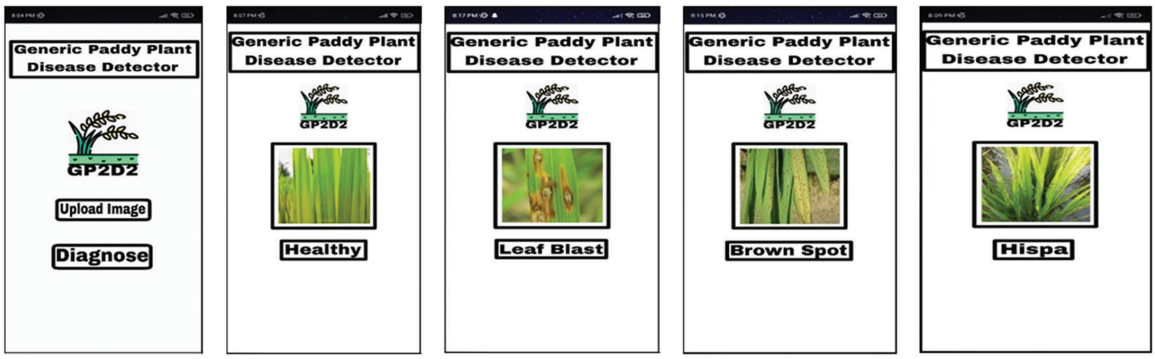


Fig. 5. Screenshots for mobile app home screen, healthy and diseased paddy leaves

5. DEVELOPMENT OF MOBILE APP: GENERIC PADDY PLANT DISEASE DETECTOR (GP2D2)

To make use of this study results a cloud based web application compatible with mobile was developed using AlexNet CNN architecture namely Generic Paddy Plant Disease Detector (GP2D2). It is an Android Mobile Application developed using Flutter with Android Studio as an Integrated Development Environment (IDE). The mobile application will help users to determine their plant health with a hassle-free process. Flutter is a free and open-source mobile UI framework created by Google for crafting natively compiled applications for mobile, web, and desktop forGoogle's Android operating system, built on JetBrains from a single codebase. Android Studio is the official IDE IntelliJ IDEA software and designed specifically for Android development. The GP2D2 mobile application allows users to capture images of the paddy plant. The user can upload the captured images to Firestore. Firestore is a cloud-hosted NoSQL database from Firebase and Google Cloud, where Android apps can access directly via native SDKs. The screenshots for the mobile app have been presented in the Fig. 5 representing home screen of the app, healthy leaves, Brown Spot, Leaf Blast and Hispa diseased leaves respectively.

6. CONCLUSION

This paper analyses the classification performance of popular CNN architectures in rice plant disease identification tasks. The results observed, AlexNet with highest accuracy of 89.4%, VGG-16, VGG-19 and ResNet-50 showed comparable accuracy. MobileNet can be considered for developing in-built mobile applications. The developed Generic Paddy Plant Disease Detector (GP2D2) application helps new-age farmers acquire the expert traditional farmer's experience in detecting the type of paddy disease digitally. Conventional methods for identification of disease in paddy plants are less effective when considered in a large area, and the developed mobile application will be more helpful in this case. Drones with cameras can be used for capturing the sample paddy images, which can be fed to the app for disease identification. This study will provide a handful of information on identifying suitable architecture for design and developing a real-time application for the identification of common diseases in paddy plants. This mobile application framework can be easily modified just by updating or replacing the existing model based on the requirements. This work will help the farmers easily identify the spread of paddy plant diseases in their early stages and take the necessary steps to avoid their spread. Successful implementation of this work will increase the Yield Per Hectare (YPH) of paddy plants and may improve the financial status. This work can be extended further to determine severity of disease conditions using semantic segmentation techniques through CNN [25].

7. REFERENCES:

[1] A Project of CGIAR, "Research Program on Rice", https://ricepedia.org/rice-as-a-crop/rice-productivity (accessed: 2023)

[2] R. Deng, M. Tao, H. Xing, C. Liu, K. Liao, L. Qi, "Automatic Diagnosis of Rice Diseases using Deep Learning", Frontiers in Plant Ccience, Vol. 12, 2021, pp. 1-15.

[3] G. Kathiresan, M. Anirudh, M. Nagharjun, R. Karthik, "Disease detection in rice leaves using transfer learning techniques", Journal of Physics: Conference Series, Vol. 1911, 2021, pp. 1-10.

[4] K. U. V. R. Teja, B. P. V. Reddy, L. R. Kesara, K. D. P. Kowshik, L. A. Panchaparvala, "Transfer Learning based Rice Leaf Disease Classification with Inception-V3," Proceedings of the International Conference on Smart Generation Computing, Communication and Networking, Pune, India, 29-30 October 2021, pp. 1-6.

[5] N. Krishnamoorthy, V. Parameswari, "Rice Leaf Disease Detection Via Deep Neural Networks with Transfer Learning for Early Identification", Turkish Journal of Physiotherapy and Rehabilitation, Vol. 32, No. 32, 2021, pp. 1087-1097.

[6] Md. A. Islam, Md. NymurRahmanShuvo, M. Shamsojjaman, ShazidHasan, Md. S. Hossain, T. Khatun, "An Automated Convolutional Neural Network based Approach for Paddy Leaf Disease Detection", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 1, 2021, pp. 280-288.

[7] R. Swathika, S. Srinidhi, N. Radha, K. Sowmya, "Disease Identification in Paddy Leaves using CNN based Deep Learning", Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, Tirunelveli, India, 4-6 February 2021, pp. 1004-1008.

[8] S. P. S. Shivam, I. Kumar, "Rice Plant Infection Recognition using Deep Neural Network Systems", Proceedings of the 3rd International Semantic Intelligence Conference, New Delhi, India, February 25-27 2021, pp. 384-393.

[9] V. K. Shrivastava, M. K. Pradhan, M. P. Thakur, "Application of Pre-Trained Deep Convolutional Neural Networks for Rice Plant Disease Classification", Proceedings of the International Conference on Artificial Intelligence and Smart Systems, Coimbatore, India, 25-27 March 2021, pp. 1023-1030.

[10] D. S. Rao, N. Kavya, S. N. Kumar, L. Y. Venkat, N. P. Kumar, "Detection and Classification of Rice Leaf Diseases using Deep Learning", International Journal of Advanced Science and Technology, Vol. 29, No. 3, 2020, pp. 5868- 5874.

[11] H. D. Prasetyo, H. Triatmoko, Nurdiansyah, I. N. Isnainiyah, "The Implementation of CNN on Website-based Rice Plant Disease Detection", Proceedings of the International Conference on Informatics, Multimedia, Cyber and Information System, Jakarta, Indonesia, 19-20 November 2020, pp. 75-80.

[12] J. Chen, J. Chen, D. Zhang, Y. Sun, Y. Nanehkaran, "using Deep Transfer Learning for Image-based Plant Disease Identification", Computers and Electronics in Agriculture, Vol. 173, No. 105393, 2020.

[13] D. Li, R. Wang, C. Xie, L. Liu, J. Zhang, R. Li, F. Wang, M. Zhou, W. Liu, "A Recognition Method for Rice Plant Diseases and Pests Video Detection Based on Deep Convolutional Neural Network", Sensors, Vol. 20, No. 3, 2020, pp. 1-21.

[14] M. KalaiPriya, S. Dhanabal, "Analyses of Nine Different Types of Diseases in Paddy with Hybrid Algorithms using Deep Learning", International Journal of Engineering Research & Technology, Vol. 8, No. 8, 2020, pp. 1-7.

[15] R. Sharma, V. Kukreja, V. Kadyan, "Hispa Rice Disease Classification using Convolutional Neural Network", Proceedings of the 3rd International Conference on Signal Processing and Communication, Coimbatore, India, 13-14 May 2021, pp. 377-381.

[16] S. Ghosal, K. Sarkar, "Rice Leaf Diseases Classification using CNN with Transfer Learning", Proceedings of the IEEE Calcutta Conference, Calcutta, India, 28-29 February 2020, pp. 230-236.

[17] S. M. M. Hossain, Md. M. M. Tanjil, M. A. B. Ali, M. Z. Islam, Md. S. Islam, S. Mobassirin, I. H. Sarker, S. M. R. Islam, "Rice Leaf Diseases Recognition using Convolutional Neural Networks", Advanced Data Mining and Applications, Springer International Publishing, 2020, pp. 299-314.

[18] S. Ramesh, D. Vydeki, "Recognition and Classification of Paddy leaf Diseases using Optimized Deep Neural network with Jaya algorithm", Information Processing in Agriculture, Vol. 7, No. 2, 2020, pp. 249-260.

[19] Wj. Liang, H. Zhang, Gf. Zhang, C. Hong-xin, "Rice Blast Disease Recognition using a Deep Convolutional Neural Network", Scientific Reports, Vol. 9, No. 2869, 2019, pp. 1-10.

[20] M. F. X. Cham, R. Tanone, H. A. T. Riadi, "Identification of Rice Leaf Disease using Convolutional Neural Network Based on Android Mobile Platform", Proceedings of the 2nd International Conference on Innovative and Creative Information Technology, Salatiga, Indonesia, 23-25 September 2021, pp. 140-144.

[21] K. Anandhan, A. S. Singh, "Detection of Paddy Crops Diseases and Early Diagnosis using Faster Regional Convolutional Neural Networks", Proceedings of the International Conference on Advance Computing and Innovative Technologies in Engineering, Greater Noida, India, 4-5 March 2021, pp. 898-902.

[22] S. Z. M. Zaki, Md. A. Zulkifley, M. Md. Stofa, N. A. M. Kamari, N. A. Mohamed, "Classification of Tomato

Leaf Diseases using Mobile Net v2", IAES International Journal of Artificial Intelligence, Vol. 9, 2020, pp. 290-296.

[23] A. A. Ahmed, G. H. Reddy, "A Mobile-Based System for Detecting Plant Leaf Diseases using Deep Learning", Agri Engineering, Vol. 3, 2021, pp. 478-493.

[24] N. Aishwarya, N. G. Praveena, S. Priyanka, J. Pramod, "Smart Farming for Detection and Identification of Tomato Plant Diseases using Light Weight Deep Neural Network", Multimedia Tools and Applications, Vol. 82, 2023, pp. 18799-18810.

[25] P. Kohli, I. Kumar, V. Vimal, "Plant Leaf Disease Identification Using Unsupervised Fuzzy C-Means Clustering and Supervised Classifiers", Cyber Security in Intelligent Computing and Communications.Studies in Computational Intelligence, Springer, Singapore, Vol 1007, 2022.

[26] K. Sivachandar, V. Thulasi Bai, L. Simone, G. Dimauro, "Semantic Segmentation of Conjunctiva Region for Non-Invasive Anemia Detection Applications", Electronics, Vol. 9, No. 8, 2020, pp. 1309-1322.

[27]D. Verma, D. Bordoloi, V. Tripathi, "Plant Leaf Disease Detection Using Mobilenetv2", Webology, Vol. 18, No. 5, 2021.

[28] P. Juyal and S. Sharma, "Detecting the Infectious Area Along with Disease Using Deep Learning in Tomato Plant Leaves", Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, Thoothukudi, India, 3-5 December 2020, pp. 328-332.

Achieving Information Security by multi-Modal Iris-Retina Biometric Approach Using Improved Mask R-CNN

Original Scientific Paper

Mohamed. A. El-Sayed
Taif University, Applied College, Department of Technology
Taif 21974, KSA
m.sayed@tu.edu.sa

Mohammed A. Abdel- Latif
South Valley University, Faculty of Science, Department of Mathematics
Qena 63514, Egypt
m_latif81@yahoo.com

Abstract *– The need for reliable user recognition (identification/authentication) techniques has grown in response to heightened security concerns and accelerated advances in networking, communication, and mobility. Biometrics, defined as the science of recognizing an individual based on his or her physical or behavioral characteristics, is gaining recognition as a method for determining an individual's identity. Various commercial, civilian, and forensic applications now use biometric systems to establish identity. The purpose of this paper is to design an efficient multimodal biometric system based on iris and retinal features to assure accurate human recognition and improve the accuracy of recognition using deep learning techniques. Deep learning models were tested using retinographies and iris images acquired from the MESSIDOR and CASIA-IrisV1 databases for the same person. The Iris region was segmented from the image using the custom Mask R-CNN method, and the unique blood vessels were segmented from retinal images of the same person using principal curvature. Then, in order to aid precise recognition, they optimally extract significant information from the segmented images of the iris and retina. The suggested model attained 98% accuracy, 98.1% recall, and 98.1% precision. It has been discovered that using a custom Mask R-CNN approach on Iris-Retina images improves efficiency and accuracy in person recognition.*

Keywords*: Biometrics, iris recognition; retina recognition; mask R-CNN*

1. INTRODUCTION

The emergence of biometric-based authentication is a consequence of the development of autonomous technologies and the growing threat to individual security. Variation in biometric behavioral and physical features facilitates the accurate recognition of individuals. The most common biometric features [1] are fingerprint [2], face [3], voice [4], iris [5], and retina [6], among others, but only medium-level security has been provided by features such as voice, face, and fingerprint due to limitations such as environmental changes and readily recordable nature. Using biometric features such as the iris and retina, a person can be identified with high certainty [7]. The iris is responsible for the amount of light that passes through the retina, which is the latter layer of the eye. Uniqueness and inaccessibility contribute to the increased robustness of these features.

To extract the features necessary for human recognition based on the iris, the inner and peripheral boundaries must be precisely localized. Various limitations, such as the reflection of light, the low resolution of image quality, and the presence of occlusions, can compromise the accuracy of recognition. These constraints must be considered for accurate iris segmentation. Two methods, namely light-based and near-infrared (NIR) based imaging, were used to acquire images of the iris; the NIR-based images were found to be of higher quality and have fewer light reflections [7]. For recognition purposes, the texture-based characteristics of the iris were extracted and used to create a template for each individual. Similarly, the segmentation of retinal blood vessels and extraction of both vascular and non-vascular features assists in the recognition of individuals.

Machine learning techniques were found to provide fast and accurate extraction of features to facilitate the

effective identification of individuals [8]. These techniques are used for segmentation in a variety of applications, including medical imaging detection for tumor locating [9], predicting tissue volumes [10], epilepsy and surgery planning, and so on [11], and interpretation of various important objects in satellite images, such as buildings, roads, forests, crops, and cloud detection [12].

The segmentation of the iris is the process of precisely recognizing and defining the boundary of the iris inside an image in biometric systems. Custom Mask R-CNN is an excellent deep-learning solution for this problem because it extends the well-known Mask R-CNN architecture by integrating a special segmentation head to segment the iris.

The iris is a distinct and dependable biometric feature that can be used to accurately identify individuals. Iris segmentation is an important stage in biometric systems because it provides an accurate representation of the iris pattern that may be used for recognition. This task can be handled with excellent accuracy and effectiveness by modern deep learning architecture Custom Mask R-CNN. [13]

Mask R-CNN is an important deep learning-based scheme that has been used for a variety of applications in the past. This included automated nucleus segmentation [14], detection of disease in lung nodules segmentation [15], segmentation of liver disease based on multimodal [16], automated counting of blood cells [17], segmentation and detection of the human face [18], detection of oral disease based on X-ray images [19], and detection and classification of breast tumors [20]. To the best of our knowledge, the Mask R-CNN method for the identification and authentication of humans from Iris and Retina images has not been investigated.

This paper describes an efficient multimodal biometric system based on iris and retinal features. This biometric recognition system integrates iris and retina features for human identification and authentication utilizing multiple segmentation approaches. Prior to enhancement, the input images are segmented. The features are extracted and then fused to create a multimodal feature vector that represents the individual's distinct characteristics. Finally, the feature vector is compared to a pre-registered database of individuals using various matching methods during the matching step. The system calculates a match score which represents the degree of similarity between the input feature vector and the registered database. This method takes advantage of the iris's and retina's unique and distinguishing traits, which are highly accurate and difficult to forge or manipulate. The proposed model achieved high recognition accuracy.

2. RELATED WORKS

Several studies used these approaches to segment the iris and retinal blood vessels for effective feature extraction [21-25]. These approaches address the constraints of classic feature extraction methods to perform authentication without reproducing the images. A unimodal biometric system is defined as the utilization of any one of the biometric features for human identification [26, 27]. This type of biometric technology appears to be compatible with the identification of a limited number of users, but it has some disadvantages when applied to a real-time scalable set of populations.

In [28], the authors proposed a multimodal biometric system based on the iris and retina for comparing spatial features. The spatial features are initially extracted from the iris and retina individually. The features are collected and stored in the database. Then Levenshtein distance (LD) algorithm is proposed to calculate the distance between the minutiae for comparison. The comparison score is evaluated for the iris and retina individually. Finally, two biometric comparison scores are fused and provide a result for authentication. They tested numerous combinations of publicly available biometric databases and found that the optimal False Acceptance Rate and False Rejection Rate indices were 0 and 3.33%, respectively.

In [29], the author used a Deep learning-based unified framework to accurate detection, recognition, and segment the iris from raw eye images. Iris-specific Mask R-CNN, a normalization layer, and feature extraction are all part of the proposed system design. Mask R-CNN is used to segment the iris. The training samples are displayed with ground truth images. Before learning the features of the iris, the normalization layer performs iris and mask normalization. After detecting the iris, a contrast enhancement process is presented to improve the image quality. FeatNet's fully connected layer is used to extract the features. The Extended Triplet Loss (ETL) function is used to learn the features. The experiments are carried out using publically available datasets.

In [30], this paper proposed an iris segmentation method for noisy images. The proposed system includes four processes: pupillary boundary marking, marking noisy regions, marking the limbic boundary, and iris boundary regularization. Shape features are extracted in the first procedure to help localize the pupillary boundary. The second procedure marks noisy spots in photos to improve performance. The third procedure uses the Gaussian smoothening filter and rank filters to mark the limbic boundary. Finally, the circular iris boundary is converted to a non-circular iris boundary by a regularization technique. Three public datasets are used to test the results.

In [31], this paper proposed retina diagnosis utilizing biometric methods with SVM and clustering methods. The main aim of this research is to extract the features from retina images. Preprocessing is the initial step in increasing image quality. It comprises denoising, contrast enhancement, and normalization. The following step is vessel segmentation, binarization, and thinning, which is accomplished using a Gaussian-matched filter with binarization and local entropy thresholding. The final step is minutiae extraction, which is accomplished

using a crossing number technique. The proposed method automatically counts the number of minutiae after extracting them, which aids in disease diagnosis. SVM is used to determine whether or not the identified eyeballs are healthy or unhealthy.

The typical datasets show good performance using the established methodologies; nonetheless, many issues have been identified, as outlined below.

In [28], features for the iris and retina are retrieved independently. The authors used the derived features directly from the images without reducing noise, which degrades the image quality and hence the process's accuracy. They do not pay attention to iris and retina localization or image quality evaluation, which increases detection time and decreases detection accuracy. The calculation of feature scores takes a lengthy time. Furthermore, it is unstable when a trained massive number of features is necessary to calculate a score for a massive number of images.

In [29], In this paper, the Iris detection is still poor due to the use of noisy images, which diminish image quality and consequently affect detection accuracy. Because Iris detection accuracy is dependent on image quality, noise in the images must be removed to obtain an accurate result. The author employed the Mask R-CNN network for iris segmentation, but it does not provide reliable localization because spatial information was neglected in favor of semantic information. Furthermore, it disregards the link between the pixels on the object's edge, resulting in poor localization.

In [30], this paper is not suitable for intensity-varying iris images due to the lack of normalization. The limbic boundary is marked using a Gaussian smoothing filter; however, it takes a long time to identify, increasing the processing cost. Also, noise factors are removed for effective segmentation, but it still has salt and pepper noise, which reduces image quality, contrast, and resolution, resulting in blurred images and lower segmentation accuracy.

In [31], The authors used SVM to determine whether the eyes are healthy or unhealthy, which is only appropriate for smaller datasets because larger datasets require more time for training and validation, increasing latency and kernel selection is required to optimize otherwise classification accuracy is reduced. The Gaussian matched filter is used for iris segmentation, which takes a long time. Because of a lack of important features such as color, statistical, texture, and form features, retina classification accuracy is still low.

In [32] A biometric method based on finger vein and face bimodal feature layer fusion is proposed, with the fusion occurring in the feature layer and using a convolutional neural network (CNN). The self-attention mechanism is employed to get the weights of the two biometrics, and the self-attention weight feature is cascaded using the bimodal fusion feature channel Concat when combined with the RESNET residual structure. AlexNet and VGG-19 network models were chosen for extracting finger vein and facial image features as inputs to the feature fusion module in the experimental phase to demonstrate the high efficiency of bimodal feature layer fusion. The experiments demonstrate that both models' recognition accuracy exceeds 98.4%.

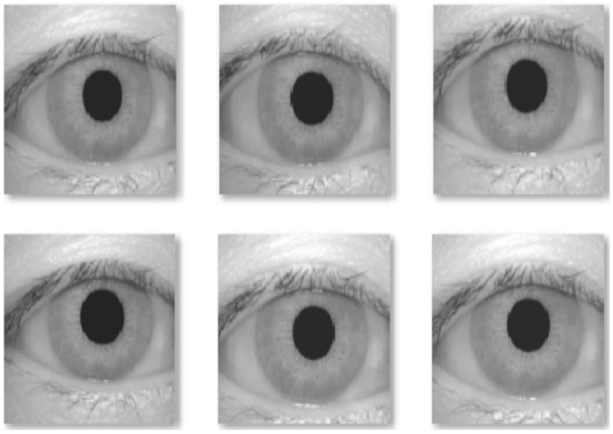
In this paper, we introduced a multimodal for identifying the same individual using two images. The first image uses customized Mask RCNN to segment the Iris region and extract informative features, while the second image uses deep neural networks based on pre-trained neural network models such as VGG16, VGG19, ResNet101, and MobileNet [33]. The fusion of the above features of a specific person is worthy of understanding the person's identity. A new technique for human recognition can be proposed using a pre-trained model and a tailored Mask-RCNN.

3. MATERIALS AND METHODS

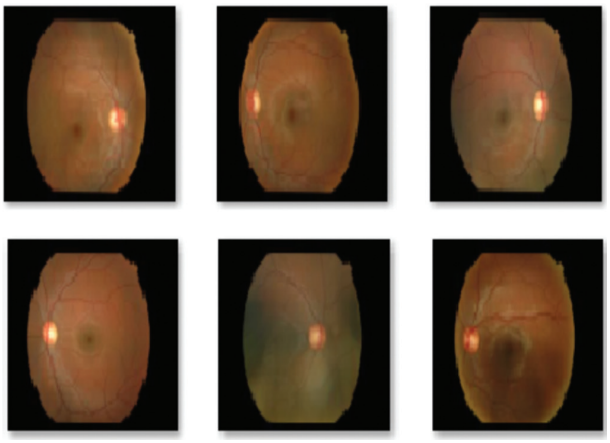
3.1. DATASET

The two main databases used in this paper comprise colored retina and iris images. Retina images are obtained by a retinograph with or without pupil dilation during repeated clinical examinations. The MESSIDOR dataset [34] is well-known from the image database. The MESSIDOR database was a financed research study introduced by the French Ministry of Research and Defense as part of the 2004 TECHNO-VISION program [34]. In the databases, we used five images for each individual. The images will be saved in uncompressed TIFF format with a resolution of 1440 960 pixels (about 4 MB for each image).

Dataset CASIA-IrisV1 database [35] contains 765 iris images from 108 images. All images are collected in BMP format with a pixels resolution of 320 × 280. The images are collected by the research group, the National Laboratory of Pattern Recognition (NLPR), Institute of Automation (IA). Chinese Academy of Sciences (CAS). which was released to provide an iris database freely for recognition researchers by captures in homemade. Fig. 1 shows simple samples of iris and retina images.



(a)



(b)

Fig. 1. Examples of images for iris CASIA-IrisV1 database (a) and retina MESSIDOR database (b).

3.2 METHODOLOGY

Fig. 2 depicts the general layout for human identification based on a biometric approach. We tested the approach on 100 individuals, each with five iris and retina images.

3.2.1. Preprocessing

Denoising. The iris and retina images contain noise and other artifacts that degrade image quality since they must be removed. Denoising is used to improve image quality, while illumination is used to remove noisy pixels from images. The noise in the iris and retina images was eliminated independently. Speckle noise in iris pictures and Gaussian noise in retina images are reduced using the Fast Guided Filter (FGF) [36], which is computationally efficient and considerably superior to a wiener filter, local means filter, and Bilateral filter.

FGF is marked by preserving the edges of the objects in the Image by using the content of the guidance image [37].

Boundary Localization and Segmentation Using Custom Mask R-CNN. Accurate iris and retina detection and localization is a critical step in achieving high detection accuracy in iris and retina-based recognition. In this article, boundary detection is a tough procedure since it varies from person to person and helps in precisely recognizing the iris. Initially, the center point and radius are retrieved from denoised or preprocessed iris images. Based on such information, the inner and outer boundaries of the iris are located and done using a customized Mask R-CNN. Traditional custom Mask R-CNN emphasized on pixel edge information and ignored object borders and shapes, resulting in poor localization. To that end, we present a custom Mask R-CNN that localizes the inner and outer boundaries by extracting local features (boundary information) and masking the boundary. Mask R-CNN enhances the mask head with local features in the Mask R-CNN. FPN (Feature Pyramid Network) is utilized here to extract and build the pyramidal features for RPN (Region Proposal Network) and proceed to the second step. The mask head learns the boundaries of the mask and the object and updates the information in the Mask R-CNN. For better segmentation results, the learning information is optimized using the Adam optimizer, which takes less memory and is more computationally efficient than gradient descent and stochastic gradient descent algorithms.

It extracts the blood vessels from the retina for segmentation. Vessel Extraction of Retinal Fundus Image Using Principal Curvature [38], which results in a segmented binary image. MATLAB 2020a can be used to obtain the proposed blood vessel segmentation methods based on adaptive principal curvature and derivative operators [39].

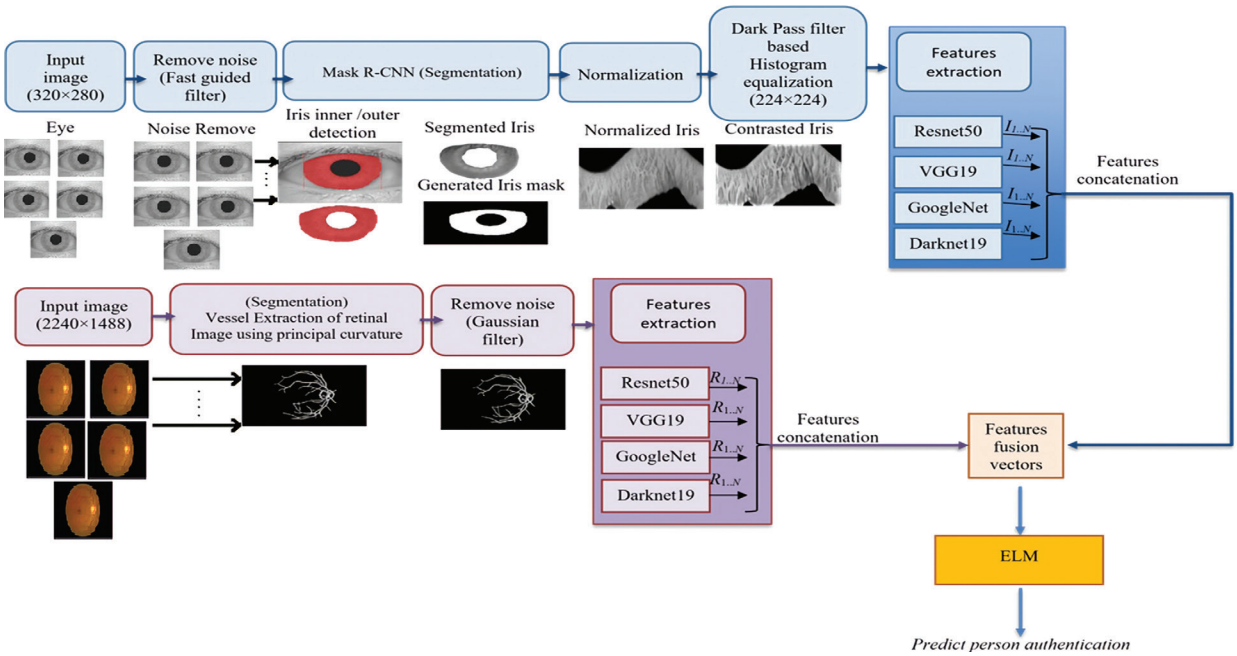


Fig. 2. The general methodology layout for human identification is based on a biometric approach.

Normalization and Contrast Enhancement. The majority of normalizing approaches, including the Daugman rubber sheet model, have been applied for iris recognition [40]. Because the size of the iris varies from person to person due to variations in intensity, the Iris image will need to be normalized. The iris is converted into a rectangular zone for this purpose, which normalizes the intensity of the Iris images.

Contrast enhancement is a method that improves the quality of normalized images by increasing brightness and increasing detection accuracy [41]. We proposed Dark Pass Filter-based Histogram Equalization for this purpose. The low-contrast image is divided into sub-histograms based on the median of each partition. The Histogram is needed for segmentation; thus, it must go across all of the pixels. Where the histogram for an image is produced and parts for the image are created based on the peak of the histogram. This method is repeated until no more pieces of an image emerge [42].

3.2.2. Feature Extraction

A lightweight deep learning model is proposed in this research to extract features from iris and retina images. We presented an Attention-based pretrained deep learning model (VGG19, ResNet 50, GoogleNet, and DarkNet 19) for feature extraction [43]. The key advantage of using a pretrained model is that it is easier to use and provides free network architecture. Furthermore, they often produce superior results for object identification and require less training. Tab 1 displays the training settings for the Custom pretrained model training in detail.

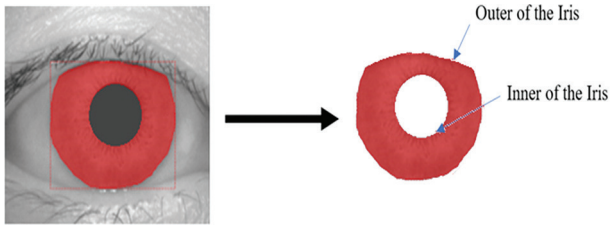


Fig. 3. Inner and Outer of Iris image detection by Custom Mask R-CNN

The computing system is configured with Windows 11 Pro, Intel Core i7, 1.6GHz, GPU Nvidia GeForce 8GB, 3295MHz DDR3 RAM memory. When the algorithm is implemented on more powerful computing platforms, the execution time will be reduced.

Table 1. Hyperparameters details

Parameter	value
Epochs	100
Learning rate	0.001
Bach size	8
Optimization	Adam
Gradient Threshold	1

In this research, we extract the feature from the fully connected layer prior to the classification process's final layer. The features are extracted from the iris and retina independently. Then concatenation of all the output features vectors from each pretrained model from both Iris and Retina $I_{N\times1}$ and $R_{N\times1}$, because it makes output features more informative for a person. For iris recognition, texture features are extracted from iris images. We extract features from both iris and retina images based on the weight values, which enhances detection accuracy.

3.2.3. Fusion Score Evaluation

The feature fusion technique is used to evaluate the feature fusion score at this step. The collected features from the segmented iris and retina are concatenated and sent into the Extreme Learning Machine (ELM) for recognition. We did feature fusion here and calculated the comparison score for that feature. We introduced the kernel aware Extreme Learning Machine (ELM) method [43] for this purpose, which randomly selects the hidden nodes and calculates the weight value of the inputs. The kernel selection procedure is critical for generating correct results, and the kernel-aware ELM method may select related kernel functions based on the inputs. As a result, based on the feature fusion score, our method achieves excellent detection accuracy, and human recognition is done.

4. RESULTS

The model was built with the Keras and TensorFlow frameworks, as well as the MATLAB toolbox for preprocessing and feature extraction. We initialized the model by preprocessing all Iris and Retina images and utilizing pretrained weights COCO with extension file *.h5. We then used transfer learning to modify the model on the segmented Iris and Retina images. For the experiment, we used a 70:30 split, done at random, and divided the images into training (70%) and test (30%) groups.

During experimentation, the performance of the proposed study is evaluated using the following metrics: Accuracy, Sensitivity/True positive rate, Specificity/True negative rate, Precision/Positive predictive value, Negative Prediction value, and F1-Score.

In this research, we use Average Precision (AP) to evaluate Mask R-CNN's performance on diverse datasets. Precision is generally defined as Eqs. (1) to (4).

$$Precision = \frac{TP}{TP+FP} = \frac{true\ positive}{total\ person\ identification} \quad (1)$$

where, True Positive (TP): A correctly identified person based on Iris and Retina traits. False positive (FP): An incorrectly identified person. True negative (TN): The correct identification of a negative person. False negative (FN): An inaccurate identification of a negative person. Average precision (AP) is frequently referred to as Average Precision for COCO datasets or COCO type datasets. To properly comprehend AP, we require Recall, Precision accuracy, and F1-score, which is defined as

$$Recall = \frac{TP}{TP+FN} = \frac{true\ positive}{total\ ground\ truth\ identification} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

The findings of the human recognition evaluation are presented in Table 2. In the first test case, we use the extracted features from the blood vessels segmentation approach from Retina images based on the adaptive principal curvature for human recognition. In the second test case, we use the extracted features from the output of the Iris images from the applied custom Mask R-CNN for identification, and in the third example, we use the extracted features from both Iris and Retina images based on the proposed system. According to the findings, the case of Iris and Retina using the proposed method may reliably identify the person with a 98% accuracy.

Table 2. Presents the human recognition evaluation results

Test Case	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Retina only	90	91.5	91.7	91
Iris only	93	92.5	92.5	92.7
Iris-Retina	98	98.1	98.1	97.7

The dataset has been examined with multiple models in order to determine the optimal model for this system. Both the concatenated features of retina and iris images. Each model is trained for 100 epochs, and the train\test accuracy graph alters the model. Fig. 3 depicts the performance versus training and test-validation set of the presented models. As seen in Fig. 3, the proposed system produces superior results.

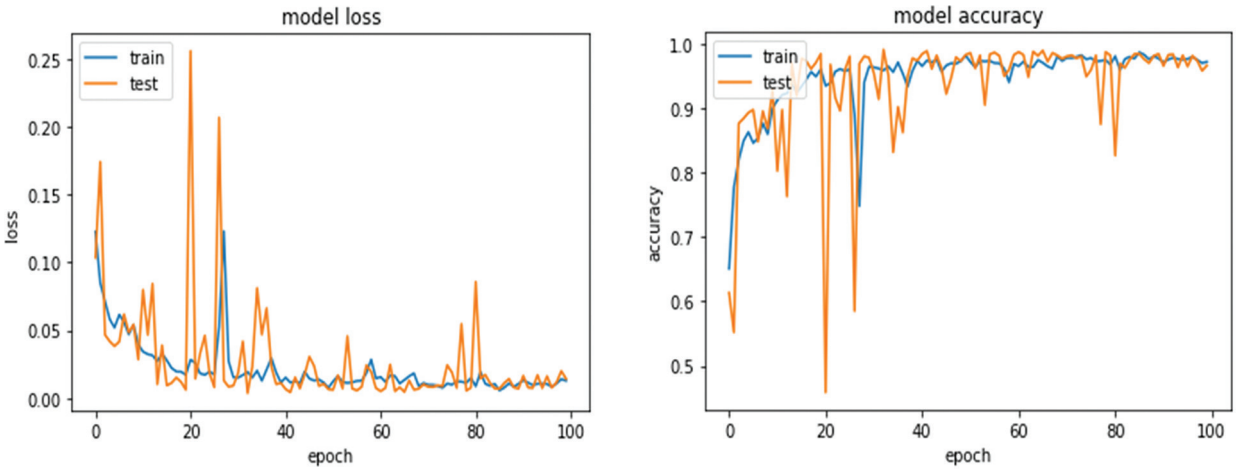


Fig. 4. Figure 3: model accuracy and loss.

5. DISCUSSION AND COMPARISONS

To achieve a high level of information security, the proposed approach employs multimodal iris-retina biometric features via an improved Mask-RCNN model. Accuracy, Precision/Positive predictive value, Negative Prediction value, and F1-Score were calculated. Human recognition attained higher accuracy in the case of iris-retina than either retina or iris alone, according to the data. The lowest recognition accuracy of the experimental results in [32], a multimodal Convolutional Neural Network Approach Based on the Fusion of Face and Finger Vein Features, was 98.84%, and the greatest recognition accuracy was 99.98%. Furthermore, in [28], The Levenshtein distance is used for spatial feature comparison in a multimodal retina-iris biometric system, and a comparison of different iris-retina multimodal techniques is shown. The systems guarantee that multimodal approaches outperform unimodal approaches in terms of accuracy. Our proposed method emphasizes the aforementioned results, where recognition accuracy exceeded 98%. In the future, we can test the proposed approach on different datasets with varied percentages of training and testing data.

6. CONCLUSIONS

We introduced a deep learning strategy for custom Mask-RCNN for precise and automatic segmentation of Iris from images and segmenting blood vessels from Retina images using adaptive principal curvature and derivative in this study. We present several powerful pretrained models for extracting informative features from segmented iris and retina images. We tested our framework on MESSIDOR and CASIA-IrisV1 databases and performed cross-dataset validation on the databases to demonstrate the model's robustness. The results show that an improved Mask-RCNN can compute deep features with effective representation of human recognition more accurately using retina and iris images for the same individual than using simply iris images, or retina images. Fusion is calculated in our work feature using ELM, which determines the feature fusion score, thereby increasing iris and retina recognition.

ACKNOWLEDGMENT

The researchers would like to acknowledge the Deanship of Scientific Research, Taif University for funding this work.

7. REFERENCES:

[1] N. Alay, H. H. Al-Baity, "Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits", Sensors, Vol. 20, No.19, 2020, pp. 5523-5540.

[2] M. Chhabra, K. K. Ravulakollu, M. Kumar, A. Sharma, A. Nayyar, "Improving automated latent fingerprint detection and segmentation using deep convolutional neural network", Neural Computing and Applications, Vol. 35, No. 9, 2023, pp. 6471-6497.

[3] M. Zulfiqar, F. Syed, M. J. Khan, K. Khurshid, "Deep face recognition for biometric authentication", Proceeding of the 1st International Conference on Electrical, Communication and Computer Engineering, Swat, Pakistan, 24-25 July 2019, pp. 1-6.

[4] A. Abozaid, A. Haggag, H. Kasban, M. Eltokhy, "Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion", Multimedia tools and applications, Vol. 78, No.12, 2019, pp. 16345-16361.

[5] M. El-Sayed, M. Abdel-Latif, "Iris recognition approach for identity verification with DWT and multiclass SVM", PeerJ Computer Science, Vol. 8, 2022, p. e919.

[6] S. Roy, P. Dutta, A. Bhowmik, B. Roy, K. Sourav, L. Kumari, "Identification of medical disorders in eye and biometric authentication analysis with iris retina scan using machine learning", Biotechnology and Biological Sciences, 1st Edition, CRC Press, 2019, pp. 29-33.

[7] D. Tirfe, V. K. Anand, "A survey on trends of two-factor authentication", Contemporary Issues in Communication, Cloud and Big Data Analytics, Springer, 2022, pp. 285-296.

[8] S. Villon, D. Mouillot, M. Chaumont, E. S. Darling, G. Subsol, T. Claverie, S. Villéger, "A deep learning method for accurate and fast identification of coral reef fishes in underwater images", Ecological informatics, Vol. 48, 2018, pp. 238-244.

[9] C. Tao, K. Chen, L. Han, Y. Peng, C. Li, Z. Hua, J. Lin, "New one-step model of breast tumor locating based on deep learning", Journal of X-ray Science and Technology, Vol. 27 No. 5, 2019, pp. 839-856.

[10] E. Eisenberg et al. "Deep learning–based quantification of epicardial adipose tissue volume and attenuation predicts major adverse cardiovascular events in asymptomatic subjects", Circulation: Cardiovascular Imaging, Vol. 13, No. 2, 2020, pp.1-9.

[11] B. Abbasi, D. M. Goldenholz "Machine learning applications in epilepsy", Epilepsia, Vol. 60, No. 10, 2019, pp. 2037-2047.

[12] A. Asokan, J. Anitha, "Machine learning based image processing techniques for satellite image analysis - A survey", Proceeding of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, Faridabad, India, 14-16 February 2019, pp. 119-124.

[13] C. D. Divya, A. B. Rajendra, "An investigation of IRIS Recognition Techniques: A Literature Survey", Applied and Computational Engineering, Vol. 2, 2023, pp. 663-669

[14] J. W. Johnson, "Automatic Nucleus Segmentation with Mask-RCNN", Advances in Intelligent Systems and Computing, Springer, 2020, pp. 399-407.

[15] E. Kopelowitz, G. Engelhard "Lung nodules detection and segmentation using 3D maskRCNN", Medical Imaging with Deep Learning, arXiv:1907.0861, 2019.

[16] S. Mulay, G. Deepika, S. Jeevakala, R. Keerthi, S. Mohanasankar, "Liver segmentation from multimodal images using HED-mask R-CNN", International Workshop on Multiscale Multimodal Medical Imaging, Springer, 2019, pp. 68–75.

[17] N. Dhieb, H. Ghazzai, H. Besbes, Y. Massoud, "An automated blood cells counting and classification framework using mask R-CNN deep learning model", Proceeding of the 31st International Conference on Microelectronics, Cairo, Egypt,15-18 December 2019, pp. 300-303.

[18] K. Lin, H. Zhao, J. Lv, C. Li, X. Liu, R. Chen, R. Zhao, "Face detection and segmentation based on improved mask RCNN", Discrete Dynamics in Nature and Society, Vol. 2020, 2020, pp. 1-11.

[19] R. Anantharaman, M. Velazquez, Y. Lee "Utilizing mask R-CNN for detection and segmentation of oral diseases", Proceeding of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, Madrid, Spain, 3-6 December 2018, pp. 2197–2204.

[20] S. Podder, S. Bhattacharjee, A. Roy, "An efficient method of detection of COVID-19 using Mask R-CNN on chest X-Ray images", AIMS Biophysics, Vol. 8, No. 3, 2021, pp. 281-290.

[21] L. Latha, S. Thangasamy, "A robust person authentication system based on score level fusion of left-right irises and retinal features", Procedia Computer Science, Vol. 2, 2010, pp. 111-120.

[22] Z. Waheed, M. U. Akram, A. Waheed, A. Shaukat "Robust extraction of blood vessels for retinal recognition", Proceeding of the Second International Conference on Information Security and Cyber Forensics, Cape Town, South Africa, 15-17 November 2015, pp. 1-4.

[23] K. S. Bhagat, P. B. Patil, R. R. Deshmukh, D. K. Kirange, S. Waghmare, "Iris recognition using radon transform and GLCM", Proceeding of the International Conference on Advances in Computing, Communications and Informatics, Udupi, India, 13-16 September 2017, pp. 2257-2263.

[24] A. Noorfizir, M. N. Rachmatullah, G. Sulong "Hybrid multilevel thresholding-otsu and morphology operation for retinal blood vessel segmentation", Engineering Letters, Vol. 28, No. 1, 2020, pp. 180-191.

[25] Y. Ma et al. "Multichannel retinal blood vessel segmentation based on the combination of matched filter and U-Net network", BioMed Research International, Vol. 2021, 2021, pp. 1-18.

[26] K. Meena, N. Malarvizhi, "An efficient human identification through multimodal biometric system", Brazilian Archives of Biology and Technology, Vol. 59, 2017, pp. 1-10.

[27] K. Su, G. Yang, B. Wu, L. Yang, D. Li, P. Sue, Y. Yin, "Human identification using finger vein and ECG signals", Neurocomputing, Vol. 332, No. 3, 2019, pp. 111-118.

[28] V. Conti et al. "A multimodal retina-iris biometric system using the Levenshtein distance for spatial feature comparison", IET Biometrics, Vol. 10, No. 1, 2021, pp. 44-64.

[29] Z. Zhao, A. Kumar, "A deep learning based unified framework to detect, segment and recognize irises using spatially corresponding features", Pattern Recognition, Vol. 93, 2019, pp. 546-557.

[30] F. Jan, N. Min-Allah, "An effective iris segmentation scheme for noisy images", Biocybernetics and Biomedical Engineering, Vol. 40, No. 3, 2020, pp. 1064-1080.

[31] M. Szymkowski et al. "A novelty approach to retina diagnosing using biometric techniques with SVM and clustering algorithms", IEEE Access, Vol. 8, 2020, pp. 125849-125862.

[32] Y. Wang, D. Shi, W. Zhou, "Convolutional Neural Network Approach Based on Multimodal Biometric System with Fusion of Face and Finger Vein Features", Sensors, Vol. 22, No. 16, 2022, pp. 1-15.

[33] H. Pan, Z. Pang, Y. Wang, Y. Wang, L. Chen, "A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects", IEEE Access, Vol. 8, 2020, pp. 119951-119960.

[34] S. S. Majeeth, C. N. K. Babu, "Gaussian Noise Removal in an Image using Fast Guided Filter and its Method Noise Thresholding in Medical Healthcare Application", Journal of Medical Systems, Vol. 43, No. 280, 2019, pp.1-9.

[35] Dataset CASIA-IrisV1 database, Http://Biometrics.Idealtest.Org./#/datasetDetail/1 (accessed: 2022)

[36] C. C. Pham, U. Ha, J. W. Jeon "Adaptive guided image filtering for sharpness enhancement and noise reduction", Advances in Image and Video Technology, Springer, 2011, pp. 323-334.

[37] D. N. Thanh, D. Sergey, V. B. Surya Prasath, N. H. Hai, "Blood vessels segmentation method for retinal fundus images based on adaptive principal curvature and image derivative operators", Proceedings of the International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, Moscow, Russia, 13-15 May 2019, pp. 211-218.

[38] A. Iroshan, "Segmentation of Blood Vessels in Retinal Fundus Images Using Maximum Principal Curvature", File Exchange - MATLAB Cen-

tral, https://de.mathworks.com/matlabcentral/fileexchange/64884-segmentation-of-blood-vessels-in-retinal-fundus-images-using-maximum-principal-curvature (accessed: 2023)

[39] L. Ma, Y. Wang, "Iris recognition based on multichannel Gabor filtering", Proceeding of the 5th Asian Conference on Computer Vision, Melbourne, Australia, 23-25 January 2002, pp.1-5.

[40] S. Perumal, T. Velmurugan, "Preprocessing by contrast enhancement techniques for medical images", International Journal of Pure and Applied Mathematics, Vol. 118 No. 18, 2018, pp. 3681-3688.

[41] B. S. Rao, "Dynamic histogram equalization for contrast enhancement for digital images", Applied Soft Computing, Vol. 89, 2020, p. 106114.

[42] Y. Abdulazeem, H. M. Balaha, W. M. Bahgat, M. Badawy, "Human action recognition based on transfer learning approach", IEEE Access, Vol. 9, 2021, pp. 82058-82069.

[43] J. Wang, S. Lu, S. H. Wang, Y. D. Zhang, "A review on extreme learning machine", Multimedia Tools and Applications, Vol. 81, 2022, pp. 41611-41660.

A Privacy-Preserving Framework Using Hyperledger Fabric for EHR Sharing Applications

Preliminary Communication

Vidhi Thakkar
FCAIT Department, GLS University, and Research Scholar, Indus University,
Ahmedabad, Gujarat, India
vidhi@glsica.org

Vrushank Shah
Electronics and Communication Department, Indus University,
Ahmedabad, Gujarat, India
ec.hod@indusuni.ac.in

Abstract *– Electronic Health Records, or EHRs, include private and sensitive information of a patient. The privacy of personal healthcare data can be protected through Hyperledger Fabric, a permissioned blockchain framework. A few Hyperledger Fabric-integrated EHR solutions have emerged in recent years. However, none of them implements the privacy-preserving techniques of Hyperledger Fabric to make transactions anonymous or preserve the transaction data privacy during the consensus. Our proposed architecture is built on Hyperledger Fabric and its privacy-preserving mechanisms, such as Identity Mixer, Private Data Collections, Channels and Transient Fields to securely store and transfer patient-sensitive data while providing anonymity and unlinkability of transactions.*

Keywords*: EHR, Permissioned blockchain, Hyperledger Fabric, Privacy-preserving Mechanisms, User anonymity*

1. INTRODUCTION

In recent years, we have witnessed a significant rise in the adoption of technology in several sectors, including healthcare. While technology has improved the productivity of healthcare organizations, technological applications such as Electronic Health Records have also raised significant privacy concerns [1, 2]. A distributed ledger technology called blockchain has emerged in recent years as a solution to address data privacy issues. Blockchain is an append-only immutable ledger made with blocks where each block points to the subsequent block using an address. If the data in a block is altered, the block's hash will change, rendering the entire blockchain invalid. Since 2016, Blockchain-based solutions have grown more popular and have found several applications across industries [3]. Integrating blockchain technology with EHR systems can overcome the challenges of centralization, data security and privacy, data integrity, interoperability, transparency, and accountability [4]. There are various permissionless and permissioned blockchain frameworks, such as Ethereum, Enterprise Ethereum, Ripple, Hyperledger Fabric, Hyperledger Iroha, Indy, Corda, Tendermint, chain, and Quorum [5] [6]. From these, Ethereum and Hyperledger Fabric blockchain frameworks are being used for healthcare organization use cases [1]. In terms of average transaction latency, throughput, scalability, and privacy, Hyperledger Fabric beats Ethereum. Hyperledger Fabric's modularity and channel design makes it more suited to businesses and use cases where trust, transparency and security are critical.

In the last few years, several Hyperledger Fabric-integrated EHR solutions have been published [7-13]. However, none of them preserves transaction privacy and user anonymity during transaction consensus. In EHR sharing systems, privacy is more crucial. User privacy and data privacy are the two notions of privacy with respect to blockchain integrated EHR solutions.

In the existing blockchain based EHR sharing use cases, data privacy concern is partially solved by using off-chain storages such as IPFS to store actual data and only metadata (hashed values and encrypted information) is stored on blockchain ledger. However, input parameters and state data of a transactions are still visible to all the network members. User privacy is concerned with user/transaction anonymity and unlinkability. Anonymity and unlinkability of the user are crucial for maintaining user privacy since they guarantee that no one can identify the user, the endorser, or the validator. In the existing Fabric-based EHR system, the user provides their identity and signature during the transaction endorsement so that the peers can later validate it. However, this disclosure makes EHR transactions linkable. All the input parameters are visible to all the peers of the channel though encrypted. Apart from that endorsement on transaction response can reveal the identity and the organization involved in the endorsement process, leading to vulnerabilities like DDoS attacks on the transactor nodes and endorsers. To overcome such challenges, this study proposes a permissioned EHR with Hyperledger Fabric's privacy preservation techniques and a few crucial developments required to protect the privacy of nodes on the channel.

The remainder of the paper is organized as follows: Section 2 discusses some significant contributions that have been made in this field. The related literature is discussed in Section 3. Section 4 provides preliminaries regarding the privacy-preserving features of Hyperledger Fabric. Section 5 presents Hyperledger Fabric Enabled Blockchain Architecture for EHR sharing systems. Section 6 provides information regarding the tools to implement the proposed blockchain-based framework. Section 7 shows a comparison of the proposed methods against other Fabric based EHR studies. Finally, Section 8 summarizes our conclusions.

2. CONTRIBUTIONS

Hyperledger Fabric is a permissioned blockchain with various privacy features for users. It differs from other permissioned frameworks because it restricts chain code execution to a group of peers, known as endorsing peers. This subset of peers on a channel executes chaincode according to the endorsement policy. Later, the validator examines the transaction during the validation phase to verify if the desired endorser signature matches the endorsement policy. EHR sharing use case requires confidentiality while executing transactions between a subset of organizations on the network. For instance, only a specific privileged subset of organizations should be aware of the transaction existence and be permitted access to its sensitive data assets. The disclosure of the endorser's signature and endorsement policy entails several security and privacy risks. To address this concern, this paper proposes privacy preserving permissioned EHR system to ensure transaction anonymity and unlinkability.

The key contributions are summarized as follows:

- The use of the Hyperledger Fabric to improve security, transparency, privacy, and interoperability.
- The use of Arweave peer-to-peer databases to save and retrieve data permanently.
- The use of Hyperledger Fabric's Java SDK and privacy-preserving features of Fabric such as Identity Mixer, private data collections, and transient field to achieve user unlinkability and transaction data privacy during consensus.

3. RELATED WORK

This section explores the latest research on Hyperledger Fabric integrated EHR sharing systems.

Huang et al. [7] suggested Medbloc, a secure EHR sharing platform built on Hyperledger Fabric. The system was created to aid the healthcare sector of New Zealand by using the Hyperledger Fabric version 1.1 and the PBFT (Practical Byzantine Fault Tolerant) consensus mechanism. Smart-contract-based access control systems have been implemented to store encrypted medical records on the blockchain.

Mahore et al. [8] suggested a permissioned blockchain paradigm to effectively maintain medical records while achieving privacy, security, scalability, and availability. By storing sensitive patient data off-chain and only retaining the hash of the data in a blockchain record, this approach solves the scalability problems that users often encounter while using the blockchain. Moreover, they employed a proxy re-encryption approach for transmitting encrypted data to the provider. They created their solution using the PBFT consensus algorithm, the solo ordering service, and Hyperledger Fabric version V1.4. They have also statistically examined their strategy using the Hyperledger Caliper to determine how effectively it works.

Later in 2020, more studies were published to create scalable and secure EHR-sharing systems using the Hyperledger Fabric. Dubovitskaya et al. [9] presented an ACTION-EHR for cancer patients. Researchers and healthcare professionals have exchanged EMR data using the prototype constructed through Hyperledger Fabric architecture. The Patient's medical information is encrypted with public-key cryptography and kept on servers off the blockchain. The system guarantees accurate access control and data confidentiality. The ACTION-EHR blockchain network is built using the Membership Service, orderer, cloud server, Hyperledger Fabric v1.4 SDK, and CouchDB for on-chain information management and permissions.

Tith et al. [10] used consortium blockchain and Hyperledger Fabric to create a distributed solution for merging existing EHRs. Encrypted data is transmitted from the patient to the physician using a centrally located server and a proxy re-encryption method. The eID has been salt hashed to prevent transactions from

being tracked on the ledger. With Hyperledger Fabric and Hyperledger Composer, their approach has achieved scalability, transparency, traceability, availability, and dependability.

Chenthara et al. [11] published a framework for protecting electronic health records using permissioned blockchain technology. Only encrypted hashes of the records are kept on the blockchain and huge EHRs have been stored on distributed IPFS. Better assurances of data quality, scalability, privacy, and interoperability are achieved in their new paradigm. The components of Fabric's MSP, CA, PBFT consensus, CouchDB, and chaincode are used to create the system. Through this approach, they were successful in increasing scalability while assuring an elevated level of security and privacy.

Stamatellis et al. [12] developed the Hyperledger Fabric-based EHR management solution PREHEALTH to protect patient privacy. The solution discusses Identity Mixer to ensure transaction anonymity and unlikability. The approach is compliant with GDPR and emphasizes Hyperledger Fabric's privacy-preserving features. However, there has been no discussion regarding the implementation of Identity Mixer in the configuration data for Hyperledger Fabric.

Mani et al. [13] introduced PCHDM, a patient-centred healthcare data management solution. Their strategy uses off-chain IPFS to store heavy files and stores hashes of the encrypted healthcare data on-chain to ensure that the healthcare records are stored securely. Byzantine Fault Tolerance (BFT) consensus for the implementation of smart-contracts inside secure containers. Thus, prominent security, privacy and scalability are possible. Hyperledger Caliper benchmarks for transaction latency, throughput, and resource utilization are used to assess performance.

The above-mentioned Fabric-integrated EHR studies focus on providing a solution for patients to have complete control over their data. Their work utilizes Fabric components such as chain codes to define access policies, plug-and-consensus PBFT, MSP (Member Service Provider), CA (Certificate Authority), CouchDB database, and off-chain storage technologies to store patient's heavy data. Below in Table 1, we have analyzed their work using parameters such as Hyperledger Fabric version, Storage, Consensus Mechanism, Ordering Service, Immutability, Performance, Data Privacy during consensus, and User Privacy. These studies address interoperability, immutability, scalability, performance, and data protection issues of permissionless blockchain-based EHR systems. However, none of these studies has considered or implemented Hyperledger Fabric's security and privacy-preserving mechanisms to further protect the health sector's sensitive data inside the permissioned network. Only Stamatellis et al. [12] have proposed a secure EHR management solution with Hyperledger Fabric's privacy-preserving features, but they have also not demonstrated any practical implementations of their work.

Table 1. Existing Fabric integrated EHR sharing studies

Reference	Year	HF Version	Storage	Consensus/ Ordering Service	P1	P2	P3
[7]	2019	V1.1	onchain	PBFT	N	N	N
[8]	2019	V1.4	cloud	-	N	N	N
[9]	2020	V1.4	cloud	PBFT	N	N	N
[10]	2020	-	cloud	BFT	N	N	N
[11]	2020	V1.3	IPFS	PBFT	N	N	N
[12]	2020	-	-	-	N	N	N
[13]	2021	-	OrbitDB+IPFS	BFT	N	N	N

*P1= Data Privacy & GDPR, *P2= Transaction data Privacy during consensus, *P3= User privacy

4. PRELIMINARIES

4.1. HYPERLEDGER FABRIC

Hyperledger Fabric is an open-source enterprise blockchain platform developed by "The Hyperledger Foundation". Fabric is a permissioned network with a modular design, higher scalability, improved confidentiality and better performance. The Fabric enables the user to perform private transactions (private data collections and transient fields) and the development of parallel and independent lightweight channels to share data with only a portion of the nodes of the blockchain system. Fabric supports three CFT consensus algorithms: Raft, Kafka (deprecated), and Solo (deprecated). The Fabric components involve channels, peers, assets, identity, membership, transactions, chaincode, ledgers, and consensus procedures as part of the network. Participants sign up through a reputable Membership Service Provider (MSP) that keeps a list of permissioned identities made by Certificate Authorities (CA).

Hyperledger Fabric Components

MSP and CA: CA assigns certificates and works as public/private key issuers. MSP is responsible for issuing node credentials for authentication and authorization. All participating entities in Fabric must use digital certificates and cryptographic processes to identify and authenticate themselves through the MSP service.

Chaincode: Chaincode is a self-executing program. They are installed on the blockchain network and update the current state of the ledger when executed. In Fabric, the chaincode is installed on the endorser nodes only. Chaincode will be executed when certain conditions are met, and the outcomes of the transaction execution are sent to the blockchain network and subsequently added to all peers' ledgers. Fabric's chaincode logic can be written using Java, Golang, Node JS or Python.

Channels: Organizations with similar goals can be grouped into a single channel. Channel enables private and secure communication between various network parties.

Peers: The peer may be a Leader Peer (distributes block to other peers), an Anchor peer (other peers can discover and communicate with it), or an Endorser, or an Orderer, or a Validator peer.

Endorsement Policy: For a smart contract to be executed, Fabric has a configurable endorsement policy that determines the minimum number of peers that must endorse the proposal.

State Database: Endorsing Peers store output data related to endorsements in local state databases like CouchDB and LevelDB. LevelDB is a key-store database used as a state database and cannot handle complex queries. CouchDB is a document store database. Here the search is performed on keys and data values rather than only keys.

Ordering Service: The ordering service has numerous OSNs (Ordering Service Nodes). These nodes establish a global order of transactions and produce blocks for the broadcast to peers. Fabric supports Raft, Solo, and Kafka are the three ordering services. Amongst Solo and Kafka are deprecated in the latest version of Fabric, only Raft is working.

Privacy-preservation Mechanisms: Privacy protection features of Fabric include the following four aspects, namely the channels, the private data collections, the zero-knowledge proofs (ZKP) and the membership service provider. The channels keep transactions hidden from the larger network whereas PDC (Private Data Collection) helps maintain data privacy between specific groups of organizations on the channel. ZKP is for anonymous client authentication with Identity Mixer support.

4.2. CHANNEL

A Channel is a secret "subnet" which allows two or more network members to perform private and confidential transactions. Channels are used to hold various information such as transactions, chaincodes, policies, membership, and configuration. This information is utilized to maintain the network's security and privacy. Every channel has a separate distributed ledger to keep digital signatures, timestamps, i/o transaction data, and transaction IDs. But disclosure of channel data to all the network members can disclose sensitive information about a transaction. To overcome this, multiple channels can be created at once. However, the creation of a network with many channels brings down decentralization, limits interoperability, and raises network overheads.

4.3. PRIVATE DATA COLLECTIONS

Using Private Data Collection (PDC) feature, a subset of organizations over the same channel can transact with one another without worrying that another business will see their confidential information. In Fabric v2.0, an implicit collection for each organization is introduced; therefore, application chain code can be used even if the collection is not defined explicitly. Peers employ local state databases like CouchDB, MongoDB, and LevelDB to store data. Figure 1 depicts the two parts of a ledger as well as the location of the private data. The one is private data, which remains confidential and passed from peer to peer using the gossip protocol to other organizations according to the specified policy. The ordering service cannot see the private data. The other element of PDC is data hash. The Hash is endorsed, ordered, and written to each peer on the channel and serves as transaction proof and is auditable.

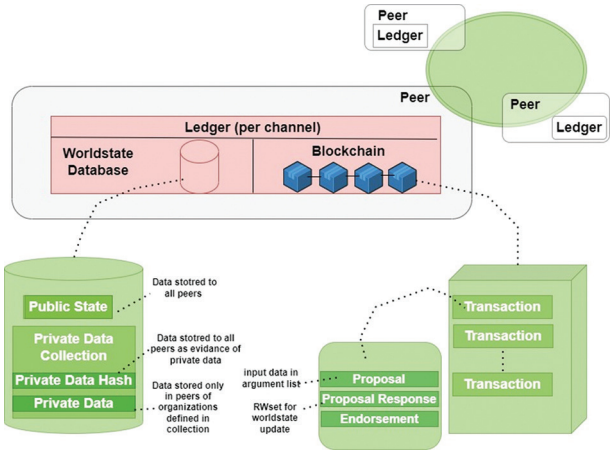


Fig 1. Ledger and location of private data

We have implemented the PDC feature for our permissioned EHR sharing application. Below is an example of creating the data definition in the chaincode. The medical data transfer sample divides the private data into three definitions and is defined in a JSON file called collection definition. The first collection is for both organizations, the second collection is only for Org1, and the third collection is only for Org2. We can check that the actual data is in the \$\$p database, whereas the hash of the actual data is in the \$\$h database.

```
Org1 and Org2 collection:
type PatientInfo struct {
PatinetID       string    `json:"patientID"`
Name string     `json:"patientname"`
Age    int       `json:"age"`
OrgName         string    `json:"Hostname"`
}
Org1 Private Details:
type Org1PrivateDetails struct {
Charges int `json:"fees"`\\fees of a doc
No_of_Patient_in_a_day int `json:"no_of_patient"`
}
Org2PrivateDetails
type Org2PrivateDetails struct {
Salary int `json:"salary"` // salary of doctor
AppraisedValue int `json:"appraisedValue"`
}
```

Discussion on PDC

DCs can be a useful way to store private and sensitive data for businesses, as the data stored on PDCs can be discarded. Unlike data on the blockchain, PDCs are not immutable, which makes them an interesting option in terms of GDPR compliance. To solve GDPR compliance with PDC blockToLive parameter can be set in config. json file.

4.4. TRANSIENT FIELD

Private Data is concerned with preserving the data privacy inside a defined subset of organizations, whereas Transient Field is an input technique for Private Data. Both are two distinct concepts from a technical standpoint, but they can be used together to achieve a certain level of security in Hyperledger Fabric Applications. The Transient Field [15] data keys will only be accessible by chain code installed on the docker container (Figure 2). The encryption/decryption key can be passed using a transient field bypassing the endorser. This key can be processed by chain code to encrypt/decrypt data and later saved to the PDC of the peer for better privacy. Transient data is passed as binary data, and base64 is encoded in the terminal. PDC and Transient Filed features can be used to achieve data confidentiality inside a Fabric network. However, using PDC for businesses has problems related to data privacy because all peers (inside and outside the subgroup) will preserve a record of private data hash as proof of data existence. There will not be any proof of transient information on the ledger.

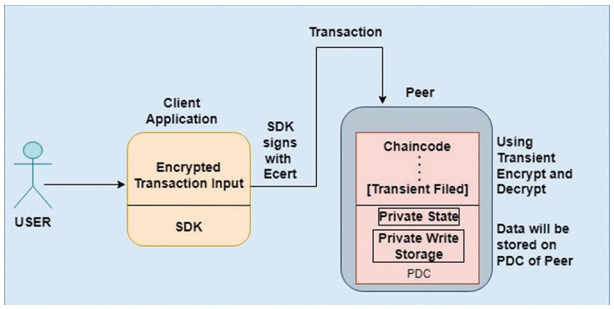


Fig 2. Example of Transient Field

4.5. IDEMIX

Idemix cryptographic protocol suite [16] uses ZKP to provide anonymity and unlinkability. It provides support for audibility and revocation. In Hyperledger Fabric, the client or user provides their identity and signature during transaction endorsement that is to be validated later by the peers. However, disclosing the identity of a user or a client to the endorsing peers for EHR transactions may compromise patient privacy and make transactions linkable. To deal with these concerns, Identity Mixer can be implemented on top of the Hyperledger Fabric. Using this Identity Mixer MSP, a user or the client can sign the transaction and remain anonymous, generating an unlinkable signature. We have implemented the Idemix concept in Hyperledger Fabric Java SDK since the support is only available for Fabric Gateway SDK for Java and Fabric Gateway Client API for Java.

Features:

Anonymity: The ability to transact without revealing the identity of the transactor.

Unlinkability: The ability of a single identity to send multiple transactions without revealing identity.

Actors:

Idemix flows involve the user, issuer and verifier as three different players. A set of user attributes are issued as a digital certificate by an issuer, also referred to as a "credential." A "zero-knowledge proof" of ownership of the credential is produced by the user, who also selectively discloses only the attributes they want to make public. With the help of the issuer's public key, the verifier can now verify the Zero Knowledge Proof provided by the user. Using the issuer's public key, the verifier verifies the proof provided by the user. The proof does not expose any further information to the verifier, issuer, or anybody else because it is zero knowledge.

Setting Up Idemix for Existing Fabric Network:

1. For Development Environments, the “idemixgen“ tool from Hyperledger Fabric Releases can be used to generate Issuer Public Key as well as Issuer Revocation Public Key.

   Steps:

   1) Generating CA Key Pair

   2) Generating Default Signer

2. For Production Environments, use “Fabric CA” as an Idemix Issuer.

   Steps:

   1) Generating CA Key Pair

   2) Generating Default Signer

Verifier:

Idemix MSP must be created using Issuers “IssuerPublicKey” & “IssuerRevocationPublicKey”. Later with the help of a “configtxgen“tool, a new Idemix Organization MSP definition in the form of JSON must be created by specifying the MSPID of the Idemix organization. Now, the Application channel configuration needs to be updated with the new Idemix organization definition. The above steps need to be done by any admin of the existing organizations in a channel. Also, a majority of the existing organizations in a channel have to approve adding a new MSP to the channel.

User:

Through Java SDK, Idemix credentials can be obtained for a user. By using this credential, users can sign and submit transactions.

Internal flow of generating Idemix credential for a user:

1. Get Certificate Authority Public Key and Revocation Authority Public Key. These are needed to request a new credential for a User. They can get from the "/cainfo"APIEndpoint of a CA.
2. This step is consisting of two steps:
   a) Registration was the same as the X509 registration.
   b) For Enrollment,
      - Get a nonce, which is required to construct a Credential Request.
      - Get an Idemix credential & CRI (Credential Revocation Information).

The process of submitting a transaction using IDEMIX

The process of submitting a transaction using Idemix Identity from CLI is shown below.

1. Generating an Idemix credential for the user

» FABRIC_PATH=~/Desktop/test

» cd $FABRIC_PATH/idemix-demo-fabric-samples/test-network/organizations/peerOrganizations/org1.example.com

» idemixgen signerconfig --ca-input=idemixmsp --admin --enrollmentId=appUser --org-unit=org1

2. Setting environmental variables for Idemix Identity

» export CORE_PEER_ADDRESS=localhost:7051

» exportCORE_PEER_TLS_ROOTCERT_FILE=$FABRIC_PATH/idemix-demo-fabric-samples/test-network/organizations/peerOrganizations/org1.example.com/peers/peer0.org1.example.com/tls/ca.crt

» export CORE_PEER_TLS_ENABLED=true

» export CORE_PEER_LOCALMSPTYPE=idemix

» exportCORE_PEER_LOCALMSPID=Org1IdemixMSP

» export CORE_PEER_MSPCONFIGPATH=$FABRIC_PATH/idemix-demo-fabric-samples/test-network/organizations/peerOrganizations/org1.example.com/idemix-config

3. Invoking the Chaincode

» cd $FABRIC_PATH/idemix-demo-fabric-samples/test-network

» export FABRIC_CFG_PATH=$PWD/./config

» peer chaincode invoke -o localhost:7050 --ordererTLSHostnameOverride orderer.example.com --tls –cafile

» ${PWD}/organizations/ordererOrganizations/example.com/orderers/orderer.example.com/msp/tlscacerts/tlsca.example.com-cert.pem -C mychannel -n basic --peerAddresses localhost:7051 –tlsRootCertFiles

» ${PWD}/organizations/peerOrganizations/org1.example.com/peers/peer0.org1.example.com/tls/ca.crt --peerAddresses localhost:9051 –tlsRootCertFiles

» ${PWD}/organizations/peerOrganizations/org2.example.com/peers/peer0.org2.example.com/tls/ca.crt -c '{"function":"TransferAsset","Args":["asset1","Vidhi"]}'

4. Query the Chaicode using Idemix Identity

» peer chaincode query -C mychannel -n basic -c '{"Args": ["GetAllAssets"]}' | jq.

Comparison with X.509

Compared to the conventional X.509, Identity Mixer signatures give advanced privacy protection. With this IDEMIX, the client can hide attributes and reveal only selected attributes. The original Enrolment id will be hidden from the orderer and other peers. We can see in Figure 3 there is an unlinkability of signatures generated as only selected attributes are disclosed.

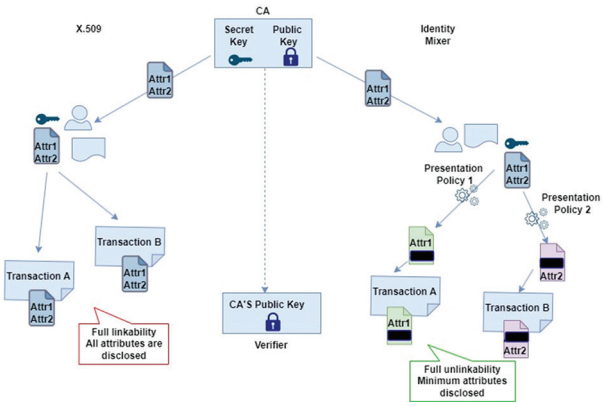


Fig. 3. Comparison with X.509

4.6 ARWEAVE DATA STORAGE

IPFS is a well-known P2P network for decentralized data exchange. Utilizing Arweave can provide permanent and tamper-proof data storage. The data is maintained and stored by a decentralized network of nodes, and they can be permanently lost if no one hosts them. We have integrated Arweave off-chain data storage technology with Hyperledger Fabric and kept encrypted data on the Arweave and the hash to the blockchain. For storing a file on Arweave using Java, read the file into a byte array and then create a transaction object with the byte array.

Example

```
Path path = Paths.get(Pathoffile);
byte[] data = Files.readAllBytes(path);
Gateway gateway = new
GatewayImpl("https://arweave.net");
Wallet wallet = Wallet.fromKeyFile(keyFilePathfromwallet);
Web3cTransaction tx = new Web3cTransaction
```

(wallet, gateway);

tx.addTag("Content-Type", mimeType);

tx.setData(data);

Upload upload = gateway.upload(tx);

5. FABRIC COMPONENTS

5.1. COMPONENTS

To develop an end-to-end blockchain system using Hyper Ledger Fabric, a complex workflow is required. The following components helped to construct the Fabric integrated permissioned EHR sharing framework [17]:

Component	Description
System	Ubuntu 18.04
Hyperledger Fabric SDK	Fabric-Java-SDK v2.4
No. of Healthcare organizations	Five organizations
No. of peers per org.	2-3
Ordering Service	Raft
World State DB	CouchDB
Concept	Docker-based
Chaincode language	Java
Front End	The web interface for the client application was created using HTML, CSS, and JavaScript
Other Integration	Hyperledger Explorer and Caliper
Access Control	It is used to restrict access to smart contract transactions based on role.
CA and MSP	X.509 certificates
Identity Mixer with ZKP	Achieve user anonymity and unlikability.
Transient field	For passing the symmetric key to chaincode through the docker as well as all the input parameters of the transaction.
Private Data Collection	For each organization to store permission lists and any private data of two or more organizations.
Chaincode	Fine-grained attribute-based access control using chaincode. 1. Create medical record 2. Save data 3. Update medical record 4. Allow doctor write 5. Update ownership (patient gives consent to another doctor) 6. Doctor fetches patient data
Encryption	Public key encryption is made possible by using Elliptical Curve Cryptography (ECC)
Docker Compose	A tool for distributing containers, which are software packages (docker-compose-ca.yaml, docker-compose-couch.yaml, docker-compose-net.yaml).

5.2 CHAINCODE SAMPLE

Create Medical Record

```
public String createMedicalRecord(Context ctx, String patientName, int age, String medicalHistory, String medications, String testResults, String owner)
{
ChaincodeStub stub = ctx.getStub();
// Generate a new ID for the medical record
String id = stub.getTxId();
// Create a new medical record object
MedicalRecord medicalRecord = new MedicalRecord(id, patientName, age, medicalHistory, medications, testResults, owner);
// Convert the medical record object to JSON
String medicalRecordJSON = gson.toJson(medicalRecord);
// Save the medical record on the blockchain
 stub.putStringState(id, medicalRecordJSON);
return id;
}
```

Fetching Patient Data

```
public String getMedicalRecord(Context ctx, String id) { ChaincodeStub stub = ctx.getStub();
// Retrieve the medical record from the blockchain
String medicalRecordJSON = stub.getStringState(id);
if (medicalRecordJSON == null || medicalRecordJSON.isEmpty())
{
throw new ChaincodeException("Medical record not found");
}
// Convert the medical record JSON to an object
MedicalRecord medicalRecord = gson.fromJson(medicalRecordJSON, MedicalRecord.class);
return medicalRecordJSON;
}
```

Update ownership of a patient's medical record

```
public void updateOwnership(Context ctx, String id, String newOwner) {
ChaincodeStub stub = ctx.getStub();
// Retrieve the medical record from the blockchain
String medicalRecordJSON = stub;
        if (medicalRecordJSON == null || medicalRecordJSON.isEmpty())
{
```

```
        throw new ChaincodeException("Medical record not found");
}

// Convert the medical record JSON to an object
MedicalRecord medicalRecord = gson.fromJson(medicalRecordJSON, MedicalRecord.class);

// Update the medical record with the new owner
medicalRecord.owner = newOwner;

// Convert the medical record object to JSON
medicalRecordJSON = gson.toJson(medicalRecord);

// Save the updated medical record on the blockchain
stub.putStringState(id, medicalRecordJSON);

}
```

6. PROPOSED ARCHITECTURE

The proposed privacy preserving EHR architecture using Hyperledger Fabric is shown in Figure 4. Our Framework has implemented privacy-preserving mechanisms of the Hyperledger Fabric to preserve data and user privacy on the network during transaction consensus. In our system, the patient has complete ownership of their healthcare data. The study offers effective access control between patients and doctors through encryption techniques. Our approach is a tamper-resistant mechanism as we have stored only metadata (hashed data and data reference URLs) for every healthcare transaction in the blockchain. Other pieces of information will be encrypted using AES before storing it to the database. Keys are encrypted using ECC (Elliptical Curve Cryptography) to fetch information. Idemix with ZKP is stacked on top of the Hyperledger Fabric to achieve unlinkability and anonymity.

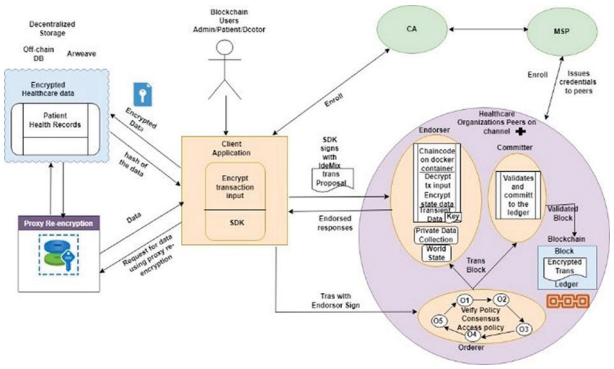


Fig 4. Hyperledger Fabric integrated Blockchain architecture for EHR sharing system

Explanation

This section describes the entire process of the client submitting the transaction to build a block on the blockchain, confirming consensus, storing EHR files to Arweave storage, and reading files from Arweave, as shown in Fig. 1. After the patient's approval, a doctor can add disease, prescription and other information. This information will be stored in the private data collection of that organization. Permission lists are stored on the peer's CouchDB and heavy files such as reports will be stored on Arweave data storage. Only hashes of the transaction will be stored on the blockchain.

The below example describes the architecture:

1. First, the client submits the transaction with encrypted input data. Suppose a patient wants to give data access consent to another doctor. This request with input data of a transaction such as patient name, doctor name, and organization name are passed through the transient field for privacy preservation purposes. Each participant will employ identity certificates associated with Idemix to carry out any action on the distributed ledger.
2. This transaction will be passed to the identified endorsers and endorsed according to the endorsement policy. The endorsement policy specifies that this transaction must be endorsed through the doctors' organization and the patient data provider organization. All Endorsers endorse transactions by executing chain code and providing consent to the doctor for a specified time request. The fetched patient data will be stored on private data collection on a particular healthcare organization's peer network. Later, the same data will be fetched using proxy re-encryption from the Arweave data storage. Endorsers apply the corresponding changes to a state database that maintains a snapshot of the current world state.
3. After endorsement, this signed transaction will be submitted to the client's application. If transient data are passed in a transaction, then only chaincode can access them, decrypt them and store them in the private data collection of the organization/s. Before sending transactions to the ordering service and adding blocks to the ledger, values within the chaincode can be encrypted using AES encryption to protect the data. So, only a person who has access to the appropriate key can decrypt encrypted data.
4. Then the client submits this transaction to the ordering service with read/write sets and signatures from the endorsers.
5. Ordering Service orders the transaction according to the channel and puts them into blocks and sends the block to peers.
6. Peers (Endorsers & Committers) validate each transaction in the block by checking whether the transaction has the necessary endorsements or not. If the transaction satisfies an endorsement policy, then the transaction will be marked as "valid" in a block and respective states will be updated in a state database by reading the write sets. If the transaction fails to satisfy the policy, then it will be marked as "invalid" in a block and there will be no state changes in the state database.

7. EVALUATION

In the last few years, only a few Hyperledger Fabric blockchain-based permissioned EHR solutions have been proposed, with each claiming that it revolutionizes the way of transaction processing along with ensuring security and privacy. However, none of them has utilized IdeMix in Fabric Java SDK. To compare our study with existing research, we analyzed four parameters: 1) Data Security and Scalability, 2) Data Privacy and GDPR, 3) Transaction Data Privacy during consensus, and 4) User Anonymity.

Data Security and Scalability

The problem with submitting confidential information to the ledger is that it will be available to everyone. Few existing studies have proposed a solution to overcome this problem. They have utilized off-chain technologies to store encrypted heavy files on a distributed network such as IPFS and put only the hash of the private data on the ledge. However, IPFS is not a truly decentralized and permanent solution. In our proposed solution, we have utilized Arweave with the Fabric-based HER system to store data permanently to make the EHR system scalable and immutable.

Data Privacy and GDPR

The existing Fabric-implemented EHR studies have neither implemented any privacy control mechanisms nor do they comply with GDPR. Few existing systems partially mitigate this limitation by storing the medical records on IPFS, which supports data deletion. However, metadata is kept on-chain. So even if the patient requests its removal, it cannot be executed. We have implemented data privacy demands using Private Data Collections and Transient Fields. By implementing this feature, organizations transact with one another on the same channel without worrying that another organization will discover their confidential information. It decreases the number of channels while enabling data privacy inside the channel. Also, we have stored data on Arweave storage where the user can delete data if no longer needed. To delete a JSON file that has been uploaded to the Arweave network, we can create and post a new transaction that references the original transaction ID and sets its data to an empty buffer.

Transaction/Data Privacy during consensus:

The notion of privacy always complicates the notion of replication and decentralization. Even though the Fabric network is permissioned, and the healthcare organization's sensitive information is stored on off-chain storage, sensitive information about a transaction can be leaked. In a Fabric-based HER system, data privacy is concerned with the privacy of input parameters of a transaction, state data privacy (output of a transaction), and smart contract privacy. While submitting a transaction, the client passes input parameters and other information. Later, the endorser executes the chain code based on the input parameters and prepares state data for the transaction. If this transaction-related information is visible to the orderer or other nodes of the channel, it may result in the leakage of private data. We have implemented chaincode to share the confidential data on-chain via a transient field and private data is passed using private data collection of the other organization.

User privacy during transaction consensus:

User privacy is concerned with user anonymity and transaction unlinkability. In Fabric, all the channel participants are aware of each other's identities, which can compromise user privacy. 1) User anonymity means hiding user-level credentials to achieve user privacy in the ledger. In Fabric, transactions are endorsed by endorsers before going through validation. During an endorsement, an endorser specifies its identity with its signature, to be later validated by other nodes. However, revealing the identity of an endorser can compromise patient privacy in transactions. Such as when patient grants read permission rights to the doctor of another organization. Revealing the identity of endorsers will leak the organization membership of their transactor, which in turn, leaks patient disease-related information through the identity of the organization. 2) Transaction unlinkability is the ability of a transactor to throw multiple transactions without being linkable. In Fabric, endorsers would know which clients had submitted transactions, and the orderers would know which endorsers had endorsed transactions. All this information could reveal sensitive and confidential information about a patient. Since the EHR system deals with patients' confidential information, this sensitive data must be kept secret for some users and hidden from other network users. For resolving these concerns, we have implemented Idemix cryptographic protocol in Hypeledger Fabric Java SDK v2.4 and achieved unlinkability and anonymity.

Below in Table 2, the proposed framework is compared to existing Fabric-integrated permissioned EHR solutions in terms of Java SDK, Raft Ordering, Transient Field for passing input parameters of a transaction, PDC for private data sharing between organizations, IdeMIX protocol for user privacy and Arweave for permeance of data. The comparison shows our study has implemented all the features to make a robust and secure EHR sharing system.

Table 2. Comparison with existing Fabric-integrated EHR systems.

Ref.	[7]	[8]	[9]	[10]	[11]	[12]	[13]	Proposed Model
Fabric Java SDK	N	N	N	N	N	N	N	Y
Raft ordering	N	N	N	N	N	N	N	Y
Transient Filed	N	N	N	N	N	N	N	Y
Private Data Collections	N	N	N	N	N	Y	N	Y
IdeMix	N	N	N	N	N	Y	N	Y
Arweave	N	N	N	N	N	N	N	Y

8. CONCLUSION

In this research study, we reviewed several existing studies that aim to address privacy of EHR data. However, we found that most of these studies failed to address user anonymity and transaction unlinkability concerns. Our proposed solution suggests a permissioned EHR that makes use of Hyperledger Fabric and its privacy-preserving features to address the critical issues regarding EHR user privacy and transaction unlinkability in the network. Our solution uses Identity Mixer using Fabric's Java SDK V2.4 to guarantee transaction anonymity and unlinkability in the network. Moreover, data confidentiality between a subset of organizations on the network using private data collections and transient fields has been implemented. Our proposed solution addresses these challenges and provides a robust method for preserving patients' data, paving the way for an improved patient experience and a digital revolution in the healthcare sector. In our future work, we will try to implement quasi-anonym endorsement to accomplish privacy and anonymity of transactions in Fabric.

9. REFERENCES

[1] C. C. Agbo, Q. H. Mahmoud, J. M. Eklund, "Blockchain technology in healthcare: a systematic review", Healthcare, Vol. 7, No. 2, 2019, p. 56.

[2] J. Kaur, R. Rani, N. Kalra, "Blockchain-based framework for secured storage, sharing, and querying of electronic healthcare records", Concurrency and Computation: Practice and Experience, Vol. 33, No. 20, 2021, p. e6369.

[3] I. Yaqoob, K. Salah, R. Jayaraman, Y. Al-Hammadi, "Blockchain for healthcare data management: opportunities, challenges, and future recommendations", Neural Computing and Applications, Vol. 34, No. 14, 2022, pp. 11475-11490.

[4] T. T. Kuo, H. E. Kim, L. Ohno-Machado, "Blockchain distributed ledger technologies for biomedical and health care applications", Journal of the American Medical Informatics Association, Vol. 24, No. 6, 2017, pp. 1211-1220.

[5] Z. Leng, Z. Tan, K. Wang, "Application of hyperledger in the hospital information systems: A survey", IEEE Access, Vol. 9, 2021, pp. 128965-128987.

[6] E. Androulaki et al. "Hyperledger fabric: a distributed operating system for permissioned blockchains", Proceedings of the Thirteenth EuroSys Conference, Porto, Portugal, 23-26 April 2018, pp. 1-15.

[7] J. Huang, Y. W. Qi, M. R. Asghar, A. Meads, Y. C. Tu, "MedBloc: a blockchain-based secure EHR system for sharing and accessing medical data", Proceedings of the 18th IEEE International Conference on Trust, Security and Privacy In Computing And Communications / 13th IEEE International Conference on Big Data Science And Engineering, Rotorua, New Zealand, 5-8 August 2019, pp. 594-601.

[8] V. Mahore, P. Aggarwal, N. Andola, S. Venkatesan, "Secure and Privacy Focused Electronic Health Record Management System using Permissioned Blockchain", Proceedings of the IEEE Conference on Information and Communication Technology, Allahabad, India, 6-8 December 2019, pp. 1-6.

[9] A. Dubovitskaya et al. "ACTION-EHR: patient-centric blockchain-based electronic health record data management for cancer care", Journal of Medical Internet Research, Vol. 22, No. 8, 2020, p. e13598.

[10] D. Tith, J. S. Lee, H. Suzuki, W. M. A. B. Wijesundara, N. Taira, T. Obi, N. Ohyama, "Application of blockchain to maintaining patient records in electronic health record for enhanced privacy, scalability, and availability", Healthcare Informatics Research, Vol. 26, No. 1, 2020, pp. 3-12.

[11] S. Chenthara, K. Ahmed, H. Wang, F. Whittaker, Z. Chen, "Healthchain: A novel framework on privacy preservation of electronic health records using blockchain technology", Plos One, Vol. 15, No. 12, 2020, p. e0243043.

[12] C. Stamatellis, P. Papadopoulos, N. Pitropakis, S. Katsikas, W. J. Buchanan, "A privacy-preserving healthcare framework using hyperledger fabric", Sensors, Vol. 20, No. 22, 2020, p. 6587.

[13] V. Mani, P. Manickam, Y. Alotaibi, S. Alghamdi, O. I. Khalaf, "Hyperledger healthchain: patient-centric IPFS-based storage of health records", Electronics, Vol. 10, No. 23, 2021, p. 3003.

[14] Private data, https://hyperledger-fabric.readthedocs.io/en/release-2.5/private-data/private-data.html (accessed: 2023)

[15] Transient Field, https://hyperledger-fabric.readthedocs.io/en/release-2.5/private-data-arch.html#how-to-pass-private-data-in-a-chaincode-proposal (accessed: 2023)

[16] MSP Implementation with Identity Mixer, https://hyperledger-fabric.readthedocs.io/en/release-2.5/idemix.htmll (accessed: 2023)

[17] A Blockchain Platform for the Enterprise, https://hyperledger-fabric.readthedocs.io/en/release-2.2/ (accessed: 2022)

Assessing the Performance of a Speech Recognition System Embedded in Low-Cost Devices

Original Scientific Paper

Fatima Barkani
University Sidi Mohamed Ben Abdellah,
Faculty of Sciences Dhar Mahraz,
Laboratory of Computer Science, Signals,
Automation and Cognitivism
Fez, Morocco
fatimabarkani4@gmail.com

Mohamed Hamidi
Mohammed First University,
Pluridisciplinary Faculty of Nador,
Team of modeling and scientific computing
Oujda, Morocco
mohamed.hamidi.5@gmail

Ouissam Zealouk
University Sidi Mohamed Ben Abdellah,
Faculty of Sciences Dhar Mahraz,
Laboratory of Computer Science, Signals,
Automation and Cognitivism
Fez, Morocco
ouissam.zealouk@gmail.com

Hassan Satori
University Sidi Mohamed Ben Abdellah,
Faculty of Sciences Dhar Mahraz,
Laboratory of Computer Science, Signals,
Automation and Cognitivism
Fez, Morocco
hassan.satori@usmba.ac.ma

Abstract *– The main purpose of this research is to investigate how an Amazigh speech recognition system can be integrated into a low-cost minicomputer, specifically the Raspberry Pi, in order to improve the system's automatic speech recognition capabilities. The study focuses on optimizing system parameters to achieve a balance between performance and limited system resources. To achieve this, the system employs a combination of Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and Mel Frequency Spectral Coefficients (MFCCs) with a speaker-independent approach. The system has been developed to recognize 20 Amazigh words, comprising of 10 commands and the first ten Amazigh digits. The results indicate that the recognition rate achieved on the Raspberry Pi system is 89.16% using 3 HMMs, 16 GMMs, and 39 MFCC coefficients. These findings demonstrate that it is feasible to create effective embedded Amazigh speech recognition systems using a low-cost minicomputer such as the Raspberry Pi. Furthermore, Amazigh linguistic analysis has been implemented to ensure the accuracy of the designed embedded speech system.*

Keywords*: Speech recognition, HMMs, GMMs, Raspberry Pi, Amazigh language*

1. INTRODUCTION

The technology of automatic speech recognition (ASR) enables the transcription of spoken messages and the extraction of their linguistic content, making it applicable to numerous ASR research applications, including various fields such as education, interactive services, messaging, machine and robot control, quality control, data entry, remote access, and more. The Raspberry Pi low-cost minicomputer has also been utilized in studies for speech detection, speaker recognition, user speech interface, and robots [1-2].

Researchers have proposed a speaker recognition method that utilizes Hidden Markov Models (HMMs) and the Google API through a Raspberry Pi board. This method is capable of recognizing and authenticating users, serving as a secure speech recognizer for automated door control, and functioning as a general voice recognizer for controlling various appliances [3]. In another study, a home automation system was developed for motion detection and image capture, utilizing a Raspberry Pi board, a connected camera, and motion sensors [4].

An algorithm for efficient home automation through email on Raspberry Pi was proposed in [5], while a method for remote controlling domestic equipment via an Android application using a Raspberry Pi card was proposed in [6]. Additionally, a speech-controlled system was developed for visually impaired individuals to control computer functions using their voice [7]. In another study, Raspberry Pi was utilized to construct an intelligent control and monitoring system for water

treatment plants [8]. Meanwhile, our lab researchers concentrated on integrating the Amazigh language into ASR systems for diverse applications [9, 10-17]. This research presents an architecture for a low-cost, speaker-independent speech recognition system based on the Amazigh language, implemented on a Raspberry Pi board. The system utilizes Amazigh voice commands to control devices connected to the Raspberry Pi. We have designed an open-source evaluation platform that combines a hybrid model of Hidden Markov Models and Gaussian Mixture Models with the Mel-Frequency Cepstral Coefficients feature extraction technique to determine the optimal values for achieving maximum performance. Our work significantly contributes to the advancement of speech recognition technologies in resource-constrained environments. This, in turn, creates opportunities for enhanced speech recognition capabilities in affordable devices. This paper is organized as follows: In Section 1, an introduction to the topic is presented. Section 2 provides an overview of the architecture and functioning of the automatic speech recognition system. Section 3 discusses the proposed work in detail. In Section 4, a discussion of the findings and their implications is presented. Finally, in Section 5, the conclusion of the study is provided.

2. ASR SYSTEM STRUCTURE

2.1. GENERAL ARCHITECTURE

Automatic Speech Recognition (ASR) is a technology that facilitates the transcription of spoken words into text through pattern recognition techniques, involving feature extraction, pattern matching, and reference model library phases [18], as illustrated in Fig. 1. Our research objective was to develop an ASR system tailored for the Amazigh language.

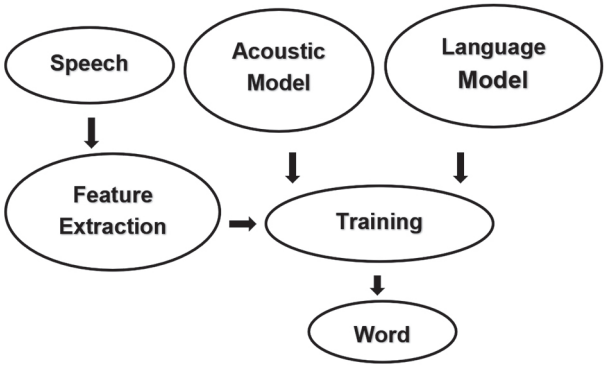


Fig. 1. ASR system architecture

2.2. FEATURE EXTRACTION

The first component of an ASR system is feature extraction, which involves capturing and analyzing speech signals. The system should extract relevant features, such as voice pitch, duration, and energy, from the voice signal. These characteristics are then transformed into a set of numerical values that can be analyzed in more detail. In this work, the feature extraction method used is the MFCC technique [19] with a 16 kHz sampling rate, 16 Kbit sample size, and a 25.6 ms Hamming window.

2.3. HIDDEN MARKOV MODEL

Hidden Markov Models (HMMs) are extensively utilized as a statistical technique for modeling and examining discrete-time series data. Their versatility has made them highly valuable across multiple domains, with a significant presence in the realm of speech processing. Specifically, HMMs have proven to be instrumental in tasks like automatic speech recognition, demonstrating their effectiveness and applicability in capturing the underlying patterns and dynamics of speech signals. The robustness and success of HMMs in these applications have solidified their status as a powerful tool in speech processing research and development and other domains. Fig. 2 presents three states of the Hidden Markov Model topology [19].

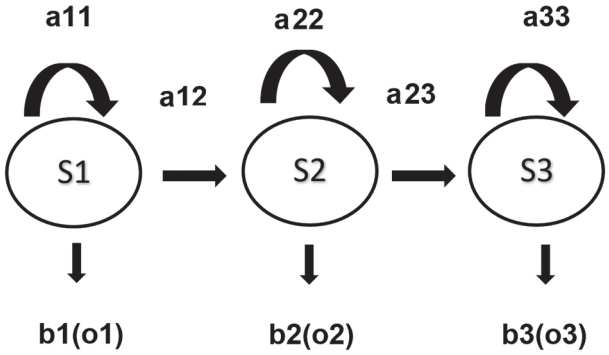


Fig. 2. The 3 states of HMM architecture

2.4. ACOUSTIC MODEL

The acoustic model is an essential element of an ASR system that represents the many sounds and words in a language. It is a statistical model that associates phonemes, syllables, and words with acoustic properties taken from speech signals. The model learns to associate linguistic units with phonemic properties during the training phase. The model is used to match the properties retrieved from the input audio signal to the corresponding linguistic units during the recognition phase. The power and accuracy of the acoustic model greatly affect the accuracy of the ASR system. Hence, one of the major areas of study in acoustics is the construction and improvement of acoustic models. This study employs 3-state and 5-state HMMs to recognize speech data. Additionally, several Gaussian mixture models with 8, 16, and 32 GMMs are used.

2.5. AMAZIGH SPEECH DICTIONARY

The dictionary is a mediator between the Acoustic Model and the Language Model. Our pronunciation dictionary includes all the Amazigh training words along with their corresponding pronunciations. The dictionary used to train our system is presented in Table 1.

Table 1. The used Amazigh Dictionary

AFLLA	AE F L AH
AFOSI	AE F AH S IY
ALNDAD	AE L N D AH D
AMAGGWAJ	AH M AE G W AH JH
ANAKMAR	AH N AE K M AH R
AWAR	AH W AO R
AZLMAD	AE Z L M AH D
DAR	D AA R
DAT	D AE T
DDAW	D AO
ELEM	E L E M
KRAD	K R A D
KOZ	K O Z
SA	S A
SEDISS	S D E SS
SMMUS	S M U S
SIN	S I N
TAM	T A M
TZA	T Z A
YEN	Y E N

Native memory refers to the memory available to a process, such as a Java process, and it is based on physical memory, disks, and other physical devices that are managed by the operating system (OS). The CPU utilizes the memory bus to access the normal memory and also predicts the instructions to execute, storing the results in registers - fast memory components that can hold CPU results. The memories that are accessed depend on the physical address size, which the CPU uses to identify physical memory. For instance, a 16-bit address can access 2^{16} (=65,536) memory locations, while a 32-bit address can access 2^{32} (=4,294,967,296) memory locations. To map the physical memory to the memory that each process can see, an operating system (OS) uses virtual memory [20].

3. THE PROPOSED SYSTEM ARCHITECTURE

The aim of this study is to build an advanced system that can recognize Amazigh speech through voice commands, utilizing a Raspberry Pi. The proposed system architecture is illustrated in Fig. 3.

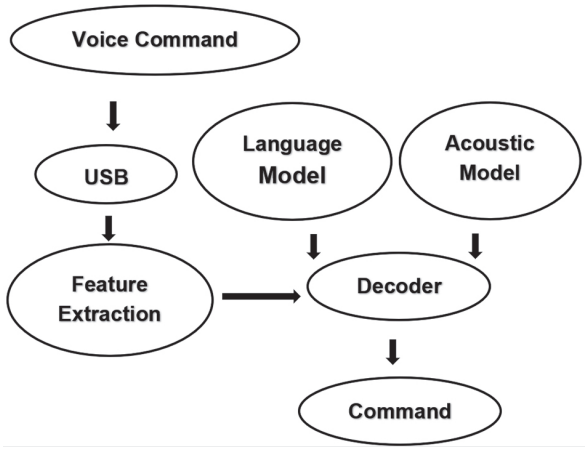


Fig. 3. Proposed System

3.1. THE RASPBERRY PI CARD

Raspberry Pi 3 is a powerful, compact computer that can perform the same tasks as a regular personal computer. In this work, we use a micro SD card compatible with Raspberry Pi, pre-loaded with 1GB of RAM. The design of this system does not include an integrated hard disk or solid-state drive for storage purposes. Instead, it relies on a 16GB SD card to serve as a boot device and internal storage. Since the Raspberry Pi does not have a built-in microphone, we use an external USB microphone in this work.

3.2. CORPUS

The speech data used in this study are unique to the laboratory and have not been previously published or shared. This dataset contains 20 Amazigh words collected from 30 native Tarifit speakers in Morocco. The dataset includes the first ten Amazigh digits (Table 2) and ten other words (Table 3). The speech was recorded using a microphone and the WaveSurfer program and saved in a ".wav" file format. Thirty percent of the data was used for testing, while the remaining 70% was used for training (refer to Table 4 for details).

Table 2. The used Amazigh digits

Amazigh digits	English	Tifinagh	Syllables
ELEM	Zero	ⵖⵎⵏⵍ	VCCV
YEN	One	ⵢⴰⵏ	CVC
SIN	Two	ⵙⵉⵏ	CVC
KRAD	Three	ⴽⵕⴰⴷ	VCCVC
KOZ	Four	ⴽⵓⵣ	CVC
SMMUS	Five	ⵙⵎⵎⵓⵙ	CCCVC
SEDISS	Six	ⵙⴹⵉⵙⵙ	CCVC
SA	Seven	ⵙⴰ	CV
TAM	Eight	ⵜⴰⵎ	CVC
TZA	Nine	ⵜⵣⴰ	CCCV

Table 3. The used Amazigh words

Amazigh Words	English	Tifinagh	Syllables
AFLLA	above	ⴰⵀⵍⵍⴰ	VCCCV
AFOSI	right	ⵀⵓⵙⵉ	VCVCV
ALNDAD	in front of	ⴰⵍⵏⴷⴰⴷ	VCCCVC
AMAGGWAJ	far	ⴰⵎⴰⴳⴳⵡⴰⵊ	VCVCCVC
ANAKMAR	near	ⴰⵏⴰⴽⵎⴰⵔ	VCVCCVC
AWAR	after	ⴰⵡⴰⵔ	VCVC
AZLMAD	left	ⴰⵣⵍⵎⴰⴷ	VCCCVC
DAR	beside	ⴷⴰⵔ	CVC
DAT	before	ⴷⴰⵜ	CVC
DDAW	down	ⴷⴷⴰⵡ	CCVC

Table 4. Corpus characteristics

Recorder type	Number of recorders used for training	Number of recorders used for testing
Amazigh digits	21	9
Amazigh words	7	3

3.3. RESULTS

The designed system was trained and generated using a laptop equipped with an Intel Core i3 CPU running at 2.4 GHz, 4 GB of RAM, and the Ubuntu 14.04 LTS operating system. After training the system, we conducted two test experiments: the first one using the same laptop. Additionally, in the second experiment, we tested the acoustic system on a Raspberry Pi board with 1 GB of RAM and the Raspbian operating system to assess its functionality on a lower-specification device. The data presented in Figs. 4 and 5 illustrate the accuracy of digit recognition achieved by using various GMMs and HMMs on a laptop. The testing corpus subset consisted of 900 tokens for all ten digits. Our results showed that when using 3 HMMs, the system achieved recognition rates of 90.33%, 91.33%, and 88.67% for 8, 16, and 32 GMMs, respectively. Meanwhile, when using 5 HMMs, the system achieved correct rates of 84.11%, 84.22%, and 81.88% for 8, 16, and 32 GMMs, respectively.

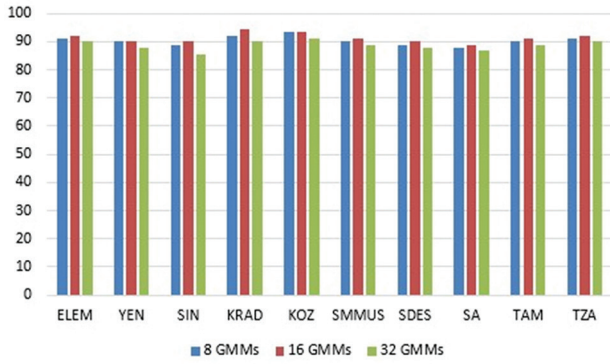


Fig. 4. Digits recognition rates by using a laptop 3 HMMs

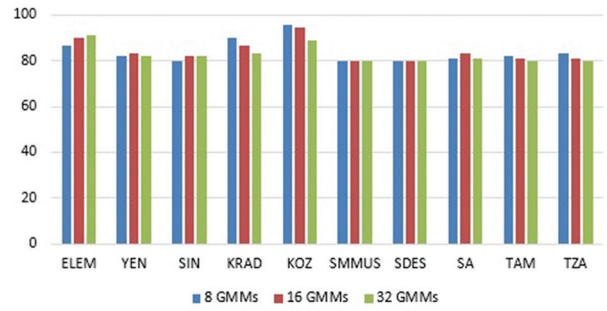


Fig. 5. Digits recognition rates by using a laptop 5 HMMs

The system aims to recognize 300 tokens of 10 Amazigh commands. When using 3 HMMs, the recognition rates achieved were 86.66%, 86.99%, and 85.33% for 8, 16, and 32 GMMs. Figures 6 and 7 present the recognition rates for Amazigh words using various GMMs and HMMs on a laptop. On the other hand, when using 5 HMMs, the system achieved recognition rates of 82.33%, 83.66%, and 83.33% for 8, 16, and 32 GMMs, respectively. The highest recognition rate was achieved using 16 GMMs with 3 HMMs. Based on these results, it can be concluded that the Amazigh digits KRAD and KOZ are the most frequently recognized using 3 and 5 HMMs, respectively. Additionally, the best-recognized Amazigh words using 3 HMMs are AFLLA, AFOSI, AZLMAD, and DDAW, while the best-recognized word using 5 HMMs is DDAW.

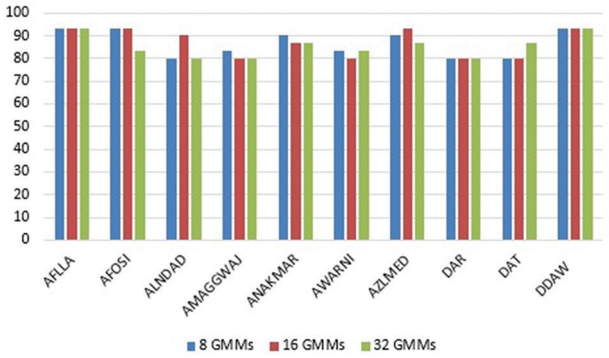


Fig. 6. Commands recognition rates by using a laptop 3 HMMs

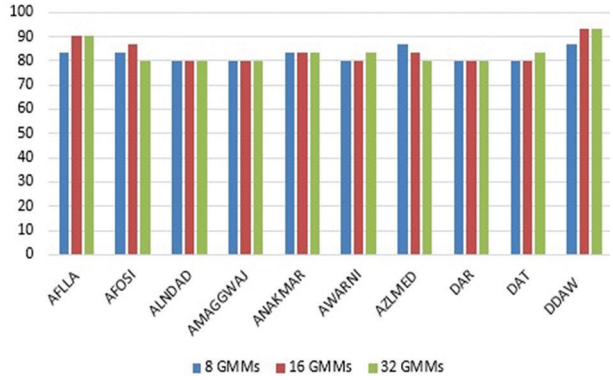


Fig. 7. Commands recognition rates by using a laptop 5 HMMs

Figs. 8 and 9 show the results of digit recognition rates obtained through various GMMs and HMMs on a Raspberry Pi board. The recognition rates obtained for 3 HMMs using 8, 16, and 32 GMMs are 89.00%, 89.67%, and 86.67%, respectively. For 5 HMMs, the recognition rates achieved using 8, 16, and 32 GMMs are 83.66%, 84.11%, and 82.77%, respectively. The highest recognition rate was obtained using 16 GMMs and 3 HMMs.

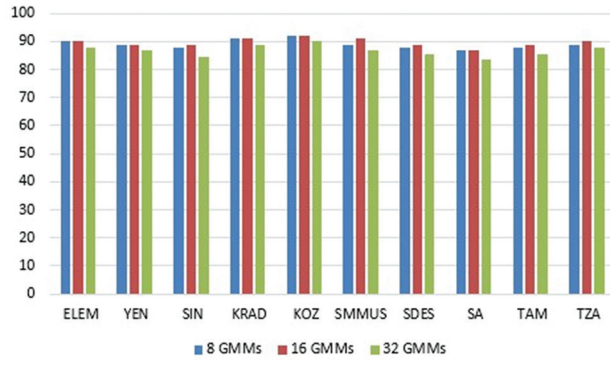


Fig. 8. Raspberry Pi board recognition rates for Amazigh Digits by using 3 HMMs

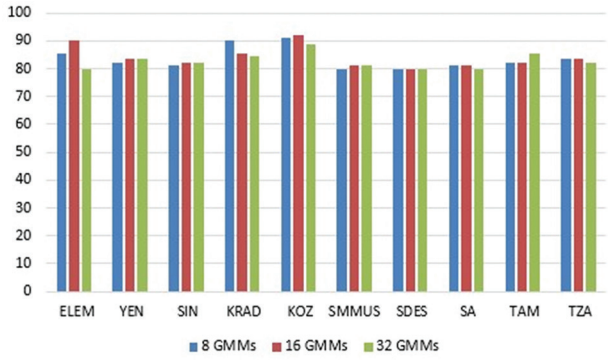


Fig. 9. Raspberry Pi board recognition rates for Amazigh Digits by using 5 HMMs

Figs. 10 and 11 present the recognition rates for Amazigh words achieved through different GMMs using a Raspberry Pi board. For 3 HMMs, the recognition rates obtained were 85.66%, 86.33%, and 83.99% when using 8, 16, and 32 GMMs, respectively. On the other hand, for 5 HMMs, the recognition rates were 82.99%, 82.66%, and 83.33% with 8, 16, and 32 GMMs, respectively. The most optimal recognition rate was achieved by using 16 GMMs with 3 HMMs. The most commonly recognized Amazigh numeral for 3 and 5 HMM is KOZ. For 3 HMM, AFLLA and DDAW are the most frequently recognized Amazigh words, while for 5 HMM, AFFLA and AFOSI are the best recognized Amazigh words. The results indicate that the recognition rates achieved using a Raspberry Pi were lower than those obtained using a laptop.

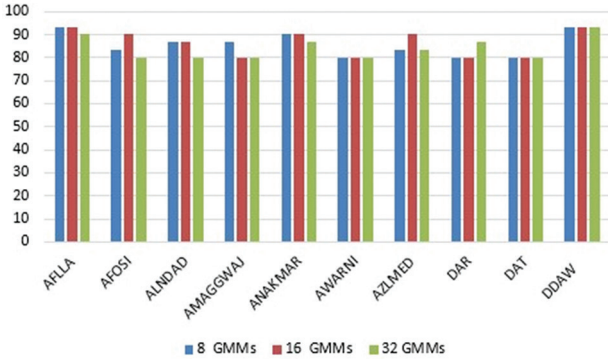


Fig. 10. Raspberry Pi board recognition rates for Amazigh-commands by using 3 HMMs

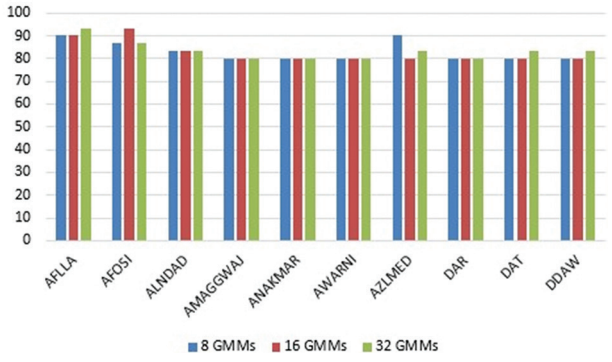


Fig. 11. Raspberry Pi board recognition rates for Amazigh-commands by using 5 HMMs

Figs. 12 and 13 show the memory consumption tests conducted on both a laptop and a Raspberry Pi board for Amazigh commands. It was found that the Raspberry Pi board consumed more memory than the laptop during the tests.

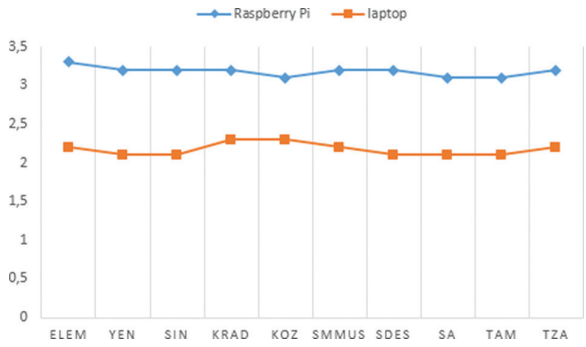


Fig. 12. System memory consumption for digits

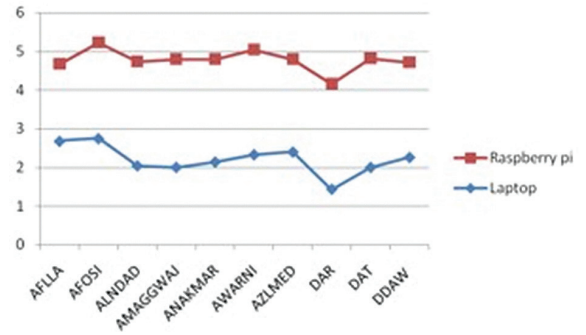


Fig. 13. System memory consumption for Amazigh words

Figs. 14 and 15 illustrate the recognition time for Amazigh digits and words, respectively, using both a laptop and a Raspberry Pi board.

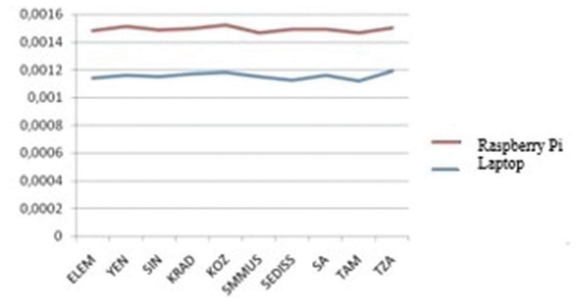


Fig. 14. System time for Amazigh digits

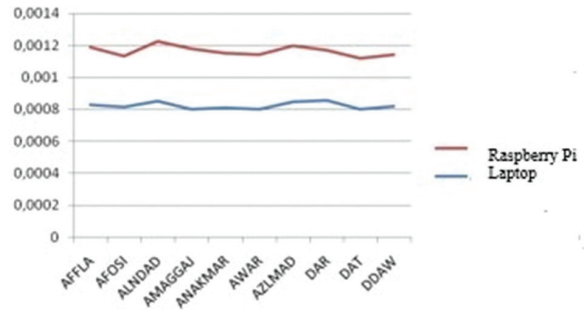


Fig. 15. System time for Amazigh words

The results indicate that the recognition time for the Raspberry Pi board was greater than that for the laptop. These results were expected, given the limitations of the Raspberry Pi hardware. Raspberry Pi has less processing power than a laptop, which can lead to slower and less accurate performance with the ASR system.

We have conducted a comparison of our proposed approach with other existing works. As presented in Table 5, our approach differs significantly from the methods employed by other researchers in several aspects, notably, our accuracy score is 89.16%.

Reference	Year	Methods	Results
[21]	2019	WAT | MFCC | SVM	100%
[22]	2020	CNNs | SVM	95.30% | 72.39%
[23]	2020	PWP | MFCC	99%
[24]	2020	SVM | DTW	97%
[25]	2021	DSP | HMM	80%
Proposed work	2023	GMM | HMM | MFCC	89.16%

4. DISCUSSION

Our tests and analysis show Our tests and analysis show the following:

- The 16 Gaussian mixture distributions yielded the best results.
- The laptop outperforms the Raspberry Pi.
- The laptop has a shorter recognition time compared to the Raspberry Pi.
- The Raspberry Pi board consumes more memory than a laptop.

Based on the results presented in Figs. 12 and 13, it is evident that the number of syllables in Amazigh words significantly impacts memory consumption. The data suggests that when the number of syllables is less than two, memory consumption is lower as well. Most Amazigh commands for computers and Raspberry Pi consist of only one syllable. Analyzing the digit consumption revealed that the KOZ digit consumed the least memory on the Raspberry Pi board, while the YEN and SIN digits consumed the least memory on a laptop. When considering word memory consumption, the DAR word consumed the least memory on both the laptop and the Raspberry Pi board. Therefore, it appears that all Amazigh commands consisting of a single syllable consume less memory. The results displayed in Figures 14 and 15 indicate that the Amazigh language commands "TAM" and "DAT" are recognized more quickly by the system, regardless of whether a computer or a Raspberry Pi is used. These two Amazigh commands have a monosyllabic structure with a consonant-vowel-consonant (CVC) pattern.

5. CONCLUSION

The aim of this research is to investigate the feasibility of utilizing Amazigh speech recognition for controlling external devices. The study involved the creation of an embedded system that utilizes isolated Amazigh words on a Raspberry Pi board. The goal was to develop a portable system that could function effectively within a limited resource environment. To achieve this, the study employed HMMs, GMMs, MFCCs, and parameter optimization to design the speech recognition system. The system was designed to recognize twenty Amazigh words, including ten words and the first ten Amazigh digits, using the open-source CMU Sphinx 4. The findings demonstrated that the optimal performance was achieved with 3 HMMs, 16 GMMs, and 39 MFCC coefficients, resulting in an accuracy rate of 89.16%.

In our future work, our efforts will be directed toward improving the performance of our system through the utilization of hybrid and deep learning techniques. We aim to explore the benefits of combining these approaches to achieve even greater results.

6. REFERENCES:

[1] F. Raffaeli, S. Awad, "Portable low-cost platform for embedded speech analysis and synthesis," Proceedings of the 12th International Computer Engineering Conference (ICENCO), IEEE, 2016, pp. 117-122.

[2] A. Mnassri, M. Bennasr, C. Adnane, "A Robust Feature Extraction Method for Real-time Speech Recognition System on a Raspberry Pi 3 Board," Engineering, Technology & Applied Science Research, Vol. 9, No. 2, 2019, pp. 4066-4070.

[3] S. Suresh, Y. S. Rao, ''Modelling of secured voice recognition based automatic control system". International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN, 2015, pp. 0976-1353.

[4] V. Patchava, H. B. Kandala, P. R. Babu, "A smart home automation technique with Raspberry Pi using IoT," Proceedings of the 2015 International Conference on Smart Sensors and Systems (IC-SSS), IEEE, 2015, pp. 1-4.

[5] S. Jain, A. Vaibhav, L. Goyal, "Raspberry Pi based interactive home automation system through E-mail," Proceedings of the 2014 International Conference on Reliability Optimization and Information Technology (ICROIT), IEEE, 2014, pp. 277-280.

[6] H. Lamine, H. Abid, "Remote control of a domestic equipment from an Android application based on Raspberry Pi card," Proceedings of the 15th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), IEEE, 2014, pp. 903-908.

[7] M. S. K. Upadhyay, M. V. N. Chavda, "Intelligent system based on speech recognition with capability of self-learning," International Journal for Technological Research in Engineering, Vol. 1, No. 9, 2014.

[8] S. S. Lagu, S. B. Deshmukh, "Raspberry Pi for automation of water treatment plant", Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2014, pp. 1999-2003.

[9] Y. E. Saady, A. Rachidi, M. Yassa, D. Mammass, "Amhcd: A Database for Amazigh Handwritten Character Recognition Research," International Journal of Computer Applications, Vol. 27, No. 4, 2011, pp. 44-48.

[10] H. Satori, F. ElHaoussi, "Investigation Amazigh speech recognition using CMU tools," International Journal of Speech Technology, Vol. 17, 2014, pp. 235-243.

[11] M. Hamidi, H. Satori, O. Zealouk, K. Satori, "Amazigh digits through interactive speech recognition system in noisy environment," International Journal of Speech Technology, Vol. 23, No. 1, 2020, pp. 101-109.

[12] H. Satori, O. Zealouk, K. Satori, F. ElHaoussi, "Voice comparison between smokers and non-smokers using HMM speech recognition system", International Journal of Speech Technology, Vol. 20, No 4, 2017, pp. 771-777.

[13] M. Hamidi, H. Satori, O. Zealouk, K. Satori, ''Speech coding effect on Amazigh alphabet speech recognition performance", J. Adv. Res. Dyn. Control Syst, Vol. 11, No 2, 2019, pp. 1392-1400.

[14] O. Zealouk, H. Satori, M. Hamidi, N. Laaidi, K. Satori, ''Vocal parameters analysis of smoker using Amazigh language", International Journal of Speech Technology, Vol. 21, 2018, p. 85-91.

[15] O. Zealouk, H. Satori, M. Hamidi, K. Satori, "Voice pathology assessment based on automatic speech recognition using Amazigh digits," Proceedings of the 2nd International Conference on Smart Digital Environment, 2018, pp. 100-105.

[16] M. Hamidi, H. Satori, O. Zealouk, K. Satori, N. Laaidi, "Interactive voice response server voice network administration using hidden Markov model speech recognition system," Proceedings of the Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), IEEE, 2018, pp. 16-21.

[17] O. Zealouk, H. Satori, M. Hamidi, K. Satori, ''Speech recognition for Moroccan dialects: feature extraction and classification methods", J. Adv. Res. Dyn. Control Syst, Vol. 11, No 2, 2019, pp. 1401-1408.

[18] X. Zhong, Y. Liang, ''Raspberry Pi: An effective vehicle in teaching the internet of things in computer science and engineering", Electronics, Vol. 5, No 3, 2016, pp. 56.

[19] H. Hermansky, ''Perceptual linear predictive (PLP) analysis of speech", the Journal of the Acoustical Society of America, Vol. 87, No 4, 1990, pp. 1738-1752.

[20] P. Vojtas, J. Stepan, D. Sec, R. Cimler, O. Krejcar, "Voice recognition software on embedded devices". In: Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018, Proceedings, Part I, Springer International Publishing, 2018, pp. 642-650.

[21] M. Walid, S. Bousselmi, K. Dabbabi, A. Cherif, ''Real-time implementation of isolated-word speech recognition system on raspberry Pi 3 using WAT-MFCC", IJCSNS, Vol. 19, No 3, 2019, pp. 42.

[22] M. S. I. Sharifuddin, S. Nordin, A. M. Ali, ''Comparison of CNNs and SVM for voice control wheelchair". IAES International Journal of Artificial Intelligence, Vol. 9, No 3, 2020, pp. 387.

[23] W. Helali, Z. Hajaiej, A. Cherif, ''Real-time speech recognition based on PWP thresholding and MFCC using SVM", Engineering, Technology & Applied Science Research, Vol. 10, No 5, 2020, pp. 6204-6208.

[24] A. Ismail, S. Abdlerazek, I. M. El-Henawy, ''Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping", Sustainability, Vol. 12, No 6, 2020, pp. 2403.

[25] A. Abdulkareem, T. E. Somefun, O. K. Chinedum, A. F. Agbetuyi, ''Design and implementation of speech recognition system integrated with internet of things", International Journal of Electrical and Computer Engineering (IJECE), Vol. 11, No 2, 2021, pp. 1796-1803.

Gradient-adaptive Nonlinear Sharpening for Dental Radiographs

Original Scientific Paper

Manoj T Joy
Sathyabama Institute of Science and Technology, Chennai, India
manojtjoy@gmail.com

B Priestly Shan
Chandigarh University, Punjab, India
priestlyshan@gmail.com

Geevarghese Titus
Amal Jyothi College of Engineering, Kanjirapally, Kerala, India
geevarghesetitus@amaljyothi.ac.in

Abstract *– Unsharp Masking is a popular image processing technique used for improving the sharpness of structures on dental radiographs. However, it produces overshoot artefact and intolerably amplifies noise. On radiographs, the overshoot artefact often resembles the indications of prosthesis misfit, pathosis, and pathological features associated with restorations. A noise-robust alternative to the Unsharp Masking algorithm, termed Gradient-adaptive Nonlinear Sharpening (GNS) which is free from overshoot and discontinuity artefacts, is proposed in this paper. In GNS, the product of the arbitrary scalar termed as 'scale' and the difference between the output of the Adaptive Edge Smoothing Filter (AESF) and the input image, weighted by the normalized gradient magnitude is added to the input image. AESF is a locally-adaptive 2D Gaussian smoothing kernel whose variance is directly proportional to the local value of the gradient magnitude. The dataset employed in this paper is downloaded from the Mendeley data repository having annotated panoramic dental radiographs of 116 patients. On 116 dental radiographs, the values of Saturation Evaluation Index (SEI), Sharpness of Ridges (SOR), Edge Model Based Contrast Metric (EMBCM), and Visual Information Fidelity (VIF) exhibited by the Unsharp Masking are 0.0048 ± 0.0021, 4.4 × 1013 ± 3.8 × 1013, 0.2634 ± 0.2732 and 0.9898 ± 0.0122. The values of these quality metrics corresponding to the GNS are 0.0042 ± 0.0017, 2.2 × 1013 ± 1.8 × 1013, 0.5224 ± 0.1825, and 1.0094 ± 0.0094. GNS exhibited lower values of SEI and SOR and higher values of EMBCM and VIF, compared to the Unsharp Masking. Lower values of SEI and SOR, respectively indicate that GNS is free from overshoot artefact and saturation and the quality of edges in the output images of GNS is less affected by noise. Higher values of EMBCM and VIF, respectively confirm that GNS is free from haloes as it produces thin and sharp edges and the sharpened images are of good information fidelity.*

Keywords*: Dental radiograph; Dental X-ray images; Image sharpening; Overshoot artefact; Unsharp Masking*

1. INTRODUCTION

1.1. BACKGROUND & PROBLEM DOMAIN

The advent of advanced image processing algorithms and machine learning techniques has revolutionized the automated analysis of dental radiographs. Many recent methods make use of image processing algorithms and machine learning techniques for the diagnosis of Periodontal Bone Loss (PBL) [1, 2], Osteoporosis [3, 4], and dental caries [5] from dental radiographs. Apart from the clinical applications, dental radiographs are used for the automated identification of humans in forensic odontology [6]. The majority of these methods used in clinical practice and forensic odontology involve the segmentation of structures like teeth and mandibles from the dental radiograph [7, 8].

Segmentation of structures from the dental radiograph is often difficult as the structures in it may not have sharp boundaries. The lack of sharpness could be due to the factors like inappropriate positioning, over-exposure, movement of patient/equipment during exposure, *etc*. Powerful post-processing algorithms are required to improve the sharpness of edges and to improve the accuracy of the segmentation of structures from dental radiographs.

1.2. REVIEW OF LITERATURE

Post-processing algorithms available in literature meant for improving the quality of dental radiographs can be classified into three categories. These categories are denoising, contrast enhancement, and sharpening algorithms. The denoising algorithms include a 2-D Butterworth low-pass filter (frequency domain) [9], Bayesian Least Squares - Gaussian Scaled Mixture (BLS-GSM) algorithm, and Total Variation (TV) filter [10]. Contrast Limited Adaptive Histogram Equalization (CLAHE) [11, 12] is an algorithm in literature meant for improving the contrast in dental radiographs. Unsharp Masking is the most popular algorithm used for improving the sharpness of dental radiographs. Yan et al. [13] proposed Bi-SCM: bidirectional spiking cortical model an adaptive unsharp mask using a bio-inspired neural network. To improve the x-ray, Skewness Reformed Complex Diffusion (SRCD) is proposed in [14] where the skewness of the images is used as a parameter to configure the smoothing filtering of the unsharp mask.

Many decades before itself, Fujita *et al*. [15] had demonstrated the prospect of Unsharp Masking for enhancing the quality of dentomaxillofacial radiographs. Harada et *al*. [16] used Unsharp Masking for sharpening the bony structure in the maxillofacial region in the three-dimensional Computed Tomography (CT) image. Couture *et al*. [17] used Unsharp Masking to enhance the texture of cortical and trabecular bones on the projection radiographs of mandibles to effectively visualize the high-frequency variations in bone mineral density. Recently, Geraets *et al*. [18] used Unsharp Masking to enhance trabecular patterns on dental radiographs towards detecting fractures. Manoj et al. [19] propose a mechanism to enhance dental radiographs corrupted by quantum noise.

1.3. LACUNA OF REVIEW AND GAPS

Despite the wide popularity and applications of Unsharp Masking, it has serious limitations also. Clark *et al.* [20] have demonstrated that Unsharp Masking enhances the perceptual quality of dental radiographs and it produces overshoot artefact and amplifies noise that will adversely influence the accuracy of diagnosis. The overshoot artefact may be misinterpreted as a prosthesis misfit or pathosis. Brettle and Carmichael [21] also demonstrated that the artefacts caused by image processing, particularly at high contrast boundaries, closely mimic the pathological features associated with restorations and invite the risk of unwanted interventions.

1.4. CONTRIBUTIONS, NOVELTY, AND HIGHLIGHTS

Contributions: To resolve the issues of noise-amplification, overshoot artefact, haloes, and discontinuity artefact in the Unsharp Masking, a novel algorithm, termed Gradient-adaptive Nonlinear Sharpening (GNS), for improving the acuity of edges in the dental radiographs, is introduced in this paper.

Novelty: In the conventional Unsharp Masking algorithm, the product of the arbitrary scalar termed as 'scale' and the difference between the input image and its Gaussian smoothed version, after a thresholding process is added back to the input image itself. The Gaussian smoothing kernel used in the Unsharp Masking is a linear filter. The weights in the kernel depend only on the spatial distance from the center. The Gaussian kernel has the same smoothing response on noise-affected pixels and edges. Hence, it cannot be expected that the difference between the input image and its Gaussian smoothed version may have higher values at the edge pixels compared to the noise-affected pixels. It is difficult to distinguish the noise-affected pixels and the edge pixels from the difference between the input image and its Gaussian smoothed version, via the thresholding operation. In GNS, the Gaussian kernel is replaced by the Adaptive Edge Smoothing Filter (AESF). AESF is a locally-adaptive 2D Gaussian smoothing kernel whose variance is directly proportional to the local value of the gradient magnitude. It smooths the edge pixels more strongly than the noise-affected pixels. Consequently, the difference between the input image and output of AESF will be significantly high at the edge pixels compared to the noise-affected pixels. The concept of the Adaptive Edge Smoothing Filter is new in image processing itself. Readers should not be confused with the edge-preserving filters. The operation of the adaptive edge smoothing filter is just opposite to that of the edge-preserving filters.

The thresholding process in the Unsharp Masking is a mathematical function with jump discontinuity. Because of the thresholding process, Unsharp Masking is prone to discontinuity artefact. Instead of the thresholding process in the Unsharp Masking, in GNS, the difference between the output of the AESF and the input image is weighted by the normalized gradient magnitude. This nonlinear weighting is a continuous function and avoids the issue of discontinuity artefact. As the thresholding process is not used in GNS, it is free from the burden of tuning the 'threshold' parameter. GNS has less number of operational parameters compared to the Unsharp Masking.

Highlights: (i) Because of the characteristics of AESF used in GNS and the nonlinear weighting process involved in its computation, it is free from noise-amplification and overshoot artefact. (ii) As no mathematical operation with jump discontinuity is involved in the GNS, it is free from the discontinuity artefact. (iii) GNS has less number of operational parameters compared to Unsharp Masking.

2. METHODS

2.1 GRADIENT-ADAPTIVE NONLINEAR SHARPENING (GNS)

In GNS, the product of the arbitrary scalar termed as 'scale' and the difference between the output of the Adaptive Edge Smoothing Filter (AESF) and the input image, weighted by the normalized gradient

magnitude is added to the input image. AESF is a locally-adaptive 2D Gaussian smoothing kernel whose standard deviation is directly proportional to the local value of the gradient magnitude. The first step in GNS is to compute the normalized gradient. The gradient magnitude is computed from the gradients along the rows (horizontal direction) and the columns (vertical direction) of the input radiograph.

$$G = \sqrt{(G_r)^2 + (G_c)^2} \tag{1}$$

In (1), is the gradient along rows (horizontal direction), and is the gradient along columns (vertical direction). G is the 2D vector comprising local values of the gradient magnitude. The gradient along rows and gradient along rows are computed by convolving the input radiograph, with respective Sobel convolution masks.

$$G_r = f ** S_r \; and \; G_c = f ** S_c \tag{2}$$

In (2), f is the input radiograph. The notion, '**' indicates 2D convolution. S_r and S_c respectively are Sobel masks along rows and columns.

$$S_r = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} and \; S_c = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{3}$$

The 2D gradient vector is normalised with the maximum value in it. The normalisation yields,

$$g = \frac{G}{G_{max}} \tag{4}$$

In (4), g is the normalized 2D gradient vector. G_{max} is the largest gradient value. In GNS, the Gaussian kernel in the Unsharp Masking is replaced by the Adaptive Edge Smoothing Filter. As pointed out, AESF is a locally-adaptive 2D Gaussian smoothing kernel whose variance is directly proportional to the local value of the gradient magnitude. It smooths the edge pixels more strongly than the noise-affected pixels. Consequently, the difference between the input image and output of AESF will be significantly high at the edge pixels compared to the noise-affected pixels. The output of the AESF at a pixel location (*r,c*),

$$\tilde{f}(r,c) = \sum_{p=-w}^{+w} \sum_{q=-w}^{+w} K(p,q) f(r+p, c+q), \;\; 1 \le r \le R \; and \; 1 \le c \le C \tag{5}$$

In (5), R and C are the number of rows and the number of columns in the input radiograph. K is the adaptive edge-smoothing Gaussian kernel. w is the radius of the adaptive Gaussian kernel. The kernel at a pixel location (*r,c*) is,

$$K_{rc}(p,q) = \frac{1}{2\pi(\sigma_{rc})^2} \, e^{-\left(\frac{p^2+q^2}{2(\sigma_{rc})^2}\right)}, -w \le p \le +w \; and \; -w \le q \le +w \tag{6}$$

In (6), σ_{rc} is the local value of the standard deviation of the adaptive Gaussian kernel. The local value of the standard deviation of the adaptive Gaussian kernel is the product of the corresponding value in the normalised gradient vector and the maximum limit of the standard deviation of the adaptive Gaussian kernel. It is computed as,

$$\sigma_{rc} = \sigma_{max}[g(r,c)] \tag{7}$$

In (7), σ_{max} is the maximum limit of the standard deviation of the adaptive Gaussian kernel. It has to be as high as possible. Its value is kept as 100 in this paper. The radius of the kernel, *w*, also needs to be high. The radius of the adaptive Gaussian kernel is set as 9 × 9, in this paper. The adaptive Gaussian kernel is normalised such that,

$$K_{rc}(p,q) = \frac{K_{rc}(p,q)}{\sum_{p=-w}^{+w} \sum_{-w}^{+w} K_{rc}(p,q)} \tag{8}$$

The difference between the input image and output of the Adaptive Edge Smoothing Filter,

$$d = f - \tilde{f} \tag{9}$$

The thresholding process in the Unsharp Masking is a mathematical function with jump discontinuity. Because of the thresholding process, Unsharp Masking is prone to discontinuity artefact. Instead of the thresholding process in the Unsharp Masking, in GNS, the difference between the output of the AESF and the input image is weighted by the normalised gradient magnitude. This nonlinear weighting is a continuous function and avoids the issue of discontinuity artefact. As the thresholding process is not used in GNS, it is free from the burden of tuning the 'threshold' parameter. As mentioned, each value in the 2D vector corresponding to the difference, d is weighted by the corresponding value in the normalized gradient vector as,

$$e = \Gamma(d,g) \tag{10}$$

The operator, Γ in (9) indicates element-wise multiplication. The sharpened radiograph,

$$\hat{f} = f + \alpha e \tag{11}$$

In (10), α is an arbitrary parameter that determines the strength of sharpening. This parameter is called ad the scale. The Schematic of the steps involved in the computation of the sharpened image in GNS described in (1) to (10) is shown in Fig. 1.

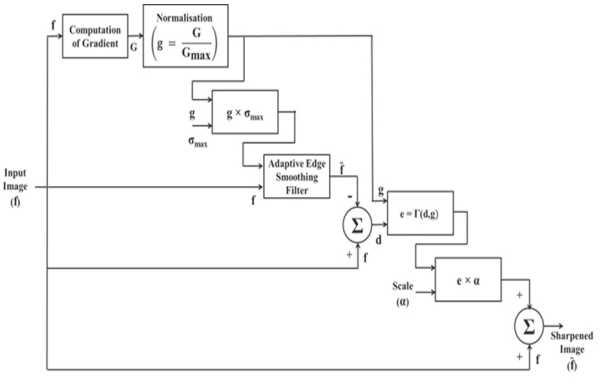


Fig. 1. Schematic of the steps involved in the computation of the sharpened image in GNS

2.2. TEST IMAGES

The dataset used in this paper is downloaded from the Mendeley data repository [22]. This is the data associated with Abdi *et al*. [23] and the dataset comprises of annotated panoramic dental radiographs of 116 patients. All images are in .png format. The images are taken by using the Soredex CranexD digital panoramic x-ray unit available at Noor Medical Imaging Centre, Qom, Iran. The subjects cover a wide range of dental conditions from healthy, to partial and complete edentulous cases.

2.3. PROTOCOL FOR PARAMETER SELECTION AND PERFORMANCE EVALUATION

Sharpening is helpful to improve the acuity of edges. However, most of the sharpening algorithms produce inadvertent artefacts like contrast overshoot, edge-widening or haloes, and noise amplification. There are no unique image quality metrics that can reflect the overall quality of the sharpened images accounting for all these quality aspects and artefacts. Hence, the performance evaluation of the sharpening techniques is a non-trivial task.

Krasula *et al*. [24] have recently proposed a framework governing the subjective evaluation of the quality of the sharpened images and have suggested certain objective measures which can be used for the quantitative performance assessment of the sharpening schemes. As an addition to the framework for evaluating the performance of the sharpening algorithms on natural-scene images reported in [24], J. Joseph and R. Periyasamy [25] have suggested four quality metrics particularly useful for evaluating the performance of sharpening algorithms on medical images. They are the Saturation Evaluation Index (SEI) [26], Sharpness of Ridges (SOR) [27], Edge Model-Based Contrast Measure (EMBCM) [28], and Visual Information Fidelity (VIF) [29]. These four quality statistics are used in this paper to aid the selection of the scale value in GNS and for comparing the performance of GNS with the Unsharp Masking.

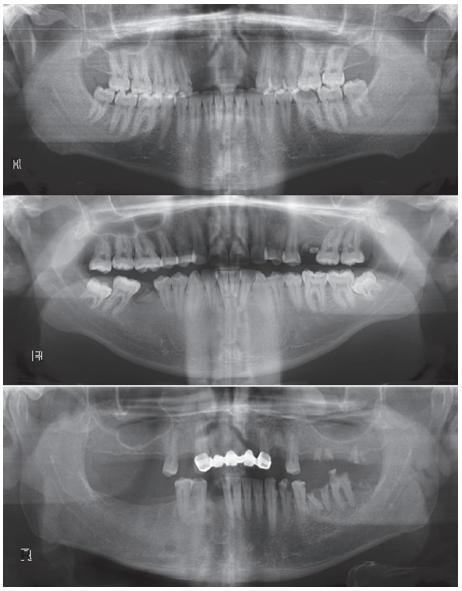
The SEI reflects the quality degradation in terms of intensity saturation caused by the contrast overshoot. The value of SEI is expected to be as low as possible, ideally 0. SOR is a metric that quantitatively shows how far the quality of edges in the sharpened images is degraded by amplified noise content. Like SEI, the value of SOR is also supposed to be as low as possible. A good sharpening, free from haloes is always expected to produce thin and sharp edges. While assessing the quality of edges, their width also needs to be taken into account. The EMBCM is the only sharpness metric that considers the width of the edges as well. EMBCM is an unbounded statistic whose value is expected to be as high as possible. VIF accounts for the overall information fidelity in the sharpened images concerning that in the input image. Like EMBCM, the value of VIF will be high if the sharpening produces artefact-free (free from overshoot, noise-amplification, haloes, and discontinuity artefact) output images with thin and sharp edges.

2.4. SYSTEM REQUIREMENTS

All experiments are conducted with Matlab® software. The version of the software is 7.12.0.635 (R2018a) with License Number: 161052. The software is installed in a desktop computer with Intel (R), Pentium (R) Processor, and CPU B950 @ 2.10 GHz. The installed Memory (RAM) is 4 GB. Operating System is Microsoft Windows 7 Version 6.1 (Build 7601: Service Pack 1), 64-bit.

3. RESULTS

The influence of the selection of the scale value in GNS on the subjective quality of the sharpened images and the objective quality measures like SEI, SOR, EMBCM, and VIF is discussed in this section. Out of the total 116 radiographs, three test images for which the pictorial results are presented are shown in Fig. 2.



(a)

(b)

(c)

Fig. 2. Test images (**a**) Test image 1 (**b**) Test image 2 (**c**) Test image 3

Enlarged portions of the output images of GNS corresponding to the input image 1, input image 2 and input image 3 for various scale values are shown in Fig. 3 to Fig. 5. When the scale is equal to 1, (Fig. 3 (b), Fig. 4 (b) and Fig. 5 (b)) significant improvement in edge strength is not evident. When the scale value increases from 1 to 4, (Fig. 3 (c) to Fig. 3 (d), 4 (c) to Fig. 4 (d) and 5 (c) to Fig. 5 (d)), the strength of edges increases to a clearly visible level. When the scale value is greater than 5, (Fig. 3 (f) to Fig. 3 (i), 4 (f) to Fig. 4 (i) and 5 (f) to Fig. 5 (i)), the noise content gets amplified slightly. However, the amplification of noise is not as critical as in the Unsharp Masking. The value of the scale in GNS is within the range of 3 to 5, the strength of edges in the dental radiographs gets enhanced to a visibly apparent level without amplifying the noise significantly. The observation is found to be consistent on all 116 test images.

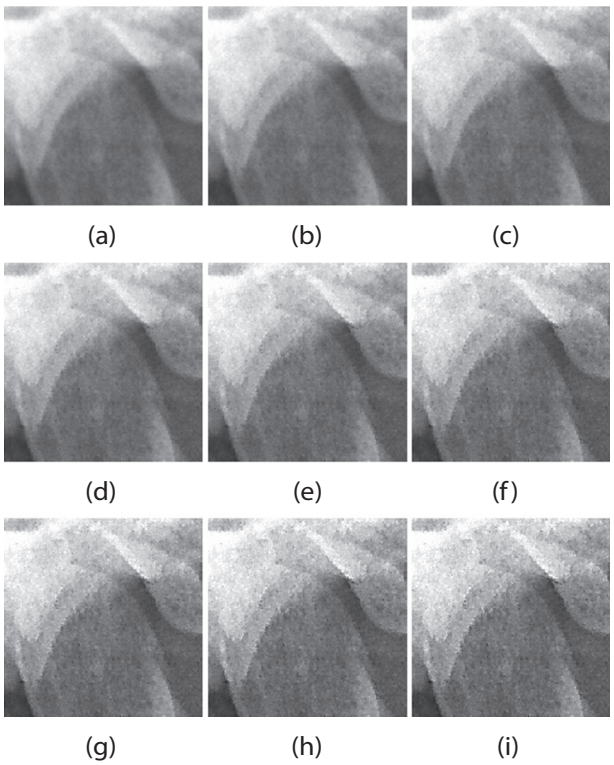


Fig. 3. Enlarged portions of the output images of GNS corresponding to the input image 1 for various scale values (**a**) Enlarged portion of the input image 1 (**b**) Scale = 1 (**c**) Scale = 2 (**d**) Scale = 3 (**e**) Scale = 4 (**f**) Scale = 5 (**g**) Scale = 6 (**h**) Scale = 7 (**i**) Scale = 8

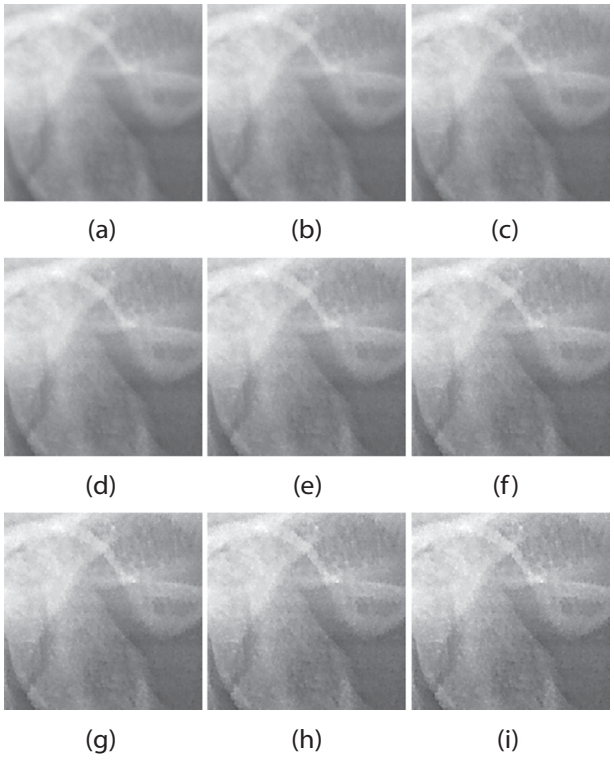


Fig. 4. Enlarged portions of the output images of GNS correrspondingto the input image 2 for various scalevalues (**a**) Enlarged portion of the input image 2 (**b**) Scale = 1 (**c**) Scale = 2 (**d**) Scale = 3 (**e**) Scale = 4 (**f**) Scale = 5 (**g**) Scale = 6 (**h**) Scale = 7 (**i**) Scale = 8

Influence of the scale value on SEI, SOR, EMBCM and VIF is illustrated in Fig. 6. The variations of the SEI with scale for three test radiographs are shown in Fig. 6 (a). The slopes of the SEI versus scale curves for all three test images are very low as seen in Fig. 6 (a). SEI does not increase critically in response to the increase in the value of scale. Even at high values of the scale, the increase in SEI is not significant. This is a clear indication that GNS produces sharpened images free from intensity saturation. The variations of the SOR with scale for three test radiographs are shown in Fig. 6 (b). Similar to the SEI versus scale curves, slopes of the SOR versus scale curves for all three test images are very low as seen in Fig. 6 (b). SOR does not increase critically, in response to the increase in the value of scale. Even at high values of the scale, the increase in SOR is not significant. This is a clear indication that even at high values of the scale, quality of the edges in the output images produced by the GNS is not significantly affected by the amplified noise content.

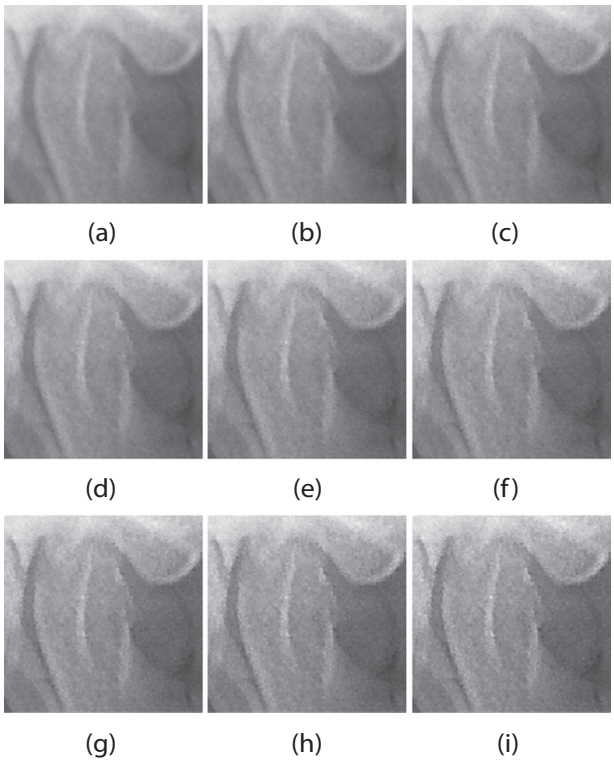
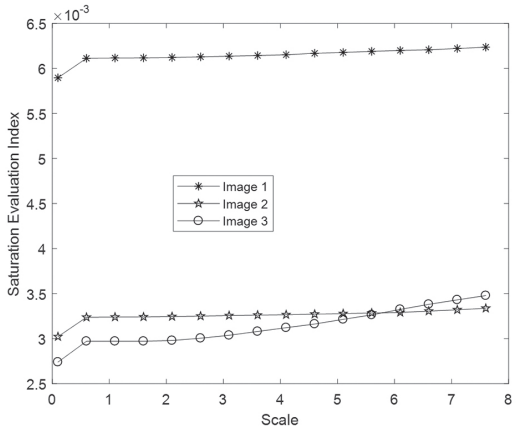


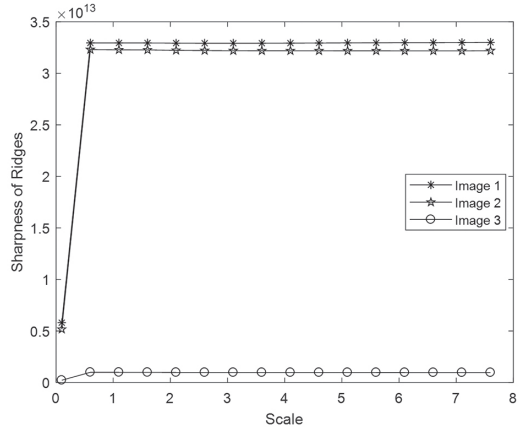
Fig. 5. Enlarged portions of the output images of GNS correrspondingto the input image 3 for various scalevalues (**a**) Enlarged portion of the input image 3 (**b**) Scale = 1 (**c**) Scale = 2 (**d**) Scale = 3 (**e**) Scale = 4 (**f**) Scale = 5 (**g**) Scale = 6 (**h**) Scale = 7 (**i**) Scale = 8

In the EMBCM versus scale curves in Fig. 6 (c), the slopes of the curves are relatively higher when the value of the scale is less than 4 compared to the slopes of the curves when the scale is above 4. As mentioned already, the EMBCM reflects the quality of the edges in terms of their sharpness and thinness. From the EMBCM versus scale curves, it can be inferred that GNS offers relatively better output images comprising sharper

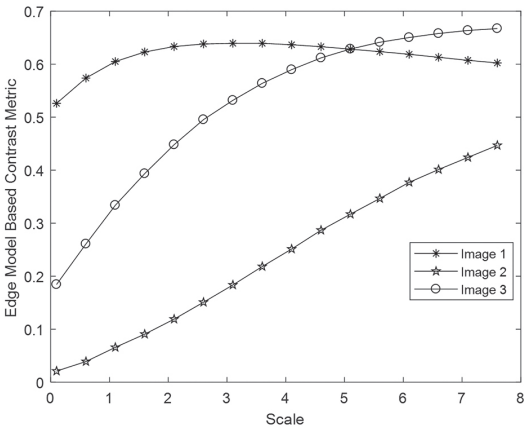
and thinner edges when the scale value is less than 4. In the VIF versus scale curves in Fig. 6 (d), VIF is at its maximum for the input image 1, when the scale value is equal to 3. For input image 2 and input image 3, the maximum values of VIF are observed at the scale values equal to 3.5 and 6.5, respectively. As per the observations drawn out from the VIF versus scale curves of 116 test radiographs, the VIF is found to be at its maximum for a range of scale values between 3 and 7. From the variations of the SEI, SOR, EMBCM, and VIF against the scale values, the practically efficient range of scale suitable for dental radiographs can be considered as between 4 and 5.



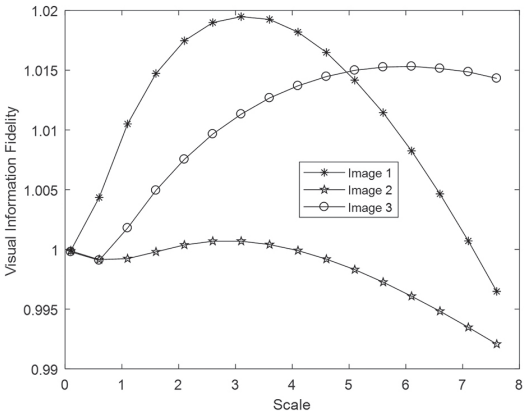
(a)



(b)



(c)



(c)

Fig. 6. Influence of the scale value on various objective image quality measures (a) SEI versus scale (b) SOR versus scale (c) EMBCM versus scale (d) VIF versus scale

4. DISCUSSIONS

In this section, the performance of the GNS is compared with the Unsharp Masking in terms of quality of the sharpened images and the values of the objective quality measures. Enlarged portions of the output images of the Unsharp Masking and GSN for three test radiographs are shown in Fig. 7 to Fig. 9.

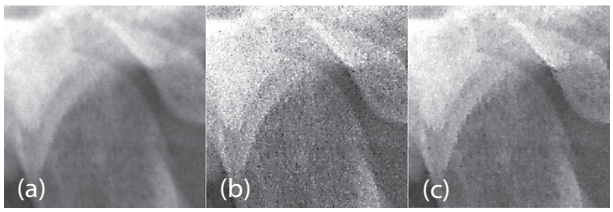


Fig. 7. Sharpened images (**a**) Enlarged portion of the input image 1 (**b**) Enlarged portion of the output of Unsharp Masking (**c**) Enlarged portion of the output of GNS

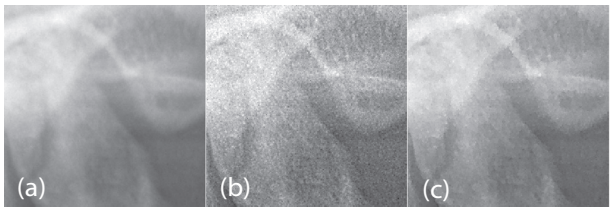


Fig. 8. Sharpened images (**a**) Enlarged portion of the input image 2 (**b**) Enlarged portion of the output of Unsharp Masking (**c**) Enlarged portion of the output of GNS

In both the Unsharp Masking and the GNS, the value of scale is set as equal to 5. The threshold in the Unsharp Masking is set equal to 0.1. It can be seen in the output images of the Unsharp Masking furnished in Fig. 7 (b), Fig. 8 (b) and Fig. 9 (b) that noise content is significantly amplified. It is evident in Fig. 7 (c), Fig. 8 (c) and Fig. 9 (c), GNS produces output images in which the edges are

sharper than those in the input images furnished in 7 (a), Fig. 8 (a) and Fig. 9 (a). Amplification of the noise content in GNS is not as critical as in the Unsharp Masking.

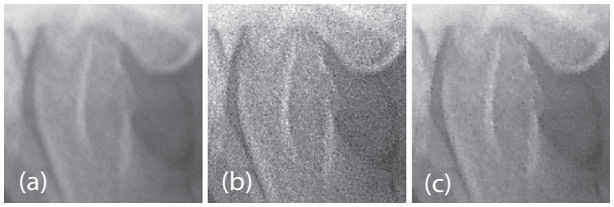


Fig. 9. Sharpened images (**a**) Enlarged portion of the input image 3 (**b**) Enlarged portion of the output of Unsharp Masking (**c**) Enlarged portion of the output of GNS

Values of SEI, SOR, EMBCM and VIF shown by the Unsharp Masking and GNS, corresponding to the output images furnished in Fig. 7 to Fig. 9 are shown in table 1 and table 2, respectively. The summary of these quality metrics on 116 test radiographs is furnished in table 3. GNS exhibited lower values of SEI and SOR and higher values of EMBCM and VIF, compared to the Unsharp Masking. Lower values of SEI and SOR, respectively indicate that GNS is free from overshoot artefact and saturation and the quality of edges in the output images of GNS is less affected by noise. Higher values of EMBCM and VIF, respectively confirm that GNS is free from haloes as it produces thin and sharp edges and the sharpened images are of good information fidelity.

Table 1. Values of quality metrics shown by the Unsharp Masking

Quality Metric	Image 1	Image 2	Image 3
SEI	0.0072	0.0036	0.0035
SOR	7.3205×10^{13}	5.7376×10^{13}	1.4223×10^{12}
EMBCM	0.5604	0.0227	0.2070
VIF	0.9757	0.9979	0.9957
Computational Time (S)	**0.37**	**0.29**	**0.19**

Table 2. Values of quality metrics shown by the GNS

Quality Metric	Image 1	Image 2	Image 3
SEI	**0.0062**	**0.0033**	**0.0032**
SOR	$\mathbf{3.2948 \times 10^{13}}$	$\mathbf{3.2183 \times 10^{13}}$	$\mathbf{9.7641 \times 10^{11}}$
EMBCM	**0.6297**	**0.3117**	**0.6259**
VIF	**1.0147**	**0.9985**	**1.0149**
Computational Time (S)	1626.05	1773.22	1595.04

Table 3. Summary of quality metrics shown by the Unsharp Masking and GNS on 116 images

Quality Metric	Unsharp Masking	GNS
SEI	0.0048 ± 0.0021	0.0042 ± 0.0017
SOR	$4.4 \times 10^{13} \pm 3.8 \times 10^{13}$	$2.2 \times 10^{13} \pm 1.8 \times 10^{13}$
EMBCM	0.2634 ± 0.2732	0.5224 ± 0.1825
VIF	0.9898 ± 0.0122	1.0094 ± 0.0094
Computational Time (S)	0.2833 ± 0.0902	1664.8 ± 95.19

Table 4. Comparison of the proposed model with other state of art architectures using Unsharp Masking mechanisms, without considering the tolerance range

Methods	SEI	SOR	EMBCM	VIF
Manoj et al. [19]	0.0044	3.7×10^{13}	0.4178	0.8990
Clark et al. [20]	0.0098	5.7×10^{13}	0.2065	0.8956
Geraets et al. [18]	0.0048	4.4×10^{13}	0.2634	0.9898
Proposed Model	**0.0042**	$\mathbf{2.2 \times 10^{13}}$	**0.5224**	**1.0094**

It is believed that noise-amplification in the Unsharp Masking can be reduced by increasing the value of the threshold. But increasing the value of the threshold may introduce discontinuity artefact. Enlarged portions of the output images of the Unsharp Masking corresponding to the three test images for a threshold value equal to 0.5 and scale equal to 5 are shown in Fig. 10. Black and white spots are visible on the output images of the Unsharp Masking. These spots are caused by the discontinuity artefact. As no thresholding process is involved in GNS, the output images produced by GNS are free from the discontinuity artefact.

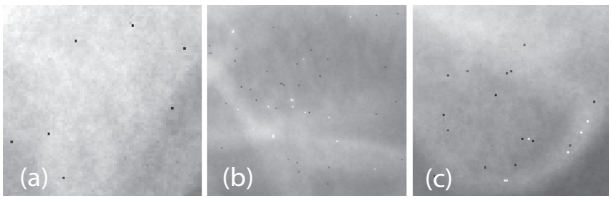


Fig. 10. Discontinuity artefact in the Unsharp Masking (**a**) Enlarged portion of input image 1 sharpened by the Unsharp Masking (**b**) Enlarged portion of input image 2 sharpened by the Unsharp Masking (**c**) Enlarged portion of input image 3 sharpened by the Unsharp Masking

To evaluate the efficacy of the proposed model, the scores of the quality metrics namely SEI, SOR, EMBCM, and VIF are compared to other architectures employed for dental radiograph enhancements. The methods have been evaluated on the Mendeley data repository [18]. The finding is illustrated in Table 4. It can be observed that the proposed model outperforms the competitors in all the metrics. One downside of the proposed model is that its computational time required is on the higher scale compared to its competitors.

5. CONCLUSION AND FUTURE SCOPE

5.1. CONCLUSION

Summary of Contributions: A noise-robust and overshoot-free alternative to the Unsharp Masking, called Gradient-adaptive Nonlinear Sharpening (GNS), for enhancing the quality of edges in dental radiographs was proposed in this paper.

Summary of Observations: GNS exhibited lower values of SEI and SOR and higher values of EMBCM and

VIF, compared to the Unsharp Masking. Lower values of SEI and SOR, respectively indicate that GNS is free from overshoot artefact and saturation and the quality of edges in the output images of GNS is less affected by noise. Higher values of EMBCM and VIF, respectively confirm that GNS is free from haloes as it produces thin and sharp edges and the sharpened images are of good information fidelity.

Advantages of GNS: Unlike the conventional Unsharp Masking, the computation of sharpened images in the proposed GNS does not involve any thresholding operation. Hence GNS is free from discontinuity artefact. GNS has less number of operational parameters compared to the Unsharp Masking. Because of the Adaptive Edge Smoothing Filter (AESF) and local gradient-based weighting of the difference image in LGUM, it is free from noise-amplification and sharpness overshoot. GNS selectively sharpen the edges without amplifying the noise content.

Applications / Commercial Viability of GNS: GNS can be used as a pre-processing algorithm in image processing tools used in clinical practice and forensic odontology for the automated analysis of dental radiographs. It helps to improve the efficiency of segmentation algorithms used in such image processing tools.

5.2. LIMITATIONS OF GNS AND FUTURE SCOPE

The GNS has three operational parameters. They are the radius of the AESF, maximum limit of the standard deviation of AESF and scale. Adaptive methods for identifying the optimum value of these operational parameters will help to eliminate the burden of tuning them. The GNS is computationally intense compared to the conventional Unsharp Masking, because of the local processing strategy adopted in it. Methods which can accelerate the computation of GNS will make it more advantageous.

6. REFERENCES:

[1] J. Krois et al. "Deep learning for the radiographic detection of periodontal bone loss", Scientific reports, Vol. 9. No. 1, 2019, p.8495.

[2] J. Kim, H. Lee, I. Song, "DeNTNet: Deep Neural Transfer Network for the detection of periodontal bone loss using panoramic dental radiographs", Nature Scientific Reports, Vol. 9, No. 1, 2019, p. 17615.

[3] M.A. Alzubaidi, M. Otoom, "A comprehensive study on feature types for osteoporosis classification in dental panoramic radiographs", Computer Methods and Programs in Biomedicine, Vol. 188, 2020, p. 105301.

[4] I. Aliaga, V. Vera, M. Vera, E. García, M. Pedrera, G. Pajares, "Automatic computation of mandibular indices in dental panoramic radiographs for early osteoporosis detection", Artificial Intelligence in Medicine, Vol. 103, 2020, p. 101816.

[5] S. Javed, M. Zakirulla, R. U. Baig, S. M. Asif, A. B. Meer, "Development of artificial neural network model for prediction of post-streptococcus mutans in dental caries", Computer Methods and Programs in Biomedicine, Vol. 186, 2020, p. 105198.

[6] A. Heinrich, F. V. Güttler, S. Schenkl, "Automatic human identification based on dental X-ray radiographs using computer vision", Nature Scientific Reports, Vol. 10, No. 1, 2020, p. 3801.

[7] H. Chen, K. Zhang, P. Lyu, "A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films", Nature Scientific Reports, Vol. 9, No. 1, 2019, pp. 1-11.

[8] S. Vinayahalingam, T. Xi, S. Bergé, "Automated detection of third molars and mandibular nerve by deep learning", Nature Scientific Reports, Vol. 9, No. 1, 2019, p. 9007.

[9] T. E. Southard, K. A. Southard, "Performance of filters for noise reduction in maxillary alveolar bone imaging", IEEE Transactions on Biomedical Engineering, Vol. 42, No. 1, 1995, pp. 13-20.

[10] I. Frosio, C. Olivieri, M. Lucchese, N. A. Borghese, P. Boccacci, "Bayesian denoising in digital radiography: A comparison in the dental field", Computerized Medical Imaging and Graphics, Vol. 37, No. 1, 2013, pp. 28-39.

[11] H. Rahmi-Fajrin, S. Puspita, S. Riyadi and E. Sofiani, "Dental radiography image enhancement for treatment evaluation through digital image processing", Journal of Clinical and Experimental Dentistry, Vol. 10, No. 7, 2018, pp. 629-634.

[12] I. Draganov, V. Gancheva. "Unsharp Masking with Local Adaptive Contrast Enhancement of Medical Images", Proceedings of International Conference on Medical Imaging and Computer-Aided Diagnosis Medical Imaging and Computer-Aided Diagnosis, Lecture Notes in Electrical Engineering, Springer, Singapore, Vol. 784, 2021, pp. 354-363.

[13] Y. Yan et al. "Bi-SCM: bidirectional spiking cortical model with adaptive unsharp masking for mam-

mography image enhancement", Multimedia Tools and Applications, Vol. 82, No. 8, 2023, pp. 12081-12098.

[14] A. Kumar, P. Kumar, S. Srivastava, "A skewness reformed complex diffusion based unsharp masking for the restoration and enhancement of Poisson noise corrupted mammograms", Biomedical Signal Processing and Control, Vol. 73, 2022, p. 103421.

[15] M. Fujita, Y. Kodera, M. Ogawa, K. Tanimoto, T. Sunayashiki, T. Wada, K. Doi, "Digital image processing of dento-maxillofacial radiographs", Oral Surgery, Oral Medicine, Oral Pathology, Vol. 64, No. 4, 1987, pp. 485-493.

[16] T. Harada, K. Nishikawa, K. Kuroyanagi, "Unsharp masking technique as a pre-processing filter for improvement of 3D-CT image of bony structure in the maxillofacial region", Oral Radiology, Vol. 14, 1998, pp. 85-94.

[17] R. A. Couture, B. R. Whiting, C. F. Hildebolt, D. A. Dixon, "Visibility of trabecular structures in oral radiographs", Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontology, Vol. 96, No. 6, 2003, pp. 764-771.

[18] W. Geraets, G. Jonasson, M. Hakeberg, "Predicting fractures using trabecular patterns on panoramic radiographs", Clinical Oral Investigations, Vol. 22, No. 1, 2018, pp. 377-384.

[19] M. T. Joy, B. P. Shan, G. Titus, "Restoration of dental radiographs corrupted by quantum noise via a noise-adaptive sub-global weighted sum of radiometrically similar pixels", International Journal of Computational Biology and Drug Design, Vol. 15, No. 3, 2023, pp. 249-266.

[20] J. L. Clark, C. P. Wadhwani, K. Abramovitch, D. D. Rice, M. T. Kattadiyil, "Effect of image sharpening on radiographic image quality", The Journal of Prosthetic Dentistry, Vol. 120, No. 6, 2018, pp. 927-933.

[21] D. Brettle, F. Carmichael, "The impact of digital image processing artefacts mimicking pathological features associated with restorations", British Dental Journal, Vol. 211, No. 4, 2011, pp. 167–170.

[22] A. H. Abdi, S. Kasaei, "Panoramic Dental X-rays with Segmented Mandibles", Mendeley Data, V1, 2017, http://dx.doi.org/10.17632/hxt48yk462.1.

[23] A. H. Abdi, S. Kasaei, M. Mehdizadeh, "Automatic segmentation of mandible in panoramic x-ray", Journal of Medical Imaging, Vol. 2, No. 4, 2015, p. 44003.

[24] L. Krasula, P. Le Callet, K. Fliegel, M. Klíma, "Quality Assessment of Sharpened Images: Challenges, Methodology, and Objective Metrics", IEEE Transactions on Image Processing, Vol. 26, No. 3, 2017, pp. 1496-1508.

[25] J. Joseph, R. Periyasamy, "Nonlinear sharpening of MR images using a locally adaptive sharpness gain and a noise reduction parameter", Pattern Analysis and Applications, Vol. 22, No. 1, 2019, pp. 273-283.

[26] J. Joseph, R. Periyasamy, "A fully customized enhancement scheme for controlling brightness error and contrast in magnetic resonance images", Biomedical Signal Processing and Control, Vol. 39, 2018, pp. 271-283.

[27] V. S. Hari, V. P. J. Raj, R. Gopikakumari, "Unsharp masking using quadratic filter for the enhancement of fingerprints in noisy background", Pattern Recognition, Vol. 46, No. 12, 2013, pp. 3198-3207.

[28] J. Guan, W. Zhang, J. Gu, H. Ren, "No-reference blur assessment based on edge modelling", Journal of Visual Communication and Image Representation, Vol. 29, 2015, pp. 1-7.

[29] H. R. Sheikh, A. C. Bovik, G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics", IEEE Transactions on Image Processing, Vol. 14, No. 12, 2005, pp. 2117- 2128.

Eight-Port Tapered-Edged Antenna Array With Symmetrical Slots and Reduced Mutual-Coupling for Next-Generation Wireless and Internet of Things (IoT) Applications

Original Scientific Paper

Bilal A. Khawaja
Faculty of Engineering, Department of Electrical Engineering,
Islamic University of Madinah,
P. O. BOX 170 Madinah, 41411, Saudi Arabia
7166@iu.edu.sa, bam.khawaja@gmail.com

Abstract *– A compact and low-cost eight-port (2x4 configuration) tapered-edged antenna array (TEAA) with symmetrical slots and reduced mutual-coupling is presented in this paper using the inset-feed technique. The 8-port TEAA is designed and simulated using CST microwave studio, fabricated using the flame-resistant (FR4) substrate having a dielectric constant (ε_r) = 4.3 and thickness (h) = 1.66mm and characterized using Keysight technologies vector network analyzer (VNA). The designed 8-port TEAA operates at the 5.05-5.2GHz frequency band. Various performance design parameters, like return-loss, bandwidth, gain, 2D/3D radiation patterns, surface current distributions, and isolation-loss, are briefly studied, and the results are summarized. The eight-port TEAA has featured the bandwidth/gain characteristic of 195MHz/10.25dB, 3dB beam-width of 52.8°, and excellent mutual-coupling (high isolation-loss) of less than -20dB, respectively. The 8-port TEAA is proposed and characterized to work for next-generation high-throughput WLANs like IEEE 802.11ax (WiFi-6E), Internet-of-Things (IoT), and the upcoming 5G wireless communication systems.*

Keywords*: Antenna array, IEEE 802.11ax, WiFi-6E, Internet-of-Things (IoT), 5G*

1. INTRODUCTION

During the last 15 years, the world has seen a rapid upsurge in the use of wearable smart gadgets and handheld wireless devices like smartwatches, smart glasses, fitness bands, etc., and mobile phones and tablets, respectively [1, 2]. The end-users are dependent on these devices for their day-to-day tasks, which vary from financial transactions to grocery shopping, online education to work-from-home business users, food delivery to live-streaming of videos and sports events, Metaverse-type online gaming as well as usage of artificial intelligence (AI) chatbots and social media applications, and smart-home monitoring to smart-farming solutions are to name the few [1-5]. This dependency is due to the advancement in the currently deployed 5G and future high-throughput wireless systems [1-3], [6]. The wireless end-users are exponentially growing, and it is estimated that by 2025, there will be 75 billion wireless devices linked to the Internet, up from the current figure of ~23 billion [7]. In order to uphold this ever-increasing demand for commercial and residential wireless end-users, IEEE has standardized an upgraded version of IEEE 802.11ac [8] Wi-Fi, which is also known as IEEE 802.11ax (WiFi-6) [8]. In comparison to 802.11ac, the 802.11ax can achieve an overall 37% increase in the data-rates, which is equivalent to a data-transmission speed of 9.6GB/s [8]. Another enhancement of WiFi-6 is the 2.4GHz and 5GHz dual frequency band operation as compared to the single-band operation of IEEE 802.11ac, which only operates at 5GHz. In the years to come, enhanced wireless systems standards like WiFi-6E [8] will also be available, which are also referred to as high-throughput wireless local area networks (WLANs) [1, 2, 6, 8]. WiFi-6E systems will operate in the triple frequency bands of 2.4GHz/5GHz and 6GHz, respectively [8].

The main reason for the technological advancement of currently available WLAN standards and the enhancement of the next-generation of high-throughput WLANs like WiFi-6E, as well as the 5G wireless systems, is the requirement for antennas that are compact, exhibit wide bandwidth, and can be easily integrated with the upcoming wireless devices and systems. In modern WLAN systems, planar antennas are typically used, which are referred to as microstrip patch an-

tennas (MPAs) [9]. The reasons MPA based on planar structures are preferred in modern wireless systems, because of low-cost, low-profile, ease-of-design, compatibility, and conformality with the printed circuit boards (PCBs) [8-10]. A conventional MPA resonates at a single frequency and offers reduced gain and narrow bandwidth. Although by the insertion of different stubs and slots to change the current path on the resonating patch, the MPA can be made to oscillate at multiple frequency bands [8-10]. Using the MPA design techniques, many researchers [10-15] have proposed different antenna array designs for the 5G wireless systems operating at both microwave and mmW frequency bands [12].

Recently, there has been much interest in the multi-port antenna array systems which can operate and cover the frequency bands of next-generation high-throughput WLANs like WiFi-6/WiFi-6E, and the upcoming 5G system [16-21]. Multi-port antenna structures are preferred and selected by a number of researchers because they offer the benefit of physical antenna size reduction and also reduce the need for antenna-feeding networks, also referred to as beam-forming networks for phased array antenna systems [22]. In [16], an eight-element single-polarization linear array antenna has been proposed to enhance the isolation by using the decoupling ground (DG) technique. The array was operating at the 4.8-5GHz band and exhibited a gain of 7.3dB. Then, authors in [17] presented a 3.3-5GHz band operation wideband eight-port antenna array for 5G new radio (5G-NR) applications by employing defected ground structure (DGS) to achieve high isolation between radiating elements. Also, researchers in [18] presented an eight-port antenna system design for future 5G mobile broadband services operating in the C-band at the 5GHz band and demonstrated a gain of 7.4dB, respectively. In [19], an 8-element array is proposed for 5G systems operating at 3.27-5.92GHz based on L-shaped open slot antennas with an increased size of 150 × 75 mm^2. An 8-element antenna array is proposed for 5G-NR and WiFi-6 applications by using the gap-coupled feeding technique and open-stub for improved bandwidth. The array achieved a gain of 4dB, which is considered relatively quite low for multi-port antenna systems [20]. Lastly, in [21], the authors presented an 8-element antenna array operating at 5.15-5.97GHz frequency band for 5G systems. The proposed antenna system uses a hybrid combination of modified Minkowski and Peano curves [14] fractal geometries. The proposed array [21] achieves good isolation of over 10dB between the antenna elements, although exhibited a reduced gain of 4dB only. The research study [16-21] shows that most of the researchers were trying to improve the bandwidth and mutual-coupling between the element of 8-port antenna arrays by using different complex techniques like DG and DGS. This leads to improved bandwidth and mutual-coupling but also results in reduced antenna array gain and larger-sized structures. Although, in this paper, the proposed 8-port array exhibits an overall size reduction of 55% with a much-improved gain as compared to other similar designs. Moreover, it also shows an excellent mutual-coupling without using complicated design techniques.

So, in this work, the authors initially present a single-element antenna with tapered edges and symmetrical slots in the oscillating patch with an inset-feed technique using the FR4 substrate for low-cost implementation of the design. All the required antenna performance parameters are then briefly studied, and the results are compared. Then, an 8-port (2x4 configuration) tapered-edged antenna array (TEAA) has been proposed and studied using the single-element antenna design. The 8-port TEAA is proposed and designed to work for the next-generation of high-throughput WLANs like WiFi-6E, sub-6GHz 5G systems, and Internet of Things (IoT) and C-band applications [3, 5, 8]. The proposed single-element antenna and 8-port TEAA are designed to operate in the 5GHz frequency band. The proposed 8-port TEAA in this work has many attractive features, such as a high-gain and maximum intensity of radiation pattern at 45°. These features make the 8-port TEAA a worthy candidate for the upcoming high-throughput WLANs, sub-6GHz 5G, IoT, and next-generation wireless applications. All the 8-elements of the 2x4 configuration TEAA are individually fed using SMA connectors. Another salient feature of the proposed 8-port TEAA is that it exhibits reduced mutual-coupling (i.e., high isolation-loss) between the antenna elements, which is one of the critical requirements for the antenna arrays designed to work for next-generation WLAN applications.

The structure of the paper is as follows: Single-element antenna design parameters, simulation, and measurement results like return-loss (S11), 2D/3D radiation patterns, and surface current distribution are studied and summarized in Section 2. In Section 3, the 8-port (2x4 configuration) tapered-edged antenna array, design, and characterization results are discussed. In addition to all of the above performance parameters, the isolation-loss is also studied for the 8-port array to study the mutual-coupling effects between the antenna elements of the array. Section 4 provides a comparison of the results with other researchers' work, and finally, Section 5 concludes the paper.

2. SINGLE-ELEMENT ANTENNA DESIGN AND CHARACTERISATION RESULTS

Initially, a single-element antenna is proposed and designed with tapered edges and symmetrical slots in the oscillating patch, as shown in Fig. 1(a–c). The inset-feed technique is also employed in this design to achieve better impedance matching between the 50Ω feed-line and the oscillating patch. The substrate used to design and realize the antenna is FR4 having the thickness (h), dielectric constant (ε_r), and loss-tangent (tan δ) of 1.6mm, 4.3, and 0.025, respectively.

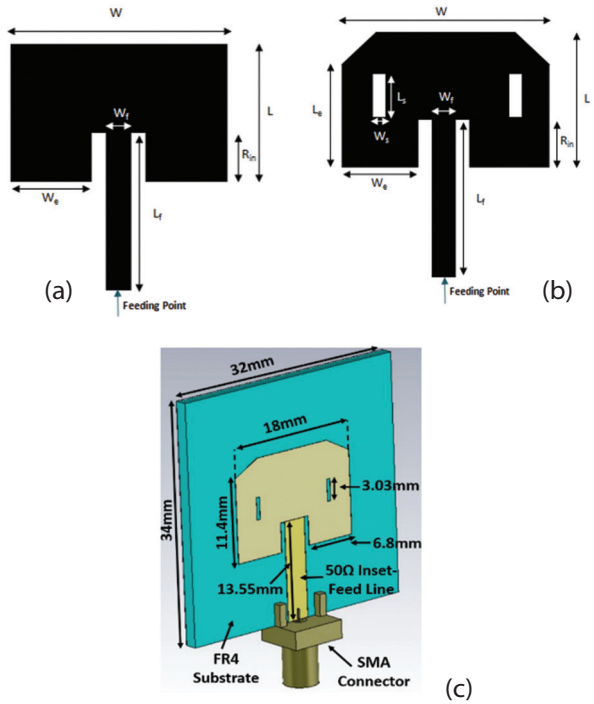


Fig. 1. (**a**) Basic rectangular MPA with inset-feed (**b**) Modified and optimized tapered-edged MPA with symmetrical slots (**c**) Layout of the final single-element antenna design from CST-MWS with dimensions

The patch antenna's initial calculations were conducted using the transmission-line model [9-10]. All the respective formulas are given below in Eq. (1) – (5):

$$L = L_{eff} - 2\Delta L \tag{1}$$

$$L_{eff} = \frac{c}{2f\sqrt{\epsilon_{reff}}} \tag{2}$$

$$\epsilon_{reff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\left[\frac{1}{\sqrt{1 + 12\frac{h}{W}}}\right] \tag{3}$$

$$\Delta L = 0.412h\frac{[\epsilon_{reff} + 0.3]\left[\frac{W}{h} + 0.264\right]}{[\epsilon_{reff} - 0.258]\left[\frac{W}{h} + 0.8\right]} \tag{4}$$

$$W = \frac{c}{2f}\sqrt{\frac{2}{\epsilon_r + 1}} \tag{5}$$

Where:

L = Actual length of the patch

W = Actual width of the patch

L_{eff} = Effective length of the patch

ϵ_{reff} = Effective dielectric constant of the patch

ΔL = Extended incremental length of the patch

CST Microwave Studio (CST-MWS) has been used to design, analyze and simulate the different antenna parameters. The overall size of the proposed antenna is 34 mm × 32 mm, as shown in Fig. 1(c).

The resonating patch of the antenna was designed to resonate at 5.2GHz. Fig. 1(c) and Table 1 shows the proposed antenna dimensions.

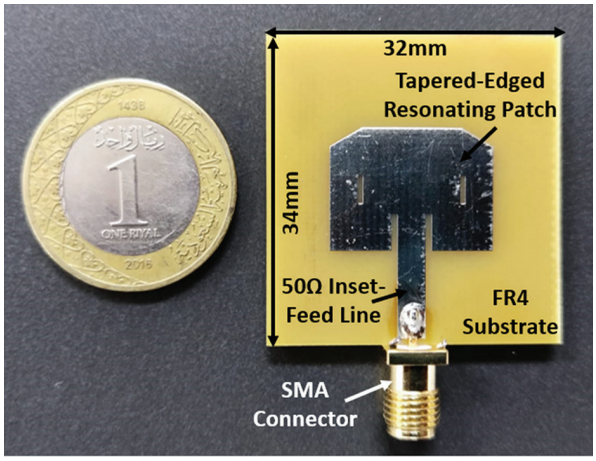
Table 1. Proposed single-element antenna dimensions.

Parameter	Dimensions (mm)
W	18
L	13.4
L_f	13.55
W_f	3.1
L_s	3.03
W_s	0.5
L_e	11.4
W_e	6.8
R_{in}	4

The final design of the single-element antenna is shown in Fig. 1(c), which is used to fabricate the antenna prototype. The FR4 substrate-based fabricated prototype of the single-element tapered-edged antenna with symmetrical slots, inset-feed and complete ground-plane is shown in Fig. 2(a). The single-element antenna was characterized using Keysight Technologies FieldFox N9916B Vector Network Analyzer (VNA), as shown in Fig. 2(b).

The antenna-under-test (AUT) connected to VNA Port-1 in Fig. 2(b) shows that the single-element antenna was resonating around the 5.2GHz frequency band. The comparison of the simulated and measured return-loss S11 [dB] curves of the proposed single-element tapered-edged antenna design is summarized in Fig. 3.

It can be seen from the simulated and measured S11 [dB] result that the simulated antenna was designed to resonate at the center frequency of 5.2GHz. Although the measured antenna response depicts a slight shift in the single-element antenna resonant center frequency which is now at 5.1GHz. This could be due to the fabrication tolerances, SMA connector soldering, and loss-tangent variation of the FR4 substrate.



(a)

(b)

Fig. 2. (**a**) Fabricated prototype of single-element antenna using FR4 substrate, and (**b**) S-parameter measurement setup showing the single-element antenna connected to Keysight VNA for S_{11} [dB] measurement

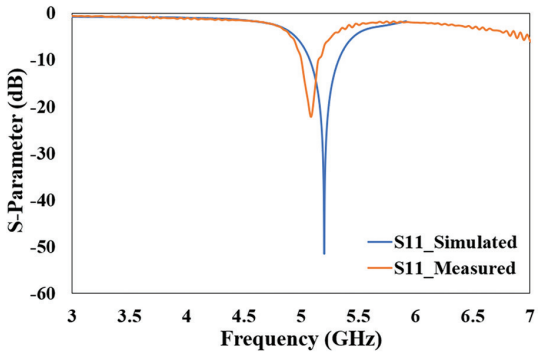


Fig. 3. Simulated and Measured S_{11} [dB] of the proposed single-element tapered-edged antenna with symmetrical slots, and inset-feed

It can also be observed from Fig. 3 that the simulated and measured S_{11} [dB] responses of the single-element antenna give the -10 dB resonant bandwidth of 248 MHz and 195 MHz, respectively. The S_{11} values also suggest good impedance matching at 50 Ω, and reasonable bandwidth. Based on the simulated and measured results, it was found that the impedance bandwidths range from 5.15-5.4 GHz and 5.05-5.2 GHz, with a voltage standing wave ratio (VSWR) that is less than 2 [9, 10]. The VSWR value of the proposed antenna indicates a stable and reliable performance.

Fig. 4 (a-c) now shows the simulated/measured 2D (E-plane/H-plane) and 3D radiation patterns (RPs) of the proposed single-element antenna, respectively. It can be seen from the E-plane and H-plane 2D RP in Fig. 4 (a-b) that the simulated and measured gain is 6.74 dB and 5 dB at 5.2 GHz, respectively. The 3D RP of the proposed single-element antenna is shown in Fig. 4(c). The antenna gain of 6.74 dB can be observed from Fig. 4(c) at the resonant frequency of 5.2 GHz. The 2D simulated and measured RP results show that the antenna radiates maximum at 0° and 10° with fewer side lobes. It is more clear in the 3D RP that the antenna radiation in the boresight with respect to the structure.

After measuring the 2D and 3D RPs, the surface current distribution of the single-element antenna is also studied, and the results are compiled in Fig. 5. The surface current distribution graph of the resonating patch antenna shows that the current density is highest close to the tapered edges and around the symmetrical slots.

3. 8-PORT TAPERED-EDGED ANTENNA ARRAY (TEAA) DESIGN AND CHARACTERISATION RESULTS

Multiple-element antenna arrays, which offer high directionality, gain, and resonant bandwidth, are one of the critical requirements for the WiFi-6E WLANs, currently deployed 5G wireless systems, and next-generation of high-throughput wireless technologies [3, 6, 8, 10, 11, 16-21]. To design the N-elements antenna array, the mathematical Eq. (6) – (8) are used. The 8-port TEAA's total field is equal to the single-element at the origin multiplied by a factor usually referred to as the array factor (AF) [9,10].

$$E_{(Total)} = [E_{(Single\ Element\ at\ Reference\ Point)}]\ x\ [AF] \quad (6)$$

For any N-element array with uniform amplitude and spacing, the AF is given by:

$$\Psi = (kd\ cos\theta\ +\ \beta) \quad (7)$$

$$AF = \left[\frac{sin\left(\frac{N}{2}\Psi\right)}{sin\left(\frac{1}{2}\Psi\right)}\right] \quad (8)$$

Where:

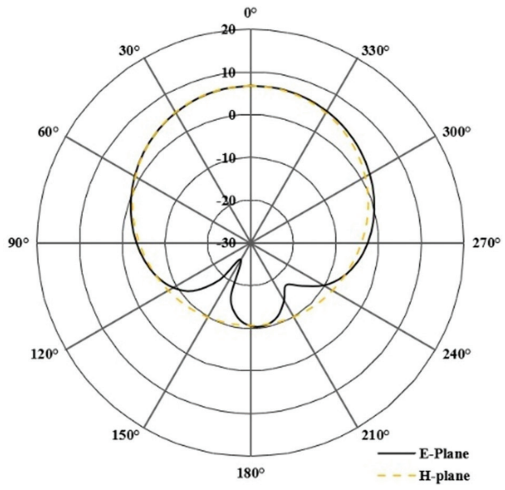
k = Wave number given by $2\pi/\lambda$,

d = Distance between the array elements,

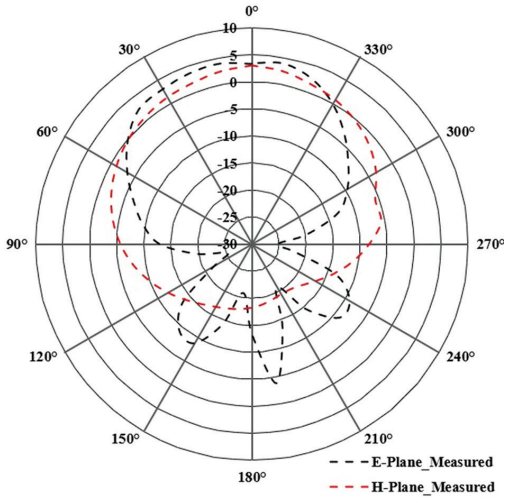
θ = Angle of the main beam of the antenna array,

β = Phase difference between individual elements,

N = No. of antenna elements in an array



(a)



(b)

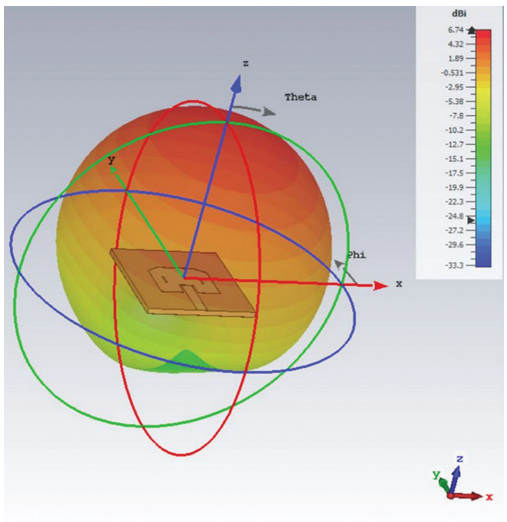


Fig. 4. (**a**) Simulated (**b**) Measured 2D (E-plane and H-plane) and (**c**) 3D Radiation Patterns of the proposed single-element tapered-edged antenna

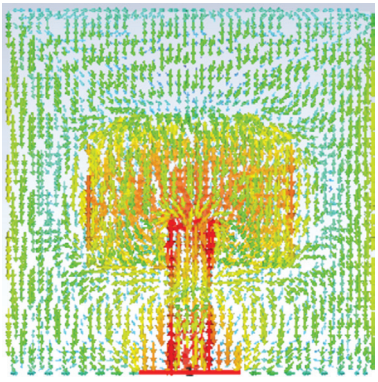
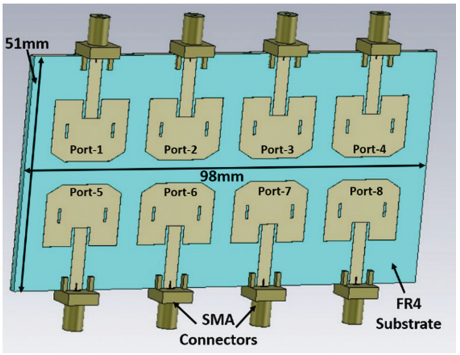
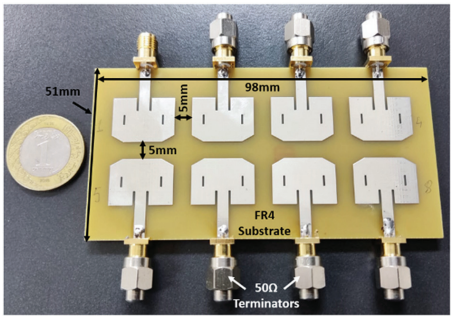


Fig. 5. Surface current distribution of the proposed single-element antenna

So, a 2x4 configuration (8-port) antenna array has been designed, fabricated, and characterized using the previously proposed single-element tapered-edged antenna, as shown in Fig. 6(a-c). The 8-port tapered-edged antenna array (TEAA) is also simulated using the CST-MWS to study its behavior, and fabricated/prototyped using FR4 substrate. The comparison of simulated and measured return-loss responses of the 8-port TEAA are summarized in Fig. 7(a-b).



(a)



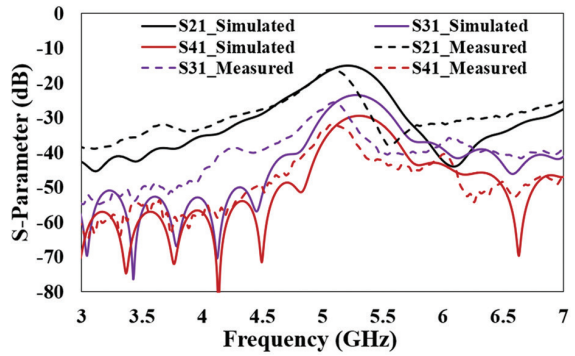
(b)

(c)

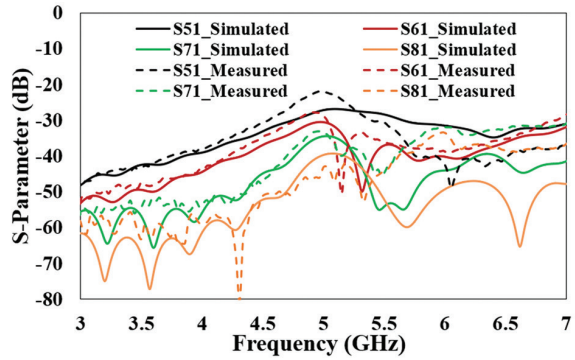
Fig. 6. (**a**) Simulated design of the 8-Port TEAA using CST-MWS showing all the ports and SMA connectors (**b**) Fabricated FR4 prototype of 8-Port TEAA with dimensions (**c**) Keysight VNA-based S-parameter measurement setup of 8-port TEAA, where VNA Port-1 is connected to one port of TEAA and all other ports are terminated using 50Ω terminators

An important performance parameter in the antenna array design is the spacing between the array elements, which is kept at 5mm [9-12]. Moreover, the center-to-center inter-element spacing is a key factor for the reduction of radiation side-lobes and grating. Therefore, in the case of the 8-port TEAA, the center-to-center inter-element spacing is kept at 0.4 λ_0, i.e., 23 mm. It can be observed from the return-loss results in Fig. 7(a-b) that the 8-port TEAA is operating in the 5.2 GHz band as desired. This makes the 8-port TEAA a worthy candidate for the currently deployed 5G wireless and IoT systems as well as for next-generation high-throughput WLAN operations. The 8-port TEAA proposed and realized in this work offers very good simulated/measured bandwidth characteristics of 248 MHz/195 MHz and simulated/measured return-loss responses in the range of -35 dB to -45 dB/-20 dB to -28 dB, at 5.2/5.1 GHz respectively.

After the return-loss results, the isolation-loss study is performed to understand the mutual-coupling behavior between the different antenna elements of the 8-port TEAA. Both simulated and measured results are compared and summarized, as shown in Fig. 7 (a-b). It can be seen from the results in Fig. 7 (a-b) that only the isolation-loss between the TEAA port-1 (antenna element-1) is measured with the rest of the antenna elements. Fig. 7 (a-b) shows very good inter-element isolation characteristics of the 8-port TEAA, and the simulated/measured isolation-loss in the range of -15 dB to -36 dB/-20 dB to -40 dB is achieved around the 5.2 GHz frequency band, respectively.



(a)



(b)

Fig. 7. Shows the simulated and measured isolation-loss of the 8-port TEAA (**a**) S_{21}-S_{41} [dB] (**b**) S_{51}-S_{81} [dB]

The isolation-loss study features the reduced mutual-coupling between the different antenna elements of the 8-port TEAA. The reduced mutual-coupling also shows that the proposed 8-port TEAA can be used as a good option for the deployment of the next-generation high-throughput WLANs and IoT systems. After performing the isolation-loss study, the simulated and measured 2D (E-plane and H-plane) RPs, 3D RPs, and the surface current distribution results of the 8-port TEAA are investigated to understand the behavior of the proposed TEAA for the upcoming next-generation high-throughput WLANs and currently deployed 5G wireless systems. Fig. 8 (a-b) compares the simulated and measured 2D E- and H-plane RPs of the 8-port TEAA at 5.2GHz, which shows that the gain is ~10.3dB and ~10dB, respectively, and the array antenna radiates maximum at 45°.

After the 2D RP, the 3D RP of the 8-port TEAA is studied at the oscillating frequency of 5.2GHz, and the results are summarized in Fig. 8 (c). It can be observed from the 3D RP in Fig. 8 (c) that at 5.2GHz resonance frequency, the TEAA exhibits a gain of 10.25dB, which demonstrates the directionality of the radiating electromagnetic waves from the proposed design. This is an important requirement for both high-efficiency next-generation WLANs and the currently deployed 5G wireless systems because it makes the smart-antenna beam-forming and beam-steering easier to implement [1-3,6, 9, 11, 12].

Fig. 9 shows the surface current distribution of the 8-port TEAA. It can be observed from the surface current distribution graph that it is high across the feeding line, edges of patches, and around the slots on the resonating patch antenna.

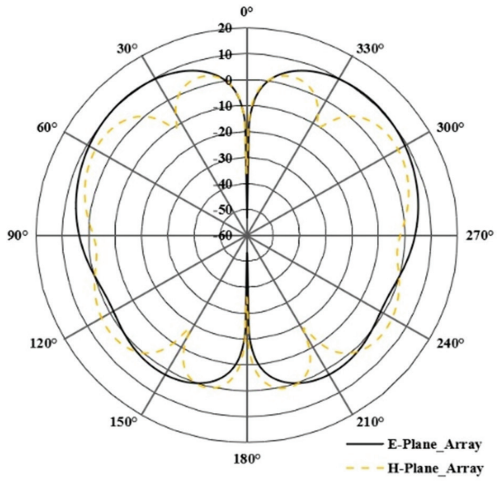
4. COMPARISON STUDY

This section will present a comparison of the 8-port TEAA presented in this paper with the similar 8-port antenna arrays proposed by other researchers in the literature [16-21]. The results are summarized in Table 2, which shows a comparison of the parameters like frequency of operation, No. of antenna elements in the array, antenna array size, gain, and mutual-coupling characteristics.

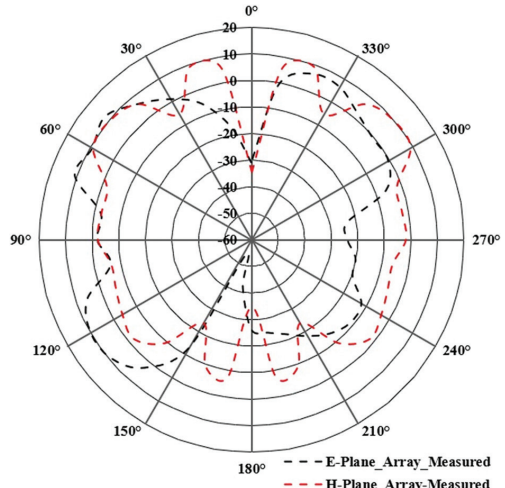
Table 2. A comparison study of the 8-port TEAA with existing designs

Ref.	No. of Array Elements	Frequency Band (GHz)	Size (mm2)	Gain (dB)	Mutual-Coupling (dB)
[16]	8	4.8-5	300 × 60	7.3	<-27
[17]	8	3.3–5.45	150 × 75	3.5	<-18
[18]	8	4.8-5	150 × 75	7.4	<-10
[19]	8	3.27–5.92	150 × 75	-	<-10
[20]	8	5.18-7.71	-	4	<-18
[21]	8	5.15-5.97	150 × 75	4	<-10
[This Work]	8	5.15-5.4	98 × 51	10.25	<-20

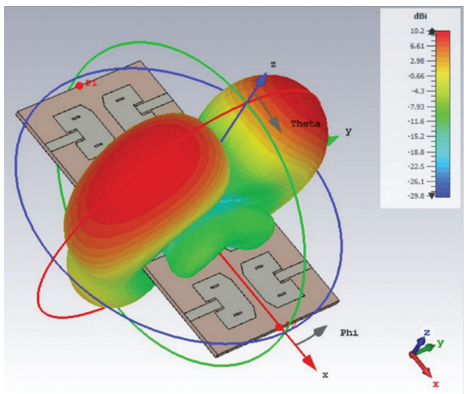
The comparison study presented in Table 2 shows that the previous designs [16-21] presented in the table are more focused on improving mutual-coupling by using different techniques, while also attempting to enhance the bandwidth without much improvement in gain. The results of the proposed 8-port TEAA design show that the size of the array antenna is reduced by 55%, while the gain is drastically improved.



(a)



(b)



(c)

Fig. 8. (**a**) Simulated (**b**) Measured 2D (E-plane and H-plane) and (**c**) 3D Radiation Patterns of the proposed 8-port TEAA at 5.2GHz

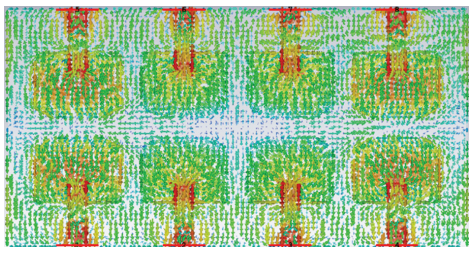


Fig. 9. Surface current distribution of the proposed 8-port TEAA at 5.2GHz

Moreover, the results demonstrate excellent reduced mutual coupling characteristics as compared to previous designs without using complicated design techniques.

5. CONCLUSION

This paper initially presented a single-element antenna with tapered edges and symmetrical slots in the oscillating patch by employing the inset-feed technique for 5.2 GHz frequency band operation. The proposed antenna has shown the bandwidth/gain characteristic of 195 MHz/6.74 dB, and the 3 dB beam-width is 84.5°. Then, by employing the single-element antenna design structure, an eight-port (2x4 configuration) tapered-edged antenna array (TEAA) has been proposed. Both single-element antenna and eight-port TEAA are realized using FR4 substrate for low-cost implementation of the design. The proposed eight-port TEAA covers the 5.05-5.2 GHz frequency band. The eight-port TEAA has also featured the bandwidth/gain characteristic of 195MHz/10.25 dB, 3 dB beam-width of 52.8°, and excellent mutual-coupling (high-isolation loss) of less than -20 dB. All of these features and the reduced-sized structure of the eight-port TEAA make it an ideal candidate for next-generation high-throughput WLANs like WiFi-6E, Internet-of-Things (IoT), and the upcoming 5G wireless communication systems.

6. REFERENCES

[1] O. Kanhere, H. Poddar, Y. Xing, D. Shakya, S. Ju, T. S. Rappaport, "A Power Efficiency Metric for Comparing Energy Consumption in Future Wireless Networks in the Millimeter-Wave and Terahertz Bands", IEEE Wireless Communications, Vol. 29, No. 6, 2022, pp. 56-63.

[2] I. F. Akyildiz, C. Han, Z. Hu, S. Nie, J. M. Jornet, "Terahertz Band Communication: An Old Problem Revisited and Research Directions for the Next Decade", IEEE Transactions on Communications, Vol. 70, No. 6, Jun. 2022, pp. 4250-4285.

[3] K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, M. Mustaqim, "Internet of Things (IoT) for Next-Generation Smart Systems: A Review of Current Challenges, Future Trends and Prospects for Emerging 5G-IoT Scenarios", IEEE Access, Vol. 8, 2020, pp. 23022-23040.

[4] S.-M. Park, Y.-G. Kim, "A Metaverse: Taxonomy, Components, Applications, and Open Challenges", IEEE Access, Vol. 10, 2022, pp. 4209-4251.

[5] S. Qazi, B. A. Khawaja, Q. U. Farooq, "IoT-Equipped and AI-Enabled Next Generation Smart Agriculture: A Critical Review, Current Challenges and Future Trends", IEEE Access, Vol. 10, 2022, pp. 21219-21235.

[6] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, G. C. Trichopoulos, "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond", IEEE Access, Vol. 7, 2019, pp. 78729-78757.

[7] Statista Research Department, "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025", www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ (accessed: 2023)

[8] E. Khorov, A. Kiryanov, A. Lyakhov, G. Bianchi, "A Tutorial on IEEE 802.11ax High Efficiency WLANs", IEEE Communications Surveys & Tutorials, Vol. 21, No. 1, 2019, pp. 197-216.

[9] C. A. Balanis, "Antenna Theory, Analysis and Design", 3rd Edition, John Wiley & Sons, 2005.

[10] H. Wong, K-M. Luk, C. H. Chan, Q. Xue, K. K. So, H. W. Lai, "Small Antennas in Wireless Communications", Proceedings of the IEEE, Vol. 100, No. 7, 2012, pp.2109-2121.

[11] M. C. Tan, M. Li, Q. H. Abbasi, M. Imran, "A Wideband Beam forming Antenna Array for 802.11ac and 4.9 GHz", Proceedings of the Antennas and Propagation 13th European Conference, Krakow, Poland, 31 March - 5 April 2019, pp. 1-5

[12] J. H. Kim, J. H. Han, J. S. Park, J. G. Kim, "Design of phased array antenna for 5G mm-wave beamforming system", Proceedings of the IEEE 5th Asia-Pacific Conference on Antennas and Propagation, Kaohsiung, Taiwan, 26-29 Jul 2016, pp. 201-202.

[13] A. K. Vallappil, B. A. Khawaja, M. K. A. Rahim, M. N. Iqbal, H. T. Chattha, M. F. bin M. Ali, "A Compact Triple-Band UWB Inverted Triangular Antenna with Dual-Notch Band Characteristics Using SSRR Metamaterial Structure for Use in Next-Generation Wireless Systems", Fractal and Fractional, Vol. 6, No. 8, 2022, p. 422.

[14] A. Karimbu Vallappil, M. K. A. Rahim, B. A. Khawaja, M. N. Iqbal, N. A. Murad, M. M. Gajibo, L. O. Nur, B. S. Nugroho, "Complementary split-ring resonator and strip-gap based metamaterial fractal antenna with miniature size and enhanced bandwidth for 5G applications", Journal of Electromagnetic Waves and Applications, Vol. 36, No. 6, 2022, pp. 787-803.

[15] B. A. Khawaja, M. A. Tarar, T. Tauqeer, F. Amir, M. A. Mustaqim, "1 × 2 Triple Band Printed Antenna Array for Use In Next Generation Flying Ad-Hoc Networks (FANETs)", Microwave and Optical Technology Letters, Vol. 58, 2016, pp. 606-610.

[16] S. Zhang, X. Chen, G. F. Pedersen, "Mutual Coupling Suppression With Decoupling Ground for Massive MIMO Antenna Arrays", IEEE Transactions on Vehicular Technology, Vol. 68, No. 8, Aug. 2019, pp. 7273-7282.

[17] J. Cai, J. Zhang, S. Xi, J. Huang, G. Liu, "A Wideband Eight-Element Antenna with High Isolation for 5G New-Radio Applications", Applied Sciences, Vol. 13, No. 1, 2022, p. 137.

[18] W. Wang, Y. Li, Y. Xu, W. Shen, "Compact 8-Port MIMO Antenna System for Dual-Band 5G Mobile Terminals", Proceedings of the International Conference on Microwave and Millimeter Wave Technology, Chengdu, China, 2018, pp. 1-3.

[19] H.-D. Chen, Y.-C. Tsai, C.-Y.-D. Sim, C. Kuo, "Broadband Eight-Antenna Array Design for Sub-6 GHz 5G NR Bands Metal-Frame Smartphone Applications", IEEE Antennas and Wireless Propagation Letters, Vol. 19, No. 7, 2020, pp. 1078-1082.

[20] K. R. Jha, N. Rana, S. K. Sharma, "Design of Compact Antenna Array for MIMO Implementation Using Characteristic Mode Analysis for 5G NR and Wi-Fi 6 Applications", IEEE Open Journal of Antennas and Propagation, Vol. 4, 2023, pp. 262-277.

[21] Y. M. Muhsin, A. J. Salim, J. K. Ali; "An eight-element multi-band MIMO antenna system for 5G mobile terminals", AIP Conference Proceedings, Vol. 2651, 2023, p. 060005.

[22] A. K. Vallappil, M. K. A. Rahim, B. A. Khawaja, N. A. Murad, M. G. Mustapha, "Butler Matrix Based Beamforming Networks for Phased Array Antenna Systems: A Comprehensive Review and Future Directions for 5G Applications", IEEE Access, Vol. 9, 2020, pp. 3970-3987.

A Comparative Experimental Investigation of MPPT Controls for Variable Speed Wind Turbines

Original Scientific Paper

[1] **Dahbi Abdeldjalil**
1. Unité de Recherche en Energies Renouvelables en Milieu Saharien(URERMS),
Centre de Développement des Energies renouvelables(CDER), 01000, Adrar , Algeria ;
2. Laboratory of sustainable Development and computing, (L.D.D.I), University of Adrar,
01000, Adrar, Algeria;
Emails:Dahbi_j@yahoo.fr

[2] **Benlahbib Boualam**
Unité de Recherche Appliquée en Energies Renouvelables, URAER,
Centre de Développement des Energies Renouvelables, CDER, 47133, Ghardaia, Algeria;
bouallam30@gmail.com

[3] **Benmedjahed Miloud**
Unité de Recherche en Energies Renouvelables en Milieu Saharien(URERMS), (CDER),
01000, Adrar , Algeria ;
benmedjahed_78@yahoo.fr

[4] **Khelfaoui Abderrahmane**
Unité de Recherche en Energies Renouvelables en Milieu Saharien(URERMS), (CDER),
01000, Adrar , Algeria ;
dihe.khelfaoui94@gmail.com

[5] **Bouraiou Ahmed**
1. Unité de Recherche en Energies Renouvelables en Milieu Saharien(URERMS), (CDER),
01000, Adrar , Algeria ;
2. Laboratory of sustainable Development and computing, (L.D.D.I), University of Adrar,
01000, Adrar, Algeria;
bouraiouahmed@gmail.com

[6] **Aoun Nouar**
Unité de Recherche en Energies Renouvelables en Milieu Saharien(URERMS), (CDER),
01000, Adrar , Algeria ;
nouar.aoun@gmail.com

[7] **Mekhilefd Saad**
Power Electronics and Renewable Energy Research Laboratory (PEARL) Department of Electrical of Engineering University of Malaya,
Kuala Lumpur, 50603, Malaysia;
saad@um.edu.my

[8] **Reama Abdellatif**
Department of Engineering System, Paris Est University, ESIEE, Paris, France;
abdel.reama@gmail.com

Abstract *– This work presents an experimental comparative investigation between Maximum power point tracking control methods used in variable speed wind turbines. In order to enhance the efficiency of the wind turbine system, the maximum power point tracking control has been applied for extracting and exploiting the maximum available wind power. Furthermore, two maximum power point tracking controls have been analyzed, developed, and investigated in real-time using Dspace. The first was optimal torque control without speed control, whereas the second was with speed control. The maximum power point tracking control performance comparison has been performed in a real-time experimental validation to illustrate the advantages of these control on the real wind energy system. The results have been achieved and discussed, where the power efficiency improvements appeared in the transit time and in the steady-state as well. In addition, the proposed optimal torque control for maximum power point tracking with speed control decreased the response time and oscillations, while it increased the power to an interval of 12,5% to 75% compared to that of strategy without speed control in the steady-state and transit state, respectively.*

Keywords: *Wind energy conversion system (WECS), Maximum power point tracking (MPPT), Optimal Torque Control (OTC), Real time control, Dspace*

1. INTRODUCTION

Nowadays, modern life and technology development has greatly increased the electrical energy demand. This is mainly due to the strong relationship between life quality and the energy consumption level. This leads to an ever-increasing of fossil sources demand. However, these later know a gradual depletion and cause environmental and pollution problems [1, 2]. Currently, there are serious large movements toward free solution energies

that are more sustainable, renewable and environment-friendly [3-7]. The wind energy takes an important class among these renewable energies [8-11]. It knows a high growth and use over the world in recent years [12-17]. A large installation capacity is noticed (733GW: onshore 698GW and 35GW offshore) and with an annual increase of 9% expected between 2021-2030 [18]. However, the intermittence criteria of this energy requests a high technology mastering to answer load needs and to benefit as far as possible from this energy [19]. Recently, the wind energy has been focused on higher efficiency and robustness with a lower cost [4, 20, 21]. Several studies have been performed to increase the efficiency of the Wind Energy Systems. For this reason, variable speed wind turbines were adopted instead of those of fixed speed to apply the Maximum power point tracking control when the available power is less than the rated generator power, and pitch angle when it becomes over the rated power to protect the generator from over power [22-24]. Thus, wind turbines can optimize the produced energy with a better quality [25-32]. This paper is focused in MPPT operation mode. Different control algorithms were applied to extract the maximum power [33-43]. They can be classified according to the control technique or to whether or not the use of the wind speed sensor. Perturb and observe (P&O) method is among the famous used techniques [44, 45]. It is easy to set up and can reach the maximum power points without the need to wind turbine parameters. However, during rapid wind speed variations, it is liable to diverge and lose the maximum power point. Even for slow variations, this command lacks precision because its power always oscillates around the maximum point, which can generate oscillations that generate noise and disturbances of the torque. In addition, this MPPT method has less efficiency and presents the difficulty of optimizing the test step [9]. Other techniques were based on the knowledge of the power curve characteristics as a function of the rotor speed. These characteristics are obtained from several tests either to store them in a 'Lookup table' or approach them with a mathematical function. However, this method loses its precision in the case of variations in turbine parameters due to climatic conditions changing on blades. Besides, the turbine characteristics are often not available (from the factory) to avoid an additional cost. For these reasons, other MPPT control strategies will be presented in this paper. MPPT based on the control of the electromagnetic torque (Optimal Torque Control (OTC) without speed control), and Optimal Torque Control with speed control which is based on the adjustment of the mechanical speed in a way to maximize the power coefficient; therefore maximizing the converted electromagnetic power; [24, 25, 39, 46-48]. Although the previous works were achieved in simulations [36,49], in this work, both later proposed MPPT controls have been analyzed, investigated, and implemented in a real-time wind turbine simulator. The experimental results will be discussed and compared to choose the best method.

Below are the main contributions of this paper:

Firstly, comparison between both proposed MPPT approaches using several tests scenarios that prove the superiority of the OTC with speed control. Secondly, confirming the comparison using real time tests and validations. Thirdly, improvement of the WECS yield and performances such as response time. Fourthly, reducing of overshoot and torque vibration in the system. The rest of this paper is organized on six sections: Section 2 starts by modelling of the wind turbine characteristics. Then the analysis of the OTC MPPT without Speed Control. Followed by the study of the OTC MPPT with Speed Control. The fifth section is reserved to the real time MPPT results; also, this part discusses results and comparisons between MPPT controls in order to improve the yield of the real wind turbine. Finally, the main findings and the suggestions for future work are outlined in Section 6.

2. WIND TURBINE MODELLING

The total wind power that flows through the blades sweep area can be written by [31, 50-53]:

$$P_m = \frac{1}{2}\rho A v_1^3 \tag{1}$$

With $$A = \pi R_t^2 \tag{2}$$

The produced mechanical power is expressed as [24,54]:

$$P_t = \frac{1}{2} C_p(\lambda) \rho A v_1^3 \tag{3}$$

The tip-speed is defined as the ratio between the linear speed before and after blades, as given [55,56]:

$$\lambda = \frac{\Omega_t R_t}{v_1} \tag{4}$$

The power coefficient value C_p is depending on the design of the wind turbine; The equation of C_p is nonlinear as it is given by [5]:

$$C_p(\lambda,\beta) = (0.5 - 0.0167(\beta - 2))\sin\left[\frac{\pi(\lambda + 0.1)}{18 - 0.3(\beta - 2)}\right] - 0.00184(\lambda - 3)(\beta - 2) \tag{5}$$

Fig. 1 shows the power coefficient as a function of the speed ratio for various pitch angle values:

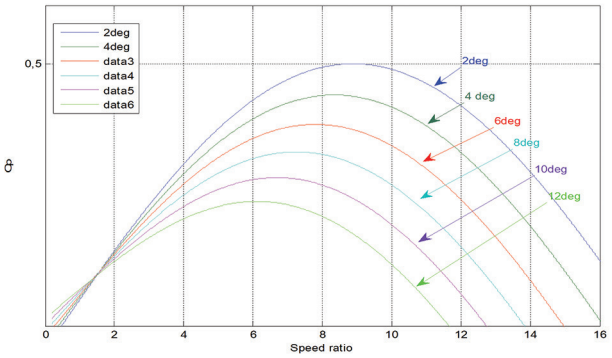


Fig. 1. $C_p(\lambda, \beta)$ Characteristics for various values of β

It can be seen that the value of the power coefficient $C_p(\lambda, \beta)$, and consequently the power are inversely proportional with the pitch angle values. For this reason, the pitch angle control increases its value only to limit the surplus power. On the contrary, when the wind power is less than the nominal power, β value should be maintained fixe at its smallest degree (β=2) to fully exploit the wind energy [2, 27, 57]. Hence, the power coefficient reaches its maximum value at a particular optimal speed ratio λ_{opt}, which is determined from the blade design. This λ_{opt} is corresponding to the maximum power value for each wind speed, Fig.1. So, as a result, to maximize the wind energy, λ should be maintained either at λ_{opt}, or regulate Ω_t at Ω_{ref} which is expressed as following [57]:

$$\Omega_{ref} = \frac{\lambda_{opt}.\ v_1}{R_t} \tag{6}$$

All these show the use interest of the Maximum Power Point Tracking (MPPT) strategy in order to reach and track the maximum power point whatever the variation of wind speed. This explains the existence of various MPPT strategies to maximise the captured wind power. Most of them are based on the optimal torque or the speed controller. In the next part, the developed Optimal Torque Control MPPT controls without and with Speed Control will be detailed and investigated with a comparison of their performances in real-time case [3].

3. OPTIMAL TORQUE MPPT CONTROL WITHOUT SPEED CONTROL

The Optimal Torque MPPT without Speed Control strategy is based on the rotor speed measurement and the subsequent determination of the desired generator power (or torque). The principle of this control is to impose and adjust the generator torque to its reference torque according to different wind speeds [25]. The OTC Method without Speed Control requires the knowledge of the optimal turbine characteristics $C_{p\text{-}Max}$ and λ_{opt}.

The main advantage of this control method is presented in its simplicity, where it avoids the use of regulators. Moreover, it does not need any wind speed measurement, which avoids the additional cost and failures of the anemometer.

This method is based on the assumption that the wind speed and consequently, the rotor turbine speed varies a bit in the steady operation. Hence, the exerted torque on the shaft is considered null in this case, [2]:

$$T_{mec} = J.\frac{d\Omega_{mec}}{dt} = 0 \tag{7}$$

Also, the resistive torque due to viscous frictions is considered null. So:

$$\begin{aligned} T_{mec} &= T_g - T_{em} - T_f = 0 \\ \Rightarrow T_{em} &= T_g = \frac{T_{aer}}{G} \end{aligned} \tag{8}$$

The reference torque leads to apply and obtain the MPPT control is given by, [2, 33, 48]:

$$T_{em-ref} = K_{opt}.\Omega_{mec}^2 \tag{9}$$

Where:

$$K_{opt} = \frac{1}{2}\frac{C_{p\max}}{\lambda_{opt}^3}\rho\pi R_t^5\frac{1}{G^3} \tag{10}$$

The goal of the Optimal Torque Control MPPT Method without Speed Control strategy is to pick up the maximum power from the wind. Hence, the curve connects the peaks of these curves that generate the maximum power point (MPP) for different wind speeds and follows the corresponding peaks of maximum power operation. It means the following of path of the power shown in Fig. 2, given by [43]:

$$P_{opt} = P_{Max} = \frac{1}{2}C_{p\_opt}(\lambda_{opt})\rho A v_1^3 \tag{11}$$

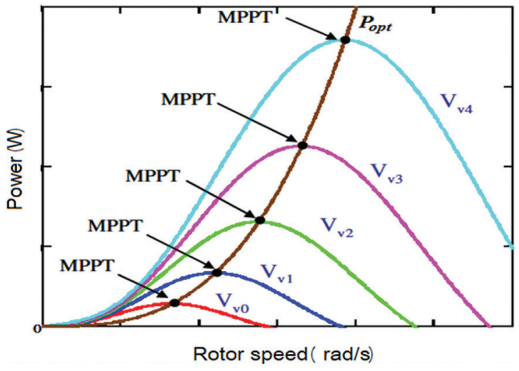


Fig. 2. Wind power characteristics

Although the simplicity of this control, its dynamic is late, especially in the case of fast wind speed variations, because it is an open-loop control. Therefore, the following control overcomes this drawback.

4. OPTIMAL TORQUE MPPT CONTROL WITH SPEED CONTROL

The Optimal Torque MPPT Control method with speed control is used to maximize the captured energy from the wind. It is based on the wind speed measurement in order to calculate the reference rotor speed. The speed controller regulates the rotor speed by controlling the generator torque (and therefore maximizing the electrical power) according to the optimal pre-specified speed λ_{opt}. In this case, the motor torque should be equal to its reference value:

$$T_{em} = T_{em-ref} \tag{12}$$

The reference of the electromagnetic torque T_{em_ref} allows making the mechanical generator speed equal to the reference speed Ω_{ref} by the relation below, [58, 59] :

$$T_{em-ref} = C_{ass}\left(\Omega_{ref} - \Omega_{mec}\right) \tag{13}$$

The reference wind turbine speed that corresponds to the optimal value of the specific speed λ_{opt} and the maximum of power coefficient $C_{p\text{-}Max}$ is given by :

$$\Omega_{tur-ref} = \frac{\lambda_{opt}.v_1}{R_t} \tag{14}$$

By developing the proportional-integral PI controller, the torque becomes:

$$T_{em-ref} = \left(\frac{K_i + K_p S}{S}\right)\left(\Omega_{ref} - \Omega_{mec}\right) \quad (15)$$

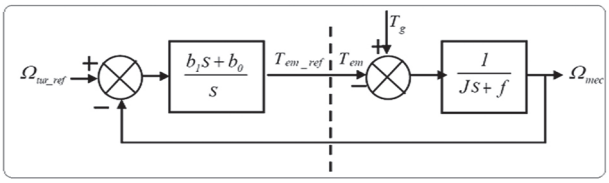


Fig. 3. Diagram block of the speed control

By neglecting losses, the electric output is considered equal to the aerodynamics power.

5. REAL TIME MPPT RESULTS

The tests were carried out in real-time on a test bench where the wind turbine was replaced by an emulator based on an asynchronous machine. This later was controlled in a way to behave such as the real wind turbine when it reacts with the wind speed. The wind turbine emulator was directly coupled with a Permanent Magnet Synchronous Generator (PMSG). The control algorithms have been implemented in a computer equipped with Control Desk software, which receives measurements (current, voltage, rotor speed…) and sends control signals into hardware devices via Dspace card based on DSP micro controller, as shown in Fig. 4 and Table.1 [60-62].

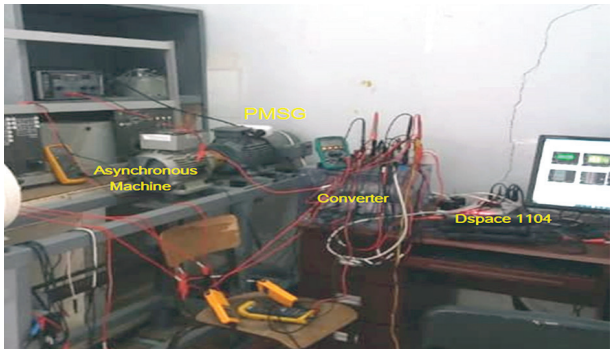
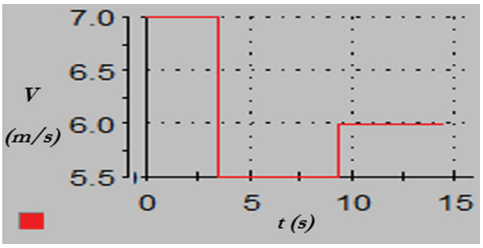


Fig. 4. Wind turbine emulator system in the test bench

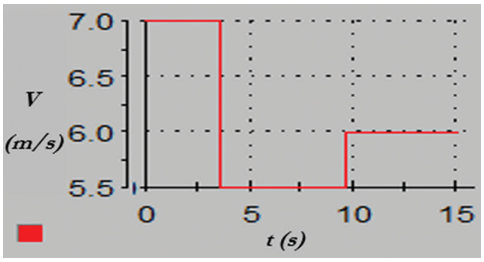
Table 1. Wind energy system parameters

Parameters	Values
Number of blades	3
Blade Radius	0.3m
Air density	ρ =1.225 [kg m-1]
PMSG nominal power	$P = 3kW$
Effective nominal power (line to line)	Vn = 400 V
Number of pole pairs	p = 1
Frequency	fs = 50Hz
Nominal torque	T = 9,5N. m
Nominal speed	3000 tr/min
Asynchronous motor nominal power	$P = 3kW$
Effective nominal power (line to line)	Vn = 400 V
Number of pole pairs	p = 2
Frequency	fs = 50Hz
Power factor	0.85
Nominal speed	1425 tr/min
Converter current RMS	30A
Converter Max DC voltage	750V
Converte Max AC voltage	440V

In order to test and choose the best MPPT strategy performances, the same step shapes of the wind profile in Fig. 5 have been applied for both OTC MPPTs.

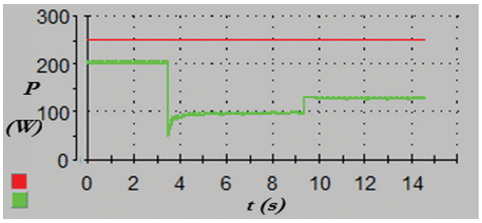


(a) OTC without speed control

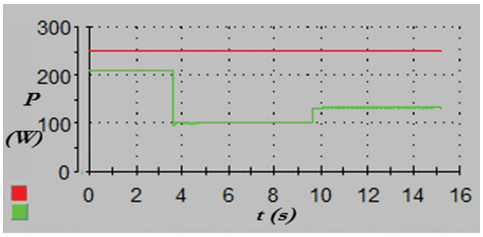


(b) OTC with speed control

Fig. 5. Wind speed profile (m/s)

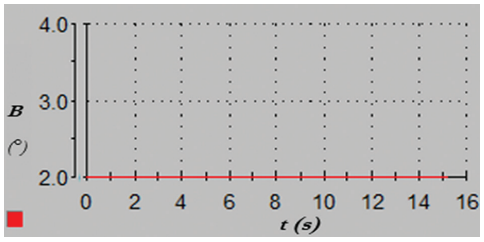


(a) OTC MPPT without with speed control

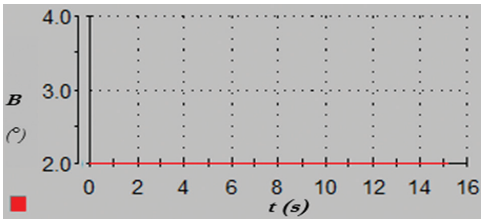


(b) OTC MPPT with speed control

Fig. 6. Wind turbine power (green color) nominal power (red)

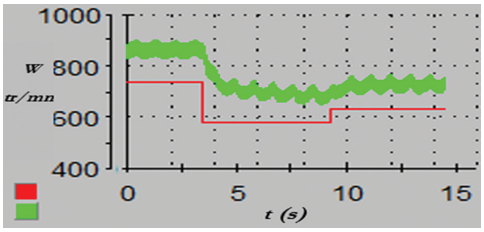


(a) OTC MPPT without speed control

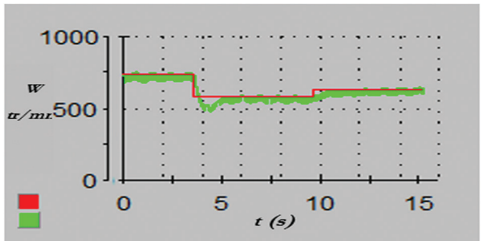


(b) OTC MPPT with speed control

Fig. 7. The Pitch angle

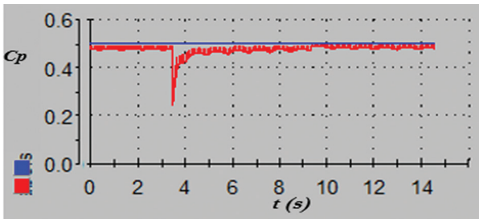


(a) OTC MPPT without speed control

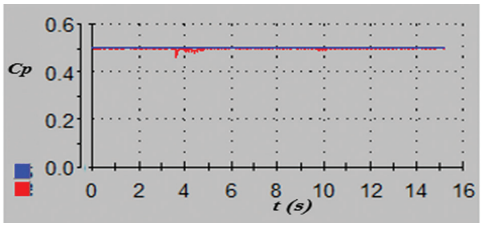


(b) OTC MPPT with speed control.

Fig. 8. Rotor speed (green color) with reference speed (red color)

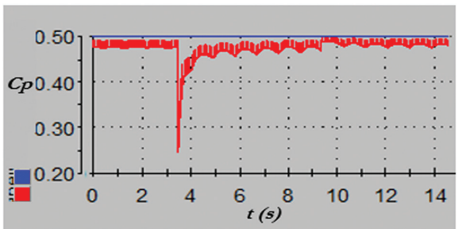


(a) OTC MPPT without with speed control

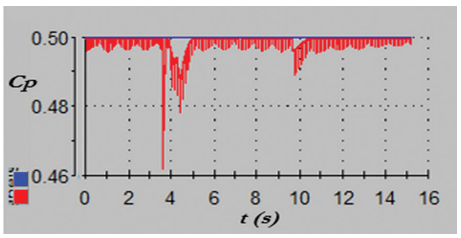


(b) OTC MPPT with speed control

Fig. 9. Power coefficient

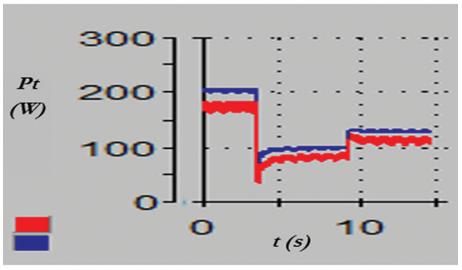


(a) OTC MPPT without with speed control

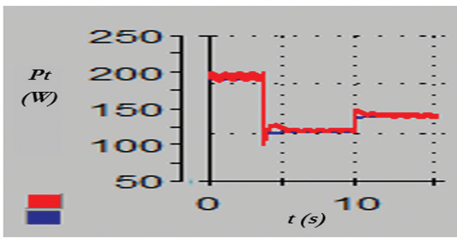


(b) OTC MPPT with speed control

Fig.10. Zoom-in of the power coefficient



(a) OTC MPPT without with speed control



(b) OTC MPPT with speed control

Fig. 11. The wind turbine power: produced (red color), maximal (blue color)

5.1. EXPERIMENTAL RESULT ANALYSIS

In order to confirm the choice of the MPPT control, the wind turbine was put under the same conditions in both cases of MPPT controls. The imposed values of the wind profile were steps of 7m/s, 5.5 m/s, and 6m/s, respectively, as is shown in figures (Fig. 5. a and Fig. 5. b).

Note that the wind turbines power is always lower than its nominal value (250W) in both cases, which requires the operation of the wind turbine at MPPT mode to harvest the maximum available power, figures (Fig. 6. a and Fig. 6. b). Therefore, the setting angle must be maintained at its minimum value ($\beta_{min} = 2$), figures (Fig. 7. a and Fig. 7. b).

Fig. 6. a and Fig. 6. b show that the produced power in the case of the speed control is more stable and

more important in transient mode, especially during the wind speed step variations.

In the case of OTC MPPT with speed control, it is noticed that the speed follows very well its reference, figure (Fig. 8. b). Therefore the power coefficient is well adjusted to its maximum value, figures (Fig. 9. b and Fig. 10. b). It can be also seen that the wind speed is variable. However, the power coefficient is virtually constant at its highest value. consequently harvesting the maximum available power as it is shown in figure (Fig. 11.b); this is from one hand. On the other hand, in the case of OTC MPPT without speed control, the rotational speed is not set at its reference value because it is an open-loop control. Moreover, the power coefficient is slightly lower than its maximum value, which leads to obtain a lower power compared to that of OTC MPPT with speed control, as it is clearly seen in the comparison between the figures (Fig. 10.a and Fig. 10.b), (Fig. 11. a and Fig. 11. b). When the wind speed is 7m/s, it is observed that the power is 175W in OTC without speed control; whereas, in OTC with speed control, the power is maximal, being 200W, which represents an improvement of 25W. When the wind speed changes from 7m/s to 5.5 m/s, Cp and the power decrease to 0.25 and 25W respectively in OTC without speed control. However, in the case of OTC with speed control Cp and the power decrease respectively only till 0.46 and 100W, an improvement of 75W. Based on the comparison, it is clear that the OTC MPPT with speed control is better than that without speed control in the transient state and even in the steady-state, which leads to harvest more power gain and justify the cost of the wind speed sensor, which can be recovered. Moreover, the oscillations during speed variations are reduced in the case of OTC MPPT with speed control. These performances can also be improved by using other speed controllers such as Fuzzy logic control [39], [53].

6. CONCLUSIONS

This paper suggested an experimental comparative investigation between two Optimal Torque MPPT Controls. The first was without Speed Control; whereas, the second was with Speed Control. In order to compare between the performances of each one, both controls have been tested under the same conditions and a similar shape of wind profile. The advantages and disadvantages of each strategy have been discussed.

The validity of the both MPPT controls has been verified by using a wind emulator controlled by Dspace in real time. It was found that the OTC without Speed Control is cheaper and simplest. However, the OTC with speed control was better in its dynamic performances, especially in the case of fast wind speed variations, which leads to harvest more energy and cover the cost of the wind speed sensor. In addition, the proposed OTC MPPT with speed controller decreased the steady-state and improved the oscillation, response time and also it increases the power to an interval of 12,5% to 75% compared to the OTC MPPT without speed control. Furthermore, it has been demonstrated that the applied control strategies can boost the effectiveness of the real wind turbine. This MPPT control encourages its implementation in a real wind turbine under a real wind speed.

7. REFERENCES

[1] A. Bouraiou et al. "Status of renewable energy potential and utilization in Algeria", Journal of Cleaner Production, Vol. 246, 2020, p. 119011.

[2] A. Dahbi, A. Reama, M. Hamouda, N. Nait-Said, N.-S. Mohamed-Said, "Control and study of a real wind turbine", Computers and Electrical Engineering, Vol. 80, 2019, p. 106492.

[3] A. Khelfaoui, M. Tamali, H. Adjlout, A. Dahbi, "Design and realization of an air solar heater and thermal measurement", Instrumentation Mesure Metrologie, Vol. 18, No. 6, 2019, pp. 595-602.

[4] M. Zadehbagheri, T. Sutikno, M. J. Kiani, "A new method of virtual direct torque control of doubly fed induction generator for grid connection", International Journal of Electrical and Computer Engineering, Vol. 13, No. 1, 2023, pp. 1201-1214.

[5] M. B. Toriki, M. K. Asyari, A. Musyafa, "Enhanced Performance of PMSG in WECS Using MPPT - Fuzzy Sliding Mode Control", Journal Européen Des Systèmes Automatisés, Vol. 54, No. 1, 2021, pp. 85-96.

[6] A. Harrag, H. Rezk, "Indirect P&O type-2 fuzzy-based adaptive step MPPT for proton exchange membrane fuel cell", Neural Computing and Applications, Vol. 33, No. 15, 2021, pp. 9649-9662.

[7] F. Tati, H. Talhaoui, O. Aissa, A. Dahbi, "Intelligent shading fault detection in a PV system with MPPT control using neural network technique", International Journal of Energy and Environmental Engineering, Vol. 13, 2022, p. 1147-1161.

[8] Y. Yang, L. Yuan, Z. Qin, H. Liu, "Wind farm resonance characteristics analysis based on harmonic impedance measurement method", Journal of Power Electronics, Vol. 20, No. 4, 2020, pp. 980-990.

[9] M. A. Abdullah, A. H. M. Yatim, C. W. Tan, R. Saidur, "A review of maximum power point tracking algorithms for wind energy systems", Renewable and Sustainable Energy Reviews, Vol. 16, No. 5, 2021, pp. 3220-3227.

[10] M. D. Pham, V. T. Hoang, H. H. Lee, "Cost-effective synchronization strategy for distributed generators in islanded microgrids", Journal of Power Electronics, Vol. 21, No. 3, 2021, pp. 583-589.

[11] A. T. Nguyen, D. C. Lee, "Advanced LVRT strategy for SCIG-based wind energy conversion systems using feedback linearization and sliding mode control", Journal of Power Electronics, Vol. 21, No. 8, 2021, pp. 1180-1189.

[12] S. M. Boudia, A. Benmansour, N. Ghellai, M. Benmedjahed, M. Abdellatif, H. Tabet, "Temporal assessment of wind energy resource at four locations in Algerian Sahara", Energy Conversion and Management, Vol. 76, 2013, pp. 654-664.

[13] M. Benmedjahed, S. Mouhadjer, "Evaluation of wind energy cost and site selection for a wind-farm in the south of Algeria", Technologies and Materials for Renewable Energy, Environment and Sustainability AIP, 2016, pp. 1-13.

[14] M. Benmedjahed, N. Ghellai, Z. Bouzid, A. Chiali, "Temporal Assessment of Wind Energy Resource in "Adrar" (South of Algeria); Calculation and Modeling of Wind Turbine Noise", Proceedings of the 2nd International Congress on Energy Efficiency and Energy Related Materials, 2015, pp. 33-42.

[15] M. H. Soulouknga, S. O. Oyedepo, S. Y. Doka, T. C. Kofane, "Evaluation of the cost of producing wind - generated electricity in Chad", International Journal of Energy and Environmental Engineering, Vol. 11, No. 2, 2020, pp. 275-287.

[16] R. V. S. M. M. I. Petra, "Artificially intelligent models for the site - specific performance of wind turbines", International Journal of Energy and Environmental Engineering, Vol. 11, No. 3, pp. 289-297.

[17] A. T. Nguyen, D. C. Lee, "Sensorless vector control of SCIG-based small wind turbine systems using cascaded second-order generalized integrators", Journal of Power Electronics, Vol. 20, No. 3, 2020, pp. 764-773.

[18] Renewable Energy Capacity Statistics 2021, https://irena.org/events (accessed: 2023)

[19] M. Benmedjahed, A. Dahbi, A. Hadidi, S. Mouhadjer, "Temperature and Wind Distribution Effects on Wind Energy Production in Adrar Region (Southern Algeria)", International Journal of Sustainable Development and Planning, Vol. 16, No. 8, 2021, pp. 1473-1477.

[20] Y. Djeriri, "Robust Second Order Sliding Mode Control of Doubly-Fed Induction Generator for Wind Energy Conversion System", Acta Electrotechnica et Informatica, Vol. 20, No. 3, 2020, pp. 30-38.

[21] S. Mouhadjer, A. Necaibia, M. Benmedjahed, "Hybrid photovoltaic-wind system for the electricity production in isolated sites", Proceedings of the International Conference of Computer Science and Renewable Energies, Agadir, Morocco, 22-24 July 2019, pp. 1-6.

[22] S. Moghaddam, M. Bigdeli, M. Moradlou, P. Siano, "Designing of stand - alone hybrid PV / wind / battery system using improved crow search algorithm considering reliability index", International Journal of Energy and Environmental Engineering, Vol. 10, No. 4, 2019, pp. 429-449.

[23] A. A. Chhipa et al. "Adaptive neuro fuzzy inference system-based maximum power tracking controller for variable speed WECS", Energies, Vol. 14, No. 19, 2021.

[24] G. B. A. K. Shivashankar, "Optimal power point tracking of solar and wind energy in a hybrid wind solar energy system", International Journal of Energy and Environmental Engineering, Vol. 13, No. 1, 2022, pp. 77-103.

[25] M. Hamidat, K. Kouzi, "An Improved Fuzzy OTC MPPT of Decoupled Control Brushless Doubly-Fed Induction Generator", Advanced Computational Techniques for Renewable Energy Systems, 2023, pp. 313-321.

[26] S. Mouhadjer, M. Benmedjahed, A. Neçaïbia, "An effective control strategy to maximize power extraction from wind turbines", AIP Conference Proceedings, Vol. 1758, No. 1, 2016.

[27] A. Dahbi, N. Nait-Said, M.-S. Nait-Said, "A novel combined MPPT-pitch angle control for wide range variable speed wind turbine based on neural network", International Journal of Hydrogen Energy, Vol. 41, No. 22, 2016.

[28] K. B. Tawfiq, A. S. Mansour, H. S. Ramadan, M. Becherif, E. E. El-Kholy, "Wind energy conversion

system topologies and converters: Comparative review", Energy Procedia, Vol. 162, 2019, pp. 38-47.

[29] Y. Zhang, L. Zhang, Y. Liu, "Implementation of maximum power point tracking based on variable speed forecasting for wind energy systems", Processes, Vol. 7, No. 3, 2019.

[30] M. Yin, W. Li, C. Y. Chung, L. Zhou, Z. Chen, Y. Zou, "Optimal torque control based on effective tracking range for maximum power point tracking of wind turbines under varying wind conditions", IET Renewable Power Generation, Vol. 11, No. 4, 2017, pp. 501-510.

[31] H. Fathabadi, "Maximum mechanical power extraction from wind turbines using novel proposed high accuracy single-sensor-based maximum power point tracking technique", Energy, Vol. 113, 2016, pp. 1219-1230.

[32] X. Yue, D. Geng, Q. Chen, Y. Zheng, G. Gao, L. Xu, "2-D lookup table based MPPT: Another choice of improving the generating capacity of a wave power system", Renewable Energy, Vol. 179, 2021, p. 625640.

[33] K. Yenduri, P. Sensarma, "Maximum Power Point Tracking of Variable Speed Wind Turbines with Flexible Shaft", IEEE Transactions on Sustainable Energy, Vol. 7, No. 3, 2016, pp. 956-965.

[34] J. G. Malar, C. A. Kumar, "Implementation of MPPT techniques for wind energy conversion system", Vol. 5, No. 3, 2018, pp. 3-6.

[35] R. Rathi, K. S. Sandhu, "Comparative analysis of MPPT algorithms using wind turbines with different dimensions & ratings", Proceedings of the 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, Delhi, India, 4-6 July 2016.

[36] O. Zebraoui, M. Bouzi, "Comparative study of different MPPT methods for wind energy conversion system", IOP Conference Series: Earth and Environmental Science, Vol. 161, No. 1, 2018.

[37] H. H. H. Mousa, A. R. Youssef, E. E. M. Mohamed, "Hybrid and adaptive sectors P&O MPPT algorithm based wind generation system", Renewable Energy, Vol. 145, 2020, pp. 1412-1429.

[38] A. R. Youssef, H. H. H. Mousa, E. E. M. Mohamed, "Development of self-adaptive P&O MPPT algorithm for wind generation systems with concentrated search area", Renewable Energy, Vol. 154, 2020, pp. 875-893.

[39] A. Dida, F. Merahi, S. Mekhilef, "New grid synchronization and power control scheme of doubly-fed induction generator based wind turbine system using fuzzy logic control", Computers and Electrical Engineering, Vol. 84, 2020.

[40] R. Syahputra, I. Soesanti, "Performance improvement for small-scale wind turbine system based on maximum power point tracking control", Energies, Vol. 12, No. 20, 2019.

[41] E. Bekiroglu, "MPPT Control of Grid Connected DFIG at Variable Wind Speed", Energies, Vol. 15, No. 9, 2022, p. 3146.

[42] N. Labed, I. Attoui, S. Makhloufi, A. Bouraiou, M. S. Bouakkaz, "PSO Based Fractional Order PI Controller and ANFIS Algorithm for Wind Turbine System Control and Diagnosis", Journal of Electrical Engineering & Technology, Vol. 18, 2022, pp. 2457-2468.

[43] S. H. Moon, B. G. Park, J. W. Kim, J. M. Kim, "Effective algorithms of a power converter for tidal current power generation system", Journal of Power Electronics, Vol. 20, No. 3, 2020, pp. 823-833.

[44] M. Karabacak, "A new perturb and observe based higher order sliding mode MPPT control of wind turbines eliminating the rotor inertial effect", Renewable Energy, Vol. 133, 2019, pp. 807-827.

[45] H. Ramadan, A. R. Youssef, H. H. H. Mousa, E. E. M. Mohamed, "An efficient variable-step P&O maximum power point tracking technique for grid-connected wind energy conversion system", SN Applied Sciences, Vol. 1, No. 12, 2019, pp. 1-15.

[46] S. El Aimani, "Modeling and control structures for variable speed wind turbine", IEEE Proceedings, Vol. 210, 2010.

[47] P. M. Habestari, A. Mehrizi-Sani, "Frequency response improvement of PMSG wind turbines using a washout filter", Energies, Vol. 13, No. 18, 2020, pp. 1-7.

[48] A. Sachan, A. K. Gupta, P. Samuel, "A review of MPPT algorithms employed in wind energy conversion systems", Journal of Green Engineering, Vol. 6, No. 4, 2017, pp. 385-402.

[49] A. Dali, S. Abdelmalek, A. Bakdi, M. Bettayeb, "A new robust control scheme: Application for MPP tracking of a PMSG-based variable-speed wind turbine", Renewable Energy, Vol. 172, 2021, pp. 1021-1034.

[50] S. M. Boudia, A. Benmansour, N. Ghellai, M. Benmedjahed, M. A. T. Hellal, "Monthly and Seasonal Assessment of Wind Energy Potential in Mechria Region, Occidental Highlands of Algeria", International Journal of Green Energy, Vol. 9, No. 3, 2012, pp. 243-255.

[51] A. Honarbari, S. Najafi-Shad, M. Saffari Pour, S. S. M. Ajarostaghi, A. Hassannia, "Mppt improvement for pmsg-based wind turbines using extended kalman filter and fuzzy control system", Energies, Vol. 14, No. 22, 2021.

[52] H. Elaimani, N. ELmouhi, A. Essadki, R. Chakib, "Comparative study of power smoothing techniques produced by a wind energy conversion system", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 1, 2022, pp. 77-86.

[53] A. Dahbi et al. "MPPT Fuzzy Logic Control for a Variable Speed Wind Turbine", Journal of Renewable Energy, Vol. 1, No. 1, 2022, pp. 73-80.

[54] A. Al-Quraan, M. Al-Qaisi, "Modelling, design and control of a standalone hybrid PV-wind micro-grid system", Energies, Vol. 14, No. 16, 2021.

[55] J. G. González-Hernández, R. Salas-Cabrera, R. Vázquez-Bautista, L. M. Ong-de-la-Cruz, J. Rodríguez-Guillén, "A novel MPPT PI discrete reverse-acting controller for a wind energy conversion system", Renewable Energy, Vol. 178, 2021, pp. 904-915.

[56] A. Aissaoui, H. Khouidmi, A. Benzouaoui, B. Bessedik, "Nonlinear Predictive Control Method for Maximizing Wind Energy Extraction of Variable Speed Wind Turbines under Turbulence", Journal Européen Des Systèmes Automatisés, Vol. 54, No. 5, 2021, pp. 661-670.

[57] M. L. Frikh, F. Soltani, N. Bensiali, N. Boutasseta, N. Fergani, "Fractional order PID controller design for wind turbine systems using analytical and computational tuning approaches", Computers and Electrical Engineering, Vol. 95, 2021, p. 107410.

[58] R. K. Behara, K. Behara, "Simulation Analysis of DFIG Integrated Wind Turbine Control System", Wind Turbines - Advances and Challenges in Design, Manufacture and Operation, IntechOpen, 2022.

[59] N. Elmouhi, A. Essadki, H. Elaimani, "Improved control for DFIG based wind power system under voltage dips using ADRC optimized by genetic algorithms", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 5, 2022, pp. 357-367.

[60] A. Dahbi, M. Hachemi, N. Nait-Said, M.-S. Nait-Said, "Realization and control of a wind turbine connected to the grid by using PMSG", Energy Conversion and Management, Vol. 84, 2014.

[61] B. Benlahbib et al. "Experimental investigation of power management and control of a PV/wind/fuel cell/battery hybrid energy system microgrid", International Journal of Hydrogen Energy, Vol. 45, No. 53, 2020, pp. 29110-29122.

[62] H. Chojaa et al. "Enhancement of Direct Power Control by Using Artificial Neural Network for a Doubly Fed Induction Generator-Based WECS: An Experimental Validation", Electronics, Vol. 11, No. 24, 2022.

A Feasible MPPT Algorithm for the DC/DC Boost Converter: An Applied Case for Stand-Alone Solar Photovoltaic Systems

Original Scientific Paper

Pham Hong Thanh
Thu Dau Mot University,
Electrical Engineering Program, Institute of Engineering and Technology
Thu Dau Mot City, Binh Duong Province, Vietnam
thanhph@tdmu.edu.vn

Le Van Dai*
Industrial University of Ho Chi Minh City,
Electric Power System Research Group, Faculty of Electrical Engineering Technology
Ho Chi Minh City, Vietnam
levandai@iuh.edu.vn
*Corresponded Author

Abstract *– One of the most promising forms of renewable energy is solar energy. However, efficient exploitation of this energy form is a topic of great interest, especially in obtaining the maximum amount of power from the solar photovoltaic (PV) system under changing environmental conditions. To solve this problem, it is necessary to propose an optimal algorithm. Therefore, this paper presents a feasible maximum power point tracking (MPPT) technique for DC/DC boost converters applied in load-connected stand-alone PV systems to extract the maximum available power. This proposed method is based on the combination of the modified perturb and observe (P&O) and fractional open circuit voltage (FOCV) algorithms. The effectiveness of the proposed technique is verified via time-domain simulation of the load-connected stand-alone PV system using PSIM software. The simulation results show a tracking efficiency with an average value of 99.85%, 99.87%, and 99.96% for tracking the MPP under varying loads, irradiation, and simultaneously varying temperature, load, and irradiation, respectively. In addition, tracking time is always stable at 0.02 sec for changing weather conditions in the large range. Therefore, the results of the proposed method indicate advantages compared to the conventional method.*

Keywords: *Maximum power point tracking (MPPT), perturb and observe (P&O), DC/DC boost converter, photovoltaic solar*

1. INTRODUCTION

According to the global market outlook, global installed PV capacity will increase to 940 GW by the end of 2021, a 22% increase from 772.2 GW in 2020, as shown in Fig. 1 [1]. Such a rapid increase is due to the ease of installation of this type of energy. For example, they can be installed in places without other uses, such as rooftops, deserts, or remote locations. Enhance, the development of PV energy has become a suitable research topic in the last decade. However, its power generation efficiency depends on the characteristics of the PV module, which vary with solar radiation level and atmospheric temperature [2]. To maximize energy from solar absorption at different radiation levels, the PV model must be driven at its maximum power point (MPP). In the past decade, a large number of MPP methods have been developed to increase the efficiency of the PV module.

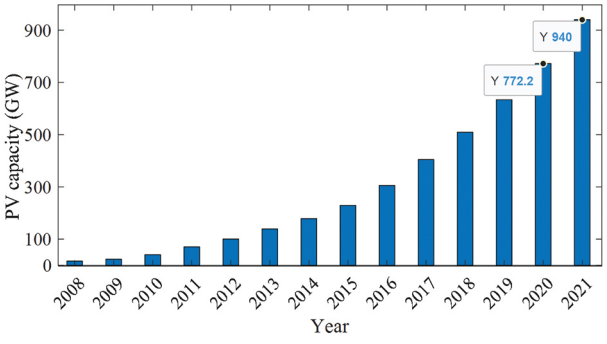


Fig. 1. Global growth in installed PV capacity 2008-2021

Currently, the PV power generation system can be divided into two types that are grid-connected and stand-alone PV systems. This paper focuses on the second type, which has been widely installed worldwide due to its low cost and high convenience in installa-

tion and use, especially for use in hard-to-connect or unconnected areas of the power grid. The stand-alone PV system refers to generating the electric energy that supplies the electrical load of the DC and AC types. This paper focuses on stand-alone PV, which is used to supply electric energy to a DC load. The architecture of the isolated-DC grid-connected PV system is proposed in this study, as shown in Fig. 2. This system can be divided into its key parts, which are the PV array, DC/DC boost converter, control unit, and load.

In this paper, the stand-alone photovoltaic solar system is considered to apply for any of the following: heating, cooking, and water pumping. For example, the authors in Ref. [3] have used this system to drive the pummel system using a brushless direct current motor. The battery is not used in this investigated system to save money and protect the environment. However, it has a significant problem in that the amount of the generated electric power depends on the weather conditions, especially solar irradiance. The efficiency of converting solar energy into electrical energy from PV panels is very low, usually in the range of 12% to 30%, due to the variations in irradiation, temperature, and load [4]. To enhance the conversion efficiency, the PV array should be tracked at the MPP. To achieve this goal, the MPP tracking (MPPT) algorithm for DC/DC converters is required. Basically, the MPPT algorithm is a power control method that adjusts the duty cycle of the DC/DC converter based on the output and input of the PV array to capture maximum power production continuously, thus achieving maximum power and supplying voltage stability under varying weather conditions.

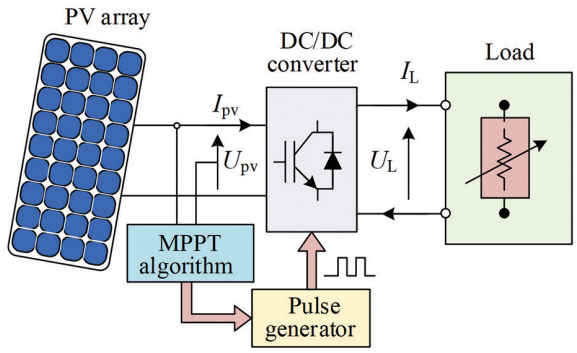


Fig. 2. The load-connected stand-alone PV system with the assistance of the MPPT algorithm

The MPPT algorithms are developed based on criteria including cost, efficiency, loss of energy, tracking time, level of oscillation, scientific tracking MPP, and type of power electric converter [5]. Considering these accounts, it has two MPPT algorithms. The first is conventional methods, which are simple and low-cost but lead to poor performance. The second has been developed using intelligent methods, which have high performance and are complex [6].

Over the last few years, there have been many proposed methods to achieve the MPPT under variable weather conditions. The most significant are methods such as perturb and observe (P&O) [7], incremental conductance (InC) [8], fractional open circuit voltage (FOCV) [9], and fractional short circuit current (FSCC) [10]. The use of these methods is effective; however, it has the problem of slow convergence and significant oscillations around the MPP. In addition, the PV array having the PV characteristic is not linear; thus, it needs to apply the MPPT control methods based on intelligent methods, including neural network (NN), fuzzy logic control (FLC), and the metaheuristic method. Authors in Ref. [11] have proposed the MPPT method based on the NN-global sliding mode for DC/DC buck-boost converter. The ANN-FL was developed by authors in Ref. [12]. The combination of the InC and FLC has been proposed by authors in [13]. These combination methods have several benefits, including being able to handle variable inputs, avoiding the requirement for precise mathematical modeling, and having self-convergence and self-learning capabilities [14]. The drawback of these methods is that the tracking performance and output efficiency are dependent on the engineer's technical knowledge. To overcome this problem, Manna, S et al. [15-18] introduced new methods based on the model reference adaptive control (MRAC) to enhance the tracking efficiency and speed of PV system under changes in ambient conditions. With these algorithms, it gives a reliable tracking efficiency and time compared to the traditional P&O, INC, FLC and ANF.

From an algorithmic point of view, even though the P&O algorithm has many benefits, a rapid change in atmospheric circumstances leads this P&O algorithm to drift away from MPP [19], and authors in Ref. [20] have provided analyses of this drift issue. In this study, the drift is clearly analyzed in terms of its potential occurrence, the movement of the operating point, and the effects of both abrupt changes in insolation and more gradual changes. As a solution to the drift issue, the authors in Ref. [21] have applied the constraint on perturbation step size (ΔD). However, the value of ΔD is high, resulting in an increase in steady-state power loss [22]. The adjustable variable step based on the Pythagorean theorem to calculate the reference voltage through the optimal value of ΔD is proposed by authors [23, 24]. However, it is manually adjusted to regulate this ΔD.

According to the literature survey, most are not interested in the self-adjusting optimal value ΔD under simultaneously varying temperature, load, and irradiation conditions. To solve this problem, this study proposes a solution that is based on the combination of the modified P&O and the FOCV algorithms. In this proposed algorithm, automatic tuning of the step size results in quick and precise tracking. Large perturbation values are better for improving dynamic performance, whereas lower values are better for improving steady-state performance. In addition, the proposed method also considers the drift problem early by setting upper and lower threshold limits for changes in power based on the slow and fast changes in the input of solar irradiance. The key contributions of this work are summarized as follows:

(i) The modeling of the load-connected stand-alone PV system designed using the PSIM environment and C++ code to assess the functionality of a PV module. This system, which consists of a PV array, a DC/DC boost converter, and an MPPT controller, can be used for any of the following: heating, cooking, and water pumping;

(ii) Establish a MPPT method based on the combination of the modified P&O and the FOCV algorithms to overcome the main drawbacks of the conventional P&O-MPPT.

(iii) The stability of the proposed method is confirmed under simultaneously changing radiation, temperature, and load.

Except for the introduction, this paper consists of four sections and is organized as follows: Section 2 covers the modeling, structure, and DC/DC boost converter. The principle of operation and the schematic diagrams of the DC/DC boost converter are presented in this section. Section 3 presents the control scheme for the DC/DC boost converter of the load-connected stand-alone PV system, followed by a recall of the conventional P&O-MPPT algorithm and a proposal for the modified P&O-MPPT algorithm. The effectiveness of the control method for the DC/DC boost converter based on the conventional and modified MPPT algorithms applied in stand-alone photovoltaic solar systems through other studied cases is verified, analyzed, discussed, and compared in Section 4. Finally, Section 5 contains the conclusions of this study, and the proposed directions for future research are presented in this section.

2. PROBLEM FORMULATION

2.1. PHOTOVOLTAIC ARRAY

The proposed single diode mode of the PV module in this study is shown in Fig. 3 and can be modeled by the relation between the output current and voltage as follows [2]:

$$I = I_{pv} - I_0\left(e^{\left(\frac{q}{\varsigma kT}(U+R_sI)\right)}-1\right)-\left(\frac{U+R_sI}{R_p}\right) \quad (1)$$

in which U is the output voltage, I is the output current, q is the electronic charge, ζ is the diode ideality factor, k is the Boltzmann constant, T is the operating temperature, R_s and R_p are respectively the series and parallel intrinsic resistances, I_{pv} is the photocurrent current and can determine by Eq. (2), and I_0 is the saturation current and can determine by Eq. (3)

$$I_{pv} = \left(I_{sc} + k_{sc}(T - T_r)\right) - \frac{\lambda}{1000} \quad (2)$$

$$I_0 = I_{rs}\left(\frac{T}{T_r}\right)^3 e^{q\frac{E_{bg}}{\varsigma k}\left(\frac{1}{T_r}-\frac{1}{T}\right)} \quad (3)$$

where λ is the illumination, E_{bg} is the band gap for silicon, k_{sc} is the short circuit factor, T_r is the reference temperature of the standard test conditions, and I_{rs} is the reverse saturation, which is given by the following equation.

$$I_{rs} = \frac{I_{sc}}{e^{\left(q\frac{U_{OC}}{\varsigma kT}\right)}-1} \quad (4)$$

where I_{sc} and U_{oc} are the short-circuit current and open-circuit voltage, respectively, which are respectively given as follows:

$$I_{sc} = \frac{G}{G_r}I_{sc.r} + k_i(T - T_r) \quad (5)$$

$$U_{oc} = U_{oc.r} + k_u(T_r - T) + \frac{\varsigma kT}{q}\ln\frac{G}{G_r} \quad (6)$$

where G and G_r are the actual solar radiation and the reference irradiance at the standard test conditions, respectively; $I_{sc.r}$ and $U_{oc.r}$ are the reference short circuit current and the reference open circuit voltage at the standard test conditions, respectively; and k_i and k_u are the temperature coefficient of I_{sc} and the temperature coefficient of U_{oc}, respectively.

For practical PV cells, the value of R_p is large leading to great influence when the PV operates in the current source region. Therefore, Eq. (1) can be reduced as follows:

$$I_{rs} = \frac{I_{sc}}{e^{\left(q\frac{U_{OC}}{\varsigma kT}\right)}-1} \quad (7)$$

In the case of a PV array having n_p parallel and n_s series of the PV cells connected together, the current can be described as follows:

$$I = n_pI_{pv} - n_pI_0\left(e^{\left(\frac{q}{\varsigma kTn_s}\left(U-\frac{n_s}{n_p}R_sI\right)\right)}-1\right) \quad (8)$$

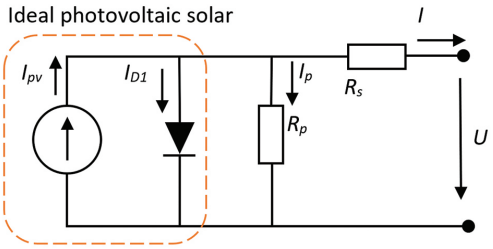


Fig. 3. The equivalent circuit of the PV module

2.2. DC/DC BOOST CONVERTER MODEL

2.2.1. Circuit description

The DC/DC converter used for the PV system includes the buck, boost, buck-boost, and single-ended primary-inductor converters. Based on their advantages, disadvantages, and applications [25, 26], the DC/DC boost converter is discussed and developed in many sectors, such as industrial drives, adaptive control, battery power applications, etc., compared to the other ones. Particularly in the case of the PV application, it not only the output voltage to the desired level but also performs the MPPT control. Therefore, DC/DC boost converter is chosen to study in this paper.

The MPPT-controlled PWM technique for the proposed stand-alone PV systems using the DC/DC boost

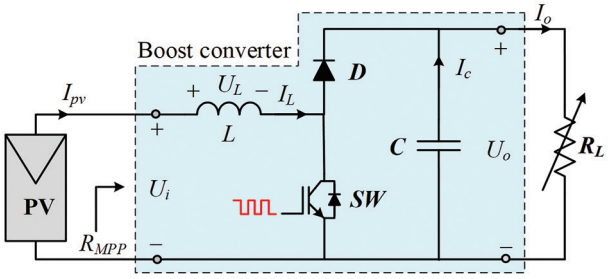
converter that is connected between the PV array and the load as shown in Fig.2. The equivalent circuit diagram is detailed in Fig. 4. The components are used, including the inductor L, power diode D_w, MOSFET *SW*, and capacitor C as shown in Fig. 4 (a).

Based on the time duration of *On* or *Off* for the *SW*, the DC/DC boost converter has two distinct modes of operation, including the continuous conduction operation (CCO) and the discontinuous conduction operation (DCO). For the CCO model, the current through L is always greater than zero, which means that the L partially discharges before the switching cycle begins. For the DCO model, the current through L goes to zero, which means that the L is fully discharged at the end of the switching cycle. Because the dynamic order of the converter is reduced, the DCO model was not selected compared to the CCO model [27]. Therefore, this study uses the CCO model for further study.

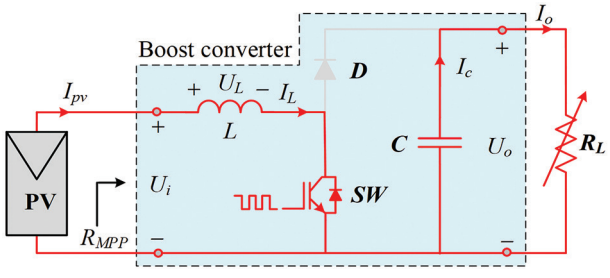
The process of recharging and discharge will constitute a switching cycle, standing for the obtained output voltage is controlled by the time duration of *On* or *Off* of *SW*. The PWM technique is applied to adjust the *On* or *Off* duration. The switching period of *SW* is T_w, the *SW* is closed with time $D_w T_w$ and open with $(1\text{-}D_w)$ T_w, in which D_w is the switching duty cycle. The performance of the boost converter depends on the input inductor and the connected load. The boost converter only operates in the case of $R_L \leq R_{MPP}$. Fig. 5 shows the tracking region of the boost converter on the *U-I* curve of the PV [28]. In order to attain the maximum power of the PV, the D_w must be changed so that the impedance values between the load and source are matched. So, the value of D_w is determined as follows [28, 29]:

$$D_w = 1 - \sqrt{\frac{R_{MPP}}{R_L}} \tag{9}$$

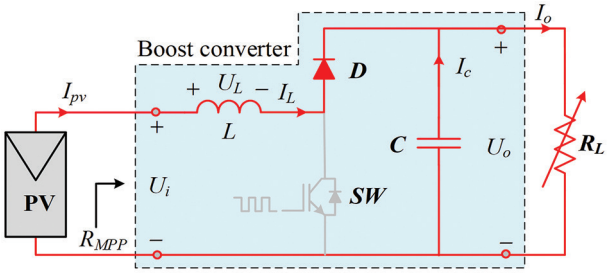
where R_{MPP} is the internal resistance of the PV array and R_L is load resistance.



(a)



(b)



(c)

Fig. 4. The DC/DC boost converter: (**a**) The equivalent circuit representation, (**b**) Equivalent circuit in the case of turned-on switch *SW*, (**c**) Equivalent circuit in the case of turned-off switch *SW*

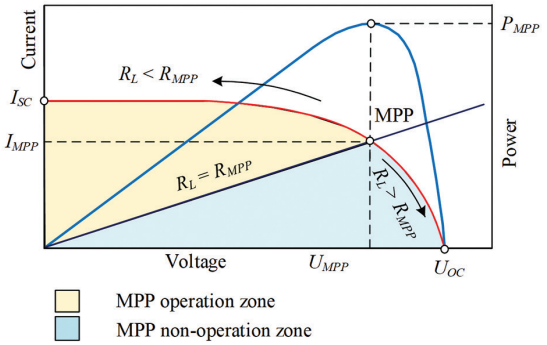


Fig. 5. Tracking region of the boost converter on the *U-I* and *P-U* curves of the PV

2.2.2. Operation analysis

The operation of this boost converter topology depends on the *On* or *Off* state of the switch *SW* and divides into two models.

Model # 1: It begins when the switch *SW* is turned on at time zero; the equivalent circuit is shown in Fig. 4 (b). During this model, the inductor L is connected to the ground, and the output voltage value is $U_o = U_i$. During this state, the inductor charged the energy. The current through the inductor L is raised and calculated by using Eq. (10). The load R_L is supplied the energy by the capacitor C. In this case, the diode current is equal to zero. The main operating waveforms of several components, in this case, are shown in the period $(0, D_w T_w)$ of Fig. 6.

$$I_L = \frac{1}{L}\int_0^{D_w T_w} U_i dt \tag{10}$$

Model # 2: It begins when the switch *SW* is turned off at the time of $D_w T_w$, the equivalent circuit is shown in Fig. 4 (c). During this model, the output voltage in the inductor L is changed and the value is $U_L = (U_o\text{-}U_i)$. During this state, the inductor discharged the energy through the diode to the load. The current through the inductor L is decayed and calculated by using Eq. (11). The main operating waveforms of several components, in this case, are shown in the period $(D_w T_w, T_w)$ of Fig. 6.

$$I_L = -\frac{1}{L}\int_{D_w T_w}^{T_w} (U_i - U_o) dt \tag{11}$$

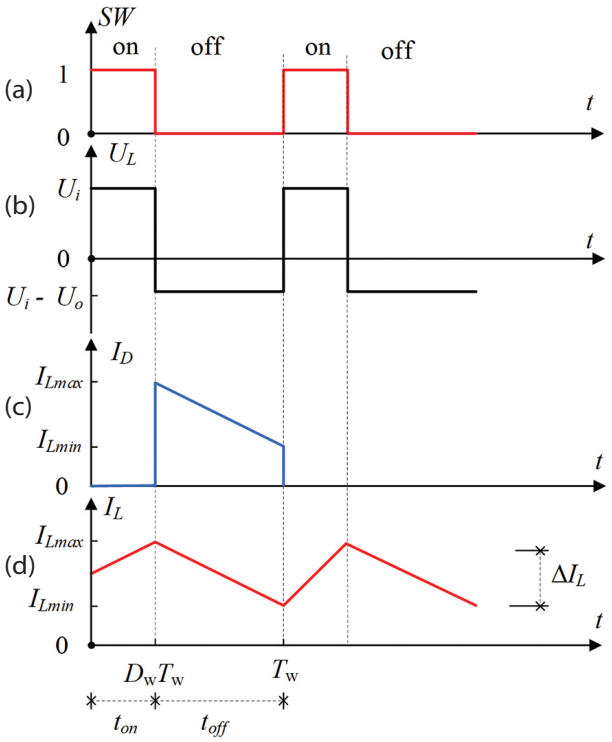


Fig. 6. Boost converter operating waveforms: (**a**) the switch *SW*; (**b**) Inductor voltage; (**c**) Diode current; (**d**) Inductor current

3. MPPT CONTROL METHOD

3.1. DC/DC BOOST CONVERTER

The equivalent circuit of the selected DC/DC boost converter is shown Fig. 4. The value of D_w is set up in the condition between zero to 1 and is considered in the condition without losses. The output voltage is calculated as follows [30]:

$$U_o = \frac{1}{1-D_w} U_i \tag{12}$$

The inductance value is determined by Eq. (13), and this value never falls to zero [31]:

$$L = \frac{(1-D_w)^2 D_w R_L}{2\Delta I_o f_w} \tag{13}$$

in which, ΔI_o is the output current ripple that is selected as 1% of the output current, and the switching frequency of f_w selected is the value of 20 kHz. The capacitance value is calculated as follows [32]:

$$\frac{U_o D_w}{U_o f_w R_L} \tag{14}$$

where ΔU_o is the output voltage ripple that is selected as 1%.

$$R_L = \frac{R_{MPP}}{(D_w - 1)^2} \tag{15}$$

where D_w is the switching duty cycle and can be determined as follows:

$$D_w = \frac{t_{on}}{(t_{on} + t_{off})} \tag{16}$$

3.2. MPPT ALGORITHMS

There are numerous MPPT algorithms for the DC/DC converter system based on solar energy systems that have been put out by numerous researchers with the shared objective of maximizing power output and operating the system at its maximum power point. The P&O is a widely popular technique for obtaining the most power from solar PV due to its ease of use and low cost in comparison to other MPPT techniques. Therefore, this paper considers this method to be improved and uses it as a new method.

3.2.1. Conventional P&O algorithm

The P&O method operates based on observing the PV power through the sensed values of the voltage and current of the PV array. Fig. 7 shows the principle of operation of this method, which depends on the calculation of the output power of the PV array based on the sensed values of the current and voltage. This power is compared to the previous one to address the direction of perturbation and, subsequently, update the switching duty cycle of the DC/DC converter as follows:

$$D_{w(k)} = D_{w(k-1)} \pm \Delta D_w \tag{17}$$

where $D_{w(k)}$ and $D_{w(k-1)}$ are the current and previous perturbations of D_w, respectively; k and (k-1) are the current and previous sampling instants.

In general, the PV array power is calculated based on the sensed values of the voltage and current. The values of voltage and power at k are stored as $P_{(k)} = U_{(k)} I_{(k)}$. Then, the power is calculated by using the previous values at (k - 1). The increment of the voltage and power of the PV array between two consecutive samples is determined as follows:

$$\begin{cases} \Delta U = U_{(k)} - U_{(k-1)} \\ \Delta P = P_{(k)} - P_{(k-1)} \end{cases} \tag{18}$$

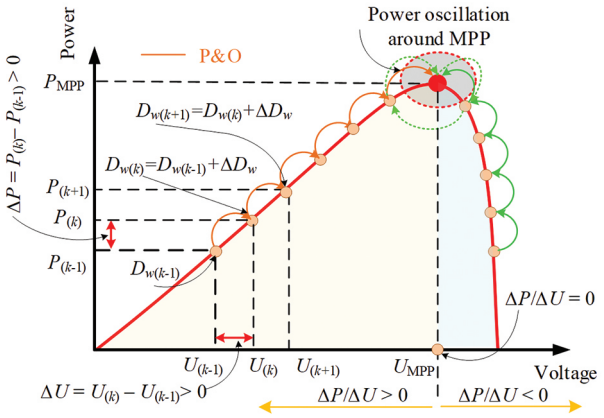


Fig. 7 . The principle of operation of the conventional perturb and observe algorithm

From Fig. 7, there are three conditions based on the fact that the slope of the power curve vs. voltage (current) of the PV array is zero at the MPP and can be described as follows:

i) $\frac{\Delta P}{\Delta U} > 0$: on the left of MPP, the voltage increases power increases;

ii) $\frac{\Delta P}{\Delta U} < 0$: on the right of MPP, power decreases with an increase in the voltage;

iii) $\frac{\Delta P}{\Delta U} = 0$: at MPP.

By comparing ΔP and ΔU, the algorithm decides whether to increase or decrease the duty cycle. If the voltage increases (positive) and the power increases (positive) in two consecutive calculation cycles, then the voltage will be driven to increase (positive) in the next cycle. If the voltage increases (positive), that leads to a decrease in power (negative), and then the voltage is controlled to decrease (negative) in the next cycle, and vice versa.

From Eq. (12), the output voltage is proportional to the D_w, which is determined by Eq. (16) and will be adjusted by increasing or decreasing a value called the "ΔD", and the updated values between two consecutive samples are determined by Eq. (17). This may be done repeatedly until the PMPP is achieved [33]. Table 1 lists the overall P&O direction characteristics, and Fig. 8 depicts its flowchart, which can be found in [34-35].

Table 1. The overall P&O direction characteristics

Voltage perturbation (ΔU)	Change in power perturbation (ΔP)	Direction of perturbation (ΔD_w)
+	+	+
+	-	-
-	+	-
-	-	+

3.2.2. Improved P&O algorithm

The conventional P&O algorithm has two main drawbacks. The first is that the ΔD_w is a fixed value, as shown in Fig. 7. This affects the process of achieving MPP because it depends on this ΔD_w jump. If this value is large enough to reach the MPP quickly, the system will fluctuate widely around the MPP. Conversely, if the offset is small, the system oscillates less around the MPP but takes longer to arrive at the MPP [36]. The second is that it depends on the measured voltage and current values, which depend on the sensors and measurement errors during the system's operation. For the measurement error, the system will measure the values n (usually choose a value from 3 to 7; if this value is too large, it is difficult to respond when environmental conditions change rapidly) times, then perform the comparison according to the P&O algorithm to find the trend in the next operating cycle [37].

To overcome these drawbacks, this study proposes a solution that the principle of operation of the proposed method is shown in Fig. 9 and the algorithm flowchart is shown in Fig. 10. In this proposed algorithm, automatic tuning of the step size results in quick and precise tracking. Large perturbation values are better for improving dynamic performance, whereas lower values are better for improving steady-state performance [37]. The current form of the generic tracking equation is presented in Eq. (17) above, in which the k^{th} optimum value of $\Delta D_{w(k)}$ should be determined as follows:

$$\Delta D_{w(k)} = M_{(k)}(grad\theta_{(k)}) \tag{19}$$

where $M_{(k)}$ is the k^{th} step size that is altered in accordance with the PV system's specifications. For this study, this value is calcaleted as follows:

$$M_{(k)} = \frac{P_{(k)}}{P_{(k-1)}} \tag{20}$$

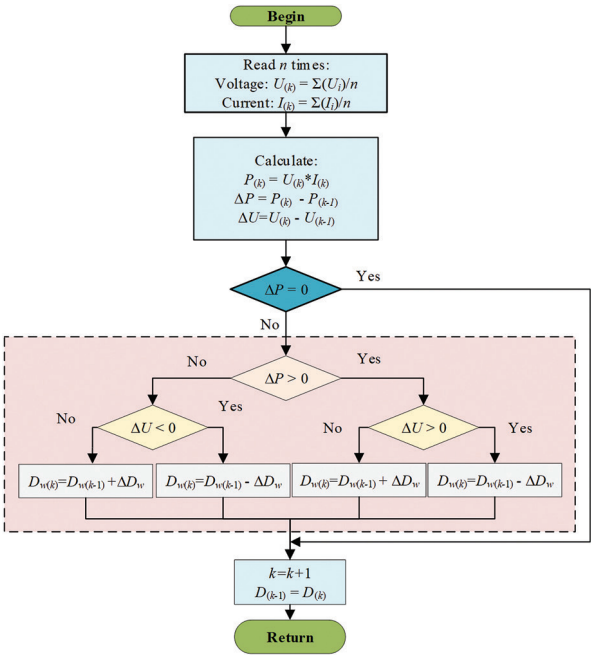


Fig. 8. The flowchart of the conventional algorithm

Corresponding to each working point of k^{th} of PV on the P-U characteristic curve as shown Fig. 9, the grad slope is determined as follows:

$$grad\theta_{(k)} = abs\left(\frac{\Delta P_{(k)}}{\Delta U_{(k)}}\right) \tag{21}$$

where $\Delta P_{(k)}$ and $\Delta U_{(k)}$ are the change in output power and working voltage of the PV module at the k^{th} step.

Substituting Eq. (19) into Eq. (17), and it can obtain as follows:

$$\begin{cases} D_{w(k)} = D_{w(k-1)} + M_{(k)}(grad\theta_{(k)}) & \text{(22a)} \\ D_{w(k)} = D_{w(k-1)} - M_{(k)}(grad\theta_{(k)}) & \text{(22b)} \end{cases}$$

It is clear from Eq. (22) that the modified automation complies with the operating point to provide a fast-tracking capability. As demonstrated in Fig. 9, when the operational point of the PV system is close to the MPP, the shift in the PV power and voltage is less significant than when the operational point is far from the MPP. As a result, the suggested approach boosts the MPPT tracker's speed during abrupt changes in the weather and lowers its oscillation during steady-state situations. Additionally, the suggested approach takes the drift issue into early consideration, Basically, the drift issue occurs when the solar

irradiation on the PV array rapidly increases by at least 10 Ws/m² [17]. The input of solar irradiance is thus dependent on the following two requirements for change in the solar irradiance perturbation ΔG_{STC} [38]

$$\begin{cases} \Delta G_{STC} < 10 \text{ is slow change} \\ \Delta G_{STC} > 10 \text{ is fast change} \end{cases} \quad (23)$$

Considering the irradiance of the PV system operating under standard test condition G_{STC} is 1000 W/m², the new conditions are obtained as follows:

$$\begin{cases} \dfrac{\Delta G_{STC}}{G_{STC}} < 1\% \text{ is slow change} \\ \dfrac{\Delta G_{STC}}{G_{STC}} > 1\% \text{ is fast change} \end{cases} \quad (24)$$

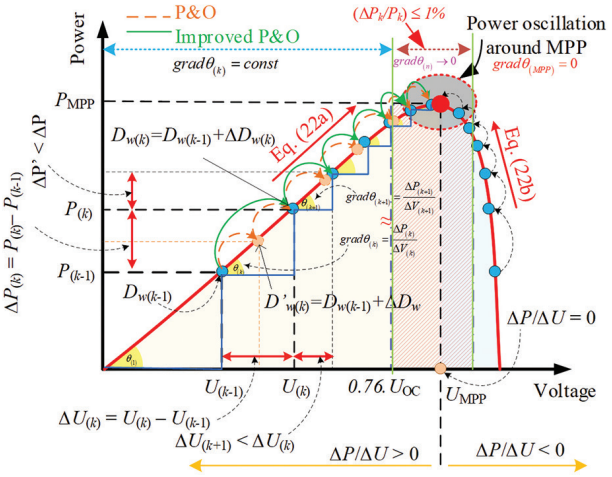


Fig. 9. The principle of operation of the proposed algorithm

The normalized change in solar irradiance is equivalent to the normalized change in power. Therefore Eq. (22) can be represented as [39]:

$$\begin{cases} \dfrac{\Delta P}{P} < 1\% \text{ is slow change} \\ \dfrac{\Delta P}{P} > 1\% \text{ is fast change} \end{cases} \quad (25)$$

where ΔP is the change in power and P represents its previous iteration. As known, if the irradiance varies and alters P's value, ΔP's value likewise changes in the same way. As a result, the value of $\Delta P/P$ remains essentially constant under a variety of environmental circumstances. Additionally, when the operation point is in the drift problem condition, this value is positive; otherwise, it is negative. In order to address the drift issue as soon as possible, a constant value of $\Delta P/P$ is inserted at the beginning of the program, as illustrated in Fig .10. In this paper, the value of $\Delta P/P$ is chosen as 0.01. Under various weather circumstances, the MPP voltage is computed at roughly 78% of the open circuit voltage. In order for the suggested method to determine the side of the operational point when the solar irradiance varies quickly, the U_{set} is applied as 76% of the open circuit voltage [39]. The operation point is to the right of the MPP if the PV voltage is greater than the U_{set}, which causes the D_w reference to decrease. If not, the D_w reference increases, and the $\Delta P/P$ shrinks dramatically when the operation point is near the MPP. As a result, the control unit enters the conventional P&O method to determine the precise optimum MPP.

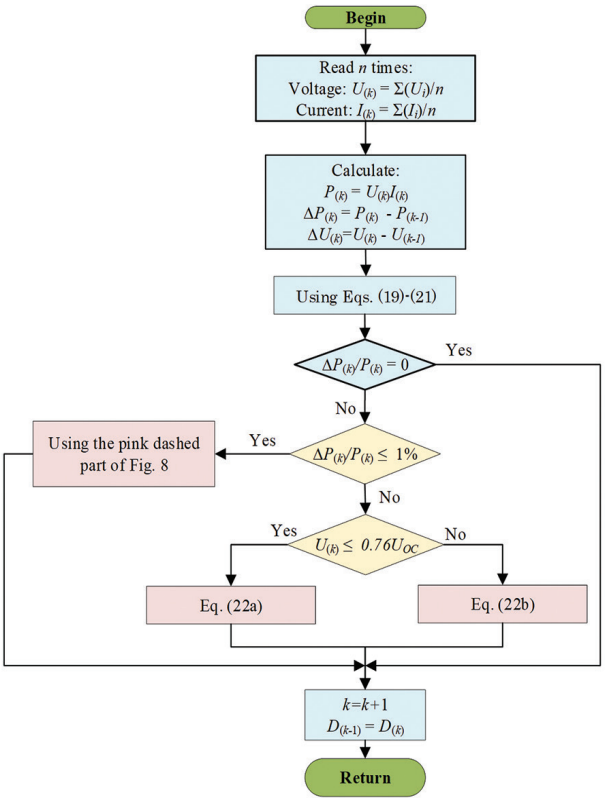


Fig. 10. The flowchart of the improved perturb and observe algorithm

4. RESULTS AND DISCUSSION

To assess the efficacy of the recommended method, a PSIM model for the proposed standalone PV system with MPPT algorithm has been developed, as shown in Fig. 11. A PV array, a DC/DC boost converter with an MPPT controller, and a resistive load are all components of the system under examination. Table 2 summarizes the electrical characteristics of the BP MSX 60 PV panel, which is used as a standard, under normal test conditions. The DC/DC boost converter design parameters are shown in Fig. 11 and were based on Section 3. The resistive load is adjustable and ranges in value from 50 Ω to 200 Ω.

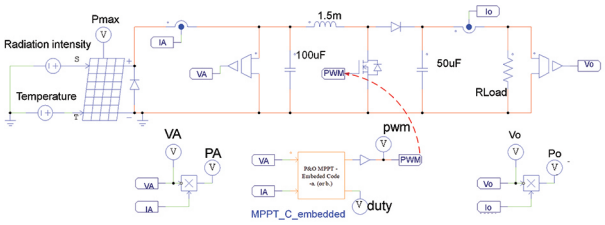


Fig. 11. Simulation model stand-alone photovoltaic system with MPPT algorithm

The *P-U* and *I-U* characteristics of a simulated BP MSX 60 PV panel for the irradiances of 200 W/m² and 1000

W/m² at 25 °C are shown in Fig.12. As a result, the MPP powers have changed from 10.66 W to 59.6 W, and the MPP voltages have changed from 16.88 V to 17.14 V corresponding to the insolation level of 200 W/m² and 1000 W/m², respectively.

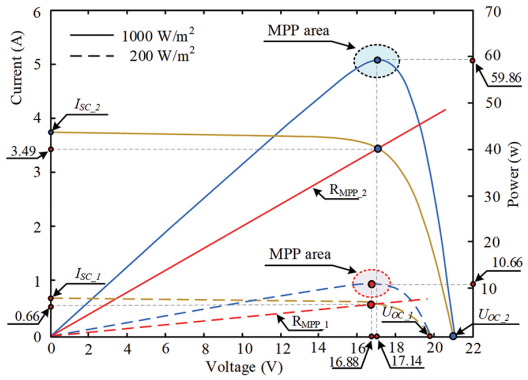


Fig. 12. The *I-U* and *P-U* characteristics for a module on the irradiances of 200 W/m² and 1000 W/m² at 25 °C

Table 2. Electrical characteristics of the PV module

Parameters	Values
The maximum power (Pmax)	60 W
The voltage at Pmax (Umpp)	17.1 V
The current at Pmax (Impp)	3.5 A
The open circuit voltage (Uoc)	21.1 V
The short circuit current (Isc)	3.8 A
The temperature coefficient of Uoc	-(80 ± 10) % V/°C
The temperature coefficient of Isc	-(0.065 ± 0.015) % V/°C
The temperature coefficient of power	-(0.5 ± 0.05) % V/°C
The nominal operating cell temperature	47 ± 2 °C
The operating temperature	25 °C

When the sun irradiation varies, it is not possible to manually adjust the load resistance with the variable value from 50 Ω to 200 Ω. Therefore, the MPPT algorithm and DC/DC boost converter have been designed in Section 3 to continuously adjust the duty cycle of the converter. Two scenarios are considered to verify the perfection of the proposed system.

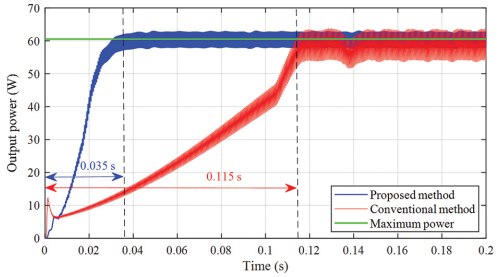
Case 1: Simulation results for the varying load under the fixed irradiation: The tested system has been simulated for two predefined load levels of 100 Ω and 150 Ω. The fixed ambient temperature of 25 °C and the fixed irradiation of 1000 W/m² are considered inputs to the PV panel. Fig. 13 shows the comparison of the output power of the PV panel of both MPPT algorithms. Observing the dynamic response, the performance and efficiency of the proposed method are better in comparison with the conventional P&O algorithm in terms of response time and output power oscillations. The proposed method has a fine response and less fluctuation around the MPP than the conventional method. In the case of R_L = 100 Ω, it takes 0.05 seconds to reach the MPP point when applying the conventional MPPT method, whereas using the proposed MPPT method, it is 0.015 seconds, as shown in Fig. 13 (a).

In addition, a simulation for the predefined varying load is tested to verify the output response for the PV according to the following structure.

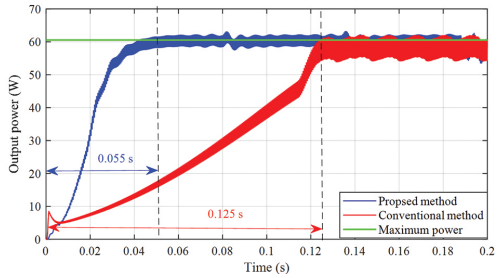
$$R_L(t) = \begin{cases} 100\Omega, & \text{for} \quad t \le 0.2 \text{ sec} \\ 50\Omega, & \text{for} \quad 0.2 \text{ sec} < t \le 0.5 \text{ sec} \\ 150\Omega, & \text{for} \quad 0.5 \text{ sec} < t \le 0.7 \text{ sec} \\ 200\Omega, & \text{for} \quad t > 0.7 \text{ sec} \end{cases} \tag{26}$$

The dynamic response is shown in Fig. 14, and the simulated results are summarized in Table 3. From this table, it can be seen that the proposed MPPT method presents better results than the conventional MPPT method in terms of response time, efficiency, and oscillations to reach the MPP point. The efficiency for tracking MPP is expressed by using Eq. (27) below, in which the maximum power is 60 W. As a result, the average efficiency and tracking time, in this case, are 99.85% and 0.0375 sec, respectively.

$$\eta = \frac{P_o}{P_{\max}} 100\% \tag{27}$$



(a)



(b)

Fig. 13. The output power of a PV panel at 25 °C with the input irradiance of 1000 W/m²: (**a**) the load of 100 Ω, (**b**) the load of 150 Ω

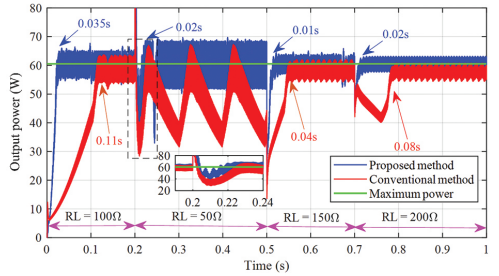


Fig. 14. The output power of a PV panel at 25°C with the input irradiance of 1000 W/m² under considering the predefined varying load

Table 3. A comparison of the properties of both methods for the predefined varying load and the fixed irradiation

Time (s)	R_L (Ω)	P_{max} (W)	Conventional algorithm			Proposed algorithm		
			P_o (W)	η (%)	Tracking time (s)	P_o (W)	η (%)	Tracking time (s)
0-0.2	50	60	49.01	81.68	0.11	59.89	99.82	0.03
0.2-0.5	100	60	58.80	98.00	0.11	59.92	99.87	0.03
0.5-0.7	150	60	58.20	97.00	0.11	59.91	99.85	0.04
0.7-1	200	60	57.64	96.07	0.09	59.92	99.87	0.05
Average	-	-	-	93.19	0.105	-	99.85	0.0375

Case 2: Simulation results for the varying irradiation under the fixed load: In this scenario, the input irradiation varies in a range of 200 W/m^2 to 1000 W/m^2 at the time from 0 to 1 second, the temperature operation is kept at 25°C, and the fixed load is 50 Ω according to the following structure.

$$\text{irradiance}(t)=\begin{cases}200\text{W/m}^2, & \text{for} \quad t\le 0.2\ \text{sec}\\ 1000\text{W/m}^2, & \text{for} \quad 0.2\ \text{sec}<t\le 0.4\ \text{sec}\\ 300\text{W/m}^2, & \text{for} \quad 0.4\ \text{sec}<t\le 0.6\ \text{sec}\\ (300\text{-}1000)\text{W/m}^2, & \text{for} \quad 0.6\ \text{sec}<t\le 0.8\ \text{sec}\\ 1000\text{W/m}^2, & \text{for} \quad t>0.8\ \text{sec}\end{cases} \quad (28)$$

The comparative output power of the two MPPT algorithms is shown in Fig. 15. According to the findings, the suggested method's power tracker addresses the input irradiance's correct direction, whereas the traditional method's tracking power does not when the input irradiation abruptly changes. Notably, the suggested approach's converter duty cycle caused the drift issue to affect the traditional method more than it did. In Fig. 16, this converter duty cycle is shown. Table 4 lists the simulated outcomes for both techniques. The efficiency for tracking MPP is expressed by using Eq. (27), in which the maximum power is defined by predefined power levels.

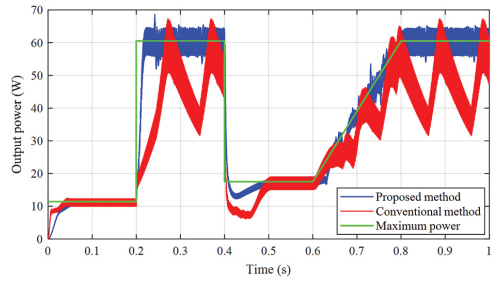


Fig. 15. The output power of a PV panel at 25°C with considering the varying irradiance

According to the results in Table 4, the proposed method's MPPT efficiency under all the different weather condition scenarios achieves an average tracking efficiency of 98.87% for the drift problem under sudden changes in weather conditions (suddenly increasing, suddenly decreasing, or linearly decreasing the input solar irradiation). The suggested strategy lowers the oscillation around the MPP under steady-state circumstances and swiftly follows the MPP during changes in weather, according to the findings of the simulations. In addition, compared to the typical approach, the output PV power is greater.

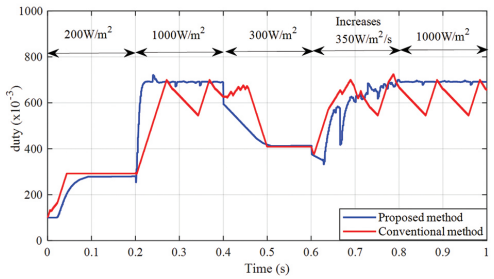


Fig. 16. The converter duty of a PV panel at 25°C with considering the varying irradiance

Case 3: Simulation results under simultaneously varying temperature, load, and irradiation: this scenario is carried out to verify the effectiveness of the proposed method for the PV system under different value range of the temperature, load, and irradiation for a period of 3 sec as shown in Fig. 17. The response of the output power of the PV system for each time when applying two method is shown in Fig. 18. The simulation results are summarized in Table 5.

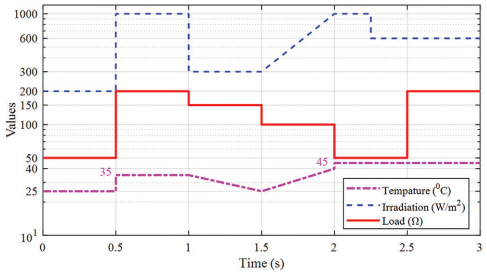


Fig. 17. The simultaneously varying temperature, load, and irradiation

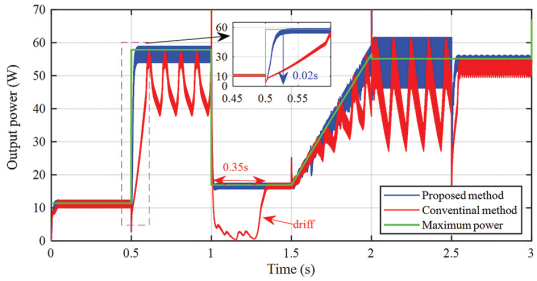


Fig. 18. The conveter duty of a PV panel under simultaneously varying temperature, load, and irradiation

Observing the obtained results shows that under different random changing conditions that affect the

survey system, the proposed method still achieves high tracking efficiency, with the lowest value of 99.71% and the highest up to 99.97%; the average efficiency in this case is 99.85%. The tracking time is always stable at 0.02 sec for condition of the large range of the temperature, load, and irradiation.

Contrary, the time is almost instantaneous (less than 0.005 sec). As a result, the proposed method gets better efficiency than the method of Ref [40]; the average efficiency under varying temperatures is 98.33 %, 96.475%, and 99.825% when applying FL, ANN, and ANN-fuzzy, respectively.

Table 4. A comparison of the properties of both methods for the fixed load and suddenly changed irradiation

Time (s)	*Irradiation* (W/m^2)	P_{max} (W)	Conventional algorithm			Proposed algorithm		
			P_o (W)	η (%)	Tracking time (s)	P_o (W)	η (%)	Tracking time (s)
0 - 0.2	200	11.4	11.35	99.56	0.09	11.38	99.82	0.04
0.2 - 0.4	1000	60	50.00	83.33	0.30	59.93	99.88	0.015
0.4 - 0.6	300	17.49	17.32	99.03	0.11	17.48	99.94	0.07
0.6 - 0.8	300 - 1000	39.02	34.18	87.60	N/A	38.99	99.87	N/A
0.8 - 1.0	1000	60	50.00	83.33	N/A	59.91	99.85	N/A
Average	-	-	-	90.57	-	-	99.87	-

Table 5. A comparison of the properties of both methods under simultaneously varying temperature, load, and irradiation

Time (s)	P_{max} (W)	Conventional algorithm			Proposed algorithm		
		P_o (W)	η (%)	Tracking time (s)	P_o (W)	η (%)	Tracking time (s)
0-0.5	11.40	11.38	99.82	0.04	11.38	99.82	0.02
0.5-1.0	57.85	49.49	85.55	0.07	57.68	99.71	0.02
1.0-1.5	17.2	17.09	99.35	0.05	17.19	99.94	0.0015
1.5-2.0	37.53	32.87	87.59	N/A	37.49	99.89	N/A
2.0-2.3	55.16	47.21	85.58	0.02	55.05	99.80	0.005
2.3-2.5	32.96	29.78	90.35	0.002	32.93	99.91	0.002
2.5-3.0	32.95	29.11	88.35	0.04	32.94	99.97	0.02
Average	-	-	90.94	-	-	99.86	-

5. CONCLUSION

The marketability of photovoltaic solar energy will be heavily influenced by its efficiency, stability, and dependability. This paper has developed the control method for the DC/DC boost converter based on the MPPT algorithm applied in a stand-alone photovoltaic solar system in order to respond well and complete other renewable energy sources from a technical aspect. The main purpose is to improve the tracking efficiency, tracking speed, and oscillation related to changing the temperature, load, and irradiation, which are the main drawbacks of the conventional P&O-MPPT.

The PV panel type of the BP MSX 60 PV is considered when developing the mathematical model. The DC/DC boost converter is designed in accordance with the MPPT algorithm with the objective of maximizing power output and operating the system at its maximum power point. The design of the DC/DC boost converter and the modification of the conventional P&O-MPPT algorithm have been explained in deep detail. For the design of the DC/DC boost converter, the CCO model is applied. The constraint on perturbation step size is selected based on the impedance value between the load and source and is considered in the condition without losses. The inductance and capacitance values are calculated based on the output current, a voltage ripple of 1%, and a switching frequency of f_w set to 20 kHz. For the MPPT algorithm, it uses the modified perturb and observe (P&O) and fractional open circuit voltage (FOCV) algorithms to determine the duty ratio with an adaptive step size. In addition, a combination of current changes as well as power and voltage changes is considered in the decision-making process to avoid the drift problem early.

The simulation validation of the proposed and conventional P&O algorithms was presented and compared using study cases of varying loads under fixed irradiation and changing irradiation under a fixed load. The simulation results show that the proposed MPPT technique achieves efficiency with an average value of 99.85%, 99.87%, and 99.96% for tracking the MPP under varying loads, irradiation, and simultaneously varying temperature, load, and irradiation, respectively. The suggested strategy lowers the oscillation around the MPP under steady-state circumstances. It swiftly tracks the MPP during weather changes, according to the findings of the simulations. In addition, the output PV power is more significant compared to the conventional approach.

6. ACKNOWLEDGMENTS

This research is funded by Thu Dau Mot University, Binh Duong Province, Vietnam under grant number DT.21.1-068. The authors also appreciate the comments of reviewers and the support of Electric Power System Research Group, Industrial University of Ho Chi Minh City, Vietnam.

7. REFERENCES:

[1] S. Michael, "Global Market Outlook for Solar Power 2020-2026", https://www.solarpowereurope.org/insights/market-outlooks/global-market-outlook-for-solar-power-2022#downloadForm (accessed: 2022)

[2] M. G. Villalva, J. R. Gazoli, E. Ruppert Filho, "Comprehensive approach to modeling and simulation of photovoltaic arrays", IEEE Transactions on Power Electronics, Vol. 24, No. 5, 2009, pp. 1198-1208.

[3] B. Abdellahi, M. E. M. Mohamed Mahmoud, N. O. Dah, A. Diakité, A. El Hassen, C. Ehssein, "Monitoring the performances of a maximum power point tracking photovoltaic (MPPT PV) pumping system driven by a brushless direct current (BLDC) motor", International Journal of Renewable Energy Development, Vol. 8, No. 2, 2019, pp. 193-201.

[4] D. Sera, R. Teodorescu, P. Rodriguez, "PV panel model based on datasheet values", Proceedings of the IEEE International Symposium on Industrial Electronics, Vigo, Spain, 2007, pp. 2392-2396.

[5] N. Karami, N. Moubayed, R. Outbib, "General review and classification of different MPPT Techniques", Renewable and Sustainable Energy Reviews, Vol. 68, 2017, pp. 1-18.

[6] R. Alik, A. Jusoh, "An enhanced P&O checking algorithm MPPT for high tracking efficiency of partially shaded PV module", Solar Energy, Vol. 163, 2018, pp. 570-580.

[7] K. H. Hussein, I. Muta, T., Hoshino, M. Osakada, "Maximum photovoltaic power tracking: an algorithm for rapidly changing atmospheric conditions", IEE Proceedings-Generation, Transmission and Distribution, Vol. 142, No. 1, 1995, pp. 59-64.

[8] B. Liu, S. Duan, F. Liu, P. Xu, "Analysis and improvement of maximum power point tracking algorithm based on incremental conductance method for photovoltaic array", Proceedings of the IEEE 7th International Conference on Power Electronics and Drive Systems, Bangkok, Thailand, 27-30 November 2007, pp. 637-641.

[9] J. Ahmad, "A fractional open circuit voltage based maximum power point tracker for photovoltaic arrays", Proceedings of the IEEE 2nd International Conference on Software Technology and Engineering, San Juan, PR, USA, 3-5 October 2010, V1-247-V1-250.

[10] M. A. Masoum, H. Dehbonei, E. F. Fuchs, "Theoretical and experimental analyses of photovoltaic systems with voltage and current-based maximum power-point tracking", IEEE Transactions on Energy Conversion, Vol. 17, No. 4, 2002 pp. 514-522.

[11] I. U. Haq, Q. Khan, S. Ullah, S. A. Khan, R. Akmeliawati, M. A. Khan, J. Iqbal, "Neural network-based adaptive global sliding mode MPPT controller design for stand-alone photovoltaic systems", Plos One, Vol. 17, No. 1,2022 e0260480.

[12] E. Karatepe, T. Hiyama, "Artificial neural network-polar coordinated fuzzy controller based maximum power point tracking control under partially shaded condition", IET Renewable Power Generation, Vol. 3 No. 2, 2009, pp. 239-253.

[13] Y. N. Anagreh, A. Alnassan, A. Radaideh, "High Performance MPPT Approach for Off-Line PV System Equipped with Storage Batteries and Electrolyzer", International Journal of Renewable Energy Development, Vol. 10, No. 3, 2021, 507-515.

[14] S. Ullah, Q. Khan, A. Mehmood, S. A. M. Kirmani, O. Mechali, "Neuro-adaptive fast integral terminal sliding mode control design with variable gain robust exact differentiator for under-actuated quadcopter UAV", ISA Transactions, Vol. 120, 2022, pp. 293-304.

[15] S. Manna, D.K. Singh, A.K. Akella, H. Kotb, K.M. AboRas, H.M. Zawbaa, S. Kamel, "Design and implementation of a new adaptive MPPT controller for solar PV systems", Energy Reports, Vol. 9, 2023, pp.1818-1829.

[16] S. Manna, A. K. Akella, D. K. Singh,"A novel MRAC-MPPT scheme to enhance speed and accuracy in PV systems", Iranian Journal of Science and Technology, Transactions of Electrical Engineering, Vol. 47, No. 1, 2023, pp.233-254.

[17] S. Manna, D. K. Singh, A. K. Akella, A. Y. Abdelaziz, M. Prasad, "A novel robust model reference adaptive MPPT controller for Photovoltaic systems", Scientia Iranica, 2022.

[18] S. Manna, A. K. Akella, D. K. Si, D. K. Singh, "Implementation of a novel robust model reference adaptive controller-based MPPT for stand-alone and grid-connected photovoltaic system", Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, Vol. 45, No. 1, 2023, pp. 1321-1345.

[19] D. P. Hohm, M. E. Ropp, "Comparative study of maximum power point tracking algorithms", Progress in photovoltaics: Research and Applications, Vol. 11, No. 1, 2003, pp. 47-62.

[20] D. Sera, L. Mathe, T. Kerekes, S. V. Spataru, R. Teodorescu, "On the perturb-and-observe and incremental conductance MPPT methods for PV systems", IEEE Journal of Photovoltaics, Vol. 3, No. 3, 2013, 1070-1078.

[21] N. Femia, G. Petrone, G. Spagnuolo, M. Vitelli, "Optimization of perturb and observe maximum power point tracking method", IEEE Transactions on Power Electronics, Vol. 20, No. 4, 2005, pp. 963-973.

[22] M. Killi, S. Samanta, "Modified perturb and observe MPPT algorithm for drift avoidance in photovoltaic systems", IEEE Transactions on Industrial Electronics, Vol. 62, No. 9, 2015, pp. 5549-5559.

[23] A. Charaabi, A. Zaidi, O. Barambones, N. Zanzouri, "Implementation of adjustable variable step based backstepping control for the PV power plant", International Journal of Electrical Power & Energy Systems, Vol. 136, p.107682.

[24] Y. Yang, F. P. Zhao, "Adaptive perturb and observe MPPT technique for grid-connected photovoltaic inverters", Procedia Engineering, Vol 23, 2011, pp. 468-473.

[25] S. Dahale, A. Das, N. M. Pindoriya, S. Rajendran, "An overview of DC-DC converter topologies and controls in DC microgrid", Proceedings of the IEEE 7th International Conference on Power Systems, Pune, India, 21-23 December 2017, pp. 410-415.

[26] M. H. Taghvaee, M. A. M. Radzi, S. M. Moosavain, H. Hizam, M. H. Marhaban, "A current and future study on non-isolated DC–DC converters for photovoltaic applications", Renewable and Sustainable Energy Reviews, Vol. 17, 2013, pp. 216-227.

[27] M. H. Rashid, "Power electronics handbook", Butterworth-Heinemann, 2017.

[28] O. P. Mahela, A. G. Shaik, "Comprehensive overview of grid interfaced solar photovoltaic systems", Renewable and Sustainable Energy Reviews, Vol. 68, 2017, pp. 316-332.

[29] P. V. Mahesh, S. Meyyappan, R. Alla, "Support Vector Regression Machine Learning based Maximum Power Point Tracking for Solar Photovoltaic systems", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 1, 2023, pp. 100-108.

[30] A. J. Mahdi, S. Fahad, W. Tang, "An adaptive current limiting controller for a wireless power transmission system energized by a PV generator", Electronics, Vol. 9, No. 10, 2020, p. 1648.

[31] S. Fahad, A. J. Mahdi, W. H. Tang, K. Huang, Y. Liu, "Particle swarm optimization based DC-link voltage control for two stage grid connected PV inverter", Proceedings of the IEEE International Conference on Power System Technology, Guangzhou, China, 6-8 November 2018, pp. 2233-2241.

[32] S. K. Kollimalla, M. K. Mishra, "A novel adaptive P&O MPPT algorithm considering sudden changes in the irradiance", IEEE Transactions on Energy Conversion, Vol. 29, No. 3, 2014, pp. 602-610.

[33] K. L. Lian, J. H. Jhang, I. S. Tian, "A maximum power point tracking method based on perturb-and-observe combined with particle swarm optimization", IEEE Journal of Photovoltaics, Vol. 4, No. 2, 2014, pp. 626-633.

[34] T. Esram, P. L. Chapman, "Comparison of photovoltaic array maximum power point tracking techniques", IEEE Transactions on Energy Conversion, Vol. 22, No. 2, 2007, pp. 439-449.

[35] S. Jain, V. Agarwal, "Comparison of the performance of maximum power point tracking schemes applied to single-stage grid-connected photovoltaic systems", IET Electric Power Applications, Vol. 1, No. 5, 2007, pp. 753-762.

[36] G. C. Mahato, T. R. Choudhury, "Study of MPPT and FPPT: A brief comparison", Proceedings of the IEEE 17th India Council International Conference, New Delhi, India, 10-13 December 2020, pp. 1-7.

[37] L. Piegari, R. Rizzo, "Adaptive perturb and observe algorithm for photovoltaic maximum power point tracking", IET Renewable Power Generation, Vol. 4, No. 4, 2010, pp. 317-328.

[38] S. D. Al-Majidi, M. F. Abbod, H. S. Al-Raweshidy, "A novel maximum power point tracking technique based on fuzzy logic for photovoltaic systems", International Journal of Hydrogen Energy, Vol. 43, No 31, 2018, pp. 14158-14171.

[39] J. Ahmad, "A Fractional Open Circuit Voltage Based Maximum Power Point Tracker for Photovoltaic Arrays", Proceedings of the 2nd International Conference on Software Technology and Engineering, San Juan, PR, USA, 3-5 October 2010, pp. 247-250.

[40] L. Hichem, O. Amar, M. Leila, "Optimized ANN-fuzzy MPPT controller for a stand-alone PV system under fast-changing atmospheric conditions", Bulletin of Electrical Engineering and Informatics, Vol. 12, No. 4, 2023, pp. 1960-1981.

Illustration of the voltage stability by using the slope of the tangent vector component

Original Scientific Paper

Agron Bislimi
UBT, University for Business and Technology,
St. Calabria n.n. 10000, Prishtina, Kosovo.
agron.bislimi@ubt-uni.net

Abstract *– This Paper is dedicated to the analysis of the evolution of the tangent vector during the Continuous Power Flow (CPF) iterations. The flow of the tangent slope (measured in degrees) is shown through the coefficient of lambda tangent vector component and the maximum voltage tangent vector component. A 17 Node Network was used for the purposes of this Paper. The system was modelled in MATLAB software. The admittance matrix of the node voltage equations was formulated and the functions in MATLAB were developed for the systematic formation of the node admittance matrix. Equations for the calculated network were generated in MATLAB. 32 Iterations were performed. Iterations and corrections of iterations were done manually. Firstly, the results for the tangent vectors calculated through the CPF program were compared to the results for the tangents directly calculated with mathematical formula for the tangent, and both results match. The chart, which contains the classical PV curve and the flow of tangent vectors during the CPF iterations, was developed based on the results obtained. The increase in the slope of the tangent in the PV diagram imposes a clear numerical stability limit by specifying an angle limit value, which can be used to trigger an alarm. In addition to the classic Power-Voltage (PV) curve, this serves as an additional indicator for ensuring voltage stability of the examined system.*

Keywords*: Voltage Stability, Tangent Slope, CPF Iterations, Tangent Vector Component*

1. INTRODUCTION

This Research differs from others because the literature uses the tangents to find the PV curves for the examined system, which serves as an indicator to ensure the voltage stability of the examined system. In addition to classical PV curves, my Research also explains the voltage stability through the slope of the tangent vector components during CPF iterations. This serves as an additional indicator for ensuring the stability of the voltage of the examined system. Using examples, it also provides a theoretical and practical explanation of the matching of directly calculated tangent results and their calculation through CPF.

The tangent vector component "t" in [1] represents the values (voltage angle δ, voltage magnitude V and load parameter lambda λ). The variable values with the largest tangent vector component from the tangent "t" are used as continuous parameter. The lambda tangent vector component ($d\lambda$) is positive for the upper part of the PV. It is zero at the critical point where the Jacobian matrix becomes singular and is negative at the bottom of the PV curve. The sign of $d\lambda$ shows us whether or not the critical point has been reached. The flow of the voltage magnitudes in view of the network load provides information about the voltage stability limit or the network load which, when exceeds, causes collapse to the voltage [1]. The most common parameterization techniques used by CPF to remove the singularity of the PF Jacobian matrix is in [2]. The sensitivity of the information discussed from the tangent vector. The continuation power flow traces the complete P-V curves by automatically changing the value of a parameter. In the local parameterization technique, a parameter change always occurs close to the Maximum Loading Point (MLP). Generally, the loading factor λ is an initially chosen parameter. Close to the MLP, it changes to the voltage magnitude that presents the largest variations and after a few points, it changes back to λ. The differential change in the voltage at each node for a given differential change in the system load is available from the tangent vector [2]. In [3] is explained the voltage stability index from the tangent vector. The ratio $Cd\lambda/dVi$, where C is constant, which is easier to be handled numerically, can be defined as voltage stability index for the entire system [3]. In each CPF iteration step, the active and reactive loads of the network are increased by a certain percentage of the base load. Voltage stability index using the tangent can be derived at [4]. The weakest node means a node which is nearer to the voltage collapse due to the lack of the reactive power. It means that the weakest node has the maximum ratio of differential change in the voltage to the differential change in the active power demand at the critical point. It represents the gradient or the tangent of a PV curve [5]. In

[6-8], it consists of the calculation of the power flow and voltages of a transmission network for specified terminal conditions. Such calculations are required for steady-state performance analysis of power systems. Thus, the tangent vector provides information about critical nodes or critical parts of the network [6-8]. Equation [$F\underline{\delta}$, $F\underline{V}$, $F\underline{\lambda}$] is the Jacobian matrix of the power flow, augmented by a column multiplied by a tangent vector required. One problem arises while solving this Equation, i.e., the additional unknowns are added to the power flow equation when the load factor λ is added to the power flow equations. The tangent vector is represented by the Equation [$d\underline{\delta}$, $d\underline{V}$, $d\underline{\lambda}$]^T, T= transpose. The critical point is reached when the load change reaches its maximum value. The change of sign of dLambda in the tangent vector is constantly checked. The CPF method, which includes the process of predictor and the process of corrector, can be applied to solve the non-linear equations. The predictor process starts with the calculation of the tangent vector. The tangent vector must be normalized in order to guarantee the non-singularity of the augmented Jacobian matrix [9-11]. The tangent vector gives the sensitivity of the parameters at a point in the PV curve where they are evaluated. The sensitivities in the entire parts of the PV curve should be evaluated [12]. The tangent vector t is used to predict initial values for the next step of CPF. The tangent size is ($6n$ + 1) for a three-phase system, where there are $6n$ voltage variables and one loading parameter (λ) variable. Prediction can be made by using tangent and secant predictors. Tangent predictors predict only by using the current solution, and the secant predictors use the current and previous solutions in order to make predictions [13]. An efficient geometric parameterization technique for the CPF is presented in [14]. Intending to reduce the CPU time, the effectiveness caused by updating the Jacobian matrix is investigated only when the system undergoes a significant change. The Paper presents in [15] a potential algorithm for Continuation Power Flow method in voltage stability assessment. [16] proposes an innovative method by modifying the Conventional Continuation Power Flow (CCPF) method. The tangent predictor is proposed to estimate the next predicted solution from two previous corrected solutions. And the corrector step is proposed to determine the next corrected solution on the exact solution. This correct solution is constrained to lie in the hyperplane running through the predicted solution orthogonal to the line from the two previous corrected solutions [16].

This field of research is currently focused on researching the voltage stability by using the slope of the tangent vector component. This Paper is structured as follows: Second chapter shows the methods employed, namely the direct calculation of tangents and tangent calculation through CPF. It consists of a research and detailed explanation of the Lambda as a continuous parameter. Moreover, it illustrates the Voltage V as a continuous parameter. Third section consists of the results. Furthermore, it compares the tangent vectors when calculated with the CPF program and when calculated directly using the mathematical formula. Finally, the chart, which contains the classical PV curve and the flow of the tangent vectors during the CPF iterations, is developed using the results. In addition to the classic Power-Voltage (PV) curve, this serves as an additional indicator for ensuring voltage stability of the examined system.

2. METHODS - DIRECT CALCULATION OF TANGENTS AND TANGENT CALCULATION THROUGH CPF. TANGENTS USING LAMBDA (λ) AND VOLTAGE V AS CONTINUOUS PARAMETERS

2.1. DIRECT CALCULATION OF TANGENTS AND TANGENT CALCULATION THROUGH CPF

2.1.1. Direct calculation of tangents

The method below was employed for the direct calculation of tangents using the tangent formula

To calculate the slope of any point on the line, we draw a tangent to it and calculate the value of tan of the angle it makes with the base. A tangent to a curve is a straight line that touches the curve at a given point and represents the gradient of the curve at that point.

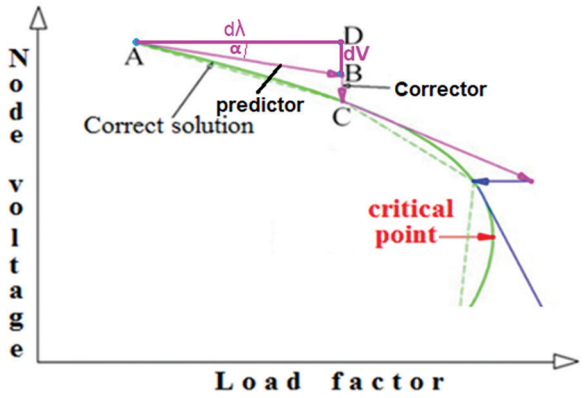


Fig. 1. Direct calculation the tangent in the 17-node network.

$$\tan(\alpha) = \frac{\sin(\alpha)}{\cos(\alpha)} = \frac{DB}{AD} = \frac{dV}{d\lambda}$$

2.1.2. Calculation of the Tangent using CPF Predictor step

The method below was employed to calculate the tangents through CPF. The system of non-linear equations looks like this after inserting the lambda parameter:

$$F(\underline{v}, \underline{\delta}, \underline{\lambda}) = 0 \quad (1)$$

The Taylor linearization of equation (1) yields, neglecting the higher-order terms:

$$d\underline{F} = \underline{F_v} d\underline{v} + F_{\underline{\delta}} d\underline{\delta} + \underline{F_\lambda} d\lambda = 0$$

$F\underline{V}$- Derivatives of F with respect to the magnitude of the voltage. $F\underline{\delta}$- Derivatives of F with respect to the voltage angle. $F\underline{\lambda}$- Derivatives from F to λ, corresponds to a column vector [1].

The first task in the predictor step is to calculate the tangent vector component.

$$\left[\underline{F\delta},\ \underline{FV},\ \underline{F\lambda}\right]\begin{bmatrix}\underline{d\delta}\\ \underline{dV}\\ \underline{d\lambda}\end{bmatrix}=0 \tag{2}$$

On the left are Jacobian matrix elements

$$\underline{t}=\begin{bmatrix}\underline{d\delta}\\ \underline{dV}\\ \underline{d\lambda}\end{bmatrix} \tag{3}$$

$\underline{t}$ is the tangent vector component

The system of equations (2) is underdetermined by the additional unknown $d\lambda$.

One, therefore, adds another equation (5) and supplements the right-hand 0-vector by:

$$\text{tk}=\pm 1 \tag{4}$$

$$\begin{bmatrix}\underline{F\delta} & \underline{Fv} & \underline{F\lambda}\\ \underline{e_k} & & \end{bmatrix}\left[\underline{t}\right]=\begin{bmatrix}0\\ \pm 1\end{bmatrix} \tag{5}$$

$\underline{e_k}$ is a vector with all elements equal to 0, except the k^{th} element, which equals 1 [3].

The value of k depends on the continuation parameter. For instance, with the continuation parameter l, k corresponds to the last position in e_k. If the continuation parameter increases over the course of the CPF iterations, t_k is set to +1, otherwise, it is set to -1. For instance, with continuation parameter l, before the reversal point of the PV curve (the stability limit) is reached, $t_k = 1$, then $t_k = -1$.

Solving equation (5) gives the tangent vector component $[t]$.

2.2. LAMBDA USED AS A CONTINUOUS PARAMETER AND VOLTAGE *V* AS CONTINUATION PARAMETER

2.2.1. lambda used as a continuous parameter

Since λ Continuation parameter is in e_k at the λ position (n-th position) is 1. As long as λ is continuation parameter, the e_k series and λ column is eliminated at the corrector step. (Since λ is fixed on the predictor value in the corrector iterations, it has to be eliminated from the corrector equation system). Chap 2. need to construct the classical PV curve. The iterations using the continuous parameter λ are performed while the tangent vector component λ is greater than the component of the high (negative) voltage tangent vector. As long as λ is a continuous parameter, λ remains constant during the corrector iterations, but the voltage angles δi and voltage values Vi change. Fig. 3 explains the tangents when the parameter λ is continuous $V=f(\lambda)$ for the 17-node network shown in Fig. 2 and explains in which range of angles the iterations span using the continuous parameters Lambda and V. For the angle range (0 → -45°), the parameter Lambda is a continuous parameter.

The largest voltage tangent vector component will be in range (0 → -1). The closer we get to the critical point, the higher the tangent angle. For the angle -44°, we gain -0.9656. For the angle -135°, the condition to change the continuous parameter V to the continuous parameter Lambda is met and the iterations are carried until the end by using the continuous parameter Lambda.

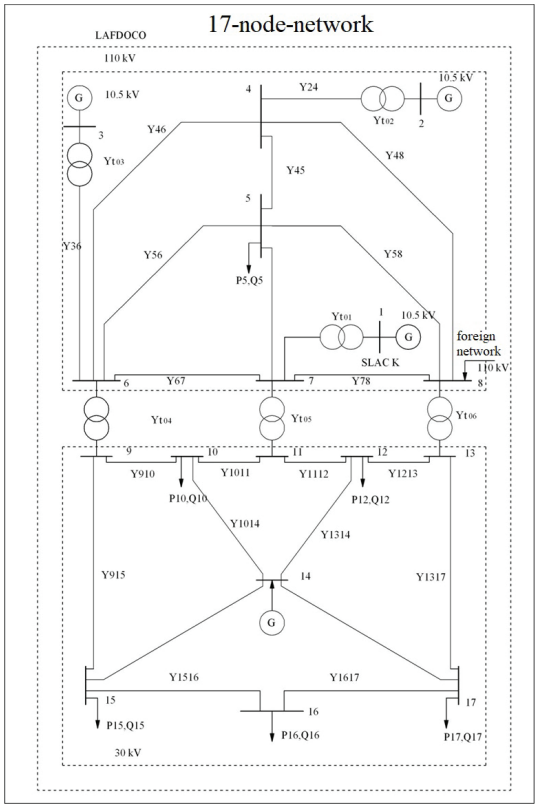


Fig. 2. 17- node network

tan (-135° → -180°) = (1→ 0). The largest (-1) voltage tangent vector component will be in the interval (-1→ 0), (the lower part of the curve).

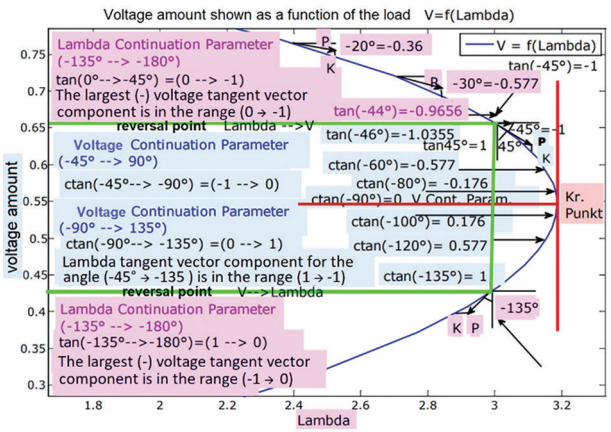


Fig. 3. Tangents using λ as a CP $V=f(\lambda)$

At the beginning of the CPF iteration, large changes in load correspond to small changes in voltage $d\lambda >> dV$. As the iterations progress, the ratio between changes in load and changes in voltage becomes smaller and

smaller. These relations are expressed by the quotient between the largest negative voltage and the lambda tangent vector component.

Fig. 3 shows that the value of the largest negative voltage tangent vector component of -0.36 corresponds to an angle of $tan^{-1}(-0,36)$ = -20°. If the condition of inverting λ to V is met, then the largest voltage tangent vector component value is -1. This corresponds to a tangent angle of -45 degrees.

Thereafter, the absolute value of the quotient between the largest voltage tangent vector component and the λ tangent vector component is greater than 1 (negative sign), resulting in a tangent angle of more than -45 degrees.

2.2.2. Voltage V as continuation parameter

As long as V is continuation parameter, the e_k- vector row and the V_5- column are eliminated at the corrector step. (V_5 is fixed at the predictor value in the corrector iterations). The differentials $dV_5 = 0$ are thus eliminated from the corrector equation system. For the illustration in Fig. 4, the tangent angle is calculated from the quotient between λ and the largest voltage tangent vector component. Since the voltage component is fixed at -1, it follows directly from the p. u. value of the lambda tangent vector component. The change from continuation parameter λ to V occurs at a voltage tangent vector component of -1.0355 p. u., which corresponds to a tangent angle of -46°. Reason: The inversion condition is that tangent angle is < -45°. The closer you get to the critical point, the more the relationship between voltage and load changes. This corresponds to a tangent angle of -90° at the critical point in Fig. 4.

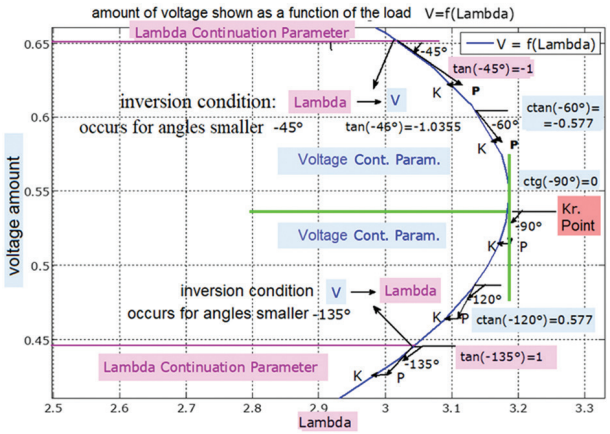


Fig. 4. $V=f(\lambda)$ using V as a continuous parameter

Thereafter, the sign of the lambda tangent vector component reverses from + to -, and the tangent angle assumes values of more than 90°. The sign remains negative, see Fig. 3, lower part of the PV curve. The values of both the lambda tangent vector component and the tangent angle increase again. Because the value of the tangent angle is negative, the slope of the tangent becomes flatter again.

For an angle of -135°, the quotient between voltage and lambda tangent vector components becomes 1 (tan(-135) = 1), and thus the condition to invert the continuation parameter from V back to λ is met. Once this condition is met, the changes in the voltage become smaller again.

Lambda tangent vector component for the angle range (-45° → -135°) will be in the range (1→ -1). For the angle -45° (tan(-45°) = -1), the condition of changing the continuous parameter Lambda to the continuous parameter V is met again.

After that, CPF iterations continue with V as continuous parameter.

- For the angle (-46°)→tan(-46°)=-1.0355.
- For the angle -80° → $ctan$(-80°) = -0.176 (The upper part of the curve)
- For the angle (-90°) → $ctan$(-90°)=0, ($tang$(-90°)=Infinite).
- For the angle (-100°) → $ctan$(-100°)=0.176 (The bottom part of the curve)

3. RESULTS: COMPARISON OF THE TANGENT VECTORS WHEN CALCULATED USING THE CPF PROGRAM AND WHEN CALCULATED DIRECTLY. COURSE OF THE QUOTIENT OF THE LARGEST (NEGATIVE) VOLTAGE TANGENT VECTOR COMPONENT AND LAMBDA TANGENT VECTOR COMPONENT.

3.1. COMPARISON OF THE TANGENT VECTORS WHEN CALCULATED USING THE CPF PROGRAM AND WHEN CALCULATED DIRECTLY

3.1.1. Direct calculation of tangents

Data for direct calculation the tangent for the first iteration.

Iteration 1: Before the predictor: V_5=0.9520
Before the predictor: λ=0
After the predictor: λ=0.15
$dV1 = V_{5\ (\text{After the predictor})} - V_{5\ (\text{Before the predictor})}$ =
=0.9520-0.9442=-0,0078
$d\lambda 1 = \lambda_{\ (\text{After the predictor})} - \lambda_{\ (\text{Before the predictor})} = 0.15 - 0 = 0.15$

The direct calculation of tangents for the first iteration is:

$$t1 = \frac{dV1}{d\lambda 1} = \frac{-0.0078}{0.15} = -0.052$$

Data for direct calculation the tangent for iteration 16.

Iteration 16: Before the predictor: V5=0.5541
After the predictor: V5=0.5441;
Before the predictor: λ=3.1854;
After the predictor: λ=3.1880;

$dV1 = V_{5\ (\text{After the predictor})} - V_{5\ (\text{Before the predictor})}$ =0.5441 - 0.5541=-0,01

$d\lambda 1 = \lambda_{\ (\text{After the predictor})} - \lambda_{\ (\text{Before the predictor})} = 3.1880 - 3.1854 = 0.0026$

$$\text{ct}\,1 = \frac{1}{\text{t1}} = \frac{\text{d}\lambda 1}{\text{dV1}} = \frac{0.0026}{-0.01} = -0.26$$

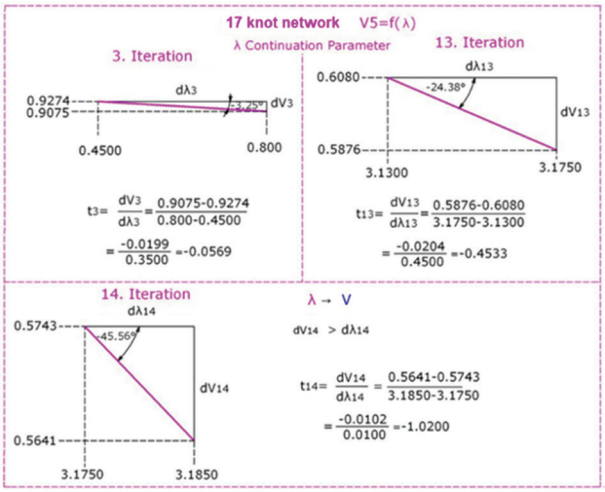


Fig. 5. Direct calculation of the tangents in the iterations 3, 13, 14 at the node 5 in the 17-node network.

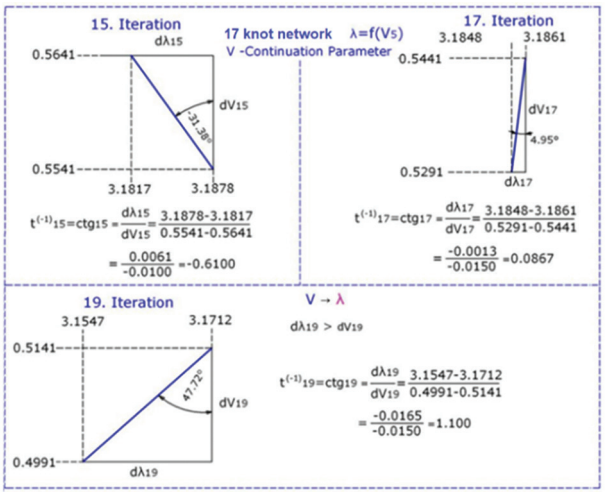


Fig. 6. Direct calculation of the tangents in the iterations 15, 17 and 19 at node 5 in the 17-node network.

3.1.2. Tangent calculation through CPF and compare

Calculation of tangents through CPF:

The tangent vector components for a 17-node network calculated using the CPF are as follows:

In the iteration 1:

t_1= [0.0031 -0.0023 -0.0318 -0.0631 -0.0370 -0.0360
-0.0303 -0.0531 -0.0573 -0.0504 -0.0562 -0.0495
-0.0578 -0.0632 -0.0724 -0.0598
V_4 V_5 V_6 V_7 V_8 V_9
-0.0303 -0.0531 -0.0573 -0.0504 -0.0562 -0.0495
V_{10} V_{11} V_{12} V_{13} V_{14} V_{15}
-0.0303 -0.0531 -0.0573 -0.0504 -0.0562 -0.0495
V_{16} V_{17} λ
-0.0562 -0.0495 **1.0000**]^T

T = transpose

In the iteration 1, the largest (negative) voltage tangent vector component occurs at node 5. The tangent inclination in direction V_5 is given by the quotient $dV_5/d\lambda$ =-0,0521/1 = -0,0521. If you compare this with the result of the direct calculation, which is 0.052 (see subchapter 3.1.1), you can see a good agreement with an error of only 0.001.

In the iteration 16, the tangent direction for the continuation parameter V_5 results from the values taken from $t16$ to $d\lambda/dV_5$ = -0.2626/1= -0.2626.

If you compare this with the result of the direct calculation, which is 0.26, you can see a good agreement with an error of only 0.0026.

Table 1. Tangents in the iterations 1-32 at the node 5 in the 17-node network calculation through CPF.

	Tangents in the iterations 1-32 at the node 5 in the 17-node network.		
Iteration	dV_5	$d\lambda$	$tangent=dV/d\lambda$
1	-0.0521	1.0000	-0.0521
2	-0.0536	1.0000	-0.0536
3	-0.0570	1.0000	-0.0570
4	-0.0616	1.0000	-0.0616
5	-0.0684	1.0000	-0.0684
6	-0.0776	1.0000	-0.0776
7	-0.0912	1.0000	-0.0912
8	-0.1103	1.0000	-0.1103
9	-0.1474	1.0000	-0.1474
10	-0.1949	1.0000	-0.1949
11	-0.2440	1.0000	-0.2440
12	-0.3115	1.0000	-0.3115
13	-0.4523	1.0000	-0.4523
14	-1.0238	1.0000	-1.0238
15	-1.0000	0.6133	-0.6133
16	-1.0000	0.2626	-0.2626
17	-1.0000	-0.0851	0.0851
18	-1.0000	-0.5960	0.5960
19	-1.0000	-1.0975	-1.0975
20	-0.6314	-1.0000	0.6314
21	-0.4746	-1.0000	0.4746
22	-0.3662	-1.0000	0.3662
23	-0.2818	-1.0000	0.2818
24	-0.2255	-1.0000	0.2255
25	-0.1827	-1.0000	0.1827
26	-0.1377	-1.0000	0.1377
27	-0.1157	-1.0000	0.1157
28	-0.1021	-1.0000	0.1021
29	-0.0927	-1.0000	0.0927
30	-0.0856	-1.0000	0.0856
31	-0.0802	-1.0000	0.0802
32	-0.0759	-1.0000	0.0759

Iteration 3, 13 and 14 using continuation parameter lambda

The tangent vector components for a 17-node network calculated using the CPF are as follows:

In the iteration 3:

t_3=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	0.0007	-0.0049	-0.0349	-0.0685	-0.0403	-0.0383
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.0326	-0.0569	-0.0612	-0.0539	-0.0600	-0.0531
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.0618	-0.0673	-0.0770	-0.0639		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.0321	**-0.0570**	-0.0294	-0.0306	-0.0368	-0.0219
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.0171	-0.0219	-0.0182	-0.0242	-0.0195	-0.0299
	V_{17}	λ				
	-0.0562	**1.0000**]^T				

T = transpose

Node 1 is the slack, nodes 2, 3 and 14 are PV nodes. Therefore, no voltage magnitude components appear for these nodes in the tangent vector.

The largest (negative) voltage tangent vector component occurs at node 5. The tangent inclination in direction V_5 is given by the quotient $dV_5/d\lambda$ = -0,0570/1 = -0,057. If you compare this with the result of the direct calculation from Fig. 5, which is 0.0569, you can see a good agreement with an error of only 0.0001.

In the iteration 13, the tangent vector components calculated using the CPF are:

t_{13}=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	-0.3010	-0.3102	-0.3580	-0.6216	-0.3646	-0.2516
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.2733	-0.4006	-0.4024	-0.3677	-0.3962	-0.3767
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.4049	-0.4184	-0.4398	-0.4080		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.2624	**-0.4523**	-0.2394	-0.2761	-0.3071	-0.14131
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.1083	-0.1684	-0.1189	-0.1725	-0.0887	-0.1120
	V_{17}	λ				
	-0.0958	**1.0000**]^T				

T = transpose

The comparison of dV_5 = -0.4523 with the value of -0.4533 from Fig. 5 shows again good value agreement.

The same applies to the iteration 14:

t_{14}=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	-0.7464	-0.7545	-0.8245	-1.4313	-0.8271	-0.5460
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.6141	-0.8796	-0.8747	-0.8036	-0.8614	-0.8276
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.8795	-0.9026	-0.9342	-0.8836		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.5932	**-1.0238**	-0.5400	-0.6248	-0.6933	-0.3131
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.2358	-0.3742	-0.2594	-0.3804	-0.1823	-0.2188
	V_{17}	λ				
	-0.2012	**1.0000**]^T				

T = transpose

dV_5 = -1.0238; from Fig. 5: -1.020. The differences in the values are a consequence of the limited numerical accuracy.

Iterations 15,17 and 19 with voltage V as continuation parameter.

The tangent vector components accounted for by the CPF in the iteration 15 are:

t_{15}=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	-0.7472	-0.7510	-0.8082	-1.4047	-0.8069	-0.5251
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.5937	-0.8493	-0.8419	-0.7749	-0.8292	-0.7993
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.8463	-0.8667	-0.8925	-0.8494		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.5789	**-1.0000**	-0.5267	-0.6098	-0.6764	-0.3025
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.2274	-0.3632	-0.2504	-0.3684	-0.1725	-0.2035
	V_{17}	λ				
	-0.1919	**0.6133**]^T				

T = transpose

The tangent direction for the continuation parameter V_5 results from the values taken from t15 to $d\lambda$ / dV_5 = -0.6133 / 1= -0.6133. From Fig. 6 with -0.610 again we get a good match. The same applies to the iterations 17 and 19, see tangent vectors below.

Tangent vector in the iteration 17:

t_{17}=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	-0.7800	-0.7756	-0.8112	-1.4159	-0.8022	-0.5063
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.5900	-0.8267	-0.8139	-0.7517	-0.8017	-0.7781
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.8178	-0.8340	-0.8498	-0.8192		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.5779	**-1.0000**	-0.5249	-0.6079	-0.6743	-0.2957
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.2214	-0.3581	-0.2442	-0.3617	-0.1615	-0.1832
	V_{17}	λ				
	-0.1825	**0.0851**]^T				

T = transpose

Tangent vector in the iteration 19:

t_{19}=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	-0.8228	-0.8050	-0.8091	-1.4284	-0.7875	-0.4723
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.5726	-0.7841	-0.7636	-0.7090	-0.7521	-0.7382
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.7665	-0.7763	-0.7770	-0.7655		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.5757	**-1.0000**	-0.5216	-0.6032	-0.6700	-0.2851
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.2120	-0.3494	-0.2342	-0.3510	-0.1449	-0.1527
	V_{17}	λ				
	-0.1682	-**1.0975**]^T				

T = transpose

Iterations 20, 25 and 32 with continuation parameter lambda

Similar calculations can also be carried out for the lower part of the PV curve with the continuation parameter lambda, which is not done here. The tangent vector components of the iterations 20, 25 and 32 given below are used to develop **Fig. 7.**

Tangent vector component in the iteration 20:

t_{20}=	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[	-0.5308	-0.5152	-0.5080	-0.9044	-0.4903	-0.2857
	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
	-0.3542	-0.4791	-0.4637	-0.4316	-0.4567	-0.4510
	δ_{14}	δ_{15}	δ_{16}	δ_{17}		
	-0.4652	-0.4694	-0.4652	-0.4638		
	V_4	V_5	V_6	V_7	V_8	V_9
	-0.3626	**-0.6314**	-0.3281	-0.3788	-0.4212	-0.1766
	V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
	-0.1308	-0.2176	-0.1446	-0.2180	-0.0863	-0.0869
	V_{17}	λ				
	-0.1017	-**1.0000**]^T				

T = transpose

In the Iteration 25:

t_{25}=

δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[-0.1583	-0.1399	-0.1148	-0.2506	-0.0972	-0.0321
δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
-0.0635	-0.0695	-0.0591	-0.0575	-0.0580	-0.0649
δ_{14}	δ_{15}	δ_{16}	δ_{17}		
-0.0588	-0.0550	-0.0424	-0.0566		
V_4	V_5	V_6	V_7	V_8	V_9
-0.1006	**-0.1872**	-0.0893	-0.0984	-0.1137	-0.0412
V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
-0.0286	-0.0519	-0.0317	-0.0518	-0.0117	-0.0016
V_{17}	λ				
-0.0174	**-1.0000**]^T				

T = transpose

In the Iteration 32:

t_{32}=

δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
[-0.0272	-0.0011	0.0340	-0.0627	0.0570	0.0569
δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}
0.0294	0.0883	0.0947	0.0846	0.0927	0.0809
δ_{14}	δ_{15}	δ_{16}	δ_{17}		
0.0951	0.1017	0.1137	0.0971		
V_4	V_5	V_6	V_7	V_8	V_9
-0.0287	**-0.0759**	-0.0246	-0.0243	-0.0300	-0.0060
V_{10}	V_{11}	V_{12}	V_{13}	V_{15}	V_{16}
-0.0021	-0.0088	-0.0024	-0.0084	-0.0061	-0.0170
V_{17}	λ				
-0.0031	**-1.0000**]^T				

T = transpose

3.2. COURSE OF THE QUOTIENT OF THE LARGEST (NEGATIVE) VOLTAGE TANGENT VECTOR COMPONENT AND LAMBDA TANGENT VECTOR COMPONENT

Fig. 7 shows the quotient between the largest negative voltage tangent vector component and the lambda tangent vector component. The scale for this quotient can be found on the ordinate of the lower part of the figure. Since it also defines the negative slope of the tangent (see Fig. 3), the figure also contains an angle scaling, namely on the abscissa of the lower part.

The upper part of the figure shows the PV curve, with the abscissa being the voltage axis and the ordinate being the lambda axis. The V-axis scaling is non-linear, so the PV curve is mapped symmetrically, putting the critical point right in the middle. (In fact, the PV curve is usually asymmetrical with respect to the critical point). The values for the development of the Fig. 6 are taken from subchapters 3.1.1. and 3.1.2.

The purple curve on the right describes CPF iterations 1 to 13. The middle blue curve applies to iterations 14-19. And, the left purple curve represents iteration 20 – 32, (see subchapter 3.1.2).

The starting point is on the right side of the Fig. 7. The continuation parameter is the load factor λ. The quotient $dV/d\lambda$ and the tangent inclination are still small for the time being. In particular, $d\lambda = +1$; $dV < 0$, $|dV| < 1$). The increase in load (λ increases) results in the increase of the absolute value of the quotient according to a tangent function, see purple curve. At point S it takes the value -1 ($d\lambda = +1$, $dV = -1$).

This corresponds to a tangent inclination of $tg^{-1}(-1)$ = -45°. If the load factor is further increased, the voltage tangent vector component becomes larger than the lambda tangent vector component (point E), and therefore continuation parameter changes from λ to V.

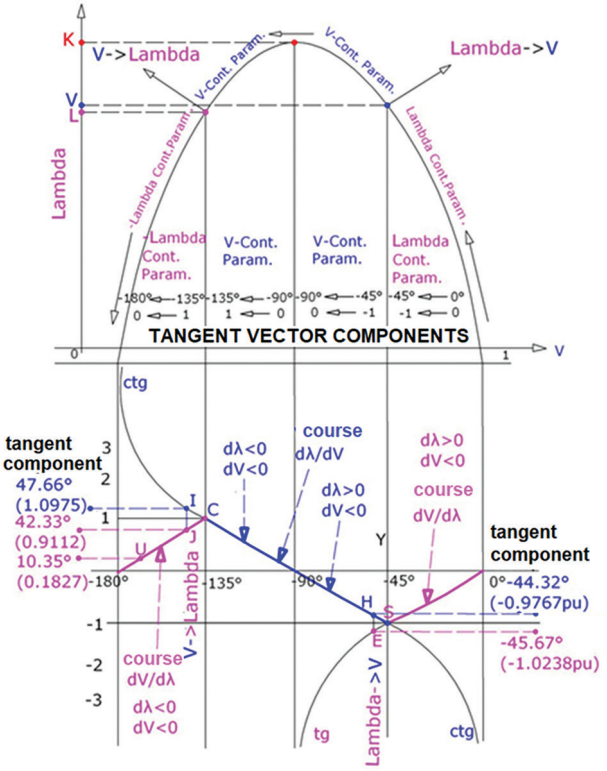


Fig. 7. The component of the tangent vector of the higher (negative) voltage compared to the lambda tangent vector component, values for the 17- node network.

In order for the quotient of the tangent vector components to remain finite, ($d\lambda/dV$) needs to get inverted. Therefore, its further course develops according to a cotangent function (blue curve). The tangent angle of -45.67° thus corresponds to a quotient of -1/1.0238 = -0.9768 (see point H on the blue curve). This is now valid for all further CPF iterations carried out with Continuation Parameter V. The critical point is reached at -90 (corresponding to $ctg(-90) = 0$ or $dV = \infty$).

At a tangent inclination of -135°, both $d\lambda$ and dV have the value -1, and the quotient of the tangent vector components becomes 1, see point C. For the development of point I, see Chap. 3.1.2 for the iteration 19.: dV = -1, $d\lambda$ = -1.0975,

$$ctg\text{-}1(d\lambda/dV) = ctg\text{-}1(1{,}0975/1) = 42{,}33°.$$

However, since the tangent slope has exceeded the critical value of -90°, a 180° transformation must be performed, which is why the final angle is: 42.33° - 180° = -137.66°. The condition for resetting the continuation parameter from V to λ is now fulfilled ($d\lambda > dV$). Accordingly, the quotient $d\lambda/dV$ is inverted to $dV/d\lambda$, and the purple tangent function shown on the left in Fig. 7 is valid for the rest of the curve.

Point I on the blue curve now corresponds to point J on the purple curve.

$(tg^{-1}(d\lambda/dV) = tg^{-1}(1/1{,}0975) = tg^{-1}(0{,}9112) = -137{,}66°)$.

For the development of a further point of the left purple curve, values from Chap. 3.3 are taken, for instance, for the iteration 25:

$d\lambda$=-1, dV= -0.1827; $tg^{-1}(dV/d\lambda) = tg^{-1}(0{,}1827) = 10{,}35°$. After transformation by 180°, the final tangent inclination is 10.35 – 180 = -169.65°. In the figure, this corresponds to point U.

4. CONCLUSIONS

This Paper illustrates the voltage stability by using the slope of the tangent vector component. It represents a research and detailed explanation of Lambda as a continuous parameter. The tangent angle was calculated from the quotient between the voltage and the largest lambda tangent vector component. Moreover, it illustrates the Voltage V as continuous parameter, where the tangent angle was calculated from the coefficient between lambda λ and the largest tangent vector component of the voltage. Furthermore, it compares the tangent vectors when calculated with the continuous power flow program and when directly calculated, for the most vulnerable network node. The highest (negative) voltage tangent vector component occurs in the node 5 in the network with 17 nodes. The results for the tangent vectors calculated through the CPF program were compared with the results directly calculated with mathematical formula and both results match. The chart, which contains the classical PV curve and the flow of tangent vectors during the CPF iterations, was developed based on the results obtained. In addition to the classic Power-Voltage (PV) curve, this serves as an additional indicator for ensuring voltage stability of the examined system. This enables easier comparison with the effectiveness of different measures regarding the improvement of the voltage stability of the examined system.

5. REFERENCES:

[1] A. Bislimi, "Influence of voltage stability problems on the safety of electrical energy networks", Institute for Electrical Systems and Energy Economics, Vienna University of Technology, Austria, PhD Thesis, 2012.

[2] V. Ajjarapy, C. Christy, "The Continuation Power Flow: A Tool for steady state Voltage stability analysis", IEEE Transactions on Power Systems, Vol. 7, No. 1 ,1992, pp. 416-423.

[3] V. Ajjarapy, "Computation Techniques for Voltage Stability Assessment and Control", Springer, New York, 2006.

[4] J. J. Sanchez Gasca, J. J. Paserba D`Aquila, W. W Price, D. B. Klapper, I. P. Hu, "Exteended-term dynamic simulation using variable time step integration", IEEE Computer Applications in Power, Vol. 6, 1993, pp. 23-28.

[5] C. W. Taylor, "Power System Voltage Stability", McGraw-Hill, Inc., 1994.

[6] P. Kundur, "Power System Stability and Control", New York: McGraw-Hill, 1994.

[7] H. Saadat, "Power system analysis", McGraw-Hill, Singapore, 1999.

[8] T. V. Cutsem, C. Vournas," Voltage stability of electric power systems", Kluwer Academic Publishers, Boston, USA, 1998.

[9] M. Z. Laton, I. Musirin, T. K. Abdul Rahman, "Voltage Stability Assessment via Continuation Power Flow Method", Journal of Electrical & Electronic Systems Research, Vol. 1, 2008, pp. 71-78.

[10] M. Singh Rawat, S. Vadhera, "Voltage Stability Assessment Techniques for Modern Power Systems", Novel Advancements in Electrical Power Planning and Performance, IGI Global, 2020, pp.128-176.

[11] W. Yi, "Voltage Instability Analysis Using P-V or Q-V Analysis", Arizona State University, AZ, USA, Master thesis, 2017.

[12] M. J. Karki, "Methods for online voltage stability monitoring", Iowa State University, IA, USA, Master thesis, 2009.

[13] P. S. Nirbhavane, L. Corson, S. M. Hur Rizvi, A. K. Srivastav, "Three-phase Continuation Power Flow Tool for Voltage Stability Assessment of Distribution Networks with Distributed Energy Resources", IEEE Transactions on Industry Applications, Vol. 57, No. 5, 2021, pp. 5425-5436.

[14] E. M. Magalhães, A. B. Neto, D. A. Alves, "A Parameterization Technique for the Continuation Power Flow Developed from the Analysis of Power Flow Curves", Mathematical Problems in Engineering, Vol. 2012, 2012.

[15] A. Bonini Neto, D. A. Alves, "An Efficient Geometric Parametrization Technique for the Continuation Power Flow through the Tangent Predictor", Trends in Computational and Applied Mathematics, Vol. 9, No. 2, 2008, pp. 185-194.

[16] L. V. Dai, N. M. Khoa, L. C. Quyen, "An Innovatory Method Based on Continuation Power Flow to Analyze Power System Voltage Stability with Distributed Generation Penetration", Complexity, Vol. 2020, 2020.

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

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