Shahid Rajaee
Teacher Training University

Journal of
Electrical and Computer Engineering Innovations (JECEI)

Vol. 13 No. 2, Summer-Fall 2025

Journal of Electrical and Computer Engineering Innovations (JECEI) Vol. 13 No. 2 , Summer- Fall 2025

JECEI

Shahid Rajaee
Teacher Training University

*p*ISSN 2322-3952
*e*ISSN 2345-3044

Journal of
Electrical and Computer Engineering Innovations (JECEI)

Semiannual Publication

Volume 13, Issue 2, Summer-Fall 2025

Journal of Electrical and Computer Engineering Innovations
http://jecei.sru.ac.ir
EISSN: 2345-3044
ISSN: 2322-3952
JECEI
SRTTU
JECEI

License Holder: Shahid Rajaee Teacher Training University (SRTTU). **Address:** Lavizan, 16788-15811, Tehran, Iran.

Journal of Electrical and Computer Engineering Innovations

Vol. 13; Issue 2: 2025

Contents

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Paying Attention to the Features Extracted from the Image to Person Re-Identification

S. H. Zahiri* *, R. Iranpoor, N. Mehrshad

Department of Electrical Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 01 July 2024
Reviewed 29 July 2024
Revised 26 September 2024
Accepted 10 October 2024

Keywords:
Person re-identification
Deep learning
Image processing
Convolutional neural network
Computer vision
Image detection

*Corresponding Author's Email Address:
hzahiri@birjand.ac.ir

Abstract

Background and Objectives: Person re-identification is an important application in computer vision, enabling the recognition of individuals across non-overlapping camera views. However, the large number of pedestrians with varying appearances, poses, and environmental conditions makes this task particularly challenging. To address these challenges, various learning approaches have been employed. Achieving a balance between speed and accuracy is a key focus of this research. Recently introduced transformer-based models have made significant strides in machine vision, though they have limitations in terms of time and input data. This research aims to balance these models by reducing the input information, focusing attention solely on features extracted from a convolutional neural network model.

Methods: This research integrates convolutional neural network (CNN) and Transformer architectures. A CNN extracts important features of a person in an image, and these features are then processed by the attention mechanism in a Transformer model. The primary objective of this work is to enhance computational speed and accuracy in Transformer architectures.

Results: The results obtained demonstrate an improvement in the performance of the architectures under consistent conditions. In summary, for the Market-1501 dataset, the mAP metric increased from approximately 30% in the downsized Transformer model to around 74% after applying the desired modifications. Similarly, the Rank-1 metric improved from 48% to approximately 89%.

Conclusion: Indeed, although it still has limitations compared to larger Transformer models, the downsized Transformer architecture has proven to be much more computationally efficient. Applying similar modifications to larger models could also yield positive effects. Balancing computational costs while improving detection accuracy remains a relative goal, dependent on specific domains and priorities. Choosing the appropriate method may emphasize one aspect over another.

Introduction

Recent advancements in deep learning techniques, coupled with increased computational power, have significantly improve the challenges associated with object identification and recognition in images. Object and person detection remain critical issues within the field of computer vision.

While object recognition is intuitive for humans (even a few-month-old child can recognize common objects), teaching computers to achieve the same level of proficiency has been a formidable challenge until the past decade [1]. The resurgence of Convolutional Neural Networks (CNNs) and deep learning for image classification has revolutionized visual perception. The

adoption of CNNs in the large-scale ImageNet Visual Recognition Challenge (ILSVRC) in 2012 by AlexNet [2] inspired further research on its applications in computer vision. Today, object detection finds applications in self-driving cars, identity verification, security, and medical contexts. In recent years, exponential growth in this field has occurred due to rapid advancements in tools and techniques.

Fig. 1: Sample images of datasets used
(a) Market-1501 dataset (b) DukeMTMC dataset (c) MSMT17 dataset.

As the subsequent challenge following general object and person detection in a scene, the problem of person re-identification (ReID) emerges. Due to its diverse applications across various domains, ReID has garnered significant attention. It serves as a fundamental and essential function in intelligent surveillance systems. The task of connecting individuals across different cameras in various locations and time frames is crucial for network-based surveillance systems. ReID is recognized as the problem of identifying individuals and forms the basis for many other important applications [3].

Current research efforts to address the ReID challenge primarily focus on two aspects: feature learning and metric learning.

Feature Learning: Developing feature representations that remain discriminative for identity while being invariant to viewpoint and lighting conditions [4].

Metric Learning: Optimizing the discriminative parameters of a ReID model using machine learning techniques [5].

In some real-world situations or images with certain limitations, the human eye may not be able to identify and distinguish the subject. In such cases, it is possible to proceed by relying on some minor features or states. In these instances, there is no need to focus on the entire image; rather, relying on specific features can be effective in achieving the desired goal more quickly and reliably.

In this work, the goal is to utilize Transformer models. These models have demonstrated significant results across various domains. However, they require large amounts of input data and powerful hardware for training. To address these challenges, we employ a convolutional neural network (CNN) as a feature extractor, using pre-trained models. Subsequently, the extracted features are fed into a smaller Transformer model, enabling attention at the feature level. This approach allows us to increase the input data to the Transformer model and reduce the data volume by leveraging features extracted from an image. Therefore, the combination of CNNs and Transformer-based models forms the foundation of our research in computer vision.

Related Works

Research in this field primarily focuses on person re-identification and object recognition, with most methods based on Convolutional Neural Networks (CNNs). A desirable and suitable approach for person re-identification involves designing an appropriate loss function for training a backbone CNN (such as ResNet [6]) used for feature extraction from images. Cross-entropy loss [7] and triplet loss [8] are commonly employed in person re-identification.

The IDE model specified in [9] is a global descriptor. For example in [10], the IDE network fine-tuned on the R-CNN model [11] is proved to be more effective than the one fine-tuned directly on an ImageNet pre-trained model. In many cases the IDE model is a commonly used baseline in deep re-ID systems.

Researchers like Luo et al. [12] proposed BNNeck to better combine cross-entropy and triplet loss functions. The main focus of the study by Sun et al. [13] was to obtain superior features using a Part-based Convolutional Neural Network (PCB). In this approach, a convolutional descriptor of the input image captures features related to different parts of the image. An improved method for part extraction (RPP) is introduced [14]. Additionally, another work presents an integrated perspective on cross-entropy and triplet loss functions [15].

Methods such as PCB [13], MGN [16], AlignedReID [17], SAN [18], and others divide the image into multiple parts and extract local features for each part. These fine-grained features are then used for information aggregation.

The Transformer model [19], originally proposed for sequential data in natural language processing (NLP), has

also shown effectiveness in computer vision tasks. Han et al. [20] and Salman et al. [21] have explored the application of Transformers in computer vision. The Vision Transformer (ViT) [22] directly utilizes pure Transformers on image patches. However, ViT requires large-scale pre-training data. To address this limitation, the DeiT framework introduces a teacher-student strategy specifically for Transformers to accelerate ViT training without the need for large-scale pre-training data [23].

He et al. [24] proposed a pure transformer-based framework for person ReID named TransReID. Specifically, they first encode an image as a sequence of patches and build a transformer-based strong baseline with a few critical improvements, which achieves competitive results on several ReID benchmarks with CNN-based methods.

Methods

Feature extraction and attention to the extracted features are fundamental to this research. To evaluate the implemented models, we must first examine the commonly used datasets in this field. Research in deep learning technology heavily relies on substantial amounts of data for model training. Subsequently, we will investigate the prevalent backbone architectures.

Datasets

To develop robust person re-identification models, it is essential to have datasets with diverse backgrounds, occlusions, and overlapping bodies [25]. While numerous datasets are available for research, some, such as VIPeR [26], GRID [27], and CUHK01 [28], are limited by the number of individuals and the small number of images per person. These datasets often rely on manual labeling methods for person identification. With the advancement of deep learning, smaller datasets are no longer sufficient for training needs. Consequently, large-scale datasets such as CUHK03 [29], Market1501 [30], DukeMTMC [31], and MSMT17 [32] have been proposed and accepted.

The Market-1501 dataset is one of the most well-known datasets for person detection and identification in images. It comprises 1501 different individuals captured in an outdoor environment, with approximately 32,217 images collected from 6 surveillance cameras equipped with various sensors. Each individual has around 6 to 20 full-body images. The dataset is divided into training and evaluation (query and gallery) sets. The training set includes images of the first 751 individuals, with approximately 12 images per person. The test set contains another 750 individuals, each having only one query image and around 4 to 18 gallery images.

The MSMT17 dataset consists of images from a 15-camera network deployed on a university campus. The camera network includes 12 outdoor cameras and 3 indoor cameras. Video collection was conducted over four days with varying weather conditions. For each day, 3 hours of footage were captured, focusing on pedestrian detection and annotation during morning, noon, and afternoon. The final raw video dataset comprises 180 hours of footage, 12 outdoor cameras, 3 indoor cameras, and 12 temporal gaps. Faster RCNN [33] was used for pedestrian bounding box detection, resulting in 126,441 annotated bounding boxes from 4,101 identities. Sample images from all three datasets are shown in Fig. 1.

The DukeMTMC dataset consists of over 2 million frames and more than 2,700 individuals. It includes eight videos, each lasting 85 minutes, recorded at 1080p quality with a frame rate of 60 frames per second. The videos were captured by eight fixed cameras placed around the Duke University campus during periods of heavy pedestrian foot traffic. Calibration data was used to determine homography between images and ground level.

Table 1: Specifications of the datasets used provides information about the datasets used, including the number of training images, the count of individuals in the training set, and the number of evaluation images, which includes query and gallery images.

Table 1: Specifications of the datasets used

Dataset	Train image	Quary image	Gallery image	Num train ID	Num Cam
Market-1501	12936	3368	15913	751	6
DukeMTMC	16522	2228	17661	702	8
MSMT17	32621	11659	82161	1041	15

Backbone Networks

Backbone networks function as initial feature extractors for object recognition and person re-identification tasks. These networks process images as input and generate corresponding feature maps as output. Most of these architectures, originally designed for object detection, typically exclude fully connected layers. Additionally, advanced versions of classification networks are available.

Considering the diverse requirements for accuracy and efficiency, individuals can opt for deeper and more compact architectures such as ResNet [34], ResNeXt [35], or lightweight networks like MobileNet [36], ShuffleNet [37], SqueezeNet [38], Xception [39], MobileNetV2 [40], and MobileNetV3 [41]. When targeting mobile devices, lightweight networks effectively meet the necessary criteria.

MobileNetV1 [36] introduced depthwise separable convolutions as an efficient replacement for traditional

convolution layers.Depthwise separable convolutions are defined by two separate layers: light weight depthwise convolution for spatial filtering and heavier 1x1 pointwise convolutions for feature generation. This method effectively factorize traditional convolution by separating spatial filtering from the feature generation mechanism.

MobileNetV2 [40] introduced the linear bottleneck and inverted residual structure in order to make even more efficient layer structures by leveraging the low rank nature of the problem. MobileNetV3 [41] use a combination of these layers as building blocks in order to build the most effective models. Layers are also upgraded with modified swish nonlinearities. Both squeeze and excitation as well as the swish nonlinearity use the sigmoid which can be inefficient to compute as well challenging to maintain accuracy in fixed point arithmetic so it is replaced by the hard sigmoid.

The nonlinearity is defined as

$$swish\ x = x\ .\sigma(x) \tag{1}$$

A nonlinearity called swish was introduced that when used as a drop-in replacement for ReLU, that significantly improves the accuracy of neural networks [42]-[44].

In MobileNetV3 the sigmoid function has been replaced with its piece-wise linear hard analog. The difference is in use ReLU6 rather than a custom clipping constant. Similarly, the hard version of swish becomes

$$h - swish\ [x] = x\ \frac{ReLU6(x+3)}{6} \tag{2}$$

MobileNetV3 is defined as two models: MobileNetV3-Large and MobileNetV3-Small. These models are targeted at high and low resource use cases respectively. In this work, according to Fig. 2, in backbone section we use both MobileNetV3 versions for feature extraction.

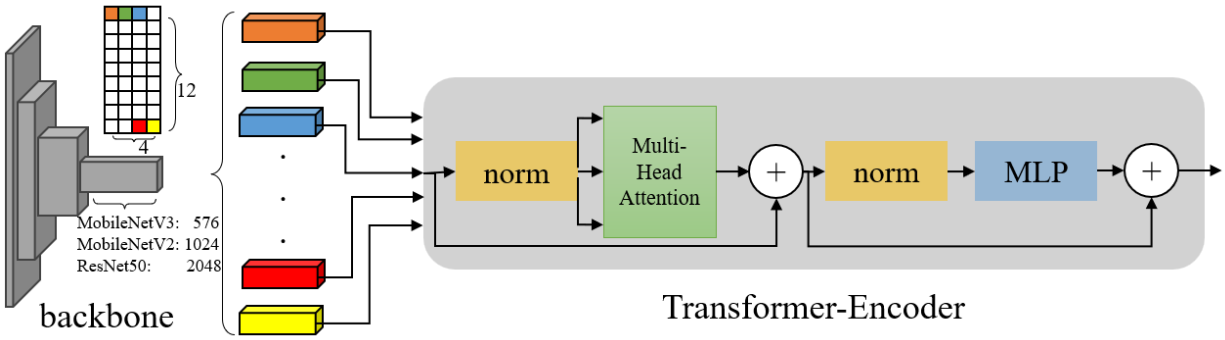


Fig. 2: Attention to the features extracted from the image in the proposed method.

Fig. 3: illustrates the output heatmap of the backbone network, serving as an example of feature extraction in the backbone architecture.

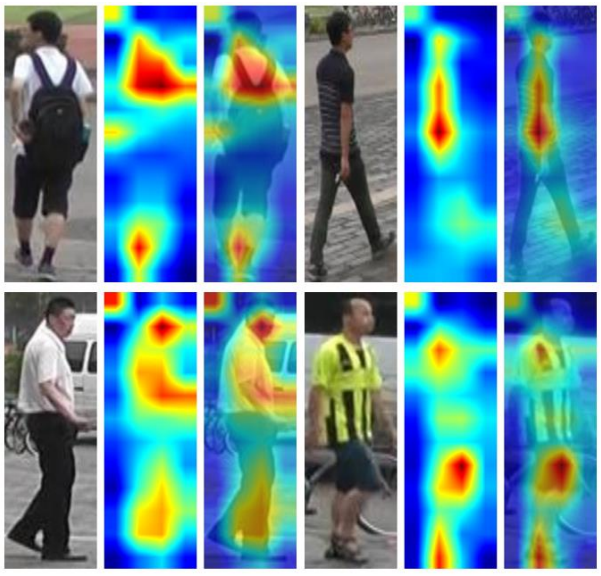


Fig. 3: MobileNetV3 extracted heatmap from the Market-1501 dataset.

Proposed Methode

In this section, the focus is on exploring efforts in person re-identification (ReID) using transformer-based methods. One of the challenges faced by this approach is its high computational cost. Therefore, achieving a balance between computational overhead and task accuracy is the primary goal. In transformer-based architectures, input images are divided into smaller patches, each treated as a patch and fed into the main network. For instance, in person re-identification, if we consider an input image with dimensions of 128x384 and a patch extraction window of 16x16, a total of 8x24 patches will be separated for input to the main network. Subsequently, 192 patches are formed, each with dimensions of 3x16x16, resulting in an array of 768x192.

The objective in this approach is to reduce the input information to the main network. Many details within an image, such as backgrounds, do not require attention mechanisms and are essentially unused information. Since networks like ViT attend to all input information, this increases computational costs. Pure transformer-based models (e.g. ViT, DeiT) split the images into non-overlapping patches, losing local neighboring structures

around the patches.

In this job, the extracted features are used instead of the non-Overlapping patches [24]. As depicted in Fig. 2, the input image is first fed into a column-based network. Depending on the functionality and complexity, various architectures can be used. For the initial step and for speed enhancement, a downsized MobileNetV3 model is employed. The output of this model, after the main layers, is a tensor of size 4x12x576. Thus, ultimately, an array of 48x576 is transferred to the main Transformer network.

In the self-attention layer, the input vector is first transformed into three different vectors: the query vector q, the key vector k. Vectors derived from different inputs are then packed together into three different matrices, namely, Q, K and V [20].

Subsequently, the attention function between different input vectors is computed by calculating scores between them using the formula $Sn = Q.K^T$. These scores are then normalized to ensure gradient stability. Finally, the softmax function is applied to translate the scores into probabilities, resulting in the weighted value matrix.

The process can be unified into a single function:

$$Attention(Q, K, V) = Softmax\left(\frac{Q.K^T}{\sqrt{d_k}}\right).V \quad (3)$$

Multi-head attention is a mechanism that used to boost the performance of the self-attention layer.

There is a residual connection to each sub -layer in the encoder and decoded. A layer-normalization [45] is followed after the residual connection. The output of these operations can be described as:

$$NormLayer(X + Attention(X)) \quad (4)$$

Here, X is used as the input of Multi Head Attention layer [20]. The output is passed through several MLP layers to get the final answer in the form of a feature vector.

During training, cross-entropy loss and the Adam optimizer are used. The execution time for each batch is also reported in the table. For comparison, Fig. 4 presents the visual output. A single query image is selected from the query set, and other images from the same class or similar images are chosen. Finally, it is determined which image belongs to the original image class. This distinction is indicated by green and red colors.

Fig. 4: The example of the final output of the architecture that receives an image from the query set and confirms the match of the person from the set of similar people in the gallery set.

After implementing the specified modifications, the experimental results are summarized in Table 2, reporting the findings related to the Market-1501, DukeMTMC, and MSMT17 datasets. The IDE and PCB-U methods are based on convolutional neural networks. The ViT method is a transformer architecture previously explained, with a patch size and a stride of 16. It has an embedding dimension of 768, a depth and number of heads of 12, and four MLP layers. The reported values for these methods are from reference articles. However, the ViT-Small and DiT-Small methods are smaller transformer models. ViT-Small has an embedding dimension of 768, a depth and number of heads of 8, and three MLP layers. DiT-Small has an embedding dimension of 384, a depth of 12, and a number of heads of 8, with four MLP layers. Additionally, the proposed M-ViT-S model uses a reduced MobileNetV3, and the M-ViT-L model uses a larger version of MobileNetV3 for feature extraction before the reduced transformer model.

The training process included a batch size of 32, and all four mentioned methods were trained under the same conditions for 100 epochs. Results for other methods are also reported, but these models were not trained under the previous fixed conditions, and their results are reported.

The reported values for the IDE, PCB-U, and ViT methods are sourced from relevant references in the table. However, other methods have also been evaluated under identical conditions. Notably, execution time for each image category is a critical consideration, particularly for real-time system performance in hardware-constrained environments. Among the methods, ViT-Small and DiT-Small exhibit the shortest execution time. However, their mAP and subsequent Rank-1 performance significantly decrease. In contrast, CNN-based methods experience increased execution time due to their more complex backbone architectures. The ViT method maintains good accuracy but at the cost of

longer execution time. Remarkably, the M-ViT-Small and M-ViT-Large approaches achieve accuracy comparable to CNN networks while reducing computation time.

Table 2: Comparison of methods for market1501 and cuhk03 datasets

Methods	Dataset	mAP	Rank-1	Time Pre Batch
IDE [9]	Market-1501	68.5	85.3	-
	DukeMTMC	52.8	72.4	
	MSMT17	-	-	
PCB-U [14]	Market-1501	77.4	92.3	0.420
	DukeMTMC	68.8	82.6	
	MSMT17	-	-	
ViT [24]	Market-1501	89.5	95.2	0.434
	DukeMTMC	82.6	90.7	
	MSMT17	69.4	86.2	
ViT-Small	Market-1501	25.6	42.2	0.218
	DukeMTMC	21.4	30.9	
	MSMT17	17.3	24.8	
DiT-Small	Market-1501	30.1	48.6	0.180
	DukeMTMC	22.5	32.5	
	MSMT17	18.6	26.2	
M-ViT-Small	Market-1501	72.1	88.0	0.234
	DukeMTMC	61.2	78.7	
	MSMT17	56.6	72.1	
M-ViT-Large	Market-1501	74.5	89.3	0.331
	DukeMTMC	64.0	79.9	
	MSMT17	58.1	74.7	

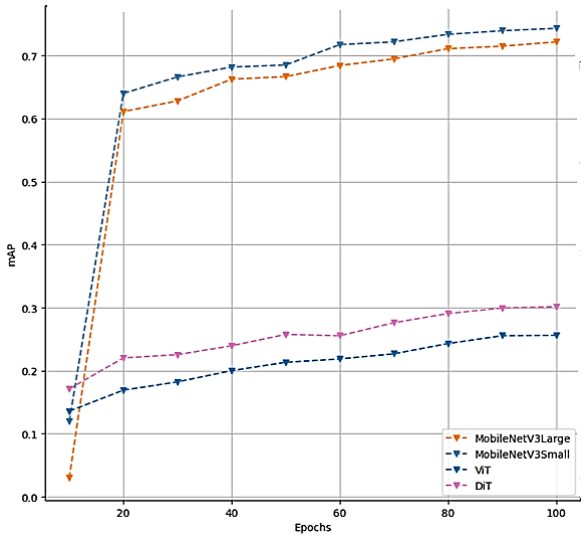


Fig. 5: mAP diagram of evaluation data.

Fig. 5: and Fig. 6: illustrates the network training process, showing the mAP and Rank-1 evaluation data for each stage.

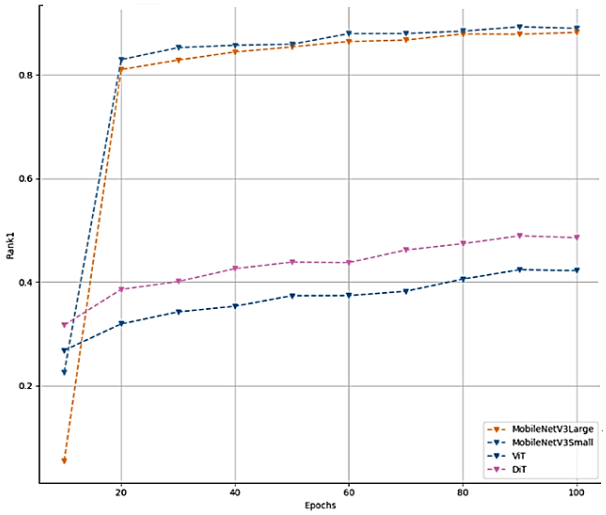


Fig. 6: Rank1 diagram of evaluation data.

Conclusion

The problem of person re-identification presents a multifaceted challenge with practical applications in our daily lives. Striking a balance between speed and accuracy is crucial when implementing a specific model for this purpose. The domain should be flexible enough to allow for the use of different methods based on specific needs.

Among other concerns in this field, data limitations pose significant challenges compared to other computer vision domains. Given the importance of input data in deep learning, tasks such as classification and identification involve extensive datasets. However, the person re-identification (ReID) problem is relatively limited in terms of available data. Another significant challenge is data quality. Since these data are often collected by cameras that lack optimal quality, methods capable of establishing meaningful connections between image components are of special importance.

By modifying existing models and leveraging essential features of each approach, the results indicate improved accuracy when using transformer models. Processing speed remains relatively unchanged due to the reduction in input information. The proposed method serves as an initial step in combining convolutional neural network (CNN) models with transformers, enhancing computational efficiency. Subsequent steps will focus on further improving models to achieve optimal results.

The adoption of newer transformer models with better computational speed, alongside more precise feature extraction, paves the way for future advancements. Data augmentation and related techniques can also contribute to enhancing performance in upcoming research areas. Furthermore, changes in models and the use of techniques to achieve faster and more appropriate responses are among important aspects to consider. Given the hardware limitations in real-time applications, reaching an optimally accurate model is crucial. This can

be achieved by reducing computational complexity in architectures, provided it does not compromise accuracy. Additionally, considering that in larger datasets similar to MSMT17, the accuracy of all models is lower, one could argue the possibility of overfitting in the models. Models may perform well only on a specific dataset and not exhibit suitable performance elsewhere. Given the limited dataset availability, this approach can be used for evaluating models in other computer vision applications, such as semantic segmentation, identification, and more, to ensure the introduced model has more precise and efficient performance.

In summary, this work aims to establish an interactive relationship between hardware constraints and sufficient accuracy. While many existing methods may achieve higher accuracy, the trade-off with increased computational demands is inevitable. This study seeks to explore specific adjustments and implementations to improve results while considering this balance between accuracy and computational cost.

Author Contributions

Dr. Zahiri has drawn the general road map. R .Iranpoor has searched for important articles in this field. Then, by checking the results and collecting the necessary data, the implementation of the proposed method has been done. Dr. Mehrshad reviewed the results and made changes in the way of implementation and final editing of the work.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

ReID	Person re-Identification
CNN	Convolutional Neural Network
ViT	Vision Transformer
DeiT	Data-efficient image transformer
PCB	Part-based Convolutional Baseline
RPP	Random Partitioning Pooling
mAP	mean Average Precision
CMC	Cumulative Matching Characteristics

References

[1] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, B. Lee, "A survey of modern deep learning based object detection models," Digital Signal Process., 126: 103514, 2022.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, 25(2), 2012.

[3] W. Wei, W. Yang, E. Zuo, Y. Qian, L. Wang, "Person re-identification based on deep learning—An overview," J. Visual Commun. Image Represent., 82: 103418, 2022.

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in Proc. 2010 IEEE computer society conference on computer vision and pattern recognition: 2360-2367, 2010.

[5] W. S. Zheng, S. Gong, T. Xiang, "Person re-identification by probabilistic relative distance comparison," in Proc. CVPR 2011: 649-656, 2011.

[6] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proc. the IEEE conference on computer vision and pattern recognition: 770-778, 2016.

[7] Z. Zheng, L. Zheng, Y. Yang, "A discriminatively learned cnn embedding for person reidentification," ACM Trans. Multimedia Comput. Commun. Appl. (TOMM), 14(1): 1-20, 2017.

[8] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, "End-to-end comparative attention networks for person re-identification," IEEE Trans. Image Process., 26(7): 3492-3506, 2017.

[9] L. Zheng, Y. Yang, A. G. Hauptmann, "Person re-identification: Past, present and future," arXiv preprint arXiv:1610.02984, 2016.

[10] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, "Person re-identification in the wild," in Proc. the IEEE conference on computer vision and pattern recognition: 1367-1376, 2017.

[11] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. the IEEE conference on computer vision and pattern recognition: 580-587, 2014.

[12] H. Luo et al., "A strong baseline and batch normalization neck for deep person re-identification," IEEE Trans. Multimedia, 22(10): 2597-2609, 2019.

[13] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in Proc. the European conference on computer vision (ECCV): 480-496, 2018.

[14] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, S. Wang, "Learning part-based convolutional features for person re-identification," IEEE Trans. Pattern Anal. Mach. Intell., 43(3): 902-917, 2019.

[15] Y. Sun et al., "Circle loss: A unified perspective of pair similarity optimization," in Proc. the IEEE/CVF conference on computer vision and pattern recognition: 6398-6407, 2020.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in Proc. the 26th ACM international conference on Multimedia : 274-282, 2018.

[17] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," Pattern Recognit., 94: 53-61, 2019.

[18] J. Qian, W. Jiang, H. Luo, H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," Meas. Sci. Technol., 31(9): 095401, 2020.

[19] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., 30, 2017.

[20] K. Han et al., "A survey on visual transformer," arXiv preprint arXiv:2012.12556, 2020.

[21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, "Transformers in vision: A survey," ACM Comput. Surv. (CSUR), 54(10s): 1-41, 2022.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, "Training data-efficient image transformers & distillation through attention," in Proc. International Conference on Machine Learning: 10347-10357, 2021.

[24] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, "Transreid: Transformer-based object re-identification," in Proc. the IEEE/CVF International Conference on Computer Vision: 15013-15022, 2021.

[25] D. Wu et al., "Deep learning-based methods for person re-identification: A comprehensive review," Neurocomputing, 337: 354-371, 2019.

[26] D. Gray, H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in Proc. 10th European Conference on Computer Vision, Part I 10: 262-275, 2008.

[27] C. C. Loy, T. Xiang, S. Gong, "Multi-camera activity correlation analysis," in Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition: 1988-1995, 2009.

[28] W. Li, R. Zhao, X. Wang, "Human reidentification with transferred metric learning," in Proc. 11th Asian Conference on Computer Vision, Part I 11: 31-44, 2013.

[29] W. Li, R. Zhao, T. Xiao, X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proc. IEEE Conf. Computer Vision and Pattern Recognition: 152-159, 2014.

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE International Conference on Computer Vision: 1116-1124, 2015.

[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. European Conference on Computer Vision: 17-35, 2016.

[32] L. Wei, S. Zhang, W. Gao, Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 79-88, 2018.

[33] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., 39(6): 1137-1149, 2017.

[34] S. Targ, D. Almeida, K. Lyman, "Resnet in resnet: Generalizing residual architectures," arXiv preprint arXiv:1603.08029, 2016.

[35] S. Xie, R. Girshick et al., "Aggregated residual transformations for deep neural networks," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 1492-1500, 2017.

[36] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[37] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in Proc. Artificial Intelligence Applications and Innovations: 97-108, 2018.

[38] F. N. Iandola, S. Han, M. W. Moskewicz et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," arXiv preprint arXiv:1602.07360, 2016.

[39] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 1251-1258, 2017.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 4510-4520, 2018.

[41] A. Howard et al., "Searching for mobilenetv3," in Proc. IEEE/CVF International Conference on Computer Vision: 1314-1324, 2019.

[42] S. Elfwing, E. Uchibe, K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," Neural networks, 107: 3-11, 2018.

[43] Y. Guo, D. Zhou, W. Li, J. Cao, "Deep multi-scale Gaussian residual networks for contextual-aware translation initiation site recognition," Expert Syst. Appl., 207: 118004, 2022.

[44] P. Ramachandran, B. Zoph, Q. V. Le, "Searching for activation functions," arXiv preprint arXiv:1710.05941, 2017.

[45] J. L. Ba, J. R. Kiros, G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.

Biographies

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Rasool Iranpoor was born on July 19, 1991. He received M.Sc. degree in Electronic Engineering from Birjand University, Birjand, Iran, in 2018. He is currently a Ph.D. student at Birjand University to receive a Ph.D. degree in Electronics Engineering. His research interests include Machine Learning, Image Processing, Computer Vision, and Deep Learning Algorithms.

- Email: rasool.iranpoor@birjand.ac.ir
- ORCID: 0000-0002-7769-259X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Nasser Mehrshad received the B.Sc. degree from Ferdowsi University of Mashhad in 1994. He completed his M.Sc. degree in Biomedical Electronics Engineering at Tarbiat Modares University in 1998 and received his Ph.D. in the same field in 2005. Currently, he serves as a full-time faculty member in the Department of Electrical and Electronic Engineering at the University of Birjand. His research interests include machine vision, digital signal processing, and biomedical engineering.

- Email: nmehrshad@birjand.ac.ir
- ORCID: 0000-0001-8678-3402
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
S. H. Zahiri, R. Iranpoor, N. Meheshad, "Paying attention to the features extracted from the image to person re-identification," J. Electr. Comput. Eng. Innovations, 13(2): 267-274, 2025.

DOI: 10.22061/jecei.2024.10968.752

URL: https://jecei.sru.ac.ir/article_2206.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Edge User Performance Improvement by Intelligent Reflecting Surface-Assisted NOMA System

F. Rahdari [1,*], ***M. Sheikh-Hosseini*** [1], ***M. Jamshidi*** [2]

[1]*Department of Computer and Information Technology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced, Technology, Kerman, Iran.*

[2]*Department of Applied Mathematics, Graduate University of Advanced Technology, Kerman, Iran.*

Article Info

Article History:
Received 17 August 2024
Reviewed 23 September 2024
Revised 29 November 2024
Accepted 02 December 2024

Keywords:
Intelligent Reflecting Surface (IRS)
Non-Orthogonal Multiple Access (NOMA)
Channel gain difference
Edge users
Optimization problem

*Corresponding Author's Email Address: *f.rahdari@kgut.ac.ir*

Abstract

Background and Objectives: This research addresses the performance drop of edge users in downlink non-orthogonal multiple access (NOMA) systems. The challenging issue is paring the users, which becomes more critical in the case of edge users due to poor signal quality as well as the similarity of users' channel gains.

Methods: To study this issue, the capabilities of intelligent reflecting surface (IRS) technology are investigated to enhance system performance by modifying the propagation environment through intelligent adjusting of the IRS components. In doing so, an optimization problem is formulated to determine the optimal user powers and phase shifts of IRS elements. The objective is to maximize the system sum rate by considering the channel gain difference constraint. Additionally, the study addresses the effect of the IRS location in the cell on system performance.

Results: The proposed approach is evaluated for various scenarios and compared with benchmarks in terms of average bit error rate (BER) and sum rate. The numerical results show that IRS-assisted NOMA improves the performance of edge users and distributes resources more fairly than conventional NOMA.

Conclusion: Simulation results demonstrate that using IRS-assisted NOMA can effectively address the issue of edge users. By modifying the channel between the BS and the edge users using IRS, the channel gain difference of the users is increased, thereby enhancing the overall system performance. Particularly, the proposed IRS-NOMA system offers a gain of about 4 dB at a BER of 10^{-2} and 3 dB at the sum rate of 10^{-1} bps/Hz compared to conventional NOMA. In addition, it was observed that the location of the IRS in the cell affects the system's performance.

Introduction

As a promising technology for 5G networks, non-orthogonal multiple access (NOMA) enables different users to multiplex signals on a shared channel within a particular domain [1]-[3]. In power-domain downlink NOMA, the base station (BS) performs superposition coding (SC) by allocating more power to the far (weak) user and less power to the near (strong) user.

Signal detection is then carried out on the user equipment (UE). In this process, the far user considers the signal of the near user as interference plus noise, while the near user uses successive interference cancellation (SIC) to decode its signal. The NOMA results in higher spectral efficiency, reduced latency, improved user fairness, and enhanced connectivity compared to the orthogonal multiple access (OMA) scheme [4], [5]. In the

NOMA system, user pairing is based on the difference between the users' channel gains. The system performance is improved when there is a significant difference between the channel gains of paired users. Hence, it is preferred to pair the near users with far users [6]. A challenging issue is associated with the paring of the middle users which causes a degradation in the performance of the NOMA system [7], [8].

The importance of this issue is heightened for edge users for two main reasons. The first reason is poor signal quality and high attenuation due to the long distance from the BS, and the second one is related to low performance due to the similarity of users' channel conditions. The approach presented in this work centers on leveraging an emerging technology for 6G known as intelligent reflecting surface (IRS) to overcome the challenge. IRS is a surface of numerous reflecting components that can be fine-tuned to steer electromagnetic waves in specific directions. This capability leads to improved coverage, higher data rates, and superior energy efficiency. As an enabler of the 6G wireless communication system, the seamless integration of IRS with other emerging technologies yields a more efficient, secure, and intelligent wireless network [9], [10]. In this way, IRS-NOMA is used in the present study to leverage the capabilities of IRS and NOMA simultaneously. In this study, we are using IRS-NOMA to take advantage of the capabilities of IRS and NOMA simultaneously. The IRS-NOMA is expected to play an essential role in developing future wireless communication systems by improving spectral efficiency, enabling more connected devices, and reducing energy consumption [11]-[14].

Numerous research works have been conducted on the benefits of IRS in conjunction with NOMA, indicating an increasing focus on IRS-NOMA applications in wireless communication. Paper [15] introduces the phase shift estimation in IRS-assisted systems under correlated Rayleigh fading. The study covers two scenarios including fully-active-IRS and hybrid-IRS, considering the number of active IRS elements. The paper presents the derivation of the maximum likelihood estimator for the channel phases of IRS elements based on the observations of active IRS elements. In [16], a low-complexity resource allocation method is introduced to maximize sum throughput by jointly optimizing phase shifts and time allocation. It first derives the problem with phase shifts as variables, and then deduces the optimization problem for the downlink wireless energy transmission process.

Authors in [17] examine the tradeoff between spectral efficiency (SE) and energy efficiency (EE) in the IRS-assisted MISO cognitive radio networks (CRN)-NOMA system. It is conducted by formulating a multi-objective optimization problem under perfect and imperfect channel state information (CSI) scenarios. An iterative block coordinate descent (BCD) algorithm is utilized to optimize the beamforming design and IRS phase shifts. The paper [18] focuses on the IRS-assisted multi-carrier (MC)-NOMA system to accommodate users in each channel and allocate the available IRS units to the respective channels. The work formulates a power minimization problem to minimize the overall transmit power while meeting the quality of service (QoS) requirements.

The IRS-NOMA network performance with two users at the border is examined in [19]. An analysis of the channel statistics of the BS-IRS-UE with Nakagami-m fading distribution is conducted and the closed-form expressions for the ergodic rate, outage probability, and approximate ratio of UEs' ergodic rate to the SINR are derived for both low and high values. Letter [20] presents an optimization algorithm in downlink IRS-assisted NOMA systems to design active and passive beamforming for the BS and IRS. The goal is to solve the interference issue among users and decrease power consumption, considering the user's QoS. The subproblems of power minimization and phase shift feasibility are defined and solved iteratively using alternating optimization.

The work [21] derives analytical expressions to evaluate the Ergodic capacity of the IRS-NOMA over the Nakagami-m faded channel by considering inter-cell interference, imperfect-CSI, and SIC. Paper [22] addresses the issue of energy efficiency optimization in IRS-assisted NOMA systems by handling passive beamforming, user clustering, and power allocation problems. The passive beamforming is tackled using the univariate search technique, the user clustering is addressed using a matching algorithm, and the power allocation is optimized using the difference of convex (DC) programming.

Most of the mentioned works focus on investigating the efficiency of IRS-NOMA with far and near users. The focus of this study is on edge users with close channel conditions. The study explores leveraging IRS capabilities to modify the transmission environment and improve system performance. In this regard, an optimization problem is presented to intelligently adjust the IRS elements' phase shifts considering the channel gain difference constraint. The system performance is evaluated under different channel conditions in terms of sum rate and average bit error rate (BER). The main contributions of the work are described as follows:

- To improve the performance of edge users in NOMA systems, IRS technology is used to modify the propagation environment by adjusting the reflection elements, thereby enhancing overall system performance.
- With considering the user channel gain conditions, a

new approach has been introduced to intelligently control IRS and allocate the power resources to edge users.
- An optimization problem is formulated to determine the optimal phase shifts of IRS elements and NOMA user powers with the aim of maximizing the system sum rate.
- The effect of the IRS location in the cell on system performance has been studied through simulation, taking into account the IRS distance from the BS and edge users.
- The performance of the IRS-assisted NOMA approach is examined in various scenarios, considering BER and sum rate as evaluation metrics.

The rest of this paper is organized as follows. Section 2 introduces the concept of the IRS-NOMA system. Section 3 presents the proposed approach for IRS control and formulates the IRS-NOMA optimization problem. Section 4 includes the simulation system setup and evaluate the proposed approach through extensive simulation tests. Finally, Section 5 presents the conclusion.

Notations: in respective order, non-boldface, boldfaced lowercase, and boldfaced uppercase letters represent the scalar, vector, and matrix.

IRS-assisted NOMA Communication System

The IRS-NOMA scheme combines IRS technology with NOMA to enhance wireless communication. The IRS involves deploying a planar array of passive reflecting elements to improve wireless communication systems' performance. It manipulates the reflection and scattering of radio signals, altering the signal's phase and controlling the propagation of electromagnetic waves. By placing the IRS between the transmitter and receiver, it can manipulate the signal to interfere with the direct signal constructively, resulting in a stronger and more reliable signal at the receiver. Additionally, the IRS can help mitigate fading effects and multipath propagation, which causes signal degradation in wireless communication systems [23]. On the other hand, NOMA is a type of multiple-access technique used in wireless communication systems, allowing multiple users to share the same radio frequency resources. Unlike traditional OMA techniques where users are assigned separate channels to avoid interference, NOMA assigns users the same frequency band and time slot but with different power and code assignments. NOMA utilizes superposition coding to encode multiple signals onto the same frequency band and time slot, thereby increasing spectral efficiency. At the receiver, the SIC technique separates individual signals from each user. This involves decoding the strongest signal first and subtracting it from the total received signal to cancel out interference. The process is repeated for each signal until all interference is removed and the original signals can be accurately decoded [24].

The IRS-NOMA offers significant benefits such as improved spectral efficiency, reduced interference, lower power consumption, reduced costs, and greater flexibility for wireless communication [25]. In IRS-NOMA, an IRS is positioned between the transmitter and the receiver to reflect the wireless signal, enhancing signal power for a specific user in the NOMA group. This enables the receiver to separate signals from different users more efficiently, resulting in higher data rates and better QoS. Additionally, IRS-NOMA allows for flexible resource allocation, which can be adjusted to meet the varying QoS requirements of different users [26].

Fig. 1 illustrates the schematic of the IRS-NOMA system with two users. In this setup, a base station (BS) communicates with the users through an IRS with reflecting elements. The signal model involves the signal sent to the IRS by the BS, which is then reflected in a controlled manner. The reflected signal is received by the user and combined with the line of sight (LOS) signal to generate a stronger and more reliable signal [27]. The signal model considers the phase, which can be adjusted by controlling the position and orientation of the reflecting elements in the IRS. By optimizing the reflection of the signal, it is possible to constructively interfere with the direct signal, leading to an enhanced signal for the user.

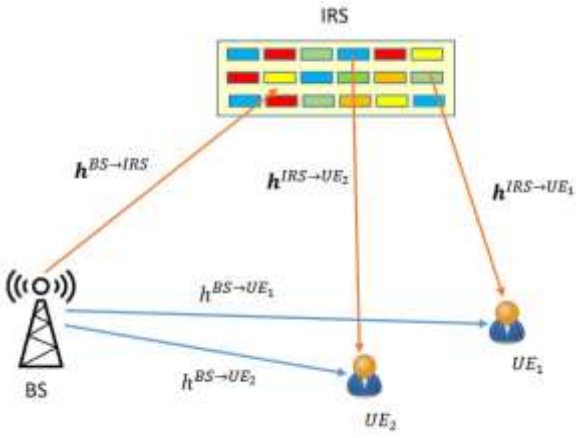


Fig. 1: IRS-NOMA system.

We consider an IRS-NOMA system consisting of a BS, an IRS, and the number of K users. It is assumed that the BS and users are equipped with one antenna for sending and receiving signals. The signal received by the k-th user could be expressed as follows [18]:

$$y_k = \tilde{h}_k s + n = \tilde{h}_k \sum_{k=1}^{K} \sqrt{p_k}\, x_k + n \quad (1)$$

in which $\tilde{h}_k$ is the effective channel between BS and k-th user [18]:

$$\tilde{h}_k = h^{BS\to UE_k} + \boldsymbol{h}^{IRS\to UE_k{}^T} \boldsymbol{\Theta} \boldsymbol{h}^{BS\to IRS} \quad (2)$$

where $h^{BS\to UE_k}$ represents the scalar quantity that models the channel between BS and k-th user. The $N \times 1$ vector $\boldsymbol{h}^{IRS\to UE_k} = [h(1)^{IRS\to UE_k} \dots h(N)^{IRS\to UE_k}]^T$ model the channel between the IRS and k-th user, and $N \times 1$ vector $\boldsymbol{h}^{BS\to IRS} = [h(1)^{BS\to IRS} \dots h(N)^{BS\to IRS}]^T$ denotes the channel between BS and IRS. The $[.]^T$ operation represents the transpose of the vector. All $h^{BS\to UE_k}$, $h(i)^{IRS\to UE_k}$ and $h(i)^{BS\to IRS}$ are defined as $CN(0,1)/\sqrt{1+d^{\alpha}}$ [28], where $CN(0,1)$ stands for a zero mean, complex normal random of variance 1 which models the small scale fading. Also, d and α represent the distance between the transmitter to the receiver terminals of the relevant link and the path loss factor, respectively. In addition, $\boldsymbol{\Theta} = diag[e^{j\theta_1} \quad \dots \quad e^{j\theta_N}]$ stands for an $N \times N$ diagonal matrix that models the IRS consisting of N reflecting elements. In this model, $\theta_i \in [0,2\pi]$ represents the phase shift of the i-th IRS element.

Moreover, $s = \sum_{k=1}^{K} \sqrt{p_k}\, x_k$ [1] denotes the signal transmitted by the BS to the users, where x_k denotes the signal intended for the k-th user, normalized to unit power ($\mathrm{E}[|x_k|^2] = 1$). The term p_k represents the power allocated by the BS for transmission of this user. Typically, more power is assigned to users with poorer channel conditions, resulting in earlier decoding during signal detection. Assuming that $|\tilde{h}_1|^2 < |\tilde{h}_2|^2 < \dots < |\tilde{h}_K|^2$, the BS distributes power among users, ensuring that $p_1 > p_2 > \dots > p_K$. It is important to note that the power allocation must satisfy the constraint $\sum_{k=1}^{K} p_k = P^{MAX}$; where P^{MAX} is the available power at the BS for each group. This approach empowers the users to recover their data symbols successfully. The last term, n, is additive white Gaussian noise (AWGN) with zero mean and variance σ^2. The data rate of the k-th user is calculated as [16]:

$$R_k = \log_2(1+\gamma_k) \tag{3}$$

where γ_k is the signal-to-interference plus noise ratio (SINR) of the k-th user equal to [16]:

$$\gamma_k = \frac{|\tilde{h}_k|^2 p_k}{|\tilde{h}_k|^2 \sum_{i=k+1}^{K} p_i + \sigma^2} \tag{4}$$

IRS-NOMA Optimization Problem

In the NOMA system, the difference in channel gains is a key factor in assigning users to common channels. However, pairing users with close channel conditions causes decoding difficulties, leading to degraded performance, especially in terms of BER. This issue is further exacerbated when edge users are accommodated in the same group in the NOMA scenario [29]. The signal decoding process encounters a high error rate, leading to a decrease in the overall performance of the NOMA system. This challenge becomes more significant as the number of users increases.

To tackle this problem, one approach involves manipulating the reflection of radio signals. This entails altering the signal phase to control the propagation of electromagnetic waves, thereby improving the performance of the users with close channel conditions. To achieve this, the capabilities of the IRS are exploited to control the propagation of electromagnetic waves by manipulating the reflection of radio signals. In this regard, the difference in user channel gains is used to evaluate the pairing quality. The aim is to manipulate the users' channel conditions by adjusting the phase shifts of IRS elements so that the channel gain difference of two users exceeds the threshold value. This section presents an optimization problem for the IRS-NOMA system that involves maximizing the system sum rate while adhering to specific constraints. The goal is to allocate available power to different users and adjust phase shifts of the IRS elements to maximize overall system performance. The optimization problem is formulated while the threshold condition for channel gains is established as a constraint:

$$\max_{\boldsymbol{p},\boldsymbol{\theta}} \sum_{k=1}^{K} \log_2(1+\gamma_k) \tag{5-a}$$

$$s.t. \sum_{k=1}^{K} p_k = P^{MAX} \tag{5-b}$$

$$p_1 > p_2 > \dots > p_K \tag{5-c}$$

$$|\tilde{h}_m|^2 - |\tilde{h}_n|^2 > g_{th}, \quad m,n \in \{1,\dots,K\}, m \neq n \tag{5-d}$$

$$\theta_i \in [0,2\pi], \forall i \in \{1,\dots,N\} \tag{5-e}$$

where $\boldsymbol{p} = [p_1 \dots p_k]$ denotes the power of the users and $\boldsymbol{\theta} = [\theta_1 \dots \theta_N]$ represents the IRS phases. γ_k is the SINR of the k-th user and P^{MAX} is the available power at the BS for each group. The terms (5-b) and (5-c) are power constraints that allocate the available power at the BS terminal among multiple users, based on the power allocation pattern of the downlink NOMA systems. Constraint (5-d) guarantees that the difference in channel gain of the users meets a minimum threshold value. Finally, the constraint (5-e) sets the permissible range for the IRS phase shifts.

To fully take advantage of IRS-NOMA, an optimization problem must be tackled to determine the optimal powers and phase shifts to maximize the sum rate of the system. The problem can be expressed as a non-convex optimization problem. Undoubtedly, inverting the problem to a convex one would provide more accurate results. However, by adding some restrictions, our attempts for this matter reached a high-order fractional problem which is again non-convex. Consequently, we decided to utilize the YALMIP toolbox within MATLAB [30]. It is a powerful toolbox supporting a wide

range of optimization problems. The advantage of this toolbox is that it allows us to work with the non-convex problem directly, without imposing any restrictions. It utilizes an appropriate nonlinear solver to address the problem.

Results and Discussion

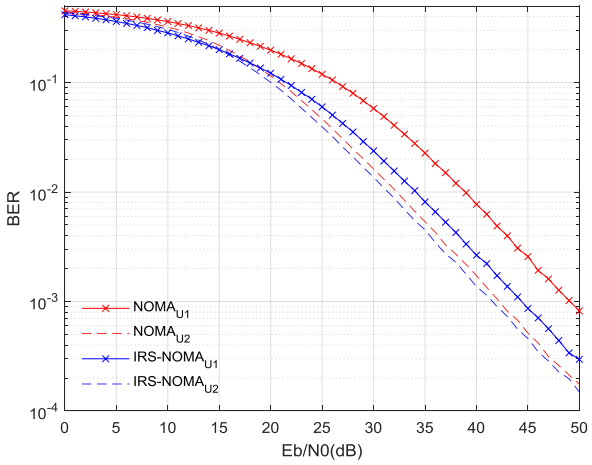
In this section, we conducted a series of simulations to evaluate the performance of the proposed approach under various scenarios, focusing on the system's average BER and sum data rate. It is necessary to clarify that the proposed method, referred to as IRS-NOMA in the figures, is compared with conventional NOMA as a benchmark. The simulation setup involves a single-cell IRS-NOMA system with a radius R, where the BS is located at the center, and users are randomly distributed along the cell's edge. The simulations were carried out using MATLAB and employed the Monte Carlo method. In these simulations, the path loss exponent (α) and the number of the IRS elements (N) are set to 2 and 5, respectively. The signal-to-noise ratio per bit (E_b/N_0) ranged from 0 to 50 (dB) and data transmission was performed using binary phase-shift keying (BPSK) modulation. Simulation parameters are detailed in Table 1.

Table 1: Details of simulation setting and parameters

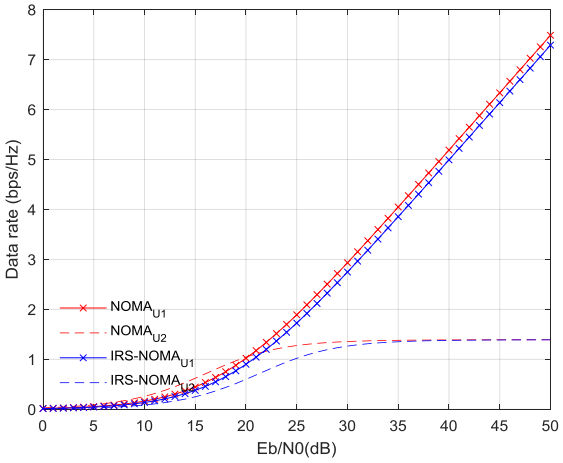
$\boldsymbol{R}$	10m
$\boldsymbol{\alpha}$	2
N	5
$\boldsymbol{E_b/N_0}$	0 to 50 (dB)
Number of users	2
Number of symbols	1e6
Modulation	BPSK
Channel realizations	1e6

Fig. 2 compares the performance of IRS-NOMA with conventional NOMA for each user in terms of BER and the data rate. As depicted, IRS-assisted NOMA enhances the system's performance for the corresponding users. Also, comparing the difference in BER and rate indicates that IRS-NOMA distributes resources (i.e., power) more fairly than NOMA. To gain more insight into the impact of the IRS on the NOMA system, the proposed IRS-NOMA method is compared with NOMA in terms of average BER and sum rate in the next set of simulations. Fig. 3 confirms that IRS-NOMA outperforms NOMA, resulting in a lower average BER and higher sum rate.

In the continuation, we will examine how the placement of an IRS affects system performance based on different scenarios shown Fig. 4. In these scenarios, the BS, edge users, and IRS are located at the three vertices of a triangle. The BS is fixed in the cell center, and the users are fixed on the circle's perimeter. The aim is to investigate the effect of the IRS position on system performance from various aspects. Given that a, b, and c are the distances from the BS to the users, BS to the IRS, and IRS to the users respectively. In the first scenario, the IRS is positioned at an equal distance from the edge users ($b = c$). The IRS is considered to be close to the BS ($b < c$) in the second scenario. Finally, in scenario 3, the IRS is located near the cell edge users ($b > c$). It's important to note that in all cases, $a > b, c$.

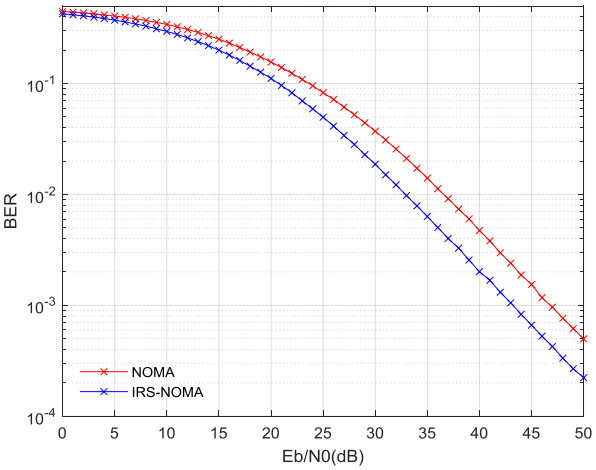


(a)

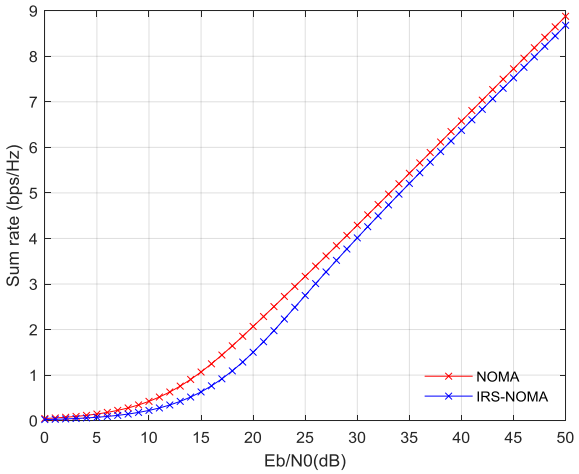


(b)

Fig. 2: Comparison of user performance in the NOMA and IRS-NOMA systems in terms of (a) BER and (b) data rate.



(a)



(b)

Fig. 3: NOMA vs. IRS-NOMA in terms of (a) average BER and (b) sum rate.

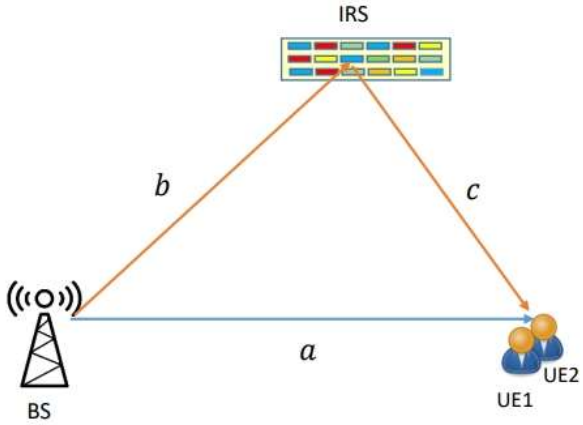
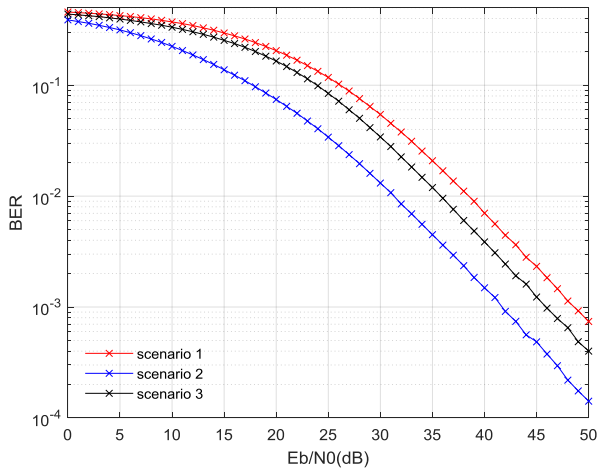
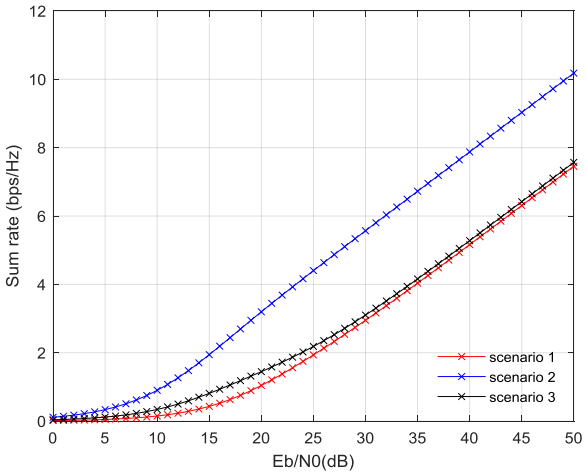


Fig. 4: Position of system components on the vertices of a given triangle.

Fig. 5 displays the simulation results for these three scenarios. Upon comparing the results, it is obvious that the second scenario demonstrates superior performance in terms of average BER and sum rate when compared to the other two scenarios. It confirms that the system performs better when the IRS position is closer to the BS.



(a)



(b)

Fig. 5: Effect of the IRS location on system performance in terms of (a) average BER and (b) sum rate.

Conclusion

This study investigates the performance challenges of edge users in the NOMA system due to poor signal quality and similar channel conditions. The aim is to utilize IRS capabilities to modify the time-varying transmission environment and improve overall system performance. To achieve this, an optimization problem is introduced to intelligently adjust the phase shifts of the IRS elements and allocate the power resources to edge users considering the channel gains constraint. The proposed IRS-assisted NOMA improves overall system performance and distributes resources more fairly than conventional NOMA. Compared to NOMA, the proposed IRS-NOMA system demonstrates a gain of about 4 dB at a BER of 10^{-2} and 3 dB at the sum rate of 10^{-1} bps/Hz. Furthermore, it has been observed that the placement of the IRS within the cell impacts system performance, suggesting that the system operates better when the IRS is positioned closer to the BS. In future work, we aim to find classical solutions for the sum rate optimization problem, particularly for groups with more than 2 users. Additionally, we plan to integrate machine learning (ML) models with IRS-NOMA systems to bring intelligence to the IRS controller.

Author Contributions

F. Rahdari and M. Sheikh-Hosseini designed the experiments. F. Rahdari and M. Sheikh-Hosseini, and M. Jamshidi formulated the optimization problem. M. Jamshidi implemented the optimization problem and obtained the results. F. Rahdari conducted the experiments and interpreted the results with M. Sheikh-Hosseini. F. Rahdari wrote the manuscript.

Acknowledgment

This research has been supported by the Institute of Science and High Technology and Environmental

Sciences, Graduate University of Advanced Technology, Kerman, Iran under grant number 01.1859.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

NOMA	Non-Orthogonal Multiple Access
BS	Base Station
SC	Superposition Coding
UE	User Equipment
SIC	Successive Interference Cancellation
OMA	Orthogonal Multiple Access
IRS	Intelligent Reflecting Surface
MISO	Multiple-Input Single-Output
SE	Spectral Efficiency
EE	Energy Efficiency
CRN	Cognitive Radio Networks
CSI	Channel State Information
BCD	Block Coordinate Descent
MC	Multi-Carrier
QoS	Quality of Service
DC	Difference of Convex
BER	Bit Error Rate
LOS	Line of Sight
SINR	Signal to Interference plus Noise Ratio
AWGN	Additive White Gaussian Noise
BPSK	Binary Phase-Shift Keying
ZC	Zero Forcing
ML	Machine Learning

References

[1] S. Patel, D. Chauhan, S. Gupta, "An overview of non-orthogonal multiple access for future radio communication," in Proc. IEEE International Conference on Intelligent Technologies (CONIT): 1-3, 2021.

[2] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, L. Hanzo, "A survey of non-orthogonal multiple access for 5G," IEEE Commun. Surv. Tutorials, 20(3): 2294-2323, 2018.

[3] M. C. Mayarakaca, B. M. Lee, "A survey on non-orthogonal multiple access for unmanned aerial vehicle networks: Machine learning approach," IEEE Access, 12: 51138-51165, 2024.

[4] M. F. Darus, F. Idris, N. Hashim, "Energy-efficient non-orthogonal multiple access for wireless communication system," Int. J. Electr. Comput. Eng., 13(2): 1654, 2023.

[5] H. Mathur, T. Deepa, "A survey on advanced multiple access techniques for 5G and beyond wireless communications," Wireless Pers. Commun., 118(2): 1775-1792, 2021.

[6] J. Li, T. Gao, B. He, W. Zheng, F. Lin, "Power allocation and user grouping for NOMA downlink systems," Appl. Sci., 13(4): 2452, 2023.

[7] F. Rahdari, M. R. Khayyambashi, N. Movahhedinia, "QoE-aware NOMA user grouping in 5G mobile communications using a multi-stage interval type-2 fuzzy set model," Ad Hoc Networks, 149: 103227, 2023.

[8] A. Akbar, S. Jangsher, F. A. Bhatti, "NOMA and 5G emerging technologies: A survey on issues and solution techniques," Comput. Networks, 190: 107950, 2021.

[9] D. Sarkar, S. S. Yadav, V. Pal, N. Kumar, S. K. Patra, "A comprehensive survey on IRS-assisted NOMA-based 6G wireless network: Design perspectives, challenges, and future directions," IEEE Trans. Netw. Serv. Manage., 21(2): 2539-2562, 2024.

[10] Z. Ding, L. Lv, F. Fang, O. A. Dobre, G. K. Karagiannidis, N. Al-Dhahir, R. Schober, H. V. Poor, "A state-of-the-art survey on reconfigurable intelligent surface-assisted non-orthogonal multiple access networks," Proc. IEEE, 110(9): 1358-1379, 2022.

[11] M. S. Gilan, B. Maham, "Performance analysis of power-efficient IRS-Assisted full duplex NOMA systems," Phys. Commun., 64: 102338, 2024.

[12] F. Naeem, G. Kaddoum, S. Khan, K. S. Khan, N. Adam, "IRS-empowered 6G networks: deployment strategies, performance optimization, and future research directions," IEEE Access, 10: 118676-118696, 2022.

[13] S. Kumar, P. Yadav, M. Kaur, R. Kumar, "A survey on IRS NOMA integrated communication networks," Telecommun. Syst., 80(2): 277-302, 2022.

[14] F. C. Okogbaa, Q. Z. Ahmed, F. A. Khan, W. B. Abbas, F. Che, S. A. R. Zaidi, T. Alade, "Design and application of intelligent reflecting surface (IRS) for beyond 5G wireless networks: A review," Sensors, 22(7): 2436, 2022.

[15] Z. Sun, Y. Jing, "On the performance of training-based IRS-assisted communications under correlated rayleigh fading," IEEE Trans. Commun., 71(5): 3117-3131, 2023.

[16] M. Ren, X. Li, X. Tian, "Throughput maximization scheme of the IRS-aided wireless powered NOMA systems," Phys. Commun., 63: 102269, 2024.

[17] Y. Wu, F. Zhou, W. Wu, Q. Wu, R.Q. Hu, K. K. Wong, "Multi-objective optimization for spectrum and energy efficiency tradeoff in IRS-assisted CRNs with NOMA," IEEE Trans. Wireless Commun., 21(8): 6627-6642, 2022.

[18] H. Al-Obiedollah, H. B. Salameh, K. Cumanan, Z. Ding, O. A. Dobre, "Competitive IRS Assignment for IRS-Based NOMA System," IEEE Wireless Commun. Lett., 13(2): 505-509, 2023.

[19] T. A. Nguyen, H. V. Nguyen, D. T. Do, S. N. Sur, "Performance analysis of two IRS-NOMA users in downlink," in Proc. International Conference on Communication, Devices and Networking, 2022.

[20] G. Li, H. Zhang, Y. Wang, Y. Xu, "QoS guaranteed power minimization and beamforming for IRS-assisted NOMA systems," IEEE Wireless Communications. Lett., 12(3): 391-395, 2022.

[21] T. Shaik, R. Bavirishetty, A. Baloju, S. Maku, "Ergodic analysis of IRS-NOMA with inter-cell interference and Imperfect-CSI and SIC over Nakagami-m faded channel," Results Eng., 23: 102464, 2024.

[22] M. Zhang, M. Chen, Z. Yang, H. Asgari, M. Shikh-Bahaei, "Joint user clustering and passive beamforming for downlink NOMA system with reconfigurable intelligent surface," in Proc. IEEE 31st Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, 2020.

[23] Y. Wang, B. Ji, D. Li, "IRS assist wireless communication: Scenarios, advantages, convergence," J. Comput. Electron. Inf. Manage., 10(3): 40-45, 2023.

[24] M. Abd-Elnaby, G. G. Sedhom, E. S. M. El-Rabaie, M. Elwekeil, "NOMA for 5G and beyond: literature review and novel trends," Wireless Networks, 29(4): 1629-1653, 2023.

[25] H. Sadia, A. K. Hassan, Z. H. Abbas, Gh. Abbas, M. Waqas, Z. Han, "IRS-enabled NOMA communication systems: A network architecture primer with future trends and challenges," Digital Commun. Networks, 10(5): 1503-1528, 2023.

[26] H. Al-Obiedollah, H. B. Salameh, K. Cumanan, Z. Ding, O. A. Dobre, "Self-sustainable multi-IRS-aided wireless powered hybrid TDMA-NOMA system," IEEE Access, 11: 57428-57436, 2023.

[27] Z. Ding, H. V. Poor, "A simple design of IRS-NOMA transmission,",IEEE Commun. Lett., 24(5): 1119-1123, 2020.

[28] Z. Ding, Z. Yang, P. Fan, H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," IEEE Signal Process. Lett., 21(12): 1501-1505, 2014.

[29] F. Rahdari, M. Sheikh-Hosseini, "Nonlinear symbolic regression for bit error rate prediction of NOMA systems in 5G cellular communications," Eng. Appl. Artif. Intell., 127: 107344, 2024.

[30] J. Lofberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in Proc. IEEE International Conference on Robotics and Automation: 284-289, 2004.

Biographies

Farhad Rahdari received his B.Sc. and M.Sc. degrees in computer engineering from the Faculty of Computer Engineering, Iran University of Science & Technology (IUST), Tehran, Iran, in 2000 and 2007. He also got his Ph.D. degree in computer engineering from the Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran in 2023. His current research interests include Resource Management in Cellular Networks, QoE Management, and Intelligent Networks.

- Email: farhadr@kgut.ac.ir
- ORCID: 0000-0001-5133-7269
- Web of Science Researcher ID: N/A
- Scopus Author ID: 55485717700
- Homepage: https://kgut.ac.ir/fa/faculty/Rahdari

Mohsen Sheikh-Hosseini received his B.S. degree in Electrical Engineering from the Shahid Bahonar University of Kerman, Iran, in 2007, and both M.Sc. and Ph.D. degrees in Telecommunications from Ferdowsi University of Mashhad, Iran, in 2009 and 2014. He is currently with the Department of Computer and Information Technology, Graduate University of Advanced Technology, Kerman, Iran. His research interests encompass (i) communications theory including designing of emerging physical layer waveforms, performance analysis of impulsive noise channels, and applications of AI in resource management of telecommunications systems; (ii) wireless and wireline communications including cellular communications (4G/5G/6G/..) and power line communications; and (iii) smart grid communications.

- Email: m.sheikhhosseini@kgut.ac.ir
- ORCID: 0000-0002-5959-7839
- Web of Science Researcher ID: N/A
- Scopus Author ID: 35276100700
- Homepage: https://kgut.ac.ir/fa/faculty/Sheikh-Hosseini

Mina Jamshidi received her B.Sc., M.Sc., and Ph.D. degrees in Mathematics from the Faculty of Mathematics, Statistics and Computer Science, Shahid Bahonar University of Kerman, Iran in 2004, 2006, and 2011, respectively. Her current research interests include Linear Algebra, Graph Theory with their applications in Data Mining. More precisely, she researches Multi-View Clustering methods based on mathematical tools.

- Email: m.jamshidi@kgut.ac.ir
- ORCID: 0000-0002-1944-4600
- Web of Science Researcher ID: N/A
- Scopus Author ID: 56285629200
- Homepage: https://kgut.ac.ir/fa/faculty/Jamshidi

How to cite this paper:
F. Rahdari, M. Sheikh-Hosseini, M. Jamshidi, "Edge user performance improvement by intelligent reflecting surface-assisted NOMA system," J. Electr. Comput. Eng. Innovations, 13(2): 275-282, 2025.

DOI: 10.22061/jecei.2024.11070.761

URL: https://jecei.sru.ac.ir/article_2233.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

RR-SFVP: A Novel Arbitration Unit Design for NoC Router, Ingeniously Fusing the Round Robin Method with Strong Fairness and Variable Priority

E. Shafigh Fard [1], M. A. Jabraeil Jamali [2,], M. Masdari [1], K. Majidzadeh [1]*

[1]*Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran.*

[2]*Department of Computer Engineering, Shabestar Branch, Islamic Azad University Shabestar, Iran.*

Article Info

Article History:
Received 31 August 2024
Reviewed 12 October 2024
Revised 10 December 2024
Accepted 15 December 2024

Keywords:
Network on Chip
Router
Round robin
Arbiter
virtual channel

*Corresponding Author's Email Address: *m_jamali@itrc.ac.ir*

Abstract

Background and Objectives: A network on Chip (NoC) is a scalable communication framework that supports several cores. In some cases, while designing a customized Network-on-Chip, the communication needs across IP cores are often uneven, resulting in imbalanced loads on the input ports of a router. The arbitration unit plays a crucial role in the design of the NoC micro-router architecture as it substantially influences the performance, chip occupancy, and power consumption of the NoC.

Methods: This article presents a router arbitration architecture that utilizes a mix of variable priority arbitration and round-robin methods. The arbitration process evaluates other channels' requests using the Round Robin index within this architectural framework. A novel approach was suggested to integrate a network router unit onto a single chip, offering several benefits compared to earlier methods. The most significant advantage is its variable priority feature, which allows inputs to be assigned different priority levels regardless of the design circuit. The system is meant to prioritize fairness across all requests by sequentially executing them. The second and primary benefit of the developed circuit is its ability to retain the previously assigned virtual channel ID. This feature preserves the provided virtual channel ID and reduces the time required to verify the requested virtual channels in the subsequent cycle.

Results: The evaluation process occurs after the flit has been requested to quit the virtual channel and the availability of the corresponding virtual channel has been verified. The simulation findings demonstrate that the RR-SFVP arbitration unit's design is 12.1% more compact in space than the standard RR approach, offering a promising solution for space-constrained designs. It exhibits 4.3% lower power consumption, a significant improvement in energy efficiency, and 55.1% reduced critical path time, enhancing the system's overall performance.

Conclusion: The RR-SFVP technique incorporates all favorable elements in the design of the arbitration unit circuit, such as variable priority and equitable arbitration. Its clear benefits make a strong case for its superiority in the field.

Introduction

Network on Chip (NoC) is a suggested architecture designed to address the issues of using a shared bus. This network employs a modular and scalable architecture rather than a conventional bus, mapping IP blocks onto the network as tiles [1]-[4]. Data is

transported across connections using a built-in router in a packed format. Unlike bass-based systems, NoC (Network-on-Chip) is an important innovation that enables increased bandwidth and improved scalability inside each tile [5], [6]. The importance of this technology lies in its capacity to surpass the constraints of shared bus systems. As a result, it has become a crucial focus of study and development in computer engineering. This field deserves our attention and active involvement due to its potential to revolutionize the future of computing. The NoC architecture typically comprises processing cores, routers, and connections. Each router is composed of a switch and many buffers [7]. NoC is a state-of-the-art on-chip interconnect network designed for packet-based communications. Network-on-Chips (NoCs) provide the advantages of reduced packet latency, increased bandwidth, improved throughput, fewer space requirements, enhanced energy efficiency, and increased fault tolerance. Network-on-chip (NoC) systems rely heavily on routers, the primary building pieces [8]. Typically, data in the network system on the NoC chip is transferred in packets, which are then subdivided into flits. The whole packet is stored in the input buffer before being sent to the wormhole switching in the store and forward architecture. The flits are then exchanged between each router. The message comprises an initial head flit, one or more data chunks, and a concluding tail flit. According to Fig. 1, the flits are 16 bits in this article. Network on Chip (NoC) is a suggested architectural solution to address the issues of using a shared bus. This network employs a modular and scalable architecture instead of a conventional bus, where IP blocks are assigned to the network as tiles [1]-[4]. Data is transported across connections using a built-in router in a packed format. Unlike bass-based systems, NoC (Network-on-Chip) is an important innovation that enables increased bandwidth and improved scalability inside each tile [5], [6]. The importance of this technology resides in its capacity to surpass the constraints of shared bus systems. As a result, it has become a crucial area of study and advancement in computer engineering. This topic requires our attention and involvement due to its potential to revolutionize the future of computing. The NoC architecture typically comprises processing cores, routers, and connections. Each router is composed of a switch and many buffers [7]. NoC is a state-of-the-art on-chip interconnect network designed for packet-based communications. Network-on-Chips (NoCs) provide the advantages of reduced packet latency, increased bandwidth, improved throughput, fewer space requirements, enhanced energy efficiency, and increased fault tolerance. Routers serve as the primary components of Network-on-Chips (NoCs) [8]. Typically, in the network system of the NoC chip, data is sent in packets, which are then fragmented into flits. In the store and forward paradigm, the whole packet is stored in the input buffer before being sent to the wormhole switching. The packet is divided into smaller units called flits and passed between each router. The message comprises an initial head flit, one or more data chunks, and a concluding tail flit. According to Fig. 1, the flits have a size of 16 bits in this article.

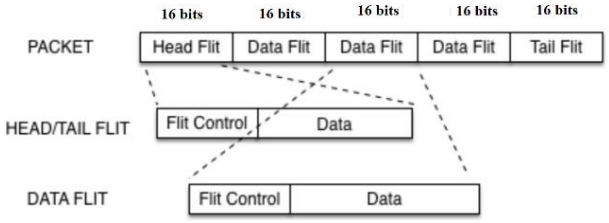
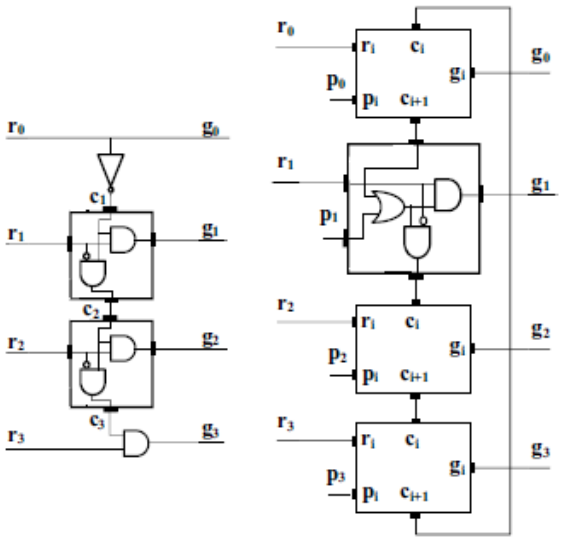


Fig. 1: Packet and flit structure [9].

A router's data pipeline comprises an input port buffer, a crossbar switch structure, and an essential component known as the arbitration unit. The arbitration unit manages traffic by identifying the virtual channel with the greatest priority for sending data under competitive settings [10]-[12]. The configuration of the arbitration unit might be intricate depending on the arbitration priority and the kind of control. The critical path delay in a Network-on-Chip (NoC) router architecture often occurs in the input ports, switches, and arbitrations. This delay is rather considerable compared to other units, primarily because of the intricate construction of the arbitration unit. Therefore, the arbitration unit circuit calculates the highest possible system speed. Hence, the choice of arbitration unit design has a crucial role in determining the performance, characteristics, speed, and power consumption of the NoC system [13]-[16]. Fig. 2 depicts two arbitration units, including four input ports.



(a) Fixed priority (b) Variable priority

Fig. 2: Arbitration architecture with four input ports [3].

These units can resolve conflicts between n requests (r0, r1... rn) using available resources criteria and indications. The winning request on each line is granted (gi). Regarding priority, the arbitration architecture may be categorized into fixed and variable architecture. In the case of an arbiter with fixed priority, the priority of requests is decided linearly. Fig. 2(b) depicts an arbiter with a predetermined priority, where r0 is assigned the most significant priority and r3 is assigned the lowest priority [17], [18]. Variable priority arbitration differs from fixed priority arbitration in that it considers additional indications while determining the allocation of resources and following sequential criteria. Arbitration may be categorized into three classes (weak, firm, or FIFO) [19] based on the fairness requirements and the form of arbitration. In weak, fair arbitration, requests are approved without considering particular criteria and priorities. In a significant and equitable arbitration process, responses to requests are determined according to priorities and criteria contingent upon the unique circumstances. Equitable demands are allocated using the initial-in-first-out (FIFO) method, which prioritizes the earliest arrival of the initial service. Furthermore, a variable priority arbitration unit may ensure equitable arbitration, as seen in Fig. 2(a). When the priority is shifted from one cell to another in a time cycle, this kind of arbitration is called round-robin arbitration, as seen in Fig. 3.

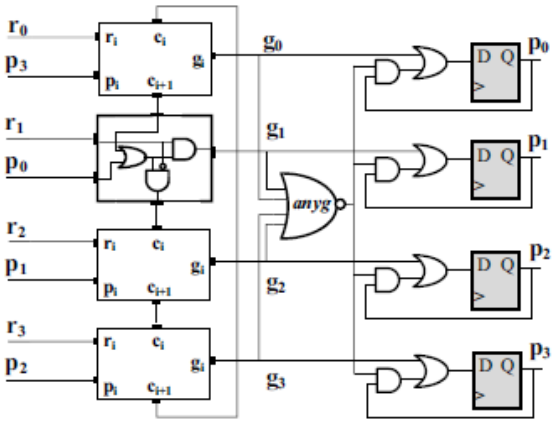


Fig. 3: Round-robin arbiter architecture with variable priority [16].

In this arbitration scenario, when the priority of g1 is high in the present cycle, P_1 is assigned a high priority in the subsequent clock cycle. Consequently, r_2 attains the most significant priority in the following clock cycle, while r_1 is assigned the lowest priority. The round-robin arbitration model [20] is a straightforward and efficient method that does not suffer from starvation. As the number of input requests increases, the round-robin arbitration structure expands, resulting in more space, greater power consumption, and longer critical route delays for big processors. This research has considered additional crucial aspects contributing to a fairer arbitration process based on the RR technique. This approach handles exit request cycles and addresses cycles when no request is generated from any channel (significant arbitration). In addition to this benefit, a circuit with a much reduced critical path latency has been created compared to earlier techniques, such as RR. The notable contributions of this study were as follows: This study effectively decreases the critical route time and power use of Network-on-Chip (NoC) by using variable and equitable priority in Round-Robin (RR) arbitration. The decrease in size allows for improved and environmentally friendly network operations, showcasing the tangible advantages of the suggested RR-SFVP arbitration architecture.

- Utilizing significant determinants to choose a robust arbitration mechanism and providing a grant to the appropriate channel in the presence of competitive situations among many virtual channels for picking an output port. The suggested design provides a cost-effective solution compared to existing arbitration systems due to its reduced hardware demand, alleviating the audience's concerns. The subsequent parts of the paper are structured in the following manner: Section 2 provides a concise overview of the relevant literature. Section 4 presents an elaborate analysis of the simulation of the RR-SFVP arbitration architecture and its corresponding outcomes. This part provides tangible proof of the architecture's efficacy, further confirming its potential. Section 5 ultimately wraps up and suggests avenues for further study, urging the audience to consider investigating the potential of this groundbreaking design.

Related Work

Dally and Towles introduced the Matrix round-robin arbitration unit design. This architecture sequentially evaluates the input requests and is considered one of the arbitration approaches that ensure essential fairness while preserving the prior award. This kind of arbitration is beneficial for a limited number of inputs. Nevertheless, the intricate hardware design and the extensive resource use have compelled researchers to choose a more efficient arbitration unit [21]. The arbitration unit is used in a router with an input port known as a link-list DAMQ (LLD), specifically called an LLD-Matrix router. This approach utilizes lists and table linkages in the input port to facilitate the reading and writing of flits in the buffer. The router equipped with the LLD input port is an option that offers a reasonably affordable hardware cost but exhibits limited performance. This method facilitates concurrent communication, which may be particularly advantageous for NoC routers—a link-list method to update five tables after each write and read operation results in significant delays. Several modifications made to the specified tables often result in substantial time delays. The latency also escalates as the input rate rises [22].

The matrix round-robin arbitration mechanism is used when a router's input port is ViChaR. In contrast to the LINK-LIST DAMQ paradigm, this technique does not use the LINK-LIST. The ViChar controller circuit is expensive in terms of hardware. This technique can accommodate virtual channels with the maximum buffer slot size, necessitating arbitration for allocating slot buffers to virtual channels and switches. Large buffer slots may create a bottleneck on critical pathways, limiting NoC. This approach has many drawbacks, including complexity, restrictions in adjustment, and lengthier pipelines in the entrance and departure of flits [23]. Fu and Ling compared two methods, RoR [19] and Matrix [21], on an FPGA platform [24], focusing on resource consumption, performance, and power consumption. They concluded that Matrix arbitration utilizes a more significant amount of resources. Both methods have equivalent power consumption; however, the Matrix method exhibits superior data processing speed compared to the RoR method. Zheng and Yang introduced a round-robin arbiter technique in which inputs are arbitrated simultaneously. The PRRA algorithm is derived from a straightforward binary search method that utilizes four inputs to enhance latency. A proposed IPRRA design aims to decrease the critical path delay in PRRA architecture [25]. The IPRRA method significantly reduces the time taken for the PRRA critical route. Lee et al. [26] suggested the Round Robin Arbiter (HDRA) method. This method employs individual filter circuits for each input, which utilize indicators within the circuit to determine the order of requests and assign grants to specific inputs. PRRA lacks fairness compared to other methods, including the discussed method. IIR arbitration is an arbitration method that outperforms similar approaches regarding power consumption and delay due to its superior architecture. This architectural design transmits requests r0, r1,… rn to the arbitration unit. This unit evaluates the requests systematically, and if they meet the criteria, a grant will be allocated to the approved request. The IRR_WF method, akin to the IRR method, does not retain the previous grant clock value. This technique is referred to as arbitration with weak fairness. The distinction between IRR and IRR_WF lies in the fact that IRR, in the absence of an exit request, retains the previous request based on the presence of REG, which is deemed equitable arbitration for the subsequent cycle. Nevertheless, in the IRR-WF approach, the absence of a register results in the loss of the previous request if all requests are zero [27]. To ensure that a high-performance network of chip switches is provided [28], an efficient arbiter is needed, especially in terms of fairness. The architecture was proposed based on a tree structure, which divided and distributed the arbitration task to separate nodes, providing high-performance arbitration with excellent scalability. The FSA (Fairness Switch Arbiter) uses a feedback-based parallel priority update mechanism to complete arbitration. The FSA method uses four inputs to achieve a critical path with only an O ($log4N$) delay. This method is similar to IPRRA in structure and performance but fairer than IPRRA.

The Weighted Round Robin Arbiter (WRRA) [29] builds on the Round Robin Arbiter's principles by focusing on fair resource allocation. Each requester receives resources proportional to their assigned weight, distinguishing it from traditional methods. For example, if two requesters have weights of 3 and 7, they would receive 30% and 70% of grants over time, respectively. The arbiter uses a credit-based mechanism to determine allocations, maintaining credit counters for each requester to indicate eligibility. A replenishment process restores credits when no requests are active. Operating on clock signals, it updates counters during grant allocation and replenishment, ensuring effective resource management. Overall, this module enhances fairness and efficiency in resource distribution.

Group decision (GD) [30] method discusses a bus arbiter designed using a group decision algorithm that integrates fixed priority, round-robin, and mixed priority systems to create a new priority sequence. This approach addresses the starvation problem in multi-master systems on a chip (SoC) by swiftly responding to masters requiring bus access, halving their waiting time. Unlike traditional methods that may prioritize specific requests, this algorithm ensures that multiple masters can receive timely responses, albeit with increased area and power requirements. The group decision algorithm offers several advantages over conventional bus arbitration methods, including fairness and improved bus usage while leveraging the benefits of various priority types. Table 1 summarizes the methods employed and their respective benefits and drawbacks, which can be found in this section.

Table 1: Advantages and disadvantages of different types of arbitration units

disadvantages	advantages	how the method works	Method name
Arbitration with fixed priority	-Delay and lower cross-section compared to other methods Simplicity - Less overhead,	All the requests enter the general multiplexer, which, by choosing a request, is registered and leads to a substantial arbitration (no request), and the counter goes to the subsequent request (if there is a request)	IRR [23]

The processing speed is lower than the other methods	The possibility of saving the current priority in the cycle without a request	Works as a variable priority, and the request chain enters the arbitration process along with the i0 and p0 priority and proceeds in order.	ROR [10]
Higher resource consumption	- Faster data processing speed -Matrix fair arbiter	Resets the bits of row i and sets the bits of column i. Matrix	Matrix [7]
Poor adjudication (failure to save current grant) in the absence of request	-faster adjudication request than the IRR -The consumption area is less compared to IRR	As the IRR method minus the presence of SF multiplexer IRR_WF	IRR_WF [23]
Poor adjudication (failure to save the current grant) in the absence of an application	Less delay in critical paths	Requests are of fixed priority type, where each request consists of a flip-flop and multiplexer, where r0 has the highest priority at the beginning, and chainwise if r0 is not requested, the priority reaches other r. HDRA	HDRA [20]
The most extended delay among all arbitration methods Poor arbitration (no current grant record) -High consumption power	Reduction of critical path delay	RR method, round trip method based on binary (parallel) search algorithm PRRA (Zheng & Yang, 2007)	PRRA [18]
High latency The largest cross-sectional area compared to other methods.	Reducing the execution time of the PRRA method	The improved model of the PRRA IPRRA method	IPRRA [18]
Combination loop	Faster than the base rra	Connecting each cell to the next (s) Cell	Timing speculative arbiter [11]
Has a long critical path due to transport propagation through fixed and variable priority cells.	However, it avoids the compound loop.	A fixed-priority cell chain replaces the connection between the last and first cells.	Acyclic arbiter [12]
Big area circuit	Shorter critical path, fairness	Tree structure search algorithm	FSA [28]
MORE POWER CONSUMED	Superior fairness in resource allocation compared to the RR.	Based on RRA with resources proportional to their assigned weight	WRRA [29]
High power consumed More area	Fairness	combines the advantages of fixed priority, round-robin,	GA [30]

The Proposed Method

There are N arbiters in an N × N fabric switch, each responsible for arbitrating requests from all input ports directed to an output port. Due to the uniform construction of an arbiter compared to others, we will investigate just one arbiter.

A. VC Arbitration

VC arbitration is a crucial organizational component of a router that significantly affects the effectiveness of a NoC system. The Arbiter conducts arbitration among the competing VCs over a singular resource, such as an output port.

This dissertation presents innovative methods for dynamic virtual channel flow control techniques and virtual channel arbitration. The first two methods rely on the adaptability of virtual channels at the router input port, enhancing the effectiveness of the network-on-chip system. In both systems, the input port consists of a centralized buffer with slots dynamically assigned to virtual channels based on real-time traffic conditions. The use of several virtual channels with low buffering resources achieves performance enhancement. The VC arbitration method relies on an efficient and rapid arbiter that operates depending on the index of its input ports (or VC requests).

In the preceding part, input-port VC arbitrations were often executed with input-arbiter modules during the switch allocator phase. Nonetheless, the NoC arbiter architecture in this approach employs VC-Selector modules at the input port for two primary reasons: a central buffer retains all VC flits from an input port, and arbitration occurs within a single clock cycle. Upon transmitting a grant signal to an input port, the read pointer is either already aligned with the victorious VC flit, or the winning flit is positioned at the output port of the buffer. The design indicates that the VC-Selector selects a VC for the arbiter while concurrently loading the flit at the buffer output. The VC_req signals in the local link data (LLD), and the request data queue (RDQ) input ports provide the read pointer and facilitate the blocking mechanism, respectively. Although one may contemplate integrating VC-Selector modules inside the switch allocator, this would establish a reliance between the input port and arbiter, sharing segments of their critical routes. To prevent this, the architecture incorporates the

input-port VC arbitration inside the input-port, obviating the need for input-arbiter modules in the switch allocator, as seen in Fig. 4.

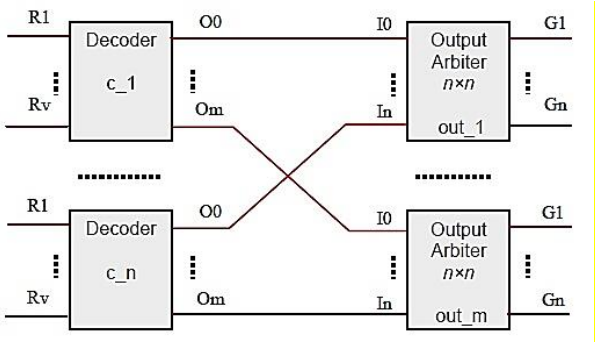


Fig. 4: *n*×*m* S.A. architectures, *n*= # of inputs, *m*= # of outputs, *v*=# of V.C.s per input-port.

The description of SA Micro-Architecture corresponds to the n input ports, as seen in Fig. 5.

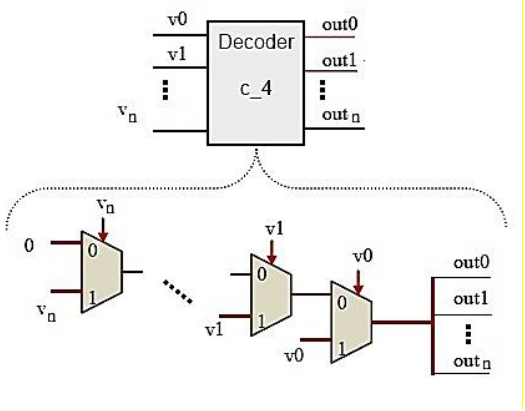


Fig. 5: Separable SA Micro-Architecture.

B. RR-SFVP Arbiter Micro-Architecture

The proposed arbitration architecture combines arbitrations with variable priority based on round robin. In the RR-SFVP method, all positive aspects have been used in the design of the arbitration unit circuit, including variable priority and fair arbitration. A particular virtual channel does not always have the highest priority, which is part of the variable priority characteristics. Referring to the previous methods of the arbitration unit, the requests are dealt with in order of fixed priority. In the RR-SFVP method, a (requested) channel does not always have the highest or lowest priority. To arbitrate more fairly, the accepted (granted) applications of the previous cycle are saved and examined in the next arbitration cycle. In the proposed method, when there is competition between several virtual channels to choose an output port, several important factors are involved in arbitration and awarding a grant to a virtual channel. The request to leave the virtual channel, the value of the previous grant cycle of the virtual channel, and requests from other channels are essential parameters in the arbitration process among virtual channels. The RR-SFVP method employs a variable priority mechanism associated with each channel, which is dynamically adjusted based on the history of requests and the current state of the channels. This design choice leads to a just and equitable system, ensuring fairness in arbitration requests and reassuring users. The RR-SFVP method reserves the details about the requests granted in the previous cycle, ensuring no requestor is indefinitely blocked. This feature, coupled with the system's ability to grant access to the shared resource over successive cycles, instills a strong sense of reliability, making the audience feel secure and confident in its performance. The RR-SFVP method allows the arbitration unit, a component responsible for managing and resolving conflicts over resource access, to simultaneously assess and make decisions for multiple channels. This parallel processing capability leads to a faster selection process, enhancing the system's ability to evaluate and make decisions for various channels simultaneously and making the audience feel the speed of the method. In the following, if r is used in the figures or text of the article, it is the abbreviation of request. Fig. 6 shows the logic circuit of the router arbiter logic block diagram.

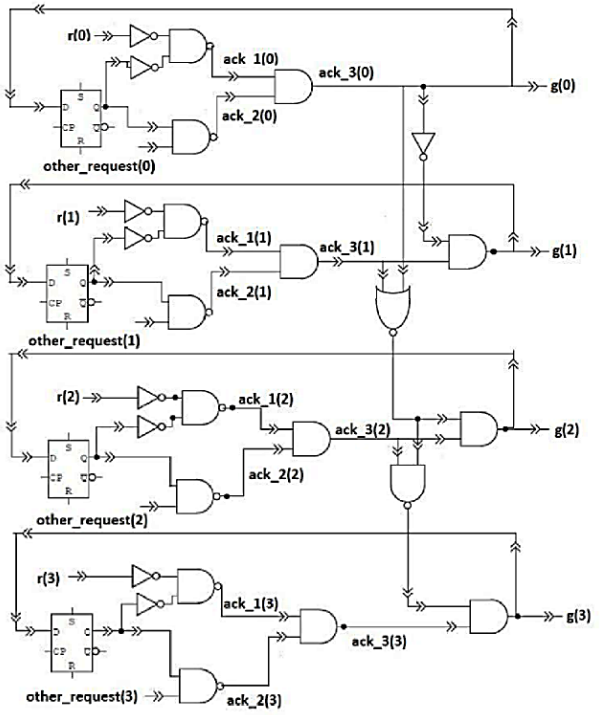


Fig. 6: RR-SFVP method arbiter circuit.

As the circuit in Fig. 6 shows, the parameters ri, Flip-Flop (i), and Other-request (i) affect the result of Gi. The flip-flops related to the circuit are initially set to 0. After each arbitration series, the rst (i) signal of the Flip-Flop (i) is adjusted by the circuit in Fig. 7.

This circuit is designed to reset a channel's flip-flop after it has been given access to avoid the indefinite blockage of other channels.

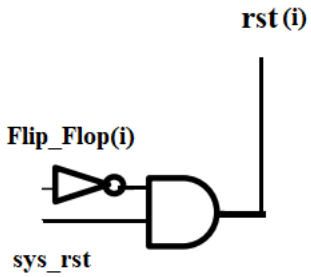


Fig. 7: Restart circuit of flip-flop.

Based on Fig. 7, any request that receives approval will not cause the flip-flop of the corresponding channel to reset during a general restart. Implementing this scenario will elevate the precedence of other virtual channels (which have not been provided). The sys_rst signal is a system-wide reset signal triggered at each arbitration process's conclusion. If the current request, rn, is active (1) and the flip-flop, fn, from the previous cycle, has a value of 0, indicating that the last cycle's gn was zero, the grant will be sent to n. This ensures that rn is not starved, even if another request, request (n), is active (equal to 1). The value of the variable "other_request" is set to 1 when at least one of the other ports of the virtual channel has submitted an exit request. In other words, the result may be expressed as the logical OR operation of all the individual exit requests (r0 OR r1 OR r2 ... OR rn-1).

If the current request, rn, is not active (0) and the flip-flop, fn, has a value of zero, regardless of the value of the parameter, other_request(n), the value of gn will be zero.

If the current request, rn, is active (1) and the flip-flop, fn, from the previous cycle, has a value of 0, indicating that the last cycle's gn was zero, then the grant will be given to n. This ensures that rn is not deprived, even if another _request (n) is active (equal to 1). The value of other_request is set to 1 when any other virtual channel ports have submitted an exit request. In other words, it may be expressed as the logical OR operation of all the exit requests (r0 or r1 or r2 ….. or rn-1).

If the rn request is disabled (0) and the flip-flop fn is set to zero, regardless of the value of the parameter other_request(n), the value of gn will be zero.

The procedure shown in Fig. 8 is used to find the appropriate factors for each channel. Upon receiving the request signal to exit channel i, the arbitration unit examines the contents of Flip-Flop (i) and the requests from other channels (i), respectively.

The suggested technique offers a distinct benefit, as seen in the algorithm depicted in *. Up to the line indicated with *, all the channels are processed simultaneously and without interdependence. However, the value of the grants is still influenced by the values of many other grants, as seen in circuit diagram 4. However, it should be noted that the principal channels do not have a high priority because the two channels have different funding.

```
Step 1: Set all Flip-Flop(N) =0
Step 2: Start
Step 3: Declare variables Flip-Flop[i],
        r[i], Other-request [i]
Step 4: Initialize variables
Step 5: Repeat the steps until i=N
   5.1: If i > N
          i <=0
        Else
read the value of Flip-Flop[i]
read the value of r [i]
read value of Other-request [i]

    ack_1 [i] = not(Flip-Flop[i]) nand
          not(r [i] )
    ack_2 [i] = Flip-Flop[i] nand
         Other-request [i])

 ack_3 [i] = ack_1 [i] and  ack_2[i]  *

        If  ack3[i]=1

          If (ack3[i-1] and ack[i-2],…)=0
          Gn=1

       Else

          Gn=0

  i<= i+1

step 6: Flip-flop(i) = Gi
step 7: Restart Flip-Flip(N)
step 8: Stop
```

Fig. 8: The algorithm of RR-SFVP method.

The conventional round-robin approach is not susceptible to hunger. However, this technique offers a benefit over the round-robin method since it does not need a sequential grant and allows for variable priority. Consequently, this method significantly improves arbitration time and throughput rate by checking all channels in every clock cycle. There is no need to verify their turn.

The suggested technique has a benefit, as seen in the algorithm depicted in Up to the line indicated with *; all the channels run concurrently and autonomously, without dependence on each other. However, following the intended line and circuit diagram 4, the magnitude of the grants is influenced by the magnitudes of the prior awards. However, it should be noted that the principal channels do not have a high priority because the two channels have distinct grants. The conventional round-

robin approach does not suffer from famine. However, this technique offers a benefit over the round-robin method since it does not need a sequential grant and allows for variable priority. Consequently, this method significantly improves arbitration time and throughput rate by checking all channels in every clock cycle. There is no need to verify their turn.

Based on the information provided in Fig. 9, two methods exist to get the virtual channel grant (G_i=1). In the first stage, Flip-Flop (i) has a value of 1, r(i) associated with exiting the virtual channel is likewise 1, and other_request(i) is 0.

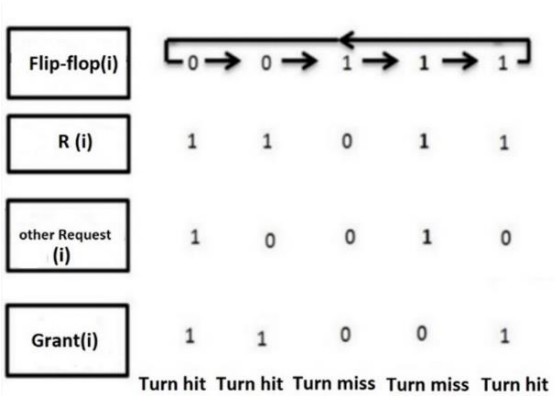


Fig. 9: Scenario of RR-SFVP method.

The second criterion for establishing the virtual channel grant is when Flip-flop(i) is equal to zero, r(i) associated with leaving the virtual channel is equal to 1, and the value of the other request is irrelevant. Furthermore, based on the shown situation in Fig. 9, two situations exist in which no grant is allocated for virtual channel I (Gi=0). The initial state occurs when all three values of signals Flip-flop (i), r (i), and Other-request (i) are simultaneously set to one. Alternatively, when Flip-flop (i) has a value of 1, the signal r(i) corresponds to the act of exiting the virtual channel, while the other request (i) is set to 0.

C. Functional and Fairness Analysis RR-SFVP

The circuit design is fair from two points of view. We consider scenarios based on the circuit in Fig.8 to prove this issue. First, I think a scenario where we want to prove that if a request is repeated repeatedly, the circuit acts fairly and repeated grants to It does not make requests and does not starve other requests; for example, in the first clock, channel number 1 only made a request and other channels did not make a request, according to the designed circuit and the logical analysis of the parameters of the zeroth virtual channel, g_0 is equal to zero, but in the circuit corresponding to the first virtual channel Considering that the previous value of the flip-flop is zero and r_1=1, the value of the virtual channel grant becomes one. In the next cycle, we assume that the first and third virtual channels reissue a request simultaneously. In this case, according to the circuit of the virtual channel, I am zero. According to logical analysis and r_0=0, the amount of grant g_0 is equal to zero, but in the first virtual channel, even though the request is issued But according to the logic analysis (other requests = 1 and the flip-flop value of virtual channel 1 = 1), therefore the grant is not assigned to the first virtual channel and the second virtual channel is the owner of the grant. This performance shows that the circuit is pretty fair and performs logical analysis in parallel, but it also pays attention to the requests of other channels.

According to the second point of view, which is a seal of approval on the fairness of the designed circuit, if no request is issued from any virtual channel with four inputs (0000), according to the circuit and digital analysis, the last flip-flop of the virtual channel, which is one, is used as ID grant. It is shown whether the virtual channel request is 1 or 0. This type of circuit operation reduces the circuit search delay and specifies the next clock from which the virtual channel ID of the arbitration operation should start.

One key factor contributing to the fairness of this arbitration unit is the reserve from the previous award cycle. When the time reaches zero, all requests are stored to enhance the evaluation efficiency and provide a more equitable selection process for the final award cycle. To better comprehend this subject matter, Fig. 10 and its corresponding timing diagram visually show the suggested approach's performance and behavior.

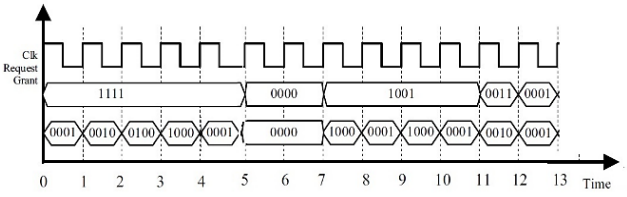


Fig. 10: Timing diagram for input request scenarios of strong fairness proposed method [26].

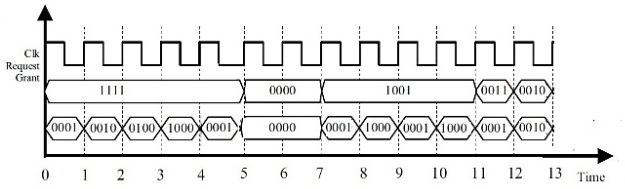


Fig. 11: Timing diagram for some input request scenarios of weak fairness arbiters [26].

Fig. 10 shows that from times 1 to 5, a constant input request of "1111" is applied, and each bit is given in every clock cycle. However, at time 5, the request will be modified to "0000", indicating that no request will be submitted. In the absence of a request, the priority of the previous request is logged and stored until a new request is initiated. For instance, at time 5, the priority of the

second bit of the save request is implemented, and at time 7, it is used for verification. Therefore, the request assumes grant ownership at time 7 in the fourth iteration. This implies that, as with other unjust systems, there is no need to initiate arbitration from the first request.

Fig. 11 indicates that approaches with inadequate fairness do not account for a scenario without requests. If, at time five, the system does not have any pending requests since the previous grant was not saved, and a request is submitted at time seven, the arbitration procedure restarts from the beginning of the relevant circuit. This condition presents a deficit in fairness. Non-compliance with the final priority, when not explicitly asked for, significantly affects the overall fairness of this suggested arbitration unit, guaranteeing a dependable and just procedure.

Fig. 10 illustrates the timing diagram for our round-robin arbiter to demonstrate its functionality and performance. A constant input request, "1111," is implemented throughout time intervals 1-5 and is partly satisfied in each clock cycle. The request is altered to "0000" at time 5, indicating no request is sent. In the absence of a request, the most recently authorized request's priority is acknowledged and included in any subsequent request. The second-bit priority of the request is recorded at time 6 and implemented at time 7. Consequently, at time 7, the fourth request is executed. We assessed our arbiter alongside many others (RoR, Matrix, PRRA, IPRRA, and HDRA) under identical testbench and request conditions. The time findings are shown in Figs. 10 and 11. The RoR, Matrix, and our IRR arbiters document the current priority shown in Fig. 8 when no request is submitted. However, the PRRA, IPRRA, and HDRA arbiters could neither exhibit the diverse waveforms seen in Fig. 11 nor document the priority. In the absence of a request, the least significant request bit for PRRA, IPRRA, and HDRA waveforms is assigned the greatest priority. The lack of a circuit to address the no-request scenario is the reason for the arbitration conducted by PRRA, IPRRA, and HDRA. The impartiality of an adjudicator is directly influenced by upholding the lowest priority under the no-request condition. The key advantage of our RR-SFVP arbitrator is its capacity to provide a more rigorous fairness arbitration.

The RR-SFVP technique utilizes parallel processing of all channels using the suggested algorithm. It also considers other requests and examines the request history of one channel. This approach ensures a fairer and more robust arbitration procedure. This form of arbitration has used the input port of the mechanism described in [34]. Upon entry, each flit is assigned to a specific virtual channel after verifying the availability of buffer slots and the emptiness of each section using the write pointer and header flit. The identifier associated with the header flit is allocated to the matching flit identifier.

To accept a new flit, it is necessary to identify the vacant slots in the shared buffer during the second step after selecting one of the virtual channels.

To assess the arbitration scheme discussed, the input port in the router must be used. As described in [31], this input port is responsible for reading and writing a portion of the input port to and from the buffer. One of the benefits of this approach is its utilization of a table and two straightforward read-and-write circuits. The input port of this approach has a hardware design that facilitates parallel processing. Simultaneously, the first vacant position in the buffer is located for writing, and straightforward and parallel hardware, as described in [28], is used to read from the address specified in the buffer.

D. Throughput Analysis RR-SFVP

Based on the circuit designed in Fig. 6, we check the throughput rate from 3 points. The first one is that since the throughput rate has an inverse relationship with the delay, the critical path delay of the circuit is less compared to other circuits, according to Table 3, which increases the throughput rate. The second case refers to the hardware structure that does not exist in the block operating circuit. As said in the fairness assessment section, every virtual channel with the highest priority in the current cycle has the lowest priority in the next cycle. This issue causes the operation Arbitration to be done faster, increasing the throughput rate. The third case goes back to the issue of the fairness of the circuit, that the ID of the last granted virtual channel is taken, considering that the corresponding flip-flop is also 1. To facilitate finding the following grant, it will be easier to find the following grant, and in a way, the time to see it will be faster, with less delay as a result. The permeability increases.

Evaluation of the Proposed Method

This part involves simulating the RR-SFVP router and comparing it with comparable scenarios based on its structure and architecture.

The primary performance indicators for the designed circuit are its speed, power consumption, and area. These metrics often quantify an arbitration circuit's speed, latency, or maximum frequency (F_{max}). The frequency of the arbitration circuit is determined by the most extended delay (critical route) between two registers at any given moment. The RR-SFVP approach is quite significant in this particular circumstance.

A. RR-SFVP Hardware Requirements

We have assessed the NoCs above based on primary hardware attributes, including power consumption, chip size, and speed, as determined by Verilog implementation using Synopsys Design Compiler. Evaluation results are obtained using ASIC technology libraries, such as 90 nm NanGate [32]. The configuration for the input-port ASIC's

power and area use the CMOS technology specifications from the Synopsys Generic 90nm Library, with a global operating voltage of 1.2V and a period of 400MHz.

Table 2 shows that the electrical characteristics of the logic gates are derived from the standard Synopsys 90 nm Digital library.

Table 2: Characteristics of gates

Gate name St.	Propagation Delay (ps)	Power Dy. (nW)	*Power (nW/MHz)*	Area (µm2)
INVX1	38	88	*12*	6.5
AND2X1	85	298	*19*	7.4
AND3X1	119	297	*34*	8.3
NAND2X1	51	336	*15*	5.5
OR2X1	85	226	*23*	7.4
OR3X1	114	250	*39*	9.2
OR4X1	137	261	*56*	10.1
NOR2X1	64	170	*15*	6.5
MUX21X1	107	815	*43*	11.1
MUX41X1	168	827	*58*	23.0
DEC24X1	119	1238	*66*	29.5
XOR2X1	133	454	*26*	13.8
DFFARX1	217	620	*100*	32.2

We conduct a hardware overhead study to evaluate the anticipated speed and hardware overhead of the previously described round-robin arbiters compared to our suggested arbiter. We do not use any algorithms to optimize the circuits like Electronic Design Automation software does.

The primary performance metrics of an arbiter circuit are speed, area, and power consumption. The standard metric for the speed of an arbiter circuit is the delay time or the maximum clock frequency (fmax). The clock frequency of an arbiter is determined by the maximum delay (critical path) between two concurrently timed registers. The circuits of 4-input arbiters are analyzed at the gate level. The electrical characteristics of the logic gates are obtained from the Synopsys 90nm Digital Standard Cell Library, as shown in Table 2.

We computed the aggregate of the areas and powers of all the cells for each arbiter to assess their power and area, as shown in Table 3.

The power encompasses both static and dynamic components. The critical route delay between two registers in each circuit is computed for speed estimate. The critical route for each circuit is shown by the numbers in parenthesis in the last column of Table 3. There is a consistent correlation between power consumption, critical route, and consumption area. These three variables are crucial in determining the most effective design of the arbitration unit. The typical metric to quantify an arbiter circuit's speed is the clock frequency's time or value (Fmax).

Table 3: Characteristics of 4-Input Arbiters based on Table 2

Type of 4-input arbiters	Area (µm2)	Power (µW)	Critical Path Delay (PS)
IRR	294	296 (282d)	625 (217+133+168+107)
RoR	328	298 (289d)	1242 (217+5*(85+85) +137+38)
Matrix	556	479 (465d)	747 (217 +2*38+3*85+114+85)
IRR_WF	280	274 (262 d)	518 (217+133+168)
HDRA	431	360 (348d)	609 (217 +64+51+85+85+107)
PRRA	510	493 (479d)	861(217+2*38+3*85+85+2*114)
IPRRA	528	488 (473d)	747 (217+2*38 +3*85+85+114)
RR-SFVP	288	285 (271d)	557 (217+38+3*51+85+64)
Arbiters	Saving	Saving	Faster
RR-SFVP / IRR	2% (better)	3.7% (better)	10.88%(better)
RR-SFVP / RoR	12.1% (better)	4.3% (better)	55.15%(better)
RR-SFVP / Matrix	48% (better)	40% (better)	25.5%(better)
RR-SFVP / IRR_WF	2.8% (worse)	4% (worse)	7.5%(worse)
RR-SFVP / HDRA	33% (better)	20% (better)	8.2%(better)
RR-SFVP / PRRA	44% (better)	42% (better)	35%(better)
RR-SFVP / IPRRA	45% (better)	41% (better)	25%(better)

B. Hardware Parameter Analysis

We evaluate the parameters of the designed circuit from two dimensions: 1- hardware and 2- network on chip. From the hardware point of view, the parameters of the circuit area, power consumption, and critical path are comparable. The type and size of the desired gate affect the area and power consumption of the circuit.

The critical path length should be shorter to reduce the delay and increase the throughput of the circuit. According to Fig. 6, gates have optimized the critical path delay. Table 3 pertains to circuit design. The first column displays the space taken up by the arbiter unit, depending on the gate used. The second column indicates the power spent, while the final column represents the delay on the critical route.

Table 3 demonstrates that the RR-SFVP technique has lower power consumption, area, and critical path than other ways, except the IRR_WF method, based on the hardware design. This technique, which incorporates weak fairness, exhibits a reduced footprint, leading to decreased power consumption and a shorter critical path than the RR-SFVP Three primary elements. This rigorous examination method guarantees the comprehensiveness of our study.

C. Performance Evaluation of RR-SFVP NoC

Latency and throughput are the primary performance metrics for assessing RR-SFVP NoC. The NoCs are built in System Verilog, and we use the ISE 14.4 simulation environment to get these performance metrics. A 8*8 mesh topology with five input/output and wormhole switching is considered. There are four VCs in every input port. According to (1), throughput is measured by the rate of receiving packets to the maximum number of packets injected at a given time. Time, which is 20ns in this evaluation as (1). The packet communication utilizes wormhole switching, with the channel width corresponding to the flit size of 16 bits. A packet has 16 flits; each input port contains a central 8-slot buffer. Each input port has four virtual channels, except for ViChaR, which includes four virtual channels corresponding to the number of buffer slots in the input port. Throughput and delay are assessed based on flit injection rates per time unit. Fig. 12 shows the Simulation waveform of the proposed arbiter architecture in the Xilinx ISE 14.4 simulator for the delayed flit departure from a router.

The suggested technique is assessed using synthetic benchmarks and actual application traffic, showcasing its potential advantages and performance enhancements.

Table 4 displays the comprehensive attributes of the redesigned NoC architecture.

Table 4: Structure of simulation parameters

Network Size	**8 x 8**
Packet/Flit/Data	16-bits
VC and I/O ports	4 VCs for each of the five ports
Switching mode	Wormhole
Topology	Mesh
Routing Algorithm	XY routing
Traffic Patterns	Tornado, Complement, Random MPEG, AV.

The different metrics, such as latency and throughput, have been measured and thoroughly evaluated using microarchitecture and Verilog simulation. We conducted measurements of both throughput and delay. Throughput is determined by calculating the rate at which packets are received compared to the maximum number of packets injected at a certain period. This may be represented as follows (1):

$$\text{Throughput}=\frac{\text{number of received packets} \times \text{ size of one packet}}{\text{number of nodes} \times \text{number of cycles}} \quad (1)$$

Equation (2) measures the average delay resulting from the average latency associated with the entrance and departure of a certain number of packets in a Network-on-Chip (NoC) during each clock cycle.

$$\text{latency}= |departure\ (time) - arrival(time)| \quad (2)$$

Fig. 12: Simulation waveform of router architecture.

The results of Fig. 12, which is done in the simulation environment of the ise 14.4 software, show the transfers of flits between the ports of different routers; for example, three flits have simultaneously requested in output virtual channel 3, which, according to the explanations about the priorities of the designed circuit, all the data enters the desired output channel with the least delay and without data loss. This circuit behavior plays a significant role in increasing the circuit's throughput.

D. RR-SFVP NoC Analysis Parameters

In this part, we will examine the circuit designed for its application in the network on the chip. The two parameters of delay and throughput are among the parameters that affect the performance of a good arbiter unit. Suppose we analyze the delay in the proposed method based on the designed circuit and the existing scenario. In that case, we will see that the delay between two flip-flops is optimal compared to the previous methods, even though the flip-flop in the circuit has a relatively high delay compared to other gates. But, considering that it increases the fairness of the arbitration unit and preserves the previous grant cycle, it is one of the advantages of this circuit.

On the other hand, considering that the ID of the last cycle grant is preserved, it saves time when searching for the next cycle. It reduces the delay and somehow increases the throughput of the arbitration unit. On the other hand, as mentioned in the hardware review section, the lower the hardware overhead, the lower the critical path delay, and as a result, according to (1), increases the throughput.

E. Evaluation of the Performance of the New Arbitration Unit in Synthetic Benchmarks

Packet communication relies on the use of wormhole switching. The channel width is equivalent to the flit size, which is 16 bits. A packet consists of 16 flits; each input port has a central buffer that can hold up to 8 slots. Each input port has four virtual channels, except for ViChaR, which has eight. The number of virtual channels in ViChaR equals the number of buffer slots in the input port. The flit injection rates per time unit are used to quantify the throughput and delay. For instance, a flit injection rate of 8 indicates that each node, namely the source core, injects eight flits every unit of time. The maximum injection rate is determined by the capacity of the NoC routers to transmit the flits. As previously stated, the arrival and departure of flits using this approach with input port [28] routers takes one cycle, but it takes two cycles for LLD and ViChaR-based routers.

Thus, considering a time unit equivalent to 16 clock cycles, the LLD and ViChaR-based sources may inject a maximum of eight flits. It is not feasible to inject more than eight flits. Nevertheless, the method-based source cores can only inject a maximum of 16 flits every time unit, making it impossible to inject more than 16 flits. Our simulation considers a maximum of 8 flit injection rates for RR-SFVP to provide a fair comparison.

The performance metrics of each Network-on-Chip (NoC) are influenced by the functional behavior of the data flow mechanism and the temporal characteristics of the router. When analyzing the performance of a Network-on-Chip (NoC), it is essential to consider the delays associated with the router on the critical path. Consequently, we evaluate these Network-on-Chips (NoCs) at various clock frequencies based on the essential path delays linked to their routers. The experiment assesses the performance parameters of the NoCs mentioned above, and the findings are shown in Figures 13, 14, and 15. There is a direct correlation between the clock rate and the performance measures. Let's consider that n packets transit through the NoC system during t, with a clock rate of f. During a period of t, the NoC system will transmit p×n packets at a clock rate of p×f. Figs. 13, 14 and 15 demonstrate the superior performance of our strategy compared to others. When the injection rates are increased, more flits are injected into the NoCs, resulting in a higher population level and more disagreement.

Regarding functionality, the LLD-RoR, LLD-Matrix, and LLD-HDRA operate similarly at the same frequency because the RoR arbiter is extremely similar to the HDRA arbiter covered in the function and fairness section. Thus, in terms of performance, the LLD-Matrix with a faster clock rate outperforms the LLD-RoR and LLD-HDRA NoCs. The ViChaR-Matrix NoC leads to the same conclusion. Consequently, four fast NoCs are chosen for assessment and comparison: LLD-Matrix, ViChaR-Matrix, EDVC-IRR, and router with input port [31] and RR-SFVP arbitration. With a 4-VC setup, the LLD-Matrix, ViChaR-Matrix, EDVC-IRR, and RR-SFVP operate at 514, 451, 820, and 1000 MHz clock, respectively. The frequencies above are derived from the critical path of the delays listed in Table 3.

The performance of a Network-on-Chip (NoC) is determined by the behavior of its data flow mechanism and its timing. Attributes of a router the evaluation of NoC performance must consider the critical path delays associated with the router. Consequently, we evaluate these Network-on-Chips (NoCs) at several clock frequencies based on the critical route.

The test evaluates the delays of their routers and the performance characteristics of the NoCs. The findings are shown in Fig. 13, Fig. 14 and Fig. 15. The performance measurements are directly proportional to the clock rate and NoC frequency.

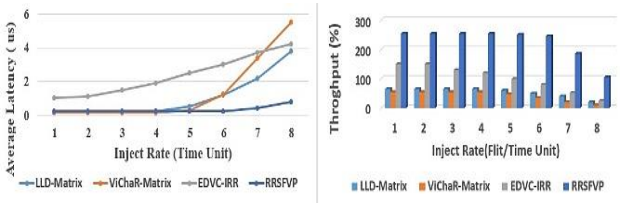


Fig. 13: Latency and average throughput for random traffic.

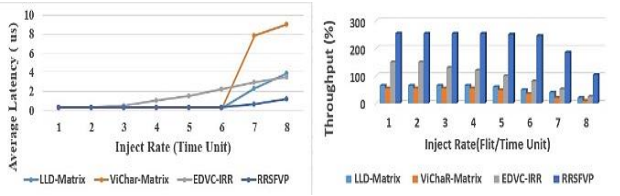


Fig. 14: Latency and average throughput for tornado traffic.

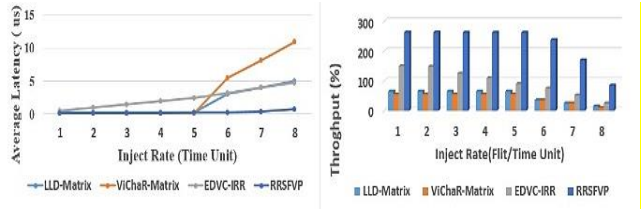


Fig. 15: Latency and average throughput for complement traffic.

Figs. 13, 14 and 15 show the mean delay and throughput requirements for the mesh topology (8×8) about the Complement and Random Tornado traffic patterns [33]-[35], respectively. Equations (3), (4) and (5) calculate the source address (Sx, Sy) and destination address (Dx, Dy) for Tornado, Complement, and random traffic patterns in a mesh topography of size m × m, where $0 < x, y \leq m-1$.

For Tornado:

$$Dx = Sx + (m/2)\text{-}1, Dy = Sy + (m/2)\text{-}1 \quad (3)$$

For Complement:

$$Dx = m\text{-}Sx\text{-}1, Dy = m\text{-}Sy\text{-}1 \quad (4)$$

For Random:

$$Dx = 1/m, Dy = 1/m \quad (5)$$

In XY routing and tornado traffic, all routers experience uniform congestion. Conversely, in complement traffic, the congestion is higher in side routers compared to middle routers. In the case of a random type, the packet is equally likely to be sent to other nodes. The experiment is conducted on an 8×8 network, where each packet comprises 16 flits. Additionally, each input port has a central buffer containing eight slots.

Arbitration systems that include diverse inputs exhibit superior efficiency compared to alternative arbiters. Typically, the chip area is smaller, the power consumption is decreased, and the critical path value is reduced. Due to fewer gates, this technique often has the lowest power consumption compared to other arbiters. Using fewer gates further streamlines the chip's architecture and arrangement.

Based on the data shown in Figs. 13, 14 and 15, the RR-SFVP's test results surpass those of other techniques, particularly at high rates. The number of flits rises with greater doses, intensifying rivalry among them.

There is a consistent correlation between the level of delay and the level of throughput. However, it should be noted that the suggested technique of fake traffic mentions at the beginning of the assessment section that the flits acceptance capacity at the entrance port of the relevant arbitration unit is twice as much as other approaches, namely two cycles. Consequently, the latency may be increased due to this factor, but this leads to a high throughput rate, as shown by expression 2.

Implementing ViChaR [23] for 4VC and 8-slots demonstrates significantly improved output NoC performance and reduced average latency compared to both Link-List and ViChaR NoC for various traffic scenarios. The acceleration is attributed to the input port's performance, directly reducing the number of executable cycles. Furthermore, the channels are simultaneously and concurrently verified by the arbitration unit in a specific section of the circuit, resulting in a better processing speed owing to the circuit's architecture. In contrast to other methods, the RR-SFVP selects the desired port in a parallel and simultaneous manner across multiple channels. This allows for efficient processing of requests and enables the system to make the best choice based on the criteria outlined in the RR-SFVP section.

F. Evaluation of the Performance of the Arbitration Unit in Real Benchmarks (Applications):

To further examine the effectiveness of the suggested technique, we assessed the proposed method's impact on the performance of two Network-on-Chip (NoC) applications, namely MPEG-4 and AV [33].

We conducted measurements of throughput and latency. Throughput was determined by calculating the rate at which packets were received and the maximum number of packets injected during a specific time frame. The average latency is determined by calculating the average time delays per clock cycle when a certain amount of packets are sent and received in the NoC. We vary the packet injection rates to assess the performance of the application-specific traffic. The rate of packet injection is modified per unit of time. The maximum bandwidth of the source cores defines the time unit. For example, the MPEG4 decoder's Source Core#8 has a maximum bandwidth of 1580 flits. Hence, the time unit will be 1580 clock cycles, assuming that each source core injects one flit each clock cycle. The AV (Audio-Video) application requires precise measurement. Specifically, Core#14 has a maximum bandwidth of 192078 flits per 192078 clock cycles.

MPEG-4 AV programs consisting of audio and video are mapped onto 2D chips with a mesh topology. The programs are mapped to dimensions of 3x4 and 4x4, respectively. Fig. 16 and Fig. 17 show diagrams of MPEG-4 and AV applications.

Packet communication utilizes wormhole routing and adheres to a specific XY routing algorithm. The arrow lines indicate the packet's route from the source cores to the destinations.

The assessment criteria in this experiment are identical to those in the previous one, with a total of 4 virtual channels and a 16-bit per flit.

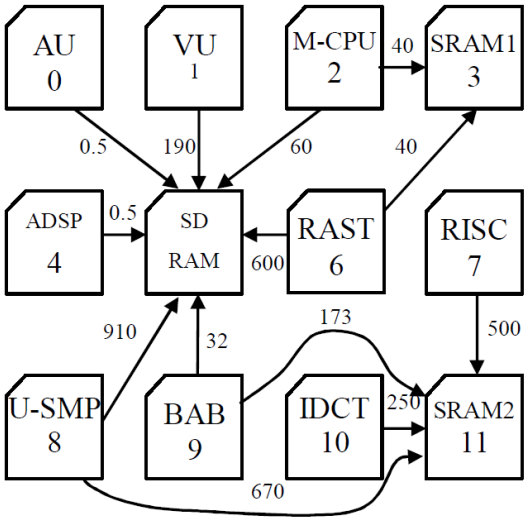


Fig. 16: Mapping of MPEG-4 core graphs to a 3×4 Mesh Topology NOC [36].

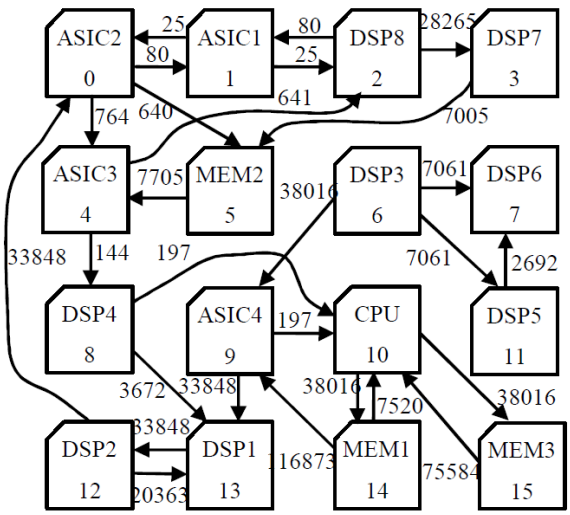


Fig. 17: Mapping of AV core graphs to 4×4 Mesh Topology NOC [36].

The delay occurs throughout the transmission of packets to all destinations, such as MPEG-4 and AV targets, with corresponding packet sizes of 55,472 and 380,128. The disparities among these three Network-on-Chip (NoC) applications are insignificant, primarily due to two specific circumstances. Initially, we configured VC-4 for every input port, surpassing the maximum number of required VCs in these applications. Fig. 16 and Fig. 17 show the most requested VCs in MPEG-4, AV, and 3 and 2. Furthermore, the packet traverses many ways.

They are not experiencing congestion. For example, the input channel on the western side of MPEG-4 router #11 has the most significant packet flow.

According to Fig. 18 and Fig. 19, the findings indicate that the delay caused by RR-SFVP is 72%, 92%, and 78% less than the average delay caused by LLD-Matrix for MPEG-4, AV, and applications.

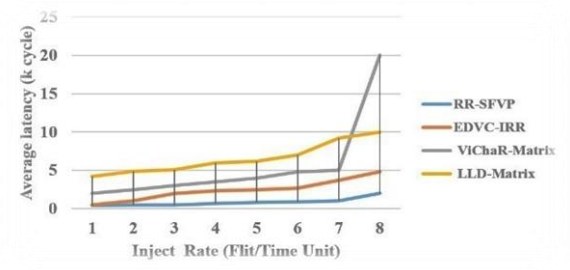


Fig. 18: Latency for MPEG-4, for 3×4 mesh NoCs.

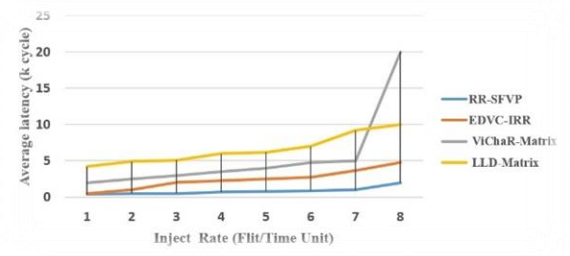


Fig. 19: Latency for AV, for 4×4 mesh NoCs.

Results and Discussion

In this article, a method for having a network router unit on a chip was proposed, which had several advantages over the previous techniques, the most important of which was its variable priority, which means that input does not always have the highest priority, even though the design circuit It is done sequentially, that is, the system is designed in such a way that fairness is considered among all requests. The second and most important advantage of the designed circuit is to save the last granted virtual channel ID, which, in addition to keeping the granted virtual channel ID, saves time in the next cycle to check the requested virtual channels. The third advantage of the designed circuit is shortening the critical path, saving area, and power consumption.

Conclusion

This study introduces the RR-SFVP microarchitecture, a modified version of the RR technique. The buffer employs an arbitration unit to pick a port from several ports based on essential priorities among various virtual channels (VCs).

This ensures that no port is deprived of resources and that no port is given greater priority than others. Assessments indicate that the RR-SFVP approach exhibits lower area and power usage than other methods. Compared to the RR arbitration unit, one of the conventional techniques, it has achieved a 55.1% reduction in critical route latency and a 12.1% drop in power consumption, improving space efficiency by 4.3%. However, the simulation findings demonstrate that the RR-SFVP approach, compared to the IRR method, a relatively recent arbitration method, exhibits a 2% reduction in area, a 3.7% decrease in power consumption, and a 10.88% decrease in critical route latency. Future work might include suggesting changes to the arbitration unit design, such as using shared comparison gates to decrease vital path delays.

Author Contributions

Elnaz Shafigh Fard conceived and designed the analysis and contributed data or analysis tools. Mohammad Ali Jabraeil Jamali performed the analysis. Mohammad Masdari collected the data. Kambiz Majidzadeh contributed to the interpretation of the results. All authors reviewed the final manuscript.

Acknowledgment

The article's authors have not received any financial support from any organization concerning this research.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

NOC	*Network On Chip*
VC	*Virtual Channel*
DAMQ	*Dynamically Allocated Multi-Queue*
ViChaR	*Virtual Channel Regulator*
EDVC	*Efficient Dynamic Virtual Channel*
FIFO	*First-in First-out*
FPGA	*Field programmable gate array*
HoL	*Head Of Line blocking problem*
HDRA	*High-speed and Decentralized Round robin*
IPRRA	*Improved parallel round-robin arbiter*
IRR_WF	*Index round-robin weak fairness*
HDRA	*High-speed and decentralized round-robin arbiter*
LLD	*Linked-List based DAMQ*
MPEG	*Moving Picture Experts Group*
AV	*Audio-Video Benchmark*

References

[1] L. Benini, G. D. Micheli, T. Tao Ye, Networks on chips, Vol. 1. Burlington: Morgan Kaufmann, 2006.

[2] R. Akbar, F. Safaei, “A novel heterogeneous congestion criterion for mesh-based networks-on-chip,” Microprocess. Microsyst., 84: 104056, 2021.

[3] A. T S, B. M, U. Sankar S M, R. Thiagarajan, A. Daltan G, P. Raja Rajeshwari, A. Sai Kumar et al., “Evaluation of low power consumption network on chip routing architecture,” Microprocess. Microsyst., 82: 103809, 2021.

[4] F.Rad, M. Reshadi, A. Khademzadeh, “A novel arbitration mechanism for crossbar switch in wireless network-on-chip,” Cluster Comput., 24: 1185-1198, 2021.

[5] S. M. Mamaghani, M. A. Jabraeil Jamali, “A load-balanced congestion-aware routing algorithm based on time interval in wireless network-on-chip,” J. Ambient Intell. Hum. Comput., 10: 2869-2882, 2019.

[6] M. Nirmala, V. Shylaja, “FPGA implementation of priority arbiter-based router design for NOC systems,” Int. J. Adv. Res. Eng. Technol. (IJARET), 10(2): 509-516, 2019.

[7] M. Rezaei-Zare, M. Fathy, A. Rezaei-Z, “Design of a high-performance router with distributed shared-buffer for load balancing for on-chip networks,” Microelectronics Journal, 132(c): 105647, 2023.

[8] B. Bahrami, M. A. Jabraeil Jamali, S. Saeidi, “ A novel hierarchical architecture for wireless network-on-chip,” J. Parallel Distrib. Comput., 120: 307-321, 2018.

[9] Y. Long Lan, V. Muthukumar, “Efficient virtual channel allocator for noc router microarchitecture,” in Proc. 30th IEEE International System-on-Chip Conference (SOCC), 2017.

[10] M. T. Balakrishnan, T. G. Venkatesh, A. Vijaya Bhaskar, “Design and implementation of congestion aware router for network-on-chip,” Integration, 88: 43-57, 2023.

[11] W. Zhou, Y. Ouyang, Y. Lu b, H. Liang, “A router architecture with dual input and output channels for Networks-on-Chip,” Microprocess. Microsyst., 90: 104464, 2022.

[12] L. Liu, Z. Zhu, D. Zhou, Y. Yang, “A fair arbitration for Network-on-Chip routing with odd-even turn model,” Microelectron. J., 64: 1-8, 2017.

[13] J. Rostami Monfared, A. Mousavi, “Design and simulation of nano-arbiters using quantum-dot cellular automata,” Microprocess. Microsyst., 72: 102926, 2020.

[14] R. Uma, H. Sarojadevi, V. Sanju, "Network-On-Chip (NoC) - routing techniques: A Study and analysis," in Proc. 2019 Global Conference for Advancement in Technology (GCAT), 2019.

[15] S. Sreekumar, “Literature survey on Network on Chips,” J. Emerging Technol. Innovative Res., 8(3): 1836-1840, 2021.

[16] W. J. Dally, B. Towles, Principles and Practices of Interconnection Networks, Morgan Kaufmann Publishers, Elsevier, 2004.

[17] J. K. Murthy, Nishanth. B. N, “Design and analysis of arbiters for the NoC's routers,” Int. J. Adv. Res. Electr., Electron. Instrum. Eng., 6(9): 6785-6792, 2017.

[18] A. Osovsky, D. Starov, O. Stukach, N. Maltseva, D. Surkov, “Crossbar switch arbitration with traffic control for NoC,” in Proc. International Siberian Conference on Control and Communications (SIBCON): 1-5, 2022.

[19] A. Toe, S. A. Dianat, Design and Verification of a Round-Robin rbiter, Theses 8-2018.

[20] T. Nadu, J. Arjana, “Low latency Noc with dynamic priority based matrix arbiter,” Indian J. Sci. Technol., 9(29), 2016.

[21] M. Evripidou, C. Nicopoulos, V. Soteriou, J. Kim, "Virtualizing virtual channels for increased network-onchip robustness and upgradeability," in Proc. IEEE Computer Society Annual Symposium on VLSI, 2012.

[22] C. A. Nicopoulos, P. Dongkook, K. Jongman, N. Vijaykrishnan, M. S. Yousif, C. R. Das, "ViChaR: A dynamic virtual channel regulator for network-on-chip routers," in Proc. 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06), 2006.

[23] Z. Fu, X. Ling, “The design and implementation of arbiters for Network-on-chips,” in Proc. 2nd Int. Conf. Industrial and Information Systems: 292-295, 2010.

[24] S. Q. Zheng, M. Yang, “Algorithm-Hardware codesign of fast parallelround-robin arbiters," IEEE Trans. Parallel Distrib. Syst., 18(1): 84-95, 2006.

[25] Y. L. Lee, J. M. Jou, Y. Y. Chen, "A high-speed and decentralized arbiter design for NoC," in Proc. IEEE/ACS Int. Conf. on Computer Systems and Applications: 350-353, 2009.

[26] M. Oveis-Gharan, G. N. Khan, "Index-based round-robin arbiter for NoC routers," in Proc. IEEE Computer Society Annual Symposium on VLSI: 62-67, 2015.

[27] O. Gharan, G. N. Khan, "Efficient dynamic virtual channel organization and architecture for NoC systems," IEEE Trans. Very Large Scale Integr. VLSI Syst., 24(2): 465-478, 2016.

[28] J. Luo, W. Wu, Q. Xing, M. Xue, F. Yu, Z. Ma, “A low-latency fair-arbiter architecture for network-on-chip switches,” Appl. Sci., 12(23): 12458, 2022.

[29] A. Mohanty, G. Chiranjeevi, “Comparative analysis of round robin arbiter and weighted round robin arbiter,” in Proc. 2024 Asia Pacific Conference on Innovation in Technology (APCIT), 2024.

[30] X. Cheng, Y. Wang, H. Jin, P. Li, “A novel bus arbiter based on group decision algorithm,” in Proc. IEEE 4th International Conference on Circuits and Systems, 2022.

[31] E. Shafigh Fard, M. A. Jabraeil Jamali, M. Masdari, K. Majidzadeh, “An efficient NoC router by optimal management of buffer read and write mechanism,” Microprocess. Microsyst., 89: 104440, 2022.

[32] S. Das, J. R. Doppa, P. P. Pande et al., “Energy-efficient and reliable 3D network-on-chip (NoC): Architectures and optimization algorithms,” in Proc. IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2016.

[33] N. Alfaraj, J. Zhang, Y. Xu, H. J. Chao, "HOPE: Hotspot congestion control for clos network on chip," in Proc. IEEE/ACM International Symposium on Networks on Chip: 17-24, 2011.

[34] Y. Xu, B. Zhao, Y. Zhang, J. Yang, "Simple virtual channel allocation for high throughput and high-frequency on-chip routers," in Proc. HPCA-16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture: 1-11, 2010.

[35] V. Dumitriu, G. N. Khan, "Throughput-oriented NoC topology generation and analysis for performance SoCs," IEEE Trans. VLSI Syst., 17(10): 1433-1446, 2009.

[36] M. Oveis-Gharan, G. N. Khan, "Statically adaptive multi fifo buffer architecture for network on chip," Microprocess. Microsyst., 39(1): 11–26, 2015.

Biographies

Elnaz Shafigh Fard received a B.Sc. degree in Software Engineering from the Azad University of Shabestar, Iran, in 2001 and an M.S. degree in Computer Systems Architecture from the Azad University of Najaf Abad branch, Isfahan, Iran, in 2014. She is currently pursuing a Ph.D. degree at the Azad University of Urmia branch, Urmia, Iran. Her research interests include embedded system design, intelligent system design and modeling, system-on-chip design, and performance and latency optimization in network-on-chip architectures.

- Email: elnazsha9427@gmail.com
- ORCID: 0009-0004-2749-4044
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mohammad Ali Jabraeil Jamali received his B.Sc. in Electrical Engineering from Urmia University, Urmia, Iran, his M.Sc. from Tabriz University, Tabriz, Iran, the M.Sc. in Computer Engineering from Islamic Azad University, Science and Research Branch, Tehran, Iran and the Ph.D. in Computer Engineering from Islamic Azad University, Science and Research Branch, Tehran, Iran, in 1994, 1997, 2003 and 2009, respectively. He is an Assistant Professor of Computer Engineering at Islamic Azad University, Shabestar branch. He is the author/co-author of over 50 technical journals and conference publications. His current research interests are processor and computer architectures, chip multiprocessors, multiprocessor systems-on-chip, networks-on-chip, ad hoc and sensor networks, security, and the Internet of Things.

- Email: m_jamali@itrc.ac.ir
- ORCID: 0000-0001-7687-5469
- Web of Science Researcher ID: NA
- Scopus Author ID: 57193977114
- Homepage: https://scimet.iau.ir/MohammadAli_JabraeilJamali?%20highlyCited

Mohammad Masdari received his B.Tech. He earned a degree in Computer Software Engineering from Islamic Azad University, Qazvin Branch, Iran, in 2001 and an M.Tech degree in Computer Software Engineering from Islamic Azad University, South Tehran Branch, Tehran, Iran, in 2003. He received his Ph.D. in Computer Software Engineering from Islamic Azad University, Science and Research Branch, Tehran, Iran, in 2014. Since 2003, he worked as a faculty member of Islamic Azad University, Urmia branch, Iran. He is an Assistant Professor in the Department of Computer Engineering of Islamic Azad University, Urmia branch, Iran. His research interests include Distributed Systems and Network Security.

- Email: mo.masdari@iau.ac.ir
- ORCID: 0000-0002-7093-2204
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://urmia.iau.ir/file/download/faculty/64f44124367a5-masdari.pdf

Kambiz Majidzadeh was born in Urmia, Iran, in 1980. He received a B.Sc. degree in Software Engineering from the Islamic Azad University of Khoy, Khoy, Iran, in 2002, M.Sc. degree in Computer Networking, and Ph.D. degree in Information Technology from Baku State University (BSU), Baku, Azerbaijan, in 2005 and 2009, respectively. His research interests are very large-scale.

- Email: k.majidzadeh@iaurmia.ac.ir
- ORCID: 0000-0001-5118-8910
- Web of Science Researcher ID: NA
- Scopus Author ID: 54880938600
- Homepage: https://urmia.iau.ir/file/download/faculty/67063cb4bdbd9-majidzadeh.pdf

How to cite this paper:
E. Shafigh Fard, M. A. Jabraeil Jamali, M. Masdari, K. Majidzadeh, "RR-SFVP: A novel arbitration unit design for NoC router, ingeniously fusing the round robin method with strong fairness and variable priority," J. Electr. Comput. Eng. Innovations, 13(2): 283-298, 2025.

DOI: 10.22061/jecei.2024.11230.779

URL: https://jecei.sru.ac.ir/article_2238.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Circuit Analog Absorber Based on a Double-Layer of Resistor-Loaded Strip Arrays with Various Bandwidths according to Selecting the Polarization

S. Barzegar-Parizi *

Department of Electrical Engineering, Sirjan University of Technology, Sirjan, Iran.

Article Info

Article History:
Received 03 September 2024
Reviewed 17 November 2024
Revised 12 December 2024
Accepted 15 December 2024

Keywords:
Resistor-loaded strips
Microwave absorber
Circuit model
TM
TE
FSS
Reflectivity

*Corresponding Author's Email Address:
barzegarparizi@sirjantech.ac.ir

Abstract

Background and Objectives: The design of the circuit analog absorbers including resistive and conductive patterns on a dielectric substrate placed above the ground plane with a free spacer is interesting for researchers in the microwave regime. Broad absorption band can be achieved by appropriately designing the structure parameters that lead to matching the input impedance of the structure with the impedance of free space over a wide operating band. In this study, a wideband circuit analogue absorber including double-layer of resistive frequency selective surfaces (FSS) is proposed.

Methods: The proposed structure is composed of two layers of periodic arrays of strips loaded with lumped resistors deposited on dielectric substrates and separated by an air spacer. Strips of each layer are orthogonal to each other. The structure is placed on a metallic back reflector with an air spacer. The bottom resistive FSS including resistor-loaded strips directed in the *x*-direction plays the effective role of producing the resonant frequencies with exciting TM polarization waves and leads to a wide high-frequency absorption band, while the top resistive FSS, including resistor-loaded strips directed in the *y*-direction plays the effective role in exciting the resonances for TE polarization that can produce a broad low frequency absorption band. Indeed, in each polarization, one of the resistive FSS acts as a resonator while the other resistive FSS acts as a transparent layer and transmits the wave. A circuit model for characterizing the proposed structure is presented for both TE and TM polarizations in the subwavelength regime, which shows good agreement with the full-wave simulations.

Results: The results demonstrate that the reflectivity below −10 dB (absorption above 90%) obtains from 3.55 to 9.82 GHz (fractional bandwidth of 93%) under normal incidence for TE polarization while with TM incident wave excitation, the absorption above 90% from 9.44 to 20.85 GHz (fractional bandwidth of 75%) can be achieved.

Conclusion: The proposed structure leads to a wideband absorber with various bandwidths corresponding to exciting TE and TM incident waves. Most of the proposed structures in the literature produce similar bandwidths for both polarizations. Therefore, a polarization-controlled wideband absorber is designed in this task.

Introduction

Electromagnetic absorbers have found many potential applications in different systems. The wide bandwidth over the operating band, and smaller thickness are the essential parameters, in the design of microwave absorbers. A resistive sheet placed a quarter-wavelength distance above the conducting plate [1]-[2] known as the Salisbury screen, was presented decades ago to reduce the reflection and result in absorbing the incident electromagnetic wave. Despite its structural simplicity,

due to the conforming of quarter-wavelength conditions at a single specified frequency, the absorption bandwidth of the Salisbury screen was relatively narrow. Jaumann absorber utilizing additional resistive layers and spacers was introduced [3]-[4] to improve the bandwidth of the Salisbury screen. However, increasing the resistive layers and spacers enhances the total thickness and it limits its scope of application. The circuit analog (CA) absorbers were proposed for achieving electromagnetic absorbers with wide bandwidth and small thickness [5]-[30]. The circuit analog absorbers are made by depositing conductive/resistive patterns or resistor-loaded patterns on a dielectric layer placed above a metallic back reflector with a free spacer. By appropriately designing the structure parameters and choosing the chip resistors, the input impedance of the structure could be matched with free space impedance over a wide operating band, and the broad absorption band occurs.

In [5]-[9], the structures based on the conductive patterns and dielectric layers have been proposed to achieve the narrow and wide absorption bands. For example, in [6], a single layer of copper FSS as swastika-like patterns has been employed to design a narrow band absorber. The multilayered structures of metallic loops and closed ring resonators have been proposed in [7] and [8], respectively, for achieving broadband and dual-band absorbers. In [9], a multilayered structure of crossed dipoles has been proposed to realize triple absorption bands. In [10]-[14], the frequency selective surfaces including resistive patterns are employed to realize the broad absorption bands. The resistive treble-square loops, resistive crisscross and fractal square patches and resistive quadruple hexagonal loops have been respectively utilized to realize broadband absorbers in [11], [12], and [13]. In [15]-[30], resistor-loaded patterns have been applied to achieve broadband absorbers. In [16] and [21]-[22], broadband absorbers based on square loops loaded by lumped-resistors have been presented. Lumped resistor loaded double octagonal rings have been employed to realize wideband absorber in [17]. In [23], a single layer of a modified circular ring and in [24], a single layer of double patterns of rectangular and ring split ring resonators loaded by lumped resistors have been applied to achieve wideband absorbers. In recent tasks, the researchers have designed absorbers consisting of multiple vertically stacked FSS layers to increase the bandwidth. In [28], a structure using a dual layer of resistor-loaded metallic strips has been proposed to achieve a wide absorption band for both TE and TM modes. The structure includes the lossy layer consisting of two orthogonal layers of dual-resistor-loaded metallic strips printed on both sides of a dielectric substrate. In [29], a polarization-insensitive wideband absorber has been proposed based on a multi-layer of square loops loaded with lumped resistors printed on the dielectric layers separated by an air spacer. In this task, the bottom resistive surface in combination with the top resistive surface, enhances the bandwidth by creating another resonance. A polarization-insensitive circuit analog absorber containing two lossy layers of a single square-loop and double-square-loop loaded with lumped resistors has been designed in [30] to obtain an ultra-wide absorption band. All of these aforementioned absorber designs are polarization-insensitive and capable of absorbing waves for both polarizations in the same absorption band. Therefore, the design of an absorber with selectivity bandwidth according to the polarization is interesting. In [31]-[34], the polarization-controlled structures that display the various absorption bands according to the selection of the polarization have been presented. However, the proposed structures present narrow absorption bands with exciting TE and TM incident waves.

In this paper, a wideband absorber is designed with various bandwidth range according to exciting each polarization. The proposed structure is composed of double layers of resistor-loaded metallic strips array printed on a dielectric substrate. The strips of one layer are orthogonal to the other layer. Two layers are separated by an air spacer, and then the bottom layer is placed above a metallic film with another air spacer. The various wide absorption bands can be achieved with the selection of polarization. The strips of the top resistive layer are arranged in y-direction while the strips of the bottom resistive layer are arranged in x-directions. Therefore, by properly designing the geometrical parameters and choosing the chip resistors of the structure, the structure can absorb the incident waves for TE polarization at low frequencies between ranges of 3.55 to 9.82 GHz, while it can absorb the incident waves for TM polarization at higher frequency between ranges of 9.44 to 20.85 GHz. An equivalent circuit model is introduced for both polarizations. Therefore, by changing the polarization, the absorption bandwidth would change. The following paper is organized as follow: the structure and analysis are presented in section 2. The proposed structure leads to various absorption bands with exciting TE and TM incident waves. The equivalent circuit model is presented for both polarizations. It's demonstrated that the top resistive layer plays the resonator role modeled as two series R-L-C branches by exciting TE mode while the situation is vice versa for TM mode. For TM polarization, the bottom resistive layer is modeled as two series R-L-C branches. Finally, Section 3 presents the main conclusions.

Structure and Design

A single unit cell of the proposed structure, which comprises two resistive frequency selective surfaces printed on dielectric layers is displayed in Fig. 1. The thickness of dielectric spacers is defined as h_{d1} and h_{d2}.

These layers are separated by an air spacer with a thickness of h_{s1} and h_{s2}. The overall structure has been terminated by a metal film acting as a back reflector.

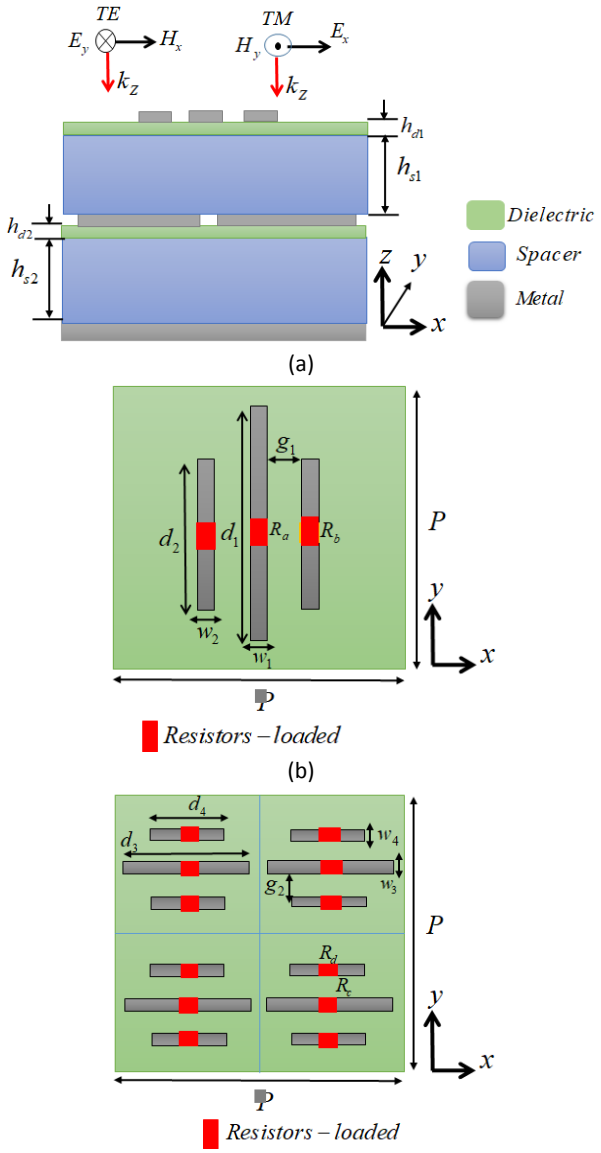


Fig. 1: (a) Perspective view of the structure including a double layer of the resistive frequency selective surfaces deposited on a dielectric substrate placed on a spacer and a metallic reflector at the bottom. (b)Top view of the top resistive FSS including three resistor-loaded strips directed in *y*-direction (c) Top view of the bottom resistive FSS including four cells of three resistor-loaded strips directed in *x*-direction.

The unit cell of the top resistive frequency selective surface (TRFSS) is built of three resistor-loaded strips directed in the *y*-direction. The length and width of the central strip is d_1, w_1 and two neighbor strips with the length and width of d_2, w_2 are placed at the right and the left of the central strip with space of g_1. Lumped resistors are placed at the center of each strip with values of R_a and R_b. The bottom resistive frequency selective surface (BRFSS) is composed of four sub unit cells of three resistor-loaded strips directed in *x*-direction. The length and width of the central strip is d_3, w_3, and two neighbor strips with the length and width of d_4, w_4 are placed at the right and the left of the central strip with space of g_2. Lumped resistors are placed at the center of each strip with values of R_c and R_d. RT/Duroid5880 with a relative permittivity of 2.2 has been used as dielectric. Copper with conductivity $\sigma = 5.8 \times 10^7$ S/m and a thickness of 0.02 mm is considered for the metal strips and the bottom metallic film. The period of the structure is supposed to be P in *x*- and *y*-directions.

Table 1: Detailed unit cell parameters of the proposed structure

Description of Parameter	Symbol	Value
Period length	*P*	*26 mm*
Thickness of top dielectric layer	h_{d1}	*0.45 mm*
Thickness of bottom dielectric layer	h_{d2}	*0.2 mm*
Thickness of top free spacer	h_{s1}	*5 mm*
Thickness of bottom free spacer	h_{s2}	*5 mm*
Central strip length of the TRFSS	d_1	*25mm*
Central strip width of the TRFSS	w_1	*0.5 mm*
smaller strip length of the TRFSS	d_2	*15 mm*
smaller strip width of the TRFSS	w_2	*0.5 mm*
Central strip length of the BRFSS	d_3	*11 mm*
Central strip width of the BRFSS	w_3	*0.4 mm*
smaller strip length of the BRFSS	d_4	*5 mm*
smaller strip width of the BRFSS	w_4	*0.4 mm*
gap between strips of the TRFSS	g_1	*3.5 mm*
gap between strips of the BRFSS	g_2	*1.5 mm*
Lumped resistor at the center of longer strips of the TRFSS	R_a	*125 Ω*
Lumped resistor at the center of smaller strips of the TRFSS	R_b	*100 Ω*
Lumped resistor at the center of longer strips of the BRFSS	R_c	*75 Ω*
Lumped resistor at the center of smaller strips of the BRFSS	R_d	*50 Ω*

The proposed structure leads to wideband absorption bands by exciting TE and TM incident waves. By exciting the TE incident wave where the electric field is in the *y*-direction and the magnetic field is in the *x*-direction, the top resistive frequency selective surface plays an essential role in exciting the resonant frequencies. It leads to a low-frequency wide absorption band. In this case, the bottom resistive FSS acts as a transparent layer and transmits the wave. At the same time, a high-frequency absorption band can be achieved by exciting TM incident waves. In this case, the bottom frequency selective surface plays an essential role in exciting the resonant frequencies, and the top resistive FSS transmits the wave. In Table. 1, the

parameters of a single unit cell of the proposed structure are presented. Fig. 2(a) and (b) show the simulated reflectivity and absorption of the proposed absorber under the normal incidence for TE polarization. The simulated results display the reflectivity below −10 dB (equal to absorption above 90%) from 3.55 to 9.82 GHz with a fractional bandwidth of 93%. Fig. 2(c) and (d) demonstrate the simulated reflectivity and absorption of the resistive absorber under the normal incidence for TM polarization. The simulated results show that the reflectivity below −10 dB (equal to absorption above 90%) occurs from 9.44 to 20.85 GHz with fractional bandwidth of 75%.

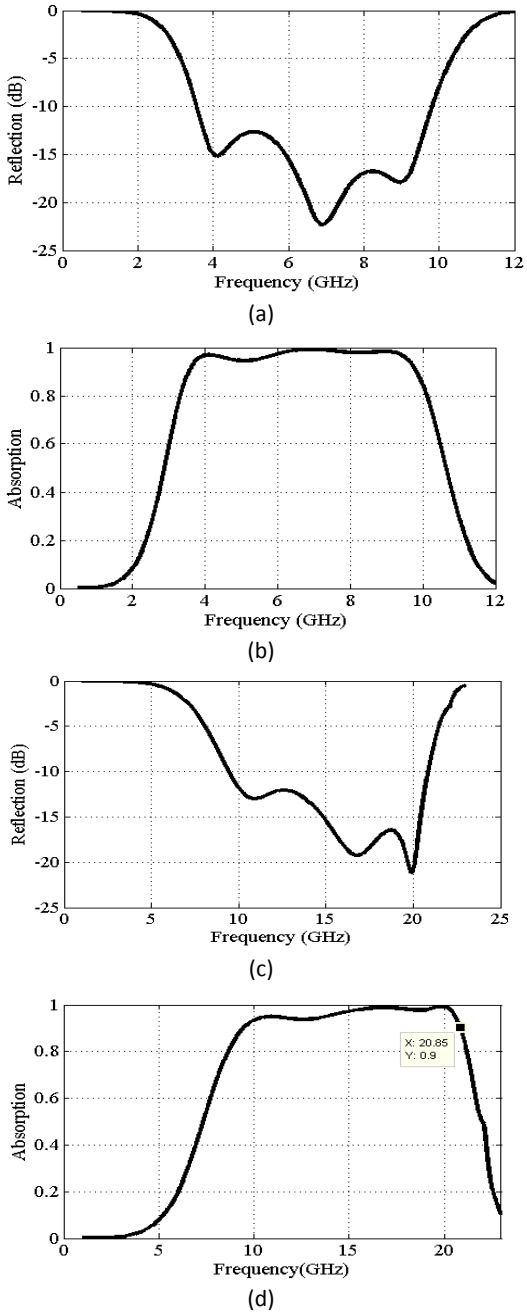


Fig. 2: Simulated (a) reflectivity (b) absorption spectra of the structure of Fig.1 for TE polarization (c) reflectivity (d) absorption spectra for TM polarization.

In Fig. 3, the variations of the lumped resistors loaded on strips are surveyed on the absorption spectra. Fig. 3(a) shows the absorption spectra of the proposed structure for TE polarization, when the lumped resistors of the top resistive layer are varied. As observed, with the selection of R_a= 100 Ω, R_b=75 Ω and values more than them, high absorption can be achieved. The variations of the lumped resistors of the bottom resistive layer on absorption spectra for TM polarization are demonstrated in Fig. 3(b).

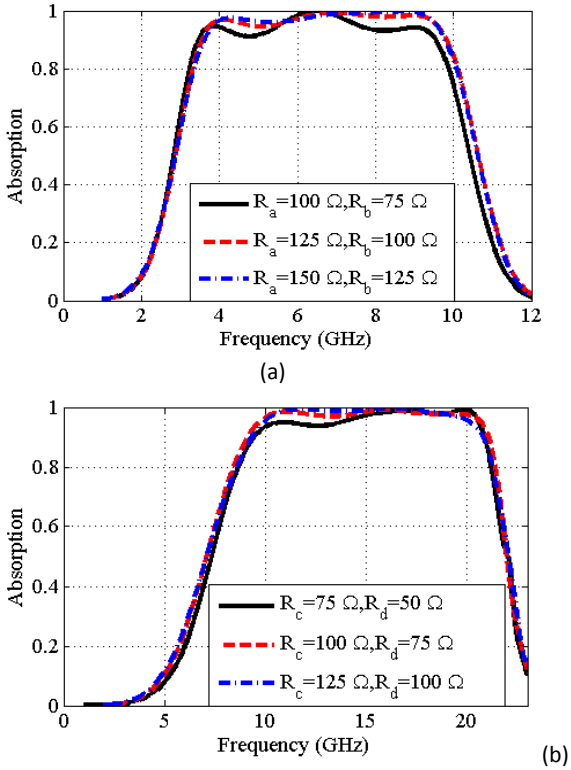


Fig. 3: Simulated absorption spectra of the structure with various lumped resistors (a) for TE polarization (b) for TM polarization.

Furthermore, to better understand the absorber behavior, the equivalent circuit models are presented for both TE and TM modes in the subwavelength regime. The circuit model corresponding to the proposed absorber is illustrated in Fig. 4(a) for the TE polarization wave and (b) for the TM polarization wave, respectively. In the TE case, the top frequency selective surface plays an essential role in exciting the resonant frequencies modeled as two branches of series resistor-inductor-capacitor (RLC) circuits connected in parallel. R_1, L_1, and C_1 show the resistance, inductance, and capacitance corresponding to the central metallic strip, respectively, while R_2, L_2, and C_2 represent those of smaller metallic strips. The values of the lumped elements corresponding to the resistor-loaded strips of top frequency selective surfaces for TE polarization are defined as: C_1=0.06 pF, L_1=19 nH, $R_1 = 230\,\Omega$, C_2=0.027 pF, L_2=16.5 nH and $R_2 = 300\,\Omega$. In this case, the bottom resistive FSS of the resistor-loaded strips acts as a transparent layer and transmits the wave. In TM case, the situation is vice versa. In this case, the

bottom frequency selective surface plays an essential role in exciting the resonant frequencies which are modeled as two branches of series resistor-inductor-capacitor (RLC) circuits connected in parallel. R_3, L_3, and C_3 show the resistance, inductance, and capacitance equivalent to the central metallic strip, respectively, while R_4, L_4, and C_4 specify those of smaller metallic strips. The values of the lumped elements [21] corresponding to the resistor-loaded strips of the bottom frequency selective surfaces for TM polarization are defined as: C_3=0.02 pF, L_3=10.5 nH, $R_3 = 230\,\Omega$, C_4=0.008 pF, L_4=9.7 nH and $R_4 = 300\,\Omega$. In this case, the top resistive FSS of the resistor-loaded strips acts as a transparent layer and transmits the wave. To demonstrate the role of the top resistive FSS as a transparent layer for TM polarization, the S-parameters of a single layer of top resistive FSS are plotted in Fig. 5 for TM polarization. Two ports are considered at the top and bottom of this layer for computing the S-parameters. The transmission and reflection coefficients are plotted in Fig. 5. As observed, this layer transmits the waves for TM polarization (S_{12} is 0 dB) and acts as a transparent layer.

The dielectric layers are modelled by the transmission lines of the characteristic admittance of $Y_d = \sqrt{\varepsilon_r} Y_0$ and the propagation constant of $\beta_d = \beta_0 \sqrt{\varepsilon_r}$ with lengths of h_{d1} and h_{d2}. The air spacers are modelled by the transmission lines of the characteristic admittance of Y_0 and the propagation constant of β_0 with lengths h_{s1} and h_{s2}. The bottom metallic film is modelled by the short circuit in the equivalent circuit model.

In TE case, the input admittance of the structure is computed as:

$$Y_{in} = Y_{sur,1} + Y_{sur,2} + Y_{slab}, \tag{1}$$

$$Y_{sur,i} = \frac{1}{R_i + j(\omega L_i - \frac{1}{\omega C_i})}, \ (i = 1, 2) \tag{2}$$

$$Y_{slab1} = j \frac{Y_d \left(Y_d \tan(\beta_d h_{d2}) - Y_0 \cot(\beta_0 h_{s2}) \right)}{Y_d + Y_0 \cot(\beta_0 h_{s2}) \tan(\beta_d h_{d2})} \tag{3}$$

$$Y_{slab2} = Y_0 \frac{\left(Y_{slab1} + jY_0 \tan(\beta_0 h_{s1}) \right)}{Y_0 + jY_{slab1} \tan(\beta_0 h_{s1})} \tag{4}$$

$$Y_{slab} = Y_d \frac{\left(Y_{slab2} + jY_d \tan(\beta_d h_{d1}) \right)}{Y_d + jY_{slab2} \tan(\beta_d h_{d1})} \tag{5}$$

where $Y_{sur,i}$ (i=1,2) is the admittance of the strips loaded with lumped resistors and Y_{slab} is the equivalent admittances of the conductor-backed dielectric slabs and air spacers.

Finally, the values of the absorption can be computed as:

$$A(\omega) = 1 - R(\omega) = 1 - \left| \frac{(Z_{in}/Z_0) - 1}{(Z_{in}/Z_0) + 1} \right|^2 \tag{6}$$

In which, $Z_{in} = Y_{in}^{-1}$ and Z_0 is the free-space impedance. When the impedance matching conditions over a specified frequency range occur, the maximum absorption can be obtained on this frequency range.

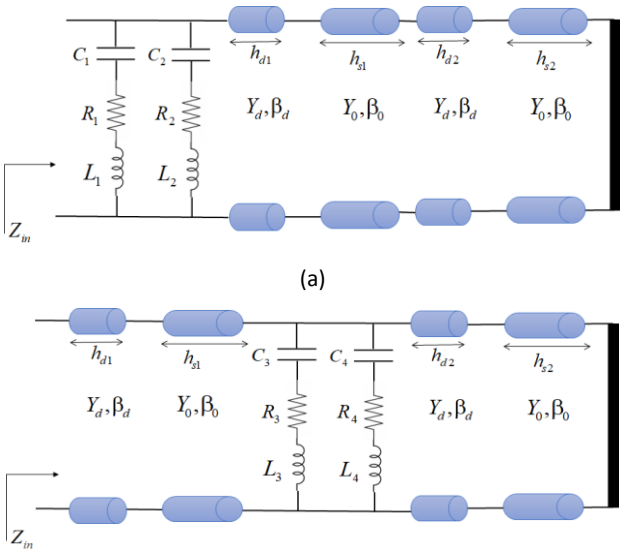


Fig. 4: The equivalent circuit model of the proposed structure for (a) TE and (b) TM polarizations in subwavelength regime.

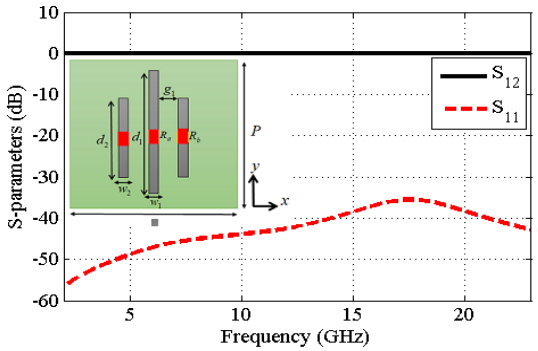


Fig. 5: The S-parameters of a single layer of top resistive FSS for TM polarization.

The comparison between the result extracted by the circuit model analysis and the results obtained by HFSS simulations are displayed in Fig. 6(a) for the proposed structure with parameters presented in Fig. 2 for TE polarization. As demonstrated, the result obtained by the circuit model is in good agreement with the full-wave simulation results.

In the TM case, the input admittance is computed as:

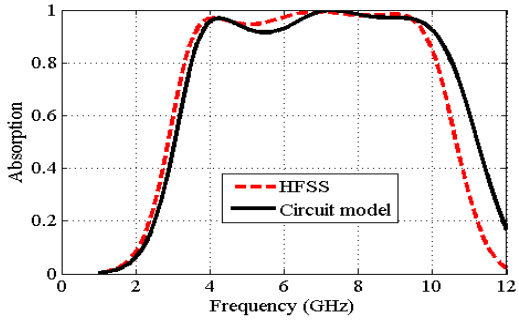
$$Y_{in} = Y_d \frac{\left(Y_{slab4} + jY_d \tan(\beta_d h_{d1}) \right)}{Y_d + jY_{slab4} \tan(\beta_d h_{d1})} \tag{7}$$

$$Y_{slab4} = Y_0 \frac{\left(Y_{slab3} + jY_0 \tan(\beta_0 h_{s1}) \right)}{Y_0 + jY_{slab3} \tan(\beta_0 h_{s1})} \tag{8}$$

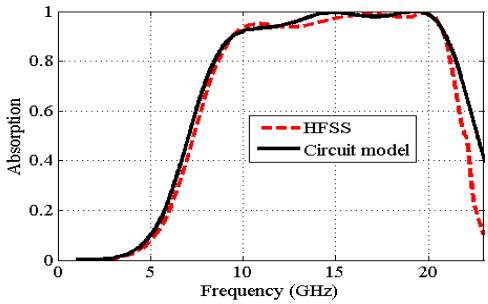
$$Y_{slab3} = Y_{sur,3} + Y_{sur,4} + Y_{slab1}, \tag{9}$$

$$Y_{sur,i} = \frac{1}{R_i + j(\omega L_i - \frac{1}{\omega C_i})}, (i = 3,4) \quad (10)$$

The result extracted by the circuit model is compared to the HFSS simulation result for TM polarization in Fig. 6(b).



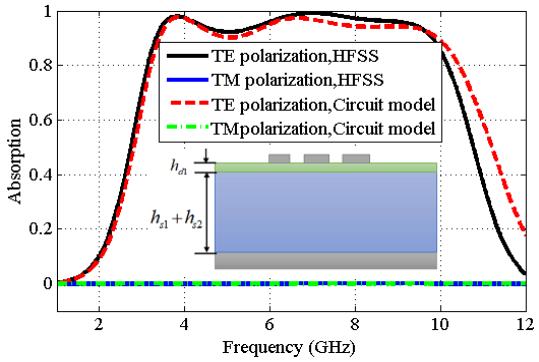
(a)



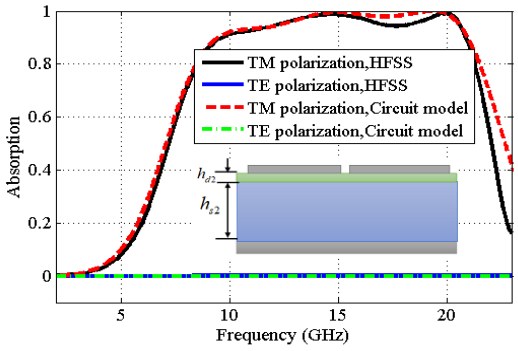
(b)

Fig. 6: the results extracted by the circuit model are compared to the HFSS simulations for the proposed structure for (a) TE polarization (b) TM polarization.

However, it is clear that without each of the resistive layers and its underlying dielectric, the absorption occurs for one of the polarizations. For example, if one puts up the top resistive layer and its underlying dielectric above the metallic film with an air spacer of thickness of $h_{s1} + h_{s2} = 10$ *mm,* the new structure absorbs the incident waves for just TE polarization. Suppose the bottom resistive layer and its underlying dielectric have been placed above the metallic film with an air spacer of thickness of $h_{s2} = 5$ *mm*. In that case, the structure absorbs the incident waves for just TM polarization. Fig. 7 shows the absorption spectra for structures without one of the resistive FSS for TE and TM polarizations.



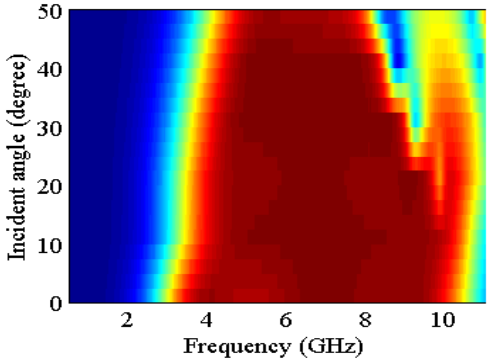
(a)



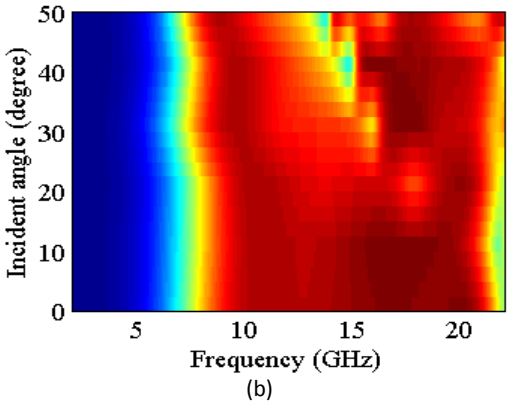
(b)

Fig.7: Simulated absorption spectra of the structure without considering (a) the bottom resistive layer and its underlying dielectric (b) the top resistive layer and its underlying dielectric and top air spacer under the normal incidence.

In the following, the performance of the structure with respect to the incident angle is investigated. As mentioned, the performance depends on polarization. For each polarization, the structure leads to a specified absorption spectra. Here, the structure's performance concerning the angle of incidence for each polarization is evaluated. Fig. 8 illustrates the absorption spectra as a function of the frequency and the incident angle up to 50° for TE and TM polarizations, respectively. For both TE and TM polarizations, absorption above 80% can be achieved for the incident angles up to 30◦. For TE polarization, the bandwidth decrease with increasing the angle of incidence. For TM polarization, the absorption values are reduced in the middle of the bandwidth.



(a)



(b)

Fig. 8 The Absorption spectra of the proposed structure as a function of different incident angles of (a) TE polarization, and (b) TM polarization.

Finally, comparisons between the designed absorber and the other absorbers are performed in the Table. 2. The comparisons are performed in terms of structure and performance. As observed, the structures proposed in [28]-[30] are symmetric and produce a wide absorption band for both TE and TM polarizations. In the present task, the proposed structure causes the absorption performance depending on the polarization. The proposed structure leads to various bandwidths with exciting polarization. Although, the structures proposed in [31]-[34] present the polarization-controlled absorbers. However, the proposed structures are narrowband absorbers, while this task presents a wideband absorber.

Table 2: Performance comparison of proposed absorber to other task

Ref	Structure	Polarization-controlled absorber	Performance
[28]	Two layers of resistor-loaded metallic strips	No	wideband for both TE and TM (3-14 GHz)
[29]	Two layers of square loops loaded with lumped resistors	No	wideband for both TE and TM (4.96 to 18.22 GHz)
[30]	Two layers of square loops loaded with lumped resistors	No	Wideband for both TE and TM (1.92 to 16.87 GHz)
[31]	A metallic FSS of a circular enclosure containing T-shaped resonator	Yes	Quad narrow bands at 9.948, 13.26, 14.92, and 15.80 GHz for TE
[32]	A metallic FSS of split –ring resonators	Yes	Triple narrow bands at 2.4, 5.2 and 5.8, GHz for TE/ Quad-bands at 4.6, 5.3, 6.5, and 6.8 GHz
[33]	A set of wires etched as Yagi-Uda shaped FSS	Yes	A single narrow band at 6.64 GHz for TM / penta-bands at 11.68, 13.58, 15.48, 17.38, and 19.28 GHz for TE
[34]	Yagi-Uda shaped FSS	Yes	Triple narrow bands at 10.64, 12.08, 14.92, and 14.09 GHz for TE
This work	Two layers of resistor-loaded strips	Yes	Wideband from 3.55 to 9.82 GHz for TE/ Wideband from 9.44 to 20.85 GHz for TM

Conclusion

In this paper, a wideband absorber with various bandwidths associated with exciting TE and TM incident waves has been designed. The proposed structure is composed of two stacked resistive layers printed on a dielectric substrate. A single unit cell of the top resistive layer includes three resistor-loaded strips directed in the *y*-direction that lead to excite the resonances for TE polarization. In contrast, the bottom resistive layer is composed of four unit cells of three resistor-loaded-strips in the *x*-direction that leads to excite the resonances for TM polarization. In each polarization, other resistive layer transmitted the wave. The structure led to absorption above 90% from 3.55 to 9.82 GHz for TE polarization, while with TM incident wave excitation, the absorption above 90% from 9.44 to 20.85 GHz has been achieved. Hence, in this task, a polarization-controlled wideband circuit analog absorber has been proposed.

Author Contributions

S. Barzegar-Parizi designed and analyzed the structure. She interpreted the results and wrote the manuscript.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The author declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the author.

Abbreviations

FSS	Frequency Selective Surface
TE	Transverse Electric
TM	Transverse Magnetic
CA	Circuit Analog
TRFSS	Top Resistive Frequency Selective Surface
BRFSS	Bottom Resistive Frequency Selective Surface

References

[1] W. W. Salisbury, "Absorbent body of electromagnetic waves," US Patent 2599944, 1952.

[2] R. L. Fante, M. T. McCormack, "Reflection properties of the Salisbury screen," IEEE Trans. Antennas Propag., 36(10): 1443-1454, 1988.

[3] L. J. Du Toit, "design of jauman absorbers," IEEE Trans. Antennas Propag. Magazin, 36(6): 17-25, 1994.

[4] E. F. Knott, J. F. Shaeffer, M. T. Tuley, "Radar cross section," SciTech, Raleigh, NC, USA, 2nd ed., 2004.

[5] T. S. Pham, H. Zheng, L. Chen, B. X. Khuyen, Y. Lee, "Wide-incident-angle, polarization-independent broadband-absorption metastructure without external resistive elements by using a trapezoidal structure," Sci. Rep., 14: 10198, 2024.

[6] S. Ghosh, S. Bhattacharyya, K. V. Srivastava, "Bandwidth-enhanced of an ultra-thin polarization insensitive metamaterial absorber," Microw. Opt. Technol. Lett., 56(2): 350-355, 2014.

[7] H. Xiong,et al., "An ultrathin and broadband metamaterial absorber using multi-layer structures," J. Appl. Phys., 114: 064109, 2013.

[8] S. Bhattacharyya, S. Ghosh, D. Chaurasiya, et al., "Bandwidth-enhanced dual-band dual-layer polarization-independent ultra-thin metamaterial absorber," Appl. Phys. A, 118: 207-215, 2015.

[9] S. Ghosh, S. Bhattacharyya, D. Chaurasiya, et al., "Polarization-insensitive and wide-angle multilayer metamaterial absorber with variable bandwidths," Electron. Lett., 51(14): 1050-1052, 2015.

[10] F. Costa, et al., "Analysis and design of ultrathin electromagnetic absorbers comprising resistively loaded high impedance surfaces," IEEE Trans. Antennas Propag., 58(5): 1551-1558, 2010.

[11] M. Li, et al., "An ultrathin and broadband radar absorber using resistive FSS," IEEE Antennas Wirel. Propag. Lett., 11: 748-751, 2012.

[12] L. K. Sun, H. F. Cheng, Y. J. Zhou, et al., "Broadband metamaterial absorber based on coupling resistive frequency selective surface," Opt. Exp., 20(4): 4675-4680, 2012.

[13] S. N. Zabri, R. Cahill, A. Schuchinsky, "Compact FSS absorber design using resistively loaded quadruple hexagonal loops for bandwidth enhancement," Electron. Lett., 51(2): 162-164, 2015.

[14] P. C. Zhang, et al., "A wideband wide-angle polarization-insensitive metamaterial absorber," in Proc. Progress in Electromagnetics Research Symp., Guangzhou, China, 941-943, 2014.

[15] M. Yoo, S. Lim, "Polarization-independent and ultrawideband metamaterial absorber using a hexagonal artificial impedance surface and a resistor-capacitor layer," IEEE Trans. Antennas Propag., 62(5): 2652-2658, 2014.

[16] J. Yang, Z. Shen, "A thin and broadband absorber using double-square loops," IEEE Antennas Wirel. Propag. Lett., 6: 388-391, 2007.

[17] S. Li, et al. "Wideband, thin, and polarization-insensitive perfect absorber based the double octagonal rings metamaterials and lumped resistances," J. Appl. Phys, 116: 043710, 2014.

[18] W. Tang, Z. Shen, "Simple design of thin and wideband circuit analogue absorber," Electron. Lett., 43 (12): 689-691, 2007.

[19] Y. Z. Cheng, et al., "Design, fabrication and measurement of a broadband polarization-insensitive metamaterial absorber based on lumped elements," J. Appl. Phys., 111: 044902, 2012.

[20] P. Munaga, et al. "A fractal-based compact broadband polarization insensitive metamaterial absorber using lumped resistors," Microw. Opt. Technol. Lett., 58(2): 343-347, 2016.

[21] Y. Shang, Z. Shen, S. Xiao, "On the design of single-layer circuit analog absorber using double-square-loop array," IEEE Trans. Antennas Propag. 61(12): 6022-6029, 2013.

[22] Y. Han, W. Che, C. Christopoulos, Y. Xiong, Y. Chang, "A fast and efficient design method for circuit analog absorbers consisting of resistive square-loop arrays," IEEE Trans. Electromagn. Compat., 58(3): 747-757, 2016.

[23] M. A. Shukoor, S. Dey, "A novel modified circular ring-based broadband polarization-insensitive angular stable circuit analog absorber (CAA) for RCS applications," Int. J. Microwave Wireless Technolog., 15: 440-453, 2022.

[24] C. Barde, et al. "Angle-independent wideband metamaterial microwave absorber for C and X band application," Int. J. Microwave Wireless Technolog., 16: 101-109, 2023.

[25] A. A. G. Amer, et al., "A wide-angle, polarization-insensitive, wideband metamaterial absorber with lumped resistor loading for ISM band applications," IEEE Access, 12: 42629-42641, 2024.

[26] Y. Zhang, et al., "Design and Analysis of a Broadband Microwave Metamaterial Absorber," IEEE Photonics J., 15(3), 2023.

[27] K. M. R. Islam, et al., "Design and experimental performance evaluation of a single-layer polarization-insensitive asymmetric microwave metasurface absorber," IEEE Trans. Antennas Propagation., 72(8): 6520-6529, 2024.

[28] M. Zhang, et al., "Design of wideband absorber based on dual-resistor-loaded metallic strips," Int. J. Antennas Propag., 1238656: 1-8, 2020.

[29] S. Ghosh, et al., "Design, characterization and fabrication of a broadband polarization insensitive multi-layer circuit analogue absorber," IET Microwaves Antennas Propag., 10(8): 850-855, 2016.

[30] J. Chen, Y. Shang, C. Liao, "Double-layer circuit analog absorbers based on resistor-loaded square-loop arrays," IEEE Antennas Wirel. Propag. Lett., 17(4): 591-595, 2018.

[31] P. Jain, et al., "Machine learning techniques for predicting metamaterial microwave absorption performance: A comparison," IEEE Access, 11: 128774-128783, 2023.

[32] Y. Wei, et al., "A multiband, polarization-controlled metasurface absorber for electromagnetic energy harvesting and wireless power transfer," IEEE Trans. Microwave Theory Tech., 70(5): 2861-2871, 2022.

[33] R. M. H. Bilal, et al., "Polarization-controllable and angle-insensitive multiband Yagi-Uda-shaped metamaterial absorber in the microwave regime," Opt. Mater. Express, 12: 798-810, 2022.

[34] J. Wang, et al., "Polarization-controlled and flexible single-/penta-band metamaterial absorber," Materials, 11: 1619, 2018.

Biographies

Saeedeh Barzegar-Parizi received the B.Sc. degree from the Iran University of Science and Technology, Tehran, Iran, in 2008, and the M.Sc. and Ph.D. degrees from the Sharif University of Technology, Tehran, in 2010 and 2015, respectively, all in electrical engineering. She has been with the Department of Electrical Engineering, Sirjan University of Technology, where she is currently an Associate Professor. Her research interests include the numerical and analytical solving of periodic structures, photonics designs/devices, plasmonics, metamaterials, optical and biomedical sensors, Microwave devices, Propagation.

- Email: barzegarparizi@sirjantech.ac.ir
- ORCID: 0000-0002-7467-7677
- Web of Science Researcher ID: AAP-2310-2020
- Scopus Author ID: 36697630200
- Homepage: https://sirjantech.ac.ir/%d8%b3%d8%b9%db%8c%d8%af%d9%87-%d8%a8%d8%b1%d8%b2%da%af%d8%b1-%d9%be%d8%a7%d8%b1%db%8c%d8%b2%db%8c/

How to cite this paper:
S. Barzegar-Parizi, "Circuit analog absorber based on a double-layer of resistor-loaded strip arrays with various bandwidths according to selecting the polarization," J. Electr. Comput. Eng. Innovations, 13(2): 299-306, 2025.

DOI: 10.22061/jecei.2024.11244.781

URL: https://jecei.sru.ac.ir/article_2239.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Public Information Precoding for MIMO Visible Light Communication System Based on Manifold Optimization

H. Alizadeh Ghazijahani [*]***, M. Atashbar***

Department of electrical engineering, Azarbaijan Shahid Madani University, Tabriz, Iran.

Article Info

Article History:
Received 04 September 2024
Reviewed 06 November 2024
Revised 10 December 2024
Accepted 16 December 2024

Keywords:
VLC
MIMO
Precoding
Public information
Optimization

[*]Corresponding Author's Email Address:
hag@azaruniv.ac.ir

Abstract

Background and Objectives: The combination of multiple-input-multiple-output (MIMO) with a Visible light communication (VLC) system leads to a higher speed of data transmission named the MIMO-VLC system. In multi-user (MU) MIMO-VLC, an LED array transmits signals to users. These signals are categorized as signals of private information for each user and signals of public information for all users.
Methods: In this research, we design an omnidirectional precoding to transmit the signals of public information in the MU-MIMO-VLC network. We aim to maximize the achievable rate which leads to maximizing the received mean power at the possible location of the users. Besides maximizing the achievable rate, we consider equal mean transmission power constraints in all LEDs to achieve higher power efficiency of the power amplifiers used in the LED array. Based on this, we formulate an optimization problem in which the constraint is in the form of a manifold, and utilize a gradient method projected on the manifold to solve the problem.
Results: Simulation results indicate that the proposed omnidirectional precoding can achieve superior received mean power besides more than 10x bit error rate reduction compared to the classical form without precoding utilization.
Conclusion: In this research, we proposed an omnidirectional precoding for transmitting the public signals in the MU-MIMO-VLC system. The proposed optimization problem maximizes the received mean power constrained with equal transmission mean power of LEDs in the array.

Introduction

Visible light communication is one of the attractive optical communication systems that utilize the visible region of optical spectrum [1]. Due to the combination of communication with lighting, VLC is one of the emerging technologies in 6G and regarded as a promising technique to provide internet in indoor wireless access [2]. Generally speaking, VLC has many significant advantages, such as license-free spectrum, high security, high data rates, low cost, and freedom from hazardous electromagnetic radiation [3], [4].

Furthermore, multiple-input-multiple-output (MIMO) techniques are used in communication systems to ensure a high data rate and reliability where transmitters and receivers use multiple antennas [5]. A special case of MIMO systems is multi-user MIMO (MU-MIMO) where the transmitter/receiver is equipped with multiple antennas and the receiver/transmitter consists of multiple users with single or multiple antennas. In this case, a base station (BS) with an antenna array supports multiple mobile stations (MS) simultaneously [6].

Recently, to jointly benefit from the advantages of VLC and MU-MIMO, MU-MIMO-VLC systems have been considered, in which a LED array is used to transmit the downlink signals to multiple users simultaneously [7], [8]. To prevent inter-user interference, the users' associated signals are precoded before transmission [9]-[18]. This type of precoding is named directional precoding where the precode matrix is designed based on the channel matrix. The authors in [9] have formulated the precoding and power allocation problems and do energy efficiency optimizations for multi-cell rate-splitting multiple access VLC systems.

In [10], C. Wang *et al.* proposed a precoding based on successive interference cancellation (SIC) to optimize the electrical/optical power of each LED and achieve the maximum sum rate. Another research designed a precoding matrix by maximizing mutual information subject to both peak and average power constraints [12]. The performance of MU-MIMO-VLC block diagonalization precoding is discussed in [13]. The results show that inter-user interference is eliminated and the complexity of users' terminals is reduced. Another research on precoding for MU-VLC is reported in [14] in which the confidentiality of users' messages has been considered.

In a MU communication system, each user has its private information. The above-mentioned and similar literature [10]-[14] employ precoding techniques to send private information to mitigate the other user's signal by maximizing the received signal strength at the intended user. Applying precoding needs to know the channel state information (CSI) of each user. Furthermore, the transmitter unit broadcasts specific public or common information concurrently to all users connected to the network. This information includes data for synchronization, medium access control frames, link recovery request, or when an IP address is dynamically assigned to a device [19]. To transmit such information, it is assumed that the user location is not known, so the user location feedback is not needed.

While the rate-splitting multiple access (RSMA) technique enables the concurrent transmission of specific common information and private data in a VLC network [20], [21].

However, in the context of RSMA, the term "common information" is defined as a message that, while meant for a particular user, must be decoded by all users. This understanding contrasts sharply with the concept of a public message, which necessitates that all users have access to the information it conveys [22]. In addition, similar to directional precoding, in RSMA, CSI is needed in the precoding of both common and private information to prevent inter-user interference.

On the other hand, the public message needs to be broadcast to all possible user locations, so inter-user interference is not a challenge here. In this case, we need an efficient precoding for the transmission signal to balance the received signal over the whole coverage area. This type of precoding is named omnidirectional precoding [23], [24]. The main differences between omnidirectional and directional precoding can be summarized as: 1) The transmit information in directional precoding is private information of that user while in omnidirectional precoding, the transmit information is the public information that all users wish to receive, 2) Directional precoding relies on the precise knowledge of the channel vector that exists between the user and the transmitter array. In contrast, omnidirectional approaches require only the channel model without the need for specific vector values, 3) in the directional precoding, the transmitted energy of LEDs is concentrated in the target point where the user is located there. In comparison, the idea in omnidirectional precoding (assuming an unknown user location) is to maximize the received energy in all potential user locations. To the best of our knowledge, there is not any study that addresses the design of omnidirectional precoding proportional to MU-MIMO-VLC systems.

In this paper, we present an omnidirectional precoding algorithm for a transmitter LED array in a MU indoor MIMO-VLC system for the transmission of public information in the network.

A comparison of the fundamental differences between introduced omnidirectional precoding for VLC and traditional directional precoding are outlined in Table 1.

To this end, while the user location is not known, we propose to maximize the achievable rate in user potential locations which leads to maximizing the received mean power at the whole area. On the other hand, to achieve higher power efficiency of the power amplifiers used in the LED array, it is needed that the mean transmission powers of the signals associated with all LEDs be the same. Consequently, we consider this constraint besides maximizing the achievable rate in our problem. This leads to a constrained optimization problem in which the constraint is in the form of manifold. Accordingly, to solve this problem, we propose a gradient method projected on the manifold.

The rest of this paper is organized as follows. The system model is described in the next section. After

that we propose our optimization problem. The simulation setup, results, and discussion are presented and finally, the paper is concluded.

Table 1: Fundamental differences between directional and proposed omnidirectional precoding for VLC

Symbol	Directional Precoding	Omnidirectional Precoding
Scope	Transmitting specific signal to each user	Transmitting signal to all users simultaneously
Data	private	public
Design requirements	CSI values for each transmitter-user pair	Channel model
Scattered Power distribution	Concentrated to the user location	Distributed over possible locations of users
Objective function	Maximizing the signal to interference plus noise ratio	Maximizing the received mean power of possible locations of users

System Model

Consider a VLC-MIMO system with a uniform rectangular LED array to broadcast public information to the single photo-diode (PD) equipped users in the coverage area shown in Fig. 1 assuming that the array of LEDs consists of M_t LEDs in a rectangular structure as $M_t = M_x \times M_y$ on the x-y plane and d_x and d_y are the distances of adjacent LEDs along the *x*-axis and *y*-axis, respectively. Note that in our model, the LED panel is placed on the ceiling with height *D* from the room's floor. Furthermore, assume that the public information signal vector $\boldsymbol{s} = [s_1, s_2, \dots s_q]^T$ sent to users, where s_i, q, and $(.)^T$ indicate the *i*-th transmitted symbol, number of symbols, and transpose, respectively. This signal vector is multiplied by the designed precoding matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \dots \boldsymbol{p}_{M_t}]^T \in \mathbb{R}^{M_t \times q}$ to generate LEDs' transmission signal vector $\boldsymbol{v} = \boldsymbol{P}\boldsymbol{s} + e$, where $\boldsymbol{v} = [v_1, v_2, \dots v_{M_t}]^T$ with v_i presents the transmit signal of *i*-th LED located in (x_i, y_i, D)

$$x_i = \left\lfloor \frac{i-1}{M_y} \right\rfloor d_x \,, y_i = (i \bmod M_y - 1) d_y \,, \quad i = 1, 2, \dots M_t \,, \tag{1}$$

in which $\lfloor . \rfloor$ denotes the floor operator sign and 'mod' indicates the reminder of deviation and *e* is the DC offset to confirm non-negative elements of $\boldsymbol{v}$.

In VLC systems, each user is equipped with a PD to receive the transmitted optical signal strength from LEDs array. Thus, the received signal at the *j*-th user in the coordinates (x_j^u, y_j^u, h) will be as follows [12]:

$$r_j = \boldsymbol{h}_j^T (\boldsymbol{P}\boldsymbol{s} + e) + n \,, \tag{2}$$

where $\boldsymbol{P}$ is the $M_t \times q$ precoding matrix, *n* is white Gaussian noise and $\boldsymbol{h}_j$ is the $M_t \times 1$ channel vector. $\boldsymbol{h}_j = [h_{j1}, h_{j2}, \dots h_{jM_t}]^T$ with h_{ji} is the VLC channel gain between the *i*-th LED and the *j*-th user.

VLC channel models are currently investigated under two categories: deterministic and stochastic models. Deterministic models aim to predict channel characteristics at a specific location of the transmitter and receiver, as well as the surrounding environment, with ray tracing, recursive, and empirical algorithms. In stochastic approaches, the impulse responses of VLC channels are defined by the law of light propagation applied to a specific geometry of transmitter, receiver, and scattered [25]. In this research, we also use deterministic channel model as commonly used in other literature.

As the line of sight (LOS) channel of the optical wireless channel contains most parts of the transmitted energy [10], we ignore the non-LOS part in this work. We use the well-known Lambertian model to estimate the path loss in VLC channel. Consequently, according to the geometry presented in Fig. 1, for the LOS path, h_{ji} is given as [26]

$$h_{ji} = \begin{cases} \frac{A_d (m_l+1)}{2\pi d_{ji}^2} \cos^{m_l}(\phi_{ji}) \cos(\theta_{ji}) TG & if \;\; 0 \le \theta_{ji} \le \psi_R \\ 0 & if \;\; \theta_{ji} < \psi_R \end{cases}, \tag{3}$$

where ϕ_{ji} denotes the emitting angle, θ_{ji} denotes the incident angle from the *i*-th LED to *j*-th user, A_d denotes the area of the receiver PD, m_l is Lambert's mode number expressing the directivity of the source beam, the T denotes the signal transmission coefficient of an optical filter, ψ_R denotes the field of view of the receiver PD, G denotes the concentrator gain, and d_{ji} is the distance between *i*-th LED to *j*-th user PD [26]. Given the significant attenuation observed in this model relative to distance, the impact of multipath reflections on the received signal is negligible and can therefore be disregarded.

Proposed Optimization

In this paper, the idea is to design an omnidirectional precoding matrix for efficient transmission of public information signals of all users distributed on the coverage area in the VLC network. To address this challenge appropriately, two limitations are considered. 1) Maximum achievable rate, 2) constant mean transmission power of array LEDs.

A. Maximum Achievable Rate

According to the channel model described in (2), the mutual information of MIMO-VLC downlink transmission after DC suppression will be as follows [27], [28]

$$I_j = \log\left(1 + \frac{1}{\delta^2}\boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j\right), \quad (4)$$

in which δ^2 indicate the variance of noise at the receiver and $(.)^H$ indicates the Hermitian. Besides, it is supposed that the mean power of the elements of the vector $\boldsymbol{s}$ is equal to one. According to (4), maximizing term $\boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j$ leads to maximum achievable rate where it is in line with maximizing received mean power at the *j*-th user. Since the users can be located at any point of the covered area, so the precoding matrix should be designed in a way that term $\boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j$ achieves its maximum value for all possible h_j values.

To this aim, first, we do sampling from the coverage area uniformly, so that the work plane seems as a grid surface with step d_g. Given this premise, $\boldsymbol{h}_j$ is defined as the channel vector between the LED array and the *j*-th point on the grid surface, and N_s denotes the total number of grid points. Then we maximize the average of received mean power (ARMP) at the sampled location points as

$$ARMP = \frac{1}{N_s M_t}\sum_{j=1}^{N_s} \boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j . \quad (5)$$

In this equation, the parameter M_t is utilized to normalize the total transmit power at the LED array.

B. Keeping Constant the Mean Transmission Power of Leds in the Array

In a similar way to the radio frequency MIMO transmission system [28], to achieve a higher power efficiency of the amplifiers used to drive the LED array, it is needed that the mean transmission powers of the signals associated with all LEDs be the same. Supposing the elements of $\boldsymbol{s}$ are independent with zero-mean and unit variance, the mean transmission power of the *i*-th LED can be stated as

$$E(\boldsymbol{v_i^2}) = E\left(\left(\boldsymbol{p_i^T s} + e\right)^2\right)$$
$$= \boldsymbol{p_i^T E(ss^T) p}_i + e^2 = \boldsymbol{p_i^T\, p}_i + e^2 , \quad (6)$$

in which $E(.)$ stands for expectation operator. In (6), the first term is the mean of AC power and the second one is DC power. To keep constant the mean transmit power of all LEDs, $\boldsymbol{p_i^T\, p}_i$ must be constant for $i = 1, 2, \ldots M_t$. Without loss of generality, we get $\boldsymbol{p_i^T\, p}_i = 1$ as a constraint of problem where can be expressed in the matrix form as follows

$$diag(\boldsymbol{P}\boldsymbol{P}^{\boldsymbol{H}}) = \boldsymbol{I}_{M_t} , \quad (7)$$

in which, $\boldsymbol{I}_{M_t}$ is the $M_t \times M_t$ identity matrix and $diag(.)$ represents a diagonal matrix whose major diagonal elements are equal to the major diagonal elements of the matrix. Accordingly, the proposed constrained optimization problem to design the precoding matrix $\boldsymbol{P}$ is

$$\max_{\boldsymbol{P}} \quad \sum_{j=1}^{N_s} \boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j$$
$$s.t. \quad diag(\boldsymbol{P}\boldsymbol{P}^{\boldsymbol{H}}) = \boldsymbol{I}_{M_t} . \quad (8)$$

As the term $\sum_{j=1}^{N_s} \boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j$ is a concave function, by choosing $(\boldsymbol{P}) \triangleq -\sum_{j=1}^{N_s} \boldsymbol{h}_j^H \mathbf{P}\mathbf{P}^H \boldsymbol{h}_j$, it transforms to a convex function. Accordingly, the optimization problem forms as

$$\min_{\boldsymbol{P}} \quad f(\boldsymbol{P})$$
$$s.t. \quad diag(\boldsymbol{P}\boldsymbol{P}^{\boldsymbol{H}}) = \boldsymbol{I}_{M_t} . \quad (9)$$

Recently, geometric solutions are used to solve various optimization problems. One kind of such solutions is manifold-based geometry which is used in constrained optimization problems [30], [31] because of its relative simplicity and optimality. The constraints in constrained optimization problems can be interpreted as isolated points in the space that are in the manifold forms such as Stiefel, Grassmann, Riemannian, etc. Consequently, the optimum points are searched in the space that is inside the manifold.

In this work, we propose a manifold-based method to solve (9). As the constraint in (9) is in the form of Grassmann manifold [30] and $f(\boldsymbol{P})$ is in the form of a quadratic function (a convex function), we use the gradient method projected on the manifold, in which, matrix $\boldsymbol{P}$ is calculated iteratively as

$$\boldsymbol{P}_{k+1} = \boldsymbol{P}_k - \mu \nabla f(\boldsymbol{P}_k) , \quad (10)$$

in which $\boldsymbol{P}_k$ is the $\boldsymbol{P}$ values in *k*-th iteration, μ is step size, and $\nabla f(\boldsymbol{P})$ is the $M_t \times q$ gradient matrix of $f(\boldsymbol{P})$. According to matrix relations on [32], we have

$$\nabla f(\boldsymbol{P}) = -2\sum_{j=1}^{N_s} \boldsymbol{h}_j \boldsymbol{h}_j^H \boldsymbol{P} . \quad (11)$$

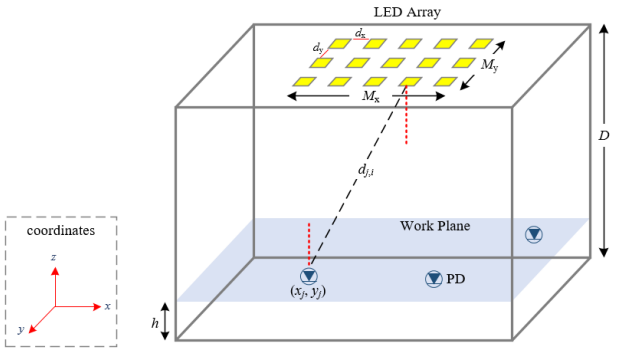


Fig. 1: Proposed system model.

In each iteration of the gradient algorithm, to ensure the constraint is established, the resulting matrix $\boldsymbol{P_{k+1}}$ is projected on the manifold. Since the constraint in (9) is in the form of the Grassmann manifold, the projection on the above manifold is as [30]

$$\boldsymbol{P}_{k+1} \leftarrow \left(diag(\boldsymbol{P}_{k+1}\boldsymbol{P}_{k+1}^{H})\right)^{-\frac{1}{2}} \boldsymbol{P}_{k+1} . \quad (12)$$

The iteration is continued until the difference of $f(\boldsymbol{P}_k)$ goes below a determined small value ε for two successive iterations to satisfy the convergence condition, as $|f(\boldsymbol{P}_{k+1}) - f(\boldsymbol{P}_k)| < \varepsilon$. The steps of the proposed algorithm to solve the optimization problem (9) are determined by the projected gradient method on the manifolds presented in Algorithm 1.

Algorithm 1: Solving the optimization problem presented in (9)

1- Initialization of matrix $\boldsymbol{P}$
2- Calculate $\nabla f(\boldsymbol{P})$ using (11)
3- Update $\boldsymbol{P}$ as (10)
4- Project $\boldsymbol{P}$ on the manifold based on (12)
5- Repeat steps 2-4 to achieve the convergence condition

Results and Discussion

In this section, we present the simulation setup and results to show how the proposed precoding algorithm for MIMO-VLC satisfies two limitations stated in section 2.

To this end, we consider optimized ARMP as an evaluation criterion in which optimized ARMP is defined as the value of ARMP based on the designed precoding matrix $\boldsymbol{P}$. To the best of our knowledge, there is not any similar study to design an omnidirectional precoding matrix for MIMO-VLC system, we choose the mean of ARMP parameters over all random precoding matrix $\boldsymbol{P}$ as a reference method to compare the results. We name this as 'classical method'. We utilize RSMA criterion to evaluate the performance of the proposed omnidirectional precoding. Since, RSMA is proportional to achievable rate parameter, we discuss the performance of the system with this criterion.

In the simulation, we consider a scenario in which a LED panel with uniform rectangular array is supposed to be installed in the center of the room ceiling.

Also, the users can be placed at all possible locations on the floor of the room and each of them receives the VLC signal emitted from all LEDs.

As mentioned in section 3, to aim for maximum achievable rate, we need to do sampling from all possible locations of users, therefore, in our simulations, the floor of the room is sampled uniformly with a distance of 0.1 in both axes. The other simulation parameters are summarized in Table 2.

Table 2: Simulation parameters

Symbol	Description	Value
-	room dimensions [width, length, height]	[5, 6, 3] m
$\boldsymbol{\psi_R}$	receiver field of view	75°
$\boldsymbol{A_d}$	receiver area	1×10^{-4} m^2
$\boldsymbol{m_l}$	Lambert's mode number	1
$\boldsymbol{T_s(\theta_T)}$	signal transmission coefficient of an optical filter	1
$\boldsymbol{g(\theta_T)}$	concentrator gain	5
$\boldsymbol{\varepsilon}$	stop parameter in iterative algorithm	10^{-4}
$\boldsymbol{q}$	number of symbols	10
$\boldsymbol{d_g}$	Work plane grid step	1 mm

A. Convergence of Optimization Problem

In the first simulation, the convergence of the gradient method projected on the manifold in solving the proposed optimization problem is investigated. In this way, a 3×3 LED rectangular array with a distance of 0.02 m between adjacent elements is considered at the ceiling and the PD of users is located on a flat surface, named work plane, with height 1m from the floor. The $f(\boldsymbol{P})$ value is calculated in each iteration. The result shown in Fig. 2 indicates that the cost function is converged in 5-th iteration. The resultant precoding matrix in 5-th iteration is as (13) showing that the constraint of proposed optimization problem

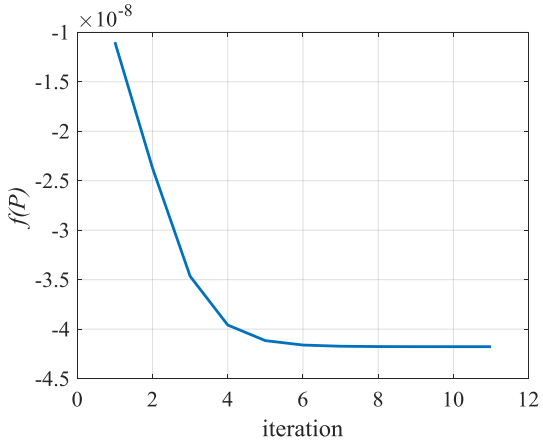


Fig. 2: Convergence of proposed algorithm over iterations.

$$\boldsymbol{P}=\begin{bmatrix} 0.485532 & 0.389968 & 0.148727 & -0.28536 & 0.032314 & 0.279723 & -0.33371 & 0.459929 & -0.06347 & 0.320037 \\ 0.063389 & -0.5254 & -0.0869 & 0.357057 & -0.25406 & -0.22072 & 0.185637 & 0.187258 & -0.33472 & -0.53857 \\ -0.29119 & -0.13221 & -0.16466 & -0.47472 & 0.395979 & 0.432586 & 0.102007 & 0.273667 & -0.21922 & -0.40984 \\ 0.031884 & -0.35761 & -0.07127 & -0.35639 & 0.424023 & -0.39309 & 0.287618 & -0.14704 & -0.4282 & 0.342038 \\ 0.243148 & -0.38668 & -0.15495 & -0.31892 & -0.39669 & -0.3259 & -0.21353 & -0.09049 & -0.49624 & 0.319427 \\ -0.37737 & -0.4311 & 0.093535 & -0.22487 & 0.246277 & 0.378553 & 0.145202 & 0.385867 & 0.450226 & -0.1892 \\ -0.45466 & -0.23233 & 0.337457 & 0.352245 & -0.43656 & 0.389548 & 0.377869 & 0.004552 & 0.058183 & 0.113256 \\ -0.46861 & 0.191308 & 0.274291 & -0.32424 & -0.51558 & -0.02095 & -0.12132 & -0.41916 & -0.05102 & 0.322735 \\ 0.087989 & 0.13417 & -0.16224 & -0.12184 & 0.365285 & 0.004531 & 0.209953 & 0.693242 & -0.44237 & 0.281573 \end{bmatrix}. \quad (13)$$

presented in (8) is satisfied. As the $f(\boldsymbol{P})$ is a well-known quadratic convex function, the convergence of Gradient descent algorithm is guaranteed. Consequently, the convergence curve of proposed algorithm presented in Fig. 2 is gained through simulation.

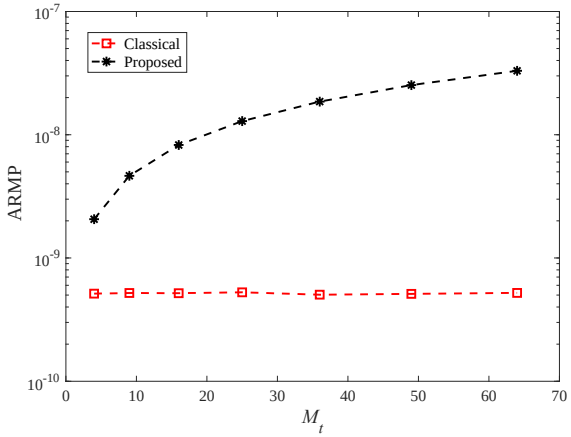


Fig. 3: The ARMP of classical and proposed methods versus number of LEDs in the array with d_x = 0.05 m.

B. Number of LED Array Elements

In the second simulation, the behavior of ARMP versus different LED numbers in the array is investigated. In this simulation the parameters are set as d_x = 0.01, h = 2.5 m, and M_t the number of LEDs varies from 4 to 64. Fig. 3 shows that the ARMP of the proposed method is improved by increasing the number of LEDs, while the classical ARMP is constant over M_t changes, as expected. This is due to the fact that as the elements of $\boldsymbol{P}$ matrix are chosen randomly and according to the law of large numbers, the ARMP is proportional to the variance of $\boldsymbol{P}$ elements. Based on the considered parameters in simulation, the ARMP value for the classical method is almost fixed at 5.2×10^{-10} for all M_t values.

This is while, for the proposed method the ARMP is 4.6×10^{-9} and 3.3×10^{-8} for M_t = 9 and 64, respectively. The received signals from LEDs at each user location sum up linearly in classical method which is not necessarily constructive while, the designed precoding leads to a constructive summation of LEDs' signals in the proposed method. By increase in M_t, the degree of freedom in the constructive summation is increased and helps to increase optimized ARMP.

C. Distance between LED Array Elements

To investigate the impact of distance between LEDs in the array, we repeated the simulation for $d_x = d_y \in \{0.01, 0.02, \ldots, 0.1\}$.

The ARMP versus different distance values is depicted in Fig. 4. In this figure, the simulation results for the proposed method are presented with three different numbers of LEDs in the array. Besides, the ARMP curve for the classical method for any arbitrary M_t is depicted versus d_x. As seen, the ARMP values remain unchanged by increasing in d_x in both methods. The constant value of ARMP is due to the low dynamic range of d_x.

The simulation is repeated for both varying d_x and d_y for proposed method under $M_t \in \{9, 25, 64\}$ and classical method. The results are shown in Fig. 5 where a near flat surface is appears for each structure with M_t LEDs.

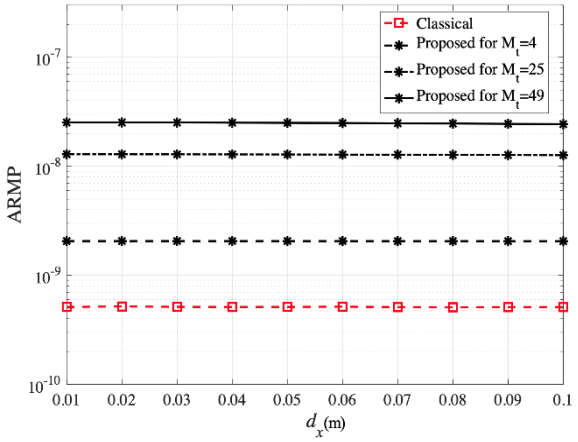


Fig. 4: The ARMP of proposed method under $M_t \in \{9, 25, 64\}$ and classical method versus low dynamic range adjacent distance between LEDs in the array.

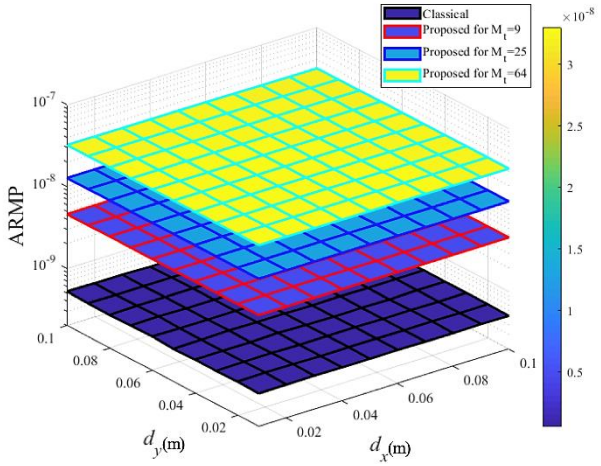


Fig. 5: The ARMP of proposed method under $M_t \in \{9, 25, 64\}$ and classical method versus low x and y axis dynamic range adjacent distances between LEDs in the array in.

To more investigation, the simulation is repeated for a d_x with a high dynamic range where the results are shown in Fig. 6. As seen, for large M_t, the ARMP curve of the proposed method falls by an increase in d_x while it is almost constant for small M_t's. This is due to the fact that when the M_t and d_x are concurrently large, the physical length of LED array (panel) goes expand over the ceiling and this makes the constructive combination of LED signals hard in most points of the work plane.

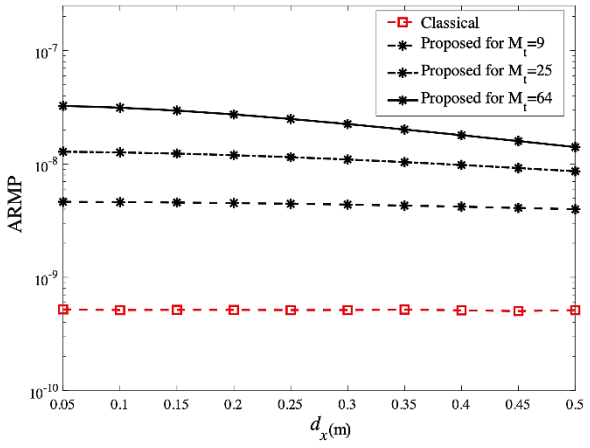


Fig. 6: The ARMP of proposed method $M_t \in \{9, 25, 64\}$ and classical method versus high dynamic range adjacent distance between LEDs in the array.

Similar to low dynamic range by varying d_x and d_y for proposed method under $M_t \in \{9, 25, 64\}$ and classical method for high dynamic range, the ARMP 3D curves are plotted in Fig. 7. As anticipated, the symmetric ARMP curves arise due to the symmetry property of the LED array in both the x and y axes.

To show the impact of LED numbers, the ARMP versus M_t is depicted in Fig. 8 for some high dynamic ranges between elements in the array. As seen, although by increasing M_t, the LED array physical length is running larger, the optimized ARMP has incremental functionality with M_t. It has resulted that in large LED array panels, considering both M_t and dx jointly, M_t has a dominant impact on the ARMP of proposed method.

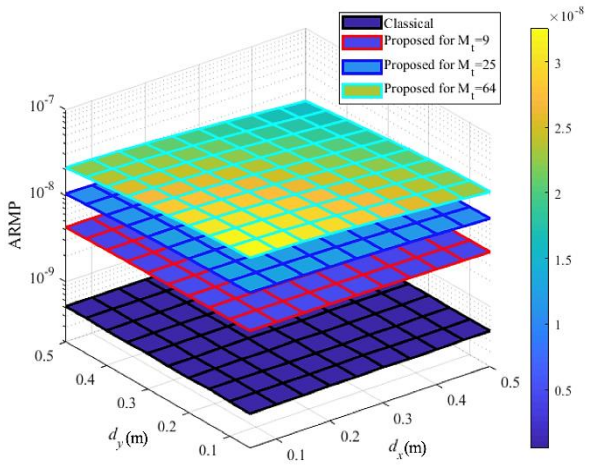


Fig. 7: The ARMP of proposed method under $M_t \in \{9, 25, 64\}$ and classical method versus high x and y axis dynamic range adjacent distances between LEDs in the array in.

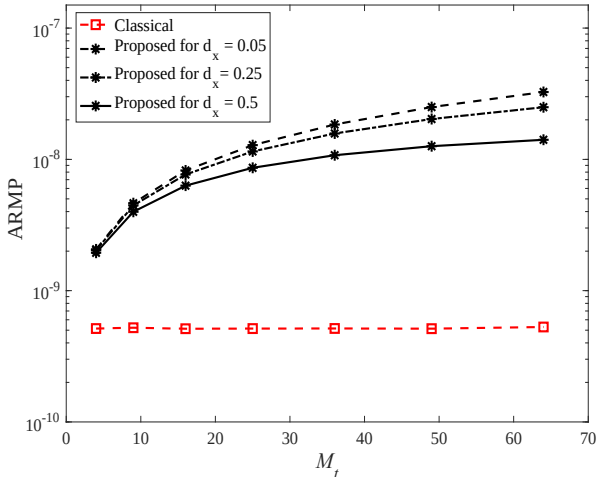


Fig. 8: The ARMP of classical and proposed algorithm versus number of LEDs in the array under some high dynamic range between elements in the array.

Finally, the performance of proposed and classical methods under different work plane heights is studied. To this end, we vary the work plane from the floor up to the height of 1m. We set d_x = 0.05m and $M_t \in \{9, 25, 64\}$ in our simulations. The result is shown in Fig. 9 in which the horizontal axis is the work plane height from the floor. As expected, by moving the work plane from the floor the ARMP for both proposed and classical methods increases. This is due to the fact that by increasing work plane height, the

distance between user locations and the LED array decreases which leads to a decrease in channel loss.

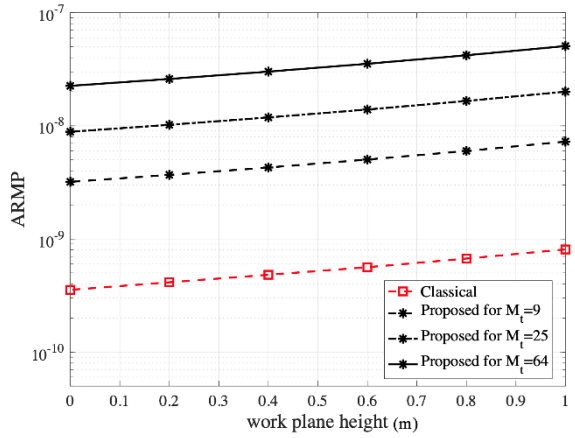


Fig. 9: The ARMP the ARMP of classical and proposed algorithm versus work plane height under different number of LEDs in the array.

D. Circular LED Array

This section aims to assess the performance of the proposed method in relation to various LED array configurations, specifically focusing on the ARMP of methods applied to circular arrays. To this end, the algorithm is tested for various values of radius $r \in \{0.5, 1, 1.5\}$. The results are depicted in Fig. 10 where plots the ARMP vs number of LEDs of the array. As seen, similar to rectangular array, the ARMP curves are rising by increase of M_t.

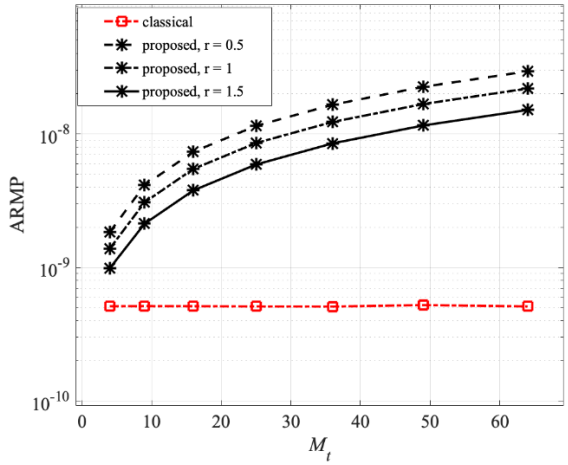


Fig. 10: The ARMP of classical and proposed algorithm versus number of LEDs in the array under some radiuses of the circle array.

E. Bit error rate analysis

For more investigation of the system performance, the bit error rate (BER) of the proposed method is compared with that of the classical method. Detection of public information at j-th user needs the value of the term $\boldsymbol{h}_j^T \boldsymbol{P}$ to be known. It is assumed that a pilot block of q symbols is broadcast at the first so the j-th user can estimate the relevant value of $\boldsymbol{h}_j^T \boldsymbol{P}$. Then, the main public information bits are modulated with usual on-off keying non-return to zero (NRZ) scheme then broadcast to all users. At the receiver side, each user detects the public information bits using the Maximum-Likelihood criterion based on the estimation of term $\boldsymbol{h}_j^T \boldsymbol{P}$ as

$$\hat{\boldsymbol{s}}_{\boldsymbol{ML}} = \underset{\boldsymbol{s}\epsilon\mathcal{S}}{\operatorname{argmin}} \left| r_j - \boldsymbol{h}_j^T \boldsymbol{P}\boldsymbol{s} \right|^2, \tag{14}$$

where $\mathcal{S}$ indicates the set of all possible values of vector $\boldsymbol{s}$. The mean BER of 15 users distributed uniformly in the work plane versus M_t with noise variance $\sigma^2 = 2 \times 10^{-7}$ is plotted in Fig. 11. As seen, the BER values of proposed method is fall with increase on M_t due to increase of ARMP in similar scenario.

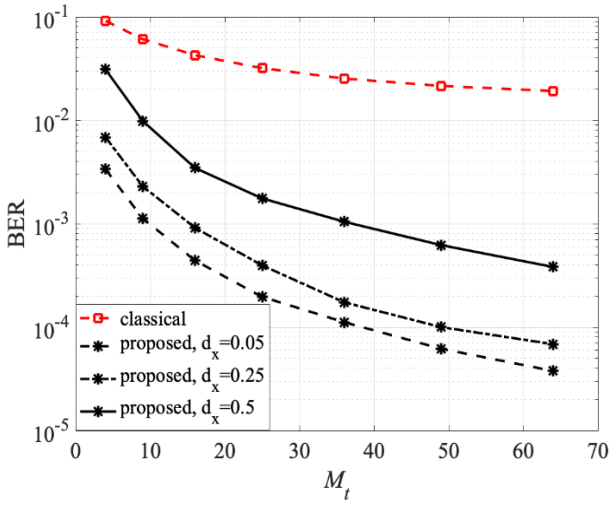


Fig. 11: The BER values versus M_t for classical and proposed method with different $d_x \in \{0.05, 0.25, 0.5\}$.

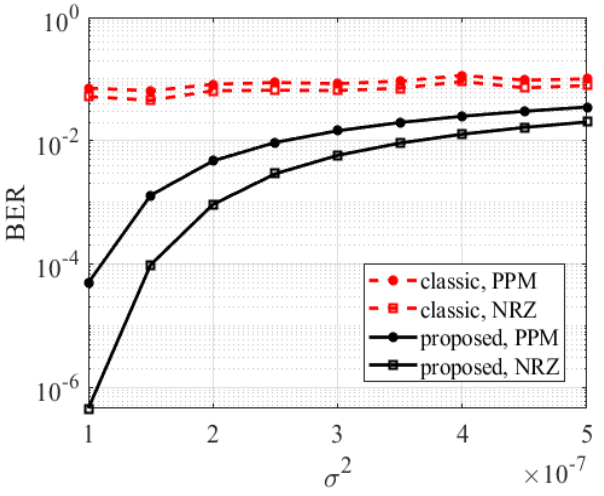


Fig. 12: The BER of proposed and classical methods versus noise variance under PPM and NRZ modulations.

Fig. 12 compares the performance of system under NRZ and pulse position modulation (PPM) versus noise variance for both proposed and classical methods under M_t = 9, d = 0.02 m. As seen, the BER values of the proposed methods encouragingly outperforms the classical method. Comparing the results, NRZ modulation outperforms than the PPM one.

Conclusion

In this research, we proposed an omnidirectional precoding for transmitting the public signals in MU-MIMO-VLC system. For this purpose, we proposed an optimization problem which maximizes the received mean power constrained with equal transmission mean power of LEDs in the array. In our formulation the constraint is in the form of manifold therefore a gradient method projected on the manifold is designed to solve the problem. We considered the ARMP parameter to investigate the performance of the system under varying some simulation parameters such as LED numbers in the array, distance between LEDs, and height of work plane from the floor. Simulation results has shown that the proposed omnidirectional precoding leads to higher ARMP values with respect to the classical method in all simulation scenarios.

Author Contributions

The present article is the outcome of a joint endeavor by H. Alizadeh and M. Atashbar. H. Alizadeh took the lead in drafting the manuscript, whereas M. Atashbar handled the execution of the simulations. Furthermore, both contributed to the analysis of the simulation outcomes.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

VLC	Visible Light Communication
MU-MIMO	Multi-User Multiple-Input-Multiple-Output
CSI	Channel State Information
ARMP	Average of Received Mean Power
LoS	Line of Sight

References

[1] S. Aboagye, A. R. Ndjiongue, T. M. Ngatched, O. A. Dobre, H. V. Poor, "RIS-assisted visible light communication systems, A tutorial," IEEE Commun. Surv. Tutorials, 25(1): 251-288, 2022.

[2] W. Jiang, F. L. Luo, 6G Key Technologies: A Comprehensive Guide, John Wiley & Sons, 2023.

[3] H. S. R Hujijo, M. Ilyas, "Enhancing spectral efficiency with low complexity filtered-orthogonal frequency division multiplexing in visible light communication system," ETRI J., 46(6): 1007-1019, 2024.

[4] L. E. M. Matheus, A. B. Vieira, L. F. Vieira, M. A. Vieira, O. Gnawali, "Visible light communication: concepts, applications and challenges," IEEE Commun. Surv. Tutorials, 21(4): 3204-3237, 2019.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, T. L. Marzetta, "Massive MIMO for next generation wireless systems," IEEE Commun. Mag., 52(2):186-195, 2014.

[6] E. Castaneda, A. Silva, A. Gameiro, M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," IEEE Commun. Surv. Tutorials, 19(1):239-284, 2016.

[7] J. Lian, M. Brandt-Pearce, "Multiuser visible light communication systems using OFDMA," J, Lightwave Technol., 38(21):6015-6023, 2020.

[8] C. Chen, W. D. Zhong, D. Wu, "On the coverage of multiple-input multiple-output visible light communications," J. Opt. Commun. Networking, 9(9): D31-D41, 2017.

[9] F. Xing, S. He, V. C. Leung, H. Yin, "Energy efficiency optimization for rate-splitting multiple access-based indoor visible light communication networks," IEEE J. Sel. Areas Commun., 40(5): 1706-1720, 2022.

[10] C. Wang, Y. Yang, Z. Yang, C. Feng, J. Cheng, C. Guo, "Joint SIC-based Precoding and Sub-connected architecture design for MIMO VLC systems," IEEE Trans. Commun., 71(2): 1044-1058, 2023.

[11] F. R Castillo-Soria, C Gutierrez, F. M Maciel-Barboza, V. I Rodriguez Abdala, J Datta, "Relay-assisted multiuser MIMO-DQSM system for correlated fading channels," ETRI J., 46(2): 184-193, 2024.

[12] F. Yang, Y. Dong, "Joint probabilistic shaping and beamforming scheme for MISO VLC systems," IEEE Wireless Commun. Lett., 11(3): 508-512, 2021.

[13] Y. Hong, J. Chen, Z. Wang, C. Yu, "Performance of a precoding MIMO system for decentralized multiuser indoor visible light communications," IEEE Photonics J., 5(4): 7800211-7800211, 2013.

[14] S. T. Duong, T. V. Pham, C. T. Nguyen, A. T. Pham, "Energy-efficient precoding designs for multi-user visible light communication systems with confidential messages," IEEE Trans. Green Commun. Networking, 5(4): 1974-1987, 2021.

[15] Q. Zhao, Y. Fan, B. Kang, "A joint precoding scheme for indoor downlink multi-user MIMO VLC systems," Opt. Commun., 403: 341-346, 2017.

[16] H. Marshoud, P. C. Sofotasios, S. Muhaidat, B. S. Sharif, G. K. Karagiannidis, "Optical adaptive precoding for visible light communications," IEEE Access, 6: 22121-22130, 2018.

[17] H. Ma, L. Lampe, S. Hranilovic, "Robust MMSE linear precoding for visible light communication broadcasting systems," in Proc. 2013 IEEE Globecom Workshops (GC Wkshps), 2013.

[18] T. V. Pham, H. Le-Minh, A. T. Pham, "Multi-user visible light communication broadcast channels with zero-forcing precoding," IEEE Trans. Commun., 65(6): 2509-2521, 2017.

[19] N. T. Le, Y.M. Jang, "Broadcasting MAC protocol for IEEE 802.15. 7 visible light communication," in Proc. 2013 Fifth International Conference on Ubiquitous and Future Networks (ICUFN), 2013.

[20] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," IEEE Commun. Surv. Tutorials, 24(4): 2073-2126, 2022.

[21] S. A. Naser, P. C. Sofotasios, S. Muhaidat, M. Al-Qutayri, "Rate-splitting multiple access for indoor visible light communication networks," in Proc. 2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2021.

[22] M. Dai, B. Clerckx, D. Gesbert, G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," IEEE Trans. Wireless Commun., 15(7): 4611-4624, 2016.

[23] L. Mavarayi, M. Atashbar, "Omnidirectional precoding for massive MIMO with uniform rectangular array in presence of mutual coupling," Digital Signal Process., 130: 103717, 2022.

[24] X. Men, T. Liu, Y. Li, M. Liu, "Constructions of 2-D golay complementary array sets with flexible array sizes for omnidirectional precoding in massive MIMO," IEEE Commun. Lett., 27(5): 1302-1306, 2023.

[25] S. Yahia, Y. Meraihi, A. Ramdane-Cherif, A. B. Gabis, D. Acheli, H. Guan, "A survey of channel modeling techniques for visible light communications," J. Network Comput. Appl., 194: 103206, 2021.

[26] Z. Ghassemlooy, W. Popoola, S. Rajbhandari, Optical wireless communications: system and channel modelling with Matlab®, CRC press, 2019.

[27] Z. Wang, Q. Wang, W. Huang, Z. Xu, Visible light communications: modulation and signal processing, John Wiley & Sons, 2017.

[28] E. Telatar, "Capacity of multi-antenna Gaussian channels," Eur. Trans. Telecommun., 10(6): 585-595, 1999.

[29] X. Meng, X. Gao, X. G. Xia, "Omnidirectional precoding based transmission in massive MIMO systems," IEEE Trans. Commun., 64(1): 174-186, 2015.

[30] P. A. Absil, R. Mahony, R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, 2008.

[31] J. H. Manton, "Geometry, manifolds, and nonconvex optimization: How geometry can help optimization," IEEE Signal Process. Mag., 37(5): 109-119, 2020.

[32] G. A. Seber, A matrix handbook for statisticians, John Wiley & Sons, 2008.

Biographies

Hamed Alizadeh Ghazijahani was born in Ghazijahan, Iran in 1988. He completed his B.Sc., M.Sc., and Ph.D. degrees in electrical engineering, with a focus on telecommunications, at the University of Tabriz, Iran, in the years 2010, 2013, and 2019, respectively. In 2020, he took on the role of assistant professor at Azarbaijan Shahid Madani University. His research primarily explores optical and wireless communication systems and networks.

- E-mail: hag@azaruniv.ac.ir
- ORCID: 0000-0002-2438-7700
- Web of Science Researcher ID: AGS-4002-2022
- Scopus Author ID: 36863455200
- Homepage: http://pajouhesh.azaruniv.ac.ir/_Pages/ResearcherEn.aspx?ID=9914

Mahmoud Atashbar was born in Marand, in the East Azarbijan province of Iran in 1979. He received his BS degree in electrical engineering from the School of Electrical Engineering, Sahand University of Technology, Tabriz, Iran, in 2003, and his MS and Ph.D. degree in telecommunication engineering from the school of electrical engineering, Iran University of Science and Technology, Tehran, Iran, in 2006 and 2013, respectively. Since 2013, he has been with the Department of Electrical Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran, where he is now an assistant professor. His research interests include wireless communication and communication systems.

- Email: atashbar@azaruniv.ac.ir
- ORCID: 0000-0003-3721-4128
- Web of Science Researcher ID: GRE-9721-2022
- Scopus Author ID: 56035037500
- Homepage: http://pajouhesh.azaruniv.ac.ir/_Pages/Researcher.aspx?ID=1040

How to cite this paper:
H. Alizadeh Ghazijahani, M. Atashbar, "A public information precoding for MIMO visible light communication system based on manifold optimization," J. Electr. Comput. Eng. Innovations, 13(2): 307-316, 2025.

DOI: 10.22061/jecei.2024.11251.782

URL: https://jecei.sru.ac.ir/article_2240.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Effective Heart Disease Prediction Model Using Deep Learning-based Dimensionality Reduction on Imbalanced Data

S. Kabirirad [1], V. Afshin [2], S. H. Zahiri [2,*]

[1]*Department of Computer Science, Faculty of Computer and Industrial Engineering, Birjand University of Technology, Birjand, Iran.*

[2]*Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.*

Article Info

Article History:
Received 25 June 2024
Reviewed 08 August 2024
Revised 03 October 2024
Accepted 15 October 2024

Keywords:
Dimensionality reduction
Imbalanced data
Heart disease prediction
Autoencoder
PCA
Information bottleneck
Feature extraction

*Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: When dealing with high-volume and high-dimensional datasets, the distribution of samples becomes sparse, and issues such as feature redundancy or irrelevance arise. Dimensionality reduction techniques aim to incorporate correlation between features and map the original features into a lower dimensional space. This usually reduces the computational burden and increases performance. In this paper, we study the problem of predicting heart disease in a situation where the dataset is large and (or) the proportion of instances belonging to one class compared to others is significantly low.

Methods: We investigated the prominent dimensionality reduction techniques, including Principal Component Analysis (PCA), Information Bottleneck (IB), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) and Variational Autoencoder (VAE) on popular classification algorithms. To have adequate samples in all classes to properly feed the classifier, an efficient data balancing technique is used to compensate for fewer positives than negatives. Among all data balancing methods, a SMOTE-based method is selected, which generates new samples at the boundary of the samples distribution and avoids the synthesis of noise and redundant data.

Results: We used UCI and Kaggle datasets to simulate and evaluate the model. The experimental results show that VAE-based method outperforms other dimensionality reduction algorithms in the performance measures. The proposed hybrid method improves accuracy to 97.7% and sensitivity to 99.4%. Also, a feature importance analysis is provided to show insights into the factors driving the predictions and help understand the underlying mechanisms of heart disease.

Conclusion: Finally, it can be concluded that the combination of VAE with oversampling algorithms can significantly enhance system performance as well as computational time.

Introduction

Heart disease or cardiovascular disease is one of the leading causes of death in humans and its early diagnosis is quite challenging. Many studies are performed to improve the early detection of heart disease and reduce mortality. These studies aim to develop computer-aided diagnostic systems using emerging technologies. These systems predict heart disease based on data classification

algorithms; thus, the application of efficient algorithms plays an essential role in their accuracy. Many researchers have employed machine learning algorithms to construct diverse models and have attained remarkable accomplishments [1], [2]. To incorporate correlation between features, dimensionality reduction methods can be used. These methods can map the initial features into a space with fewer dimensions and extract effective features to feed the classification models. Many researchers have emphasized that feature reduction can improve performance and lead to faster processing per record.

Recently, AEs have excelled in unsupervised machine learning works for denoising data, compression and feature reduction. These networks can represent features in complex and large datasets with exceptional performance [3]. AEs can be considered as feedforward networks that their hidden layers have fewer neurons than the input and output layers. An AE is an encoder-decoder pair that generates an encoded representation and then reconstructs the input with encoded knowledge.

Usually, most dataset instances are normal and only a small percentage of them are related to abnormal or patient cases, as a result, the lack of patient instances may cause the model to not be properly fed and fully trained to recognize patients. Therefore, we use a data balancing phase to compensate for fewer patient instances than normal ones.

The Synthetic Minority Oversampling Technique (SMOTE) has promising results in addressing imbalanced data [4]. However, SMOTE has limitations, as it can generate noise and redundant data that do not significantly enhance the performance parameters. To overcome these limitations, improved versions of SMOTE, such as the Borderline Synthetic Minority Oversampling Technique (BSM) are proposed [5]. This technique focuses on generating samples at the boundary of the sample distribution to avoid the synthesis of noise and redundant samples.

When the training data has a large volume or high sample dimensions, there are problems such as the feature redundancy or feature irrelevance. In such a situation, SMOTE-based sampling methods lead to failure. Therefore, dimensional reduction methods can be helpful to implement sampling methods in low-dimensional space. The traditional dimension reduction method creates a great deal of redundancy in the feature space and the distribution of samples between the categories is mixed. This is a challenge for data synthesis with edge samples.

In this paper, a hybrid system is proposed that uses dimensionality reduction techniques namely, PCA, Information bottleneck (IB), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) and variational AE (VAE) to incorporate the correlation between features and extract the most essential features. Then, new samples are synthesized using BSM, especially at the boundary of the sample distribution. Finally, the combined samples are applied to train classification algorithms, including MLP, SVM and Logistic regression (LR) algorithms. We analyze the impact of dimensionality reduction and data balancing techniques on the performance of the classification algorithms. The experimental results show that VAE outperforms PCA and IB, besides, PCA has better computational time than VAE and IB. Also, data augmentation improves performance metrics. It can be concluded that the use of deep learning methods increases performance and efficiency, especially in large data sets. The results show that the proposed model using AE-based dimensionality reduction and BSM oversampling methods provides better performance, accuracy of 97.7% and sensitivity of 99.4%. The main contributions of the paper are as follows:

- Investigating the impact of applying three dimensionality reduction methods, PCA, IB, t-SNE, UMAP and VAE, on several classification algorithms using performance measures (accuracy, sensitivity, F1-score, precision, ROC- AUC score).
- Applying an improved SMOTE algorithm, BSM, after dimensionality reduction. This has a significant effect on the performance in two ways: First, essential features are restored and the synthesized data is generated based on these features. Second, after reducing the dimension, the problem of synthesis of noise data is solved.
- Proving the higher performance of VAE rather than the other dimensionality reduction techniques.
- Studying the effect of dimensionality reduction on computational time of large datasets.
- Proposing a hybrid model based on VAE and BSM with high accuracy and sensitivity 97.7% and 99.4%.

The rest of the paper is organized as follows: Section "Methodology" reviews the building blocks of the proposed model including dimensionality reduction techniques, oversampling methods and machine learning algorithms. The proposed model is described in Section "Architecture of the Model". Section "Experimental Results" shows experimental results and performance analysis. Finally, the paper concludes in Section "Conclusion".

Related Work

Bhatt et al. [6] examined the efficacy of several machine learning algorithms in predicting heart disease. They proposed a k-mode clustering algorithm that utilizes random forest, decision tree, multilayer perceptron, and XGBoost. Khan et al. [7] presented a hybrid machine learning method and performed experimental analysis.

Hassan et al. [8] proposed a system with combining a pre-trained Deep Neural Network (DNN) for feature extraction, Principal Component Analysis (PCA) for dimensionality reduction, and Logistic Regression (LR) for classification. The system demonstrated accuracy rates of 91.79% and 93.33% on the Cleveland dataset. In [9], authors developed a system based on machine learning and feature selection algorithms to achieve acceptable results. In [10], a system was developed that combines ensemble deep learning and feature fusion methods. This system utilized two algorithms, information gain and conditional probability, to reduce the number of features and assign specific weights to heart disease features. Following this, an ensemble deep learning classifier was trained to forecast heart disease in patients.

PCA is a common statistical technique that has found applications for finding patterns in high-dimensional data.

Results of recent research demonstrate that utilization of deep learning methods enhances the accuracy of predictions. For example, in [11], some machine learning techniques, including logistic regression (LR), SVM, deep neural network, decision tree, Nave Bayes, random forest and k-nearest neighbor are investigated and it concluded that DNN had the best performance with 98.15% accuracy and 98.68% sensitivity. Deep learning has been used successfully in various fields, especially in image analysis, visualization and working with large volumes of data. It is an evolving technique that is capable of representation of multi-level records [12]. DNN is a complex neural network with several hidden layers between the input and output layers. The input data is converted to nonlinear or activation functions to generate classes. In [2], a hybrid DNN is proposed to utilize convolutional neural network (CNN) and long short-term memory jointly. This method can predict heart disease with an accuracy of 93.7%.

For example, in [13], the authors used various feature selection techniques to forecast heart disease. In particular, they employed an SVM classifier for forward feature extraction, along with back-elimination feature selection. Their results demonstrated a reduction in the number of input variables, leading to an improvement of accuracy up to 85%. In another paper, Shao et al. [14] proposed a rough set strategies and multivariate adaptive regression splines to optimize the number of descriptive features and achieve an accuracy of 82.14%.

In recent research, applying newer feature selection algorithms such as fuzzy-based systems or DNN has significantly improved performance metrics. In [15], a hybrid model based on Fuzzy C-means and ANN along with PCA was proposed. PCA was used to select important features of the dataset. The extracted data from PCA was clustered using fuzzy C-means and finally, ANN was applied to predict cardiovascular disease. Its simulation results showed the effectiveness of the method with an accuracy value 99.55%, however, the precision is 33.27% and it requires a significant improvement.

In [16], the authors proposed a two-phase method in which the first phase involved sparse AE training to learn the best representation of training data. The second phase utilized ANN to predict health status based on trained records. Its experimental results showed that the model's accuracy is 90%, which shows a better performance than some traditional machine learning and neural network approaches. Authors of [17] proposed a system based on two deep neural networks that consist of one PCA and four deep learning models, including two variational AE and two DNN models. Ebiaredoh-Mienye et al. [18] proposed a model consisting of feature selection and classification phases that integrate an improved sparse AE and Softmax regression. They showed that the model has a robust feature learning algorithm and a high-performance classification.

Methodology

In this subsection, we briefly review the algorithms used in the proposed model, including data balancing techniques, dimensionality reduction methods and machine learning algorithms.

A. Data Balancing

The classification of imbalanced datasets is a challenging issue. When imbalanced data appear in the classification, the problem of overfitting arises and the result will be biased toward the majority class. In such situations, the data should be balanced either by oversampling or undersampling techniques to improve the performance. An oversampling technique increases the number of samples of the minority class, such as ADYSAN, SMOTE, SMOTE-TOMKE, etc. An undersampling technique reduces the number of samples of the majority class, like Dense Nearest Neighbor, Edited Nearest Neighbor and so on. We can apply a hybrid method such that oversampling is applied to the minority class to improve the model detection for the minority class samples, and undersampling is performed on the majority class to reduce the bias in the majority class samples.

SMOTE is an extended method that builds upon the random oversampling algorithm. Initially, it computes the Euclidean distance between K neighboring samples belonging to the same category surrounding each sample x_i in the minority class. Subsequently, a neighbor is randomly chosen, and a synthetic sample is created with a probability that falls on the line connecting the sample and its neighbor. The formula for synthesis can be represented as below:

$$x_{new} = x_i + r * (\hat{x_i} - x_i) \quad (1)$$

where r is a random number between [0,1], x_i is a sample

to be oversampled and $\hat{x}_i$ is a random neighbor sample.

The SMOTE method is one of the widely used methods for data synthesis, for which many improvements have been proposed so far. But most of them are either complicated or only focus on one of SMOTE's weakness. Among the proposed methods, the BSM approach is an approach that, in addition to removing noise data and detecting main features, also considers border data.

BSM categorizes the samples into safe, noisy and dangerous samples, where the dangerous samples are those that are on the boundary of the distribution. By generating synthetic data for these samples, the ability of the method to identify patient samples is significantly increased. Applying BSM can help the method to predict heart disease to a great extent, especially when minority samples are difficult to detect.

B. Dimensionality Reduction

Dimensionality reduction is a preprocessing step that reduces high-dimensional data to a controllable size while retaining the original information intact. It is a common step used for pattern recognition, classification applications and compression schemes. The dimensionality reduction has been effective in multiple aspects: first, the reduced representation combines different features of the records. Second, reducing the dimension speeds up the execution of the algorithm and improves the performance of the system in some cases. In this paper, several common dimensionality reduction methods have been used: PCA, IB and AE. AE-based model is shown to provide better performance while PCA-based model improves speed compared to IB and AE.

I) Principal Component Analysis (PCA)

PCA is a linear transformation that reduces the dimensionality of the input data, keeping its most significant parts. To achieve this, one must calculate the eigenvalues and eigenvectors of the data covariance matrix, then arrange the eigenvectors based on the eigenvalues in a descending manner and ultimately project the original data onto the directions of the eigenvectors. This method is suitable for fully correlated data. In practice, the only important PCA parameter that needs to be adjusted is the dimension of the projection space. This can be conveniently determined by examining the variance ratios of the principal components. Several types of improvements have been introduced for PCA. For example, possible principal component analysis (PPCA) was introduced to address the problem of missing values of features [19] or an extended PCA [20] was presented for applying on big data.

II) Information Bottleneck (IB)

IB introduced as an information-theoretic principle for extracting a compressed representation of the input data that maximizes a target prediction. It can be considered as an optimization problem that minimizes the mutual information I(Z; X) between the input variable X and its latent representation Z and it maximizes the mutual information I(Z;Y) between the output variable Y and the latent representation Z. In other words, it intends to maximize the following objective function:

$$\phi_{IB}^{\Theta} = I(Z;Y|\vartheta) - \beta I(Z;X|\vartheta) \quad (2)$$

where β ∈ [0, 1] manages the size of IB and θ is a Lagrange multiplier.

III) t-SNE

t-SNE represents a non-linear, unsupervised, and manifold-based feature extraction technique. It can map the high-dimensional data into a lower-dimensional space, typically comprising two or three dimensions, while maintaining the significant structure of the original data. Its primary application lies in the realms of data exploration and visualization. Although various feature extraction algorithms exhibit robust performance, they often struggle with visualizing high-dimensional data effectively and frequently fail to maintain both local and global data structures. In this context, t-SNE proves to be an advantageous tool for visualizing high-dimensional data by preserving the important structural attributes. The process begins with the application of Stochastic Neighbor Embedding (SNE), which transforms high-dimensional Euclidean distances into conditional probabilities that denote similarities between each pair of data points. Subsequently, a student t-distribution with one degree of freedom, similar to Cauchy distribution, is utilized to derive the second set of probabilities in the lower-dimensional space. Consequently, t-SNE aims to minimize the divergence between these two sets of probabilities across the high-dimensional and low-dimensional spaces [21].

IV) UMAP

The UMAP algorithm stands out as a strong competitor to t-SNE in terms of visualization quality, often demonstrating a greater ability to maintain global structure while offering enhanced computational efficiency. Additionally, UMAP imposes no limitations on the embedding dimension, rendering it a versatile option for dimension reduction in machine learning applications. While UMAP is similar to t-SNE, it also possesses significant differences that have led many practitioners to favor it for dimension reduction tasks. UMAP optimizes performance utilizing cross-entropy as the loss function in contrast to t-SNE's use of KL divergence, and employing stochastic gradient descent to optimize the cost function rather than the more time-consuming gradient descent method.

V) Autoencoder (AE)

The methods such as PCA may not fully succeed in

extracting the complex features of nonlinear datasets. In order to address this issue, AE as a deep learning model can be used. AE is trained to learn how to generate the original input with a minimum reconstruction error. It comprises two steps: the encoder, which transforms the d-dimensional input data into a latent representation, and the decoder, which reconstructs the representation to a vector resembling the original input. This process is known as reconstruction, with the difference between the decoder's output and the original input termed as reconstruction error.

Node layers identify input data patterns and use them to generate encrypted data representations. The network training algorithm adjusts the behavior of each node to be close to the configuration of the input data. If a linear activation function is applied, the AE becomes similar to a simple linear regression or PCA. But a nonlinear activation function, such as a rectified linear unit (ReLU) or a sigmoid function, makes the AE different from the PCA. The multiple types of AEs can be combined or modified to obtain new models for various applications. Among their widely used types, we can mention the types of variational AE (VAE), denoising AE (DAE) and sparse AE (SAE). VAE is enhanced with variational inference and parameterization to increase the model's ability in feature extraction and retain the diversity of the generated data. DAE takes a noisy input while training to recover the original undistorted input. By this means, the encoder can extract the most essential features and learn a robust representation of the input data. In SAE a sparsity constraint is imposed on the hidden nodes to mine essential information and avoid redundancy in large-scale datasets.

In this paper, we use a hybrid model based on VAE to enhance the model's ability in feature extraction while preserving the diversity of the generated data. The experimental results indicate that even with a 3 layered VAE, the model outperforms both IB and PCA.

C. Classifier Techniques

In the following, we review some common classifier methods that have high performance results, including MLP, SVM and Logistic Regression (LR) algorithms.

I) Multi-Layer Perceptron (MLP)

MLP is an artificial neural network that consists of an input layer, an output layer and multiple hidden layers instead of a single hidden layer. It is a feedforward network, meaning that each layer feeds the subsequent layer through a series of weights. MLP uses the backpropagation technique which is a supervised learning method. It has the capability to learn nonlinear models. Its multiple hidden layers and nonlinear activation function differentiate it from a linear perceptron. For applying MLP, several hyperparameters such as the number of hidden neurons, layers, and iterations must be adjusted.

II) Support Vector Machine (SVM)

SVM is a supervised machine learning technique used for classification regression and outliers detection. In SVM, a hyper-plane is created for separating different types of data. One of the advantages of SVM is that its training is computationally simple and unlike neural networks, it does not suffer from the problem of a local minimum. To accurately control the error rate, the kernel function and C parameter should be chosen correctly.

III) Logistic Regression

Logistic regression is a statistical method to classify an observation into one of two classes, or into one of many classes. It models the relationship between the independent features and the binary dependent variable (target) using the logistic function.

Architecture of the Model

Our proposed model combines feature reduction and data balancing. First, the initial data is preprocess to normalize. Then, the prepared data is employed in training the VAE, (t-SNE, UMAP, IB or PCA). After training, VAE (t-SNE, UMAP, IB or PCA) can differentiate between classes in the latent space, utilize BSM for interpolation of latent variables and synthesis new data. Finally, combination of the original and synthesized data is used for classification. The general architecture of the model is demonstrated in Fig. 1.

Since, VAE is a generative model, it can provide better performance and contribute for generating new samples using the latent variables. Therefore, we input the latent variables to the decoder to synthesis data. Subsequently, the decoder is eliminated and the encoder output is connected to the classifier and a combined network is created. The original data along with synthesized data are used to train the network. Algorithm 1 shows the pseudo code of the model based on VAE and BSM.

Algorithm 1. The proposed model based on VAE and BSM.

Input: Training data $T = \{t_1, t_2, \dots, t_n\}$;
Test data $S = \{s_1, s_2, \dots, s_m\}$

Output: predicated labels: $P = \{p_1, p_2, \dots, p_m\}$

1. Initialize VAE network;
2. Data preprocessing: remove missing data and normalize features to $[0,1]$. Output T' and S';
3. Feed T' to VAE's encoder and output $Z = \{z_1, z_2, \dots, z_n\}$.
4. Run BSM using input Z and output Z'.
5. Feed Z' to VAE's decoder and output new samples T_{new};
6. $T' = T' \; U \; T_{new}$;
7. Train the combined classification network by T'
8. For each $s_i, i = 1,2, \dots, m$:
   Return the prediction p_i;

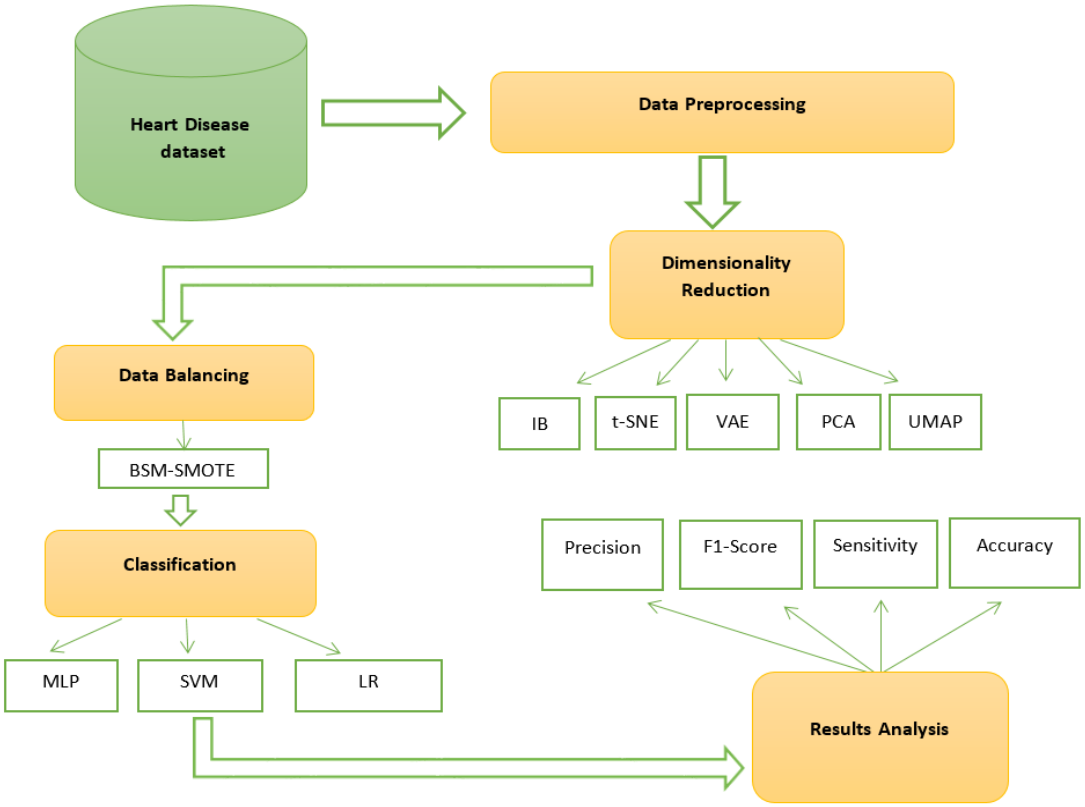


Fig. 1: Diagram of the proposed model.

In detail, the proposed model has mainly four phases:

1. Data preprocessing phase: this phase includes handling missing values, normalizing and shuffling data.

The dataset contains missing values that are handled by K-Nearest Neighbors imputation technique [22]. In the splitting data stage, we utilized 70% − 30% train-test data partitioning approach. The dimensionality reduction and data balancing steps are applied only to the training data.

2. Dimensionality reduction: In this phase, several dimensionality reduction methods, including PCA, IB, t-SNE, UMAP and VAE are applied to incorporate correlation between features. These dimensionality reduction methods are selected according to their performance and efficiency on heart disease prediction problem and the type of dataset.

3. Data balancing: Of all instances, 3596 are negative and 644 are positive. The lack of negative instances leads to low accuracy in predicting these cases (i.e., a high number of false negatives). Thus, we use an oversampling technique to generate samples of the minority class. However, most of these methods have limitations, as it can generate noise and redundant samples. Therefore, we use BSM which generates samples at the boundary of the sample distribution and avoids the synthesis of noise and redundant samples.

After balancing data, the performance of the models is improved. Since in most oversampling methods, the classification type data is discarded or examined separately, we balance the data after the dimensionality reduction.

4. Classification: At the end of the prediction process, the combined data is used as input to classification models. Among the classification methods, we selected three widely used categories that had higher performance than the others, MLP, SVM and LR.

5. Evaluation of methods: The performance of the models is evaluated and compared by the evaluation measures: accuracy, sensitivity, precision, F1-score. These measures are defined as follows:

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

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (5)$$

$$F1 = \frac{2*precision*sensitivity}{precision+sensitivity} \quad (6)$$

TN and TP denote true negative and true positive, i.e., they represent the number of patients and normal instances that are classified correctly. While FP and FN denote false positive and false negative, i.e., they represent the number of patients and normal instances that are incorrectly predicted.

Experimental Results

In this section, we use two publicly available datasets,

Kaggle heart dataset [23] and UCI heart disease repository [24]. First dataset consists of 4238 samples and 16 features. The second dataset consists of 597 samples and 13 features. Every feature is a risk factor that may be behavioral, demographic or medical. The features include age, knee joint pain status, waist circumference, neutral fat, BMI, SBP, TC, obesity status, frequency of eating out, HDL, weight change in one-year status, and marital status.

The heart disease database includes 15 features as input and its output is classified into two groups patient and normal. Since there is no general rule to adjust the parameters such as the number of hidden layers and the number of neurons in various layers, it is vital to obtain a good network structure with optimal performance.

D. Experiment Setup

We implemented the proposed model and determined the values of the parameters which provide optimal performance as follows:

VAE: The number of layers and neurons in the VAE is chosen based on a grid search over batch size (20, 30, 40), epochs (25, 50, 100, 200), neural network depth (2, 3, 4) and the dimensionality of the first hidden layer (40, 30, 20,12). we use Tanh and ReLU as activation functions and consider reduction size 4, 6, 8, 10 and 8. Also, we assume that the learning rate is 0.01, and ‘‘Adam’’ is used as the gradient descent algorithm. For example, we find the best architectures with two hidden layers listed in Table 1.

Table 1: The network structure parameters for VAEs with reduction sizes 4, 6, 8, 10

Reduction size	Architecture
4	15-30-10-4-10-40-15
6	15-20-8-6-8-20-15
8	15-20-10-8-10-20-15
10	15-20-12-10-12-20-15

IB: we consider the size of IB and Lagrange multiplier (θ) are 8 (size of reduction) and 0.95, respectively. Our implementation of IB is based on the Neural Network model for nonlinear information bottleneck [25].

PCA: it is enough to determine the optimal size to reduce the dimension.

t-SNE: We consider number of iterations and the value of α are 1000, respectively. Also, the perplexity is set to 30 to determines the number of nearest neighbors considered.

UMAP: We consider number of neighbors is 6 to balances local versus global structure in the data. Also, minimum distance is set to 0.3 to controls the minimum distance between points in the low-dimensional embedding.

MLP network: we apply five hidden layers with sizes 24, 30, 20, 15, 10, respectively. Also, the activation function and solver are selected Tanh and Adam, respectively.

However, in this method, adjusting the parameters is difficult and requires trial and error.

SVM: we use RBF as kernel function and consider parameters C = 1, gamma = 100000. With these selections, good results have been achieved.

LR algorithm: we apply it with the training parameter ridge estimator.

E. Performance Evaluation

In the following, we show performance measures with respect to the possible dimensionality reduction methods, data balancing, classification algorithms and the size of reduction. Observing the results of the methods with various reduction sizes, it was found that reduction size 8 has best performance measures values. Therefore, we have shown the performance parameters results of algorithms for reduction size 8.

Table 2-4 show the results of the performance measures obtained from models based on various dimensionality reduction methods, PCA, IB, t-SNE, UMAP and VAE, before and after data balancing for each UCI dataset and Kaggle dataset. We have shown the names of different methods as a combination of the dimension reduction method and classifier, so the possible methods will be the combination of PCA, IB, t-SNE, UMAP and VAE with MLP, PCA and LR.

Table 2 demonstrates the results using the MLP network. It can be seen, applying dimensionality reduction methods on MLP not only does not increase the performance, but also has a negative effect on it. However, data balancing improved performance significantly and the t-SNE-based method is better than other dimensional reduction algorithms. Table 3 indicates that the VAE-based method improves performance metrics of SVM algorithm, with an accuracy of 81.3, while the other dimensionality reduction algorithms, PCA and IB, have a negative effect on the performance. The BSM data balancing method, due to the synthesis of minority class samples in the boundary, reduces the value of FN and significantly strengthens the performance parameters.

As can be seen in Table 4, the effect of dimensionality reduction on LR method is similar to SVM, and VAE-based method improves the performance metrics. Also, applying data balancing on VAE-LR increases performance measures such as accuracy to 91.7 and sensitivity to 97.3. Comparing all the results in Table 2-4, shows that the best performance is achieved with the VAE-SVM method after data balancing. It improves accuracy to 97.1% and sensitivity to 99.2%.

Table 2: Performance comparison between methods based on different dimension reduction techniques before and afterdata balancing while we use MLP classification and reduction sizes 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively

Dataset		Methods	Accuracy (%)	F1-score (%)	Precision (%)	ROC-AUC score (%)	Sensitivity (%)
Kaggle	Original data	PCA-MLP	77.1	49.06	46.2	58.9	52.3
		IB-MLP	76.0	48.33	45.3	57.6	51.8
		tSNE-MLP	84.1	24.5	33.1	50.4	32.4
		UMAP-MLP	83.1	21.8	28.1	48.2	30.7
		VAE-MLP	79.0	50.18	47.1	60.4	53.7
		MLP	86.0	70.26	80.91	68.38	62.08
	Balanced data	PCA-MLP	88.5	89.29	86.3	83.4	92.5
		IB-MLP	86.6	82.07	77.2	74.1	87.6
		tSNE-MLP	95.5	95.1	95.1	95.5	95.1
		UMAP-MLP	89.1	88.3	95.4	89.5	79.1
		VAE-MLP	89.2	88.7	94.6	89.0	92.8
UCI	Original data	PCA-MLP	73.2	47.9	45.8	58.2	51.8
		IB-MLP	72.1	46.5	44.6	54.6	50.7
		tSNE-MLP	83.2	23.5	32.7	49.6	31.2
		UMAP-MLP	82.6	21.1	27.6	48.1	30.2
		VAE-MLP	77.3	48.2	47.3	58.3	52.2
		MLP	82.7	69.3	78.5	67.9	61.0
	Balanced data	PCA-MLP	87.8	89.1	84.6	85.3	90.5
		IB-MLP	86.2	81.5	76.7	75.3	86.5
		tSNE-MLP	944.4	93.9	93.8	93.9	94.5
		UMAP-MLP	88.6	87.5	93.1	88.5	78.0
		VAE-MLP	87.9	86.5	86.8	86.9	91.1

Table 3: Performance comparison between methods based on different dimension reduction techniques before and after data balancing while we use SVM classification and reduction sizes 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively

Dataset		Methods	Accuracy (%)	F1-score (%)	Precision (%)	ROC-AUC score (%)	Sensitivity (%)
Kaggle	Original data	PCA-MLP	77.1	49.06	46.2	58.9	52.3
		IB-MLP	76.0	48.33	45.3	57.6	51.8
		tSNE-MLP	84.1	24.5	33.1	50.4	32.4
		UMAP-MLP	83.1	21.8	28.1	48.2	30.7
		VAE-MLP	79.0	50.18	47.1	60.4	53.7
		MLP	86.0	70.26	80.91	68.38	62.08
	Balanced data	PCA-MLP	88.5	89.29	86.3	83.4	92.5
		IB-MLP	86.6	82.07	77.2	74.1	87.6
		tSNE-MLP	95.5	95.1	95.1	95.5	95.1
		UMAP-MLP	89.1	88.3	95.4	89.5	79.1
		VAE-MLP	89.2	88.7	94.6	89.0	92.8
UCI	Original data	PCA-MLP	73.2	47.9	45.8	58.2	51.8
		IB-MLP	72.1	46.5	44.6	54.6	50.7
		tSNE-MLP	83.2	23.5	32.7	49.6	31.2
		UMAP-MLP	82.6	21.1	27.6	48.1	30.2
		VAE-MLP	77.3	48.2	47.3	58.3	52.2
		MLP	82.7	69.3	78.5	67.9	61.0
	Balanced data	PCA-MLP	87.8	89.1	84.6	85.3	90.5
		IB-MLP	86.2	81.5	76.7	75.3	86.5
		tSNE-MLP	944.4	93.9	93.8	93.9	94.5
		UMAP-MLP	88.6	87.5	93.1	88.5	78.0
		VAE-MLP	87.9	86.5	86.8	86.9	91.1

Table 4: Performance comparison between methods based on different dimension reduction techniques before and after data balancing while we use LR classification and reduction sizes 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively

Dataset		Methods	Accuracy (%)	F1-score (%)	Precision (%)	ROC-AUC score (%)	Sensitivity (%)
Kaggle	Original data	PCA-LR	77.9	47.69	47.1	48.0-	48.3
		IB-LR	76.3	45.75	45.7	46.8	45.8
		tSNE-LR	68.8	78.3	73.2	50.0	69.9
		UMAP-LR	71.8	80.1	74.8	50.5	71.8
		VAE-LR	79.6	54.6	55.0	54.1	54.2
		LR	78.5	55.98	61.6	55.3	53.0
	Balanced data	PCA-LR	83.2	84.65	99.8	78.5	73.5
		IB-LR	81.6	87.65	82.8	77.6	93.1
		tSNE-LR	85.4	84.7	83.2	73.9	85.1
		UMAP-LR	86.4	85.1	84.6	75.0	86.4
		VAE-LR	**91.7**	**93.07**	**89.2**	**89.4**	**97.3**
UCI	Original data	PCA-LR	75.8	46.2	45.	46.2	46.2
		IB-LR	74.5	44.2	43.9	44.8	44.1
		tSNE-LR	71.3	80.0	73.9	50.0	70.9
		UMAP-LR	79.2	54.1	54.0	53.9	54.1
		VAE-LR	79.2	53.8	54.2	53.6	53.0
		LR	77.1	53.5	60.2	53.5	51.7
	Balanced data	PCA-LR	81.2	83.1	99.1	77.3	71.8
		IB-LR	80.8	86.5	81.3	76.2	91.8
		tSNE-LR	85.1	83.6	82.9	73.3	84.5
		UMAP-LR	85.8	84.7	84.5	74.2	86.1
		VAE-LR	90.5	92.1	88.4	88.3	96.0

Also, it can be seen, in all experiments, performance parameters were enhanced after data balancing. The reason is that, besides increasing the samples of the minority class, the applied data balancing algorithm (BSM) does not consider the noise data. Furthermore, applying the model to both datasets has the similar effect on performance. But since the number of samples of Kegel dataset is more than other dataset, the results are more reliable.

Fig. 2 shows comparison of accuracy values between various sizes of dimension reduction while VAE algorithm is used. It can be found that the accuracy of all methods is greatly enhanced after increasing the reduction size to 4 and it is maximized in 8.

F. Time Complexity

In this section, we provide experiments for computational efficiency. For each combination of the dimensionality reduction methods with the mentioned ML algorithms, the computational time of the network training is calculated. We use a computer with this specification: Intel Core i7 7700HQ, 2.60GHz, and 8GB RAM and also, we utilize Python 3.9 as the programming language. We also emphasized that the execution conditions are the same for all methods.

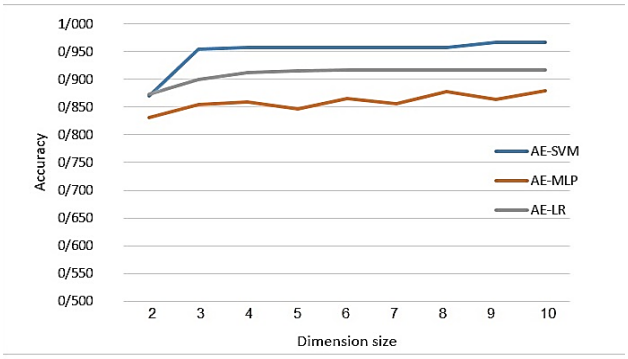


Fig. 2: Comparison of accuracy measure between various sizes of dimension reduction using VAE-based method for Kaggle dataset.

As can be seen in Fig. 4, the computational time for the LR in all cases is lower than the other classification methods. In addition, reducing dimension by using VAE usually improves computational time of the training while also increasing the performance.

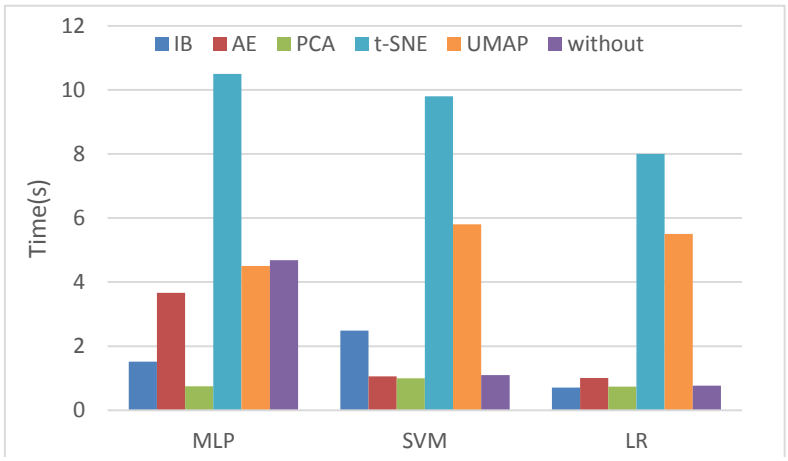


Fig. 4: Evaluation of time efficiency in the proposed model using various classifiers for Kaggle dataset while reduction sizes are 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively.

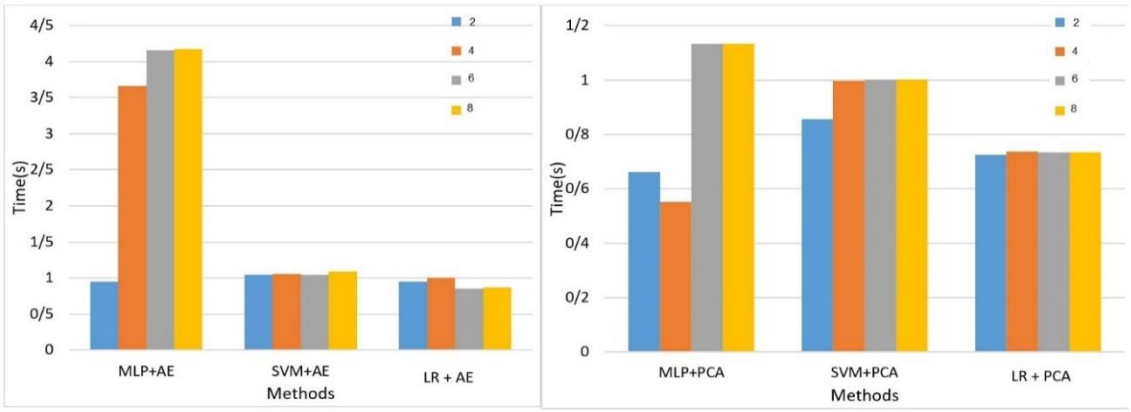


Fig. 3: The impact of reduction size on computational time using VAE (left) and PCA (right) for Kaggle dataset.

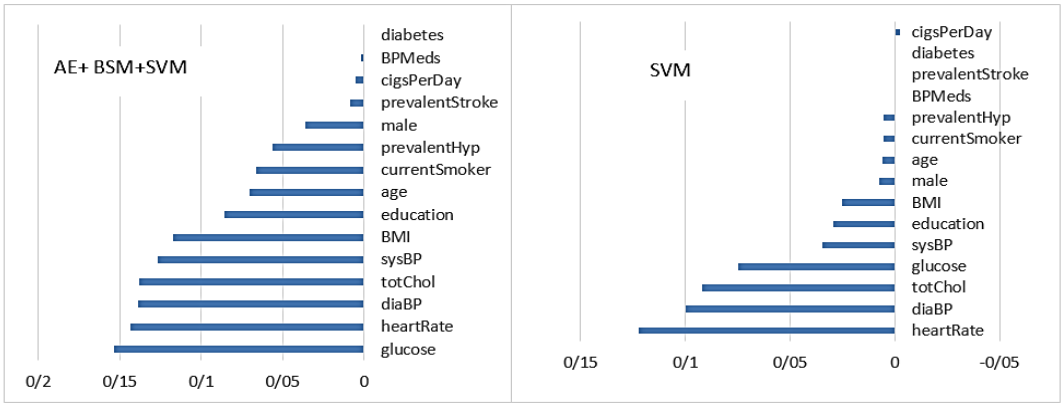


Fig. 5: Drop-column importance analysis on the proposed model for Kaggle dataset.

For large training sets, it was found that the difference in processing time is considerable. Also, t-SNE greatly increases the computational time.

Fig. 3 shows the impact of reduction size on computational time and indicates that in general, reducing the dimension decreases computational time, especially when the dataset is large.

I) Interpretability

Interpretability is a crucial aspect when it comes machine learning algorithms, especially deep learning algorithms, as they often produce models that are difficult to understand. These models, commonly known as black-box models, offer improved performance at the expense of complexity, making it challenging to comprehend the underlying mechanisms. Without it, even if accuracy is enhanced, the lack of transparency and accountability in the model may not be acceptable in medical settings. Research on interpretability has evolved

significantly due to the intricate nature of deep learning models, with various methods being employed to shed light on how these models operate. These methods include estimating feature importance, analyzing feature interactions, determining the contribution of specific layers or neurons, and interpreting models using high-level concepts that are more understandable to humans than low-level input features. We employed drop-column importance values to interpret the importance of the features in the proposed model, which has provided essential insights into the underlying mechanisms of disease prediction.

This information has the potential to assist clinicians in developing personalized treatment plans and risk management strategies for patients, ultimately leading to improved clinical outcomes. Fig. 5 shows the visualization results through drop-column importance on the VAE-based method and SVM method. The results indicate key features such as glucose, heartRate, diaBP, totChol and sysBP possess the highest importance value in VAE-based method.

It indicates that these variables play a crucial role in predicting heart disease. Similarly, variables heartRate, diaBP, totChol, glucose and sysBP have high importance value in SVM method. Therefore, the feature importance analysis discovered a consistent set of top 6 features, namely, glucose, heartrate, diaBP, totChol and sysBP which were very important in the prediction process. The results of the test indicate that there is no notable variance among the algorithms tested, as the dataset is limited and the supervised algorithms used are effective in yielding similar results.

Comparison

Now, we performed comparison study of our proposed model with the results from previous studies. The comparison results of the proposed model compared to the results given in other similar studies on Kaggle dataset is shown in Table 5. It can be seen that the proposed model demonstrated the high accuracy compared to previous studies results. In conclusion, our method outperformed most of these studies in accuracy, sensitivity, precision, F1-score, and AUC.

Table 5: Comparison the proposed model with other methods in recent studies

Authors	Approach	Accuracy	Precision	Sensitivity	F1-score	AUC
Saqlain et al. [26]	MFSFSA SVM	81.19	-	72.92	0.85	0.83
Mohan et al. [27]	HRFLM	88.4	90.1	92.8	90.0	-
Gupta et al. [28]	FAMD –RF	93.44	-	89.28	92.59	0.93
Fitriyani et al. [29]	DBSCAN-SMOTEE- XGBOOST	98.40	98.57	98.33	98.32	1.00
Bharti et al. [30]	DL-based Classifier	94.2	93.1	82.3	-	-
Hossain et al. [31]	Hybrid CNN-LSTM	74.15	81.82	72.04	76.62	73.95
Manikandan et al. [32]	Boruta feature selection	88.52	87.88	90.62	89.23	-
Proposed model	VAE- BSM - SVM	97.7	95.8	99.4	97.5	96.3

It is important to highlight that a direct comparison of the results may not be accurate due to application of different data pre-processing and training/testing methods. Moreover, the effectiveness of the prediction model is influenced by various factors including feature selection, data types and size, noise reduction, hyperparameters, data sampling, and model selection. Therefore, the overall comparison provided in Table 5 should not be solely relied upon to assess the performance of the prediction models. Instead, it can serve as a general comparison between the proposed model and previous research studies.

Conclusion and Future Work

Dimensionality reduction is a feature selection method that usually increases performance measures and computational speed of training. In this paper, we investigate the impact of some dimensionality reduction methods, including PCA, IB and VAE, on several machine learning algorithms in terms of performance measures and computational time. After implementing the model and reviewing the obtained results, we found that deep learning methods such as VAEs enhance the efficiency and the performance of the system. However, the effect of applying feature reduction on performance is negligible in some models. In addition, applying dimensionality reduction sometimes improves speed up to five times and sometimes does not affect. In an effort to better balance our training data, we use BSM data augmentation method. Finally, the hybrid model based on VAE and SVM achieves accuracy and sensitivity of 97.7% and 99.4% using Kaggle dataset.

In future works, the performance of the method can be enhanced to handle huge numbers of features and large volume of records. Additionally, the increasing emphasis on privacy, security, and time-sensitive applications shows the need to explore deeper into edge computing in order to enhance medical clinical decision support system.

Author Contributions

Vahidreza Afshin simulated the proposed method in Python. Seyed Hamid Zahiri and Saiedeh Kabirirad supervised and consulted in the design, implementation and results of this research. All authors discussed important sections and contributed to the final text.

Acknowledgment

We sincerely thank the respected referees for their accurate review of this paper.

Also, we sincerely thank the ICT Research Institute and Connectivity and Communication Technologies Development Headquarters and Dr. Kharrat.

Conflict of Interest

The authors announce no potential conflict of interest regarding the publication of this paper. Also, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission and redundancy have been completely witnessed by the authors.

Abbreviations

PCA	Principal Component Analysis
IB	Information Bottleneck
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder
SMOTE	The Synthetic Minority Oversampling Technique
BSM	the Borderline Synthetic Minority Oversampling Technique
LR	Logistic regression
DNN	Deep Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
RF	Random Forest
DT	Decision Tree
AUC	Area Under the ROC Curve
CM	Confusion Matrix
ROC	Receiver Operating Characteristic
DCNN	Deep CNN
MLP	Multi-Layer Perceptron
AE	Autoencoder
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative

References

[1] U. Kose et al., "A practical method for early diagnosis of heart diseases via deep neural network," in Deep Learning for Medical Decision Support Systems, pp. 95-106, 2021.

[2] A. A. Ali, H. S. Hassan, E. M. Anwar, A. Khanna, "Hybrid technique for heart diseases diagnosis based on convolution neural network and long short-term memory," in Applications of Big Data in Healthcare: Elsevier, pp. 261-280, 2021.

[3] D. Pratella, S. Ait-El-Mkadem Saadi, S. Bannwarth, V. Paquis-Fluckinger, S. Bottini, "A survey of autoencoder algorithms to pave the diagnosis of rare diseases," Int. J. Mol. Sci., 22(19): 10891, 2021.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., 16: 321-357, 2002.

[5] H. Han, W. Y. Wang, B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Proc. International Conference on Intelligent Computing: 878-887, 2005.

[6] C. M. Bhatt, P. Patel, T. Ghetia, P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, 16(2): 88, 2023.

[7] A. Khan, M. Qureshi, M. Daniyal, K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," Health Social Care Community, 2023(1): 1406060, 2023.

[8] D. Hassan, H. I. Hussein, M. M. Hassan, "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis," Biomed. Signal Process. Control, 79: 104019, 2023.

[9] S. Kabirirada, H. Kardanmoghaddamb, V. Afshin, "Heart disease prediction by using artificial neural networks," Int. J. Comput. Sci. Inf. Secur., 14(1), 2016.

[10] A. Ahmed, S. A. Hannan, "Data mining techniques to find out heart diseases: an overview," Int. J. Innovative Technol. Exploring Eng. (IJITEE), 1(4): 18-23, 2012.

[11] S. I. Ayon, M. M. Islam, M. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," IETE J. Res., 68(4): 2488-2507, 2022.

[12] K. Li, A. Zhu, W. Zhou, P. Zhao, J. Song, J. Liu, "Utilizing deep learning to optimize software development processes," arXiv preprint arXiv:2404.13630, 2024.

[13] S. Shilaskar , A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," Expert Syst. Appl., 40(10): 4146-4153, 2013.

[14] Y. E. Shao, C. D. Hou, C. C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification," Appl. Soft Comput., 14: 47-52, 2014.

[15] R. R. Ema, P. C. Shill, "Integration of fuzzy C-Means and artificial neural network with principle component analysis for heart disease prediction," in Proc. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT): 1-6, 2020.

[16] I. D. Mienye, Y. Sun, Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," Inf. Med. Unlocked, 18: 100307, 2020.

[17] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, Y. Piao, K. H. Ryu, "An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets," IEEE Access, 9: 135210-135223, 2021.

[18] S. A. Ebiaredoh-Mienye, E. Esenogho, T. G. Swart, "Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis," Electronics, 9(11): 1963, 2020.

[19] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis," Physica A, 482: 796-807, 2017.

[20] T. Zhang, B. Yang, "Big data dimension reduction using PCA," in Proc. 2016 IEEE International Conference on Smart Cloud (SmartCloud): 152-157, 2016.

[21] F. Anowar, S. Sadaoui, B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)," Comput. Sci. Rev., 40: 100378, 2021.

[22] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," J. Syst. Software, 85(11): 2541-2552, 2012.

[23] "Kaggle Cardiovascular Disease Dataset," Accessed 1 November 2022.

[24] "UCI Machine Learning Repository. Uci.edu.," Accessed 14 June 2022.

[25] A. Kolchinsky, B. D. Tracey, D. H. Wolpert, "Nonlinear information bottleneck," Entropy, 21(12): 1181, 2019.

[26] S. M. Saqlain et al., "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowl. Inf. Syst., 58: 139-167, 2019.

[27] S. Mohan, C. Thirumalai, G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, 7: 81542-81554, 2019.

[28] A. CS, S. Lal, V. PRABHU GURUPUR, P. P. Saxena, "Multi-modal medical image fusion with adaptive weighted combination of NSST bands using chaotic grey wolf optimization," IEEE Access, 7: 40782-40796, 2019.

[29] N. L. Fitriyani, M. Syafrudin, G. Alfian, J. Rhee, "HDPM: an effective heart disease prediction model for a clinical decision support system," IEEE Access, 8: 133034-133050, 2020.

[30] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," Comput. Intell. Neurosci., 2021(1): 8387680, 2021.

[31] M. M. Hossain et al., "Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI," Inf. Med. Unlocked, 42: 101370, 2023.

[32] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. Karthikeyan, R. Manikandan, A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," Inf. Med. Unlocked, 44: 101442, 2024.

Biographies

Saiedeh Kabirirad is an Assistant Professor in the Department of Computer Science, at Birjand University of Technology. She received her M.Sc. and Ph.D. in Computer Science from Shahid Beheshti University in 2010 and 2019, respectively. Her research area includes cryptography, information security, and data mining.

- Email: kabiri@birjandut.ac.ir
- ORCID: 0000-0003-1503-138X
- Web of Science Researcher ID: NA
- Scopus Author ID: 57204363739
- Homepage: NA

Vahidreza Afshin received the B.S. and M.S. degrees in Electrical Engineering in 2005 and 2012 from Islamic Azad University respectively. He is currently pursuing the Ph.D. degree in the Department of Electronics Engineering, University of Birjand, Iran. His research interest includes soft computing and optimization of control systems by machine learning algorithms.

- Email: vahidreza.afshin@birjand.ac.ir
- ORCID: 0000-0003-2621-515X
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

How to cite this paper:
S. Kabirirad, V. Afshin, S. H. Zahiri, “An effective heart disease prediction model using deep learning-based dimensionality reduction on imbalanced data,” J. Electr. Comput. Eng. Innovations, 13(2): 317-330, 2025.

DOI: 10.22061/jecei.2024.10847.742

URL: https://jecei.sru.ac.ir/article_2208.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

FATR: A Comprehensive Dataset and Evaluation Framework for Persian Text Recognition in Wild Images

Z. Raisi* *, V. M. Nazarzehi Had, E. Sarani, R. Damani

Electrical Engineering Department, Chabahar Maritime University, Chabahar, Iran.

Article Info

Article History:
Received 06 September 2024
Reviewed 11 Nvember 2024
Revised 27 December 2024
Accepted 30 December 2024

Keywords:
Persian scripts
Scene text recognition
Real-World datasets
Synthetic images
Deep learning
Farsi

*Corresponding Author's Email Address: *zobeir.raisi@cmu.ac.ir*

Abstract

Background and Objectives: Research on right-to-left scripts, particularly Persian text recognition in wild images, is limited due to lacking a comprehensive benchmark dataset. Applying state-of-the-art (SOTA) techniques on existing Latin or multilingual datasets often results in poor recognition performance for Persian scripts. This study aims to bridge this gap by introducing a comprehensive dataset for Persian text recognition and evaluating SOTA models on it.

Methods: We propose a Farsi (Persian) text recognition (FATR) dataset, which includes challenging images captured in various indoor and outdoor environments. Additionally, we introduce FATR-Synth, the largest synthetic Persian text dataset, containing over 200,000 cropped word images designed for pre-training scene text recognition models. We evaluate five SOTA deep learning-based scene text recognition models using standard word recognition accuracy (WRA) metrics on the proposed datasets. We compare the performance of these recent architectures qualitatively on challenging sample images of the FATR dataset.

Results: Our experiments demonstrate that SOTA recognition models' performance declines significantly when tested on the FATR dataset. However, when trained on synthetic and real-world Persian text datasets, these models demonstrate improved performance on Persian scripts.

Conclusion: Introducing the FATR dataset enhances the resources available for Persian text recognition, improving model performance. The proposed dataset, trained models, and code is available at https://github.com/zobeirraisi/FATDR.

Introduction

Text is a crucial source of visual information in our daily lives. It can be found everywhere, from documents and images to street signs, billboards, house numbers, and license plates. These texts provide vital details about location and identity and have various applications in real life [1]-[4]. Identifying text from input images involves two primary steps: first, accurately localizing the text instance (scene text detection), and second, converting the detected regions into word or character strings (scene text recognition).

Fig. 1: The challenges of Persian scripts in the wild images. (a) the same character with identical font, as seen inside of the red box, can appear in different shapes according to its position in the word instances, (b) a high degree of overlap in characters, and (c) The first three characters of the word "ی" are distinct characters, distinguished by the arrangement of small dots known as "Noghteh".

Table 1: Persian characters with similar body shapes

آ ا	ب پ ت ث	ج چ ح خ
د ذ	ر ز ژ	س ش
ص ض	ط ظ	ع غ
ف	ق	ک گ
ل	م	ن
ه	و	ی

Detecting and recognizing text in images with diverse characteristics such as color, font, orientation, language, and scene complexity is challenging. Traditional classical machine learning methods [5], [6] often struggle with complex scenarios. In contrast, recent deep learning-based approaches [7]-[15] have shown promising results in detecting and recognizing text even under hostile conditions. However, the majority of recent scene text detection and recognition has been conducted on Latin scripts, resulting in the development of multiple benchmark datasets for this purpose. This focus on Latin scripts has created a significant gap in text detection and recognition for non-Latin languages that use right-to-left scripts, such as Persian, Arabic, and Urdu. These languages have unique features that distinguish them from Latin writing systems, highlighting the importance of addressing this challenge with more attention.

Persian text recognition in the wild is a more challenging task due to its unique characteristics that differ significantly from Latin scripts. The complexity of this challenge is illustrated in Fig. 1 As seen, these challenges are connected letters from different positions (front, back, or side), diacritical marks, and the same character appearing differently in different positions of word instances (as shown in Fig. 1(a)), overlapping characters (shown in Fig. 1(b)). In addition, the Persian script is full of another specific challenge different from Latin text instances and that is the appearance of identical shapes characters with a different number of placement of dots (as shown in Fig. 1(c)) that causes problems in recognizing of these characters, which are illustrated in Table 1. These challenges independently pose a significant obstacle to current SOTA text detection and recognition (TDR) methods, which are mainly designed for Latin scripts. Persian letters are either horizontally or vertically oriented, with horizontal letters playing a crucial role in connectivity.

In contrast to Arabic and Urdu scripts, publicly available datasets for Persian scene text recognition are limited (See Section 2.2 for more details). While Persian scripts are similar to Arabic and Urdu, using existing Arabic or Urdu benchmark datasets may lead to poor performances. For instance, as seen in Table 2, the most recent Urdu scene text recognition model [16] that includes all classes of Persian alphabets in Table 3, still falls short of expected levels of word recognition accuracy.

Table 2: Comparing the word recognition accuracy (see Section 4.2.1) performance of the Urdu language model proposed in [16] on Urdu and Persian datasets. We used the test set of the cropped word images of our proposed dataset. Some sample images of both languages are provided in Fig. 2

Model	Urdu Dataset	Persian Dataset
Urdu-Large	92.97	**38.37**

Fig. 2: Comparison of (a) Urdu script and (b) Farsi script, where the model in Table 2 (Urdu-Large [16]) successfully recognized all the images in (a) while failing on images in (b).

As seen from Table 2 the UTR-Net with a WRA of 92.97% declines significantly ~50% on Persian scripts. Fig. 2 demonstrates some sample images of both languages with similar characters but different styles and fonts that the model in [16] successfully recognized the images in Fig. 2(a) while failing or missing some characters in Fig. 2(b) . Therefore, introducing or preparing a unique dataset for the Persian language is essential. Furthermore, the earlier proposed dataset for Persian scripts focused on only offering a synthetic dataset as in [17], focusing on single task detection as in [18], [19] or recognition or a specific kind of text instances like documents as in [20], [21], [63].

To address the mentioned problems, we introduce a new dataset for detecting and recognizing Persian text in real-world situations. The proposed FATR dataset is designed to be comprehensive and a good benchmark for measuring the robustness and generalizability performance of current and future models. To prepare this dataset, we captured a diverse collection of in-the-wild images tailored to the unique features of Persian script. We also built a large-scale synthetic Persian text dataset that can be used for training and evaluating Persian scene text recognition models. By addressing the

challenges of real-life scenarios, our study advances the field of text recognition, bridging the gap between Table 3.

Table 3: The Persian characters with their appearance in scripts

Persian Letter	Symbol	beginning	middle	end
همزه	ء أ	ئـ	ـئـ	ـئ ـأ ـؤ
الف	ا	آ		
بِ	ب	بـ	ـبـ	ـب
پِ	پ	پـ	ـپـ	ـپ
تِ	ت	تـ	ـتـ	ـت
ثِ	ث	ثـ	ـثـ	ـث
جیم	ج	جـ	ـجـ	ـج
چِ	چ	چـ	ـچـ	ـچ
حِ	ح	حـ	ـحـ	ـح
خِ	خ	خـ	ـخـ	ـخ
دال	د			ـد
ذال	ذ			ـذ
رِ	ر			ـر
زِ	ز			ـز
ژِ	ژ			ـژ
سین	س	سـ	ـسـ	ـس
شین	ش	شـ	ـشـ	ـش
صاد	ص	صـ	ـصـ	ـص
ضاد	ض	ضـ	ـضـ	ـض
طا	ط	طـ	ـطـ	ـط
ظا	ظ	ظـ	ـظـ	ـظ
عین	ع	عـ	ـعـ	ـع
غین	غ	غـ	ـغـ	ـغ
فِ	ف	فـ	ـفـ	ـف
قاف	ق	قـ	ـقـ	ـق
کاف	ک	کـ	ـکـ	ـک
گاف	گ	گـ	ـگـ	ـگ
لام	ل	لـ	ـلـ	ـل
میم	م	مـ	ـمـ	ـم
نون	ن	نـ	ـنـ	ـن
واو	و			ـو
هِ	ه	هـ	ـهـ	ـه
یِ	ی	یـ	ـیـ	ـی

The main contributions are summarized as follows:

1. We propose a Persian text recognition dataset. To the best of our knowledge, this is the first publicly available dataset that contains various text instances captured in wild images from different environments, considering all the challenges in Latin benchmark datasets. This dataset can be used as a benchmark for future research.
2. We also present a large-scale synthetic dataset for Persian scent text recognition of about 200K cropped word images.
3. We review the past and recent advancements in scene text recognition for Latin and non-Latin scripts.
4. We train six well-known SOTA scene text recognition model, and evaluate, compare, and analyze quantitatively and qualitatively their performances on the proposed FATR dataset.

Related Work

A. Scene Text Recognition

In scene text recognition, the main objective is identifying the characters or words present within the detected text regions in the given input images. This task is more complex than recognizing printed scanned documents because real-world images pose various challenges, such as low resolution, extreme lighting, diverse fonts, orientations, languages, and lexicons compared to the clean background scanned or printed documents. To address these difficulties, researchers have proposed different methods based on both classical machine learning techniques such as [22]-[24], and deep learning techniques such as [10]-[13], [25].

Classical machine learning-based methods [22], [26], [27] typically use features like HOG [28] or SIFT [29] in combination with classifiers such as SVM [30]. These methods either adopt a bottom-up approach, where classified characters are linked into words, or a top-down approaches that directly recognize entire words from the image [31]. However, these methods often struggle to recognize new words that are not part of the training dataset, and they have limited capabilities in representing features that are essential for real-world scenarios. Additionally, classical methods are often unable to recognize input word images that are multi-oriented or curved, which are common in wild images.

On the other hand, recent scene text recognition methods utilize deep learning architectures to address the challenges of complex real-world scenarios. Inspired by speech recognition, many recent methods model scene text as a sequence of characters [10]-[12], which are called sequence-based methods. These methods leverage techniques like Connectionist Temporal Classification (CTC) [32] to predict character sequences. However, these methods are designed for 1-dimensional (1D) sequences, and converting 2D image features to 1D leads to information loss, hindering the recognition of irregular text. To address this, researchers proposed a 2D-CTC [33] technique that directly operates on 2D probability distributions, achieving better recognition accuracy.

The attention mechanism initially used for machine translation has also been adopted for scene text recognition [11], [34]. Attention allows the model to focus on specific image regions during decoding, enhancing the

recognition of irregular text. Different attention-based frameworks have been proposed, ranging from basic 1D-attention models to more complex methods that employ rectification or character-aware techniques to handle various text distortions. However, some methods have difficulty recognizing images with complex backgrounds or high computational costs. With the advancement of the transformer architecture [35], many recent scene text recognition models [36], [37] have utilized the transformer in their pipeline and achieved SOTA performance in several benchmark datasets with complex and challenging word images.

B. Datasets

Latin Scripts: The scene text recognition benchmarks can be categorized into two general categories: regular text datasets, including ICDAR13 [38], III5k [39], and SVT [23], which contain primarily horizontal text instances, and irregular text datasets, including ICDAR15 [40], CUT80 [41], SVT-P [42], and COCO-Text [43], which contain challenging multi-oriented and curved text instances.

Researchers also pre-trained their models on synthetic images to achieve a more general and higher accuracy performance. SynthText (ST) [44] and MJSynth (MJ) [45] synthetic datasets are two datasets that have been used extensively for the pre-training purposes of scene text detection and recognition algorithms.

Multi-Lingual Scripts: Researchers used several multi-lingual text datasets to measure the performance of their models. ICDAR17-MLT [46] and ICDAR19-MLT [47] are two examples of multilingual datasets that contain the following languages: Arabic, Latin, Chinese, Japanese, Korean, Bangla, and Hindi. There are also some other datasets [48]-[51] that have scripts, mostly in English and Chinese, that are designed for the specific purposes of text recognition.

Right-to-Left Scripts: Arabic, Urdu, and Persian are three languages that use similar letter scripts but are distinct when spoken. Different from left-to-right languages such as Latin and Chinese, finding publicly available benchmark datasets specifically designed for these languages can be challenging. However, numerous conventional and modern techniques have been developed on private datasets for these languages. In this case, Arabic and Urdu are better suited than Persian. For instance, ICDAR17-MLT [46] and ICDAR19-MLT [47] are two publicly available benchmarks that contain Arabic scripts and can be utilized for training and evaluating Arabic text recognition. ARASTEC [52] and ARASTI [53] are two real-world datasets used for Arabic character and word recognition in natural images, respectively. However, these datasets are not available to the public. For Urdu text, IIITH [54] and UPTI [55] are two well-known Urdu script datasets that include real-world and synthetic text instances. Recently, a work by Rahman et al. [16] introduced the UTRSet-Real and UTRSet-Synth datasets, which are publicly available.

Regarding Persian, the only real-word dataset currently is PESTD [18]. This dataset is unique in that it is a Persian English dataset with images captured in the wild. However, it is only designed for traffic sign detection and mainly has images of road traffic signs. Moreover, it has yet to consider the recognition task, which is the most challenging aspect. Furthermore, the dataset is private for testing. A publicly available dataset contains synthetic images, namely ITDR-Synth [17], designed for both detection and recognition. This dataset contains 6,100 and 40,220 images for detection and recognition, respectively.

Farsi Text Recognition (FATR) Dataset

In this section, we present our comprehensive Persian language dataset tailored specifically for investigating and analyzing Persian text recognition challenges. For this purpose, a synthetic dataset for the recognition task that contains text instances of captured images of both indoor and outdoor environments. Table 4 shows more details of the proposed FATR dataset. The indoor images contain images of dense and small text instances taken from indoor store signs and products. All the outdoor images consist of a wide variety of challenging cases of urban landscapes, such as storefronts, street signs, wall signs, and traffic signs.

Table 4: The proposed Farsi (Persian) text dataset

Text Category		Recognition	
		Train	Test
Indoor Text	Product	648	158
	Lobbies	2080	748
Outdoor Text	Storefronts	7796	1951
	Street signs	261	81
	Graffiti	346	92
	Traffic signs	1804	501
All		12935	3529

Fig. 3 shows different sample images of FATR that are taken with different camera phones.

Fig. 3: Sample real-world images of the proposed FATR dataset.

A. Real-World Text Recognition

We converted the quadrilateral boxes into rectangular boxes and cropped them to prepare the recognition dataset. The recognition dataset consists of 12935-word images cropped for training and 3529 for testing. In Fig. 4, you can see some sample images and their corresponding strings.

Fig. 4: Example of real-world cropped word patch images of the proposed FATR dataset used for training and evaluation of recognition model taken from different indoor and outdoor places.

We consider various challenges when preparing the recognition images of FATR datasets. Fig. 5 illustrates some sample images of these challenges including partially occluded text, rotated text, illumination variation, low resolution, English text, image blurriness, complex background, difficult fonts, and special characters (*e.g.*, /,() -,:, @,#,…).

We provide a probability distribution of the cropped word images in the FATR dataset, focusing on image height, width, and character length. Fig. 6 illustrates these computations. As shown in Fig. 6(a) and Fig. 6(b), the FATR dataset contains word images of varying resolutions, with a minimum width of 4 pixels a minimum height of 5 pixels, and a maximum width of 4391 pixels, and a maximum height of 2505 pixels.

As seen in Fig. 6(c), the length of the word instances ranges from 1 character to 22 characters, with an average length of approximately 5 characters per word. It is worth mentioning that the dataset contains a total of 5795 unique word instances.

B. Synthetic Persian Dataset

The SOTA scent text techniques are trained on a combination of large synthetic cropped word images of SynthText [44] and MJ-Synth [45], and they achieved good performances on real-world benchmark datasets. In this paper, we also introduce FATR-Synth, a synthetic dataset of ~200K word images of Persian scripts, and the real-world images of FATR for training the SOTA scene text recognition models. Inherited from [16], these images are created using different text attributes like font, size, and color and include over 200 Persian fonts. It addresses the scarcity of Persian words and numerals in existing datasets by incorporating sufficient samples and provides a vocabulary of 200,000 words and ~50000 unique Farsi words collected from the Internet with an average length of 8 characters for synthetic text generation. The produced dataset is also publicly available for further research. Fig. 7 illustrates some examples of the prepared synthetic dataset.

Fig. 5: Challenges in real-world images from the FATR dataset. The identified challenges include OC (occluded text), MO (multi-oriented), IV (illumination variation), LR (low resolution), ML (multi-language), IB (image blurriness), CB (complex background), DF (difficult fonts), and SC (special characters).

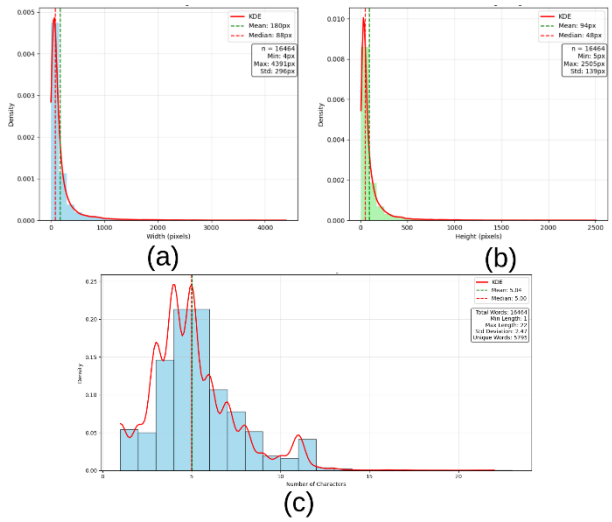


Fig. 6: Probability distribution of (a) word width in pixels, (b) word height in pixels, and (c) the number of characters in a word image computed from the FATR dataset. Best viewed when zoomed.

Fig. 7: Sample synthetic word images of the proposed FATRSynth dataset. This dataset is only used as training. Each cropped word image mainly contains one-word instances.

Experimental Results

This section presents our comprehensive evaluation for investigating some selected SOTA text recognition models including CRNN [10], STAR-Net [12], ROSETTA [13], CLOVA [14], and UTR-Net [16] on the proposed FATR dataset. Table 4 shows the number of test images used for the evaluation of these models.

A. Implementation Details

All models used in this paper are trained and tested on a machine equipped with an NVIDIA GPU with plate number RTX-3090. To ensure a fair comparison, we train all the models in comparisons on a similar dataset. For the recognition task, we follow the same settings described in [14] to train the models. We use a combination of FATR-Synth and FATR real-world images for training. To generate the images of FATR-Synth, we follow the settings provided in [16][1].

In this paper, we only focus on evaluating Persian characters. Therefore, we only consider Persian text instances during inference and ignore English text. We evaluate the selected recognition models using the Word Recognition Accuracy (WRA) and Normalized Edit Distance (NED) evaluation metrics. We train all recognition models, except UTRNeT [16], for 200,000 iterations, while the UTRNeT model is trained for 50 epochs. We use the following 35 characters for training all the evaluation models: ا, آ, ب, پ, ت, ث, ج, چ, ح, خ, د, ذ, ر, ز, ژ, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ک, گ, ل, م, ن, و, ه, ء, ی, ئ.

B. Evaluation Metrics

For the recognition task, given a set of cropped word images, we use two metrics for the evaluation of recognition models: Word Recognition Accuracy (WRA) and Normalized Edit Distance (NED). WRA is mainly used to evaluate the accuracy of scene text recognition schemes [10], [12]-[14]., which can be calculated as:

$$\mathrm{WRA} = \frac{\#\ \text{accurately Recognized Words}}{All\ the\ word\ instances} \times 100 \tag{1}$$

The NED metric is defined as follows [47]:

$$Norm = 1 - \frac{1}{N}\sum_{i=1}^{N} D(w_i, w_i') / \max(w_i, w_i') \tag{2}$$

where Levenshtein Distance is shown by $D(:)$ [56]. w_i and w_i' are ground truths corresponding to the text regions and predicted word strings, respectively.

C. Quantitative Results

We conducted several experiments on the recognition using pre-training models. Persian scripts' characters are fundamentally different from those of left-to-right languages, and using a SOTA right-to-left language model would result in a significantly low WRA margin (See Table 2 in Section 1). Therefore, we selected some of the best and most well-known recognition models [10], [12]-[14], [16], trained them in a similar setting, and evaluated them on the FATR dataset. Table 5 shows the quantitative results of our experiments. We first trained these models on synthetic images in the ITDR-Synth dataset, which resulted in poor performance. However, when we used our proposed FATR-Synth dataset, the models' WRA performance improved significantly. We further enhanced the models' performance by combining natural and synthetic images of the proposed FATR dataset. Among these models, CLOVA [14] achieved the best performance regarding both WRA and edit distance (ED) for all our experiments. Our quantitative results confirm that training the models on the proposed FATR dataset can significantly improve the recognition performance compared to the only publicly available synthetic Persian dataset [17].

Table 5: Experimental Results of the select scene text recognition models [10], [12]-[14], [16] on the proposed FATR dataset. The results trained on the synthetic images of ITDR-Synth proposed in [17] and our proposed FATR-Synth are shown with blue and red colors, respectively. All the model results trained on the combination of our proposed synthetic and real-world images of FATR are shown in black color. The WRA and NED denote the word recognition accuracy and the normalized edit distance.

Model	Trained Dataset	WRA	NED
CRNN [10]	IDTR-Synth	22.28	0.45
	FATR-Synth	45.65	0.75
	FATR-Synth+FATR-Real	53.04	0.78
ROSETTA [13]	IDTR-Synth	19.26	0.44
	FATR-Synth	36.84	0.72
	FATR-Synth+FATR-Real	65.7	0.85
STARNET [12]	IDTR-Synth	27.17	0.48
	FATR-Synth	64.66	0.85
	FATR-Synth+FATR-Real	68.74	0.86
CLOVA [14]	IDTR-Synth	27.68	0.50
	FATR-Synth	64.94	0.85
	FATR-Synth+FATR-Real	69.24	0.87
UTRNet [16]	IDTR-Synth	24.72	0.49
	FATR-Synth	52.98	0.77
	FATR-Synth+FATR-Real	66.93	0.86

D. Qualitative Results

We tested the models listed in Table 5 to demonstrate their performance on real-world images. To that effect, we evaluated the qualitative results on various cropped word images from the FATR dataset, as presented in Fig. 8. The output strings of Fig. 8(a)-(c) demonstrate that the chosen models can accurately recognize regular text

[1] https://github.com/abdur75648/urdu-synth/

instances with horizontal or near-horizontal orientation. However, some selected models failed or produced the wrong output for one or two characters when evaluated on challenging examples, such as rotated text or text with complicated font styles, as shown in Fig. 8(d)-(g). Ultimately, all the models we evaluated produced false recognition when the text was vertically oriented, partially occluded, or used a complex font in adverse situations.

Fig. 8: Qualitative results among the selected recognition models [10], [12]-[14], [16] on some images of the FATR dataset. Each output strings stand for the following models: I) CRNN [10], II) ROSETTA [13], III) STAR-Net [12], IV) CLOVA [14], and V) UTRNet [16]. These models are trained on the combination of real and synthetic images of the FATR dataset. The green and red colors denote the accurate and inaccurate characters predicted by the models.

E. Discussion and Future Work

Persian text recognition in the wild presents significant challenges, one of which is the time-consuming and costly annotation process. Recent advancements in artificial intelligence, such as DALLE [57], ChatGPT [58], and Gemini [59], have made it possible to generate images using text prompts, providing a way to address this issue. Another approach to tackle the annotation problem is automating the process, which can be achieved using the latest model [60]. However, detecting and recognizing Latin text, as well as text images taken from the wild, still pose various challenges, such as orientation, occlusion, and degradation in image quality. These challenges remain unsolved problems in the computer vision community. To overcome these challenges, techniques like augmentation, compositionality [61], and masked autoencoder transformers [62] combined with AI modules can be used to assist the models. As future work, we also aim to design a deep learning-based architecture by utilizing the above-mentioned advancement in computer vision and natural language processing to tackle the shortcomings of the current state-of-the-art models and capture and recognize challenging word images from wild images.

Conclusions

This paper highlights a critical research gap in right-to-left text recognition in wild images, specifically for Persian scripts. We have introduced a comprehensive Persian text recognition dataset to address this issue, which provides real-world and synthetic images to evaluate SOTA text recognition models. We have evaluated several scene text recognition models on the proposed FATR dataset. Our experimental results have shown that the current multilingual text dataset can still perform well on Persian scripts. For Persian scent text recognition, a specialized dataset is essential to accurately recognize text instances in wild images.

Acknowledgment

The authors express their acknowledgement to the JECEI reviewers, editors, and editorial board for their constructive feedback, valuable suggestions, and professional support, which improved the quality of our paper.

Author Contributions

Z. Raisi collected the data, implemented the code, carried out the analysis, and wrote the paper. The other authors also equally collected the data, edited the paper and interpreted the results.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication or falsification, double publication and, or submission, and redundancy, have been completely witnessed by the authors.

Abbreviations

FATR	Farsi (Persian) text recognition
SOTA	State-of-The-Art
WRA	Word Recognition Accuracy
NED	Normalized Edit Distance
CTC	Connectionist Temporal Classification

References

[1] Y. Zhu, C. Yao, X. Bai, "Scene text detection and recognition: Recent advances and future trends," Front. Comput. Sci., 10(1): 19-36, 2016.

[2] H. Lin, P. Yang, F. Zhang, "Review of scene text detection and recognition," Arch. Comput. Methods Eng., 27: 433-454, 2020.

[3] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, J. Zelek, "Text detection and recognition in the wild: A review," arXiv preprint arXiv:2006.04305, 2020.

[4] Z. Raisi, J. Zelek, "Text detection and recognition for robot localization," J. Electr. Comput. Eng. Innov., 12(1): 163-174, 2024.

[5] K. Wang, B. Babenko, S. Belongie, "End-to-end scene text recognition," in Proc. 2011 International Conference on Computer Vision: 1457-1464, 2011.

[6] A. Bissacco, M. Cummins, Y. Netzer, H. Neven, "PhotoOCR: Reading text in uncontrolled conditions," in Proc. 2013 IEEE International Conference on Computer Vision: 785-792, 2013.

[7] Z. Raisi, V. M. Nazarzehi, “A transformer-based approach with contextual position encoding for robust persian text recognition in the wild,” J. AI Data Min., 12(3): 455-464, 2024.

[8] Z. Raisi, G. Younes, J. Zelek, “Arbitrary shape text detection using transformers,” in Proc. 2022 26th International Conference on Pattern Recognition (ICPR): 3238-3245, 2022.

[9] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman., “Deep structured output learning for unconstrained text recognition,” arXiv:1412.5903v5, 2015.

[10] B. Shi, X. Bai, C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Trans. Pattern Anal. Mach. Intell., 39(11): 2298-2304, 2016.

[11] B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, “Robust scene text recognition with automatic rectification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 4168- 4176, 2016.

[12] W. Liu, C. Chen, K. Y. K. Wong, Z. Su, J. Han, “STARNet: A spatial attention residue network for scene text recognition,” in Proc. British Machine Vision Conference (BMVC): 43.1-43.13, 2016.

[13] F. Borisyuk, A. Gordo, V. Sivakumar, “Rosetta: Large scale system for text detection and recognition in images,” in Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining: 71-79, 2018.

[14] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, H. Lee, “What is wrong with scene text recognition model comparisons? Dataset and model analysis,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV): 4715-4723, 2019.

[15] C. Ma, L. Sun, J. Wang, Q. Huo, “Dq-detr: Dynamic queries enhanced detection transformer for arbitrary shape text detection,” in Proc. International Conference on Document Analysis and Recognition: 243-260, 2023.

[16] A. Rahman, A. Ghosh, C. Arora, “Utrnet: Highresolution urdu text recognition in printed documents,” in Proc. International Conference on Document Analysis and Recognition: 305-324, 2023.

[17] F. Alimoradi, F. Rahmani, L. Rabiei, M. Khansari, M. Mazoochi, “Synthesizing an image dataset for text detection and recognition in images,” J. Inf. Commun. Technol., 53(53): 78, 2023 [In Farsi].

[18] A. Rashtehroudi, A. Ranjkesh, A. Shahbahrami, "PESTD: a large-scale Persian-English scene text dataset," Multimedia Tools Appl., 82: 34793-34808, 2023.

[19] S. Kheirinejad, N. Riaihi, R. Azmi, “Persian text-based traffic sign detection with convolutional neural network: A new dataset,” in Proc. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE): 060- 064, 2020.

[20] M. Rahmati, M. Fateh, M. Rezvani, A. Tajary, V. Abolghasemi, “Printed persian ocr system using deep learning,” IET Image Process., 14(15): 3920-3931, 2020.

[21] A. Fateh, M. Rezvani, A. Tajary, M. Fateh, “Persian printed text line detection based on font size,” Multimedia Tools Appl., 82(2): 2393-2418, 2023.

[22] T. E. De Campos, B. R. Babu, M. Varma, et al., “Character recognition in natural images,” in Proc. Fourth International Conference on Computer Vision Theory and Applications (VISAPP), 7: 273-280, 2009.

[23] K. Wang, S. Belongie, “Word spotting in the wild,” in Proc. European Conference on Computer Vision: 591-604, 2010.

[24] L. Neumann, J. Matas, “Real-time scene text localization and recognition,” in Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition: 3538-3545, 2012.

[25] F. Zhan, S. Lu, “Esir: End-to-end scene text recognition via iterative image rectification,” in Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition: 2059-2068, 2019.

[26] M. Sawaki, H. Murase, N. Hagita, “Automatic acquisition of context-based images templates for degraded character recognition in scene images,” in Proc. 15th International Conference on Pattern Recognition (ICPR), 4: 15-18, 2000.

[27] Y. F. Pan, X. Hou, C. L. Liu, “Text localization in natural scene images based on conditional random field,” in Proc. 2009 10th International Conference on Document Analysis and Recognition: 6-10, 2009.

[28] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1: 886-893, 2005.

[29] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. of Comp. Vision, 60(2): 91-110, 2004.

[30] J. A. Suykens, J. Vandewalle, “Least squares support vector machine classifiers,” Neural Process. Lett., 9(3): 293-300, 1999.

[31] J. Almazan, A. Gordo, A. Forn´ es, E. Valveny, “Word´ spotting and recognition with embedded attributes,” IEEE Trans. Pattern Anal. Mach. Intell., 36(12): 2552-2566, 2014.

[32] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proc. 23rd International Conference on Machine Learning: 369-376, 2006.

[33] Z. Wan, F. Xie, Y. Liu, X. Bai, C. Yao, “2D-CTC for scene text recognition,” arXiv:1907.09705v1, 2019.

[34] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, “Aster: An attentional scene text recognizer with flexible rectification,” IEEE Trans. Pattern Anal. Mach. Intell., 41(9): 2035-2048, 2018.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need,” in Proc. 31st Conference on Neural Information Processing Systems (NIPS 2017): 5998-6008, 2017.

[36] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, J. Zelek, “2lspe: 2d learnable sinusoidal positional encoding using transformer for scene text recognition,” in Proc. 2021 18th Conference on Robots and Vision (CRV): 119-126, 2021.

[37] Z. Qiao, Z. Ji, Y. Yuan, J. Bai, “Decoupling visual semantic features learning with dual masked autoencoder for self-supervised scene text recognition,” in Proc. International Conference on Document Analysis and Recognition: 261-279, 2023.

[38] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. De Las Heras, “ICDAR 2013 robust reading competition,” in Proc. 2013 12th International Conference on Document Analysis and Recognition: 1484-1493, 2013.

[39] A. Mishra, K. Alahari, C. V. Jawahar, “Scene text recognition using higher order language priors,” in Proc. BMVC, 2012.

[40] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al., “ICDAR 2015 competition on robust reading,” in Proc. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015.

[41] A. Risnumawan, P. Shivakumara, C. S. Chan, C. L. Tan, “A robust arbitrary text detection system for natural scene images,” Expert Syst. with Appl., 41(18): 8027- 8048, 2014.

[42] T. Quy Phan, P. Shivakumara, S. Tian, C. Lim Tan, “Recognizing text with perspective distortion in natural scenes,” in Proc. IEEE International Conference on Computer Vision (ICCV): 569-576, 2013.

[43] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, "Microsoft coco: Com-´ mon objects in context," in Proc. 13th European Conference on Computer Vision: 740-755, 2014.

[44] A. Gupta, A. Vedaldi, A. Zisserman, "Synthetic data for text localisation in natural images," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 2315-2324, 2016.

[45] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," arXiv preprint arXiv:1406.2227, 2014.

[46] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, D. Karatzas, "ICDAR2017 robust reading challenge on omnidirectional video," in Proc. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1: 1448-1453, 2017.

[47] Y. Sun, Z. Ni, C. K. Chng, Y. Liu, C. Luo, C. C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al., "ICDAR 2019 competition on large-scale street view text with partial labeling– RRC-LSVT," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.

[48] W. Wu, Y. Zhao, Z. Li, J. Li, M. Z. Shou, U. Pal, D. Karatzas, X. Bai, "Icdar 2023 competition on video text reading for dense and small text," in Proc. International Conference on Document Analysis and Recognition: 405–419, 2023.

[49] R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, et al., "ICDAR 2019 robust reading challenge on reading Chinese text on signboard," in Proc. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.

[50] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, J. Liu, D. Karatzas, C. Seng Chan, L. Jin, "Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art," in Proc. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.

[51] Z. Wan, J. Zhang, L. Zhang, J. Luo, C. Yao, "On vocabulary reliance in scene text recognition," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 11425-11434, 2020.

[52] M. Tounsi, I. Moalla, A. M. Alimi, F. Lebouregois, "Arabic characters recognition in natural scenes using sparse coding for feature representations," in Proc. 2015 13th International Conference on Document Analysis and Recognition (ICDAR): 1036-1040, 2015.

[53] M. Tounsi, I. Moalla, A. M. Alimi, "Arasti: A database for arabic scene text recognition," in Proc. 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR): 140-144, 2017.

[54] M. Jain, M. Mathew, C. Jawahar, "Unconstrained ocr for urdu using deep cnn-rnn hybrid networks," in Proc. 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR): 747- 752, 2017.

[55] N. Sabbour, F. Shafait, "A segmentation-free approach to arabic and urdu ocr," in Proc. Document recognition and retrieval XX, 8658: 215-226, 2013.

[56] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in Soviet physics doklady, 10: 707-710, 1966.

[57] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, 1(2): 3, 2022.

[58] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[59] G. Team, R. Anil, S. Borgeaud, Y. Wu, J. B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., "Gemini: a family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.

[60] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, et al., "Segment anything," in Proc. the IEEE/CVF International Conference on Computer Vision: 4015- 4026, 2023.

[61] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," arXiv preprint arXiv:2006.15538, 2020.

[62] Z. Raisi, J. Zelek, "Occluded text detection and recognition in the wild," in Proc. 2022 19th Conference on Robots and Vision (CRV): 140-150, 2022.

[63] A. Faraji, M. Saeed, H. Nezamabadi-pour, "Introducing a database for Farsi document image understanding and segmentation," J. Mach. Vision Image Process., 10(2): 31-46, 2023 [In Persian].

Biographies

Zobeir Raisi received his Ph.D. degree in 2022 from the Vision Image Processing Lab (VIPLab) at the Systems Design Engineering Department, University of Waterloo, Waterloo, Ontario, Canada. Currently, he is an Assistant Professor in the Department of Electrical Engineering at Chabahar Maritime University, Iran. His research interests include computer vision, artificial intelligence, and robotics.

- Email: zobei.raisi@cmu.ac.ir
- ORCID: 0000-0002-1591-4492
- Web of Science Researcher ID: GLV-1410-2022
- Scopus Author ID: 54897975500
- Homepage: https://www.cmu.ac.ir/staff/zraisi

Valimohammad Nazarzehi Had received his Ph.D. degree in 2016 from the University of New South Wales, Australia. Currently, he is an Assistant Professor in the department of Electrical Engineering, Chabahar Maritime University, Iran. His research interests include decentralized control, marine control systems, and control of mobile robots, robotics, and image processing.

- Email: v.nazarzehi@cmu.ac.ir
- ORCID: 0000-0003-3261-6320
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.cmu.ac.ir/staff/vnazarzehi

Esmaeil Sarani is a Ph.D. graduate in Electrical Power Engineering from the University of Tehran and currently serves as an Assistant Professor in the Department of Electrical Engineering at the Chabahar Maritime University, Iran. His research interests include the design of electrical machines, fault detection in electrical machines, renewable energy development with a focus on wave energy harvesting, and the application of artificial intelligence in various domains.

- Email: sarani@cmu.ac.ir
- ORCID: 0000-0001-6598-2410
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.cmu.ac.ir/staff/sarani

Rasoul Damani received the B.Sc, M.Sc and Ph.D. degrees from Sharif University of Technology (SUT), Tehran, Iran in 1998, 2000 and 2015 respectively, all in Electrical Engineering. He is currently an Assistant Professor at Chabahar maritime university, Chabahar, Iran. His research interests include the areas of optical communication and underwater communication systems.

- Email: damani@cmu.ac.ir
- ORCID: 0000-0002-4748-0684
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.cmu.ac.ir/staff/rdamani

How to cite this paper:
Z. Raisi, V. M. Nazarzehi Had, E. Sarani, R. Damani, "FATR: A comprehensive dataset and evaluation framework for persian text recognition in wild images," J. Electr. Comput. Eng. Innovations, 13(2): 331-340, 2025.

DOI: 10.22061/jecei.2024.11256.784

URL: https://jecei.sru.ac.ir/article_2253.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Advanced Race Classification Using Transfer Learning and Attention: Real-Time Metrics, Error Analysis, and Visualization in a Lightweight Deep Learning Model

M. Rohani, H. Farsi, S. Mohamadzadeh *

Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 29 September 2024
Reviewed 07 December 2024
Revised 06 January 2025
Accepted 10 January 2025

Keywords:
Race classification
Attention module
Efficient-Net network
Transfer learning
Real-time performance

*Corresponding Author's Email Address:
s.mohamadzadeh@birjand.ac.ir

Abstract

Background and Objectives: Recent advancements in race classification from facial images have been significantly propelled by deep learning techniques. Despite these advancements, many existing methodologies rely on intricate models that entail substantial computational costs and exhibit slow processing speeds. This study aims to introduce an efficient and robust approach for race classification by utilizing transfer learning alongside a modified Efficient-Net model that incorporates attention-based learning.

Methods: In this research, Efficient-Net is employed as the base model, applying transfer learning and attention mechanisms to enhance its efficacy in race classification tasks. The classifier component of Efficient-Net was strategically modified to minimize the parameter count, thereby enhancing processing speed without compromising classification accuracy. To address dataset imbalance, we implemented extensive data augmentation and random oversampling techniques. The modified model was rigorously trained and evaluated on a comprehensive dataset, with performance assessed through accuracy, precision, recall, and F1 score metrics.

Results: The modified Efficient-Net model exhibited remarkable classification accuracy while significantly reducing computational demands on the UTK-Face dataset. Specifically, the model achieved an accuracy of 88.19%, reflecting a 2% enhancement over the base model. Additionally, it demonstrated a 9-14% reduction in memory consumption and parameter count. Real-time evaluations revealed a processing speed 14% faster than the base model, alongside achieving the highest F1-score results, which underscores its effectiveness for practical applications. Furthermore, the proposed method enhanced test accuracy in classes with approximately 50% fewer training samples by about 5%.

Conclusion: This study presents a highly efficient race classification model grounded in a modified Efficient-Net architecture that utilizes transfer learning and attention-based learning to attain state-of-the-art performance. The proposed approach not only sustains high accuracy but also ensures rapid processing speeds, rendering it ideal for real-time applications. The findings indicate that this lightweight model can effectively rival more complex and computationally intensive recent methods, providing a valuable asset for practical race classification endeavors.

Introduction

The advances of recent years in the field of artificial intelligence (AI) and deep learning (DL) have significantly improved the accuracy of facial recognition, image classification, and object detection. Among these advancements, race classification (RC) from facial images remains a critical and inherently challenging task due to the subtle differences in facial features across various

racial groups and the extensive diversity among human faces [1], [2]. This capability holds immense potential for applications in security, human-computer interaction, and social and demographic analysis.

RC is not only an academic problem but also has significant real-world implications. For instance, in security and surveillance, accurate RC can enhance monitoring and identification [3], [4]. In personalized user experiences such as augmented reality (AR) and virtual reality (VR), understanding racial features can improve user interaction [5]. In healthcare, accurate RC can aid in providing tailored medical advice and interventions, as certain medical conditions are prevalent in different racial groups [6]. However, ethical considerations are crucial to address the potential biases and fairness issues in RC.

In addition to the major benefits such as model accuracy and efficiency, this study emphasizes the importance of lightweight models in resource-constrained environments. Lightweight architectures are especially useful for deployment on devices with limited computational resources, such as mobile phones and embedded platforms [7], [8]. This makes advanced RC features feasible and applicable across various domains, from consumer electronics to remote sensing technologies. The novelty of this research lies in its holistic approach, integrating real-time performance metrics, error analysis, and detailed visualization to provide a comprehensive understanding of model behavior and identify areas for improvement. Real-time performance metrics are crucial for applications demanding instant results, while error analysis can pinpoint specific challenges and potential biases in the classification process. Visualization techniques offer intuitive insights into how the model perceives diverse racial features, aiding in the refinement of the model.

This study demonstrates the powerful combination of convolutional neural network (CNN) architectures and transfer learning techniques when applied to complex classification tasks like race recognition [9]-[12]. The findings confirm the high accuracy and efficiency of the model, establishing a foundational benchmark for further research and development in AI and deep learning. This research lays the groundwork for future endeavors in developing robust and ethical racial classification systems applicable across diverse technological and societal contexts. The quest for high-performance, ethical, and practical racial classification models represents a critical and continuously evolving frontier in the field of artificial intelligence. By addressing the complexities of race classification, this work contributes to a more equitable and responsible application of AI technologies.

Related Work

In recent years, race recognition has become a prominent topic in facial recognition and image processing [13]-[15]. Numerous studies have been conducted to improve the accuracy and efficiency of various methods for this purpose. Al-Azani and El-Alfy (2019) examined race recognition methods in challenging conditions using Histogram of Oriented Gradients (HOG) features [16]. Their research demonstrated that HOG features could be used for race recognition under various lighting and background conditions [17]. However, a major drawback of this approach is its sensitivity to changes in scale and angle, which can result in decreased accuracy when dealing with noisy images or unexpected variations due to its reliance on low-level features. In a comparative study between machine learning and deep learning methods for age, gender, and race recognition, Hamdi and Moussaoui found that deep learning methods generally performed better than machine learning methods, especially in race recognition [18]. Krishnan et al. investigated the fairness of gender classification algorithms across different gender-race groups [19]. The results indicated that there were performance disparities among these algorithms across different groups, highlighting the need to address such disparities in the development of future algorithms. Ahmed et al. utilized deep networks for race estimation. Their study showed that deep networks could achieve high accuracy in race estimation, particularly when leveraging diverse data combinations [20]. Belcar et al. focused on race recognition using Convolutional Neural Networks (CNNs) and the middle part of the face [21]. Their results indicated that utilizing specific facial regions could enhance the accuracy of race recognition algorithms.

However, reliance on specific facial parts may lead to decreased overall model performance in scenarios where these parts are not fully visible or affected by external factors [22]. Patel et al. introduced a shift-invariant deep neural network for tri-fold classification [23]. This network demonstrated strong performance against spatial variations in input data, providing notable results. Lastly, Wirayuda et al. proposed a compact-fusion feature framework for race recognition, which improved the accuracy of recognition algorithms by using combined features [24]. However, the high complexity and need for combined processing of this framework could impact its real-time performance. Despite these advancements, there remains a need for improvements in race recognition accuracy in real-world conditions and in environments with limited computational resources. This research aims to address this gap by presenting an optimized model that reduces resource consumption and parameters while maintaining high accuracy. Our model, utilizing advanced optimization techniques and novel methods, is designed to offer robust performance under varying and challenging conditions while minimizing computational demands.

Proposed Method

In this section, proposed method is presented for the classification of races from facial images using transfer learning based on state-of-the-art deep learning architectures. At the center of the approach is the Efficient-Net model, which is well known for its trade-off between top results and low computation cost. Our method addresses many built-in difficulties of RC, such as subtle differences in facial features among racial groups and the need for a balanced, diverse dataset to train the model effectively. We first used Efficient-Net as the base model since it is already known for its success in image classification works [25]. The advantage of this architecture is that it adopts a compound-scaling approach, where all network dimensions are scaled up uniformly to bring about improved performance while consuming less computation cost. However, since RC is very specific in nature, using Efficient-Net directly would not be enough. We will make use of transfer learning to fine-tune the pre-trained Efficient-Net model to our targeted RC task. By this method, we retrain the upper layers to adapt the model with the unique characteristics of racial features. To enhance the preprocessing step, we employed the Multi-task Cascaded Convolutional Networks (MTCNN) for accurate face detection and alignment [26]. MTCNN is effective in extracting facial regions from images, ensuring that the input to the model focuses solely on the relevant facial features. This step is crucial for improving the overall accuracy of the RC process by eliminating background noise and variations in face alignment. Fig. 1 illustrates the progression of the proposed method.

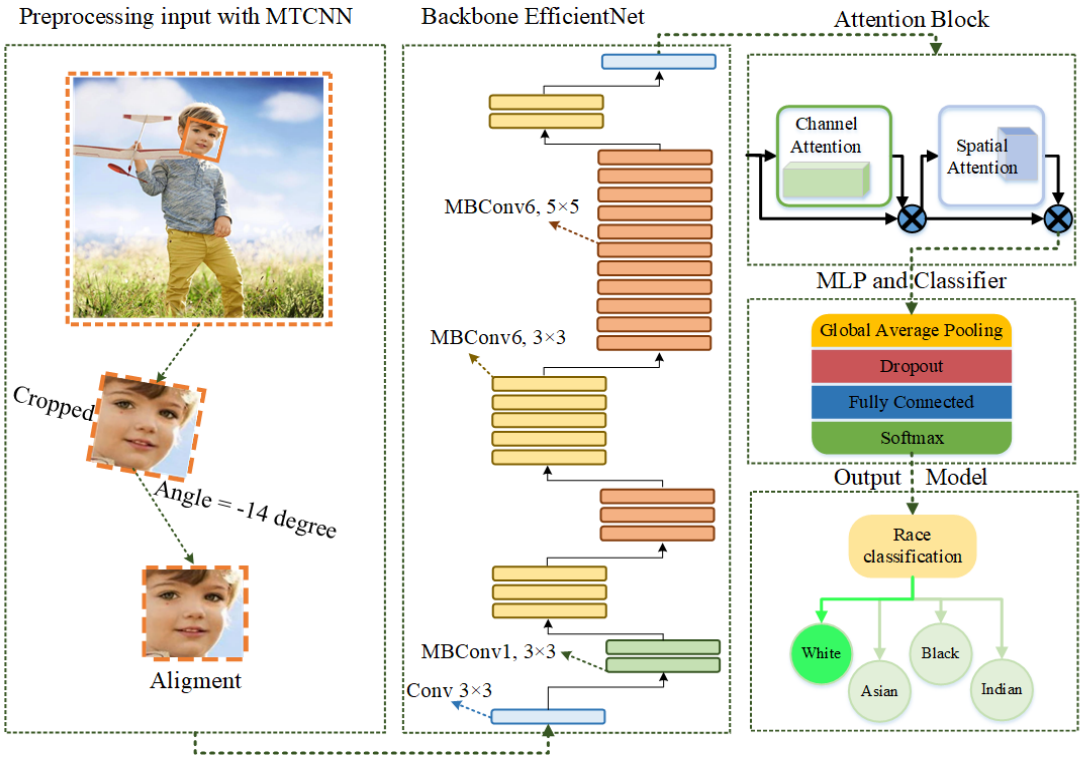


Fig. 1: An overview of proposed method.

A. Attention Mechanism

In the proposed methodology, the integration of the convolutional block attention module (CBAM) into the Efficient-Net architecture, specifically positioned after the convolutional layers, introduces a refined approach to enhancing feature extraction for race classification. This attention mechanism enables the network to dynamically focus on both salient channels and critical spatial regions within the feature maps, which is vital for effectively distinguishing subtle racial characteristics. In race classification, where minor facial feature variations across different ethnicities are key, conventional convolutional layers may fail to sufficiently capture these nuanced differences. CBAM addresses this limitation by applying dual attention mechanisms, improving the network's sensitivity to discriminative features [27]. CBAM module shows in Fig. 2.

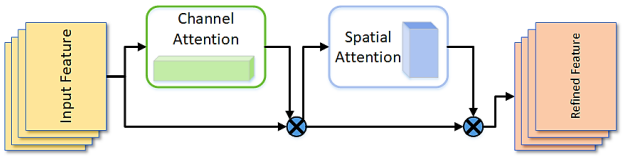


Fig. 2: Schematic representation of the CBAM architecture.

Formally, given an intermediate feature map FM, which belongs to a three-dimensional space denoted $H \times W \times C$, where H represents the height of the feature map, W represents its width, and C indicates the number of channels, CBAM first applies channel attention followed by spatial attention. In this context, the dimensions H, W, and C correspond to the spatial and depth characteristics of the feature map extracted from the convolutional layers, with the height and width representing the two-dimensional spatial resolution and the channel reflecting the depth or number of filters applied in the convolutional process as expressed in (1).

$$CA(FM) = \sigma(MLP(AvgPool(FM)) + MLP(MaxPool(FM))) \quad (1)$$

where MLP denotes a multi-layer perceptron, and σ represents the sigmoid activation function. This operation selectively enhances important channels by considering both average and max-pooled representations of the feature map. Subsequently, spatial attention is applied as expressed in (2).

$$SA(FM) = \sigma(Conv_{7\times7}(AvgPool(FM); MaxPool(FM)) \quad (2)$$

where $Conv_{7\times7}$ signifies a convolutional layer that processes the concatenation of average and max-pooled feature maps across the channel dimension, thereby refining spatial feature selection. The incorporation of CBAM in this manner allows for a more targeted feature representation, capturing both global dependencies and localized variations in facial structures pertinent to racial differentiation. By combining channel and spatial attention, the proposed approach offers a novel enhancement to Efficient-Net, enabling the model to better distinguish subtle racial traits, thereby improving classification accuracy in datasets with complex race-related features.

B. Data Balancing

Finally, the issue of data imbalance is addressed, which makes the classification task challenging with race datasets. An imbalanced dataset can further lead to biased models, performing poorly on the under-represented classes. In order to cope with that, we used a lot of data augmentation and resampling strategies. Data augmentation will create different training examples for rotations, scaling, flipping, and so on. Resampling can be implemented using techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or oversampling for a certain group to ensure its representation in the dataset. Table 1 presents the pseudocode of the proposed method.

In the process of machine learning with imbalanced datasets, the SMOTE (Synthetic Minority Over-sampling Technique) is utilized as an advanced method for balancing the number of samples across classes.

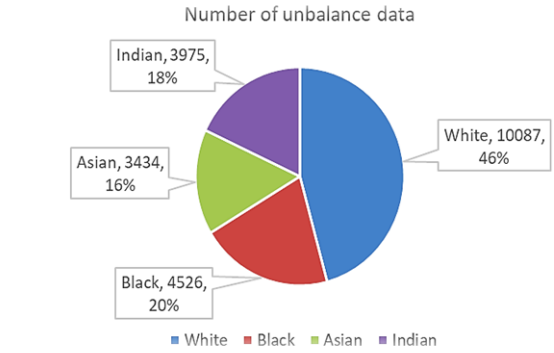
Table 1: Pseudocode of proposed method

- Start
1. # Load and preprocess the data
2. Data_list = Load data (dataset_path)
3. Extract face images by ***MTCNN*** **()**
4. Split data into ***training*** and ***testing*** sets
5. Balance ***training*** data with ***SOMTE*** method
6. Split balance data into ***training*** and ***validation*** sets
7. # Define the Efficient-Net model
8. Initialize ***Efficient-NetB3*** with ***Image-Net*** weights
9. Add ***Attention CBAM*** to ***base_model***
10. Add ***GlobalAveragePooling*** layer to ***base_model***
11. Add ***Dropout*** layer with a dropout rate of ***0.5***
12. Add ***Dense layer*** with ***softmax*** activation function
13. Compile model using ***'Adam'*** optimizer and ***'sparse_categorical_crossentropy'*** loss function
14. # Train the model
15. Trained_model = *Train model* (model, training_set_images)
16. # Metrics: ***Accuracy***, ***Precision***, ***Recall***, and ***F1 score***
17. # Evaluate the model
18. Validation_metrics = *EvaluateModel* (trained_model, validation_set_images)
19. # Test the model
20. Test_metrics = *TestModel* (trained_model, test_set_images)
- return ***Validation_metrics, Test_metrics***

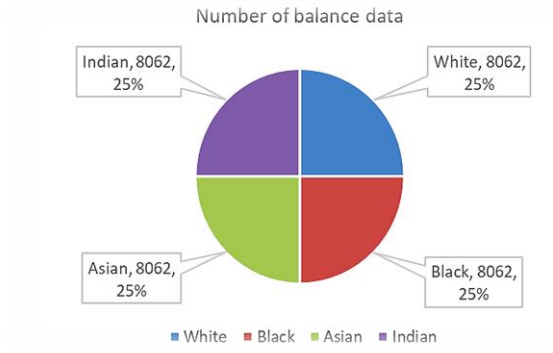
This technique generates new data by creating samples based on the existing minority class samples rather than merely duplicating them randomly. This approach not only enhances the diversity of the data but also mitigates the bias resulting from class imbalance. Consequently, the SMOTE technique has been employed in this study to address the issue at hand.

In Fig. 3, the total number of images is 22,013, whereas the total should amount to 22,022 based on the individual class sample counts. During the filtration process, nine samples were removed from the data frame due to being corrupted. For the training and processing of deep learning models, the dataset consisting of 22,013 images with four distinct racial labels (White, Black, Asian, and Indian) was utilized. To achieve balance in data distribution and enhance model quality, the dataset was initially divided into two main sections: training (train) and testing (test), with 80% of the data allocated for training and 20% for testing. Subsequently, the training data was further divided into two subsets for model validation: final training (train final) and validation. Accordingly, 90% of the training data was designated for final training and 10% for validation. This strategic division ensured that the racial distribution was preserved in each subset, allowing the model to be trained effectively on

balanced data. Finally, after applying the SMOTE technique to equalize the number of samples in each class, the new distribution resulted in 8,062 images for each racial category.



(a)



(b)

Fig. 3: Distribution of Images by racial labels before (a) and after (b) data filtration.

Results and Discussion

This section presents the main findings from our experiments, showcasing the performance of the proposed method across various evaluation metrics. The results are summarized in tables and figures to provide a clear and concise overview of the data. These findings will, in the following sections, serve as the basis for discussing their significance and relevance to existing research in the field.

A. Criterion

In evaluating the performance of machine learning models, four primary metrics are commonly used: accuracy, precision, recall, and the F1-score. These metrics provide a comprehensive assessment of a model's effectiveness [28].

Accuracy: This metric simply represents the ratio of correctly predicted instances to the total number of predictions. In other words, in (3) accuracy measures the overall correctness of a model by considering both true positives and true negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

Precision: Precision indicates the proportion of positive predictions that are actually correct. This metric is particularly important when the cost of false positives is high in (4).

$$Precison = \frac{TP}{TP + FP} \tag{4}$$

Recall: Recall measures in (5) the percentage of actual positive instances that are correctly identified by the model. It is crucial when the cost of missing positive cases is high.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both metrics in (6). It is especially useful when dealing with imbalanced class distributions.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

RAM: Memory consumption (RAM) refers to the amount of memory required for storing data and model parameters during the training and evaluation of machine learning models. This metric can be calculated using (7).

$$RAM_{MB} = \frac{P \times Size\_of\_data\_type}{1024^2} \tag{7}$$

where P is the total number of parameters in the network and $Size\_of\_data\_type$ is the size of the data type (typically 4 bytes for float32). The number of parameters includes weights and biases across all layers, including convolutional and dense layers. Specifically, in a convolutional layer with K filters, C input channels, and kernel dimensions $H \times W$, the number of parameters is calculated in (8).

$$P_{conv} = K \times (H \times W \times C + 1) \tag{8}$$

B. Dataset

UTK-Face provides over 20,000 facial images with very wide coverage of racial backgrounds, age groups, and both genders. This will come in quite handy for research into racial classification, where the dataset is diverse and comes with a detailed label including race, age, and gender for each image [29]. The dataset allows for the making of a model capable of precise and fair RC from facial features, thus being important for its real-world application. Fig. 4 displays a diverse set of facial images representing different racial categories included in the dataset.

Fig. 4: Sample images from the UTK-Face dataset [29].

Table 2 provides a comparison between the base model and the proposed model in terms of parameters and memory usage. The base model comprises a total of 12,324,539 parameters, of which 12,237,236 are trainable, resulting in an approximate RAM consumption of 47.01 MB, indicating a higher resource requirement. In contrast, proposed model 1 demonstrates a reduced parameter count of 10,789,683, with 10,702,380 trainable parameters, leading to a lower memory consumption of 41.16 MB. Proposed model 2, although slightly more resource-intensive than Proposed model 1, still presents an improvement over the base model, featuring 11,379,605 total parameters and a RAM requirement of 43.41 MB. This reduction in both total and trainable parameters, along with decreased memory consumption, highlights the greater efficiency of the proposed models while ensuring a significant number of trainable parameters, which is essential for maintaining model performance. Additionally, the lower memory footprint of these models makes them particularly well-suited for deployment in resource-constrained environments, enabling wider implementation in scenarios where computational resources are limited.

Table 2: Model parameters and memory usage overview

Model	Parameters		
	Total	Trainable	RAM (MB)
Base model	12,324,539	12,237,236	47.01
Proposed model 1	10,789,683	10,702,380	41.16
Proposed model 2	11,379,605	11,292,302	43.41

Table 3 provides a comprehensive analysis of the performance metrics for the training and testing phases of the models evaluated on an NVIDIA Tesla T4 with 16GB GDDR6 RAM. As indicated in Fig. 3, the base model, trained on 22,013 images, achieved a training duration of 2249 seconds over 50 epochs. In contrast, both Proposed model 1 and Proposed model 2, which were trained on an expanded dataset of 32,248 images, required slightly longer training times of 2275 seconds and 2282 seconds, respectively.

In terms of testing efficiency, the base model demonstrates a processing time of 0.35 milliseconds per image, resulting in a frame rate of 2827 frames per second. Proposed model 1 shows improved performance with a reduced image processing time of 0.28 milliseconds and an increased frame rate of 3571 frames per second. Proposed model 2 also performs well, processing each image in 0.30 milliseconds and achieving a frame rate of 3352 frames per second. These results indicate that while the proposed models have a higher training time, they capitalize on a larger dataset to enhance their efficiency during testing. The substantial parameter reductions observed in the proposed models, as outlined in Table 2, coupled with their improved testing speeds, suggest that these models are not only more resource-efficient but also better suited for applications that demand high-speed image processing. This efficiency in real-time applications is paramount, highlighting the potential for deploying the proposed models in scenarios requiring rapid processing capabilities.

Table 3: Performance metrics for model training and testing on an INVIDIA Tesla T4 with 16GB GDDR6 RAM

Method	Training time (50 epoch) Second	Test (time per image) Millisecond	Test frame per second
Base model	**2249**	0.35	2827
Proposed model 1	2275	**0.28**	**3571**
Proposed model 2	2282	0.30	3352

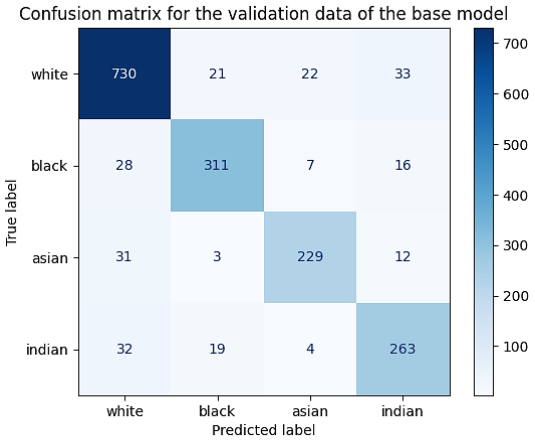
Table 4 provides a detailed comparison of model performance metrics across different racial categories, revealing significant insights into the effectiveness of the proposed methodologies. The base model without balance and attention modules shows robust accuracy, particularly in the White and Black categories, with notable recall for White (0.94). However, it struggles with the Indian category, exhibiting lower accuracy (0.85) and precision (0.68). Proposed Method 1, which employs balanced data, demonstrates improvements in precision for the White category (0.92) but experiences a decline in Indian category performance, achieving 0.74 accuracy. In contrast, Proposed Method 2, which integrates both balanced data and an attention module, achieves the highest overall metrics, particularly excelling in the Asian category (0.90 accuracy) and significantly improving Indian classification metrics to 0.86 accuracy and 0.76 precision. This comprehensive analysis underscores the importance of data balancing and advanced modeling techniques in enhancing classification accuracy, particularly for underrepresented groups, thereby demonstrating the proposed methods' superior capability in addressing the challenges faced by the base model.

Table 4: Detailed performance metrics for RC models

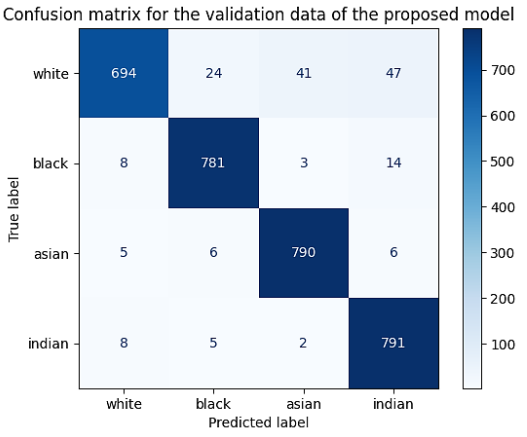
Method	Race	Criterion				
		Accuracy	Precision	Recall	F1-Score	Support-test
Base model without balance and attention module	White	0.86	0.84	**0.94**	0.89	2016
	Black		0.88	0.83	0.85	905
	Asian		0.88	0.84	0.86	687
	Indian		0.85	0.68	0.76	795
Proposed method_1 with balance data (Efficient-Net-BD)	White	0.87	**0.92**	0.88	0.90	2016
	Black		**0.88**	0.85	0.87	905
	Asian		0.88	0.87	0.87	687
	Indian		0.74	**0.85**	0.79	795
Proposed method_2 with balance data and attention module (Efficient-Net-BD-AM)	White	**0.88**	0.89	0.93	**0.91**	2016
	Black		0.87	**0.88**	**0.87**	905
	Asian		**0.90**	**0.89**	**0.90**	687
	Indian		**0.86**	0.76	**0.81**	795

Fig. 5 presents the confusion matrices for the validation dataset, featuring two distinct matrices that demonstrate the advantages of using a balanced dataset as well as the strengths of the proposed model in developing effective classification models. Fig. 5(a) illustrates the performance of the base model, while Fig. 5(b) pertains to proposed model 2. In Fig. 5(a), the unbalanced nature of the dataset is prominently displayed, highlighting its impact on classification accuracy across various racial categories. This matrix clearly shows the detrimental effect of imbalance, particularly in classes with fewer samples. In contrast, Fig. 5(b) showcases the application of a balanced dataset, emphasizing the strengths of the proposed method. For instance, the number of samples in the Indian class has increased from 318 to 806. This increase leads to a significant reduction in misclassifications, with errors decreasing from 55 to 15 during the validation phase. Moreover, similar trends can be observed across other classes, indicating a systematic improvement in classification performance. The results underscore the importance of data representation in training processes, suggesting that such enhancements can lead to a more robust and reliable model capable of better generalization to unseen data.

Fig. 6 presents the confusion matrices for the test dataset, consisting of two separate matrices. Since the test data were partitioned at the beginning of the process, these data were kept untouched for the testing phase in this research. Fig. 6(a) illustrates the performance of the base model, while Fig. 6(b) corresponds to proposed model 2. In Fig. 6(a), the performance of the base model on the test data is observed. In contrast, Fig. 6(b) shows the confusion matrix for the test data using proposed model 2. In these matrices, it can be seen that in matrix Fig. 6(a), the base method performed better in the white class compared to the proposed model. However, in this study, the classification of the white race has less challenge for the models, given that the database for the white class, as shown in Fig. 3, has greater diversity and accounts for 46% of the data. Therefore, improvements are needed in other classes. In Fig. 6(b), it is observed that the black class has 20 fewer error samples compared to the base model, the Asian class shows an improvement of 30 samples, and finally, the Indian class improved by 72 samples. Thus, the strength of proposed model 2 is demonstrated in other classes with fewer sample sizes. As a result, this model can be confidently utilized for RC applications.



(a)



(b)

Fig. 5: Confusion matrices for the validation dataset, illustrating the performance of the base model (a) and the proposed model 2 (b).

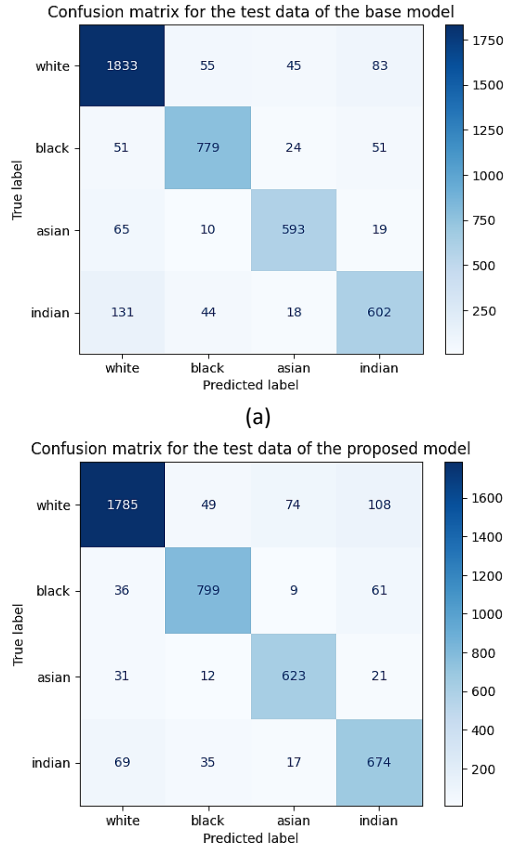


Fig. 6: Confusion matrices for the test samples, illustrating the performance of the base model (a) and the proposed model 2 (b).

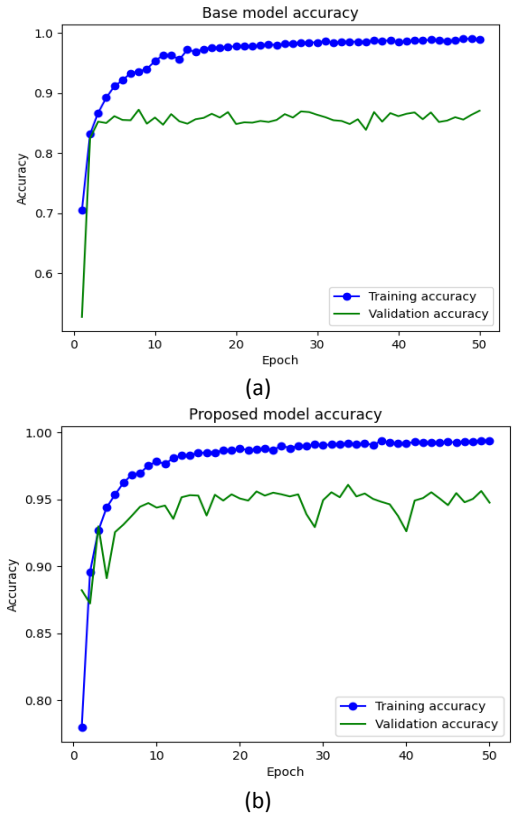


Fig. 7: Epoch-wise accuracy progression for RC model. (a): base model and (b) proposed model 2.

Fig. 7 illustrates the accuracy progression over training epochs for both the base model and proposed model 2. The x-axis represents the number of epochs, while the y-axis shows the model's accuracy across RC categories. In Fig. 7(a), which corresponds to the base model, the accuracy stabilizes at around 87% by the end of the training process. In contrast, Fig. 7(b) presents the performance of proposed model 2, where the accuracy reaches a significantly higher value of 95%. Additionally, it is evident that proposed model 2 not only achieves a better final accuracy but also starts with a higher accuracy at the beginning of the training process compared to the base model. This demonstrates the superior learning capability and faster adaptation of the proposed model in comparison to the base model.

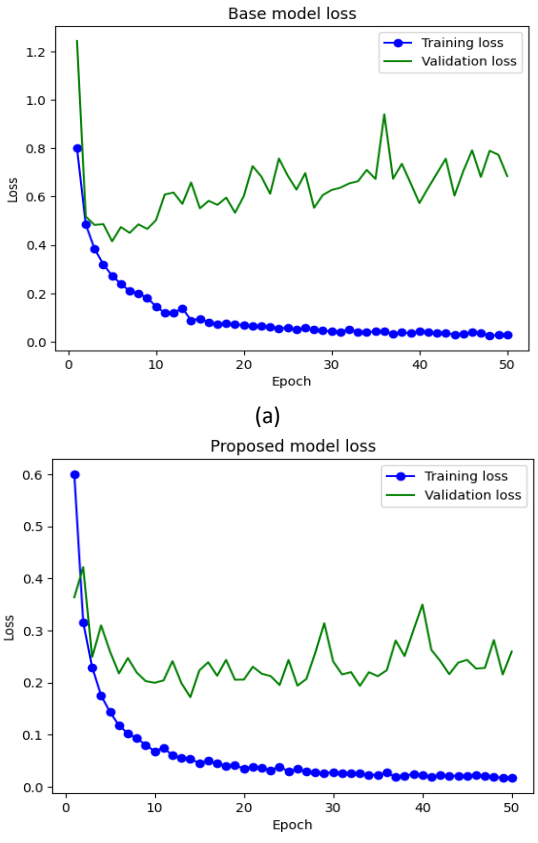


Fig. 8: Epoch-wise Loss trend for RC model. (a): base model and (b) proposed model 2.

Fig. 8 presents the loss progression over the training epochs for both the base model and proposed model 2, offering a detailed comparison of their performance. The x-axis represents the number of epochs, while the y-axis reflects the corresponding loss values during both the training and validation phases. In Fig. 8(a), which represents the base model, the initial loss is nearly double that of proposed model 2, as shown in Fig. 8(b). This disparity suggests that the base model faces greater difficulty in learning at the start of training. As training proceeds, the final validation loss for the base model

remains approximately three times higher than that of proposed model 2, underscoring a significant difference in convergence between the models. Moreover, the base model demonstrates increasing loss during later stages, indicating instability and a lack of convergence. In stark contrast, proposed model 2 exhibits a much more stable and lower loss curve, reflecting a more efficient learning process. Furthermore, the base model's loss curve shows greater variability, with frequent fluctuations throughout the training process, signaling potential issues with optimization. Conversely, proposed model 2 maintains a smoother and more consistent loss trajectory, indicating better control and robustness in learning. This clear improvement in stability and overall performance highlights the advantages of the proposed model in handling the classification task effectively. In Fig. 9 the plot illustrates the precision, recall, and F1-score for each class based on the model's performance. Each point represents the respective metric's score for the individual racial classes (White, Black, Asian, and Indian). Additionally, the dashed line indicates the overall accuracy of the model, providing a reference point to assess how well the model performs across different classes relative to its general performance. Fig. 9 illustrates the performance of the model using both macro and weighted averages. The macro-average reflects the model's performance equally across all classes, without considering the size of each class, providing insight into how the model performs on average for each class, regardless of the number of samples.

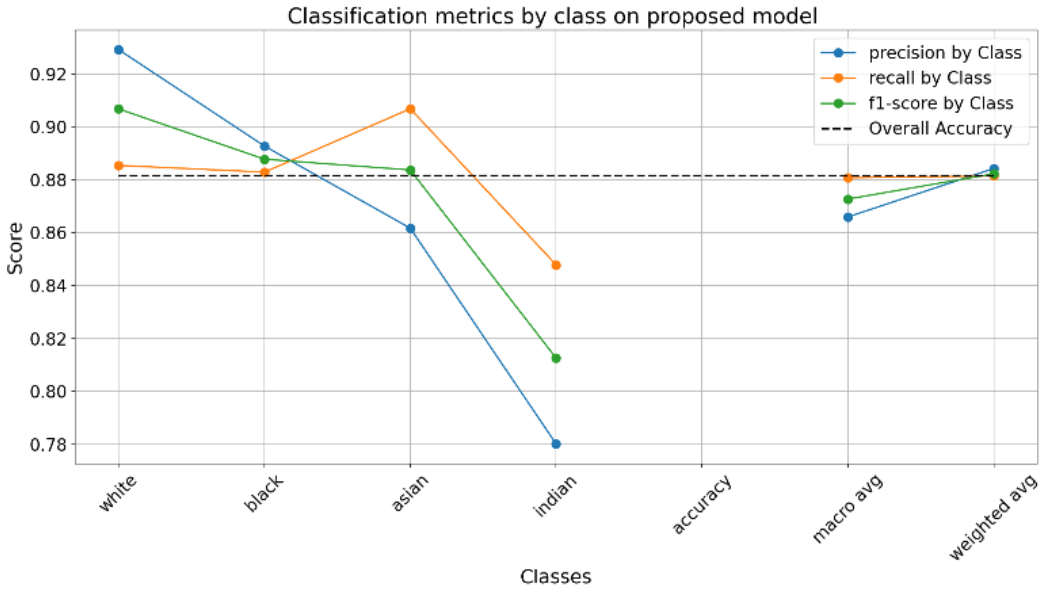


Fig. 9: Illustration of precision, recall, and F1-score for each class based on the proposed model 2 performance.

This allows for a balanced view of performance across minority and majority classes. In contrast, the weighted average evaluates the model's performance based on the sample size of each class, assigning greater importance to larger classes. This metric offers a more comprehensive understanding of the model's overall effectiveness, particularly in handling the class imbalance present in the dataset, by reflecting the influence of uneven class distributions on the final outcomes.

Table 5 reflects the continuous advancements in RC methodologies, particularly in the context of challenging and real-world conditions, as analyzed through the UTK-Face dataset. Al-Azani and El-Alfy (2019) laid the foundation by using Histogram of Oriented Gradients (HOG) features, achieving an accuracy of 69.68%. Despite the utility of HOG for handling variations in lighting and background, this approach showed limitations due to its sensitivity to scale and angle changes, leading to performance degradation in noisy environments.

Following this, Hamdi and Moussaoui (2020) demonstrated the superiority of deep learning methods over traditional machine learning approaches, achieving a higher accuracy of 78.88%. Similarly, Krishnan et al. (2020) highlighted performance disparities across different gender-race groups, reaching an accuracy of 79.49% but revealing the necessity of more equitable and robust models. Ahmed et al. (2022) advanced the field with deep networks, optimizing the use of diverse data combinations to achieve a notable accuracy of 77.50%. Meanwhile, Belcar et al. (2022) refined race recognition by concentrating on specific facial regions like the middle part of the face, reaching an accuracy of 80.34%. However, this approach's reliance on specific regions introduced vulnerabilities when those parts were occluded or affected. Deviyani (2022) marked a significant leap in accuracy, achieving 87.20% by employing StarGAN, and analyzing multiple datasets comprehensively. This method's versatility made it one of

the most thorough analyses in the domain. Patel et al. (2023) and Wirayuda et al. (2023) added further refinements with accuracies of 76.22% and 82.19%, respectively, employing shift-invariant architectures and compact-fusion frameworks, though these methods faced challenges in terms of complexity and real-time applicability.

Table 5: Numerical results on the UTK-Face dataset

Method	Accuracy (%)
Al-Azani and El-Alfy (2019) [16]	69.68
Hamdi and Moussaoui (2020) [18]	78.88
Krishnan et al. (2020) [19]	79.49
Ahmed et al. (2022) [20]	77.50
Belcar et al. (2022) [21]	80.34
Deviyani (2022) [30]	87.20
Patel et al. (2023) [23]	76.22
Wirayuda et al. (2023) [24]	82.19
Base model	86.46
Proposed model1	86.82
Proposed model2	**88.19**

In comparison, our base model shows substantial improvement, achieving 86.46% accuracy. The proposed model 1 slightly surpasses this with 86.82%, while proposed model 2 demonstrates the highest accuracy at 88.19%, outperforming all prior models. Notably, proposed model 2 leverages advanced optimization techniques and superior feature extraction, making it highly efficient under diverse and challenging conditions. Additionally, it maintains low computational costs, addressing the gap highlighted in previous studies regarding real-time performance and resource efficiency. This underlines the robustness and generalizability of the proposed model for race recognition tasks across different demographic groups.

Fig. 10 showcases a selection of correctly classified test samples, highlighting the model's capability to accurately predict various racial categories across a diverse set of facial images. The displayed images, along with their corresponding true and predicted labels, demonstrate the strong alignment between the model's predictions and actual labels. Notably, the proposed model exhibits robust performance across a wide range of racial categories and age groups, effectively handling the inherent diversity in both age and race within the dataset. This indicates the model's adaptability and precision in classifying facial images under varying demographic conditions, further validating its strength in real-world applications.

Fig. 11 presents a facial sample alongside its incorrect prediction in comparison to the actual label. The inaccuracies in test predictions can be attributed to three main factors: low image quality, which hampers the model's ability to distinguish subtle facial features; labeling errors, which result in incorrect associations between images and their racial labels—an example being a White male misclassified as Black. Additionally, the similarity among racial classes poses a challenge, as overlapping features can complicate differentiation. Another potential source of error arises from the data augmentation methods employed during the balancing process; if erroneous data exists in the dataset, it can propagate errors further. Addressing these issues through improved image quality, accurate labeling, and enhanced model sensitivity is crucial for bolstering prediction accuracy.

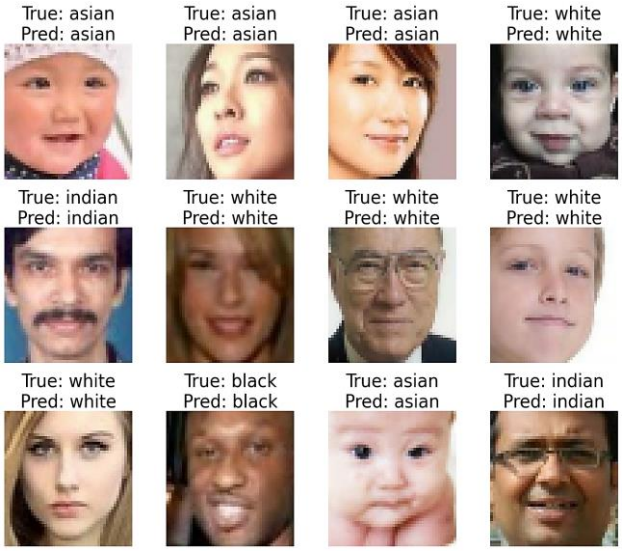


Fig. 10: Examples of true predicted test samples.

Fig. 11: Examples of incorrectly predicted test samples.

Conclusion

This study presents an innovative solution to the race classification (RC) problem by utilizing the Efficient-Net model in conjunction with transfer learning techniques to enhance performance. A key strength of this approach lies in the preprocessing of input images, achieved through the Multi-task Cascaded Convolutional Network (MTCNN) for accurate face detection and alignment. This preprocessing not only isolates facial features but also minimizes background noise, thereby establishing a robust foundation for effective classification. Addressing the issue of data imbalance was another crucial aspect of

our methodology. We implemented sophisticated techniques such as data augmentation and oversampling to generate a diverse set of training samples. Data augmentation involved applying transformations like rotation and scaling, which aided in enriching the training dataset. Oversampling was employed to mitigate class imbalance, particularly for racial categories with fewer training samples. This focus on enhancing dataset quality was effective in improving the model's generalization across different racial groups. Additionally, one of the key advantages of proposed model 2 is its ability to significantly reduce error rates compared to the base model in classes with limited data, positively impacting accuracy in these categories. The use of Efficient-Net, recognized for its optimal balance between accuracy and computational efficiency, has been specifically tailored for RC tasks. This adjustment enables the model to effectively capture subtle variations in facial features among different racial groups, thereby contributing to improved accuracy. Evaluation demonstrated that this model exhibits remarkable real-time performance. Assessment metrics, including accuracy, precision, recall, and F1 score, indicated overall high performance, although some variability was observed among racial categories. This variability highlights ongoing challenges related to classification accuracy, particularly in classes with limited training data. Factors such as low image quality, erroneous labeling, and similarities among specific racial features have been identified as contributors to classification errors. Future efforts should focus on enhancing data quality, correcting labeling inaccuracies, and refining the model to address these issues.

Author Contributions

M. Rohani designed the experiments and developed the overall methodology, wrote manuscript, conducted the data analysis, statistical evaluations and collected and preprocessed the dataset. H. Farsi interpreted the results and has drawn the general road map. S. Mohamadzadeh edited and revised the manuscript.

Acknowledgment

The authors wish to express their profound gratitude to the esteemed reviewers and editors of JECEI for their meticulous review, constructive feedback, and invaluable suggestions, which have significantly enhanced the quality of this article. Furthermore, the authors extend their sincere appreciation to the editorial board for their professional guidance and exemplary handling of the manuscript during the review process.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

AI	Artificial Intelligence
DL	Deep Learning
RC	Race Classification
AR	Augmented Reality
VR	Virtual Reality
CNN	Convolutional Neural Network
MTCNN	Multi-task Cascaded Convolutional Networks
SMOTE	Synthetic Minority Over-sampling Technique
HOG	Histogram of Oriented Gradients
CA	Channel attention
SA	Spatial attention
FM	Feature map
AvgPool	Average pooling
MaxPool	Maximum pooling
CBAM	Convolutional block attention module

References

[1] E. Ghasemi Bideskan, S. M. Razavi, S. Mohamadzadeh, M. Taghippour, "Facial expression recognition through optimal filter design using a metaheuristic kidney algorithm," J. Electr. Comput. Eng. Innovations (JECEI), 12(2): 425-438, 2024.

[2] M. Rohani, H. Farsi, S. Mohamadzadeh, "Deep multi-task convolutional neural networks for efficient classification of face attributes," Int. J. Eng., 36(11): 2102-2111, 2023.

[3] A. Nieves Delgado, "Race and statistics in facial recognition: Producing types, physical attributes, and genealogies," Social Stud. Sci., 53(6): 916-937, 2023.

[4] M. Rohani, H. Farsi, S. Mohamadzadeh, "Facial feature recognition with multi-task learning and attention-based enhancements," Iran. J. Energy Environ., 16(1): 136-144, 2025.

[5] D. M. Hilty, A. M. P. Schmid, R. E. Holbrook, J. P. Greer, "A review of telepresence, virtual reality, and augmented reality applied to clinical care," J. Technol. Behav. Sci., 5(1): 178-205, 2020.

[6] C. Lu, R. Ahmed, A. Lamri, S. S. Anand, "Use of race, ethnicity, and ancestry data in health research," PLOS Global Public Health, 2(9): 1060-1076, 2022.

[7] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, "Biometrics recognition using deep learning: A survey," Artif. Intell. Rev., 56(8): 8647-8695, 2023.

[8] I. Adjabi, A. Ouahabi, A. Benzaoui, A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," Electronics, 9(8): 1188-1202, 2020.

[9] K. Weiss, T. M. Khoshgoftaar, D. Wang, "A survey of transfer learning," J. Big Data, 3(1): 1-40, 2016.

[10] S. Zahiri, R. Iranpoor, N. Mehrshad, "Paying attention to the features extracted from the image to person re-identification," J. Electr. Comput. Eng. Innovations (JECEI), 13(2): 267-274, 2025.

[11] Z. Ghasemi-Naraghi, A. Nickabadi, R. Safabakhsh, "Multi-Task learning using uncertainty for realtime multi-person pose estimation," J. Electr. Compu. Eng. Innovations (JECEI), 12(1): 147-162, 2024.

[12] M. Rohani, H. Farsi, S. H. Zahiri, "Statistical analysis and comparison of the performance of meta-heuristic methods based on their powerfulness and effectiveness," J. Inf. Syst. Telecommun. (JIST), 10(37): 49-59, 2022.

[13] M. Wang, W. Deng, "Deep face recognition: A survey," Neurocomputing, 429(1): 215-244, 2021.

[14] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, "Biometrics recognition using deep learning: A survey," Artif. Intell. Rev., 56(8): 8647-8695, 2023.

[15] M. J. A. Dujaili, "Survey on facial expressions recognition: databases, features and classification schemes," Multimedia Tools Appl., 83(3): 7457-7478, 2024.

[16] S. Al-Azani, E. S. El-Alfy, "Ethnicity recognition under difficult scenarios using HOG," J. Electr. Eng. Autom., 1(1): 1-10, 2019.

[17] M. Ruhani, H. Farsi, S. Mohamadzadeh, "Object tracking in video with correlation filter and using histogram of gradient feature," J. Soft Comput. Inf. Technol., 9(4): 43-55, 2020.

[18] S. Hamdi, A. Moussaoui, "Comparative study between machine and deep learning methods for age, gender, and ethnicity identification," in Proc. 2020 4th International Symposium on Informatics and its Applications (ISIA):1-6, 2020.

[19] A. Krishnan, A. Almadan, A. Rattani, "Understanding fairness of gender classification algorithms across gender-race groups," in Proc. 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA): 1028-1035, 2020.

[20] M. A. Ahmed, R. D. Choudhury, K. Kashyap, "Race estimation with deep networks," J. King Saud Univ. Comput. Inf. Sci., 34(7): 4579-4591, 2022.

[21] D. Belcar, P. Grd, I. Tomičić, "Automatic ethnicity classification from middle part of the face using convolutional neural networks," Informatics, 9(1): 18-32, 2022.

[22] S. Li, W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affective Comput., 13(3): 1195-1215, 2020.

[23] S. Patel, V. Srivastava, A. Bajpai, "Three fold classification using shift invariant deep neural network," in Proc. 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS): 787-791, 2023.

[24] T. A. B. Wirayuda, R. Munir, A. I. Kistijantoro, "Compact-fusion feature framework for ethnicity classification," Informatics, 10(2): 51-84, 2023.

[25] M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. International Conference on Machine Learning: 6105-6114, 2019.

[26] X. Li, Z. Yang, H. Wu, "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks," IEEE Access, 8: 174922-174930, 2020.

[27] S. Woo, J. Park, J. Y. Lee, I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. European Conference on Computer Vision (ECCV): 3-19, 2018.

[28] R. Yacouby, D. Axman, "Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models," in Proc. First Workshop on Evaluation and Comparison of NLP Systems: 79-91, 2020.

[29] Z. Zhang, Y. Song, H. Qi, "Age progression/regression by conditional adversarial autoencoder," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 5810-5818, 2017.

[30] A. Deviyani, "Assessing dataset bias in computer vision," arXiv preprint arXiv: 2205.01811, 2022.

Biographies

Mehrdad Rohani received his M.Sc. degree in Telecommunications Engineering from Birjand University, Birjand, Iran, in 2018. He is currently pursuing a Ph.D. degree in Electrical Engineering with a focus on Telecommunications at Birjand University. His research interests encompass Machine Learning, Image Processing, Computer Vision, and Deep Learning Algorithms.

- Email: m.ruhani@birjand.ac.ir
- ORCID: 0000-0003-2930-019X
- Web of Science Researcher ID: LQJ-4143-2024
- Scopus Author ID: NA
- Homepage: NA

Hasan Farsi received his B.Sc. and M.Sc. degrees in Communication Engineering from Sharif University of Technology, in 1992 and 1994, and his Ph.D. in Communication Engineering from the University of Surrey, UK, in 2003. He is currently a Professor in the Department of Electrical and Computer Engineering at the University of Birjand. His research interests include deep learning, image processing, signal processing.

- Email: hfarsi@birjand.ac.ir
- ORCID: 0000-0001-6038-9757
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/hasanfarsi/fa

Sajad Mohamadzadeh received his B.Sc. degree in Communication Engineering from the University of Sistan and Baluchestan, Iran, in 2010, and his M.Sc. and Ph.D. degrees in Communication Engineering from the University of Birjand, Iran, in 2012 and 2016, respectively. He is currently an Associate Professor in the Department of Electrical and Computer Engineering at the University of Birjand. His research interests include image processing, deep neural networks and deep learning.

- Email: s.mohamadzadeh@birjand.ac.ir
- ORCID: 0000-0002-9096-8626
- Web of Science Researcher ID: NA
- Scopus Author ID: 57056477500
- Homepage: https://cv.birjand.ac.ir/mohamadzadeh/

How to cite this paper:
M. Rohani, H. Farsi, S. Mohamadzadeh, "Advanced race classification using transfer learning and attention: real-time metrics, error analysis, and visualization in a lightweight deep learning," J. Electr. Comput. Eng. Innovations, 13(2): 341-352, 2025.

DOI: 10.22061/jecei.2025.11318.793

URL: https://jecei.sru.ac.ir/article_2258.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Nonlinear Filter-Based Estimation of Wheel-Rail Contact Forces and Related Considerations using Inertial Measurement Unit

M. Moradi, R. Havangi *

Faculty of Electrica l Engineering and Computer, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 08 September 2024
Reviewed 28 November 2024
Revised 11 January 2025
Accepted 15 January 2025

Keywords:
Adhesion coefficient traction system
Wheel slip
Nonlinear filter
Vehicle dynamics

Corresponding Author's Email Address:
Havangi@Birjand.ac.ir

Abstract

Background and Objectives: Rail vehicle dynamics are significantly influenced by the forces at the wheel-rail contact interface, particularly the wheel-rail adhesion force, which is critical for effective braking and acceleration. Continuous monitoring of this force is essential to prevent infrastructure damage and enhance transportation efficiency. Given the challenges of directly measuring adhesion force, alternative methods using state observers have gained prominence. The choice of model and estimator efficacy are vital for accurate variable estimation.
Methods: In this study, the dynamics of the wheelset is simulated in the presence of irregularities that can be encountered in the railroad. Estimation of wheel-rail adhesion force is done indirectly by nonlinear filters as estimators and their accuracies in the estimation are compared to identify the better one. Meanwhile, inertial sensors (accelerometer and gyroscope) outputs are used as measuring matrix and employed to simulate actual situation and evaluate the estimators' performances. The proposed approach is implemented in MATLAB to assess the accuracy and effectiveness of these estimators in determining states and variables.
Results: The proposed method effectively utilizes longitudinal, lateral, and torsional dynamics to estimate wheel-rail adhesion force across varying conditions. Experimental results demonstrate high precision, rapid convergence, and low error rates in the estimations.
Conclusion: In this study, the identification of the wheel and rail contact conditions is carried out by analyzing the dynamic characteristics of the railway wheelset. The results of proposed method can lead to decreasing wheel deterioration and operational costs, minimizing high creep levels, maximizing the use of already-existing adhesion, and improving the frequency of service. It is worth noting that the proposed method is beneficial for both conventional railway transport and automated driverless trains.

Introduction

Modeling wheel-rail interactions is inherently complex, particularly when diverse track conditions are considered. As model complexity increases, the computational load also rises, leading to extended response times. Consequently, to maintain computational efficiency, it is essential to focus on the most significant and influential components that impact the wheel-rail forces. The tangential forces at the wheel-rail interface are the key components in the wheel-and-rail interface and are caused by wheel and rail relative motion. In fact, the available evidence suggests that the motion is defined by a gradual sliding phenomenon at the contact surface known as creepage. The forces caused by creepage are denoted as creep forces, and they control how well a rail vehicle accelerates and brakes. Adhesion is determined as the ratio of the tangential frictional force between the wheel and rail to the load of the wheel. The friction

coefficient is defined as the ratio between the friction force and the normal force at the contact surface [1]. The friction coefficient always limits the adhesion coefficient [2]. As a result, there may be differences between the adhesion and friction coefficients.

In [3] the importance of friction in determining wheel and rail adhesion is discussed. Estimating adhesion in the wheel and rail contact region is a complex procedure because it depends on a number of operational variables, including the operational mechanism of rail self-cleaning, axle load distributions, track irregularities, and processes occur at the wheel and rail nonlinear contact interface. Effective and continuous monitoring of the adhesion coefficient is necessary for estimating the maximum adhesion force and preserving a satisfactory braking and acceleration performance, but measuring the adhesion coefficient with a conventional physical sensor is difficult [4]. The adhesion coefficient is highly dependent on any materials that are present at the wheel and rail interface, including water [5], leaves [6], [7], snow, oil, and grease. The wheel and rail adhesion characteristics have different behaviour under large sliding conditions and as the slip ratio rises, the adhesion coefficient keeps rising after it reaches the saturation point instead of decreasing [8]. Numerous researchers attempted to solve the adhesion problem, and various approaches, including statistical, genetic, and mathematical control theory, were put forth and applied [9], [10]. Two key elements influencing the railway surfaces are train velocity and contact area temperature [11]. The maximum adhesion coefficient is reached at higher values of both the adhesion coefficient and the slip velocity.

As such, determining the adhesion level is a crucial task for a rail vehicle to operate properly. In [12], a novel method for figuring out the adhesion coefficient between the wheel and rail was presented. Furthermore, a different adhesion control method based on tracking the adhesion status between the wheel and rail is presented in another research paper [13]. Traction power in trains can be efficiently utilized when optimal adhesion control is achieved [14], [15]. It is noteworthy to mention that in order to prevent wheel slippage or slide, the creep velocity of the train within the stable region must be limited in accordance with the changes seen in the adhesion coefficient characteristic curve. In [16] re-adhesion control is used to bring the trains back to the stable region by quickly identifying instances of wheel slide and adjusting the torque precisely. The correct selection of the initial model by considering the most important factors related to adhesion force and the selection of an estimator that is compatible with the structure of the system under study can create a more reliable and accurate output. Due to the nonlinear nature of the adhesion coefficient, the use of nonlinear types of Kalman filters has been of great interest. An innovative method that estimates the wheel and rail states using the Kalman-Bucy filter (KBF) approach is suggested in [17] to predict the wheel and rail wear, regions of adhesion variations or low adhesion, and the development of rolling contact fatigue. Additionally, the lateral creep force is detected for the purpose of determining the local adhesion condition using the KBF [18]. In [19], a model-based approach utilizing Extended Kalman Filter (EKF) is presented to estimate the adhesion force in the wheel and rail contact surface. But the strategy is not evaluated on every track circumstance. In [20], an EKF based estimation method was proposed for estimating the slip, creep force, and friction coefficient between the wheel and rail surface using the induction motor current, stator voltage, and speed. Using multi-rate EKF state identification is an alternate method for detecting slip velocity. This method determines the traction motor load torque accurately by combining the EKF method with the multi-rate technique. Faster slip detection, enhanced dependability, and better traction performance are the benefits of this approach [21].

The adhesion coefficient was found as a function of slip velocity in [22]. To estimate the slip velocity, the measured wheel velocity was fed into the EKF. In order to attain the best outcome, various EKF configurations were examined and adjusted in this study using system and measurement noise covariance matrices. Real-time wheel-rail contact force and moment estimation based on an EKF estimator under typical driving circumstances is represented by the researchers in [23]. EKF uses a Jacobian matrix in states estimation which is an error-prone process [24]. To overcome these problems, an unscented transformation proposed in [25]. A model-based approach using Unscented Kalman Filter (UKF) is proposed in [26] for estimation of friction coefficient, creep force, and creepage. The UKF faces challenges with numerical stability when applied to high-dimensional systems due to the fact that the central stem (mean) of the sigma points carries a heavier weight, often negative, in such systems. Nevertheless, estimators appear to be unreliable in certain crucial track conditions, so more work is required to more effectively monitor these wheel-rail parameters in real time. Meanwhile, after reviewing the literature on railway wheelset dynamics condition monitoring, it is found that more effort and improvement need to be done to solve the issue of analyzing wheelset conditions and updating them to the desired situation in order to meet the global transportation vehicle expectations of extremely fast, high comfort, increased safety, and cost-effectiveness.

For optimal operation, the Kalman filter requires a system model. Also, mathematical models of important processes are necessary for the methodical process of

adhesion estimation. In addition to the mentioned methods, adhesion condition was estimated [27] and wheel-rail contact force was predicted [28] using the artificial neural network approach. These methods, however, ignore a number of important factors, including changes in the wheel and rail profiles and friction levels. Use of the traditional wheel-rail contact algorithms [29], which are employed in multi-body software packages and yield accurate results, is preferable in this situation. Nevertheless, the low computational speed of the classical contact models makes them unsuitable for real-time implementation. The fast approximation model seems to offer satisfactory precision in order to achieve real-time simulation by fulfilling the criteria specified in the literature [30]. However, it is not capable of taking contact profile changes into account and requires user-defined coefficients in models.

The aim of this research is to assess the adhesion force and slip in the contact region of the wheel and rail with accuracy by using nonlinear filters approach. It is worth noting that in this process EKF and UKF are employed for estimation as nonlinear filters. The main goal of this experiment is to determine, whether employing UKF in the system allows to achieve better results in respect to the EKF. Analysis of the measured inertial sensors values is used in estimation process. To evaluate the observer's performance, a dynamic model is constructed, comprising lateral, longitudinal, and yaw dynamics of the wheelset. In summary, the manuscript introduces a novel, comprehensive approach by employing nonlinear filtering that integrates Inertial Measurement Units (IMUs) data to estimate wheel-rail contact forces in real-time. The Polach model is utilized to explain the wheel-rail contact conditions. The rest of this research is organized into four parts. First, the details of lateral, longitudinal and yaw dynamical model of the wheelset are explained. Then, the process of estimator design is outlined. This is followed by an in-depth discussion of the experimental results. Finally, the conclusion is presented.

Lateral, Longitudinal and Yaw Dynamical Model of the Wheelset

If the rail is considered to be rigid, the wheelset has three degrees of freedom namely longitudinal, lateral, and yaw motions. Compared to longitudinal displacement, yaw and lateral displacement are very small but they have the key role in stability and ride comfort of the vehicle. The interaction between the wheel and rail contact area influences the dynamic performance of the rail vehicle. In order to design dynamic control systems and monitor the situation more efficiently, it is very important to know the nature of the contact force. Also, to prevent the wheel from slipping during traction and sliding during braking, it is important to know the adhesion not only in normal operating conditions but also during traction and braking. Estimation of adhesion coefficient, slip ratio, and lateral dynamics of rail vehicle in traction and braking modes are essential for travel safety and passenger comfort. Wheel-rail adhesion mechanism is shown in Fig. 1. A more complex model leads to an increase in computational load and an increase in response time. As a result, in order to maintain computational efficiency, it is necessary to focus on the most important and influential components affecting the wheel and rail forces in the estimation process. This approach allows for a more balanced trade-off between model accuracy and processing speed and facilitates practical real-time applications.

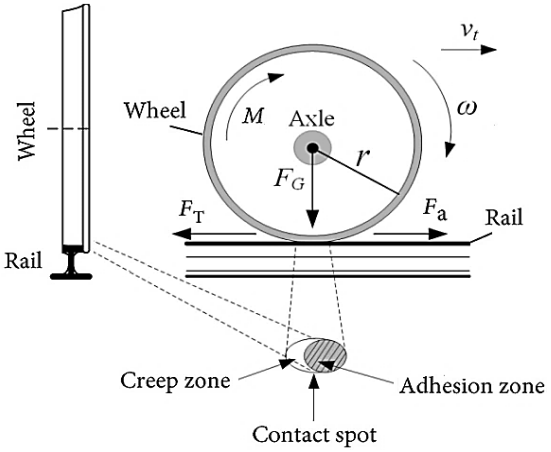


Fig. 1: Wheel-rail adhesion mechanism.

Estimating the dynamics of the wheelset is a complex process because the wheel and rail interface is an open loop system with variable external conditions. A novel model-based methodology has been devised in this study to estimate the most important dynamics of the wheelset in various contact conditions. Since the Kalman filter (KF) is not suitable estimator for the nonlinear contact system of wheel and rail, therefore, EKF is employed to estimate the adhesion coefficient, slip ratio, and lateral dynamics of the wheelset. The system utilized in this research is shown in Fig. 2, which consists of two wheels and an axle.

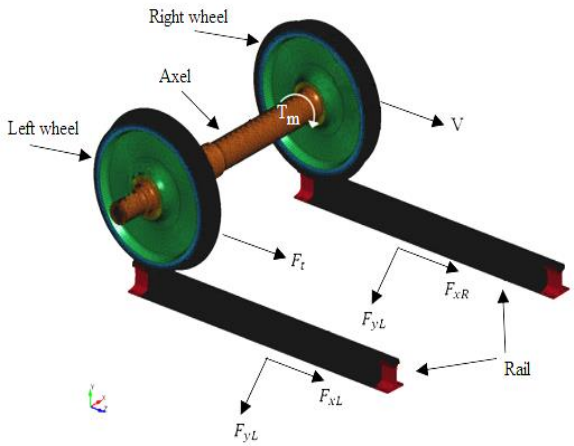


Fig. 2: Three-dimensional wheel-rail system model.

There is a direct correlation between the lateral and yaw dynamics and track irregularities. The left and right wheels' linear speeds will differ if the wheelset moves sideways from its initial position. These speeds are obtained from the following relations:

$$V_{Rw} = \omega_R.[r - \kappa_w(y - y_d)] \quad (1)$$

$$V_{Lw} = \omega_L.[r + \kappa_w(y - y_d)] \quad (2)$$

where V_{Rw} and V_{Lw} are longitudinal velocity of the right and left wheels, ω_R and ω_L are angular velocity of the right and left wheels, r is wheel radius, $\kappa_w$is wheel conicity, y is lateral movement, and y_d is track irregularities in lateral direction.

The wheelset's dynamic characteristics are also influenced by the creep forces that arise in the contact zones between the wheel and rail. These creep forces, which can be classified as longitudinal creepage (ξ_x) and lateral creepge (ξ_y) depending on the direction of movement, are brought on by the creeps that arise from the wheels' relative speed to the rail. The creepage of both wheels in the wheelset in the lateral and longitudinal directions are displayed in equations (3)-(5).

$$\xi_{Rx} = \frac{r\omega_R - V}{V} - [\frac{S\dot{\psi}}{V} + \frac{\kappa_w(y - y_d)}{r}] \quad (3)$$

$$\xi_{Lx} = \frac{r\omega_L - V}{V} + \frac{S\dot{\psi}}{V} + \frac{\kappa_w(y - y_d)}{r} \quad (4)$$

$$\xi_y = \xi_{Ly} = \xi_{Ry} = \frac{\dot{y}}{V} - \psi \quad (5)$$

where ξ_{Rx} and ξ_{Lx} are longitudinal creepage of the right and left wheels, ξ_{Ry} and ξ_{Ly} are lateral creepage of the right and left wheels, S is half gauge of track, $\dot{\psi}$ is yaw rate, $\dot{y}$ is lateral velocity, ψ is yaw angle, and V is train longitudinal velocity.

In equations (3) and (4), the expressions $\frac{r\omega_R - V}{V}$ and $\frac{r\omega_L - V}{V}$ do not include dynamics related to y and ψ, therefore can be ignored to simplify longitudinal creepage equations. In addition, the dynamics consist of lateral displacement and yaw rate are sufficient in identifying alterations in the wheel-rail contact conditions. The simplified longitudinal creepage equations are as follows:

$$\xi_{Rx} = -\frac{S\dot{\psi}}{V} - \frac{\kappa_w(y - y_d)}{r} \quad (6)$$

$$\xi_{Lx} = \frac{S\dot{\psi}}{V} + \frac{\kappa_w(y - y_d)}{r} \quad (7)$$

The total slip ξ_j is a combination of longitudinal ξ_{jx} and lateral ξ_{jy} slips and obtained from the following equation:

$$\xi_j = \sqrt{\xi_{jx}^2 + \xi_{jy}^2} \qquad j\text{=L or R} \quad (8)$$

The creep force F_j can be expressed as a nonlinear function of the slip, which is determined by utilizing the given equation in (9).

$$F_j = \mu_j F_{Nj} \qquad j\text{=L or R} \quad (9)$$

In normal conditions and for small amounts of creepage (microslip), F_j changes linearly with creepage. As the sliding speed increases, the creep force changes nonlinearly and reaches the maximum value (saturation), and if the increase in sliding speed continues, it begins a downward trend. In general, to describe wheelset stable and unstable behaviors, the adhesion-slip curves can be divided into three areas. The initial section displays a nearly linear pattern, followed by a nonlinear segment known as the high slip ratio region, and concluding with a negative slope indicating the unstable zone of the curve. Fig. 3 shows the details.

The nonlinear region is strongly influenced by factors such as pollution and weather and it results in large and uncertain changes in the creep force. The analysis of contact force distribution in both the longitudinal and lateral orientations was thoroughly studied in [31]. These forces can be calculated using equation (10).

$$F_{ji} = F_j \frac{\xi_{ji}}{\xi_j} \qquad j\text{=L or R} \quad \& \quad i\text{= } x \text{ or } y \quad (10)$$

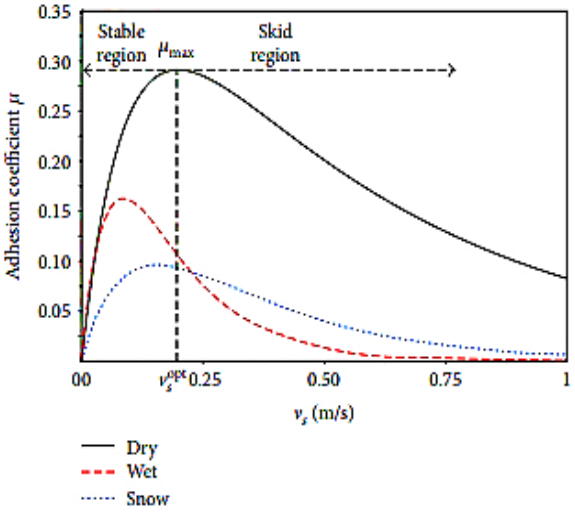


Fig. 3: Adhesion curves [9].

An entire wheelset model encompasses every aspect of the interactions between wheels and rails, facilitating the analysis of wheelset dynamics. The given equations below represent the motion of the wheelset at any location along the creep curve for yaw, rotational, torsional, lateral, and longitudinal dynamics.

$$\ddot{x} = \frac{F_{Rx} + F_{Lx}}{M_t} \quad (11)$$

$$\ddot{y} = \frac{-F_{Ry} - F_{Ly} + F_C}{m_w} \quad (12)$$

$$\ddot{\psi} = \frac{F_{Rx}S - F_{Lx}S - K_w\psi}{J_w} \quad (13)$$

$$T_s = t_s\theta_s + C_{vis}(\omega_R - \omega_L) \quad (14)$$

$$\theta_s = \int(\omega_R - \omega_L)dt \quad (15)$$

$$\dot{\omega}_L = \frac{T_s - T_L}{J_L} \quad (16)$$

$$\dot{\omega}_R = \frac{T_m - T_s - T_R}{J_R} \quad (17)$$

where F_{Rx} and F_{Lx} ara right and left wheel creep forces in longitudinal direction, M_t is rail vehicle mass, F_{Ry} and F_{Ly} ara right and left wheel creep forces in lateral direction, m_w is total weight of wheel with induction motor, F_c is centrifugal force, K_w is yaw stiffness, J_w is wheelset moment of inertia, T_s is torsional torque, t_s is torsional Stiffness of axle, θ_s is twist angle, C_{vis} is viscous material damping of the shaft, T_R and T_L are right and left wheel tractive torques, T_m is motor torque, and J_R and J_L are inertias of right and Left wheel.

In equation (12) F_C is considered when the wheels travel along a crooked railroad track. In equation (14) C_{vis} can be disregarded because it is typically very small. In the above equations, F_{xR}, F_{xL}, F_{yR}, F_{yL}, T_m, T_L, and T_R are defined as follows:

$$F_{Rx} = \frac{F_R}{\xi_R}\left[\left[\frac{r\omega_R - V}{V} - \left[\frac{S\dot{\psi}}{V} + \frac{\kappa_w(y - y_d)}{r}\right]\right]\right] \quad (18)$$

$$F_{Lx} = \frac{F_L}{\xi_L}\left[\frac{r\omega_L - V}{V} + \frac{S\dot{\psi}}{V} + \frac{\kappa_w(y - y_d)}{r}\right] \quad (19)$$

$$F_{Ry} = \frac{F_R}{\xi_R}\left[\frac{\dot{y}}{V} - \psi\right] \quad (20)$$

$$F_{Ly} = \frac{F_L}{\xi_L}\left[\frac{\dot{y}}{V} - \psi\right] \quad (21)$$

$$T_m = \mu M_t g r \quad (22)$$

$$T_L = r F_{Lx} \quad (23)$$

$$T_R = r F_{Rx} \quad (24)$$

As can be seen from the above equations, the dynamics of the wheelset are complex and all the movements of the wheelset are interdependent. Due to the powerful interactions exist between different wheel movements in lateral and longitudinal directions, it is of great importance to utilize a general model that encompasses all the motions associated with contact forces in the investigation of the wheelset dynamics. Wheels directly interact with the rail, as a result, any alterations in contact conditions will affect the wheelset dynamics.

This study uses a model-based methodology to estimate the variables related to wheel-rail contact. Model-based estimation utilizes the system's information through a mathematical framework and measured responses to the input to estimate the state variables of the system in real time. For practical purposes, the estimator design should be as simple as possible while taking into account the wheelset dynamics that are associated with the contact conditions. Therefore, the wheelset model [32] is simplified first.

$$\ddot{y} = -\frac{F_R}{m_w \xi_R}\left[\frac{\dot{y}}{V} - \psi\right] - \frac{F_L}{m_w \xi_L}\left[\frac{\dot{y}}{V} - \psi\right] \quad (25)$$

$$\ddot{\psi} = \frac{F_R}{J_w \xi_R}\left[-\frac{S\dot{\psi}}{V} - \frac{\kappa_w(y - y_t)}{r}\right]S - \frac{F_L}{J_w \xi_L}\left[\frac{S\dot{\psi}}{V} + \frac{\kappa_w(y - y_t)}{r}\right]S - \frac{K_w \psi}{J_w} \quad (26)$$

The simplified model has numerous benefits. The primary advantage lies in the straightforward estimator design with the fewest number of states, which enables the fast convergence of the estimator. In addition, in the simplified model, no input torque is required for the estimator, and yaw and lateral dynamics are affected by track disturbances. Since the relation between the adhesion coefficient μ and slip ξ is nonlinear, assuming that the wheels on both sides, i.e. the left and right, have the same contact conditions, first, equations (25) and (26) are arranged and the lateral and yaw dynamic models of the wheelset are derived.

$$\ddot{y} = -\frac{1}{V m_w}\left(\frac{F_R}{\xi_R} + \frac{F_L}{\xi_L}\right)\dot{y} + \frac{1}{m_w}\left(\frac{F_R}{\xi_R} + \frac{F_L}{\xi_L}\right)\psi \quad (27)$$

$$\ddot{\psi} = -\frac{S^2}{V J_w}\left(\frac{F_R}{\xi_R} + \frac{F_L}{\xi_L}\right)\dot{\psi} - \frac{S}{r J_w}\kappa_w\left(\frac{F_R}{\xi_R} + \frac{F_L}{\xi_L}\right)y + \frac{S}{r J_w}\kappa_w\left(\frac{F_R}{\xi_R} + \frac{F_L}{\xi_L}\right)y_t - \frac{K_w}{J_w}\psi \quad (28)$$

In the second step, by considering the equalities in (29) and replacing them in equations (27) and (28), equations (30) and (31) are obtained.

$$F_R = F_L = F_a \quad \& \quad \xi_L = \xi_R = \xi \quad (29)$$

$$\ddot{y} = -\frac{2F_a}{\xi m_w}\left(\frac{\dot{y}}{V} - \psi\right) \quad (30)$$

$$\ddot{\psi} = -\frac{2F_a S^2}{\xi V J_w}\dot{\psi} - \frac{2F_a S}{\xi r J_w}\kappa_w(y - y_t) - \frac{K_w}{J_w}\psi \quad (31)$$

Other process variables are defined as follows:

$$\xi = \sqrt{\left(\frac{\kappa_w(y - y_d)}{r} + \frac{S\dot{\psi}}{V}\right)^2 + \left(\frac{\dot{y}}{V} - \psi\right)^2} \quad (32)$$

$$\mu = \mu_0((1 - D)e^{-B\xi V} + D) \quad (33)$$

$$F_a = \frac{2F_N\mu}{\pi}\left(\frac{k_A\varepsilon}{1 + (k_A\varepsilon)^2} + \arctan(k_S\varepsilon)\right) \quad (34)$$

$$\varepsilon = \frac{2\pi a^2 bc}{3F_N\mu_{k-1}}\xi_{k-1} \quad (35)$$

where F_N is the normal force, D and B are reduction factors associated with distinct friction coefficients, G is shear module, a and b are the semi-axis length of the ellipse in contact zone, and C_{11} is the Kalker coefficient.

In the following, the design of the estimator is discussed.

Nonlinear Filter-Based Estimation of Wheel-Rail Contact Forces

The details of the filters used for estimation of wheel-rail lateral dynamics can be found in the following subsections. The discrete-time nonlinear model is presented as follows:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k) + w_k \\ z_k &= h(x_k) + v_k \end{aligned} \quad (36)$$

where $f(.)$ represents the dynamics of wheelset, h(.) is the relationship between the observation z_k and the state

vector x_k, $u_k$refers to the input vector, while w_k and v_k represent the vectors of noise that affect the process and measurement respectively. The state variables used to create the EKF algorithm process matrix include lateral velocity ($\dot{y}$), yaw rate ($\dot{\psi}$), slip ratio (ξ), friction coefficient (μ), and adhesion force (F_a). Besides, lateral acceleration ($\ddot{y}$) and yaw rate are considered to create the measurement matrix.

$$x = \begin{bmatrix} \dot{y} & \dot{\psi} & \xi & \mu & F_a \end{bmatrix}^T$$
$$z = [\ddot{y} \quad \dot{\psi}] \tag{37}$$

To design nonlinear filter, the model used for estimation must also be discrete. Therefore, equations (30)-(34) should be discretized, the result of which is given below.

$$\dot{y}_k = \dot{y}_{k-1} - \frac{2\tau F_{ak-1}}{\xi_{k-1} m_w}\left[\frac{\dot{y}_{k-1}}{V} - \psi\right] \tag{38}$$

$$\dot{\psi}_k = \dot{\psi}_{k-1} - \frac{2\tau S F_{ak-1}}{\xi_{k-1} J_w}\left[\frac{\kappa_w (y - y_d)}{r} + \frac{S\dot{\psi}_{k-1}}{V}\right] - \frac{K_w \psi}{J_w} \tag{39}$$

$$\xi_k = \sqrt{\left(\frac{\kappa(y-y_d)}{r} + \frac{S\dot{\psi}_{k-1}}{V}\right)^2 + \left(\frac{\dot{y}_{k-1}}{V} - \psi\right)^2} \tag{40}$$

$$\mu_k = \mu_0\left((1-D)e^{-B\xi_{k-1}V} + D\right) \tag{41}$$

$$F_{ak} = \frac{2F_N \mu_{k-1}}{\pi}\left(\frac{k_A \varepsilon}{1+(k_A\varepsilon)^2} + \arctan(k_S \varepsilon)\right) \tag{42}$$

The components of the measurement matrix are as follows:

$$\ddot{y}_k = -\frac{2F_a}{\xi m_w}\left(\frac{\dot{y}_{k-1}}{V} - \psi\right) \tag{43}$$

The second component of the measurement matrix, i.e. $\dot{\psi}$, is obtained as equation (39).

A. Extended Kalman Filter

The EKF is an improved version of the conventional KF designed to handle nonlinear systems. The primary objective of this research is to identify the best estimation for the state vector of the wheelset. The EKF algorithm can be given by the following equations:

$$P_{k+1|k} = F_K P_k {F_K}^T + Q \tag{44}$$

$$K_K = P_{k+1|k} {H_k}^T \left(H_k P_{k+1|k} {H_k}^T + R\right)^{-1} \tag{45}$$

$$\hat{x}_{k+1|k} = f\left(\hat{x}_{k|k}.u_k\right) \tag{46}$$

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_K\left(z_k - h\left(\hat{x}_{k+1|k}\right)\right) \tag{47}$$

$$P_{k+1|k+1} = (I - K_K H_k) P_{k+1|k} \tag{48}$$

where $P_{k+1|k}$ is the priori prediction error covariance matrix, $P_{k+1|k+1}$ is the posteriori prediction error covariance matrix, K_K is the Kalman gain, $\hat{x}_{k+1|k}$ is the priori state prediction vector, $\hat{x}_{k+1|k+1}$is the posteriori state prediction vector, Q and R are the covariance matrixes of process and measurement noise, I is the unit matrix symbol, and F_K and H_k are the Jacobians of the system and the measurement equations defined as follows:

$$F_K = \begin{bmatrix} \frac{\partial \dot{y}_k}{\partial \dot{y}_k} & \frac{\partial \dot{y}_k}{\partial \dot{\psi}_k} & \frac{\partial \dot{y}_k}{\partial \xi_k} & \frac{\partial \dot{y}_k}{\partial \mu_k} & \frac{\partial \dot{y}_k}{\partial F_{ak}} \\ \frac{\partial \dot{\psi}_k}{\partial \dot{y}_k} & \frac{\partial \dot{\psi}_k}{\partial \dot{\psi}_k} & \frac{\partial \dot{\psi}_k}{\partial \xi_k} & \frac{\partial \dot{\psi}_k}{\partial \mu_k} & \frac{\partial \dot{\psi}_k}{\partial F_{ak}} \\ \frac{\partial \xi_k}{\partial \dot{y}_k} & \frac{\partial \xi_k}{\partial \dot{\psi}_k} & \frac{\partial \xi_k}{\partial \xi_k} & \frac{\partial \xi_k}{\partial \mu_k} & \frac{\partial \xi_k}{\partial F_{ak}} \\ \frac{\partial \mu_k}{\partial \dot{y}_k} & \frac{\partial \mu_k}{\partial \dot{\psi}_k} & \frac{\partial \mu_k}{\partial \xi_k} & \frac{\partial \mu_k}{\partial \mu_k} & \frac{\partial \mu_k}{\partial F_{ak}} \\ \frac{\partial F_{ak}}{\partial \dot{y}_k} & \frac{\partial F_{ak}}{\partial \dot{\psi}_k} & \frac{\partial F_{ak}}{\partial \xi_k} & \frac{\partial F_{ak}}{\partial \mu_k} & \frac{\partial F_{ak}}{\partial F_{ak}} \end{bmatrix} \tag{49}$$

$$H_k = \begin{bmatrix} \frac{\partial \ddot{y}_k}{\partial \dot{y}_k} & \frac{\partial \ddot{y}_k}{\partial \dot{\psi}_k} & \frac{\partial \ddot{y}_k}{\partial \xi_k} & \frac{\partial \ddot{y}_k}{\partial \mu_k} & \frac{\partial \ddot{y}_k}{\partial F_{ak}} \\ \frac{\partial \dot{\psi}_k}{\partial \dot{y}_k} & \frac{\partial \dot{\psi}_k}{\partial \dot{\psi}_k} & \frac{\partial \dot{\psi}_k}{\partial \xi_k} & \frac{\partial \dot{\psi}_k}{\partial \mu_k} & \frac{\partial \dot{\psi}_k}{\partial F_{ak}} \end{bmatrix} \tag{50}$$

The Jacobian matrices mentioned in (49) and (50) are used to specify the process and measurement matrices, which are shown in (51) and (52) respectively.

$$F_k = \begin{bmatrix} 1 - a_{11} & 0 & -\frac{F_{ak-1}}{\xi_{k-1}} a_{13} & 0 & a_{13} \\ 0 & 1 - \frac{m_w S^2}{J_w} a_{11} & \frac{F_{ak-1}}{\xi_{k-1}} a_{23} & 0 & -a_{23} \\ -\frac{m_w}{2\tau V} a_{13} & \frac{J_w}{2\tau V} a_{23} & 0 & 0 & 0 \\ 0 & 0 & a_{43} & 0 & 0 \\ 0 & 0 & a_{53} & a_{54} & 0 \end{bmatrix} \tag{51}$$

in which matrix elements are as follows:

$$a_{11} = \frac{2\tau F_{ak-1}}{\xi_{k-1} m_w V}$$

$$a_{13} = -\frac{2\tau}{\xi_{k-1} m_w}\left[\frac{\dot{y}_{k-1}}{V} - \psi\right]$$

$$a_{23} = \frac{2\tau S}{\xi_{k-1} J_w}\left[\frac{\kappa_w (y - y_d)}{r} + \frac{S\dot{\psi}_{k-1}}{V}\right]$$

$$a_{43} = -BV\mu_0\left((1-D)e^{-B\xi_{k-1}V}\right.$$

$$a_{53} = \frac{4a^2 bc}{3}\left(\frac{k_A\left(1-(k_A\varepsilon)^2\right)}{\left(1+(k_A\varepsilon)^2\right)^2} + \frac{k_S}{1+(k_S\varepsilon)^2}\right)$$

$$a_{54} = \left(k_A \frac{F_N \varepsilon}{\pi}\right)^3 \left(\frac{2\pi}{F_N\left(1+(k_A\varepsilon)^2\right)}\right)^2 + \frac{2F_N}{\pi} arctg k_s \varepsilon - \frac{2k_S F_N \varepsilon}{\pi\left(1+(k_S\varepsilon)^2\right)}$$

$$H_k = \begin{bmatrix} -\frac{1}{\tau} a_{11} & 0 & -\frac{F_{ak-1}}{\tau \xi_{k-1}} a_{13} & 0 & \frac{1}{\tau} a_{13} \\ 0 & 1 - \frac{m_w S^2}{J_w} a_{11} & \frac{F_{ak-1}}{\xi_{k-1}} a_{23} & 0 & -a_{23} \end{bmatrix} \tag{52}$$

In estimating adhesion based on longitudinal, lateral and yaw dynamics, the measurement matrix includes lateral acceleration and yaw rate. In addition, inertial sensors are used to measure lateral acceleration and yaw rate.

Finally, the outputs obtained from the measurements of the sensors (accelerometer and gyroscope) and the predicted measurements are used to estimate the states. Fig. 4 shows the process.

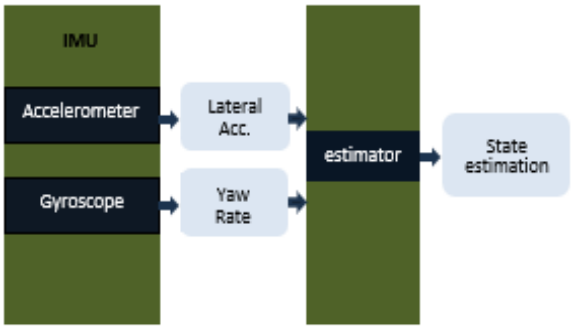


Fig. 4: State estimation overview.

Typically, the extended Kalman filter is not considered to be an optimal estimator and still has some shortcomings such as:

(1) Utilizing with the highly nonlinear system can be quite challenging.
(2) Since it needs Jacobian matrices for linearization, analytical derivation of this matrices is difficult and numerical derivation may impose a higher computational cost.
(3) The linearization introduces approximation errors that are not accounted for in the prediction and update steps.
(4) Due to the uncertainty surrounding the values of Q and R, they are acquired through trial-and-error approaches, resulting in a laborious and time-consuming process.

Referring to the EKF deficiencies, it is necessary to use an estimator that does not have such drawbacks in the estimation process. In the following, alternative estimators are investigated and their performances are compared.

B. Unscented Kalman Filter

The UKF is an alternative approach to linearization. While EKF treats the nonlinearity using analytical linearization, the UKF performs statistical linearization based on a set of rules. The approximation errors are consequences of linearization, which lead the EKF to underestimate state uncertainties. The UKF is formulated through the integration of the unscented transformation (UT) method, which is for calculating the statistics of a random variable that undergoes a nonlinear transformation. It is assumed that the wheelset system is in discrete-time nonlinear form with the state vector $\hat{x}_k$, the input vector u_k, and the observation vector z_k.

$$x_{k+1} = f(\hat{x}_k, u_k) + w_k \qquad w_k \sim (0, Q_k) \tag{53}$$

$$z_k = h(\hat{x}_k, u_k) + v_k \quad v_k \sim (0, R_k) \tag{54}$$

where Q and R are the system and observation noise covariance respectively.

At the beginning of the UKF implementation to estimate the state variables of the wheelset, a set of $2n_x + 1$ weighted samples or sigma points are determined as follows:

$$\chi_{k|k}^0 = \hat{x}_{k|k} \qquad i = 0 \tag{55}$$

$$\chi_{k|k} = \hat{x}_{k|k} + \left(\sqrt{(n_x + \lambda)P_{k|k}}\right)_i \qquad i = 1, \dots, n_x$$
$$\chi_{k|k} = \hat{x}_{k|k} - \left(\sqrt{(n_x + \lambda)P_{k|k}}\right)_i \qquad i = n_x + 1, \dots, 2n_x$$

$$w_m^{(0)} = \frac{\lambda}{\lambda + n_x} \tag{56}$$

$$w_c^{(0)} = \frac{\lambda}{\lambda + n_x} + 1 - \alpha^2 + \beta \tag{57}$$

$$w_c^{(i)} = w_m^{(i)} = \frac{\lambda}{2(\lambda + n_x)} \qquad i = 1, \dots, 2n_x \tag{58}$$

where $\hat{x}_{k|k}$ is the mean of x_{k+1}, $(\sqrt{(n_x + \lambda)P_{k|k}})_i$ is the ith-column of the matrix square root, $P_{k|k}$ is the covariance of x_{k+1}, n_x is the dimension of the state variables. The weights w_m and w_c are utilized for determining the mean and covariance respectively. α is employed to regulate the distribution of the sigma points around $\hat{x}_{k|k}$ and usually set to a small positive value between 0 and 1. β is a non-negative term utilized to incorporate prior knowledge of the distribution of x_{k+1}. Finally, $\lambda = \alpha^2(n_x + \rho) - n_x$ is a scaling parameter in which ρ is a secondary scaling parameter usually set to 0. It should be noted that in this study, the mentioned parameters are set as follows:

$$\alpha = 1, \quad \beta = 0, \quad \rho = 1$$

Sigma points $\chi_{k|k}$ are propagated through the nonlinear equations of the wheelset system. The transformed sigma points are assessed for each of the 0 to $2n_x$ points in the manner outlined below:

$$\chi_{k+1|k}^{(i)} = f(\chi_{k|k}^{(i)}, u_k) \tag{59}$$

The mean and covariance of the priori state estimation at time k are obtained by the following equations:

$$\hat{x}_{k+1|k} = \sum_{i=0}^{2n_x} w_m^{(i)} \chi_{k+1|k}^{(i)} \tag{60}$$

$$P_{k+1|k} = \sum_{i=0}^{2n_x} w_c^{(i)} (\chi_{k+1|k}^{(i)} - \hat{x}_{k+1|k})(\chi_{k+1|k}^{(i)} - \hat{x}_{k+1|k})^T + Q_k \tag{61}$$

In order to implement the measurement update, the equations (62)-(69) will be utilized. The transformed sigma points can be utilized to predict the measurements through the known nonlinear measurement equation. After rearranging the weighted sigma points, the covariance of the predicted measurement can be estimated. To consider the measurement noise, the covariance matrix R_k should be incorporated. Following that, the cross covariance can be estimated as per equation (65).

$$\chi_{k|k}^{(i)} = \left[\hat{x}_{k|k}^{(i)} \quad \hat{x}_{k|k}^{(i)} \pm \left(\sqrt{(n_x + \lambda)P_{k|k}^{(i)}}\right)_i\right] \tag{62}$$

$$\xi_{k+1|k}^{(i)} = h_{k+1}(\chi_{k+1|k}^{(i)}, U_{k+1}) \tag{63}$$

The expected measurement $\hat{z}_{k+1|k}$ is as:

$$\hat{z}_{k+1|k} = \sum_{i=0}^{2n_x} w_m^{(i)} \xi_{k+1|k}^{(i)} \tag{64}$$

Using the predicted sigma points, $P^{xz}_{k+1|k}$and $P^{zz}_{k+1|k}$ also determines as follows:

$$P^{zz}_{k+1|k} = \sum_{i=0}^{2n} \omega_i^{(c)} (\xi^{(i)}_{k+1|k} - \hat{z}_{k+1|k})(\xi^{(i)}_{k+1|k} - \hat{z}_{k+1|k})^T + R_k \quad (65)$$

$$P^{xz}_{k+1|k} = \sum_{i=0}^{2n} \omega_i^{(c)} (\chi^{(i)}_{k+1|k} - \hat{x}_{k+1|k})(\xi^{(i)}_{k+1|k} - \hat{z}_{k+1|k})^T \quad (66)$$

The mean and square root of covariance for the states are recalculated based on the actual measurement.

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}(z_{k+1} - \hat{z}_{k+1|k}) \quad (67)$$

$$P_{k+1|k+1} = P_{k+1|k} - K_{K+1} P^{zz}_{k+1|k} K^T_{k+1} \quad (68)$$

$$K_{k+1} = P^{xz}_{k+1|k} (P^{zz}_{k+1|k})^{-1} \quad (69)$$

From the above equations shown for UKF, it can be concluded that this filter has two main advantages compared to EKF, firstly, there is no need for Jacobians in UKF implementation, and secondly, UKF can estimate the mean and covariance of the states accurately the second order for any nonlinearity.

Results

In this part, wheel and rail adhesion force is estimated based on the lateral, longitudinal, and yaw dynamics of the wheelset. The values of the parameters mentioned in the equations of the previous sections are given in Table 1 and Table 2. It is worth noting that all simulations are done in MATLAB environment.

Table 1: Polach model parameters under different friction condition

Model parameter	Wheel–rail conditions			
	Dry	Wet	Low	Very Low
k_A	1	1	1	1
k_S	0.4	0.4	0.4	0.4
D	0.6	0.2	0.2	0.1
B	0.4	0.4	0.4	0.4

Table 2: Parameter values used in the simulation

$K_s(\frac{N}{m})$	6063260	$K_w\ (\frac{N}{rad})$	5×10^6
$r\ (m)$	0.5	$S(m)$	0.75
$J_R(Kgm^2)$	134	$\kappa_w\ (rad)$	0.15
$J_L(Kgm^2)$	64	$M_t\ (Kg)$	15000
$J_w(Kgm^2)$	700	FN (KN)	60
$m_w\ (Kg)$	1250	G $(\frac{N}{m^2})$	8.4×10^{10}

The values of friction coefficients and other required parameters used in equations are as follows:

$$\mu_0 = \begin{cases} 0.55 & t < 10 \\ 0.3 & 10 \le t < 20 \\ 0.06 & 20 \le t < 30 \\ 0.03 & 30 \le t < 35 \end{cases}$$

a = 0.0015 m, b = 0.0075 m, C_{11} =4.12, V= 15 $\frac{m}{s}$

Matrices Q and R are as follows:

Q = diag([5× 10^{-14},1× 10^{-14},1× 10^{-14},1 × 10^{-14} , 1× 10^{-14}])

R= diag([1× 10^{-1}, 1× 10^{-1}])

At the beginning of the simulation, Fig. 5 is developed based on the equations (11)-(17). In addition, a random input y_d is created to simulate the dynamics of the wheelset in the presence of irregularities that can be encountered in the railroad. In this model, the dynamics of each wheel in the right and left sides are shown separately.

Finally, the output of right and left wheel blocks attached to the lateral acceleration and yaw rate blocks. The main goal of developing this simulink model is to simulate the outputs of accelerometer and gyroscope sensors. In estimation process these outputs along with the predicted variables of the same type are used to estimate the state variables. Figs. 6 and 7 show the results of measuring lateral acceleration and yaw rate, respectively which are obtained from simulink execution.

In Fig. 8 the diagrams of lateral speed, yaw rate, slip, adhesion coefficient, and adhesion force of the wheelset are shown. These outputs show the simulation of the actual conditions of the system. In the estimator evaluation stage, the trajectories of these graphs are used as a pattern and the estimator's compliance in following the relevant pattern is used as a criterion to check the accuracy of the estimator in the estimation of state variables.

In Figs. 9-13 the diagrams of yaw rate, lateral speed, slip ratio, adhesion coefficient and adhesion force of the wheelset are shown in three situations, UKF-baesd and EKF-based estimated, and actual.

In Fig. 9, the actual and estimated trajectories of yaw rate change approximately between 0.8 and -0.8 and from the beginning the convergence of the estimated trajectories to the actual one is evident.

In Fig. 10, the actual and estimated trajectories of lateral velocity are depicted. As can be seen, the UKF-based estimated lateral velocity converges to the actual one in less than 1 second but this convergence in EKF-based estimation occurs after 5 seconds.

Therefore, the UKF estimator has provided an acceptable results regarding these two variables. In Figs. 11-13 which are related to the slip, adhesion coefficient and adhesion force respectively, the outputs of the two estimators are drawn and compared with real variables of the same type.

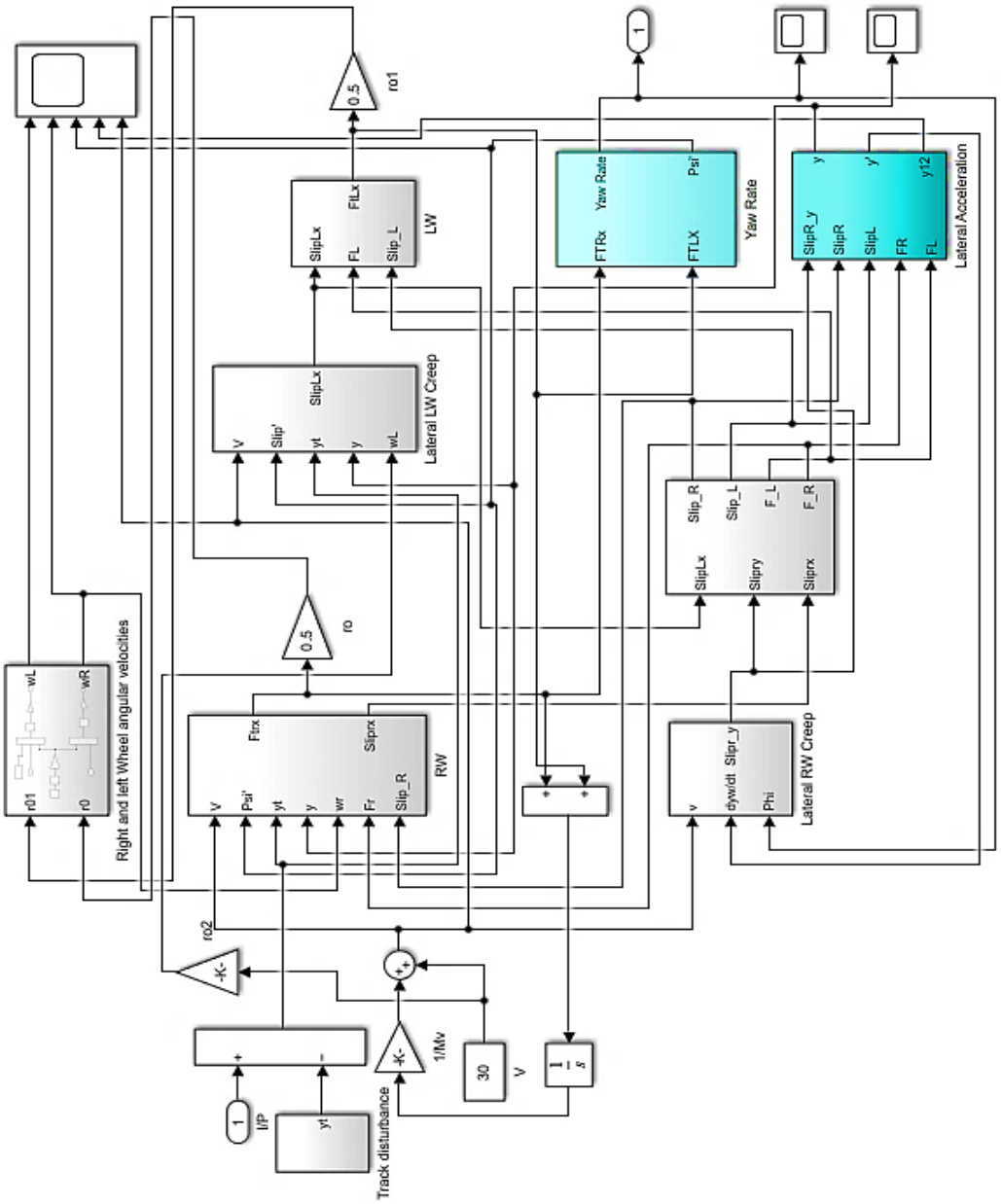


Fig. 5: Wheelset dynamics simulink model.

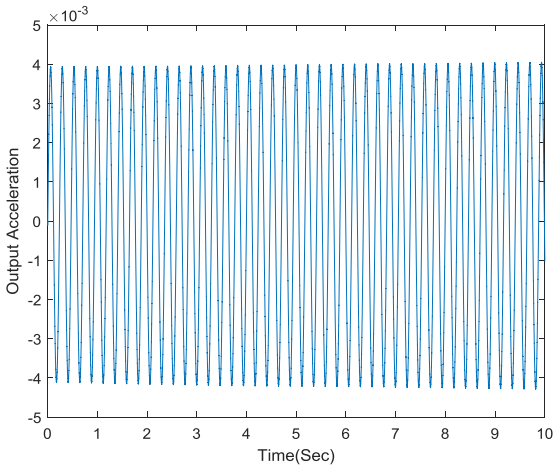


Fig. 6: Output diagram of accelerometer.

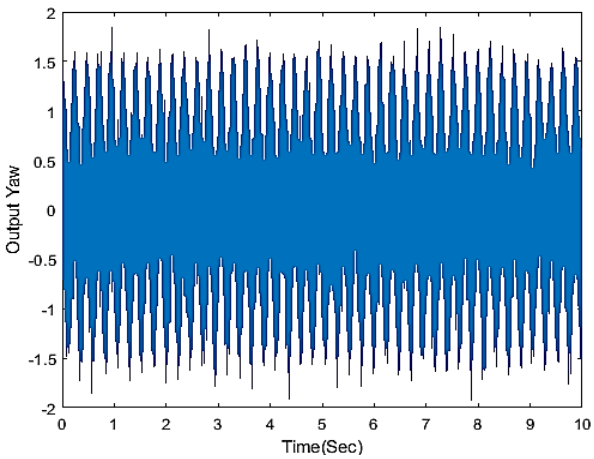


Fig. 7: Output diagram of gyroscope.

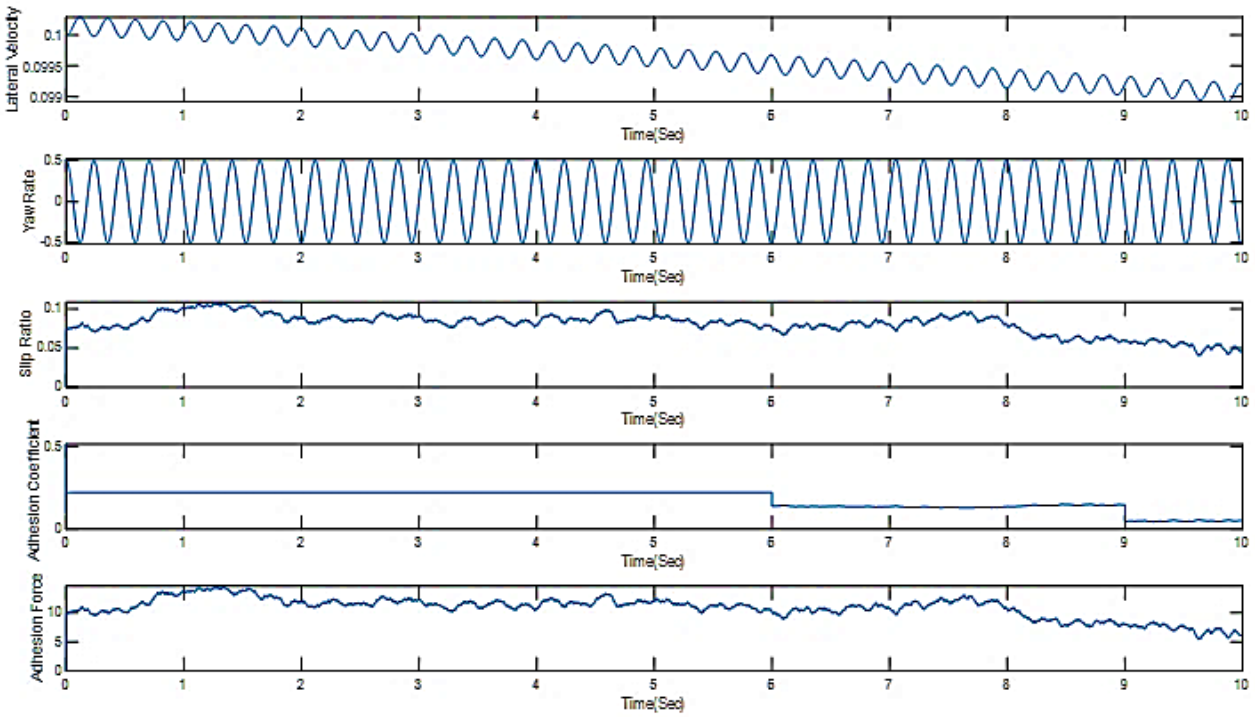


Fig. 8: dynamics of the wheelset.

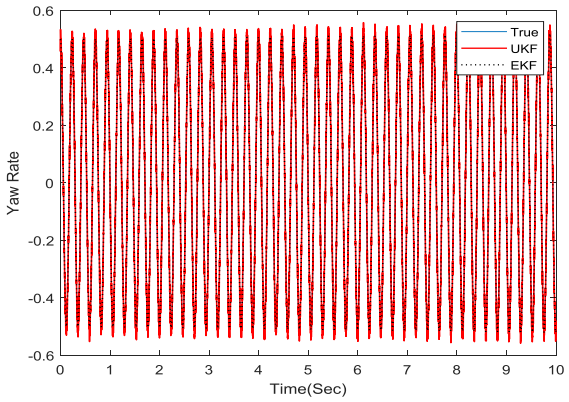


Fig. 9: Estimated, and actual trajectories of yaw rate.

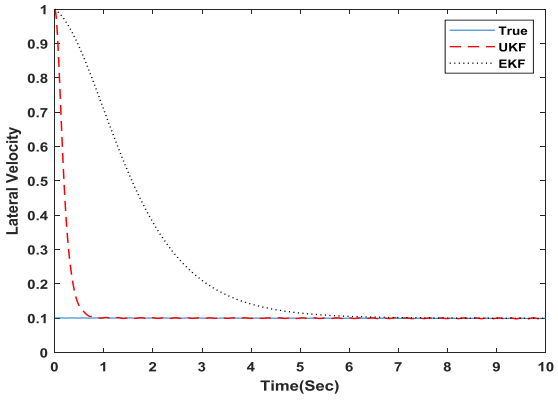


Fig. 10: Estimated, and actual trajectories of lateral velocity.

In Figs. 11-13 the simulation is carried out for 10 seconds to calculate slip ratio, adhesion coefficient, and adhesion force. All mentioned variables are estimated by EKF and UKF estimators.

Due to the irregularities exist in the lateral direction, fluctuations in the graphs are inevitable.

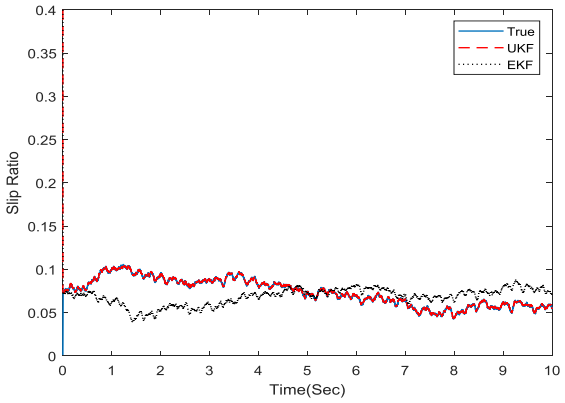


Fig. 11: Estimated, and actual trajectories of slip.

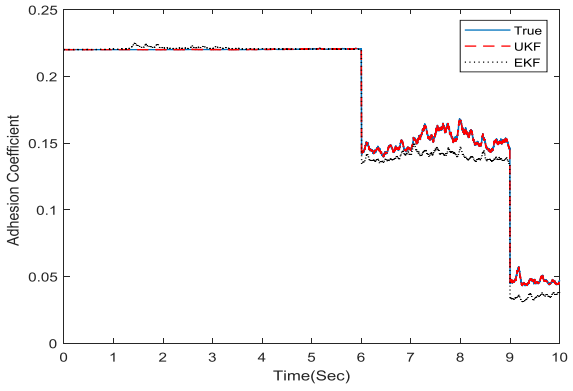


Fig. 12: Estimated and actual trajectories of adhesion coefficient.

In all three figures, UKF-based estimated outputs follow the actual trajectories of the variables with high convergence and accuracy but there is no necessary

convergence in the estimation of the mentioned variables based on EKF, which is more evident in estimating slip ratio and adhesion force. As can be seen in Fig. 12, estimation of adhesion coefficient with EKF has acceptable output up to 6 second but non-convergence after 6 second leads to ignoring this estimator as an ideal one. Therefore, in addition to the successful performance in estimating lateral velocity and yaw rate, UKF also shows a favorable performance in estimating slip ratio, adhesion coefficient, and adhesion force.

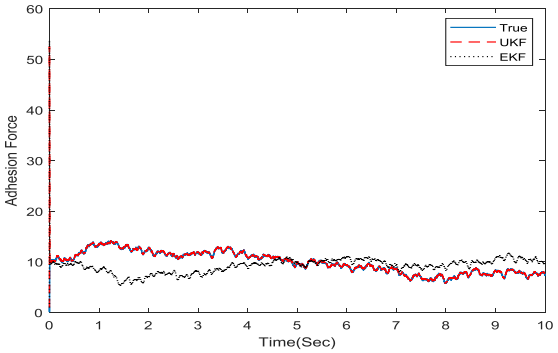


Fig. 13: Estimated and actual trajectories of adhesion force.

Conclusion and Future Work

The performance of railway operation mainly is affected by wheel-rail contact forces but it is not possible to measure these contact forces and interrelated dynamics directly, therefore it is necessary to estimate these wheelset dynamics through state of art technique. In this research paper, a railway wheelset model and a novel observer-based estimator are developed in Simulink/MATLAB to calculate and estimate nonlinear wheelset dynamics. The estimators based on the EKF and UKF are used to estimate adhesion coefficient, slip ratio, and yaw rate effectively in dry, wet, greasy and extremely slippery track conditions. The performances of the UKF and EKF algorithms are assessed and compared with each other. The UKF estimator not only verified excellent performance in the normal operation of a railway vehicle on a normal track but equally depicted robustness in traction and braking modes of the vehicle in wet, oily, and extremely slippery track conditions. The validity of the estimator is also checked in the transition of adhesion conditions from dry to extremely slippery and vice-versa during the simulation. In the future, this approach will be implemented on Field Programmable Gate Arrays (FPGA) platform for real-time condition monitoring of wheelset dynamics to avoid the accidents and derailment of railway vehicle.

Author Contributions

M. Moradi collected the data, carried out the analysis and wrote paper, R. Havangi wrote the paper, interpreted the results and supervised the research.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

a and b	Semi-axis length of the contact patch
B and D	Reduction factors
c	Contact shear stiffness coefficient
F_c	Centrifugal force
F_N	Normal force between the wheel and rail
k_A	Reduction factor in the adhesion area
k_S	Reduction factor in the slip area
M_t	Rail vehicle mass
n_i	Gear reduction ratio
r	Wheel radius
R_r and R_s	Rotor and stator resistance
S	Half gauge of track
T_m	Motor torque
T_L	Load torque
V	Longitudinal velocity
V_{wR}	Longitudinal velocity of the right wheel
V_{wL}	Longitudinal velocity of the left wheel
y	Lateral movement
y_d	Track irregularities in lateral direction
ϵ	Gradient of tangential stress
κ_w	Wheel conicity
ψ	Yaw angle
ξ	Total creepage between the wheel and rail
ξ_x	Longitudinal creepge
ξ_y	Lateral creepage
μ_f	Friction coefficient
μ_0	Maximum friction coefficient
ω_R	Angular velocity of the right wheel
ω_L	Angular velocity of the left wheel

References

[1] E. E. Magel, "A survey of wheel/rail friction," (No. DOT/FRA/ORD - 17-21), Federal Railroad Administration. Office of Research, Development, and Technology, Washington, DC, United States, 2017.

[2] U. Olofsson, "17 Adhesion and friction modification," in Wheel–Rail Interface Handbook, ELSEVIER, UK, pp. 510-527, 2009.

[3] Z. Yuan, M. Wu, C. Tian, J. Zhou, "A review on the application of friction models in wheel-rail adhesion calculation," Urban Rail Transit, 7: 1-11, 2021.

[4] Z. Shi, K. Wang, L. Guo, Z. Chen, "Effect of arc surfaces friction coefficient on coupler stability in heavy haul locomotives: simulation and experiment," Veh. Syst. Dyn., 55(9): 1368-1383, 2017.

[5] L. Buckley-Johnstone, G. Trummer, P. Voltr, K. Six, R. Lewis "Full scale testing of low adhesion effects with small amounts of water in the wheel/rail interface," Tribol. Int., 141: 105907, 2020.

[6] R. Lewis, G. Trummer, K. Six, J. Stow et al., "Leaves on the line: characterising leaf based low adhesion on railway rails," Tribol. Int., 185: 108529, 2023.

[7] H. Chen, "Wheel slip/slide and low adhesion caused by fallen leaves," Wear, 203187: 446-447, 2020.

[8] J. Zhou, M. Wu, C. Tian, Z. Yuan, C. Chen, "Experimental investigation on wheel–rail adhesion characteristics under water and large sliding conditions," Ind. Lubr. Tribol., 73(2): 366-372, 2021.

[9] K. Zhao, P. Li, Ch. Zhang, J. He, Y. Li, T. Yin, "Online accurate estimation of the wheel-rail adhesion coefficient and optimal adhesion antiskid control of heavy-haul electric locomotives based on asymmetric barrier lyapunov function," J. Sensors, 2740679: 1-12, 2018.

[10] R. Bibi, B. S. Chowdry, R. A. Shah, "PSO based localization of multiple mobile robots employing LEGO EV3," in Proc. 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET): 1-5, 2018.

[11] K. Ishizaka, B. White, M. Watson, S. R. Lewis, R. Lewis, "Influence of temperature on adhesion coefficient and bonding strength of leaf films: a twin disc study, Wear, 203330: 454-455, 2020.

[12] S. Shrestha, Q. Wu, M. Spiryagin, "Review of adhesion estimation approaches for rail vehicles," Int. J. Rail Transp., 7(2): 79-102, 2019.

[13] X. Fang, S. Lin, Z. Yang, F. Lin, H. Sun, L. Hu, "Adhesion control strategy based on the wheel-rail adhesion state observation for high-speed trains," Electronics, 7(5): 70, 2018.

[14] B. Liu, T.X. Mei, S. Bruni, "Design and optimisation of wheel–rail profiles for adhesion improvement," Veh. Sys. Dyn., 54(3): 429-444, 2016.

[15] Y. Chen, H. Dong, J. Lu, X. Sun, L. Guo, "A super-twisting-like algorithm and Its application to train operation control with optimal utilization of adhesion force," IEEE Trans. Intell. Transp. Syst., 17(11): 3035-3044, 2016.

[16] M. Yamashita, T. Soeda, "Anti-slip re-adhesion control method for increasing the tractive force of locomotives through the early detection of wheel slip convergence," in Proc. 17th European Conference on Power Electronics and Applications: 1-10, 2015.

[17] P. D. Hubbard, C. Ward, R. Dixon, R. Goodall, "Verification of model based adhesion estimation in the wheel-rail interface," Chem. Eng. Trans., 33: 757-762, 2013.

[18] C. P. Ward, R. M. Goodall, R. Dixon, G. A. Charles, "Adhesion estimation at the wheel–rail interface using advanced model-based filtering," Veh. Syst. Dyn., 50: 1797-1816, 2012.

[19] S. Strano, M. Terzo, "On the real-time estimation of the wheel-rail contact force by means of a new nonlinear estimator design model," Mech. Syst. Signal Process., 105: 391-403, 2018.

[20] Y. Zhao, B. Liang, "Re-adhesion control for a railway single wheelset test rig based on the behaviour of the traction motor," Veh. Syst. Dyn., 51(8): 1173-1185, 2013.

[21] S. Wang, J. Xiao, J. Huang, H. Sheng, "Locomotive wheel slip detection based on multi-rate state identification of motor load torque," J. Franklin Inst., 353(2): 521-540, 2016.

[22] P. Pichlik, J. Zdenek, "Extended Kalman filter utilization for a railway traction vehicle slip control," in Proc. International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), Intl Aegean Conference on Electrical Machines and Power Electronics (ACEMP): 869-874, 2017.

[23] S. Strano, M. Terzo, "On the real-time estimation of the wheel-rail contact force by means of a new nonlinear estimator design model," Mech. Syst. Signal Process., 105: 391-403, 2018.

[24] S. J. Julier, J. K. Uhlmann, "Unscented filtering and nonlinear estimation," Proc. IEEE, 92: 401-422, 2004.

[25] S. J. Julier, J. K. Uhlmann, H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in Proc. American Control Conference - ACC'95, Autom Control Council: 1628-1632, 1995.

[26] Y. Zhao, B. Liang, S. Iwnicki, "Friction coefficient estimation using an unscented Kalman filter," Int. J. Veh. Mech. Mobility, 52(1): 220-234, 2014.

[27] T. Gajdar, I. Rudas, Y. Suda, "Neural network based estimation of friction coefficient of wheel and rail," in Proc. IEEE International Conference on Intelligent Engineering Systems: 315-318, 1997.

[28] A. Shebani, S. Iwnicki, "Prediction of wheel and rail wear under different contact conditions using artificial neural networks," Wear, 406-407: 173-184, 2018.

[29] S. Z. Meymand, A. Keylin, M. Ahmadian, "A survey of wheel–rail contact models for rail vehicles," Int. J. Veh. Mech. Mobility, 54: 386-428, 2016.

[30] J. Belanger, P. Venne, J. N. Paquin, "The what, where and why of real-time simulation," Transient Analysis of Power Systems: Solution Techniques, Tools, and Applications, 37-49, 2011.

[31] O. Polach, "Creep forces in simulations of traction vehicles running on adhesion limit," Wear, 258: 992-1000, 2005.

[32] G. Charles, R. Goodall, R. Dixon, "Model-based condition monitoring at the wheel-rail interface," Veh. Syst. Dyn., 46: 415-430, 2008.

Biographies

Maryam Moradi received her M.S. degrees in Telecommunications Engineering from the Faculty of Engineering, University of Sistan and Baluchestan, Zahedan, Iran, in 2015, and received her Ph.D. degree in Control Engineering in Faculty of Engineering, University of Birjand, Birjand, Iran, in 2024. In 2019, she joined University of Applied Sciences & Technology as a Teacher. Her research interests are neural network, estimation and filtering.

- Email: m_moradi@birjand.ac.ir
- ORCID: 0009-0007-0635-6867
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Ramazan Havangi received his M.S. and Ph.D. degrees from the K.N. Toosi University of Technology, Tehran, Iran, in 2003 and 2012, respectively. He is currently an Associate Professor of control systems with the Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. His main research interests are inertial navigation, integrated navigation, estimation and filtering, evolutionary filtering, simultaneous localization and mapping, fuzzy, neural network, and soft computing.

- Email: Havangi@Birjand.ac.ir
- ORCID: 0000-0001-5711-3127
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/havangi/fa

How to cite this paper:
M. Moradi, R. Havangi, "Nonlinear filter-based estimation of wheel-rail contact forces and related considerations using inertial measurement unit," J. Electr. Comput. Eng. Innovations, 13(2): 353-364, 2025.

DOI: 10.22061/jecei.2025.11176.772

URL: https://jecei.sru.ac.ir/article_2260.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Damping Critical Electromechanical Oscillations via Generators Redispatch Considering ZIP Load Model and Transmission Lines Resistance

M. Setareh [*]***, A. Moradibirgani***

Faculty of Electrical Engineering, Shahid Beheshti University, Tehran, Iran.

Article Info

Article History:
Received 24 October 2024
Reviewed 12 December 2024
Revised 18 January 2025
Accepted 22 January 2025

Keywords:
Generators redispatch
Oscillatory modes
Remedial action
Sensitivities
Transmission line resistance
ZIP load model

[*]Corresponding Author's Email Address: *m_setareh@sbu.ac.ir*

Abstract

Background and Objectives: This paper proposes a novel formula to calculate the sensitivities of electromechanical modes with respect to generators active power changes. The generic ZIP load model is considered and the effect of various types of loads on the best paradigm of generators redispatch (GR) is investigated. Furthermore, transmission lines resistance are modeled in the proposed formulae; and the best GR schemes to improve the power system damping with considering and neglecting transmission lines resistance are compared.

Methods: Four energy functions are defined and the quadratic eigenvalue problem is applied to construct the framework of the proposed formula. The dynamic equation of the classical model of synchronous generators along with algebraic equations of power network considering transmission lines resistance and ZIP model of power system loads are written in a systematic manner using the partial differential of the energy functions. Then, set of equations of the power system are linearized and sensitivity factors are calculated using power system model parameters and power flow variables, which can be either obtained via state estimation or measured directly by phasor measurement units.

Results: The 39-bus New England power system is used to calculate sensitivities. The value of Sensitivity factors in conditions of considering transmission lines resistance and neglecting ones are compared and then the best GR plan to improve critical damping is determined. If all the loads are assumed to be in constant power mode, then for two modes with and without considering transmission lines resistance, generators pairs (9,1) and (5,2) are the best redispatch plans to damp oscillations. However, If all the loads are assumed to be in constant current mode, the best generators pair without considering transmission lines resistance mode does not change, although, it changes to generators pair (5,1) for the mode of considering transmission lines resistance.

Conclusion: Using the classical model of synchronous generators does not give information about the damping-ratio of the inter-area mode and only estimates its frequency well. Besides, considering the load model and resistance of transmission lines change the best paradigm of GR to suppress oscillations.

Introduction

Maintaining the small signal stability is a vital issue in the operation of power systems. Fig. 1 portrays the classification of strategies, methods, tools, and criteria for improving the small signal stability of power system. As seen in this figure, the corresponding strategies are

divided into twofold: 1) equipment-based [1] and 2) remedial actions [2]. Their primary distinction is that the first category controllers are always in service, whereas corrective actions are applied to the system as needed.

Besides, remedial actions usually impose additional costs on the operating power system, whereas the costs associated with applying the first strategy pertain to the expansion planning of the power system.

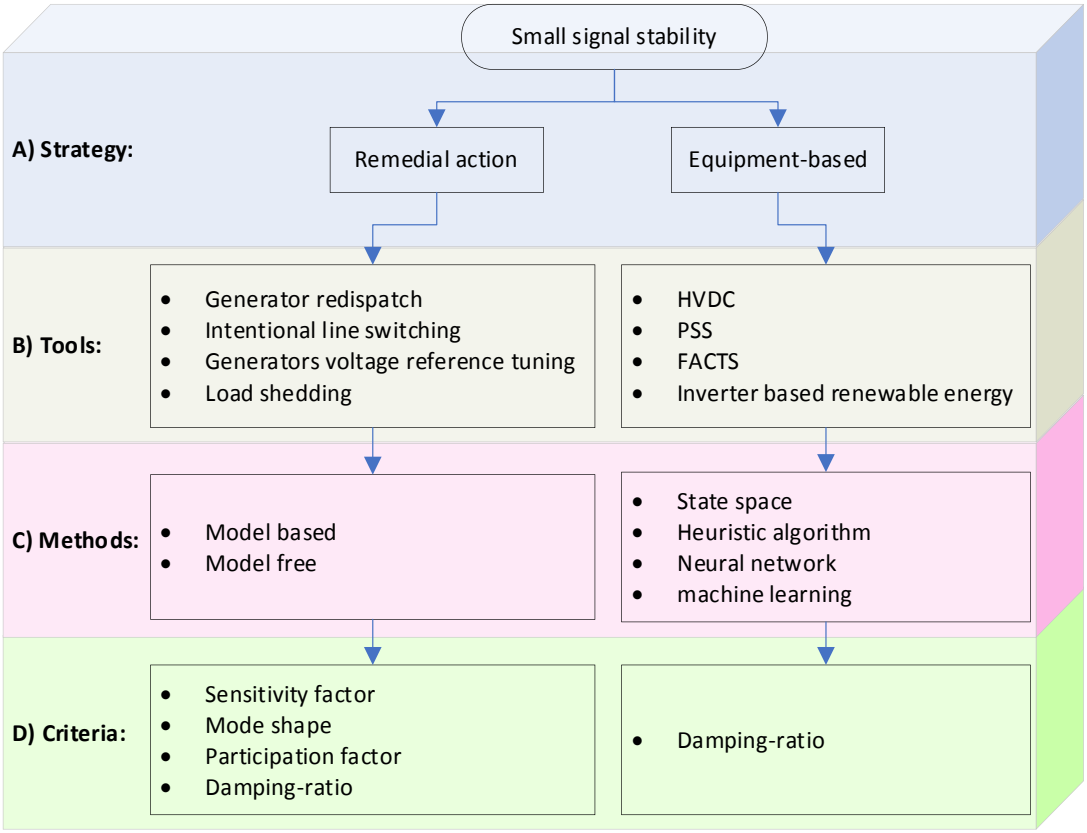


Fig. 1: Power system small signal stability improvement. A) strategies, B) tools, C) methods, and D) criteria.

In the equipment-based category, various mathematical methods such as state space, transfer function, and heuristic algorithms have been applied to tune the parameters of controllers. The most popular device in this category is the power system stabilizer (PSS) installed next to the generator excitation system [3]. This electronic equipment has lead-lag blocks to provide phase compensation. Setting the parameters of these blocks requires detailed studies, and therefore it is accomplished offline. Meta-heuristic algorithms such as genetic algorithm [4], particle swarm optimization [5], chaotic bat algorithm [6], and adaptive rat swarm [7] were applied to determine the PSS parameters. The industrial application of this equipment is mainly limited to damp local electromechanical oscillations. However, synchronized wide-area signals were used to extend the performance of PSS for damping inter-area oscillations [8], [9]. Nevertheless, power network controllers i.e. HVDC, and flexible AC transmission system (FACTS) equipped with a converter are more efficient for suppressing inter-area oscillations. Because they can be installed in the substations connected to transmission tie-lines.

The authors in [10] used model predictive control for tuning parameters of HVDC controllers, and in [11], augmented random research was applied to do so. Static var compensator (SVC) [12], static compensators (STATCOM) [13], and thyristor Controlled Series Capacitor (TCSC) [14] are more applicable FACTS devices to improve power system small signal stability. SVC and TCSC provide instantaneous reactive power support to the connected power system and help to dampen electromechanical oscillations by controlling the voltage of the point of common coupling with the power grid. However, STATCOM can suppress power oscillations by exchanging active and reactive powers with the power grid.

With the proliferation of renewable sources in smart power systems, their penetration has reached such a level that they can play an important role in power system stability. These resources are mainly connected to the power network through a converter. Therefore, applying auxiliary power oscillation damping controllers which produce a proper modulation signal for the grid-connected Inverter of renewable resources has been given special attention [15]-[17]. Authors in [18] designed

optimal probabilistic robust damping controllers to suppress multiband oscillations in the power system integrated with wind farms using adaptive compass search. The authors in [19] developed a coordinated control scheme for utility-scale photovoltaic and wind power plants to suppress electromechanical oscillations while maintaining voltage stability.

The controllers in the equipment-based strategy are tuned for a specific power system operating point. Therefore, their efficacy drastically deteriorates at out-of-range operating points, and remedial actions in this situation can be applied. Because optimal remedial actions are determined based on power system operation conditions and can be applied whenever needed.

The frequency of electromechanical oscillations is usually in the range of 0.1 Hz to 1.5 Hz [2]. Therefore, there is enough time to perform remedial measures to improve the damping of the power system. The application of online remedial measures in nowadays modern power systems is growing due to the proliferation of measurement instrument installation i.e. phasor measurement units (PMU), as well as the emergence of more powerful processors. PMU can send measured data at a rate of several tens of Hz to the power system control center, and therefore proper situational awareness of the real-time condition of the power system can be gained. [20], [21]. Therefore, proactive remedial measures to improve the operating conditions of the power system can be used online [22]. Nonetheless, remedial actions can be used either preventively or correctively. If appropriate measures are taken to improve the operational conditions before any event occurs, remedial measures are preventive. The measures taken after the incident to rescue the system from instability are referred to as corrective measures. Since the power system is non-linear, its small signal characteristics vary when the operating point changes via applying remedial actions. GR [23]-[29], transmission lines switching [30], load shedding [25], and voltage reference tuning of generators [31] are remedial actions introduced in the literature. However, GR is a more applicable remedial action to suppress low-damp electromechanical oscillations.

The best remedial actions to achieve the preset aim have been determined in the literature through two methods: 1) model-based [32] and 2) model-free [33]. In the first method, sensitivity factors of the power system with respect to remedial actions are determined analytically. Direct calculation of the amplitude of remedial measures is the main pros of these methods. However, the need for an accurate system model is a disadvantage. In model-free methods, measurement data is used, and therefore, the need for a valid database is one of the limitations of these methods. Besides, they cannot estimate sensitivity factors directly. The usually use criteria such as participation factors and mode shape to specify the best remedial actions [34].

In [23], generators active power sensitivity factors were used to determine optimal power flow (OPF). To do so, first- and second-order sensitivities of critical eigenvalues with respect to OPF variables were calculated. Authors in [24] used aperiodic small signal rotor angle stability assessment strategy to detect unstable power system operation, and GR was applied such that the critical generator backs into the secure region. In [25], OPF for achieving an appropriate security level in terms of small-signal rotor angle stability was applied. To do so, sensitivity factors with respect to GR have been calculated by a repetitive approach. Critical incidents were recognized in terms of security margin and the operating points of energy resources and consumers were determined considering the cost of generation variations and load decrement.

Authors in [26] processed synchronized wide area signals using independent component analysis along with random decrement technique to estimate damping-ratio of critical oscillatory modes. An iterative algorithm was used to determine required changes of generators active power for improving the critical inter-area mode damping. Mode shape index was applied to specify the most effective generators pair for redispatching. Then, their active power was changed step by step. At each step, the damping-ratio of the critical electromechanical mode was estimated using the algorithm of random decrement. This process is repeated until the desired damping is reached. In [27], analytical formulae were proposed for model-based modal analysis. Second-order synchronous generators dynamic model was used and corresponding equations were organized in the form of the quadratic eigenvalue problem (QEP). Lossless transmission lines and constant active power loads were considered, and sensitivity factors related GR for improving damping-ratio of oscillatory modes were calculated.

Iterative-based method for GR sensitivity factors calculation was applied in [28]. Sensitivity factors were specified by standard modal analysis in both conditions of before and after changing generators active power. At each iteration, the active power of generators has been changed optimally using sensitivity factors. This process continues until the minimum damping constraint is met. Likewise, authors in [29] proposed a sequential GR to damp low-frequency oscillations. An analytical model-based manner based on normalized participation factors (NPF) was presented to determine an effective GR scheme for improving power system damping. NPF was calculated as an indicator of sensitivity factors.

In [30] a list of effective transmission lines switching was arranged offline. To do so, modal analysis was performed after each transmission line switching, and the

damping-ratio changes of critical electromechanical modes were calculated. Since transmission line switching may cause a violation of the system security constraints, an optimization algorithm for determining the required generators active power and load shedding was used.

Voltage reference tuning of generators is a cost-free remedial action and its efficacy in suppressing electromechanical oscillations was investigated in [31]. However, this study needs the synchronous generator dynamic model comprising the field winding and the excitation system.

The implementation of online remedial measures is a relatively new strategy that has become feasible by the intelligentization of power systems. This advancement has spurred researchers to calculate various indices for identifying the effective remedial actions on enhancing the performance of the power system. Sensitivity coefficients represent accurate information for this purpose. In this paper, we develop our previous research in [35] to take into account the generic ZIP load model along with the resistance of transmission lines for calculating sensitivity factors of electromechanical modes using a systematic method. The dynamic-algebraic equations (DAE) of the power system are arranged in the form of the quadratic eigenvalue problem. Then, they are linearized, and sensitivity factors of oscillatory modes related to the GR remedial action are calculated. The main contributions of this paper are summarized as follows:

- Proposing four energy functions to model the power system linearized DAE by a regular method in the form of QEP.
- Modeling ZIP load model and transmission lines resistance to calculate the sensitivities.
- Scrutinizing the impact of load model and transmission lines resistance to specify the most effective generators redispatches.

The rest of the paper continues as follows. In the next section, four energy functions are proposed to arrange dynamic-algebraic equations of the power system into the form of the quadratic eigenvalue problem. Then, the differential equations are taken and the closed-form equations of sensitivity factors of the electromechanical modes are presented. In the simulation section, IEEE 39 bus test system is used to determine critical mode sensitivities. Finally, the last section concludes the paper and discusses directions for future work.

Power System Dynamic-Algebraic Equations

Considering the second-order model of synchronous generators and the complete model of the power network, the model of the power system for studying small signal stability is illustrated as shown in Fig. 2. Assuming that the power system has m generators, buses 1 to m are internal buses of the generators and other buses $m+1$ to n be load buses.

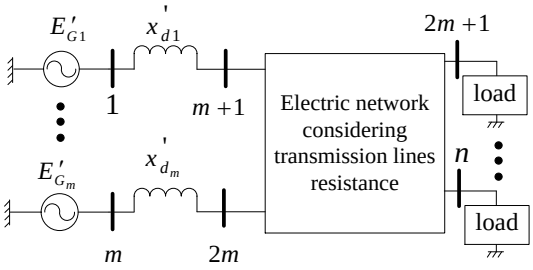


Fig. 2: Power system model.

Equations (1) and (2) exhibit internal buses and other buses active power balance, respectively. Besides, Equation (3) represents reactive power balance equations for non-generator buses [36].

$$\frac{2h_i}{\omega_0}\ddot{\delta}_i+\frac{d_i}{\omega_0}\dot{\delta}_i+G_{ii}V_i^2+V_i\sum_{\substack{j=1\\ j\neq i}}^{n}G_{ij}V_j\cos\left(\delta_i-\delta_j\right)+V_i\sum_{\substack{j=1\\ j\neq i}}^{n}B_{ij}V_j\sin\left(\delta_i-\delta_j\right)=P_{mech_i}\qquad i=1,\ldots,m \tag{1}$$

$$G_{ii}V_i^2+V_i\sum_{\substack{j=1\\ j\neq i}}^{n}G_{ij}V_j\cos\left(\delta_i-\delta_j\right)+V_i\sum_{\substack{j=1\\ j\neq i}}^{n}B_{ij}V_j\sin\left(\delta_i-\delta_j\right)=P(V_i)\quad i=m+1,\ldots,n \tag{2}$$

$$\sum_{\substack{j=1\\ j\neq i}}^{n}G_{ij}V_j\sin\left(\delta_i-\delta_j\right)-\sum_{\substack{j=1\\ j\neq i}}^{n}B_{ij}V_j\cos\left(\delta_i-\delta_j\right)=B_{ii}V_i+\frac{Q_i(V_i)}{V_i}\qquad i=m+1,\ldots,n \tag{3}$$

where h_i and d_i are inertia constant in second and damping constant in pu torque/pu speed of the i^{th} generator. G_{ii} and B_{ii} are real and imaginary parts of entry i,j of Y_{bus} matrix, respectively. Furthermore, P_{mech_i}, $P(V_i)$ and $Q(V_i)$ are the input mechanical power of i^{th} generator and net active and reactive power injections at bus i, respectively.

Here, power system loads are modeled by the static model which is known as ZIP model. The generic formulation for ZIP load model is equal to [37]:

$$P(V)=P_0\left(k_{PI}\left(\frac{V}{V_0}\right)+k_{PZ}\left(\frac{V}{V_0}\right)^2+k_{PC}\right) \tag{4}$$

$$Q(V)=Q_0\left(k_{QI}\left(\frac{V}{V_0}\right)+k_{QZ}\left(\frac{V}{V_0}\right)^2+k_{QC}\right) \tag{5}$$

where k_{PI}, $k_{PZ,}$ and k_{PC} are coefficients of the proportion of constant current, impedance, and power of the active power load. Similarly, k_{QI}, $k_{QZ,}$ and k_{QC} are the coefficients related to the reactive power load. It is noteworthy that the sum of the coefficients of the different portions of the active power must be equal to 1. The same is true for the reactive power coefficients.

Organizing Power System Equations in QEP Model

QEP Model is equivalent to a generalized eigenvalue problem. This is a more comprehensive form than the standard state space method. The system equations can be written as a mixture of zero to second-order equations next to each other [38], [39].

Equations (6) to (9) show four proposed energy functions to arrange the power system DAE into QEP model:

$$R^{Pbus} = -\sum_{i=1}^{m} P_{mech_i}\delta_i - \sum_{i=m+1}^{n} P(V_i)\delta_i \tag{6}$$

$$R^{Qbus} = -\sum_{i=m+1}^{n}\left(\begin{array}{l}\frac{1}{2}b_{ii}V_i^2 + \\ \mathrm{Q}_{i_0}\left(k_{QI}\frac{V_i}{V_{i_0}} + \frac{1}{2}k_{Qz}(\frac{V_i}{V_{i_0}})^2 + k_{QC}\ln V_i\right)\end{array}\right) \tag{7}$$

$$R^{B} = -\sum_{\substack{i,j \\ i\neq j}} B_{ij}V_iV_j\cos\left(\delta_i-\delta_j\right) \tag{8}$$

$$R^{G} = -\sum_{\substack{i,j \\ i\neq j}} G_{ij}V_iV_j\cos\left(\delta_i-\delta_j\right) \tag{9}$$

Now, the system equations can be rewritten as:

$$\frac{2h_i}{\omega_0}\ddot{\delta}_i + \frac{d_i}{\omega_0}\dot{\delta}_i + G_{ii}V_i^2 + \frac{\partial^2 R^G}{\partial\delta_i^2} + \frac{\partial R^{Pbus}}{\partial\delta_i} + \frac{\partial R^B}{\partial\delta_i} = 0 \qquad i=1,...,m \tag{10}$$

$$G_{ii}V_i^2 + \frac{\partial^2 R^G}{\partial\delta_i^2} + \frac{\partial R^{Pbus}}{\partial\delta_i} + \frac{\partial R^B}{\partial\delta_i} = 0 \qquad i=m+1,...,n \tag{11}$$

$$\frac{\partial^2 R^G}{\partial V_i\partial\delta_i} + \frac{\partial R^{Qbus}}{\partial V_i} + \frac{\partial R^B}{\partial V_i} = 0 \qquad i=m+1,...,n \tag{12}$$

The above equations are non-linear, and their linearization leads to the following quadratic differential equations:

$$\frac{2h_i}{\omega_0}\Delta\ddot{\delta}_i + \frac{d_i}{\omega_0}\Delta\dot{\delta}_i + \sum_{j=1}^{2n-m}\left(\frac{\partial L_{i,j}^G}{\partial\delta_i} + L_{i,j}^B + L_{i,j}^{Pbus}\right)\Delta z_j = 0 \qquad i=1,...,m \tag{13}$$

$$\left(2G_{ii}V_i\right)\Delta V_i + \sum_{j=1}^{2n-m}\left(\frac{\partial L_{i,j}^G}{\partial\delta_i} + L_{i,j}^B + L_{i,j}^{Pbus}\right)\Delta z_j = 0 \quad i=m+1,...,n \tag{14}$$

$$\sum_{j=1}^{2n-m}\left(\begin{array}{l}\frac{\partial L_{i,j}^G}{\partial\delta_{i-n+m}} \\ +L_{i,j}^B + L_{i,j}^{Qbus}\end{array}\right)\Delta z_j = 0 \qquad i=n+1,...,2n-m \tag{15}$$

where $\mathbf{z}$ is the state vector of the system. It involves the voltage angle of all buses and voltage amplitudes of non-generator buses as follows. Matrices $\mathbf{L}^B, \mathbf{L}^G, \mathbf{L}^{Pbus}$, and $\mathbf{L}^{Qbus}$ are hessian matrices of energy functions R^B, R^G, R^{Pbus}, and R^{Qbus}, respectively. Dimensions of all matrices equal $(2n-m)\times(2n-m)$.

$$\mathbf{z} = \begin{bmatrix} z_1 & ... & z_{2n-m}\end{bmatrix} = \begin{bmatrix}\delta_1 & .... & \delta_n & V_{m+1} & ... & V_n\end{bmatrix} \tag{16}$$

A. Computing Hessian Matrices

For the convenience of calculating matrices $\mathbf{L}^B$ and $\mathbf{L}^G$, the following variable transformation is used:

$$\boldsymbol{\theta} = \left[\theta_1,...,\theta_l\right] = \left[\delta_1,...,\delta_n\right]\times\mathbf{A} \tag{17}$$

$$\boldsymbol{\upsilon} = \left[\upsilon_1,...,\upsilon_l\right] = \left[\ln V_1,...,\ln V_n\right]\times|\mathbf{A}| \tag{18}$$

where A matrix is the incidence matrix of the power system, and l is the number of transmission lines.

Using (17) and (18), the energy functions R^B and R^G are rewritten in terms of the new state variables as follows, and then all hessian matrices are calculated as (21)-(22).

$$R^B = -\sum_{k=1}^{l} B_k e^{\upsilon_k}\cos\left(\theta_k\right) \tag{19}$$

$$R^G = \sum_{k=1}^{l} G_k e^{\upsilon_k}\cos\left(\theta_k\right) \tag{20}$$

$$\mathbf{L}^B = \mathbf{H}^T\begin{bmatrix}\frac{\partial^2 R^B}{\partial\theta^2} & \frac{\partial^2 R^B}{\partial\theta\,\partial\upsilon} \\ \frac{\partial^2 R^B}{\partial\upsilon\partial\theta} & \frac{\partial^2 R^B}{\partial\upsilon^2}\end{bmatrix}\mathbf{H} + \begin{bmatrix}\mathbf{0}_{n\times n} & \mathbf{0}_{n\times(n-m)} \\ \mathbf{0}_{(n-m)\times n} & \mathbf{L}^{Bdiag}\end{bmatrix} \tag{21}$$

$$\mathbf{L}^G = \mathbf{H}^T\begin{bmatrix}\frac{\partial^2 R^G}{\partial\theta^2} & \frac{\partial^2 R^G}{\partial\theta\,\partial\upsilon} \\ \frac{\partial^2 R^G}{\partial\upsilon\,\partial\theta} & \frac{\partial^2 R^G}{\partial\upsilon^2}\end{bmatrix}\mathbf{H} + \begin{bmatrix}\mathbf{0}_{n\times n} & \mathbf{0}_{n\times(n-m)} \\ \mathbf{0}_{(n-m)\times n} & \mathbf{L}^{Gdiag}\end{bmatrix} \tag{22}$$

where:

$$\frac{\partial^2 R^B}{\partial\theta^2} = -\frac{\partial^2 R^B}{\partial\upsilon^2} = diag\left\{B_1E_1,...,B_lE_l\right\} \tag{23}$$

$$\frac{\partial^2 R^B}{\partial\theta\,\partial\upsilon} = \frac{\partial^2 R^B}{\partial\upsilon\,\partial\theta} = diag\left\{B_1F_1,...,B_lF_l\right\} \tag{24}$$

$$\left[\mathbf{L}^{Bdiag}\right]_{i,j} = \begin{cases}\sum_{k=1}^{l}\frac{|A(i+m,k)|}{V_{i+m}^2}\times B_kE_k & i=j=1,...,n-m \\ 0 & \text{otherwise}\end{cases} \tag{25}$$

$$\frac{\partial^2 R^G}{\partial\theta^2} = -\frac{\partial^2 R^G}{\partial\upsilon^2} = -diag\left\{G_1E_1,...,G_lE_l\right\} \tag{26}$$

$$\frac{\partial^2 R^G}{\partial\theta\,\partial\upsilon} = \frac{\partial^2 R^G}{\partial\upsilon\,\partial\theta} = -diag\left\{G_1F_1,...,G_lF_l\right\} \tag{27}$$

$$\left[\mathbf{L}^{Gdiag}\right]_{i,j} = \begin{cases}-\sum_{k=1}^{l}\frac{|A(i+m,k)|}{V_{i+m}^2}\times G_kE_k & i=1,...,n-m \\ 0 & \text{otherwise}\end{cases} \tag{28}$$

$$E_k = e^{\upsilon_k}\cos(\theta_k) \tag{29}$$

$$F_k = e^{\upsilon_k}\sin(\theta_k) \tag{30}$$

$$\mathbf{H}=\begin{bmatrix}\mathbf{A}^T & \mathbf{0}_{l\times(n-m)}\\ \mathbf{0}_{l\times n} & \begin{bmatrix}\frac{|A_{m+1,1}|}{V_{m+1}} & \cdots & \frac{|A_{m+1,l}|}{V_{m+1}}\\ \vdots & \cdots & \vdots\\ \frac{|A_{n,1}|}{V_n} & \cdots & \frac{|A_{n,l}|}{V_n}\end{bmatrix}^T\end{bmatrix} \tag{31}$$

where $\mathbf{L}^{Pbus}$ and $\mathbf{L}^{Qbus}$ are simplicity obtained using second-order derivatives in terms of the original state variables as follows:

$$\left[L^{Pbus}\right]_{i,j}=\begin{bmatrix}\mathbf{0}_{(2n-m)\times m} & \left[\frac{\partial^2 R^{Pbus}}{\partial \mathbf{z}^2}\right]^T & \mathbf{0}_{(2n-m)\times(n-m)}\end{bmatrix}^T$$
$$=\begin{cases}\left(\frac{-P_{i_0}}{V_{i_0}}\right)\times\left(k_{PI}+2k_{PZ}\frac{V_i}{V_{i_0}}\right) & i=m+1,\ldots,n\\ & j=n+1,\ldots,2n-m\\ 0 & \text{otherwise}\end{cases} \tag{32}$$

$$\left[L^{Qbus}\right]_{i,j}=\begin{bmatrix}\mathbf{0}_{(2n-m)\times n} & \left[\frac{\partial^2 R^{Qbus}}{\partial \mathbf{z}^2}\right]^T\end{bmatrix}^T$$
$$=\begin{cases}\left(-B_{i-n+m,i-n+m}+\right. & \\ \left. Q_{(i-n+m)_0}\left(-\frac{k_{QZ}}{V^2_{(i-n+m)_0}}+\frac{k_{QC}}{V^2_{i-n+m}}\right)\right) & i=j=\\ & n+1,\ldots,2n-m\\ 0 & \text{otherwise}\end{cases} \tag{33}$$

B. Modal Analysis using Quadratic Polynomial Matrix (QPM)

Finally, if the set of equations are arranged in the matrix form, then the following equation is obtained.

$$\mathbf{M}\Delta\ddot{\mathbf{z}}+\mathbf{D}\Delta\dot{\mathbf{z}}+\left(\mathbf{L}^V+\mathbf{L}^{G\delta}+\mathbf{L}^B+\mathbf{L}^{Pbus}+\mathbf{L}^{Qbus}\right)\Delta\mathbf{z}=\mathbf{0} \tag{34}$$

where:

$$\mathbf{M}=\frac{diag\left\{2h_1,\ldots,2h_m,\mathbf{0}_{1\times(2n-2m)}\right\}}{\omega_0} \tag{35}$$

$$\mathbf{D}=\frac{diag\left\{d_{G_1},\ldots,d_{G_m},\mathbf{0}_{1\times(2n-2m)}\right\}}{\omega_0} \tag{36}$$

$$\mathbf{L}^V=diag\left\{\mathbf{0}_{1\times m},2G_{mm+1}V_{m+1},\ldots,2G_{nn}V_n,\mathbf{0}_{1\times(n-m)}\right\} \tag{37}$$

$$\mathbf{L}^{G\delta}=\left[\left(\mathbf{L}_1^{G\delta 1}\right)^T,\ldots,\left(\mathbf{L}_n^{G\delta 1}\right)^T,\left(\mathbf{L}_{n+1}^{G\delta 2}\right)^T,\ldots,\left(\mathbf{L}_{2n-m}^{G\delta 2}\right)^T\right]^T \tag{38}$$

$$\mathbf{L}_i^{G\delta 1}=\begin{bmatrix}\frac{\partial L^G_{i,1}}{\partial\delta_i} & \cdots & \frac{\partial L^G_{i,2n-m}}{\partial\delta_i}\end{bmatrix}\quad i=1,\ldots,n \tag{39}$$

$$\mathbf{L}_i^{G\delta 2}=\begin{bmatrix}\frac{\partial L^G_{i,1}}{\partial\delta_{i-n+m}} & \cdots & \frac{\partial L^G_{i,2n-m}}{\partial\delta_{i-n+m}}\end{bmatrix}\quad i=n+1,\ldots,2n-m \tag{40}$$

The length of vectors $\mathbf{L}_i^{G\delta 1}$ and $\mathbf{L}_i^{G\delta 2}$ are $2n-m$. These vectors are calculated using the chain differential rule. According to this rule, the differential in terms of δ is obtained in terms of θ as follows:

$$\frac{\partial}{\partial\delta_i}=\sum_{k=1}^{l}\frac{\partial}{\partial\theta_k}\frac{\partial\theta_k}{\partial\delta_i}=\sum_{k=1}^{l}H^T(i,k)\times\frac{\partial}{\partial\theta_k} \tag{41}$$

Consequently, $\mathbf{L}_i^{G\delta 1}$ and $\mathbf{L}_i^{G\delta 2}$ are obtained as follows:

$$\mathbf{L}_i^{G\delta 1}=\mathbf{H}^T(i,:)\left(\sum_{k=1}^{l}H^T(i,k)\times\mathbf{L}_k^{Gq}\right)\mathbf{H} \tag{42}$$

$$\mathbf{L}_i^{G\delta 2}=\mathbf{H}^T(i,:)\left(\sum_{k=1}^{l}H^T(i-n+m,k)\times\mathbf{L}_k^{Gq}\right)\mathbf{H}+\mathbf{L}_i^{Gdiag\_\delta} \tag{43}$$

$\mathbf{L}_i^{Gq}$ and $\mathrm{L}_{\mathrm{i}}^{\mathrm{Gdiag\_}\delta}$ are the $2l\times 2l$ matrix and the $1\times 2(n-m)$ vector as:

$$\left[L_k^{Gq}\right]_{i,j}=\begin{cases}G_kF_k & i=j=k\\ -G_kE_k & i=j-l=k\\ -G_kE_k & i=j+l=k+l\\ -G_kF_k & i=j=k+l\\ 0 & \text{otherwise}\end{cases} \tag{44}$$

$$\left[L_i^{Gdiag\_\delta}\right]_{1,j}=\begin{cases}\frac{\sum_{k=1}^{l}A(i-n+m,k)G_kF_k}{V^2_{i-n+m}} & i=j\\ 0 & \text{otherwise}\end{cases} \tag{45}$$

QPM is characterized as (46). The i^{th} eigenvalue of the system λ_i, and the corresponding right and left eigenvectors, $\mathbf{X}$ and $\mathbf{W}$, are calculated by (47) and (48), respectively.

$$\mathbf{Q}(\lambda_i)=\mathbf{M}\lambda_i^2+\mathbf{D}\lambda_i+\mathbf{L}^V+\mathbf{L}^{G\delta}+\mathbf{L}^B+\mathbf{L}^{Pbus}+\mathbf{L}^{Qbus} \tag{46}$$

$$\mathbf{Q}(\lambda_i)\times\mathbf{X}=\mathbf{0}_{(2n-m)\times 1} \tag{47}$$

$$\mathbf{W}^T\times\mathbf{Q}(\lambda_i)=\mathbf{0}_{1\times(2n-m)} \tag{48}$$

Entries of the right and left eigenvectors are defined by (49) and (50), respectively. In the right eigenvector, entries x_1 to x_n and x_{n+1} to x_{2n-m} are related to voltage angle state variables and amplitude of load bus voltage state variables, respectively. Similarly, this is true for the entries of the left eigenvector.

$$\mathbf{X}=\begin{bmatrix}x_{\delta_1} & \cdots & x_{\delta_n} & x_{V_{m+1}} & \cdots & x_{V_n}\end{bmatrix}^T$$
$$=\begin{bmatrix}x_1 & \cdots & x_{2n-m}\end{bmatrix}^T \tag{49}$$

$$\mathbf{W}=\begin{bmatrix}w_{\delta_1} & \cdots & w_{\delta_n} & w_{V_{m+1}} & \cdots & w_{V_n}\end{bmatrix}^T$$
$$=\begin{bmatrix}w_1 & \cdots & w_{2n-m}\end{bmatrix}^T \tag{50}$$

Sensitivity Factors of the Electromechanical Mode

To calculate the sensitivities of eigenvalues concerning state variables, both sides of (47) is pre-multiplied by the left eigenvector $\mathbf{W}^T$, and the differential is taken from the resultant equation. By doing so, the following equations

is obtained:

$$d\lambda_i = -\frac{\mathbf{W}^T d\left(\mathbf{L}^{total}\right)\mathbf{X}}{\alpha} \tag{51}$$

where:

$$\mathbf{L}^{total} = \mathbf{L}^{B} + \mathbf{L}^{Pbus} + \mathbf{L}^{Qbus} + \mathbf{L}^{V} + \mathbf{L}^{G\delta} \tag{52}$$

$$\alpha = 2\lambda_i \mathbf{W}^T \mathbf{M}\mathbf{X} + \mathbf{W}^T \mathbf{D}\mathbf{X} \tag{53}$$

As can be seen in (51), the denominator is a constant value, whereas the numerator includes the differential of matrix $\mathbf{L}^{total}$ that comprises five differential terms. The following subsections will provide corresponding formulae to calculate them.

C. Calculation of $\boldsymbol{W}^T d(\boldsymbol{L}^B)\boldsymbol{X}$

For the sake of simplicity, right and left eigenvectors are defined in terms of new state variables as follows [27]:

$$\mathbf{X}' = \left[x_{\theta_1},...,x_{\theta_l},x_{\upsilon_1},...,x_{\upsilon_l}\right]^T = \mathbf{H}\times\mathbf{X} \tag{54}$$

$$\mathbf{W}' = \left[w_{\theta_1},...,w_{\theta_l},w_{\upsilon_1},...,w_{\upsilon_l}\right]^T = \mathbf{H}\times\mathbf{W} \tag{55}$$

Using the above equations, $\boldsymbol{W}^{\mathrm{T}}\mathrm{d}(\boldsymbol{L}^{\mathrm{B}})\boldsymbol{X}$ is determined in the following:

$$\mathbf{W}^T d\mathbf{L}^B\mathbf{X} = \sum_{k=1}^{l} B_k \begin{Bmatrix} \left(w_{\upsilon_k}x_{\upsilon_k} - w_{\theta_k}x_{\theta_k} - C_{L_k}\right)F_k \\ +\left(w_{\theta_k}x_{\upsilon_k} + w_{\upsilon_k}x_{\theta_k}\right)E_k \end{Bmatrix} d\theta_k$$
$$+\sum_{i=m+1}^{n}\sum_{k=1}^{l}\left|A(i,k)\right| B_k \left\{ \begin{bmatrix} -\left(w_{\upsilon_k} - w_{V_i}^{\ln}\right)\times \\ \left(x_{\upsilon_k} - x_{V_i}^{\ln}\right) \\ +C_{L_k} + w_{\theta_k}x_{\theta_k} \\ -w_{V_i}^{\ln}x_{V_i}^{\ln} \end{bmatrix} E_k + \begin{bmatrix} w_{\theta_k}\left(x_{\upsilon_k} - x_{V_i}^{\ln}\right) \\ +x_{\theta_k}\left(w_{\upsilon_k} - w_{V_i}^{\ln}\right) \end{bmatrix} F_k \right\} dV_i^{\ln} \tag{56}$$

where:

$$dV_i^{\ln} = \frac{dV_i}{V_i} \tag{57}$$

$$w_{V_i}^{\ln} = \frac{w_{V_i}}{V_i} \tag{58}$$

$$x_{V_i}^{\ln} = \frac{x_{V_i}}{V_i} \tag{59}$$

$$C_{L_k} = \sum_{i=m+1}^{n}\left|A(i,k)\right|\left(w_{V_i}^{\ln}x_{V_i}^{\ln}\right) \tag{60}$$

D. Calculation of $\boldsymbol{W}^T d(\boldsymbol{L}^{Pbus})\boldsymbol{X}$

After conducting some calculations, the following equation is derived.

$$\mathbf{W}^T d\mathbf{L}^{Pbus}\mathbf{X} = -2\sum_{i=m+1}^{n} P_{i_0}\left(\frac{V_i}{V_{i_0}}\right)^2 k_{Pz}w_{\delta_i}x_{V_i}^{\ln}dV_i^{\ln} \tag{61}$$

E. Calculation of $\boldsymbol{W}^T d(\boldsymbol{L}^{Qbus})\boldsymbol{X}$

$\mathbf{W}^T d(\mathbf{L}^{Qbus})\mathbf{X}$ is equal to:

$$\mathbf{W}^T d\mathbf{L}^{Qbus}\mathbf{X} = -2\sum_{i=m+1}^{n} Q_{i_0}k_{QC}w_{V_i}^{\ln}x_{V_i}^{\ln}dV_i^{\ln} \tag{62}$$

F. Calculation of $\boldsymbol{W}^T d(\boldsymbol{L}^{V})\boldsymbol{X}$

$\mathbf{W}^T d(\mathbf{L}^{V})\mathbf{X}$ is obtained as follows:

$$\mathbf{W}^T d\mathbf{L}^{V}\mathbf{X} = 2\sum_{i=m+1}^{n} G_{ii}x_{\delta_i}w_{\delta_i}V_i dV_i^{\ln} \tag{63}$$

G. Calculation of $\boldsymbol{W}^T d\left(\boldsymbol{L}^{G\delta}\right)\boldsymbol{X}$

$\mathbf{L}^{G\delta}$ consists of sub-matrices $\mathbf{L}^{G\delta 1}\mathrm{L}_{\mathrm{i}}^{\mathrm{G\delta 1}}$ and $\mathbf{L}^{G\delta 2}$. Therefore $\boldsymbol{W}^T d\left(\boldsymbol{L}^{G\delta}\right)\boldsymbol{X}$ is obtained as shown in the following:

$$\mathbf{W}^T d\mathbf{L}^{G\delta}\mathbf{X} = \sum_{i=1}^{n} w_{\delta_i}d\mathbf{L}_i^{G\delta 1}\mathbf{X} + \sum_{i=n+1}^{2n-m} w_{V_{i-n+m}}d\mathbf{L}_i^{G\delta 2}\mathbf{X} \tag{64}$$

where:

$$w_{\delta_i}d\mathbf{L}_i^{G\delta 1}\mathbf{X} = w_{\delta_i}\sum_{k=1}^{l} G_k\left|A(i,k)\right|\left\{x_{\theta_k}E_k + x_{\upsilon_k}F_k\right\}d\theta_k +$$
$$w_{\delta_i}\sum_{j=m+1}^{n}\sum_{k=1}^{l} G_k\left|A(i,k)A(j,k)\right|\left\{\begin{pmatrix} x_{\theta_k}F_k \\ -\left(x_{\upsilon_k} - x_{V_j}^{\ln}\right)E_k \end{pmatrix}\right\}dV_j^{\ln} \tag{65}$$

$$w_{V_{i-n+m}}d\mathbf{L}_i^{G\delta 2}\mathbf{X} =$$
$$w_{V_{i-n+m}}^{\ln}\sum_{k=1}^{l} A(i-n+m,k)G_k\begin{pmatrix} x_{\theta_k}E_k + \\ \left(x_{\upsilon_k} - 2x_{V_{i-n+m}}^{\ln}\right)F_k \end{pmatrix}dV_{i-n+m}^{\ln}$$
$$+w_{V_{i-n+m}}^{\ln}\sum_{k=1}^{l} A(i-n+m,k)G_k\begin{pmatrix} x_{\theta_k}F_k - \\ \left(x_{\upsilon_k} - x_{V_{i-n+m}}^{\ln}\right)E_k \end{pmatrix}d\theta_k \tag{66}$$
$$-w_{V_{i-n+m}}^{\ln}\sum_{j=m+1}^{n}\left\{\sum_{k=1}^{l}\begin{pmatrix} \left|A(j,k)\right|\times A(i-n+m,k)\times \\ G_k\begin{pmatrix} x_{\theta_k}E_k + \\ \left(x_{\upsilon_k} - x_{V_j}^{\ln} - x_{V_{i-n+m}}^{\ln}\right)F_k \end{pmatrix} \end{pmatrix}\right\}dV_j^{\ln}$$

H. Calculation of Sensitivity Factors

In the previous sections, the changes in the electromechanical mode were obtained according to the changes in the state variables. With respect to (34), in steady-state condition, changes in state variables can be calculated in terms of power changes of generators using (67).

Because net active power injection of slack bus is included in (34), pseudo inverse of $\mathbf{L}^{total}$ matrix i.e. $(\mathbf{L}^{total})^{\dagger}$ is used

$$d\mathbf{z}^T = \left(\mathbf{L}^{total}\right)^{\dagger} d\left[P_{G_1},...,P_{G_m},\mathbf{0}_{1\times 2(n-m)}\right]^T \tag{67}$$

By integrating all obtained equations in matrix form, $d\lambda_i$ in (34) can be expressed in terms of generators active power as follows:

$$\Delta\lambda_i = \frac{-1}{\alpha}\left[S_{\theta_1},...,S_{\theta_l},S_{V_{m+1}},...,S_{V_n}\right]\times \begin{bmatrix}\mathbf{A}^T & \mathbf{0}_{l\times(n-m)}\\ \mathbf{0}_{(n-m)\times l} & \mathrm{I}_{n-m}\end{bmatrix}\times \left(\mathbf{L}^{total}\right)^{\dagger} d\left[P_{G_1},...,P_{G_m},\mathbf{0}_{2(n-m)\times 1}\right] = \left[S_{P_1} \quad ... \quad S_{P_m}\right] d\left[P_{G_1},...,P_{G_m}\right]^T \tag{68}$$

where S_{P_1} to S_{P_m} are sensitivity factors of the electromechanical mode related to variations of generators active power.

If the critical electromechanical mode is $\lambda_i = \sigma + j\omega$ then its damping-ratio is defined as follows:

$$\zeta = \frac{-\sigma}{\sqrt{\sigma^2+\omega^2}} \tag{69}$$

Partial differential equation of the electromechanical mode damping-ratio with respect to its arguments i.e. σ and ω is equal to:

$$d\zeta = -\frac{\omega}{|\lambda|^3}\left(\omega d\sigma - \sigma d\omega\right) \tag{70}$$

Separating imaginary and real portions of (68) and substituting those into (70) yields:

$$d\zeta = -\frac{\omega}{|\lambda|^3}\left(\sum_i^m \begin{pmatrix}\omega\times real\left[S_{P_i}\right]\\ -\sigma\times imag\left[S_{P_i}\right]\end{pmatrix} dP_{G_i}\right) \tag{71}$$

Finally by defining new index Sen_{damp_i}, (71) can be rewritten as:

$$d\zeta = \sum_i^m Sen_{damp_i}\, dP_{G_i} \tag{72}$$

where:

$$Sen_{damp_i} = -\frac{\omega}{|\lambda|^3}\times\left(\omega\times real\left[S_{P_i}\right] - \sigma\times imag\left[S_{P_i}\right]\right) \tag{73}$$

Sen_{damp_i} index represents the damping-ratio sensitivity with respect to active power changes of i^{th} generator. The larger absolute value of Sen_{damp_i}, the effect of changes in the active power of i^{th} generator on variation of the electromechanical mode damping-ratio is greater. Therefore, the generators can be sorted according to the damping-ratio sensitivity factor values, and the two generators that have the largest difference are selected for applying generation redispatch.

Generators Redispatch Plan

Fig. 3 shows the proposed iterative plan for applying generators redispatch remedial action. At first, the initial operating points of generators active power are received from the load flow of the power system. The network losses are determined and its deviation is set to zero. Next, generators are redispatched using sensitivity coefficients to satisfy the permissible damping-ratio limit DR_{sch}. To take economic considerations into account, since redispatching imposes additional costs on the operation of the system, variations in generators active power must be minimized.

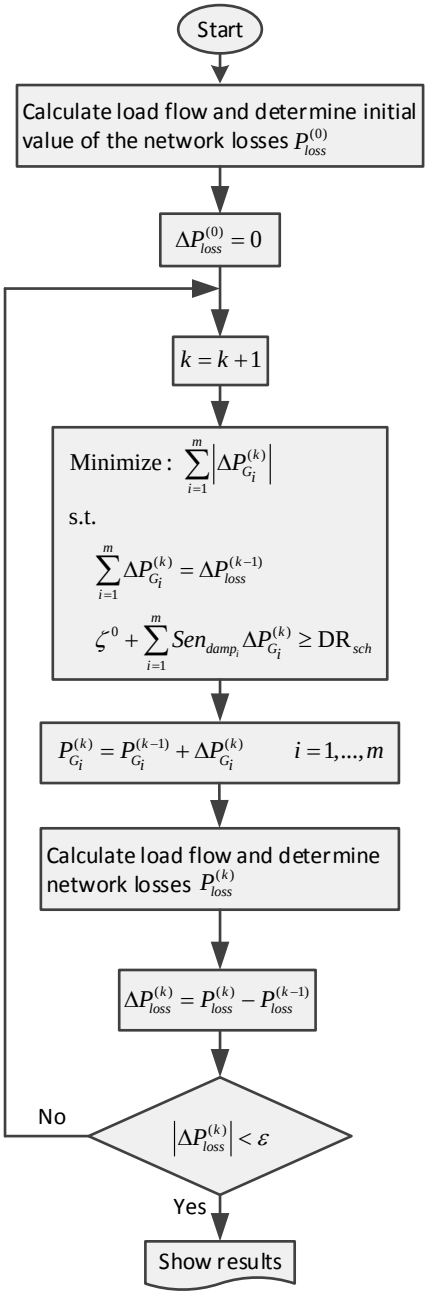


Fig. 3: Overview of the proposed generators redispatch scheme.

The greater the damping-ratio sensitivity factor associated with a generator, its active power changes lead to greater variations in the power system damping. Accordingly, generators with larger damping-ratio sensitivity factor are selected for redispatching in the proposed algorithm to satisfy the constraint of minimum power system damping with the least variations of

generators active power. Furthermore, since the resistance of the transmission lines is included in the modeling, the changes of network active power losses are considered in the generation redispatches to satisfy the balance of production and consumption in the system. To do so, the total active power variations of the generators must be equal to the variation of the network losses which is obtained from the previous iteration. After the optimization stage, the operating point of active power generators is updated and load flow is executed again. The amount of power network losses is calculated as the difference between the output active power of the generators and the consumption power of the loads. If the absolute value of the difference in power network losses between two consecutive steps is less than ε, the iteration process ends. Otherwise, the value of ΔP_{loss} is updated, and the process returns to the optimization step.

The bottleneck of real-world case studies is the need of transmission lines, synchronous generators, and load model data for modeling the proposed method. The exact amount of these data is usually not available and therefore a pre-processing step is needed to estimate them.

Results and Discussion

In this section, the IEEE 39-bus test power system [40] shown in Fig. 4 is applied to evaluate the proposed formulae for determining the sensitivities of an electromechanical mode. The classical model of synchronous generators is used to investigate the effect of the load model and transmission lines resistance on the value of sensitivities. Two combinations of ZIP load model factors are tested and sensitivity factors along with optimal generators rediptach scheme are calculated.

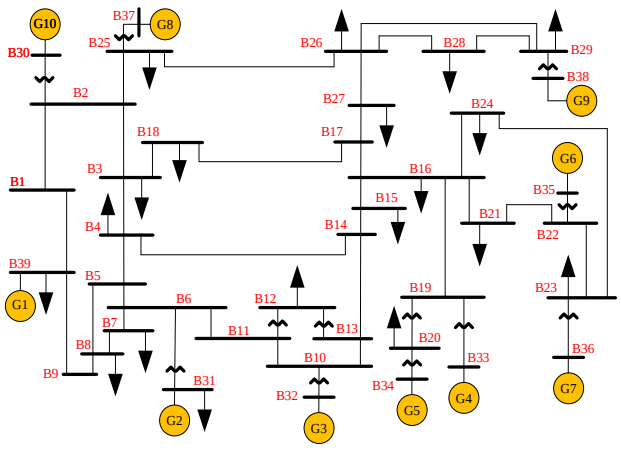


Fig. 4: IEEE 39-bus test power system.

I. Constant Power Load

The test power system has an inter-area mode which is $-0.0014 + 4.2218i$. For this oscillatory mode, generators 2 to 10 oscillate against generators 1. Table 1 lists sensitivity factors of the inter-area mode related to generators active power changes with considering and also neglecting the resistance of transmission lines. It is evident that changes in the active power of generators only influence the frequency of the inter-area mode and do not affect its real part. Nonetheless, changes in the frequency of the oscillatory mode, while keeping the real part constant, result in variations of its damping-ratio.

Table 1: Sensitivity factors corresponding to the electromechanical mode of interest considering constant power loads

Sensitivities	Considering transmission lines resistance	Neglecting transmission lines resistance
S_{P1}	$3.7\times10^{-5}+0.1063i$	$-9.7\times10^{-7}+0.0407i$
S_{P2}	$-4.68\times10^{-5}-0.2612i$	$4.3\times10^{-6}-0.0513i$
S_{P3}	$-3.99\times10^{-5}-0.0795i$	$4.6\times10^{-7}-0.0065i$
S_{P4}	$-7.91\times10^{-5}-0.166i$	$-6.8\times10^{-6}-0.0320i$
S_{P5}	$-7.99\times10^{-5}-0.162i$	$-9.8\times10^{-6}-0.0322i$
S_{P6}	$-5.58\times10^{-5}-0.2656i$	$4.1\times10^{-6}-0.0726i$
S_{P7}	$-6.12\times10^{-5}-0.1483i$	$-9.3\times10^{-7}-0.0321i$
S_{P8}	$-1.04\times10^{-4}-0.1714i$	$-8.2\times10^{-6}-0.0204i$
S_{P9}	$-1.2\times10^{-4}-0.2089i$	$-8.4\times10^{-6}-0.0228i$
S_{P10}	$8.9\times10^{-6}+0.0269i$	$-2.2\times10^{-6}+0.0098i$

Table 2 depicts Sen_{damp} indices. As seen in this table, their value with considering and neglecting transmission lines resistance are totally different. Moreover, Sen_{damp_2} and Sen_{damp_6} in the condition of considering transmission lines resistance are antiphase with their values in the condition of neglecting the resistances. Based on the calculated sensitivity factors considering transmission lines resistance, it is concluded that increasing the power of generators 2 and 6 lead to improve power system damping. However, neglecting lines resistance leads to an incorrect inference.

Table 3 arranges the values of the damping-ratio sensitivities in descending order. Let the redispaching plan is applied using a pair of generators (*i*,*j*) to improve power system damping.

Table 2: Electromechanical mode damping-ratio sensitivity factors corresponding to the electromechanical mode of interest considering constant power loads

No.	Considering transmission lines resistance	Neglecting transmission lines resistance
1	-1.71×10^{-5}	-0.29×10^{-5}
2	3.11×10^{-5}	-0.638×10^{-5}
3	1.55×10^{-5}	0.039×10^{-5}
4	3.14×10^{-5}	0.408×10^{-5}
5	3.13×10^{-5}	0.482×10^{-5}
6	3.35×10^{-5}	-0.417×10^{-5}
7	2.58×10^{-5}	0.269×10^{-5}
8	3.78×10^{-5}	0.353×10^{-5}
9	4.58×10^{-5}	0.376×10^{-5}
10	-0.418×10^{-5}	-0.023×10^{-5}

Table 3: Priority list of generators redispatches considering constant power loads

Rank	Considering transmission lines resistance	Neglecting transmission lines resistance
1	G9	G5
2	G8	G4
3	G6	G9
4	G4	G8
5	G5	G7
6	G2	G3
7	G7	G10
8	G3	G1
9	G10	G6
10	G1	G2

In the model of neglecting transmission lines resistance, the active power changes of both generators must be equal and in opposite directions to satisfy load balance constraint. Therefore, the damping-ratio variation is obtained as follows:

$$d\zeta = \left(Sen_{damp_i} - Sen_{damp_j}\right) dP_{G_i}$$

The greater the difference in the ranks of the selected generators, the absolute value of $(Sen_{damp_i} - Sen_{damp_j})$ is greater, and therefore, redispatching the generators pair has more effect on improving the electromechanical mode damping-ratio. It can be inferred from Table 3 that if transmission lines resistance is neglected, generators pair 5 and 2 are the best options for redispatching. To do this, the active power of generator 5 must be increased, while the active power of generator 2 must be decreased. The active power changes of the pair of generators are not equal when the resistance of transmission lines is considered. Nonetheless, the best choice for redispatching the pair of generators that have the greatest difference in their damping sensitivity coefficients. To do this, the active power of generator 9 must be increased, while the active power of generator 1 must be decreased. The proposed algorithm shown in Fig. 3 must be applied to determine new operating point of the generators pair active power.

In order to execute generators rediptach optimization problem which is shown in Fig. 2, it is assumed that generators can only change their initial power by 50%. Fig. 5 depicts the results of the proposed strategy with considering transmission lines resistance. The aim is to apply optimal generators redispatch to improve the damping-ratio of the mode of interest by 50%. The base power is 100 MVA. Active power of generator 9 is increased and generator 1 is decreased. The sum of absolute values of generators active power changes is equal to 5.09 pu.

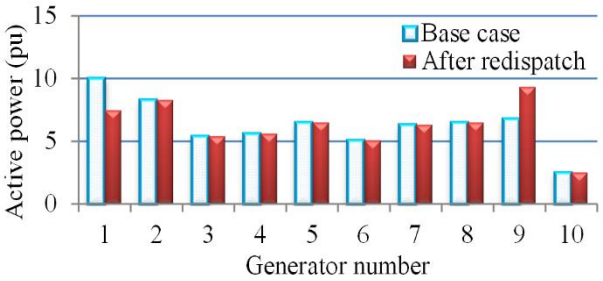


Fig. 5: Optimal generators redispatches plan considering constant power loads with considering transmission lines resistance.

It is important to emphasize that there is no feasible solution for achieving the goal when ignoring the resistance of transmission lines. Therefore, it is concluded that considering transmission lines resistance has a significant impact on the accuracy of sensitivities when the classical model of synchronous generators is used. In order to show quantitative insights into the effect of transmission lines resistance, it is assumed that the aim is applying optimal generators redispatch to improve the damping-ratio of the mode of interest by 25% of the initial value.

Table 4 compares required variations of generators active power with considering and neglecting transmission lines resistance. As seen in this table, sum of absolute values of generators active power changes with considering and neglecting transmission lines resistance are equal to 4.1 pu and 17.44 pu, respectively. Therefore, it is concluded that modeling the resistance of transmission lines has a significant impact on determining the pattern and amplitude of generators redispatch.

Table 4: Generators active power variations

Gen.	Considering transmission lines resistance	Neglecting transmission lines resistance
1	-2.03 pu	-2.04 pu
2	0	-4.14 pu
3	0	0
4	0	2.8 pu
5	0	3.24 pu
6	0	-2.54 pu
7	0	0
8	0	0
9	2.07 pu	2.68 pu
10	0	0
Sum of absolute deviations	4.1 pu	17.44 pu

J. Constant Current Load

To simulate constant current loads, the coefficients of ZIP load model in (4) and (5) are set as follows:

$$\begin{cases} k_{PI} = k_{QI} = 1 \\ k_{PC} = k_{PZ} = k_{QC} = k_{QZ} = 0 \end{cases}$$

In the corresponding modeling, the mode of interest is obtained as $-0.0014 + 4.0019i$. Comparing the calculated value of the electromechanical mode with its value in the condition of all loads are constant power type, it is concluded its frequency is decreased and its real part is not changed.

Tables 5 and 6 depict sensitivity factors and Sen_{damp} indices where all loads are constant current type. The priority list of generators active power redispatches for enhancing damping-ratio of the electromechanical mode is shown in Table 6.

Comparing Tables 2 and 6, it is inferred that the type of load model affects damping-ratio sensitivities. For example, the sign of Sen_{damp} index for generators 2 and 3 in the condition of constant current load with considering transmission lines resistance are different from ones in the condition of constant power load. Therefore, increasing active powers of generators 2 and 3 leads to improve damping-ratio of the electromechanical mode in condition of all loads are constant power, although these generators redispatches attenuate the mode damping.

Now, let's neglect transmission lines resistance, if all loads are constant power type, concerning the results shown in the right column of Table 3, redispatching generators pair 5 and 2 has the greatest effect on the variation of the mode damping-ratio.

Table 5: Sensitivity factors corresponding to the electromechanical mode of interest considering constant current loads

Sensitivities	Considering transmission lines resistance	Neglecting transmission lines resistance
S_{P1}	$7.27\times10^{-5} + 0.1588i$	$6.96\times10^{-6} + 0.0618i$
S_{P2}	$3.23\times10^{-5} - 0.0839i$	$4.18\times10^{-5} - 0.04004i$
S_{P3}	$1.76\times10^{-5} + 0.0393i$	$4.7\times10^{-7} + 0.01282i$
S_{P4}	$-1.46\times10^{-5} - 0.024i$	$-1.44\times10^{-5} - 0.0203i$
S_{P5}	$-1.64\times10^{-5} - 0.022i$	$-1.78\times10^{-5} - 0.0204i$
S_{P6}	$1.91\times10^{-5} - 0.0996i$	$2.97\times10^{-6} - 0.06801i$
S_{P7}	$-9.85\times10^{-7} - 0.0178i$	$-8.58\times10^{-6} - 0.023i$
S_{P8}	$-2.66\times10^{-5} - 0.0046i$	$-1.079\times10^{-5} + 0.0038i$
S_{P9}	$-3.8\times10^{-5} - 0.0182i$	$-1.08\times10^{-5} + 0.0042i$
S_{P10}	$4.92\times10^{-5} + 0.1004i$	$-6.1\times10^{-7} + 0.0269i$

Table 6: Electromechanical mode damping-ratio sensitivity factors corresponding to the electromechanical mode of interest considering constant current loads

No.	Considering transmission lines resistance	Neglecting transmission lines resistance
1	-3.22×10^{-5}	-0.734×10^{-5}
2	-0.064×10^{-5}	-0.691×10^{-5}
3	-0.789×10^{-5}	-0.128×10^{-5}
4	0.582×10^{-5}	0.5488×10^{-5}
5	0.607×10^{-5}	0.634×10^{-5}
6	0.402×10^{-5}	-0.133×10^{-5}
7	0.1822×10^{-5}	0.4244×10^{-5}
8	0.706×10^{-5}	0.2367×10^{-5}
9	1.109×10^{-5}	0.2343×10^{-5}
10	-2.12×10^{-5}	-0.2287×10^{-5}

However, in the condition where all loads are constant current type, with respecting the right column of Table 7, redispatching generators pair 5 and 1 has the greatest effect.

Therefore, the best pair of generators for redispatches is changed by varying the load model.

Table 7: Priority list of generators redispatches considering constant current loads

Rank	Considering transmission lines resistance	Neglecting transmission lines resistance
1	G9	G5
2	G8	G4
3	G5	G7
4	G4	G8
5	G6	G9
6	G7	G3
7	G2	G6
8	G3	G10
9	G10	G2
10	G1	G1

Fig. 6 shows the results of the proposed strategy for applying optimal generators redispatch to improve the mode of interest damping-ratio by 50%, in the condition that all leads are constant current type. As seen in this figure, active powers of generators 8 and 9 increase while that of generator 1 decreases. The sum of absolute values of generators active power changes is equal to 8.38 pu, while it is equal to 5.09 pu in the condition that all loads are constant power type. Furthermore, generators 1, 8, and 9 are used for the optimal redispatching plan where all loads are constant current, while only generators 1 and 9 participate in the optimal redispatches plan where all loads are constant power. Therefore, the load model plays an outstanding role in determining sensitivity factors and the amplitude of generation redispatches, while it does not have a fundamental effect on changing the pattern of generators redispatches.

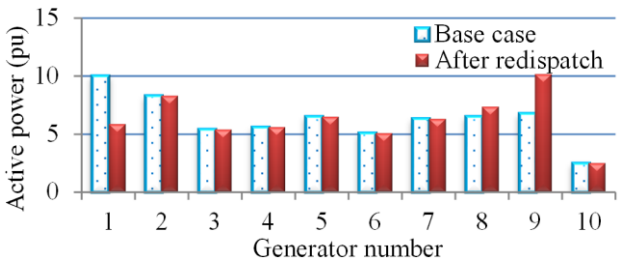


Fig. 6: Optimal generators redispatches plan considering constant current loads with considering transmission lines resistance.

Conclusion

This paper introduces a novel formula for determining the sensitivity of the damping-ratio of oscillatory modes with respect to active power of generators considering ZIP load model and resistance of power network transmission lines. The proposed formulae incorporate the inertia and damping constants of generators, transmission line parameters, and both magnitude and angle of bus voltages. Since power system losses change with redispatches of generators, the iterative algorithm was suggested to achieve the optimal redispatch plan when transmission lines resistance is modeled. The effectiveness of the proposed method for enhancing the damping-ratio of electromechanical modes is demonstrated through simulations of the 39-bus New England test system. The results show that load model and also considering transmission lines resistance are effective to determine the best generators redispatch scheme.

This study serves as a preliminary exploration of the calculating damping-ratio sensitivities considering transmission lines resistance. Future research should be conducted considering more accurate generator models to calculate the precise effects of corrective actions on the damping changes of the system using the proposed model. Therefore, extending the proposed formula to encompass more comprehensive synchronous generator dynamic model, and power system security constraints in the generator redispatch scheme are future research issues.

Nowadays, most loads are connected to the electric network through inverters. Therefore, Modeling of WECC composite load model by the proposed formulae must be examined. Furthermore, considering the proliferation of renewable energy resources (RES) and their role in power system stability, the development of the proposed formulae for modeling RES to identify optimal corrective actions are important topics for future researches.

Author Contributions

M. Setareh: Conceptualization, Investigation, Methodology, Simulation, Validation, Formal analysis, Writing manuscript.

A. Moradibirgani: Simulation, Formal analysis, Writing, Review & Editing.

Acknowledgment

We sincerely thank the respected referees for their accurate review of this paper.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

GR	Generator Redispatch
HVDC	High Voltage Direct Current

PSS	Power System Stabilizer
FACTS	Flexible Alternating Current Transmission Systems
SVC	Static Var Compensator
STATCOM	Static Compensators
TCSC	thyristor Controlled Series Capacitor
OPF	Optimal Power Flow
PMU	Phasor Measurement Units
QEP	Quadratic Eigenvalue Problem
NPF	Normalized Participation Factors
DAE	Dynamic-Algebraic Equations
PU	Per Unit
QPM	Quadratic Polynomial Matrix
RES	renewable Energy Resources

References

[1] M. J. Gibbard, P. Pourbeik, D. J. Vowles, Small-signal stability, control and dynamic performance of power systems, University of Adelaide press, 2015.

[2] N. Hatziargyriou et al., "Definition and classification of power system stability - revisited & extended," IEEE Trans. Power Syst., 36(4): 3271-3281, 2021.

[3] P. M. Anderson, A. A. Fouad, Power system control and stability, second edition. John Wiley & Sons, 2002.

[4] M. G. Jolfaei, A. M. Sharaf, S. M. Shariatmadar, M. B. Poudeh, "A hybrid PSS-SSSC GA-stabilization scheme for damping power system small signal oscillations," Int. J. Electr. Power Energy Syst., 75: 337-344, 2016.

[5] D. Wang, N. Ma, M. Wei, Y. Liu, "Parameters tuning of power system stabilizer PSS4B using hybrid particle swarm optimization algorithm," Int. Trans. Electr. Energy Syst., 28(9): e2598, 2018.

[6] M. Tadj et al., "Improved chaotic Bat algorithm for optimal coordinated tuning of power system stabilizers for multimachine power system," Sci. Rep., 14(1): 15124, 2024.

[7] A. T. Moghadam, M. Aghahadi, M. Eslami, S. Rashidi, B. Arandian, S. Nikolovski, "Adaptive rat swarm optimization for optimum tuning of svc and pss in a power system," Int. Trans. Electr. Energy Syst., 2022, 4798029, 2022.

[8] J. Zhou, D. Ke, C. Y. Chung, Y. Sun, "A computationally efficient method to design probabilistically robust wide-area PSSs for damping inter-area oscillations in wind-integrated power systems," IEEE Trans. Power Syst., 33(5): 5692-5703, 2018.

[9] T. Prakash, V. P. Singh, S. R. Mohanty, "A synchrophasor measurement based wide-area power system stabilizer design for inter-area oscillation damping considering variable time-delays," Int. J. Electr. Power Energy Syst., 105: 131-141, 2019.

[10] H. Zhao, Z. Lin, Q. Wu, S. Huang, "Model predictive control based coordinated control of multi-terminal HVDC for enhanced frequency oscillation damping," Int. J. Electr. Power Energy Syst., 123, 106328, 2020.

[11] W. Gao, R. Fan, R. Huang, Q. Huang, W. Gao, L. Du, "Augmented random search based inter-area oscillation damping using high voltage DC transmission," Electr. Power Syst. Res., 216, 109063, 2023.

[12] R. Huang, W. Gao, R. Fan, Q. Huang, "A guided evolutionary strategy based-static var compensator control approach for interarea oscillation damping," IEEE Trans. Ind. Informatics, 19(3): 2596-2607, 2023.

[13] W. Peres, N. N. da Costa, "Comparing strategies to damp electromechanical oscillations through STATCOM with multi-band controller," ISA Trans., 107: 256-269, 2020.

[14] R. Huang, W. Gao, R. Fan, Q. Huang, "Damping inter-area oscillation using reinforcement learning controlled TCSC," IET Gener. Transm. Distrib., 16(11): 2265-2275, 2022.

[15] N. Nikolaev, K. Dimitrov, Y. Rangelov, "A comprehensive review of small-signal stability and power oscillation damping through photovoltaic inverters," Energies, 14(21): 7372, 2021.

[16] J. L. Rodriguez-Amenedo, S. A. Gomez, "Damping low-frequency oscillations in power systems using grid-forming converters," IEEE Access, 9: 158984–158997, 2021.

[17] W. Li, D. Cai, S. Wu, G. Zhang, F. Zhang, "The impact of supplementary active power control of wind turbine on power system low frequency oscillations," Electr. Power Syst. Res., 224, 109746, 2023.

[18] C. Yan, W. Yao, J. Wen, J. Fang, X. Ai, J. Wen, "Optimal design of probabilistic robust damping controllers to suppress multiband oscillations of power systems integrated with wind farm," Renew. Energy, 158: 75-90, 2020.

[19] M. Basu, J. Kim, R. M. Nelms, E. Muljadi, "Coordination of utility-scale PV plant and wind power plant in interarea-oscillation damping," IEEE Trans. Ind. Appl., 59(4): 4744-4751, 2023.

[20] M. Hojabri, U. Dersch, A. Papaemmanouil, P. Bosshart, "A comprehensive survey on phasor measurement unit applications in distribution systems," Energies, 12(23): 4552, 2019.

[21] A. D. Femine, D. Gallo, C. Landi, M. Luiso, "A design approach for a low cost phasor measurement unit," in Proc. 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC): 1-6, 2019.

[22] J. Giri, "Proactive management of the future grid," IEEE Power Energy Technol. Syst. J., 2(2): 243-52, 2015.

[23] J. E. Condren, T. W. Gedra, "Expected-security-cost optimal power flow with small-signal stability constraints," IEEE Trans. Power Syst., 21(4): 1736-1743, 2006.

[24] T. Weckesser, H. Johannsson, J. Ostergaard, "Real-time remedial action against aperiodic small signal rotor angle instability," IEEE Trans. Power Syst., 31(1): 387-396, 2016.

[25] R. Zárate-Miñano, F. Milano, A. J. Conejo, "An OPF methodology to ensure small-signal stability," IEEE Trans. Power Syst., 26(3): 1050-1061, 2011.

[26] L. Yazdani, M. R. Aghamohammadi, "Damping inter-area oscillation by generation rescheduling based on wide-area measurement information," Int. J. Electr. Power Energy Syst., 67: 138-151, 2015.

[27] S. Mendoza-Armenta, I. Dobson, "Applying a formula for generator redispatch to damp interarea oscillations using synchrophasors," IEEE Trans. Power Syst., 31(4): 3119-3128, 2016.

[28] T. J. M. A. Parreiras, S. Gomes, G. N. Taranto, K. Uhlen, "Closest security boundary for improving oscillation damping through generation redispatch using eigenvalue sensitivities," Electr. Power Syst. Res., 160: 119–127, 2018.

[29] H. Golzari-Kolur, S. M.-T. Bathaee, T. Amraee, "A sequential generation redispatch algorithm to ensure power system small signal stability under low-frequency oscillations," Int. Trans. Electr. Energy Syst., 2023: 1-15, 2023.

[30] M. Khaji, M. R. Aghamohammadi, "Emergency transmission line

switching to suppress power system inter-area oscillation," Int. J. Electr. Power Energy Syst., 87: 52-64, 2017.

[31] M. Setareh, M. Parniani, "Sensitivity-based optimal remedial actions to damp oscillatory modes considering security constraints," Int. J. Electr. Power Energy Syst., 135, 107580, 2022.

[32] Y. Li, G. Geng, Q. Jiang, W. Li, X. Shi, "A sequential approach for small signal stability enhancement with optimizing generation cost," IEEE Trans. Power Syst., 34(6): 4828-4836, 2019.

[33] H. Panahi, M. Abedini, M. Sanaye-Pasand, "Enhancing situation awareness by determining critical intra-area and interarea transmission lines," IEEE Syst. J., 17(4): 6192-6201, 2023.

[34] K. N. Hasan, R. Preece, J. V. Milanovic, "Priority ranking of critical uncertainties affecting small-disturbance stability using sensitivity analysis techniques," IEEE Trans. Power Syst., 32(4): 2629-2639, 2017.

[35] M. Setareh, M. Parniani, "Sensitivity-based generators redispatch to improve electromechanical mode damping considering transmission lines resistance," in Proc. 27th Iranian Conference on Electrical Engineering: 491-496, 2019.

[36] P. W. Sauer, M. A. Pai, J. H. Chow, Power System Dynamics and Stability: With Synchrophasor Measurement and Power System Toolbox 2e. Chichester, UK: John Wiley & Sons, Ltd, 2017.

[37] M. Bircan, A. Durusu, B. Kekezoglu, O. Elma, U. S. Selamogullari, "Experimental determination of ZIP coefficients for residential appliances and ZIP model based appliance identification: The case of YTU Smart Home," Electr. Power Syst. Res., 179, 106070, 2020.

[38] K. Veselić, Damped oscillations of linear systems: A mathematical introduction, in The Model: 1-13, Berlin: Springer Science & Business Media, 2011.

[39] F. Tisseur, K. Meerbergen, "The quadratic eigenvalue problem," SIAM Rev., 43(2): 235-286, 2001.

[40] I. Hiskens, "39-bus system (New England Reduced Model)," IEEE PES Task Force on Benchmark Systems for Stability Controls, Tech. Rep., PES TR18, 2013.

Biographies

Mohammad Setareh received the B.Sc. degree from Iran University of Science and Technology, Tehran, Iran, in 2011, the M.Sc. degree from University of Tehran, Tehran, Iran, in 2013, and the Ph.D. degree from Sharif University of Technology, Tehran, Iran, in 2019, all in electrical power engineering. He is currently an Assistant Professor with the Faculty of Electrical Engineering at Shahid Beheshti University, Tehran, Iran. His research interests include power system stability, optimal power management in smart grid, wide area monitoring and control, and restructuring in power systems.

- Email: m_setareh@sbu.ac.ir
- ORCID: 0000-0002-1278-9876
- Web of Science Researcher ID: AAF-1981-2019
- Scopus Author ID: 57202689177
- Homepage: https://cpe.sbu.ac.ir/~m_setareh

Alireza MoradiBirgani received the B.Sc. degree in Electrical Power Engineering from Shahid Chamran University of Ahvaz, Ahvaz, Iran, in 2023. He is currently working toward the M.Sc. degree with the Faculty of Electrical Engineering at Shahid Beheshti University, Tehran, Iran. His research interests include power systems, stability, power electronics, and electric vehicles.

- Email: a.moradibirgani@mail.sbu.ac.ir
- ORCID: 0009-0002-4801-1385
- Web of Science Researcher ID: KXQ-5530-2024
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
M. Setareh, A. Moradibirgani, "Damping critical electromechanical oscillations via generators redispatch considering ZIP load model and transmission lines resistance," J. Electr. Comput. Eng. Innovations, 13(2): 365-378, 2025.

DOI: 10.22061/jecei.2025.11106.766

URL: https://jecei.sru.ac.ir/article_2277.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Improved Correlation Coefficient Sparsity Adaptive Matching Pursuit in Noisy Condition

A. Vakili [1], *M. Shams Esfand Abadi* [1], *M. Kalantari* [2,*]

[1] *Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

[2] *Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

Article Info

Article History:
Received 14 December 2024
Reviewed 12 January 2025
Revised 18 January 2025
Accepted 29 January 2025

Keywords:
Compressed sensing (CS)
Greedy sparse recovery algorithm
Sparsity Adaptive Matching Pursuit
Sparsity estimation

*Corresponding Author's Email Address: *mkalantari@sru.ac.ir*

Abstract

Background and Objectives: In the realm of compressed sensing, most greedy sparse recovery algorithms necessitate former information about the signal's sparsity level, which may not be available in practical conditions. To address this, methods based on the Sparsity Adaptive Matching Pursuit (SAMP) algorithm have been developed to self-determine this parameter and recover the signal using only the sampling matrix and measurements. Determining a suitable Initial Value for the algorithm can greatly affect the performance of the algorithm.

Methods: One of the latest sparsity adaptive methods is Correlation Calculation SAMP (CCSAMP), which relies on correlation calculations between the signals recovered from the support set and the candidate set. In this paper, we present a modified version of CCSAMP that incorporates a pre-estimation phase for determining the initial value of the sparsity level, as well as a modified acceptance criteria considering the variance of noise.

Results: To validate the efficiency of the proposed algorithm over the previous approaches, random sparse test signals with various sparsity levels were generated, sampled at the compression ratio of 50%, and recovered with the proposed and previous methods. The results indicate that the suggested method needs, on average, 5 to 6 fewer iterations compared to the previous methods, just due to the pre-estimation of the initial guess for the sparsity level. Furthermore, as far as the least square technique is integrated in some parts of the algorithm, in presence of noise the modified acceptance criteria significantly improve the success rate while achieving a lower mean squared error (MSE) in the recovery process.

Conclusion: The pre-estimation process makes it possible to recover signal with fewer iterations while keeping the recovery quality as before. The fewer the number of iterations, the faster the algorithm. By incorporating the noise variance into the accept criteria, the method achieves a higher success rate and a lower mean squared error (MSE) in the recovery process.

Introduction

Recently, there has been a growing focus among researchers on sub-Nyquist sampling methods, due to their application in numerous industrial products, such as image encryption, radars, mm-wave body scanners and pocket handheld ultrasound scanners.

These applications almost utilize the wideband signals, which when sampled at their traditional Nyquist rate, a vast number of samples would be generated, leading to significant challenges related to processing and storage. Sub-Nyquist sampling methods, such as modulated wideband converter (MWC), offer solutions to overcome

these issues by reducing the sampling rate down to the Landau rate [1]-[5].

The key concept behind these techniques is compressed sensing [6], in which data is captured in a compressed format. The three main stages of compressed sensing are sparsification, compressed sampling, and recovery. Sparsification transforms the original signal into a sparse format. Compression involves taking measurements by multiplying the sparse signal by a sensing matrix that has specific characteristics. Finally, in the recovery phase, the signal is reconstructed from the measurements employing the sparse recovery algorithms.

These algorithms can be divided into three primary categories [7]. The first group includes methods based on convex relaxation, such as BP (Basis Pursuit) and LASSO, which attempt to find the solution by shaping linear programming (LP) problems [8], [9]. Although these methods are considered accurate, regarding their high computational complexity, they might not be applicable to practical real large-scale problems.

The second group consists of non-convex methods that rely on statistical approaches, such as BCS (Bayesian compressed sensing) [10].

The third group comprises greedy algorithms that implement the recovery process through iterative steps [4]. It should be noted that these algorithms are the most popular practical techniques because of their lower implementation complexity. Among these, Matching Pursuit (MP) algorithms are the most widely used and practical, known for their performance. In each iteration of the simple OMP, the column of the sensing matrix related to the highest correlation value with the samples is selected. This atom selection process is irreversible, and there is no chance to correct for incorrectly selected atoms [11], [12].

Other algorithms, such as CoSaMP and IHT, incorporate a backtracking approach that means in addition to selecting a certain number of atoms, they are capable of removing excessively selected ones by applying a threshold in each iteration [13]-[16]. However, these greedy algorithms need former knowledge about the signal's sparsity level, which may not always be available in practical situations.

To address this, a sub-category of greedy algorithms, known as Sparsity Adaptive Matching Pursuit (SAMP), has been developed to estimate the sparsity level as well as the recovered signal [17]-[19].

These methods start with an initial value of the sparsity level and gradually adjust it in each iteration with a specific step size. Various versions of SAMP have been developed to enhance performance in both speed and accuracy. The fixed step size is used in the basic SAMP [13], while in FSAMP, the step size increases linearly [20]. In SAMPVSS and IGSAMP, exponential functions are offered to increase the step size [21]-[24]. In Some recent algorithm, such as CCSAMP, the step size is adjusted based on the correlation coefficient calculated in each iteration [25], [26]. It is notable that the initial value selection for the step size significantly impacts the algorithm's performance. The small step size will increase the number of iterations, while the big value might lead to the overestimation of the sparsity level.

The main contribution of this work is the integration of a pre-estimation phase to determine the optimal initial step size for CCSAMP, leading to a reduced number of required iterations. Additionally, we have established a different termination criterion, significantly increasing the success rate of CCSAMP under noisy conditions.

This article is organized as follows: the second section provides an overview of the compressed sensing and SAMP algorithms. The third section introduces the pre-estimation phase and the new termination criteria. The implementation results, validating the performance of the presented work, are presented in the fourth section.

Overview of Compressed Sensing and SAMP

Consider a signal $\boldsymbol{\theta}\epsilon\mathbb{R}^{N\times 1}$ with length of N which is targeted to be compressed to a measurement signal $\mathbf{y}\epsilon\mathbb{R}^{M\times 1}$ with length of M, where M is far smaller than N ($M << N$). But as far as compressed sensing concepts are only applicable to the either sparse or compressible signals, as shown in below equation, in case of having a non-sparse original signal, $\boldsymbol{\theta}$ should be represented in terms of the sparse basis of $\boldsymbol{\Psi}$ and the sparse signal $\mathbf{x}\epsilon\mathbb{R}^{N\times 1}$, such that $\mathbf{x}$ contain only K non-zero elements.

$$\boldsymbol{\theta} = \boldsymbol{\Psi}\,\mathbf{x} \tag{1}$$

In different cases, matrix $\boldsymbol{\Psi}$ may vary. Depending on the type of application, it can be constructed utilizing Fourier transform, Discrete Cosine Transform (DCT), Wavelet transform, or other similar transforms. After a sparse representation of the signal, the compressed measurement $\mathbf{y}_{M\times 1}$ is calculated by multiplying the matrix $\boldsymbol{\Phi}_{M\times N}$ that decreases the dimension from N to M, as shown below:

$$\mathbf{y} = \boldsymbol{\Phi}\,\boldsymbol{\theta} = \boldsymbol{\Phi\Psi}\,\mathbf{x} = \boldsymbol{A}\mathbf{x} \tag{2}$$

where $\boldsymbol{A} = \boldsymbol{\Phi\Psi}$. It is noted that $\boldsymbol{A}$ is named the sensing matrix throughout this article, and to guarantee a successful recovery process of the original signal from the measurements, this matrix must satisfy special characteristics, specifically the Restricted Isometry Property (RIP). This condition is met only if the below equation is satisfied with constant $\delta_k \in (0,1)$ [27].

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \le \|\boldsymbol{A}\mathbf{x}\|_2^2 \le (1+\delta_k)\|\mathbf{x}\|_2^2 \tag{3}$$

One method of recovering the original signal from the samples $\mathbf{y}$ is solving the l_0-norm minimization problem, as

show in (4). The goal is to find the sparsest signal that, when sampled with $\boldsymbol{A}$, it produces the samples $\mathbf{y}$.

$$\hat{\mathbf{x}} = \arg\min\|\mathbf{x}\|_0 \quad s.t. \mathbf{y} = \boldsymbol{A}\mathbf{x} \tag{4}$$

In this formula $\| \|_0$ denotes the l_0 −norm and reflects the signal's sparsity level. However, this method is an NP-hard problem, and its complexity increases significantly as the dimension grows. So, it is recommended to approximate it with a l_1-norm minimization problem, as shown below [28].

$$\hat{\mathbf{x}} = \arg\min\|\mathbf{x}\|_1 \quad s.t. \mathbf{y} = \boldsymbol{A}\mathbf{x} \tag{5}$$

Although this method can recover the signal with high accuracy, greedy algorithms are often preferred due to their advantage of lower implementation complexity. Among these, certain algorithms, based on the SAMP algorithm, do not require any former information about the signal's sparsity level. The pseudo-code related to the basic SAMP is presented in Algorithm 1.

Algorithm 1: Sparsity Adaptive Matching Pursuit (SAMP)

Input params: measurement signal $\mathbf{y}$, sensing matrix $\boldsymbol{A}$, initial step-size s_0
Output: recovered signal $\hat{\mathbf{x}}$
Initial: $\hat{\mathbf{x}} = 0;\ \mathbf{r}_0 = \mathbf{y};\ F_0 = \emptyset; L = s_0; t = 1;$
while $(true)$
$B_t = \max(\boldsymbol{A}^H \mathbf{r}_{t-1}, L)$
$C_t = F_{t-1} \cup B_t$
$F = \max(\boldsymbol{A}_{C_t}^{\dagger}\mathbf{y}, L)$
$\mathbf{r} = \mathbf{y} - \boldsymbol{A}_F \boldsymbol{A}_F^{\dagger}\mathbf{y}$
if $\|\mathbf{r}\|_2 < \varepsilon$
break;
else if $\|\mathbf{r}\|_2 \geq \|\mathbf{r_{t-1}}\|_2$
$L = L + s_0$
else
$F_t = F; \mathbf{r}_t = \mathbf{r}; t = t + 1;$
end
$\hat{\mathbf{x}} = \boldsymbol{A}_F^{\dagger}\mathbf{y}$
end while

In each iteration, firstly, the candidate set C_t is calculated by finding the L indices corresponding to the largest correlation between columns of the sensing matrix $\boldsymbol{A}$ and the previous residual signal. The function $\max(\mathbf{temp}, L)$ returns the index set associated with the L highest value of the input vector of temp. Then, the signal is temporarily recovered related to the union set of F_{t-1} and C_t. Then, as the backtracking stage, the final index set is created by selecting its L largest value.

Based on this set of indices, the residual $\mathbf{r}$ is updated. During each iteration. If the correct atoms are selected, the norm of $\mathbf{r}$ tends to decrease.

Otherwise, it indicates that the chosen sparsity level L is insufficient and must be increased by the step size s_0. The repetition of the algorithm continues until the norm of $\mathbf{r}$ becomes smaller than a predefined epsilon. Meanwhile, the algorithm might stop unsuccessfully if L exceeds M or if the iteration counter t exceeds the maximum number of iterations.

As far as the SAMP algorithm uses a fixed step size s, it is prone to either overestimate or underestimate the correct sparsity level. In different modifications of the SAMP algorithm, others have made attempts to make the step size variable.

For instance, in the CCSAMP method, a varying step size is introduced that adjusts based on the correlation between $\mathbf{y}_{C_t}$ and $\mathbf{y}_F$ obtained through the Least Squares technique for the candidate set and the final set, as illustrated in (6) [25]. Step size adjustment is achieved through a multilevel decision-making process. Low correlation indicates a significant change in each iteration. So, the step size must grow. In other words, the lower the correlation, the higher the step size. As the correlation converges to 1, the step size must be selected more cautiously with small values.

$$\begin{aligned} \mathbf{y}_{C_t} &= \boldsymbol{A}_{C_t}(\boldsymbol{A}_{C_t}{}^{T}\boldsymbol{A}_{C_t})^{-1}\boldsymbol{A}_{C_t}{}^{T}\mathbf{y} \\ \mathbf{y}_F &= \boldsymbol{A}_F(\boldsymbol{A}_F{}^{T}\boldsymbol{A}_F)^{-1}\boldsymbol{A}_F{}^{T}\mathbf{y} \\ \rho_t &= corr(\mathbf{y}_{C_t}, \mathbf{y}_F) \\ s &= \begin{cases} s_0 + 10 * (1 - \rho_t) & \rho_t < 0.9 \\ s_0 & \rho_t < 1 - 10^{-6} \\ 1 & otherwise \end{cases} \end{aligned} \tag{6}$$

The Proposed Method

A. Pre-estimation Phase

The performance of CCSAMP slightly varies with different initial values of the sparsity level. This section discusses the details of ICCSAMP, specifically how the initial sparsity level is calculated through a pre-estimation phase by implementing a matching test [21], [29].

To estimate the initial sparsity level, the index set S_0 is first calculated with $L_0 = 1$, as below:

$$S_0 = \max(\boldsymbol{A}^H\mathbf{y}, L_0) \tag{7}$$

Then, the following expression is evaluated to check its correctness:

$$\|\boldsymbol{A}_{S0}^{T}\mathbf{y}\|_2 < \frac{1 - \delta_s}{\sqrt{1 + \delta_s}}\|\mathbf{y}\|_2 \tag{8}$$

Here, the constant δ_s is limited between 0 and 1. If the condition is satisfied, the L_0 is incremented by one, and S_0 is updated respectively. Otherwise, the L_0 is considered as the initial sparsity level. This value helps to reduce the number of iterations of the algorithm, leading to speed enhancement.

The pseudo-code related to this algorithm is presented below:

Algorithm 2: Improved Correlation Coefficient Sparsity Adaptive Matching Pursuit (ICCSAMP)

Input params: measurement signal $\mathbf{y}$, sensing matrix $\boldsymbol{A}$, initial step-size s_0
Output: recovered signal $\hat{\mathbf{x}}$
Initial: $\hat{\mathbf{x}} = 0;\ C_0 = \emptyset;\ F_0 = \emptyset; L_0 = 1;\ t = 1;$
//pre-estimation phase
while(true)
$S_0 = \max(\boldsymbol{A}^H \mathbf{y}, L_0)$
  if $(\|\boldsymbol{A}_{S0}^T \mathbf{y}\|_2 < ((1-\delta_s)/\sqrt{1+\delta_s})\|\mathbf{y}\|_2)$
$L_0 = L_0 + 1;$
else $break;$
end while
  $L = L_0;$
$\mathbf{r_0} = \mathbf{y} - \boldsymbol{A}_{S_0} \boldsymbol{A}_{S_0}^{\dagger} \mathbf{y}$
//body
while $(true)$
$S_t = \max(\boldsymbol{A}^H \mathbf{r}_{t-1}, L)$
$C_t = F_{t-1} \cup S_t$
$F = \max(\boldsymbol{A}_{C_t}^{\dagger} \mathbf{y}, L)$
$\mathbf{y}_{C_t} = \boldsymbol{A}_{C_t} (\boldsymbol{A}_{C_t}{}^T \boldsymbol{A}_{C_t})^{-1} \boldsymbol{A}_{C_t}{}^T \mathbf{y}$
$\mathbf{y}_F = \boldsymbol{A}_F (\boldsymbol{A}_F{}^T \boldsymbol{A}_F)^{-1} \boldsymbol{A}_F{}^T \mathbf{y}$
$\rho_t = corr(\mathbf{y}_{C_t}, \mathbf{y}_F)$
if$(\rho_t < 0.9)$ **then**
  $s = s_0 + 10 * (1 - \rho_t)$
else if $(\rho_t < 1 - 10^{-6})$ **then**
  $s = s_0$
else
  $s = 1$
$\mathbf{r} = \mathbf{y} - \boldsymbol{A}_F \boldsymbol{A}_F^{\dagger} \mathbf{y}$
  if $\|\mathbf{r}\|_2 < \varepsilon$
    break;
  else if $\|\mathbf{r}\|_2 \geq \|\mathbf{r_{t-1}}\|_2$
    $L = L + s$
  else
    $F_t = F; \mathbf{r}_t = \mathbf{r}; t = t + 1;$
end while
$\hat{\mathbf{x}} = (\boldsymbol{A}_{F_t}{}^T \boldsymbol{A}_{F_t})^{-1} \boldsymbol{A}_{F_t}{}^T \mathbf{y}$
end

B. Acceptance Criteria in Noisy Condition

Given the sensing matrix $\boldsymbol{A}$, the measurement vector $\mathbf{y}$, and the set of indices S, when the number of measurements M is much smaller than the length of the signal N, in each iteration, reconstructing the signal at specific indices leads to a set of underdetermined equations. Thus, least-squares techniques play a key role in the projection of the samples onto the signal domain.

An important issue that should not be neglected is the performance of least-squares techniques under noisy conditions. In the presence of measurement noise, the observation model is given by

$$\mathbf{y} = \boldsymbol{A}\mathbf{x} + \mathbf{n} \tag{9}$$

where $\mathbf{n}$ is additive zero-mean Gaussian noise with variance σ^2, same as below:

$$\mathbf{n} = \mathcal{N}(0, \sigma^2 \boldsymbol{I}) \tag{10}$$

In the Least square technique, the target is to find $\hat{\mathbf{x}}$ which can minimize the residual error $\mathbf{r}$.

$$\mathbf{r} = \mathbf{y} - \boldsymbol{A}\hat{\mathbf{x}} \tag{11}$$

However, the variance of the residual $\mathbf{r}$, $var(\mathbf{r})$, is often found to be greater than the variance of the noise σ^2. In other words, when applying the Least Mean Square method for finding $\hat{\mathbf{x}}$, there is always some stable error remaining which prevents the Mean Square Error (MSE) from becoming lower than the noise variance [30].

As mentioned before, the stopping condition in CCSAMP checks whether the residual norm is lower than a predefined fixed epsilon, typically equal to 10^{-6}. This fixed threshold can cause the algorithm to fail under noisy conditions. However, if the threshold is chosen based on the noise variance, the success rate of the algorithm will increase.

Experimental Results

All evaluation procedures in this article were performed using simulations in MATLAB R2021a.

To ensure a meaningful comparison among different algorithms, a unique test condition was established. Specifically, a K-sparse signal $\mathbf{x}$ was created by generating a random Gaussian signal with a length of N=256 and retaining only K randomly located elements.

To compress this signal by a compression ratio of 50%, a sensing matrix $\boldsymbol{A}$ of size 128×256 was generated, with its elements drawn from a Gaussian distribution. This structure, with high probability, ensures that this sensing matrix satisfies the RIP condition.

The measurement vector $\mathbf{y}$ is created by taking samples from $\mathbf{x}$ by multiplying it with $\boldsymbol{A}$. Given $\mathbf{y}$ and $\boldsymbol{A}$, In this section, the target is to compare the performance of recovery algorithms for estimating $\mathbf{x}$.

One measure to evaluate the performance of the algorithms is the success rate. In order to calculate this measure, for different values of the sparsity K, each recovery algorithm was repeated 500 times with different sparse signals, and the percentage of successful recovery was recorded. The result of recovery is considered successful whenever the norm of residual becomes less than a predefined epsilon.

In the first experiment, signals with different sparsity levels K, ranging from 10 to 50, were generated. The test was repeated for 512 times, and as shown in Fig. 1, the pre-estimation phase in ICCSAMP could decrease the number of iterations in average, while maintaining the success rate unchanged.

However, a slight difference in the success rate under highly sparse conditions ($K = 10$) is due to the value of the estimation parameter δ_0. The effect of this parameter on the success rate is illustrated in Fig. 2. In the case of selecting a small δ_0, the estimation of the sparsity level fails. From this figure, it can be concluded that values above 0.3 perform well for different sparsity levels.

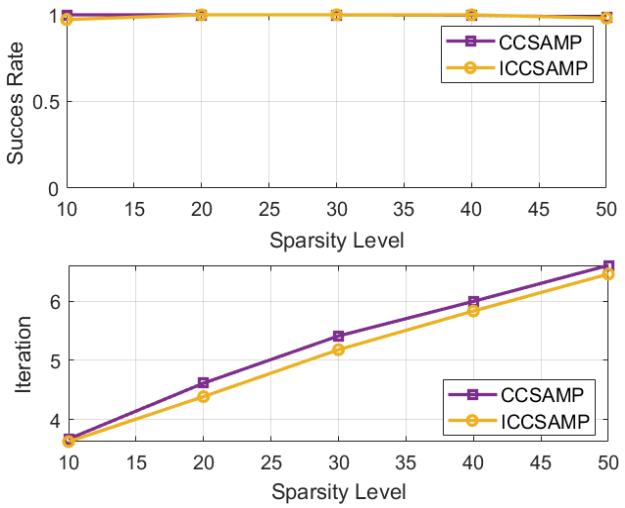


Fig. 1: Performance comparison of the CCSAMP and ICCSAMP without noise. (a) Success Rate (b) number of required iterations.

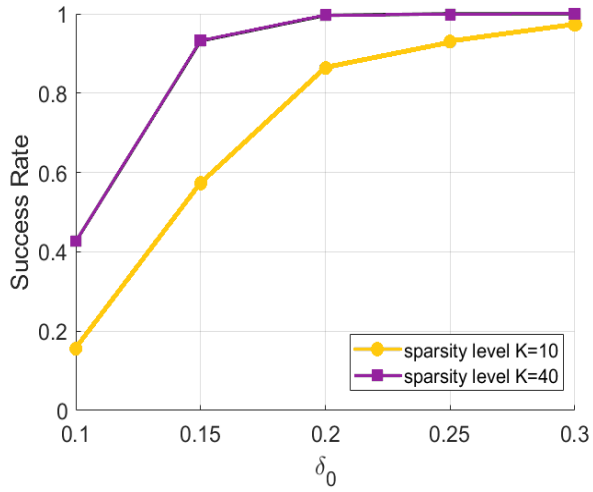


Fig. 2: Effect of δ_0 value on the success rate for sparsity levels of 10 and 40.

As is depicted in Fig. 3, it is noted that in the absence of measurement noise, both of methods, CCSAMP and ICCSAMP, reach the same level of the MSE. This can be concluded that the pre-estimation phase can increase the speed of algorithm, retaining the recovery error the same.

In another experiment, the number of iterations of the SAMP, SAMPVSS, CCSAMP, and ICCSAMP were compared in terms of different sparsity levels. In this test, the initial sparsity level $s_0 = 4$ was selected to be the same for the four algorithms. For SAMPVSS, the parameters $\alpha = 3$ and $\beta = 2$ were used. As illustrated in Fig. 4, it is evident that the ICCSAMP algorithm, which utilized the pre-estimation phase, requires fewer iterations than the other algorithms, resulting in an increase in the overall speed of the recovery process.

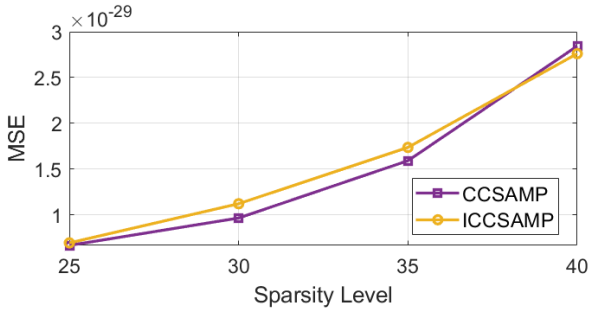


Fig. 3: MSE comparison of the CCSAMP and ICCSAMP without noise.

In another experiment, the number of iterations of the SAMP, SAMPVSS, CCSAMP, and ICCSAMP were compared in terms of different sparsity levels. In this test, the initial sparsity level $s_0 = 4$ was selected to be the same for the four algorithms. For SAMPVSS, the parameters $\alpha = 3$ and $\beta = 2$ were used. As illustrated in Fig. 4, it is evident that the ICCSAMP algorithm, which utilized the pre-estimation phase, requires fewer iterations than the other algorithms, resulting in an increase in the overall speed of the recovery process.

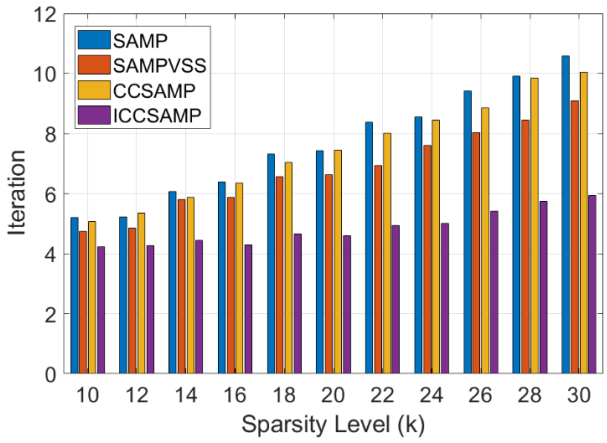


Fig. 4: Comparison of the number of required iterations in different sparsity levels for SAMP, SAMPVSS, CCSAMP, and ICCSAMP.

In the final experiment, our aim was to verify the effect of the presented modified acceptance criteria by comparing the performance of CCSAMP and ICCSAMP under noisy conditions. Meanwhile, to verify that the performance of the presented algorithm remains consistent across signals with different lengths, we conducted this test using an input signal of length 512.The success rate, number of iterations, and the mean squared error (MSE) of the recovery error are illustrated in the Fig. 5.

It is evident that in the presence of Gaussian measurement noise with various SNR levels, the CCSAMP algorithm with fixed stop criterion of 10^{-5} was unable to perform, having a success rate below 0.6 in various SNR levels. In contrast, the ICCSAMP algorithm could successfully recovered the signal. As shown, for ICCSAMP, both the MSE and the number of iterations decrease as the SNR increases. Specifically, in SNR higher than $20dB$, the MSE of the CCSAMP is nearly zero which proves that the signal could be recovered with high accuracy. This behavior can be explained by the fact that greedy algorithms operate iteratively. If the acceptance criterion is not chosen appropriately, the algorithm is likely to fail in detection of the correct sparsity level. Neglecting noise variance and using a fixed value for epsilon increases the likelihood of failure, particularly in low SNR conditions.

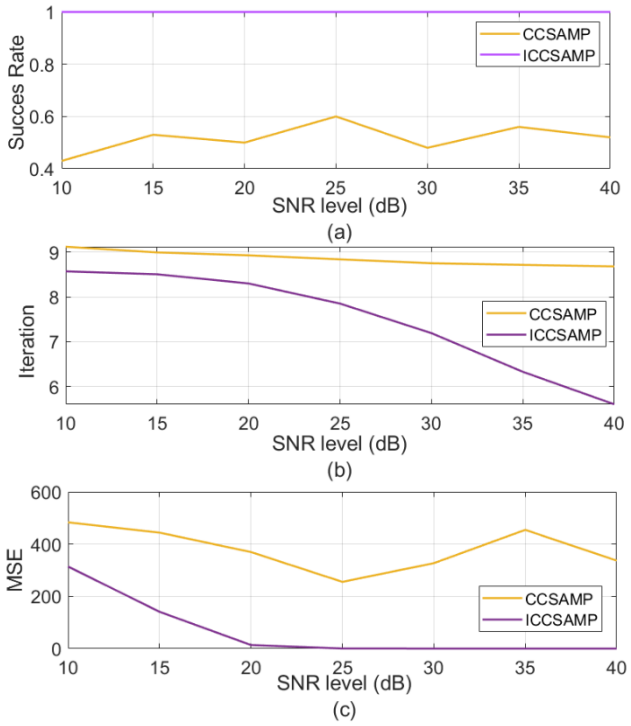


Fig. 5: Performance comparison of CCSAMP and ICCSAMP in different SNRs. (a) success rate (b) number of required iterations (c) MSE.

Discussion

All the experiments in this article are conducted using a white Gaussian signal as the input, which follows a normal distribution. It is important to note that if other distributions, such as Cauchy sequences with outlier values, are used, the performance may be affected. During the sparsification process if outlier values are selected, the recovery process becomes more challenging. This is because greedy algorithms operate by selecting columns of the sampler matrix that have higher correlation with the samples. In this phase, outliers tend to dominate, which can significantly affect the atom selection process.

In an experiment, we generated two sets of input based on the Gaussian and Cauchy distribution, and repeated the algorithm for 512 times. The below table also implies on the above-mentioned discussion.

Table 1: ICCSAMP performance with two different inputs

Input distribution	Success rate	MSE
Gaussian	0.88	42.407
Cauchy	0.51	1.59×10^{5}

Summary and Conclusion

In sub-Nyquist sampling methods such as MWC, sparse recovery algorithms are essential for the blind recovery process. Although most greedy recovery algorithms require knowing the sparsity level in advance, a group of algorithms, known as SAMP algorithms, can adaptively recover the signal by adjusting the sparsity level in each iteration. The initial value of the sparsity level in SAMP algorithms can highly affect their performance due to either overestimation or underestimation. Also, the way the step-size is adjusted can change the computational time of the algorithm. In our proposed method, we not only reduced the number of required iterations by integrating a pre-estimation phase, but also increased the success rate of the algorithm in noisy conditions by implementing a modified stopping criterion, based on the variance of the white gaussian noise.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contribution

Azadeh Vakili: Conceptualization, Investigation, Software, Methodology, Validation, Writing - original draft. Mohammad Shams Esfand Abadi: Project administration, Supervision, Validation, Writing - review & editing. Mohammad Kalantari: Project administration, Supervision, Validation, Writing - review & editing.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Abbreviation

$\mathbf{A}$	*Sensing matrix*
$(.)^{H}$	*Hermitian transpose of argument matrix*

$(.)^{\dagger}$	*Pseudo inverse of argument matrix*
$\boldsymbol{A}_F$	*Columns of* $\boldsymbol{A}$ *corresponding to index set* F
$\hat{\boldsymbol{x}}$	*Estimated recovered signal*
$\| \|_0$	l_0 *−norm*
$\|.\|_2$	*Euclidean norm*
$corr(\boldsymbol{x}, \boldsymbol{y})$	*Correlation between* $\boldsymbol{x}$ *and* $\boldsymbol{y}$
$var(.)$	*Variance of elements of argument vector*
RIP	*Restricted Isometry Property*
$\boldsymbol{n}$	*Measurement noise*
$\boldsymbol{y}$	*Compressed samples*
MSE	*Mean Square Error*
K	*True sparsity level*
L	*Estimated sparsity level*
SNR	*Signal to Noise Ratio*
SAMP	*Sparsity Adaptive Matching Pursuit*
CCSAMP	*Correlation Calculation SAMP*

References

[1] J. Kang, H. Yoon, C. Yoon, S. Emelianov, "High-frequency ultrasound imaging with sub-nyquist sampling," IEEE Trans. Ultrason. Ferroelectr. Freq. Control, 69(6): 2001-2009, 2022.

[2] D. Cohen, Y. C. Eldar, "Sub-nyquist radar systems: Temporal, spectral, and spatial compression," IEEE Signal Process. Mag., 35(6): 35-58, 2018.

[3] C. Wang, S. Ling, "An image encryption scheme based on chaotic system and compressed sensing for multiple application scenarios," Inf. Sci., 642, 119166, 2023.

[4] Y. C. Eldar, Compressed Sensing. in: Sampling Theory Beyond Bandlimited Systems, Cambridge University Press. 608-628, 2015.

[5] D. L. Donoho, "Compressed sensing," IEEE Transa. Inf. Theory, 52(4): 1289-1306, 2006.

[6] E. J.Candes, J. Romberg, T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory, 52(2): 489-509, 2006.

[7] E. C. Marques, N. Maciel, L. Naviner, "A review of sparse recovery algorithms," IEEE Access, 7: 1300-1322, 2019.

[8] S. Chen, D. L. Donoho, M. A. Saunders, "Atomic decomposition by Basis Pursuit," SIAM J. Sci. Comput., 20(1): 33-61, 1999.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R. Stat. Soc., 58(1): 267-288, 1996.

[10] S. Ji, Y. Xue, L. Carin, "Bayesian compressive sensing," IEEE Trans. Signal Process., 56(6): 2346-2356, 2008.

[11] S. G. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries," IEEE Trans. Signal Process., 41(12): 3397-3415, 1993.

[12] Y. Pati, R. Rezaifar, P. Krishnaprasad, "Orthogonal matching Pursuit: Recursive function approximation with applications to wavelet decomposition," in Proc. Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, 1993

[13] W. Dai, O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," IEEE Trans. Inf. Theory, 55: 2230-2249, 2009.

[14] D. Needell, J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," Appl. Comput. Harmon. Anal., 26(3): 301-321, 2008.

[15] T. Blumensath, M. Davies, "Gradient pursuits," IEEE Trans. Signal Process., 56(6): 2370-2382, 2008.

[16] Q. Wang, G. Qu, "A new greedy algorithm for sparse recovery," Neurocomputing, 275: 137-143, 2018.

[17] T. Do, L. Gan, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in Proc. Conference on Signals, Systems, and Computers, 2008.

[18] R. Yan, Q. Li, H. Xiong, "Mitigating impulsive noise in airborne PLC: Introducing the S-SAMP-PV algorithm for MIMO OFDM systems," Signal Process., 230, 109798, 2025.

[19] Z. B. Lahaw, H. Seddik, "A new greedy sparse recovery algorithm for fast solving sparse representation," Visual Comput., 38: 2431-2445, 2022.

[20] S. Yao, Q. Guan, S. Wang, "Fast sparsity adaptive matching pursuit algorithm for large-scale image reconstruction," EURASIP J. Wireless Commun. Networking, 78, 2018.

[21] Z. Liquan, M. Ke, J. Yanfei, "Improved generalized sparsity adaptive matching pursuit algorithm based on compressive sensing," J. Electr. Comput. Eng., 2020, 2782149, 2020.

[22] Y. Zhang, Y. Liu, X. Zhang, "A variable step-size sparsity adaptive matching pursuit algorithm," IAENG Int. J. Comput. Sci. 48(3), 2021.

[23] C. Wang, Y. Zhang, L. Sun, "Improved sparsity adaptive matching pursuit algorithm based on compressed sensing," Displays, 77, 102396, 2023.

[24] Y. Fu, S. Liu, C. Ren, "Adaptive step-size matching pursuit algorithm for practical sparse reconstruction," Circuits Syst. Signal Process., 36: 2275-2291, 2017.

[25] Y. Li, W. Chen, "A correlation coefficient sparsity adaptive matching pursuit algorithm," IEEE Signal Process. Lett., 30: 190-194, 2023.

[26] X. Wang, Y. Jiang, G. Ding, "An improved variable step SAMP method based on correlation principle," Electronics,13(22),4502, 2024.

[27] E. J. Candes, T. Tao, "Decoding by Linear Programming," IEEE Trans. Inf. Theory, 51(12): 4203-4215, 2005.

[28] J. Tropp, S. Wright, "Computational methods For sparse solution of linear inverse problems," Proc. IEEE, 98(6): 948-958, 2010.

[29] X. Zhang, Y. Liu, X., Wang, "A sparsity pre-estimated adaptive matching pursuit algorithm," J. Electr. Comput. Eng., 2021, 5598180, 2021.

[30] M. H. Hayes, Statistical Digital Signal Processing and Modeling, John Wiley and Sons, Canada, 1996.

Biographies

Azadeh Vakili received her B.Sc. degree in Computer Hardware Engineering and her M.Sc. in Mechatronic Engineering from K. N. Toosi University of Technology (KNTU), Tehran, Iran in 2011 and 2013, respectively. After 10 years of industrial experience on research and development of electronical devices, she is currently Ph.D. candidate in Electronic Engineering at Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran. Her area of interest includes signal processing, sampling theory, compressed sensing, and Internet of Things.

- Email: azadehvakili@sru.ac.ir
- ORCID: 0009-0009-4103-3208
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mohammad Shams Esfand Abadi received the B.S. degree in Electrical Engineering from Mazandaran University, Mazandaran, Iran and the M.Sc. degree in Electrical Engineering from Tarbiat Modares University, Tehran, Iran in 2000 and 2002, respectively, and the Ph.D. degree in Biomedical Engineering from Tarbiat Modares University, Tehran, Iran in 2007. Since 2004 he has been with the Department of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran, where he is currently a Professor. His research interests include digital filter theory, adaptive distributed networks, and adaptive signal processing algorithms.

- Email: mshams@sru.ac.ir
- ORCID: 0000-0002-9856-6592
- Web of Science Researcher ID: Y-7686-2019
- Scopus Author ID: 7006167272
- Homepage: https://www.sru.ac.ir/shams/

Mohammad Kalantari received B.Sc. degree in Computer Engineering from Iran University of Science and Technology (IUST), Tehran, Iran and M.Sc. and Ph.D. in Computer Engineering from Amirkabir University of Technology (AUT), Tehran, Iran in 2001 and 2009 respectively. He is currently working as Assistant Professor at Signal Processing Laboratory in Computer Engineering Department at Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran. His area of interest includes, Statistical signal processing, Spherical array processing, Sampling theory, and Compressed sensing.

- Email: mkalantari@sru.ac.ir
- ORCID: 0000-0002-6852-9344
- Web of Science Researcher ID: HZJ-9229-2023
- Scopus Author ID: 55893680300
- Homepage: https://www.sru.ac.ir/kalantari/

How to cite this paper:
A. Vakili, M. Shams Esfand Abadi, M. Kalantari, "Improved correlation coefficient sparsity adaptive matching pursuit in noisy condition," J. Electr. Comput. Eng. Innovations, 13(2): 379-386, 2025.

DOI: 10.22061/jecei.2025.11542.813

URL: https://jecei.sru.ac.ir/article_2278.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Ensemble Learning Algorithm for Power Transformer Health Assessment Using Dissolved Gas Analysis

K. Gorgani Firouzjah [*], ***J. Ghasemi***

Department of Electrical Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran.

Article Info

Article History:
Received 12 November 2024
Reviewed 25 December 2025
Revised 26 January 2025
Accepted 29 January 2025

Keywords:
Power transformer
Health assessment
Ensemble learning
Dissolved gas analysis
Predictive maintenance

[*]Corresponding Author's Email Address: *k.gorgani@umz.ac.ir*

Abstract

Background and Objectives: Power transformer (PT) health assessment is crucial for ensuring the reliability of power systems. Dissolved Gas Analysis (DGA) is a widely used technique for this purpose, but traditional DGA interpretation methods have limitations. This study aims to develop a more accurate and reliable PT health assessment method using an ensemble learning approach with DGA.

Methods: The proposed method utilizes 11 key parameters obtained from real PT samples. In this way, synthetic data are generated using statistical simulation to enhance the model's robustness. Twelve different classifiers are initially trained and evaluated on the combined dataset. Two novel indices (a risk index and an unnecessary cost index) are introduced to assess the classifiers' performance alongside traditional metrics such as accuracy, precision, and the confusion matrix. An ensemble learning method is then constructed by selecting classifiers with the lowest risk and cost indices.

Results: The ensemble learning approach demonstrated superior performance compared to individual classifiers. The learning algorithm achieved high accuracy (99%, 92%, and 86% for three health classes), a low unnecessary cost index (6%), and a low misclassification risk (16%). This result indicates the effectiveness of the ensemble approach in accurately detecting PT health conditions.

Conclusion: The proposed ensemble learning method provides a reliable and accurate assessment of PT health using DGA data. This approach effectively optimizes maintenance strategies and enhances the overall reliability of power systems by minimizing misclassification risks and unnecessary costs.

Introduction

Power transformers (PTs) are considered one of the essential elements in power supply systems. The main task of this equipment is to manage voltage levels to ensure compatibility between generation sources and various electrical loads [1]. In this regard, the reliable performance of PTs essential to ensure a continuous power supply. This equipment (as a critical network asset) to reduce the risk of unexpected failures requires regular maintenance and condition assessment. In fact, carrying out these preventive measures improves the reliability of the network. They also reduce downtime and enhance the overall stability of the power system [2]. In order to protect personnel and equipment, it is essential to identify PT issues in their early phases, thereby preventing costly repairs and mitigating potential safety hazards. Modern diagnostic tools and intelligent decision-making procedures drive predictive maintenance techniques, which are essential for ensuring a continuous energy supply and improving the functionality of PTs. Asset managers that proactively identify potential problems and meticulously plan maintenance measures can

successfully reduce downtime, extend PT operating lifespans, and significantly reduce overall operational expenses. In addition to enhancing grid stability, this innovative approach fosters a more robust and sustainable energy system. The literature on PT fault diagnosis provides most of the techniques for identifying early faults and preventing catastrophic failures. The researchers employ a variety of methods in this procedure. Among these are vibration data analysis, thermographic image processing [3], dissolved gas analysis (DGA), and acoustic emission analysis [4]. However, each of these diagnostic techniques has its own unique advantages and drawbacks [5]. One popular and successful method for diagnosing PT faults is DGA [6].

This strategy is based on the fact that different types of PT faults produce distinct gases in the oil. Accordingly, assessing the quantities of these gases can identify the kind and extent of the fault [7]. Many metrics obtained from dissolved gases in oil have led to the identification of key gases. These include carbon monoxide (CO), carbon dioxide (CO_2), hydrogen, methane (CH_4), ethylene (C_2H_4), ethane (C_2H_6), and acetylene (C_2H_2) [8]. However, we have also exploited other practical features like the oil's moisture content and its insulation breakdown voltage level. As indicated before, DGA is a robust oil assessment instrument. Key gas method (KGM) is a popular DGA data analysis tool [9]. KGM, a DGA subsidiary, specializes on fault-related gases. This strategy improves fault type and severity identification. Measurements of dissolved gases in PT insulating oil indicate fault occurrence and type. This method leverages the fact that every inaccuracy leads to different ratios of oil to dissolved gas. It is possible to ascertain the nature and severity of the fault by comparing the gas concentrations to the threshold values [10]. The characteristics of these gases can identify various faults. For instance, internal faults can result in the production of hydrogen, CO, CO_2, and CH_4 d. However, cellulose faults have the ability to produce other gases, such as CH_4, C_2H_6, and C_2H_4 [10]. On the other hand, higher hydrocarbon temperatures can lead to higher CH_4 and C_2H_6 concentrations. Moreover, studies in [4] indicate that an electrical arc or partial discharge may raise the concentration of hydrogen. The IEEE C57.104-2008 standard provides useful guidelines for testing, interpreting, and decision-making of various PT faults. The aim of this standard is to standardize and improve the accuracy and reliability of the DGA method. This standard, by establishing uniform procedures and specific requirements, helps increase the efficiency and accuracy of the DGA method for detecting faults and assessing the health status of PTs.

In the literature, the DGA process includes various steps, the first of which is data preparation. This involves collecting oil samples and conducting necessary tests. The procedure is done to measure the concentration of dissolved gases. After data preparation, the concentration of gases (such as hydrogen, methane, ethylene, etc.) is identified. Then, the PT fault type is determined using various analytical methods (such as key gas ratios, Rogers ratios, IEC ratios, and the Duval triangle [11]). In the next step, the obtained results from the analytical methods are compared with real data to evaluate the accuracy and efficiency of the method. Eventually (based on the obtained results), decisions will be made regarding the health status of the PT and the necessary maintenance actions. Gas concentrations in oil must be within limits. The references [12]-[14] provide the concentration limits for gases in the DGA method. These boundaries are used for fault diagnosis and health assessment. For example, if the hydrogen concentration in the oil exceeds the permissible limit, it may indicate windings or core fault . However, interpreting DGA parameters takes careful consideration of several aspects. In fact, environment, oil type, age, and loads affect gas concentrations and fault detection. Furthermore, the relation between the concentrations of different gases and the types of fractures can be complex and nonlinear. For this reason, the use of advanced statistical methods and mathematical models is essential for the accurate interpretation of results.

Recent years have seen extensive research into advanced DGA interpretation methods. Researchers used neural networks (NNs) [15]-[17], genetic algorithms [18], and fuzzy logic [19], [20]. These approaches are promising PT fault detectors because they can learn complex data patterns and correlations [21]. Traditional DGA interpretation relies on empirical rules and expert knowledge. While these methods have been helpful in many circumstances, they may not be enough to reliably estimate transformer health under all operating conditions. Many studies have shown that classifier algorithms including support vector machines (SVM) [22], [23], k-nearest neighbors (KNN) [24], [25], random forests, decision tree, and Naïve Bayes can categorize DGA data and discover different faults. For instance, utilizing DGA data, Benmahamed et al. [26] suggested a unique method for improving the precision of transformer problem identification. They use a bat algorithm for parameter optimization, a Gaussian classifier, and a SVM classifier in their method. The proposed method were able to accurately classify six different types of faults than with traditional DGA methods. This was done by optimizing the SVM parameters and using the concentration of five combustible gases as input. In [27], they used SVM and the optimization procedure to enhance the model parameters and increase the accuracy of fault detection. Haque et al.

[28] proposed a novel method for fault diagnostics in PTs, employing DGA and a Random Forest classifier. They classified several fault types with excellent accuracy using their approach, which combines a modified Duval pentagon method with Euclidean distance characteristics and density-based clustering. Using DGA data, a study [29] assessed the diagnostic performance of Naive Bayes and the KNN algorithm for transformer oil insulation states. An approach for finding faults in oil-immersed PTs uses DGA, a mixed KNN algorithm, and a decision tree [24]. The literature review demonstrates that feature selection enhances detection process. However, some issues with the literature may make it less useful and accurate in practice. The limitations of prior studies:

- Some research has mostly looked at small datasets or certain kinds of mistakes. This problem makes the model less reliable and useful in the real world.
- Some studies have not used ensemble learning methods to improve the accuracy and stability of the models (by combining the predictions of models).
- Many studies have ignored the importance of mistakes and the costs associated with erroneously recognizing issues. This problem may lead to the selection of models that are extremely accurate but carry serious risks and high costs. For example, misdiagnosing a malfunctioning PT as functional might result in irreparable damage, whereas misdiagnosing a working PT as a challenge in high maintenance costs.

This research has presented a novel ensemble learning method to address the limitations of previous studies in PT health status assessment using DGA data. This algorithm uses a big dataset and combines 8 successful classifiers in an ensemble learning method. It aims to develop a model to determine the PT health across various scenarios. This approach considers diagnostic accuracy and error costs to make cost-effective decisions. This generates synthetic data and standardizes it to improve performance. Additionally, in order to measure the PT's health status, limit high-risk mistakes, and save unnecessary costs, we have developed two risk and unnecessary cost indicators. The technologies used in this study should improve PT condition monitoring systems and reduce unexpected breakdowns. Consequently, we have organized the paper as follows: Section 2 explains the suggested approach. Section 3 compares the performance of the proposed technique with other methods, showcasing the implementation details and evaluation results on real datasets. Results and suggestions for future work are summarized in Section 4.

Proposed Method

This section outlines a proposed methodology that employs DGA data and machine learning algorithms to assess the health status of transformers. The primary objective of this method is to address the shortcomings of conventional techniques and to enhance the precision and dependability in assessing the health condition of the transformer. The suggested technique categorizes DGA data into three clusters: Healthy, needs retesting in the future and needs immediate retesting. This classification allows power system operators to evaluate the PT health status accurately. Also, they make appropriate decisions for PT maintenance and repair. Fig. 1 shows the flowchart of the proposed algorithm. As shown, the algorithm includes multiple steps as follows. At statistical distribution extraction step, the appropriate distribution for DGA parameters (gas concentrations and their ratios) in each of the three classes is determined. The algorithm generate synthetic DGA parameters to increase the of training data size and enhance the performance of machine learning models. These data are generated using the obtained statistical distributions. The algorithm normalize both real and synthetic DGA data using an appropriate method to ensure uniformity in their scale and measurement units. The data is randomly shuffled and assigned to training and testing sets for machine learning analysis. In order to establish an effective model, twelve classifiers are trained and assessed. Consequently, an ensemble approach is then employed to further enhance the accuracy and reliability of health predictions. The following sections will provide complete descriptions of the steps.

A. Step 1: Retrieval of Initial Data

This paper outlines a method for evaluating the health of transformers using historical DGA data. These data are collected through the analysis of oil samples extracted from operational PTs and stored in a dedicated database. These include the concentrations and ratios of each of the gases, which collectively provide unique insights into the transformer's state. The eleven important parameters are then reviewed and explained in terms of their technical importance. The average breakdown voltage measures the dielectric strength of the oil, which shows how well it can handle electrical breakdowns when voltage stress is applied. The drop in breakdown voltage shows that the oil insulation is breaking down. This can be caused by contamination, oxidation, or the buildup of breakdownproducts. This makes it more likely that there will be partial discharges and, eventually, insulation failure. Moisture in transformer oil, even in small amounts (ppm), can make it less effective at insulating and can greatly reduce the oil's dielectric strength.

This can speed up the breakdown of the paper insulation, cause acids and sludge to form, and further weaken the transformer's integrity. Carbon monoxide solubilized in oil serves as the primary indication of thermal stress on cellulose insulation (paper). The elevation in CO levels signifies that the insulation is experiencing overheating, either attributable to overload, inadequate cooling, or isolated hot spots within the PT.

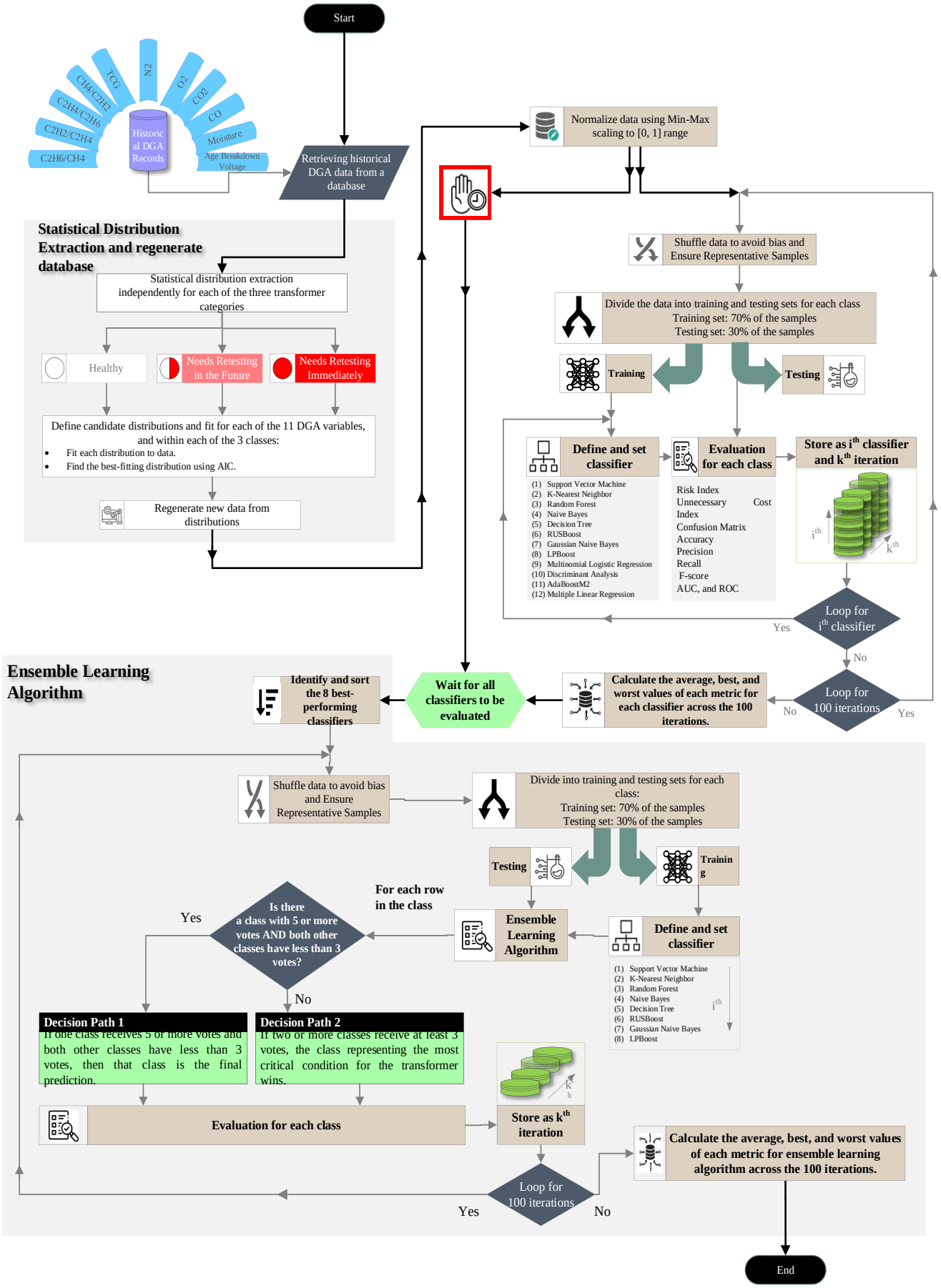


Fig. 1: Flowchart of the proposed transformer health assessment algorithm.

When paper insulation degrades, it releases CO_2 (a substance more difficult to analyze than CO). Taking into account the CO_2/CO ratio can aid in accurately determining the type and severity of the insulation issue. The oil containing oxygen poses a significant issue as it accelerates oxidation, leading to the production of acids and sludge. When oxygen enters the system, it typically indicates an air leak in the generator or conservator tank, necessitating immediate repair. While nitrogen typically serves as a background gas, it can occasionally serve as a diagnostic indicator. A high nitrogen level could mean that air is getting into the oil or that nitrogenous molecules are breaking down. Total Combustible Gases (TCG) measures the amount of combustible gas released during partial discharges, overheating, and arcing. To ascertain the nature of the mistake, the program must thus probe further into the TCG level spike. Gas ratios (CH_4/C_2H_2, C_2H_6/CH_4, C_2H_4/C_2H_6, and C_2H_2/C_2H_4) offer a detailed understanding of fault situations. Analyzing the relative concentration of different gases allows for the inference of the fault's nature (e.g., arcing, overheating, or partial discharge) and its severity. The literature frequently used these ratios in conjunction with recognized diagnostic criteria such as the Duval Triangle or IEC 60599 to assess DGA results effectively. The historical DGA data (which builds up over time and includes data from many transformers) helps learn more about how transformers work. It makes preventative maintenance easier and lowers the chance of failures, which keeps the power grid running smoothly.

B. Step 2: Statistical Distribution Extraction

In this step, the statistical distribution governing each of the 11 DGA variables is extracted separately for each of the three transformer health classes ("Healthy," " needs retesting in the future," and "needs immediate retesting "). The purpose of this procedure is to create synthetic data and increase the database's size. Synthetic data is generated to address limitations in the amount and diversity of real-world data while also reducing the cost and time required for data acquisition. With the increase in the volume of training data, various patterns are better learned by the model. As a result, the accuracy and generalizability of the model increase. Various operational conditions and different types of errors is simulated by generating synthetic data. In this way, the model is helped to perform acceptably under various conditions and to be more resilient against new data. The process of extracting statistical distributions is such that several probable statistical distributions (such as normal, log-normal, Weibull, gamma, beta, and exponential) are applied to the data of each variable in each class. Then, using the Akaike Information Criterion (AIC), the best distribution governing the data is selected. The AIC criterion, by simultaneously considering the goodness of fit of the distribution to the data and the complexity of the model, helps in selecting the best distribution. Each variable is analyzed to determine its statistical distribution within each of the three classes. This technique yields 33 unique distributions, given 11 variables and 3 classes. New DGA data is produced with the Monte Carlo approach with these 33 distributions. The data are incorporated into the primary database to augment the training dataset, hence enhancing the efficacy of machine learning models in assessing the health status of transformers.

C. Step 3: Data Normalization

The third step of the suggested algorithm is to normalize the data. This is done after getting the statistical distribution, simulating synthetic data. This step prepares data for the learning algorithm to improve detection. For this purpose, the Statnorm normalization method is used. It transforms the data based on the standard normal distribution (with a mean of zero and a standard deviation of one). Statnorm is a powerful normalization technique specifically suitable for data that do not follow a normal distribution. This method transforms the data into a standard normal distribution by using the rank transformation and then applying the inverse cumulative normal distribution function. This approach uses Statnorm's rank-based outlier elimination. This minimizes algorithm outlier sensitivity, which is especially important for DGA data (with outlier values). In addition, the data normalizing into a standard normal distribution improves the performance of some machine learning algorithms (such as SVM, k-NN, and logistic regression). These algorithms perform more effectively for data that follows a normal distribution. In order to prevent the introduction of bias, the normalization of both the training and test datasets should be same.

As mentioned above, the method builds a synthetic dataset to enhance the transformer health detection algorithm through three steps: statistical distribution extraction, synthetic data production, and data normalization. This method improves PT health assessment models by adding training data. In addition, it improves these models by conducting sensitivity analysis and simulating different situations., useful, and reliable. Transformer condition assessment improves, resulting in fewer failures.

D. Step 4: Data Preparation

This step is very important for getting data ready for programs that use machine learning. The chosen method reduces possible errors and improves the model's performance with new data (by randomizing and dividing it up). There are two important parts to the data preparation step. In this regard, the suggested method mixes up the data in a way that gets rid of any bias that

might come from the order of the original dataset. This issue guarantees that both the training and testing sets (accurately) represent the entire data distribution. Secondly, it partitions the data into training and testing sets, often according to a 70/30 ratio. All three categories—"healthy," "requiring future retesting," and "requiring immediate retesting"—utilize this procedure.

E. Step 5: Classifier Training and Evaluation

At this step, the algorithm focuses on the precise training and evaluation of various classifiers to detect the health status of the transformer. The goal of this step is to select the best algorithm for classifying DGA data and accurately diagnosing the PT health. The proposed PT health assessment system uses 12 well-known machine learning classifiers [30]. The approach optimizes classifiers via hyper-parameter tweaking. It finds optimal hyper-parameter combinations using grid search. In SVM, grid search optimizes the kernel type (linear or polynomial), penalty parameter (C), and kernel coefficient (gamma). Adaptive Boosting (AdaBoostM2), Linear Programming Boosting (LPBoost), and Random Under-sampling Boosting (RUSBoost) all have multi-class versions. The number of weak learners (decision trees) is an important hyper-parameter that is optimized using cross-validation.

The cross-validation procedure optimizes the number of nearest neighbors (k) in the KNN. In addition, the algorithm adjusts the number of decision trees in a random forest based on cross-validation to achieve a balance between model accuracy and complexity. This paper evaluates many classifiers and chooses the optimal model for transformer health status detection using a cross-validation approach with 100 random repeats. The method is accurate and reliable. The method randomly splits data into 70% training and 30% testing per iteration. The results are independent of the data separated by this approach. Using training data, each classifier is trained and the grid search technique optimizes the model's hyper-parameters. The suggested method uses testing data to evaluate the accuracy, precision, and confusion matrix of the trained model. It randomly divides data between training and testing sets in each iteration . The procedure's primary loop repeats data partitioning, training, and evaluation 100 times. The primary loop calculates the mean and standard deviation of the evaluation metrics for each classifier. This procedure is done to evaluate the system performance. The advantages of this technique encompass a reduction in the variance of model performance estimate, an evaluation of model stability against fluctuations in training and testing data, and the identification of optimal model parameters and data partitioning configurations.

This paper uses two new evaluation metrics to assess the performance of classifiers in diagnosing the health status of transformers in addition to traditional criteria.

- Risk Index (primary priority):

This metric addresses the reduction of errors that pose significant risks. In practice, misidentifying a faulty transformer as healthy is a critical error. This issue has the potential to cause catastrophic consequences. The value of this metric is derived from the sum of the following two values. First, the number of PTs that truly need immediate retesting but are mistakenly classified as healthy or needing retesting in the future is determined. Then, this number is divided by the total number of PTs that truly need immediate retesting This value (in percentage) allows the algorithm to model the percentage of high-risk errors. Second, the algorithm identifies the number of transformers incorrectly classified as "healthy" but actually requiring retesting in the future. Then, the algorithm divides this number by the total number of transformers that need retesting in the future (convert to percentage). Finally, the two obtained percentages are summed to derive the error metric.

- Unnecessary Cost Index (secondary priority):

The purpose of this criterion is to reduce unnecessary expenses. In fact, unnecessary expenses are imposed on the system when a healthy transformer is wrongly classified as defective. Then, this number is divided by the total number of healthy transformers (then converted to a percentage). This method makes it possible to determine the percentage of unnecessary expenses. Additionally, the number of transformers that actually need retesting in the future but have been mistakenly classified as needing immediate retesting is determined. Then, this number is divided by the total number of transformers that actually need to be retested in the future (then converted to a percentage). Finally, we add the two obtained percentages together to derive the cost metric. In these criteria, different errors are weighted according to their importance. For instance, the error criterion assigns more weight to errors that result in misclassifying faulty transformers as healthy. The algorithm supplements conventional metrics with these criteria to evaluate the performance of classifiers from various aspects. These two new criteria allow the algorithm to select classifiers that not only have high accuracy in detecting transformer health status but also minimize high-risk errors and unnecessary costs.

F. Ensemble Learning Algorithm

At this step, the program applies an ensemble learning strategy. This procedure is a very efficient machine learning method that applies multiple classifiers at once. In fact, it combines multiple classifiers into a single, more accurate, and reliable forecast. The proposed ensemble learning based algorithm implements eight classifiers. It trained these classifiers and optimized their parameters using grid search methods and cross-validation in the

preceding phases. The following are included: SVM, KNN, Random Forest, Naive Bayes, Decision Tree, RUSBoost, Gaussian Naive Bayes, and LPBoost The proposed algorithm selected these classifiers due to their exceptional performance and diversity in the classification of DGA data. The algorithm receives DGA data as input for each case. Each of the eight classifiers independently predicts the health status of the transformer based on the input DGA data and categorizes it into one of three classes. The algorithm calculates class votes (predictions) from eight classifiers. For choosing the winning class (final transformer health diagnostic), two choice paths are considered:

- Decision path 1:

  If it gets 6 out of 8 votes, a class wins and becomes a transformer.
- Decision path 2:

  If both classes receive at least three votes each, the class indicating a more critical condition for the transformer will be declared the winner. In other words, the priority order will be "needs immediate retesting" followed by " needs retesting in the future" and finally "healthy".

The proposed algorithm uses various metrics to evaluate the ensemble learning performance. These include accuracy, precision, and confusion matrix. These metrics encompass error rate metrics, risk factors, and unnecessary cost indexes. The algorithm The algorithm also utilized these metrics in step 5, which involves training and evaluating classifiers. The algorithm conducts the evaluation procedure 100 times to enhance confidence in the outcomes. Subsequently, it computes the mean and standard deviation of the evaluation metrics. The proposed ensemble learning technique enhances the evaluation of PT health by merging the predictions of the top eight classifiers and accounted for more significant factors in the final decision-making process. Employing this technique can diminish maintenance expenses and prolong the longevity of PTs.

Numerical Results and Analysis

This section explores the numerical outcomes of implementing the method from Section 2 in real-world scenarios. This section demonstrates how the proposed method accurately categorizes PT health using DGA data. Statistical distribution assessment, classical classifier evaluation, and group modeling are key components of this technique. Section 3.1 will carefully evaluate the statistical distributions of critical DGA parameters for each transformer health class to find data trends. This information should be utilized to generate synthetic data. Section 2 proposes a 100-fold random cross-validation method to compare the performance of twelve different machine learning classifiers, including SVM, KNN, and Random Forest. Thus, accuracy, precision, confusion matrix, and unnecessary cost indices are used. This section will test the ensemble learning model's transformer health state classification accuracy and dependability. The ensemble learning model uses eight top classifier predictions. The proposed method is compared to existing DGA interpretation methods for advantages and disadvantages.

A. Evaluation of Data Preprocessing Steps

This section will address the numerical evaluation of the first three steps of the proposed algorithm, which include "statistical distribution extraction," "synthetic data generation," and "data normalization." The goal of this step is to identify an appropriate statistical distribution for each of the 11 DGA parameters (including breakdown voltage, moisture, gas concentrations, and their ratios) within each of the three transformer health classes. The information regarding the transformer under study is presented as follows. Fig. 2 to Fig. 5 present the statistical distribution of the key features of the studied transformers. Fig. 2 shows the statistical distribution of transformer lifespans. As can be seen, most transformers have a lifespan of between 5 to 10 years (with a frequency of about 38%). Additionally, there are a few transformers with a lifespan of over 35 years. This indicates that the data includes transformers with a variety of ages, from new to old. Fig. 3 shows the statistical distribution of transformer capacities. The majority of transformers (about 70%) have a capacity between 30 to 40 MVA. A few transformers with lower capacities (10-20 MVA) and (20-30 MVA) are also present in the data. Fig. 4 shows the statistical distribution of transformer oil weights. Transformers with an oil weight of 12 to 16 tons (approximately 48%) exhibit the highest frequency. Additionally, transformers with lower oil weights (4-8 tons) and 8-12 tons are also present in the test samples. Fig. 5 shows the statistical distribution of the type of oil used in transformers. Most transformers (over 90%) use IEC-296- type oil. Only a small percentage of transformers (less than 10%) use Nynas oil. These figures show that the transformers under study cover a wide range of age, capacity, oil weight, and type of oil. It is also worth mentioning that all the transformers studied in this research have a voltage level of 63/20 kV.

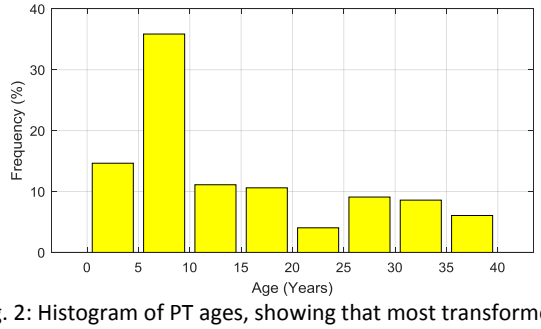


Fig. 2: Histogram of PT ages, showing that most transformers are between 5 and 10 years old.

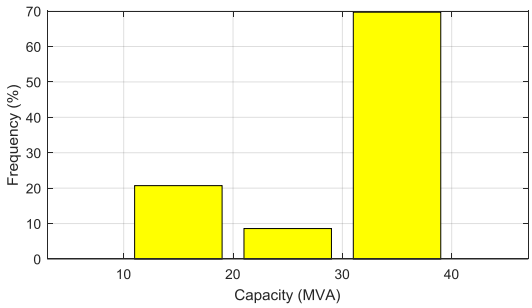


Fig. 3: Histogram of PT capacities, showing that the majority of transformers have a capacity between 30 and 40 MVA.

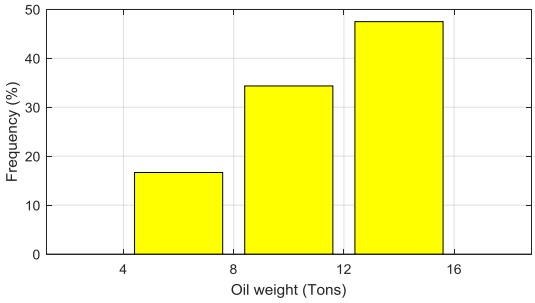


Fig. 4: Histogram of PT oil weights, showing that most transformers have an oil weight between 12 and 16 tons.

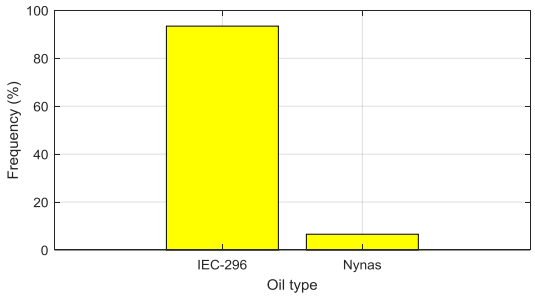


Fig. 5: Bar chart of PT oil types, showing that over 90% of transformers use IEC-296 oil.

Fig. 6 to Fig. 10 show scatter plots and histograms of DGA parameters for the three health classes of transformers so that you can look at how the data is spread out and check how well the statistical distribution extraction is working. The distributions of breakdown voltage and humidity, CO and CO_2, O_2 and N_2, and TCG and are shown in Fig. 6 to Fig. 9 for the three classes. Fig. 10 also displays the three classes' gas ratios (CH_4/C_2H_2, C_2H_6/CH_4, and C_2H_2/C_2H_4). Parameter transformer classes have varied DGA parameter distributions, as shown in these figures. Fig. 6 demonstrates a different distribution of "breakdown voltage" in the "healthy" class compared to the other two classes. Fig. 7 illustrates the distinct relationship between the "CO concentration" and "CO_2 concentration" across the three classes. Accurate modeling of key data requires identifying the appropriate statistical distribution for each. Each DGA parameter is analyzed across three transformer health classes. This ensures the synthetic data accurately represents the different health conditions of PTs.

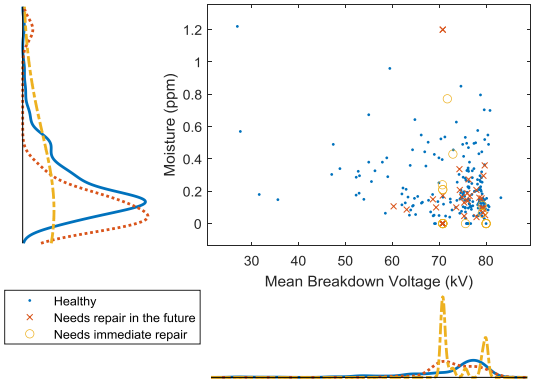


Fig. 6: Visualization of DGA data for breakdown voltage and moisture across three PT health Class.

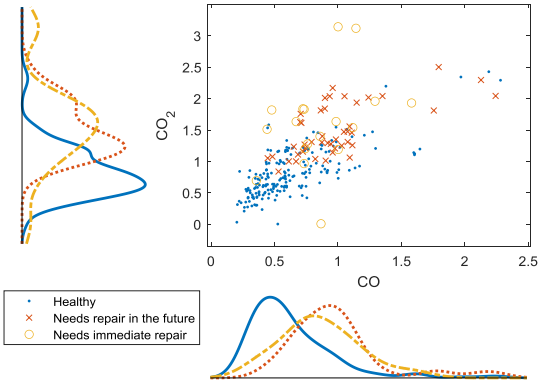


Fig. 7: Distribution of CO and CO_2 concentrations according across three PT health Class.

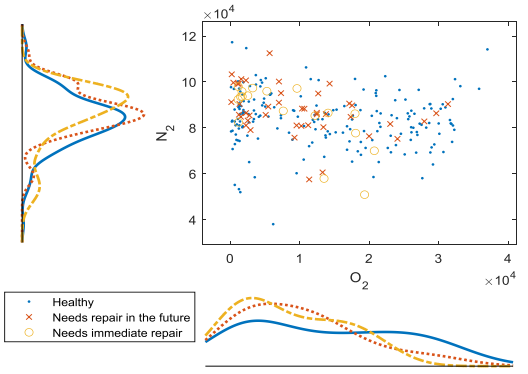


Fig. 8: Scatter plot and histogram depicting O_2 and N_2 concentrations across three PT health Class.

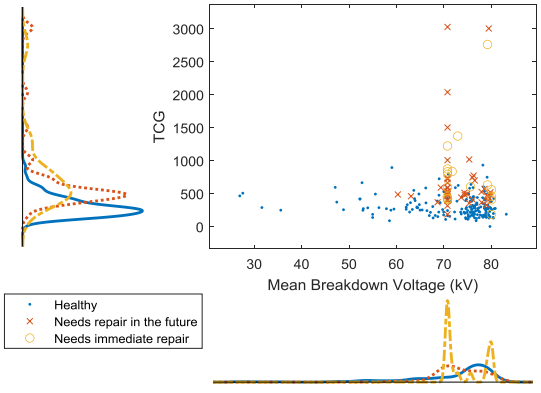


Fig. 9: Visualization of DGA data for TCG and breakdown voltage across three PT health Class.

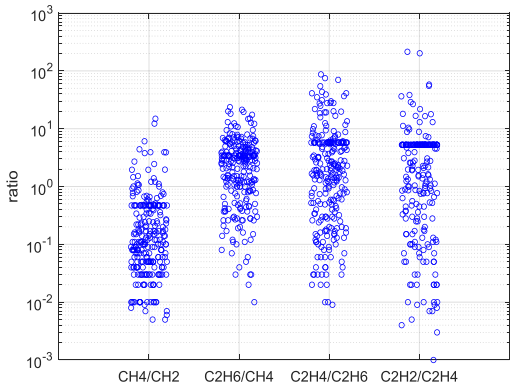


Fig. 10: Distribution of gas ratios (CH_4/C_2H_2, C_2H_6/CH_4, C_2H_4/C_2H_6, C_2H_2/C_2H_4) in DGA data.

These statistical distributions provide synthetic data and enable statistical analysis. The systematic technique encompassed the subsequent methods to ascertain the statistical distribution of each parameter: The normal distribution, log-normal distribution, Weibull distribution, gamma distribution, beta distribution, and exponential distribution were used to model the DGA data. Each class parameter's data was fitted to one of these distributions. The maximum likelihood estimation technique is used to estimate parameters. Each distribution's AIC is obtained after fitting to the data.

The exponential distribution is the best distribution for the "average breakdown voltage" parameter since it has the lowest AIC value in all classes. Similar analysis was done on the remaining other DGA parameters, identifying the optimum statistical distribution. Table 1 shows that DGA parameters in different classes have varied statistical distributions. The "Parameter" column of Table 1 indicates the 11 DGA metrics employed to evaluate transformer health. The parameters encompass breakdown voltage, moisture levels, concentrations of CO, CO2, O2, and N2 gases, along with the ratios of CH4/C2H2, C2H6/CH4, C2H4/C2H6, and C2H2/C2H4 gases ratios. The "Class" column indicates the three health classes of the transformer, which are: Healthy (H): The transformer is in a healthy condition. Needs retesting in the future (FR): The transformer is currently healthy, but it requires retesting in the future. Immediate retest required (IR): The transformer is in a critical condition and requires an immediate retest. The AIC criterion selects the best statistical distribution for each parameter in each class, as shown in the "Distribution" column. Using the Monte Carlo method, parameters for each distribution, like A (lower bound), B (upper bound), μ (mean), and π (standard deviation), are used to make fake data. As observed in the table, the statistical distribution of DGA parameters varies across different transformer health classes. This indicates that the health status of the transformer affects the statistical distribution of DGA parameters. As mentioned, this paper illustrates the impact of transformer health on the statistical distribution of DGA parameters. The parameter "CO concentration" has a log-normal distribution in the "healthy" class and a gamma distribution in the "needs immediate retesting" class. Table 1 shows that exponential and log-normal distributions are the most common for DGA parameters. The following are the optimal distributions for seven parameters in a variety of classes. Additionally, this implies that these distributions may be capable of modeling DGA data. "Data normalization" and "synthetic data generation" steps will be implemented subsequent to this.

B. Evaluation of Classifiers

This section evaluates all 12 classifiers utilized in Fig. 1. All classifiers were trained and tested with preprocessed DGA data during the evaluation process, utilizing various metrics to measure the performance of each classifier. To ensure the reliability and robustness of the assessment, a 100-fold randomized cross-validation technique was utilized.

Table 1: Statistical distributions of data across three transformer health classes

Oil Property/ Characteristic	Class	Best-Fit Distribution	Estimated Parameters of Best-Fit Distributions	
Average breakdown voltage 1 to 6	1	Weibull	A=75.33	B=12.1104
	2	Weibull	A=75.3733	B=19.4471
	3	Lognormal	μ= 4.30501	σ=0.0560327
Moisture	1	Exponential	μ=0.233556	
	2	Exponential	μ=0.145379	
	3	Exponential	μ=0.113236	
Average concentration of CO	1	Lognormal	μ=-0.576039	σ=0.463825
	2	Lognormal	μ=-0.0754555	σ= 0.341141
	3	Gamma	α=8.02887	β=0.107086
Average concentration of CO_2	1	Weibull	A=0.925048	B=2.19319
	2	Lognormal	μ=0.365075	σ =0.270727
	3	Normal	μ=1.59517	σ=0.74024
Average concentration of O_2	1	Exponential	μ=13764.9	
	2	Exponential	μ=9888.07	
	3	Exponential	μ=8486.44	
Average concentration of N_2	1	Normal	μ=83619.9	σ=11446.2
	2	Normal	μ=87500.9	σ=10307.7
	3	Weibull	A=90734.2	B=9.71992
Average TCG	1	Gamma	α=4.18118	β=73.8387
	2	Lognormal	μ=6.36514	σ=0.538263
	3	Lognormal	μ=6.46998	σ=0.588093
CH_4/CH_2 present in the oil	1	Exponential	μ=0.540598	
	2	Lognormal	μ=-2.66182	σ=1.63765
	3	Exponential	μ=0.12083	
C_2H_6/CH_4 ratio in the oil	1	Weibull	A=2.63218	B=0.853263
	2	Lognormal	μ= 1.4266	σ=0.596508
	3	Exponential	μ=5.73942	
C_2H_4/C_2H_6 ratio in the oil	1	Exponential	μ=7.13276	
	2	Exponential	μ=1.45367	
	3	Exponential	μ=1.31425	
C_2H_2/C_2H_4 ratio in the oil	1	Exponential	μ=6.48753	
	2	Exponential	μ=1.28539	
	3	Exponential	μ=1.68514	

The classifier was trained and evaluated on the training and testing sets by employing random data splitting for each cross-validation cycle. To obtain a reliable estimate of the classifier's performance, the results were subsequently averaged over 100 iterations. Each classifier's performance was measured using a variety of metrics, including accuracy, precision, and the confusion matrix. In addition to these conventional metrics, two new metrics were developed to address the importance of misclassifications in transformer fault diagnosis: the Risk Index and the Unnecessary Cost Index. The evaluation results are given in Table 2. This table shows the average and standard deviation of each assessment measure across 100 cross-validation iterations for each of the 12 classifiers. In the next subsections, we do a thorough examination of each classifier's performance, assessing its strengths and limitations across many evaluation measures. This investigation will provide useful insights into each classifier's suitability for the task of transformer defect diagnostics, as well as guidance in selecting the most successful algorithm for this application.

Table 2: Statistical distributions of data across three transformer health classes

Algorithm		Avg (%)		Max (%)		Min (%)		Avg Accuracy (%)	Confusion Matrix									Class	Average per class (%)	
									Avg			worst case			Best case					
		Risk	Cost	Risk	Cost	Risk	Cost												Accuracy	Precision
Ensemble Learning	Used in ensemble learning?	16	6	26	13	7	2	97	582	4	2	583	3	2	583	3	2	1	99	99
									3	122	7	5	118	9	5	118	9	2	92	92
									2	6	46	3	9	42	3	9	42	3	83	86
Support Vector Machine	Yes	23	6	34	11	13	2	97	582	4	2	585	2	1	585	2	1	1	99	99
									4	121	6	6	119	7	6	119	7	2	92	92
									3	7	44	7	9	38	7	9	38	3	85	81
K-Nearest Neighbor	Yes	24	7	34	12	13	3	96	577	9	2	579	9	0	579	9	0	1	99	98
									5	120	7	6	118	8	6	118	8	2	88	91
									3	8	43	6	10	38	6	10	38	3	83	79
Random Forest	Yes	24	2	33	5	14	0	97	586	1	1	587	1	0	587	1	0	1	99	100
									6	124	2	7	124	1	7	124	1	2	94	94
									3	8	43	5	10	39	5	10	39	3	94	80
Naive Bayes	Yes	38	35	61	53	22	18	87	552	20	16	557	17	14	557	17	14	1	98	94
									10	84	38	17	89	26	17	89	26	2	71	64
									3	14	37	5	21	28	5	21	28	3	42	69
Gaussian Naive Bayes	Yes	38	35	61	53	22	18	87	552	20	16	557	17	14	557	17	14	1	98	94
									10	84	38	17	89	26	17	89	26	2	71	64
									3	14	37	5	21	28	5	21	28	3	42	69
Decision Tree	Yes	42	9	62	18	22	4	93	575	8	5	572	13	3	572	13	3	1	97	98
									10	113	9	18	108	6	18	108	6	2	85	85
									6	13	35	8	18	28	8	18	28	3	72	66
RUSBoost	Yes	43	47	63	64	28	28	86	571	6	10	570	10	8	570	10	8	1	97	97
									14	60	58	20	63	49	20	63	49	2	77	45
									6	12	36	6	20	28	6	20	28	3	35	67
LPBoost	Yes	59	47	194	196	0	0	76	476	52	60	588	0	0	588	0	0	1	97	81
									10	86	36	124	8	0	124	8	0	2	NaN	65
									5	23	26	45	9	0	45	9	0	3	NaN	48
Multinomial Logistic Regression	No	83	9	95	15	68	4	90	576	10	2	578	8	2	578	8	2	1	97	98
									15	109	9	13	115	4	13	115	4	2	71	82
									5	34	15	7	39	8	7	39	8	3	60	28
Discriminant Analysis	No	86	14	101	21	71	7	89	577	10	1	580	8	0	580	8	0	1	95	98
									25	92	16	33	84	15	33	84	15	2	71	70
									8	28	18	6	35	13	6	35	13	3	53	33
Multiple Linear Regression	No	93	10	106	16	83	5	88	559	29	0	562	26	0	562	26	0	1	98	95
									10	115	7	15	111	6	15	111	6	2	62	87
									3	43	8	3	48	3	3	48	3	3	52	14
AdaBoostM2	No	119	1	130	2	112	0	89	583	5	0	583	5	0	583	5	0	1	94	99
									25	107	0	39	93	0	39	93	0	2	70	81
									12	42	0	9	45	0	9	45	0	3	NaN	0

SVM:

The aim of the SVM approach is to determine the best decision boundary that differentiates between various classes. The Radial Basis Function (RBF) kernel is used in this model. This method is a non-linear kernel that allows the model to learn non-linear decision boundaries. Also, Bayesian optimization modifies hyper-parameters of the model via the penalty parameter and kernel coefficient. In this manner, the optimal values for these parameters are selected to reduce the classification error on the validation data. Also, the data are normalized (given a zero mean and unit variance) before the SVM model is trained to lessen the effect of feature scale on model performance. A binary classifier is trained for each pair of classes using the "one-vs-one" method. In reality, the coding method is also regarded as a hyperparameter for issues involving more than two classes. Consequently, the aforementioned information are employed to train the algorithm on the training data. In 100 iterations, the SVM model obtained an average of 23% risk and 6% cost, as demonstrated by the Table 2 available. The risk associated with this method is 13% in the best-case scenario and 34% in the worst-case scenario. Additionally, the accuracy of this method is 99%, 92%, and 85% for the three classes, respectively. This procedure outperforms other methods in terms of accuracy and risk, as indicated by these findings. As shown by the average confusion matrix of this method, 44 of the 54 cases classified in class 3 were correctly identified, while 7 and 3 cases were classified in lesser classes, respectively. This demonstrates that the likelihood of this method misdiagnosing class 3 samples as classes 1 and 2 is relatively low. In addition, this method has a low risk of misdiagnosing class 2 samples as class 1, as only four of the 132 cases classified in class 2 were classified in class 1. The SVM method's overall risk and accuracy are 23% and 97% respectively, which are acceptable.

KNN:

The KNN model is being developed using training data. This function places training samples in the feature space. In order to forecast the class of the new sample, K samples in the feature space that are close to it are referenced. The distance (such as the Euclidean distance) is computed between the new sample and the training samples. Presumably, the prevailing class among the K neighbors is the expected label for the new sample. For this model, a range of options for k were assessed; cross-validation was used to identify the ideal value. For this model, various values for k were tested, and the best value was obtained using cross-validation. It is obvious from the Fig. 11 that the KNN method achieves the lowest level of risk when the number of neighbors (k) is equal to 1. So, the optimal number of neighbors in this method was determined to be k = 1.

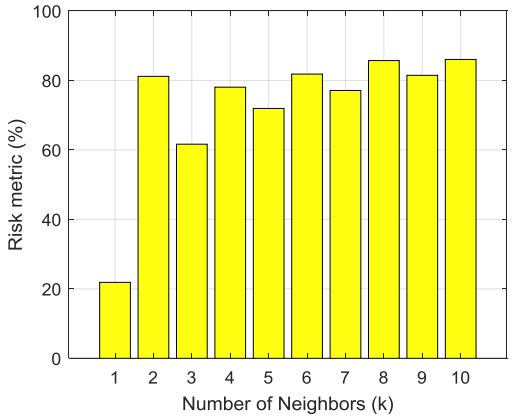


Fig. 11: Risk assessment of the KNN algorithm with different numbers of neighbors.

That is, the class of a new sample is predicted solely by utilizing its adjacent neighbor in the feature space. According to Table 2, this model has attained an average accuracy of 96%, a risk of 24%, and a cost of 7% over 100 iterations. This method carries a risk of 13 % in the best-case scenario and 34% in the worst-case scenario. In addition, the accuracy of this method for the three classes is 99%, 88%, and 83%, respectively. Based on these findings, this procedure outperforms other methods in terms of accuracy and risk. Based on the average confusion matrix of this method, 43 of the 54 cases in class 3 were correctly identified, while 8 and 3 cases were classified into lesser classes, respectively. This matter implies that the likelihood of this method misclassifying class 3 samples as classes 1 and 2 is relatively low. In addition, the method's minimal risk of misclassifying clas s 2 samples as class 1 is supported by the fact that only 5 of the 132 cases in class 2 were classified as class 1. The KNN method with k = 1 has an overall risk of 24%, which is deemed satisfactory.

Random Forest:

This study employs the random forest method to classify PT health. In the training phase, the algorithm generates many decision trees and determines a class based on the average of these classifications or the mean prediction (regression) of the individual trees. The classifier is trained with a random forest model including 100 decision trees. Each decision tree is trained using random features and a random subset of the training data. Bagging is the technique employed to do this. This technique enhances tree diversity and mitigates overfitting. hence, it strengthens the model and increases its applicability across diverse scenarios. According to Fig. 1, a 100-fold randomized cross-validation method is used to evaluate the random forest model performance. The results, shown in the Table 2, show that the model was accurate 97% of the time, with a risk index of 24% and an unnecessary cost index of 2%. The model showed that the

worst-case risk index is 33%, and the best-case risk index is 14%. The model is 99% accurate for the "Healthy" class, 94% accurate for the " needs retesting in the future" class, and 94% accurate for the "Needs Immediate Retesting" class. Due to its poor performance compared to SVM and KNN, this method placed third on the Risk Index. The average confusion matrix shows that out of 54 cases in Class 3, 43 were correctly identified as being in that class, 8 were mistakenly put in Class 2, and 3 were mistakenly put in Class 1. This means that the system did a pretty good job of finding major faults (Class 3). Also, out of 132 cases in Class 2, 124 were correctly classified. Six were wrongly labeled as Class 1, and two were wrongly labeled as Class 3. This indicates a minimal likelihood of erroneously classifying transformers that require retesting in the future. The Random Forest model achieves an adequate equilibrium among accuracy, risk, and cost. Nonetheless, its comparatively elevated risk score compared to SVM and KNN indicates that it has the potential to more effectively identify defective transformers.

Other classifiers:

In this part, we look at how well the classifiers used in this study worked. These are Naive Bayes, Decision Tree, RUSBoost, Gaussian Naive Bayes, LPBoost, Multinomial Logistic Regression, Discriminant Analysis, AdaBoostM2, and Multiple Linear Regression. By examining the Table 2, we can analyze the performance of these classifiers in relation to the risk index, the unnecessary cost index, and their corresponding confusion matrices. For instance, the risk index of the Naive Bayes classifier is 38%, whereas the unnecessary cost index is 35%. On the other hand, the decision tree classifier displays an unnecessary cost index of 9% and a danger index of 42%. Furthermore, a review of the confusion matrices reveals that certain classifiers do better in accurately categorizing transformers as belonging to class 3. The SVM, KNN and Random Forest classifiers notably identify a greater number of Class 3 instances than the Naive Bayes classifier. The classifier selection relies on the application and the relative relevance of accuracy, risk, and cost. For example:

- A classifier with a lower risk index should be preferred to reduce the danger of misclassifying damaged transformers as healthy.
- Choose a classifier with a lower unnecessary cost index to reduce costs from misclassifying healthy transformers as faulty.

The following points can be made based on the data of Table 2.

- Risk:
  - Risk classifiers between 25% and 50%:
    - Naive Bayes (38%)
    - Gaussian Naive Bayes (38%)
    - Decision Tree (42%)
    - RUSBoost (43%)
  - Risk classifiers between 50% and 75%:
    - LPBoost (59%)
  - Classifiers with a risk greater than 75%:
    - Multinomial Logistic Regression (83%)
    - Discriminant Analysis (86%)
    - Multiple Linear Regression (93%)
    - AdaBoostM2 (119%)
- Cost:
  - Classifiers with less than 10% cost:
    - Random Forest (2%)
    - Decision Tree (9%)
    - Multinomial Logistic Regression (9%)
    - Multiple Linear Regression (10%)
  - Classifiers with more than 10% cost:
    - Discriminant Analysis (14%)
    - Gaussian Naive Bayes (35%)
    - Naive Bayes (35%)
    - LPBoost (47%)
    - RUSBoost (47%)
- Confusion matrix (for class 3):
  - Naive Bayes:
    - Out of 54 samples in class 3, 37 are correctly identified, and 14 are classified in class 2 and 3 in class 1.
  - Decision Tree:
    - Out of 54 samples in class 3, 35 items are correctly identified, 13 items are classified in class 2, and 6 items are classified in class 1.
  - RUSBoost:
    - Out of 54 samples in class 3, 36 are correctly identified, and 12 are classified in class 2 and 6 in class 1.
  - LPBoost:
    - Out of 54 samples in class 3, 26 are correctly identified, and 23 are classified in class 2 and 5 in class 1.
  - Multinomial Logistic Regression:
    - Out of 54 samples in class 3, 15 cases are correctly identified, 34 cases are classified in class 2, and 5 cases are in class 1.
  - Discriminant Analysis:
    - Out of 54 samples in class 3, 18 items are correctly identified, 28 items are classified in class 2, and 8 items are classified in class 1.
  - AdaBoostM2:
    - Out of 54 samples in class 3, 0 items are correctly identified, 42 items are classified in class 2, and 12 items are classified in class 1.
  - Multiple Linear Regression:
    - Out of 54 samples in class 3, 8 cases are correctly identified, 43 cases are classified in class 2, and 3 cases are in class 1.

Ensemble learning based classifiers:

This section shows the performance of the suggested ensemble learning model, which uses the top eight classifiers. This model combines classifier predictions to improve detection. This approach uses SVM, KNN with k=1, Random Forest, Naive Bayes, Decision Tree, RUSBoost, Gaussian Naive Bayes, and LPBoost. The selection of these models was based on their exceptional performance and diversity in categorizing DGA data.

In the first subfig. 1 of Fig. 12, the risk indices of the 12 used classifiers are shown. This figure shows a significant break in the risk chart after the eighth classifier, indicating an increase in risk values to high and unacceptable levels. In other words, the first eight classifiers have significantly lower risk compared to the next four classifiers. For this reason, the top 8 classifiers have been selected as the best classifiers and have been used in the ensemble learning algorithm. So, it is possible to create an accurate and reliable detection model by combining the predictions of these low-risk classifiers. Moreover, eliminating four high-risk classifiers mitigates the excessive escalation of model complexity and preserves its speed and efficiency. Table 2 shows that the model reduced risk by 7% relative to the best single classifier (23%) over 100 iterations, achieving a risk level of 16%. The proposed method exhibits risk index in range of 7% in the best-case scenario to 26% in the worst-case scenario. The proportion of unnecessary expenses is observed to range from 2% to 13%, yielding an average value of 6%. In addition to the fact that the ensemble method offers less risk compared to the results of the individual method, it achieves an average accuracy of 97%. The accuracy attained in class 3 of the ensemble technique exceeds that of the individual classifier methods. This issue stems from prioritizing method selection intended to reduce the risk index. In the typical confusion matrix for this approach, 54 occurrences in class 3 were assessed, leading to 46 instances being correctly identified. In contrast, 6 cases were wrongly put in class 2, and 2 cases were wrongly put in class 1. This proves that the suggested method works very well at finding samples that belong to class 3. Only 3 of the 132 class 2 cases were misclassified; the other 121 were correctly classified. This demonstrates that using this method, there is a very low risk of incorrectly identifying class 2 samples. The ensemble learning method has the lowest risk, at 16% (14% for class 3 and 2% for class 2), of all the methods this study looked at. In addition, the unnecessary cost index that is related to this method works better than other methods, except for Random Forest and Decision Tree. Because this method has a low-risk rating and doesn't cost too much, it can be assumed that it is a trustworthy way to check the health of transformers. Fig. 12 shows how well the suggested ensemble learning model works compared to different algorithms. Looking at other methods, the ensemble learning approach has a relatively lower risk level. While reducing the chance of high-risk mistakes, this study shows that the suggested method is both efficient and effective at accurately figuring out the health status of PT.

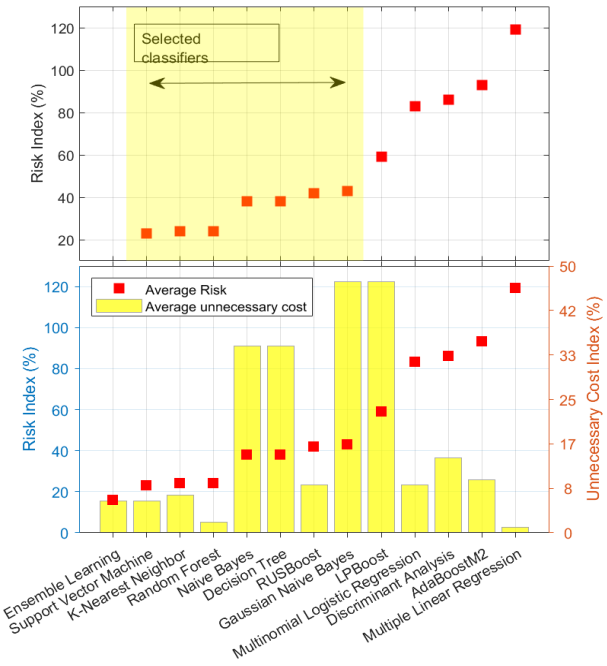


Fig. 12: Performance evaluation of the ensemble learning model compared to other methods.

Furthermore, the unnecessary cost index associated with this method is maintained at a level that is deemed acceptable when compared to the majority of alternative methods. This outcome demonstrates its efficacy in minimizing excessive expenditures. Overall, the ensemble learning model demonstrates commendable performance regarding the risk index and unnecessary cost index, indicating its efficacy in diagnosing the health status of the transformer.

Conclusion

A new ensemble learning technique for PT health assessment (dividing them into three groups: "healthy," "needs retesting in the future," and "needs immediate retesting.") using DGA data has been proposed in this paper. The algorithm combines the performance of eight classifiers (such as: SVM, KNN, Random Forest, Naive Bayes, Decision Tree, RUSBoost, Gaussian Naive Bayes, and LPBoost) and considers errors and costs to create a transformer health diagnosis model with high accuracy and reliability. The proposed approach is evaluated by the risk index, unnecessary cost index, and accuracy, which demonstrate superior performance compared to previous techniques across these metrics. The risk index has decreased as a result. This finding suggests that it can identify defective transformers more accurately. This

model can minimize transformer maintenance expenses and extends their lifespan. This study found overall accuracy 97%, average risk index 16%, unnecessary cost index 6%, best-case risk 7%, and worst-case risk 26%. The findings show that the suggested approach efficiently evaluates transformer health, reducing the expenses and dangers of incorrect diagnosis. The optimal result of single classifier (SVM), is an average risk index of 23. The comparison of these values with the average risk index of 16% in the proposed ensemble learning technique demonstrates its efficiency and effectiveness in reducing misdiagnosis. In addition, the method improved the accuracy in the worst-case class. This improvement is particularly significant due to the importance of this class. It achieved an accuracy of 86%, compared to the SVM and KNN techniques (with recorded accuracy of 81% and 79%, respectively). Hence, the results show that the ensemble learning method works better at finding healthy transformers than single-classifier methods. This approach can be further improved through the incorporation of additional data and the application of sophisticated machine learning techniques. This methodology is capable of identifying additional transformer and electrical network anomalies.

Future Work

This study provides a number of directions for further investigation. The authors suggest that the following should be accounted for in future work:

- In this paper, a simple classification method with three classes was used. In future works, hierarchical classification methods can be used for a more precise categorization of transformer health status. For example, the class "Need for Retesting" can be divided into two subclasses: "Need for Retesting in the Near Future" and "Need for Retesting in the Distant Future".
- The DGA data used in this paper includes a part of the total technical information regarding transformers. In the future works, data can be collected with more detailed technical information, such as the type of transformer, the age of the transformer, and environmental conditions. This work can help improve the accuracy of machine learning models.
- In the proposed method, 11 DGA parameters were used to train machine learning models. Accordingly, the impact of feature selection on parameters can be examined for the future works. For example, methods such as PCA or mutual information-based approaches can be used to select the most important features.
- The proposed method is established based on traditional machine learning algorithms such as SVM and KNN. The deep learning algorithms, such as convolutional neural networks (CNN), can be used for fault detection in transformers in future works. These algorithms can improve fault detection accuracy (due to their high capability in learning complex features from data).
- The proposed method in this paper is established based on the certainty and accuracy of the initial data and the correct evaluation by the expert. In future works, the impact of uncertainty in DGA data on the performance of machine learning models can be examined. Additionally, methods based on uncertainty learning can be used to improve the accuracy of models.

Author Contributions

K. Gorgani Firouzjah and J. Ghasemi implemented the methods and evaluated their performance. Also, they wrote the paper, coordinated the study and contributed to the analysis of the results. They edited the manuscript. In addition, authors revised and discussed the results and approved the final manuscript.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

AdaBoostM2	Adaptive Boosting
AIC	Akaike Information Criterion
CH_4	Methane
C_2H_2	Acetylene
C_2H_4	Ethylene
C_2H_6	Ethane
CNN	Convolutional Neural Networks
CO_2	Carbon Dioxide
CO	Carbon Monoxide
DGA	Dissolved Gas Analysis
H_2	Hydrogen
KGM	Key Gas Method
KNN	K-Nearest Neighbors
LPBoost	Linear Programming Boosting
NN	Neural Network
O_2	Oxygen
PT	Power Transformer
RBF	Radial Basis Function
RUSBoost	Under-Sampling Boosting
SVM	Support Vector Machines
TCG	Total Combustible Gases

References

[1] Y. Wang, S. Gong, S. Grzybowski, "Reliability evaluation method for oil-paper insulation in power transformers," Energies, 4(9): 1362-1375, 2011.

[2] D. G. T. Da Silva, H. J. B. Da Silva, F. P. Marafão, H. K. M. Paredes, F. A. S. Gonçalves, "Enhanced health index for power transformers diagnosis," Eng. Fail. Anal., 126: 105427, 2021.

[3] C. Xia, M. Ren, B. Wang, M. Dong, G. Xu, J. Xie, C. Zhang, "Infrared thermography-based diagnostics on power equipment: State-of-the-art," High Volt., 6(3): 387-407, 2021.

[4] F. M. Laburú, T. W. Cabral, F. V. Gomes, E. R. de Lima, J. C. S. S. Filho, L. G. P. Meloni, " New insights into gas-in-oil-based fault diagnosis of power transformers," Energies , 17(12): 2889, 2024.

[5] S. Yu, D. Zhao, W. Chen, H. Hou, "Oil-immersed power transformer internal fault diagnosis research based on probabilistic neural network," Procedia Comput. Sci., 83: 1327-1331, 2016.

[6] L. Cheng, T. Yu, "Dissolved gas analysis principle-based intelligent approaches to fault diagnosis and decision making for large oil-immersed power transformers: A survey," Energies, 11(4): 913, 2018.

[7] Z. A. A. Etman, D. E. A. Mansour, N. H. El-Amary, "Performance evaluation of dissolved gas analysis techniques against measurement errors," in Proc. 2017 IEEE 19th International Conference on Dielectric Liquids, 2017.

[8] V. Tra, B. P. Duong, J. M. Kim, "Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data," IEEE Trans. Dielectr. Electr. Insul., 26(4): 1325-1333, 2019.

[9] H. C. Sun, Y. C. Huang, C. M. Huang, "A review of dissolved gas analysis in power transformers," Energy Procedia, 14: 1220-1225, 2012.

[10] M. S. Ali, A. H. Abu Bakar, A. Omar, A. S. Abdul Jaafar, S. H. Mohamed, "Conventional methods of dissolved gas analysis using oil-immersed power transformer for fault diagnosis: A review," Electr. Power Syst. Res., 16: 109064, 2023.

[11] M. Duval, "The duval triangle for load tap changers, non-mineral oils and low temperature faults in transformers," IEEE Electr. Insul. Mag., 24(6): 22-29, 2008.

[12] N. Poonnoy, C. Suwanasri, T. Suwanasri, "Fuzzy logic approach to dissolved gas analysis for power transformer failure index and fault identification," Energies, 4(1): 36, 2021.

[13] O. E. Gouda, S. H. El-Hoshy, S. S. M. Ghoneim, "Enhancing the diagnostic accuracy of DGA techniques based on IEC-TC10 and related databases," IEEE Access, 9: 118031-118041, 2021.

[14] A. M. Aciu, C. I. Nicola, M. Nicola, M. C. Nițu, "Complementary analysis for dga based on duval methods and furan compounds using artificial neural networks," Energies, 14(3): 588, 2021.

[15] Hartono, Y. Muharni, C. Adipura, W. Martiningsih, M. Otong, M. Irvan, "Analysis of Power Transformator Conditions Using Dga Method Using Artificial Neural Network in Krakatau Electrical Power Company," Int. J. Eng. Technol. Manag. Res., 7(6): 77-88, 2020.

[16] D. A. Barkas, S. D. Kaminaris, K. K. Kalkanis, G. C. Ioannidis, C. S. Psomopoulos, "Condition assessment of power transformers through DGA measurements evaluation using adaptive algorithms and deep learning," Energies, 16(1): 54, 2023.

[17] S. J. T. Shahrabad, V. Ghods, M. T. Askari, "Power transformer fault diagnosis using DGA and artificial intelligence," Recent Adv. Comput. Sci. Commun., 3(4): 579-587, 2019.

[18] J. Li, Q. Zhang, K. Wang, J. Wang, T. Zhou, Y. Zhang, "Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine," IEEE Trans. Dielectr. Electr. Insul., 23(2): 1198-1206, 2016.

[19] S. Genc, S. Karagol, "Fuzzy logic application in DGA methods to classify fault type in power transformer," in Proc. IEEE 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA 2020), 2020.

[20] A. Abu-Siada, S. Hmood, S. Islam, "A new fuzzy logic approach for consistent interpretation of dissolved gas-in-oil analysis," IEEE Trans. Dielectr. Electr. Insul., 20(6): 2343-2349, 2013.

[21] Y. Zhang, Y. Tang, Y. Liu, Z. Liang, "Fault diagnosis of transformer using artificial intelligence: A review," Front. Energy Res., 10: 1006474, 2022.

[22] H. Zheng, Y. Zhang, J. Liu, H. Wei, J. Zhao, R. Liao, "A novel model based on wavelet LS-SVM integrated improved PSO algorithm for forecasting of dissolved gas contents in power transformers," Electr. Power Syst. Res., 155: 196-205, 2018.

[23] S. Souahlia, K. Bacha, A. Chaari, "SVM-based decision for power transformers fault diagnosis using Rogers and Doernenburg ratios DGA," in Proc. IEEE 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), 2013.

[24] O. Kherif, Y. Benmahamed, M. Teguar, A. Boubakeur, S. S. M. Ghoneim, "Accuracy improvement of power transformer faults diagnostic using KNN classifier with decision tree principle," IEEE Access, 9: 81693-81701, 2021.

[25] Z. B. Sahri, R. B. Yusof, "Support vector machine-based fault diagnosis of power transformer using k nearest-neighbor imputed DGA dataset," J. Comput. Commun., 2(9): 22-31, 2014.

[26] Y. Benmahamed, O. Kherif, M. Teguar, A. Boubakeur, S. S. M. Ghoneim, "Accuracy improvement of transformer faults diagnostic based on DGA data using SVM-BA classifier," Energies, 14(10): 2970, 2021.

[27] U. M. Rao, I. Fofana, K. N. V. P. S. Rajesh, P. Picher, "Identification and application of machine learning algorithms for transformer dissolved gas analysis," IEEE Trans. Dielectr. Electr. Insul., 28(5): 1828-1835, 2021.

[28] N. Haque, A. Jamshed, K. Chatterjee, S. Chatterjee, "Accurate sensing of power transformer faults from dissolved gas data using random forest classifier aided by data clustering method," IEEE Sens. J., 22(6): 5902-5910, 2022.

[29] Y. Benmahamed, M. Teguar, A. Boubakeur, "Diagnosis of power transformer oil using PSO-SVM and KNN classifiers," in Proc. IEEE 3rd International Conference on Electrical Sciences and Technologies in Maghreb, CISTEM 2018, 2018.

[30] H. Yousefpour, J. Ghasemi, "Ensemble-based detection and classification of liver diseases caused by hepatitis c," Contrib. Sci. Technol. Eng., 1(1): 33-43, 2024.

Biographies

Khalil Gorgani Firouzjah was born in Babolsar, Mazandaran, Iran. He received his B.S. degree in Electrical Engineering from Ferdowsi University of Mashhad in 2005, and his M.S. and Ph.D. degrees in Electrical Engineering from the University of Mazandaran in 2008 and 2013, respectively. He served as a Researcher at Mazandaran Regional Electric Company from 2009 to 2011. He then joined the University of Mazandaran in 2014 as an Assistant Professor in the Electrical Engineering Department, and was promoted to Associate Professor in the same department in 2018. He continues to serve as an Associate Professor at the University of Mazandaran. His research focuses on power systems, energy management, and the application of machine learning and optimization algorithms within the power systems.

- Email: k.gorgani@umz.ac.ir
- ORCID: 0000-0002-7132-6759
- Web of Science Researcher ID: AAN-9973-2020
- Scopus Author ID: 55364934100
- Homepage: https://rms.umz.ac.ir/~kgorgani/

Jamal Ghasemi received his Bachelor's, Master's, and Ph.D. degrees in Electrical Engineering from the University of Mazandaran in 2004, 2008, and 2012, respectively. Since 2012, he has been a faculty member at the University of Mazandaran, where he currently serves as an Associate Professor. His research interests include fuzzy theory, evidence theory, image processing, and pattern recognition.

- Email: j.ghasemi@umz.ac.ir
- ORCID: 0000-0002-9573-0107
- Web of Science Researcher ID: AAN-1682-2020
- Scopus Author ID: 26032012000
- Homepage: https://rms.umz.ac.ir/~jghasemi/

How to cite this paper:
F K. Gorgani Firouzjah, J. Ghasemi, "Ensemble learning algorithm for power transformer health assessment using dissolved gas analysis," J. Electr. Comput. Eng. Innovations, 13(2): 387-402, 2025.

DOI: 10.22061/jecei.2025.11450.805

URL: https://jecei.sru.ac.ir/article_2279.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Enhancing Multi-Entity Detection and Sentiment Analysis in Financial Texts with Hierarchical Attention Networks

L. Hafezi, S. Zarifzadeh*, M. R. Pajoohan

Computer Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran.

Article Info

Article History:
Received 18 November 2024
Reviewed 21 January 2025
Revised 04 February 2025
Accepted 05 February 2025

Keywords:
Hierarchical attention network
Entity detection
Sentiment analysis

*Corresponding Author's Email Address:
szarifzadeh@yazd.ac.ir

Abstract

Background and Objectives: Detecting multiple entities within financial texts and accurately analyzing the sentiment associated with each is a challenging yet critical task. Traditional models often struggle to capture the nuanced relationships between multiple entities, especially when sentiments are context-dependent and spread across different levels of a document. Addressing these complexities requires advanced models that can not only identify multiple entities but also distinguish their individual sentiments within a broader context. This study aims to introduce and evaluate two novel methods, ENT-HAN and SNT-HAN, built upon the Hierarchical Attention Networks, specifically designed to enhance the accuracy of both entity extraction and sentiment analysis in complex financial documents.

Methods: In this study, we design ENT-HAN and SNT-HAN methods to address the tasks of multi-entity detection and sentiment analysis within financial texts. The first method focuses on entity extraction, where capture hierarchical relationships between words and sentences. By utilizing word-level attention, the model identifies the most relevant tokens for recognizing entities, while sentence-level attention helps refine the context in which these entities appear, allowing the model to detect multiple entities with precision. The second method is applied for sentiment analysis, aiming to classify sentiments into positive, negative, or neutral categories. The sentiment analysis model employs hierarchical attention to identify the most important words and sentences that convey sentiment about each entity. This approach ensures that the model not only focuses on the overall sentiment of the text but also accounts for context-specific variations in sentiment across different entities. Both methods were evaluated on FinEntity dataset, and the results demonstrate their effectiveness, with significantly improving the accuracy of both entity extraction and sentiment classification tasks.

Results: The ENT-HAN and SNT-HAN demonstrated strong performance in both entity extraction and sentiment analysis, outperforming the methods they were compared against. For entity extraction, ENT-HAN was evaluated against RNN and BERT models, showing superior accuracy in identifying multiple entities within complex texts. In sentiment analysis, SNT-HAN was compared to the best-performing method previously applied to FinEntity dataset. Despite the good performance of the existing methods, SNT-HAN demonstrated superior results, achieving a better accuracy.

Conclusion: The outcome of this research highlights the potential of the ENT-HAN and SNT-HAN for improving entity extraction and sentiment analysis accuracy in financial documents. Their ability to model attention at multiple levels allows for a more nuanced understanding of text, establishing them as a valuable resource for complex tasks in financial text analysis.

Introduction

In the realm of natural language processing (NLP), sentiment analysis has progressed from broad evaluations of textual sentiment to more precise, entity-level sentiment analysis, reflecting the growing need for detailed insights that businesses, social platforms, and

researchers seek to better understand public opinion. This granular approach is particularly crucial in the financial domain, where sentiment analysis has become an indispensable tool for deciphering market trends, investor behavior, and overall economic sentiment [1]. By focusing on specific financial entities such as companies, stocks, or market indices within a text, entity-level sentiment analysis provides a more nuanced understanding than traditional document-level or sentence-level analyses, offering critical insights into how individual entities are perceived [2]. These insights are essential for predicting market movements and evaluating investor sentiment with a greater accuracy.

The challenge in this domain is twofold: first, accurately identifying financial entities within complex and often ambiguous language; second, determining the sentiment directed at each entity, which can vary significantly within different contexts. Traditional methods, which often rely on rule-based approaches or classical machine learning, struggle to capture the intricate dependencies and nuances present in financial texts, especially when multiple entities are discussed with differing sentiments. For example, in this paragraph: "Among the highlights of this week, Macy's raised its annual profit forecast, easing some investor worries over consumer spending after recent disappointments on the earnings front from Walmart and other big names.", several financial entities are referenced, but their sentiments differ, with a positive sentiment for Macy's and a negative sentiment for Walmart.

To address the above challenges, this paper proposes utilizing the Hierarchical Attention Network (HAN) [3], suggesting the ENT-HAN method for multi-entity detection and the SNT-HAN method for sentiment analysis. These methods are designed to model hierarchical structures in the text, capturing both word-level and sentence-level dependencies. The model consists of two key attention mechanisms: one at the word level and another at the sentence level. At the word level, the attention mechanism identifies the most relevant words within each sentence, assigning higher weights to words that contribute more significantly to the task at hand, such as entity extraction or sentiment detection. These weighted word representations are then aggregated into a sentence representation. At the sentence level, the attention mechanism operates similarly, focusing on the most informative sentences within the text, producing a document-level representation. This hierarchical structure allows the models to focus on the most critical parts of a text, enabling them to handle longer and more complex documents more effectively. The models are particularly suited for tasks involving multiple entities or sentiments, as they capture context at both granular and holistic levels.

Entity extraction plays a pivotal role in the sentiment analysis process, as it enables the precise identification of specific entities within text, allowing for a more targeted and accurate assessment of sentiment. By isolating entities, sentiment analysis can be more effectively tailored to evaluate opinions and emotions associated with particular individuals, organizations, or products, thus enhancing the overall reliability and relevance of the analysis [4]. When the critical step of accurately identifying multiple entities within a text is successfully completed, it becomes feasible to proceed to the sentiment analysis of each entity with greater ease and confidence. Successfully addressing the first challenge of entity extraction lays a strong foundation, allowing the second challenge—analyzing the sentiment associated with each entity—to be tackled more effectively and with higher reliability.

By applying ENT-HAN and SNT-HAN, we aim to enhance the precision of entity extraction and enhance the accuracy of sentiment analysis at the entity level. The proposed models are evaluated on FinEntity dataset[1] [5], with a comparison to existing methodologies, demonstrating its superior performance in extracting multiple entities dynamics and sentiments associated with each entity in financial texts.

Related Work

This section primarily reviews the most relevant papers on financial sentiment analysis and event-based sentiment analysis.

A. Financial Sentiment Analysis

Sentiment analysis has evolved an essential tool within the financial industry, providing valuable insights into market trends, investor behavior, and economic outlooks [6]. Unlike general sentiment analysis, which focuses on broader textual sentiment, financial sentiment analysis specifically targets the emotions and opinions expressed about financial entities, such as companies, stocks, and market indices. The unique characteristics of financial texts—often laden with technical jargon, abbreviations, and context-dependent phrases—demand specialized approaches that can accurately capture and interpret sentiment within this domain. Over the past decade, numerous studies have explored various methods for extracting and analyzing sentiment in financial contexts, from traditional machine learning techniques to more recent advancements in natural language processing and deep learning. This section examines the most influential works that have shaped the

[1] The FinEntity dataset is publicly accessible at https://github.com/yixuantt/FinEntity

current landscape of financial sentiment analysis, highlighting key methodologies and their impact on the field.

One of the pioneering studies applying deep learning techniques to financial polarity analysis was conducted by Kraus and Feuerriegel [7]. They employed a Long Short-Term Memory (LSTM) neural network to analyze ad-hoc company announcements and predict stock market movements, demonstrating that their method outperformed conventional machine learning approaches. Several other research efforts have explored diverse neural network architectures for financial sentiment analysis. Sohangir et al. [8] tested diverse neural network models on a StockTwits dataset and discovered that Convolutional Neural Networks (CNNs) yielded the highest level of performance. Lutz et al. [9] utilized doc2vec to generate sentence embeddings from company-specific announcements and employed multi-instance learning to forecast stock market results. Additionally, Maia et al. [10] combined text simplification with LSTM networks to classify sentences from financial news by sentiment, attaining cutting-edge outcomes in the Financial PhraseBank dataset.

While advanced deep learning approaches and specialized language models have been widely adopted to enhance sentiment analysis in finance, tailored versions of BERT for specific fields, such as FinBERT [11], have markedly enhanced financial sentiment analysis by being finely tuned for financial texts, leading to improved reliability and accuracy. Fatouros et al. [12] used a zero-shot prompting method, they evaluate various ChatGPT prompts on a carefully selected dataset of forex news headlines for sentiment classification. Moreover, they investigate the relationship between predicted sentiment and market returns as an additional evaluation metric. In comparison to FinBERT, a well-regarded model for financial sentiment analysis, ChatGPT demonstrated roughly 35% better performance in sentiment classification and a 36% stronger correlation with market returns. Ardekani et al. [13] developed FinSentGPT, a financial sentiment forecasting model built on a refined version of ChatGPT. Evaluating it on U.S. media news and a multilingual dataset from the European Central Bank, they found that FinSentGPT matches the performance of a top English finance sentiment model, outperforms a traditional machine learning model, and accurately predicts sentiment across different languages. This suggests that sophisticated large language models can provide adaptable and context-sensitive financial sentiment analysis across languages.

Luo and Mo in their paper [14] investigated sentiment towards the 45th President of the United States in news articles employing a novel entity sentiment analysis model known as the Negative Sentiment Smoothing Model (NSSM). The NSSM model adjusts sentiment scores by accounting for Negative Associated Entities (NAEs), that are entities linked to negative sentiments within the data. Three versions of the NSSM model (NSSM-A, NSSM-B, and NSSM-C) were developed using a smoothing algorithm. The study focused on "Trump" as the target entity and assessed the effectiveness of the NSSM models on a dataset comprising 10,993 paragraphs of news related to the target entity, gathered from CNN, FOX, and NPR over a three-month span from July 1, 2019, to September 30, 2019. The highest accuracy was achieved by NSSM-B, with an accuracy rate of 85.96%.

B. Entity-based Sentiment Analysis

Effective entity extraction is crucial in the sentiment analysis process, as it enables the precise identification of the subjects or entities to which sentiments are directed. By accurately extracting entities, sentiment analysis can yield more targeted and contextually relevant insights, thereby improving the overall accuracy and reliability of the analysis. This section will examine the most influential works in the domain of entity-based sentiment analysis.

Poria et al. [15] proposed an innovative deep learning approach for entity extraction in opinion mining, a critical task in sentiment analysis that focuses on identifying specific targets of opinions within a text, such as the attributes of a product or service being evaluated. They utilized a 7-layer deep convolutional neural network to classify each word in opinionated sentences as either an aspect or a non-aspect word. To further refine this process, they incorporated a set of linguistic patterns into the neural network. This hybrid classifier, coupled with a word-embedding model, was designed for sentiment analysis. The dataset used for training and testing spanned two domains: Laptop and Restaurant. The entity extraction framework achieved F-scores of 82% for the laptop domain and 87% for the restaurant domain.

Zhao et al. [16] introduce a method for sentiment analysis and key entity detection that utilizes BERT, tailored specifically for mining financial texts and analyzing public opinion on social media platforms. Their approach begins by using a pre-trained model to perform sentiment analysis, after which key entity detection is approached as a sentence matching or Machine Reading Comprehension (MRC) task at various levels of granularity, with a primary focus on identifying negative sentiment. Their approach employed RoBERTa as the pre-training model. Furthermore, they found that fine-tuning the pre-trained model yields better results than using it to generate sentence-level vectors for downstream models in their specific task. In the end, incorporating ensemble methods and focal loss further enhances performance to a certain degree. The experimental results show F1 scores of 95% and 85% on two publicly available financial negative entity recognition datasets, indicating that their

method significantly outperforms traditional approaches. Additionally, this method demonstrates 96% accuracy in sentiment analysis.

Li et al. [17] introduce an innovative approach called the Twin Towers End-to-End model (TTEE) to address the Target and Aspect-Sentiment Detection (TASD) challenge. The TTEE model simplifies the complex TASD task by employing an end-to-end multi-task framework that concurrently handles target detection and aspect-sentiment classification. It utilizes a twin towers architecture, based on BERT or its advanced variants, to effectively separate the context from the given aspects, thereby reducing redundant calculations and significantly enhancing computational efficiency. This approach provides a distinct advantage in identifying implicit target entities and their associated aspects and sentiments within the context, without the need for additional model architectures. The highest F1 measure achieved for entity extraction with this model is 68.39%.

Wan et al. [18] propose a Span-based Multi-Modal Attention Network (SMAN) to address the joint task of entity and relation extraction. Although this study performs entity extraction on the data, the results are not utilized for sentiment analysis. Nonetheless, due to its focus on entity extraction, it has been included in this section. In the Entity Recognition stage, the model identifies two types of features: entity features (spans) represented by span units and contextual features (tokens) derived from the surrounding text. During the Relation Extraction (RE) stage, the model incorporates three types of features: entity features (spans), entity type features (labels), and contextual features (tokens). To enhance the modeling of interactions between these modalities, SMAN first generates unique representations for span units while capturing high-dimensional contextual features using a cloze mechanism. This mechanism allows the model to mask central span units, effectively learning contextual dependencies. Furthermore, entity type labels assist in refining relationship predictions, simulating the human reasoning process of associating entity types with potential relationships.

The proposed SMAN architecture includes a Modal-Enhanced Attention (MEA) module designed to model the contextual dependencies within single-modal data and facilitate fine-grained interactions among multi-modal data. By stacking MEA modules, the model captures enhanced representations of text and relationships. Extensive experiments on public datasets, including SciERC, ADE, and CoNLL04, demonstrate the superior performance of SMAN. The model delivers state-of-the-art performance, demonstrating significant improvements in F1 scores for entity detection across various datasets. On the ADE dataset, which comprises 4,272 samples extracted from medical reports and includes overlapping entities, SMAN achieves impressive F1 scores of 90.95%, highlighting its strong capability to manage complex overlapping scenarios effectively.

Our Methodology

This section begins by introducing the concepts and terminology used throughout this study, as well as the specific problem that our research addresses. The following provides a detailed explanation of the ENT-HAN and SNT-HAN models used for entity extraction and sentiment analysis.

A. Terminology and Problem Statement

This paper aims to solve two distinct tasks: entity extraction and sentiment analysis. A financial dataset consisting of 979 news articles is utilized, where all entities within each article are annotated along with their corresponding sentiments. The goal is to accurately identify the entities mentioned in the text and determine the sentiment (positive, negative, or neutral) associated with each entity.

The key to achieving accurate entity sentiment analysis lies in precisely determining the boundaries of the text segment where sentiment-related words are connected to and influence the target entity. In news articles, using the paragraph as the unit boundary for entity sentiment analysis proves to be practical, as sentiment words within a paragraph are generally associated, either explicitly or implicitly, with the target entity [14].

For both tasks, the appropriate paragraph p is first selected as the input unit. From the chosen paragraph, relevant sentences s are identified, followed by extracting the most relevant words w from those sentences. These selected units are then transformed into a three-dimensional vector representation, (p, s, w). Actually, for each entity i, the paragraph p_i is selected in which the entity appears and then choose the m surrounding sentences related to that entity, $p_i = (s_1, s_2, \dots, s_m)$, and each sentence comprises n words, $s_j = (w_1, w_2, \dots, w_n)$. In fact, word k from sentence j in paragraph i is positioned at the coordinates (i, j, k) within the three-dimensional input vector. This representation allows the model to effectively capture the hierarchical relationships between paragraphs, sentences, and words, ensuring that each word's context within a sentence and its broader role in the paragraph are preserved during both entity extraction and sentiment analysis.

Entity detection is a sequence labeling task. We formulate entity detection as a binary classification problem, the output Y_i is typically defined as a single value that represents one of the two possible classes. Given the word w, if w is an entity, it is labeled as 1; otherwise, it is labeled as 0.

$$Y_i=\begin{cases}0 & if\ w_i is\ not\ an\ entity\\ 1 & if\ w_i is\ an\ entity\end{cases} \quad (1)$$

To formulate the sentiment analysis problem, which consists of three classes, i.e. positive, negative, and neutral, the output is computed according to the following equation:

$$Y_i = \begin{cases} 0 & if\ sentiment\ is\ neutral\\ 1 & if\ sentiment\ is\ positive\\ 2 & if\ sentiment\ is\ negative\end{cases} \quad (2)$$

In this formulation, the model assigns a probability score to each sentiment class based on the features extracted from the input text. The class with the highest probability is selected as the predicted sentiment. This approach ensures that the classification process accounts for the nuanced differences between the three sentiments categories, leading to a more accurate representation of the emotional tone of the text.

B. Architecture Overview

The proposed ENT-HAN and SNT-HAN are built upon Hierarchical Attention Network (HAN), an advanced deep learning model designed to capture hierarchical structures in sequential data, making it particularly effective for tasks involving long and complex texts. These methods operate through two distinct layers of attention: word-level and sentence-level. First, an embedded representation of each word is generated, typically using methods like GloVe or BERT. The word-level attention mechanism then calculates a relevance score for each word in a sentence, using a combination of bidirectional GRUs and attention weights to emphasize words that are crucial for the task, such as identifying entities or sentiment indicators. These weighted word vectors are aggregated into a sentence vector, effectively summarizing the sentence.

At the next level, the sentence-level attention mechanism processes these sentence vectors, again applying a bidirectional GRU and attention mechanism to assign importance to specific sentences within a document. This enables the model to prioritize sentences that are more informative or contextually important for the task. The final output is a document-level representation that captures both fine-grained (word-level) and broader (sentence-level) context, making ENT-HAN and SNT-HAN well-suited for tasks multi-entity extraction and sentiment analysis in domains like financial text, where contextual nuances play a critical role.

The models' architecture consists of two layers. In each layer, inputs are transformed into one-dimensional vectors using a sequence encoder. An attention mechanism is then applied to these vectors, assigning higher weights to the most informative inputs. These weighted inputs are subsequently passed to the next layer. Ultimately, a Sigmoid function is utilized to produce the final output for entity extraction problem, indicating whether a word is likely to be an entity or not, and a Softmax function is used to produce the final output for the sentiment analysis problem, indicating whether an entity is positive, neutral or negative. The architecture of ENT-HAN model is depicted in the Fig. 1, and the subsequent sections will provide a more detailed explanation of this method. The architecture of the SNT-HAN method is similar to the ENT-HAN, with the primary difference being the function used in the output layer. While the ENT-HAN method addresses a binary classification problem, the SNT-HAN deals with a multi-class classification task. This distinction requires adjustments in the final layer, which are discussed in greater details in the following section.

ENT-HAN and SNT-HAN models comprise multiple sections: a word sequence encoder, a word-level attention layer, a sentence encoder, a sentence-level attention layer. The following sub-sections explain each of these parts.

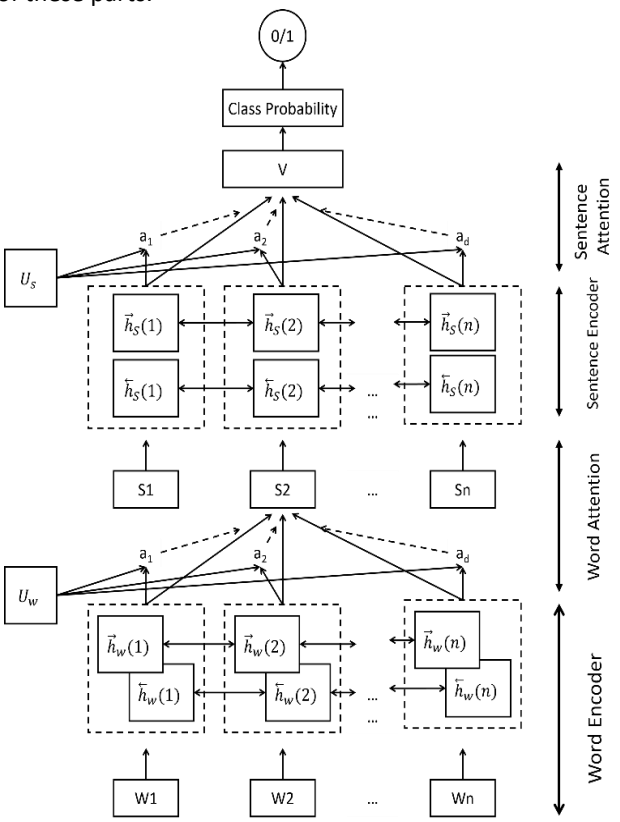


Fig. 1: ENT-HAN architecture.

The flowchart presented in Fig. 2 illustrates the step-by-step process of the ENT-HAN model for the entity detection. Each component in the flowchart represents a crucial stage in the model's operation. The process applied for sentiment analysis follows a similar approach, adhering to the same structure. This ensures consistency in how sentiments are extracted and classified across different entities and contexts within the text.

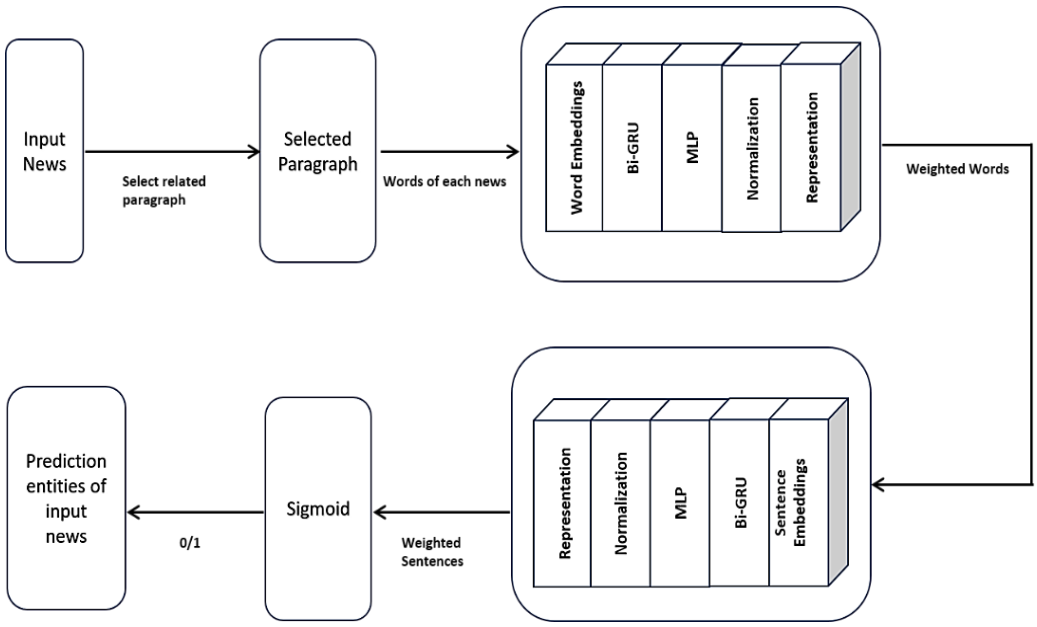


Fig. 2: Flowchart of ENT-HAN model.

Further, Algorithm 1 and Algorithm 2 outline the detailed steps of the ENT-HAN and SNT-HAN model, respectively, summarizing their core operations and logic.

Algorithm 1: Entity extraction process with ENT-HAN method

Input: a set of news paragraphs (P) belonging to all news documents, where each paragraph p_i contains sentences, and each sentence s_j contains words.

#Step 1: Preprocess inputs
$data = []$
$for\ each\ paragraph\ p_i\ in\ P$
  $for\ each\ sentence\ s_i\ in\ p_i$
    $for\ each\ word\ w_k\ in\ s_i$
      #Convert w_k into unique numerical representation
      $data[i, j, k] = tokenizer(w_k)$
#Step2: Map each word w_k to word embedding using a pre-trained embedding model
$embed_{sea} = GloVe(w_k)$
#Step 3: Word Encoding
$encoder_{wrd} = bidirectional\left(GRU\left(embed_{seq}\right)\right)$
$attention_{wrd} = attention\_layer(encoder_{wrd})$
#Step 4: Sentence Encoding
$encoder_{snt} = bidirectional\left(GRU(attention_{wrd})\right)$
$attention_{snt} = attention\_layer(encoder_{snt})$
#Step 5: Classification
$vector_{doc} = dence\_laye(attention_{snt})$
$entity = sigmoid(vector_{doc})$
$return\ \ entity\ \ (0\ or 1)\ \#return\ 0\ if\ the\ token\ is\ not\ an\ entity, otherwise\ return\ 1$

Output: predicted entities

Algorithm 2: Sentiment Analysis process with SNT-HAN method

Input: a set of paragraphs (P) with specified entities

#Step 1: Preprocess inputs
$data = []$
$for\ each\ entity$ in paragraph p_i
  $select\ sentences\ \{1..15\}\ and\ words\ \{1..120\}$
$for\ selected\ sentence\ s_j\ in\ paragraph\ p_i$
  $for\ selected\ word\ w_k\ in\ sentence\ s_j$
    #Convert $w_k$into unique numerical representation
    $data[i,j,k] = tokenizer(w_k)$
#Step 2: Map each word w_k to word embedding using a pre-trained embedding model
$embed_{seq} = Glove\ or\ BERT(w_k)$
#Step 3: Word Encoding
$encoder_{wrd} = bidirectional\left(GRU\left(embed_{seq}\right)\right)$
$attention_{wrd} = attention\_layer(encoder_{wrd})$
#Step 4: Sentence Encoding
$encoder_{snt} = bidirectional\left(GRU(attention_{wrd})\right)$
$attention_{snt} = attention\_layer(encoder_{snt})$
#Step 5: Classification
$vector_{doc} = dence\_laye(attention_{snt})$
$sentiment = softmax(vactor_{doc})$
$return\ 0\ or\ 1\ or\ 2\ \ \#Return\ 0\ for\ neutral, 1\ for\ positive, and\ 2\ for\ negative\ entities$

Output: sentiment (positive, negative, neutral) of each entity

1. Word Embedding

Word embedding is a technique in natural language processing where words or phrases are represented as vectors in a continuous vector space. This enables the model to position semantically similar or related words closer together, based on patterns learned from the training data [19]. GloVe and BERT represent two different approaches to word embeddings, each with distinct advantages. GloVe (Global Vectors for Word Representation) is a static embedding model, which means that each word is assigned a single vector based on co-occurrence statistics across a large corpus. This approach captures general word meanings well, but it cannot handle context-specific variations in meaning. On the other hand, BERT (Bidirectional Encoder Representations from Transformers) generates dynamic embeddings, where a word's representation depends on the surrounding context in the sentence. BERT, through its deep bidirectional architecture, captures nuanced, context-dependent meanings of words, making it more suitable for tasks like entity extraction and sentiment analysis in complex language structures. However, BERT is computationally more intensive compared to GloVe, which is faster but less capable of understanding context.

Hence, pre-trained embedded vectors are used, using GloVe in one instance and BERT in another, to provide the model with an additional advantage in terms of performance [20].

2. Word Sequence Encoder

Assume that a selected paragraph contains m sentences, and n indicates the number of words in each sentence. The word in the i^{th} sentence is denoted as w_{ij} where j$\epsilon[1,n]$. Initially, the words are transformed into vectors through an embedding matrix W_e, resulting in $x_{ij} = W_e w_{ij}$. Next, word vectors are obtained from both directions using a bidirectional GRU, which processes input sequences in both backward and forward directions. In the forward direction, the input sequence is processed from the first word to the last word, with each word's representation being influenced by the preceding words in the sequence. Simultaneously, in the backward direction, the input sequence is processed from the last word to the first word, where each word's representation is influenced by the subsequent words in the sequence. This bidirectional approach captures contextual information from both preceding and subsequent words for each word in the sequence.

The Gated Recurrent Unit (GRU) is a gating mechanism introduced in 2014 by Cho et al. [21], and further developed by Chung et al. [22]. It offers the advantage of faster computation compared to many other recurrent neural network models, which are especially effective for handling sequential data. The GRU uses two gates—the reset gate and the update gate—that control the flow of information within each unit. The update gate, denoted as z_t^j, intuitively allows the model to regulate how much of the past information from the previous state should be retained and transmitted to the new state. This gate is computed by taking a linear combination of the previous hidden state and the current input, which is then processed through a Sigmoid function:

$$z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j \tag{3}$$

where x_t is the sequence vector at time t. The reset gate, denoted as r_t^j, determines how much of the previous hidden state should be forgotten or reset. This is achieved by taking a linear combination of the previous hidden state h_{t-1} and the current input x_t , and then passing the result through an activation function. The reset gate is computed as:

$$r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j \tag{4}$$

here, W_r represents the weight matrix, and σ is the Sigmoid activation function that ensures the gate's output remains between 0 and 1, effectively controlling the degree of resetting the previous hidden state.

The new state h_t^j at time t is a linear interpolation between the current new state $\hat{\mathrm{h}}_t^j$ and the previous state h_{t-1}^j:

$$h_t^j = (1 - z_t^j) h_{t-1}^j + z_t^j \hat{\mathrm{h}}_t^j \tag{5}$$

The candidate state $\hat{\mathrm{h}}_t^j$ is computed as follows:

$$\hat{\mathrm{h}}_t^j = \tanh(W x_t + U(r_t \odot h_{t-1}))^j \tag{6}$$

where $\odot$ is an element-wise multiplication. If r_t is zero, the model disregards the previous state.

The bidirectional nature of the above approach allows the model to gain a deeper understanding of the context in which a word appears, thereby enhancing its ability to grasp semantic relationships and meanings. This dual perspective provides a richer representation of the word, improving the model's capacity to acquire the full context of the text. Thus, we have:

$$\begin{aligned} x_{ij} &= W_e w_{ij}, j\epsilon[1,n], i\epsilon[1,m] \\ \vec{h}_{ij} &= \overrightarrow{GRU}(x_{ij}), j\epsilon[1,n], i\epsilon[1,m] \\ \overleftarrow{h}_{ij} &= \overleftarrow{GRU}(x_{ij}), j\epsilon[n,1], i\epsilon[m,1] \end{aligned} \tag{7}$$

here, x_{ij} represents the transformation of the input w_{ij} using a weight matrix W_e and $\vec{h}_{ij}$ and $\overleftarrow{h}_{ij}$ indicate the forward and backward GRU operation on x_{ij}, respectively.

3. Word Attention

In the Hierarchical Attention Network model, certain words are more critical to a sentence's meaning. To effectively combine the representations of these informative words into a sentence vector, the ENT-HAN and SNT-HAN models employ an attention mechanism. Here's a detailed breakdown of the process:

a. Bidirectional GRU (Bi-GRU):

The sentence is initially expressed as a sequence of word vectors. This sequence is processed through a bidirectional GRU, which handles the words in both their original order and reverse order. This bidirectional processing captures contextual information from both future and past words, resulting in a sequence of hidden states where each hidden state encapsulates the contextual information of a word within the sentence.

b. One-Layer MLP:

The hidden states from the Bi-GRU are fed into a one-layer Multilayer Perceptron (MLP). This layer performs a linear transformation followed by a non-linear activation function to each hidden state to calculate the importance scores for each word. Specifically, the importance score u_{ij} is computed as:

$$u_{ij} = \tanh(W_w h_{ij} + b_w) \tag{8}$$

where W_w is the weight matrix and b_w is the bias term.

c. Importance Measurement:

The output of the MLP represents the importance scores for each word in the sentence. These scores indicate the relative importance of each word within the context of the entire sentence.

d. Normalization (Softmax):

To convert the importance scores into a normalized probability distribution, the scores are passed through a Softmax function. This function ensures that the importance weights sum to 1, with higher scores translating to higher weights:

$$\alpha_{ij} = \frac{\exp(u_{ij}^T u_w)}{\sum_p \exp(u_{ip}^T u_w)} \tag{9}$$

here, u_w is a vector representing the weights, and the denominator normalizes the weights across all words in the sentence.

e. Sentence Representation:

The normalized importance weights α_{ij} are used to compute the sentence representation by taking a weighted sum of the word vectors. The sentence vector s_i is given by:

$$s_i = \sum_p \alpha_{ip} h_{ip} \tag{10}$$

where h_{ip} represents the word vectors, and α_{ip} are the normalized importance weights. This aggregation

focuses on the most relevant words, resulting in a sentence vector that effectively captures the key information.

This mechanism enables the model to concentrate on the most significant words when generating a representation for the entire sentence.

4. Sentence Sequence Encoder

The same procedure used for encoding words is applied to the derived sentence vectors to create the document vector. A bidirectional GRU is employed to encode the sentences:

$$\begin{aligned}\vec{h}_i &= \overrightarrow{GRU}(s_i), i\epsilon[1,m]\\ \overleftarrow{h}_i &= \overleftarrow{GRU}(s_i), i\epsilon[m,1]\end{aligned} \quad (11)$$

For getting an annotation of sentence i, $\vec{h}_i$ and $\overleftarrow{h}_i$ must be concatenated, i.e., $h_i = [\vec{h}_i, \overleftarrow{h}_i]$. h_i encapsulates the neighboring sentences surrounding sentence i while still concentrating on this sentence.

In this part of ENT-HAN and SNT-HAN model, GRU-based sequence encoder is the same as the one applied in the word encoder.

5. Sentence Attention

Each sentence in a news article conveys a distinct semantic meaning, hence it is essential to calculate attention weights for different sentences individually to emphasize those that are more critical for event detection. To compute the document vector v_k, which summarizes all the information from the sentences in a paragraph of a news article, the following formulas are used:

$$\begin{aligned}u_i &= \tanh(W_s h_i + b_s)\\ \alpha_i &= \frac{\exp(u_i^T u_s)}{\sum_p \exp(u_p^T u_s)}\\ v &= \sum_p \alpha_p h_p\end{aligned} \quad (12)$$

A transformation is applied to the hidden state h_i using learnable parameters W_s and b_s, producing an intermediate representation u_i. Then, attention weights α_i are computed by measuring the similarity between u_i and a context vector u_s, normalized over all inputs. By following this process, the document vector v effectively captures the key information from the paragraph, highlighting the sentences that contribute most to the event detection and sentiment analysis tasks.

6: Prediction in Entity Extraction

To predict the probability of binary classification (with only two classes, entity and non-entity), the Sigmoid function is applied. The Sigmoid function transforms the model's output into a probability value between 0 and 1.

$$\rho = Sigmoid(W_p d + b_p) \quad (13)$$

Specifically, Sigmoid function takes any real-valued input and transforms it into a value in range 0 and 1, allowing the model to interpret the output as the probability of a given class. If the output of the Sigmoid function is closer to 1, the model predicts the positive class (e.g., entity present); if it's closer to 0, it predicts the negative class (e.g., entity absent). This makes the Sigmoid function ideal for binary classification problems, as it converts raw predictions into easily interpretable probabilities. To optimize the model during training, the cross-entropy loss function is employed. Cross-entropy quantifies the disparity between the predicted probabilities $q(x)$ and the actual labels $p(x)$. It assesses how effectively the predicted probability distribution corresponds with the actual distribution of the labels. The cross-entropy loss function CE is defined as:

$$CE = -\sum_x p(x) \log_2 q(x) \quad (14)$$

7. Prediction in Sentiment Analysis

For sentiment analysis, however, the Softmax function is employed, as it is designed for multi-class classification with three sentiment categories: positive, negative, and neutral. The Softmax function assigns a probability to each class, ensuring that the sum of probabilities across the three sentiment classes equals 1, facilitating a more accurate sentiment prediction.

$$\rho = Softmax(W_p d + b_p) \quad (15)$$

To optimize the SNT-HAN model during training, similar to ENT-HAN, the cross-entropy loss function is employed.

Experiments

In this study, the FinEntity dataset is used, a comprehensive collection of financial texts annotated with sentiment labels. The dataset consists of 979 example paragraphs, featuring a total of 2,131 entities classified into three sentiment categories: Positive, Negative, and Neutral. Notably, approximately 60% of the paragraphs in the dataset contain multiple entities, making it particularly challenging and relevant for tasks involving complex entity extraction in financial contexts. It is possible that in a sentence containing multiple entities, the sentiment associated with each entity may differ, presenting a complex challenge that required careful attention to address. This variability in sentiment adds an additional layer of difficulty to the analysis, as it requires distinguishing between the emotional tones linked to each individual entity within the same textual context.

To extract entities from this dataset and to identify the sentiments associated with each of the entities, we employed ENT-HAN and SNT-HAN methods, respectively and we found them highly effective for performing entity extraction and sentiment analysis tasks. The ENT-HAN method demonstrated strong performance in accurately identifying and classifying entities within the financial texts of the FinEntity dataset. Its ability to handle the

intricacies of financial language, including the detection of multiple entities within the same paragraph, proved to be particularly beneficial. Additionally, SNT-HAN method shows to be highly effective in analyzing the sentiments associated with these entities, delivering strong performance in accurately capturing and classifying their emotional context.

A. Dataset

This description provides an overview of the FinEntity dataset, which is a collection of paragraphs focused on financial text. Here's a summary of the key details:

- Total Examples: 979 paragraphs.
- Total Entities: 2,131 entities classified into three sentiment categories:
  - Positive Entities: 503 entities (approximately 24% of the total).
  - Negative Entities: 498 entities (approximately 23% of the total).
  - Neutral Entities: 1,130 entities (approximately 53% of the total).
- Sentiment Label Distribution: The distribution across Positive, Negative, and Neutral entities is fairly balanced, though Neutral entities form the majority.
- Entity Presence in Text: About 60% of the financial text contains multiple entities, indicating that the dataset often features paragraphs with more than one entity labeled with a sentiment.

B. Settings

Testing different values for model parameters and selecting the optimal ones is a crucial aspect of model development and fine-tuning. In this study, we systematically explored various parameter settings to identify the configurations that yield the best performance for our model.

This process involved adjusting and evaluating multiple hyper-parameters to optimize the model's accuracy and effectiveness. The selected hyper-parameters that were found to produce the most favorable results are detailed in Table 1.

Table 1: The setting of hyper-parameters

Hyper-parameter	ENT-HAN	SNT-HAN
Batch Size	56	56
Max Sentence Length	100	120
Max Sentence Number in a Paragraph	7	15
Embedding Dimensions	100	100
Validation Split	20%	20%
Epochs	10	10
GRU Dimensions	50	50
Word Dimensions	100	100
Sentence Dimensions	100	100

The evaluation metric for assessing ENT-HAN model performance is the F1 score, which provides a balanced measure of recall and precision, ensuring that our model performs well across both aspects. The evaluation metric for assessing SNT-HAN model is accuracy.

C. Experimental Results of Entity Detection

For comparison purposes, a comparative analysis is conducted between ENT-HAN method, the BERT model—a widely recognized benchmark in natural language processing tasks—and the Recurrent Neural Network (RNN) model. By evaluating our approach against these well-established models, we aim to assess the relative strengths and weaknesses of ENT-HAN method in terms of entity extraction within financial texts. This comparison provides an understanding of how our method performs in relation to both the cutting-edge BERT model and the traditional RNN approach.

BERT (Bidirectional Encoder Representations from Transformers) [23] is a transformative method in natural language processing that has significantly advanced the state of the art in various tasks, including entity extraction. Unlike traditional models that process text sequentially, BERT employs a bidirectional approach, allowing it to capture the context of a word based on both its preceding and following words. This deep contextual understanding enables BERT to more accurately identify and classify entities within a text. Its pre-training on vast amounts of text data and subsequent fine-tuning for specific tasks have made BERT particularly effective in extracting nuanced and context-sensitive entities, which is crucial for applications in complex domains like finance literature.

Recurrent Neural Networks (RNNs) [24] are a type of neural network specifically designed to handle sequential data by retaining a memory of prior inputs. This capability enables RNNs to capture temporal dependencies, making them well-suited for tasks such as language modeling, time series forecasting, and sequence classification.

The Decoding Enhanced BERT with Disentangled Attention (DeBERTa) [25], introduced by Microsoft, represents an advanced variation of the BERT architecture that has achieved significant benchmarks across numerous natural language processing tasks. This model excels particularly in tasks such as entity extraction and sentiment classification, thanks to its ability to capture intricate linguistic structures and contextual relationships within text. Unlike its predecessors, DeBERTa employs a disentangled attention mechanism, which uses two separate vectors for each token—one to represent its semantic content and the other to encode its positional information. This disentangled representation allows the model to better understand the nuanced interactions between words in various contexts.

Furthermore, the model enhances its pre-training process through an improved mask decoder, optimizing its performance on masked language modeling tasks by providing more accurate predictions for masked tokens. Collectively, these innovations enable DeBERTa to deliver state-of-the-art performance in understanding and processing complex textual data. The primary metrics utilized for comparison in our study on the entity extraction problem include F1 score, precision, recall, and accuracy [26]. These metrics are widely recognized in classification tasks, with the F1 score providing a balanced measure of recall and precision, precision evaluating the proportion of correctly identified positive instances, recall measuring the proportion of actual positive instances correctly identified, and accuracy reflecting the overall correctness of predictions. Precision refers to the ratio of true positive predictions among the total number of positive predictions made by the model. It answers the question: "Out of all the entities the model identified as relevant, how many were actually relevant?"

On the other hand, recall is the ratio of true positive predictions to the total number of actual positive instances in the dataset. It addresses the question: "Out of all the relevant entities in the dataset, how many did the model successfully identify?". The F1 score is then defined as:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + recall} \quad (16)$$

This metric is particularly valuable in situations where there is a discrepancy in the number of relevant and irrelevant cases, as it provides a more thorough assessment of model performance by taking into account both false positives and false negatives. Focusing on the F1 score ensures that our comparison captures the overall effectiveness of the models in accurately identifying and classifying entities. Accuracy measures the proportion of correctly classified instances (both positive and negative) to the total number of instances. It reflects the overall effectiveness of a model in making correct predictions. The formula is:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (17)$$

The comparative analysis evaluated the performance of ENT-HAN method against the RNN, BERT and DeBERTa model, specifically, the study focuses on the F1 score (f1) as the key metric while also incorporating precision (Prec.), recall (Rec.), and accuracy to provide a comprehensive evaluation of model performance. The results in Table 2 show that ENT-HAN model surpasses RNN, BERT, and DeBERTa by achieving higher scores across all key metrics. This demonstrates the model's ability to balance precision and recall effectively in entity extraction tasks, accurately identifying relevant entities while minimizing false positives and false negatives. The ENT-HAN model's superior F1 score highlights its enhanced capability in capturing the context-sensitive nature of financial entities compared to BERT and DeBERTa. Leveraging its advanced attention mechanism, the model prioritizes the most critical sentences and words within the content, significantly improving entity recognition performance.

Table 2: Comparison of our model with baseline model

Method	Prec.	Rec.	F1	Accuracy
RNN	73.9	80.6	77.1	60.24
BERT	85.9	92.8	89.2	75.44
DEBERTA	90.8	96.5	93.6	93.35
ENT-HAN	89.2	96.7	92.8	93.90

The performance of DeBERTa and ENT-HAN model, which exhibit similar behavior in entity extraction, was compared in terms of computational time complexity and memory usage during execution. The results of this comparison, calculated for each epoch, are presented in the Table 3. The findings indicate that while DeBERTa demonstrates high predictive accuracy, it comes with significantly higher time and space complexity compared to ENT-HAN. This trade-off highlights the potential limitations of DeBERTa for scenarios requiring efficient computation and resource utilization.

Table 3: Comparison of DeBERTa and ENT-HAN complexity

Method	Time Complexity (Sec)	Space Complexity (MB)
DeBERTa	2925.9576	9050.84
ENT-HAN	1304.4537	4548.47

This part aims to evaluate our approach against those introduced in the literature, despite the fact that none of these approaches used the same dataset, and the datasets and domains they addressed differ significantly. While this makes a precise comparison challenging, it may still provide a general perspective on the effectiveness of our method.

Table 4: Comparison of different methods evaluated on different datasets

Method	Dataset	F1 Score
CNN+Linguistic Patterns [15]	LAPTOP REVIEWS	82.32%
	Restaurant Reviews	87.17%
Ensemble-Focal-Loss [16]	2019 CCF BDCI	95.25%
	2019 CCKS	85.05%
TTEE [17]	RES15	58.94%
	Res16	68.39%
	ACOS_LAPTOP	43.94%
SMAN [18]	ADE	90.95%
ENT-HAN	FinEntity	92.80%

Importantly, as shown in Table 4, our approach demonstrates a high level of accuracy in correctly identifying multiple entities within a single sentence—a capability that is not highlighted in any of the other methods. This advantages of accurately extracting multiple entities further underscores the robustness of our method, even if direct comparisons are limited by differences in datasets and application domains.

D. Experimental Results of Sentiment Analysis

In the case of the sentiment analysis problem, we compare our findings with those presented in the original paper [5] that introduced the dataset. In that study, six different methods were applied to the dataset, which is briefly described below. The methods include BERT, BERT-CRF, FinBERT, FinBERT-CRF, ChatGPT (zero-shot), and ChatGPT (few-shot). The BERT method, as previously explained, serves as the baseline for comparison. FinBERT is a specialized variant of the BERT model, designed by Yang et al. in 2020 [27] to address the unique challenges of natural language processing within the financial domain. Unlike standard BERT, which is trained on general language corpora, FinBERT is pre-trained on large datasets of financial texts, such as earnings reports, news articles, and financial statements, enabling it to better understand domain-specific language, terminology, and context. For each token's hidden output, FinBERT applies a linear layer to perform tasks like sentiment analysis. This model is particularly effective in tasks that require a deep understanding of financial jargon and context, and when combined with Conditional Random Field (CRF) [28] layers (as in FinBERT-CRF), it further improves performance in sequence-based tasks such as sentiment classification. Few-shot [29] and zero-shot learning [30] of ChatGPT are used to perform tasks with minimal or no task-specific training data. In zero-shot learning, the model is given a task without any prior examples or training for that specific task. It relies entirely on its general knowledge and pre-training to generate a response. For instance, in sentiment analysis, a zero-shot approach would attempt to identify the sentiment of entities based solely on its understanding of language, without seeing any labeled examples beforehand. In few-shot learning, the model is provided with a small number of examples (few-shot examples) to learn the task before making predictions. This technique allows the model to better understand the structure or rules of the task with minimal data, improving its performance compared to zero-shot. In both approaches, ChatGPT leverages its extensive pre-training, but few-shot learning typically yields more accurate and reliable results, especially in complex tasks like sentiment analysis.

This study employs micro-avg, macro-avg, and weighted-avg methods for prediction, three commonly used averaging methods for precision, recall, and F1-score provide different perspectives on performance. Micro-Averaging method aggregates true positives, false positives, and false negatives across all classes and computes the metrics globally. It is effective for datasets with class imbalance, as it gives equal weight to each instance. Macro-Averaging approach calculates the metric for each class independently and then averages them. It treats all classes equally, regardless of their size, which can highlight performance disparities between minority and majority classes. Weighted Averaging technique calculates metrics for each class and averages them, weighted by the number of instances in each class. It balances the influence of each class based on its prevalence, making it useful for understanding the model's overall performance in imbalanced datasets.

Among the models applied in the aforementioned paper, the FinBERT-CRF model achieved the best performance, with a macro average of 85%. The ChatGPT method does not perform particularly well in sentiment analysis tasks, it has the lowest level of accuracy. Despite its strong language modeling capabilities, its performance in accurately classifying sentiments, especially in domain-specific texts like financial documents, tends to lag behind more specialized models. This is likely due to the lack of fine-tuning for the specific nuances and subtleties present in sentiment expressions, which can lead to lower accuracy in identifying the correct sentiment class. As a result, while ChatGPT can provide reasonable outputs, it often struggles to match the precision and reliability of models specifically trained for sentiment analysis, such as FinBERT or other fine-tuned approaches. However, our proposed SNT-HAN method demonstrates superior accuracy in sentiment analysis, outperforming the existing models. The detailed results of the comparison are provided in the Table 5.

Table 5: Comparison of baseline method and our method

Method	MicroAvg	Macro Avg	Weighted Avg
BERT	80	80	80
BERT-CRF	81	81	81
ChatGPT (zero-shot)	59	56	59
ChatGPT (few-shot)	67	65	67
FinBERT	83	83	83
FinBERT-CRF	84	85	84
SNT-HAN with GloVe Embedding	87	86	84
SNT-HAN with BERT Embedding	89	87	85

Our results highlight the effectiveness of the SNT-HAN, which enables more precise sentiment analysis. By

capturing both word and sentence-level context, the SNT-HAN model has proven to be more adept at handling complex financial text, particularly when multiple entities are present within a single sentence. Overall, these findings suggest that SNT-HAN is a highly effective approach for multi-entity sentiment analysis, outperforming traditional methods in this domain. It is also evident that using BERT for word embedding, rather than GloVe, leads to an improvement in accuracy, highlighting the impact of context-aware embeddings on model performance. The results confirm that integrating attention mechanisms significantly enhances sentiment detection accuracy in financial texts.

Conclusion

In conclusion, entity extraction is a critical component of natural language processing, particularly in domains like finance where precise identification and classification of entities are essential for accurate sentiment analysis and decision-making. ENT-HAN method has proven to be highly effective in this regard. By leveraging its advanced capabilities, ENT-HAN successfully captures the complex, context-dependent relationships between entities in text, leading to superior performance as evidenced by its higher F1 score compared to other models like RNN and BERT. The results highlight the robustness of deep learning methods in addressing the challenges of entity extraction, making ENT-HAN method a valuable tool for tasks that demand high accuracy and reliability.

The SNT-HAN proves to be a highly effective approach for sentiment analysis, particularly in scenarios involving complex texts with multiple entities. One of the key strengths of SNT-HAN lies in its ability to model attention at both word and sentence levels, allowing the model to focus on the most relevant sections of text when determining sentiment. This hierarchical structure enhances the model's capacity to capture context-dependent sentiments, leading to more accurate sentiment classification, even when multiple entities are present within a single document. Furthermore, SNT-HAN's ability to account for the varying importance of words and sentences makes it well-suited for tasks requiring nuanced understanding, such as financial sentiment analysis. Overall, SNT-HAN's superior performance compared to traditional models highlights its robustness and adaptability in handling intricate sentiment analysis tasks.

Author Contributions

L. Hafezi and S. Zarifzadeh conceptualized and designed the study and conducted data analysis. L. Hafezi contributed to data collection, conducted statistical analyses, and wrote the initial draft of the manuscript. S. Zarifzadeh and M. Pajoohan supervised the project and provided critical feedback during manuscript development. All authors reviewed and approved the final manuscript.

Acknowledgment

We sincerely appreciate the insightful comments and constructive suggestions provided by the anonymous reviewers, which have greatly contributed to enhancing the quality and clarity of this manuscript. We are also grateful to the editor for their careful oversight throughout the review process. Furthermore, we acknowledge the editorial team of *JECEI* for their dedication and professionalism in managing the submission and publication process.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

The following abbreviations are used throughout this document for clarity and conciseness.

HAN	Hierarchical Attention Network
LSTM	Long Short-Term Memory
NSSM	Negative Sentiment Smoothing Model
TASD	Target and Aspect Sentiment Detection
TTEE	Twin Towers End-to-End
MEA	Model Enhanced Attention
GloVe	Global Vectors for Word Representation
BERT	Bidirectional Encoder Representation from Transformers
GRU	Gated Recurrent Unit
Bi-GRU	Bidirectional GRU
M LP	Multi-Layer Perceptron
RNN	Recurrent Neural Network
DeBERTa	Decoding Enhanced BERT with disentangled Attention
CRF	Conditional Random Field

References

[1] K. Du, F. Xing, e "Financial sentiment analysis: Techniques and applications," ACM Comput. Surv., 56(9): 1-42, 2024.

[2] J. Yu, J. Jiang, R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," IEEE/ACM Trans. Audio Speech Lan. Process., 28: 429-439, 2020.

[3] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, "Hierarchical attention networks for document classification," in Proc. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 1480-1489, 2016.

[4] L. Deng et al., "Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models," in Proc. 2015 Conf. Empirical Methods in Natural Language Processing: 179-189, 2015.

[5] Y. Tang et al., "FinEntity: Entity-level sentiment classification for financial texts," in Proc. 2023 Conference on Empirical Methods in Natural Language Processing: 15465-15471, 2023.

[6] C. Liu et al., "Large language models and sentiment analysis in financial markets: A review, datasets, and case study," IEEE Access, 12: 134041-134061, 2024.

[7] M. Kraus, S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," Decis. Support Syst., 104: 38-48, 2017.

[8] S. Sohangir et al., "Big data: Deep learning for financial sentiment analysis," J. Big Data, 5(1), 2018.

[9] B. Lutz et al., "Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning," arXiv preprint arXiv:1901.00400, 2018.

[10] M. Maia, et al., "FinSSLx: A sentiment analysis model for the financial domain using text simplification," in Proc. 12th International Conf. Semantic Computing (ICSC): 318-319, 2018.

[11] Z. Liu et al., "FinBERT: A pre-trained financial language representation model for financial text mining," in Proc. International Joint Conf. Artificial Intelligence: 4513-4519, 2021.

[12] G. Fatouros et al., "Transforming sentiment analysis in the financial domain with ChatGPT", Mach. Learn. Appl., 14, 100508, 2023.

[13] A. M. Ardekani et al. "FinSentGPT: A universal financial sentiment engine," int. Rew. Financ. Anal., 94, 103291, 2024.

[14] M. Luo, X. Mu, "Entity sentiment analysis in the news: A case study based on Negative Sentiment Smoothing Model (NSSM)," Int. J. Inf. Manage. Data Insights, 2(1), 2022.

[15] S. Poria, E. Cambria, A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," Knowledge-Based Syst., 108: 42-49, 2016.

[16] L. Zhao, L. Li, X. Zheng, J. Zhang, "A BERT based sentiment analysis and key entity detection approach for online financial texts," in Proc. CSCWD 2021: 1233-1238, 2021.

[17] Z. Li, Y. Song, X. Lu, M. Liu, "Twin towers end to end model for aspect-based sentiment analysis," Expert Syst. Appl., 249, 2024.

[18] Q. Wan et al., "A span-based multi-modal attention network for joint entity-relation extraction," Knowledge-Based Syst., 262, 2023.

[19] A. Thomas et al., "Investigating the impact of pre-trained word embeddings on memorization in neural networks," in Proc. Text, Speech, and Dialogue, Lecture Notes in Computer Science, 2020.

[20] C. Wang et al., "A comparative study on word embeddings in deep learning for text classification," in Proc. 4th Int. Conf. Natural Language Processing and Information Retrieva: 37-46, 2021.

[21] K. Cho et al.,"On the properties of neural machine translation: encoder-decoder approaches," in Proc. English Workshop on Syntax, Semantics and Structures in Statistical Translation, 2014.

[22] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Workshop on Deep Learning, 2014.

[23] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1: 4171-4186, 2019.

[24] D. E. Rumelhart, J. L. McClelland, "Learning internal representations by error propagation," in Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, MIT Press, 318-362, 1987.

[25] P. He, X. Liu, J. Gao, W. Chen, "DeBERTa: Decoding enhanced BERT with Disentangled Attention," in Proc. ICLR, 2021.

[26] H. Dalianis, "Evaluation metrics and evaluation," Clinical Text Mining, Springer, Cham, 2018.

[27] Y. Yang, M. C. Uy, A. Huang, "FinBert: A pretrained language model for financial communications," arXiv:2006.08097, 2020.

[28] C. Sutton, A. McCallum, "An introduction to conditional random fields," Found. Trends Mach. Learn., 4(4): 267-373, 2011.

[29] U. Lee et al., "Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education," Educ. Inf. Technol., 29: 11483-11515, 2024.

[30] F. H. Shubho et al., "ChatGPT-guided Semantics for Zero-shot Learning," in Proc. 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA): 418-425, 2023.

Biographies

Leila Hafezi is a Ph.D. candidate in Software Engineering at Yazd University, Yazd, Iran. She received her B.Sc. degree in Computer Software Engineering from Yazd University and her M.Sc. degree in Software Engineering from Isfahan University of Technology, Isfahan, Iran. Her research interests include Natural Language Processing, Deep Learning, and Data Science.

- Email: leilahafezi@stu.yazd.ac.ir
- ORCID:0009-0009-3794-9985
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Sajjad Zarifzadeh is an Associated Professor in the Department of Software Engineering at Yazd University, Yazd, Iran. He earned his Ph.D. from University of Tehran. His research interests span Big Data Engineering, Web Mining, Internet Services and Data Analysis.

- Email: szarifzadeh@yazd.ac.ir
- ORCID: 0000-0001-5627-0542
- Web of Science researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://pws.yazd.ac.ir/zarifzadeh/en

Mohammad Reza Pajoohan is an Associated Professor in the Department of Software Engineering at Yazd University, Yazd, Iran. He earned his Ph.D. from Universiti Sains Malaysia. His research interests include Data Mining and Knowledge Discovery, Medical Data Processing, Databases and Big Data, Privacy Preservation Data Publication and Business Process Reengineering.

- Email: pajoohan@yazd.ac.ir
- ORCID: 0000-0002-7848-2025
- Web of Science researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://pws.yazd.ac.ir/pajoohan/

How to cite this paper:
L. Hafezi, S. Zarifzadeh, M. R. Pajoohan, "Enhancing multi-entity detection and sentiment analysis in financial texts with hierarchical attention networks," J. Electr. Comput. Eng. Innovations, 13(2): 403-416, 2025.

DOI: 10.22061/jecei.2025.11294.790

URL: https://jecei.sru.ac.ir/article_2302.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Hybrid Fine-Tuning of Large Language Models Using LoRA: Enhancing Multi-Task Text Classification through Knowledge Sharing

A. Beiranvand, M. Sarhadi, J. Salimi Sartakhti *

Department of Computer Engineering, University of Kashan, Kashan, Iran.

Article Info

Article History:
Received 02 November 2024
Reviewed 13 January 2025
Revised 12 February 2025
Accepted 02 March 2025

Keywords:
Large language model
Fine-Tuning
PEFT
LoRA
Knowledge sharing
Attention mechanism

*Corresponding Author's Email Address: Salimi@kashanu.ac.ir

Abstract

Background and Objectives: Large Language Models have demonstrated exceptional performance across various NLP tasks, especially when fine-tuned for specific applications. Full fine-tuning of large language models requires extensive computational resources, which are often unavailable in real-world settings. While Low-Rank Adaptation (LoRA) has emerged as a promising solution to mitigate these challenges, its potential remains largely untapped in multi-task scenarios. This study addresses this gap by introducing a novel hybrid approach that combines LoRA with an attention-based mechanism, enabling fine-tuning across tasks while facilitating knowledge sharing to improve generalization and efficiency. This study aims to address this gap by introducing a novel hybrid fine-tuning approach using LoRA for multi-task text classification, with a focus on inter-task knowledge sharing to enhance overall model performance.

Methods: We proposed a hybrid fine-tuning method that utilizes LoRA to fine-tune LLMs across multiple tasks simultaneously. By employing an attention mechanism, this approach integrates outputs from various task-specific models, facilitating cross-task knowledge sharing. The attention layer dynamically prioritizes relevant information from different tasks, enabling the model to benefit from complementary insights.

Results: The hybrid fine-tuning approach demonstrated significant improvements in **accuracy** across multiple text classification tasks. On different NLP tasks, the model showed superior generalization and precision compared to conventional single-task LoRA fine-tuning. Additionally, the model exhibited better scalability and computational efficiency, as it required fewer resources to achieve comparable or better performance. Cross-task knowledge sharing through the attention mechanism was found to be a critical factor in achieving these performance gains.

Conclusion: The proposed hybrid fine-tuning method enhances the accuracy and efficiency of LLMs in multi-task settings by enabling effective knowledge sharing between tasks. This approach offers a scalable and resource-efficient solution for real-world applications requiring multi-task learning, paving the way for more robust and generalized NLP models.

Introduction

Large language models (LLMs) have become essential in artificial intelligence, especially for natural language processing (NLP) and various other applications. These models, characterized by their sophisticated architectures and deep neural networks, have fundamentally transformed NLP by demonstrating unparalleled capabilities in both generating and comprehending human language. The impact of LLMs extends beyond NLP [1], influencing fields such as machine translation, sentiment analysis, and even creative writing. Despite their transformative potential,

fully fine-tuning these models presents significant challenges. The primary obstacle lies in the sheer number of parameters, often reaching billions, which necessitates substantial computational resources and advanced hardware. This complexity not only increases the cost and time required for fine-tuning but also raises concerns about energy consumption and environmental impact. Consequently, researchers are exploring alternative approaches such as transfer learning, parameter-efficient tuning, and the development of more efficient model architectures to mitigate these challenges.

To utilize an LLM for various tasks, a common approach is to fine-tune a pre-trained model on the specific task data [2], [3]. Full fine-tuning of a language model can be computationally intensive, typically requiring the update of all parameters in the pre-trained model, and the fine-tuned model may end up with as many parameters as the original model [4]. To overcome this issue, parameter-efficient fine-tuning methods like Low-Rank Adaptation (LoRA) [5] enable fine-tuning a pre-trained model by introducing small LoRA modules for different tasks. In these methods, the main parameters of the pre-trained model remain fixed, and only the weights of the two low-rank matrices in LoRA are updated, which are significantly fewer in number compared to the main parameters of the pre-trained model.

LoRA significantly reduces the computational resources required and enables the fine-tuning process across various tasks. For example, thousands of LLaMA models [6], fine-tuned using LoRA, are available on Hugging Face Hub [7]. These practical applications demonstrate that LoRA is not only widely used for fine-tuning tasks in LLMs but also achieves model accuracy comparable to other full-weight fine-tuning methods. The lightweight nature of LoRA-based fine-tuning allows for training multiple LoRA modules on a single GPU. LoRA-based fine-tuning systems, such as Alpaca-LoRA [8], primarily focus on optimizing single-task fine-tuning and have not fully explored efficient strategies for multi-task fine-tuning.

Despite the successes achieved, the majority of existing research and systems have focused predominantly on single-task fine-tuning, with limited exploration of efficient strategies for multi-task fine-tuning. In this paper, we introduce a hybrid model that fine-tunes large language models using the LoRA method, enhancing model accuracy by enabling simultaneous learning across multiple tasks. This hybrid approach employs an attention mechanism to integrate the outputs of various tasks, yielding superior performance in diverse text classification tasks.

Existing parameter-efficient fine-tuning techniques, such as Low-Rank Adaptation (LoRA), have shown promise in reducing computational requirements. However, their applications have been largely limited to single-task learning, leaving multi-task scenarios underexplored. Multi-task learning, with its potential for inter-task knowledge sharing, offers significant advantages in terms of generalization and resource efficiency, yet it poses unique challenges in balancing task-specific requirements. To address these challenges, we propose a hybrid fine-tuning approach that enhances multi-task text classification by leveraging LoRA alongside an attention mechanism for effective knowledge sharing.

The main contributions of this paper are as follows:

- **Hybrid Fine-Tuning Approach**: This paper introduces a hybrid fine-tuning approach that fine-tunes large language models (LLMs) using Low-Rank Adaptation (LoRA). Unlike traditional fine-tuning methods that focus on single tasks, this hybrid approach enables simultaneous fine-tuning across multiple tasks. The central innovation lies in leveraging knowledge sharing between tasks, allowing the model to learn from multiple tasks concurrently and enhance its overall performance. By sharing task-specific knowledge, the model improves generalization and accuracy across diverse text classification challenges.
- **Advanced Attention Mechanism**: The model incorporates an attention mechanism that facilitates cross-task knowledge integration. This attention layer intelligently combines outputs from different tasks, allowing the model to dynamically focus on the most relevant information from each task. As a result, the model benefits from a broader understanding of the data, as insights gained from one task can enhance the performance on others. This inter-task knowledge sharing is a key driver of the model's superior accuracy and efficiency.
- **Improved Computational Efficiency**: While enhancing accuracy through multi-task knowledge sharing, the proposed method also significantly optimizes computational resources. By using LoRA, the model reduces the number of trainable parameters, allowing it to operate efficiently on a single GPU without compromising on performance. This combination of enhanced learning and reduced computational demands makes the approach highly suitable for practical, large-scale deployments.
- **Comprehensive Experimental Validation:** To evaluate the proposed approach, we conduct extensive experiments on multiple benchmark datasets for diverse text classification tasks. Our results demonstrate that the hybrid method achieves improved accuracy and efficiency compared to both single-task fine-tuning and the base model.

Preliminaries

This section delves into the preliminary concepts,

including the fine-tuning of large language models, parameter-efficient fine-tuning methods, and the attention mechanism.

A. Large Language Models

Large Language Models represent some of the most cutting-edge advancements in artificial intelligence, characterized by their ability to generate text with high precision and quality. These models leverage complex architectures and deep neural networks, which enable them to produce text that is both coherent and contextually appropriate across a diverse array of topics. The foundational architecture of these models is built upon transformers, a breakthrough introduced in 2017 that rapidly became a cornerstone in natural language processing due to its unparalleled efficacy [9]. Key applications of these models include automated content generation, machine translation, natural language processing, and question-answering systems. LLMs can analyse vast amounts of data and learn linguistic patterns, allowing them to generate sentences that are logical and meaningful, often indistinguishable from text written by humans [2].

Prominent examples of LLMs include GPT [10] by OpenAI, BERT [3], and T5 [11] by Google. These models, through the analysis of vast amounts of data, have learned intricate linguistic patterns, enabling them to generate sentences that are both logical and contextually rich, making it challenging to differentiate their output from text authored by humans. The primary advantage of these models lies in their exceptional ability to understand and generate high-quality text across multiple languages, as well as to provide accurate responses to questions of varying complexity. As a result, LLMs have found extensive applications in areas such as machine translation, chatbots, automated content creation, and even recommendation systems. Given these capabilities, LLMs are not only powerful tools for natural language processing but have also become foundational pillars in the development of AI-driven technologies. However, to fully harness their potential, these models often require fine-tuning for specific tasks. This fine-tuning is essential for optimizing their performance in particular applications.

In this research, we adopt a hybrid approach using Low-Rank Adaptation (LoRA) to fine-tune large language models (LLMs) for multi-task text classification. The hybrid approach is designed to optimize computational efficiency while enhancing model accuracy, primarily by facilitating inter-task knowledge sharing. This approach enables the model to benefit from the learning outcomes of different tasks simultaneously, allowing it to leverage relevant information gained across tasks to improve overall performance. In this context, our hybrid approach ensures that the advantages of multi-task learning—particularly the ability to transfer knowledge across tasks—are fully realized. The dynamic sharing of task-specific knowledge allows the model to become more robust, mitigating issues such as overfitting to a particular dataset while also enhancing its ability to adapt to new, unseen tasks. By fine-tuning different tasks in parallel and allowing for cross-task information flow, our model achieves higher performance metrics than traditional fine-tuning approaches that isolate task learning.

B. Fine-Tuning Language Models

Training a large language model (LLM) from scratch requires significant time and financial resources. The use of thousands of GPUs can take several days [12] and demands substantial financial investment [13]. Fine-tuning pre-trained models has emerged as an efficient way to leverage the benefits of LLMs: fine-tuning is the process of adapting a pre-trained model to a specific task by training it further on task-specific data, thereby improving its performance on that task. This approach has gained widespread acceptance, as it allows researchers to utilize general-purpose pre-trained models and tailor them to meet specific needs. Many organizations, such as Meta with their LLaMA models [6], make their pre-trained models publicly available. These publicly available pre-trained models can be fine-tuned for various downstream tasks, making fine-tuning the most practical way to capitalize on the benefits of LLMs. However, fully fine-tuning large language models is computationally expensive, as it requires updating all the parameters of the model.

C. Parameter-Efficient Fine-Tuning Methods

Full fine-tuning of large pre-trained models generally requires updating all model parameters, often resulting in substantial computational costs [4]. In contrast, parameter-efficient fine-tuning (PEFT) methods [14] selectively adjust only a small number of additional parameters, leading to a significant reduction in computational and memory resources. One of the most advanced PEFT techniques is the Low-Rank Adaptation (LoRA) method [5], which achieves efficient fine-tuning by keeping the pre-trained model entirely frozen and applying weight updates through a trainable low-rank decomposition matrix, as illustrated in Fig. 1.

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

where x represents the input data from the target task, $W_0 \in R^{d\times k}$ are the weights of the pre-trained model that remain fixed, $B \in R^{d\times r}$ and $A \in R^{r\times k}$ and are the trainable low-rank decomposition matrices, where $r \ll \min(d, k)$ is the rank of the decomposition. Fig. 1 illustrates the LoRA method, which we employed as the primary tool for fine-tuning large language models in this research.

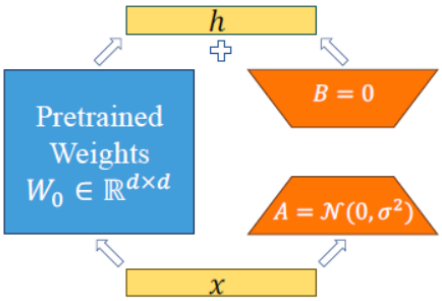


Fig. 1: Reparameterization in LoRA Method [5]. This method only trains A and B. The Pretrained Weight of model were frozen.

By utilizing LoRA, our research has been able to significantly reduce computational costs while simultaneously improving the accuracy and efficiency of the models across various text classification tasks.

D. Attention Mechanism

The attention mechanism is one of the key innovations in modern AI architectures, particularly in transformer-based models. This mechanism allows models to focus more on the relevant and important information within the data. Essentially, the attention mechanism assigns different weights to different inputs, helping the model to select the most important parts of the data for processing, thereby improving the overall performance of the model.

In transformer-based models, the attention mechanism is implemented as self-attention, where each word in a sentence attend to all other words in the same sentence. This mechanism consists of three matrices: query, key, and value. For each token, the model assigns different weights to each of these matrices based on their similarity with other tokens. The basic formula for the attention mechanism is given as (2).

$$Attention(Q, K, V) = softmax\left(\frac{QK}{\sqrt{d_k}}\right)V \quad (2)$$

where Q is the matrix of queries, K is the matrix of keys, and V is the matrix of values. The term d_k is a normalization factor that prevents the resulting values from becoming too large due to the dot product of the queries and keys. This process enables the model to identify and focus on the most relevant and important parts of the data [9].

To further enhance the model's ability to capture complex relationships within the data, the Multi-Head Attention mechanism is introduced. In Multi-Head Attention, instead of computing a single attention function, the model computes multiple attention functions (heads) in parallel. Each head independently performs the attention operation, capturing different aspects of the relationships between words. The outputs of these attention heads are then concatenated and linearly transformed to form the final output. The formula for Multi-Head Attention is as follows:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^0 \quad (3)$$

where each $head_i$ is calculated as:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (4)$$

here, W_i^Q, W_i^K and W_i^V are the weight matrices for the queries, keys, and values for the i-th attention head, and W^0 is the output weight matrix. The Multi-Head Attention mechanism allows the model to jointly attend to information from different representation subspaces at different positions, providing a more comprehensive understanding of the input data [9].

The attention mechanism is not only used as a tool to enhance the performance of language models, but it can also be leveraged to aggregate and combine the outputs of multiple models. This application is particularly useful in complex systems that require the integration of information from various sources. In such scenarios, the attention mechanism can assist the model in intelligently determining which parts of the different outputs should be given more focus, ultimately producing a higher-quality combined output. In this approach, the outputs of several models are fed as inputs to an attention layer. This allows the models to flexibly utilize the knowledge and information from multiple sources, thereby achieving higher accuracy and efficiency in solving complex tasks [15].

For instance, in a multilingual machine translation system, the outputs of different models trained for various languages can be combined using the attention mechanism to create more accurate and reliable translations. Similarly, in recommendation systems or data analysis applications, the outputs of multiple models can be aggregated using attention to achieve better results [16]. The effectiveness of using the attention mechanism for output aggregation lies in its ability to automatically learn which outputs and their components are more significant under different circumstances, thereby improving the final output and enhancing decision-making quality.

In this study, we employed an attention mechanism to intelligently combine the outputs of fine-tuned multi-task models, enabling more effective cross-task knowledge sharing. By integrating outputs from multiple tasks, this approach allows language models fine-tuned using the Low-Rank Adaptation method to draw on the relevant insights learned from different tasks concurrently. Instead of treating each task as isolated, the attention mechanism dynamically identifies and emphasizes critical features from each task, thereby enhancing the model's ability to generalize across diverse datasets.

Related Work

This section reviews related works in the field, starting with a number of studies that have full fine-tuned language models, adjusting all model parameters. Following this, we introduce several studies that have utilized parameter-efficient fine-tuning (PEFT) methods. Finally, we discuss works that have employed hybrid fine-tuning approaches.

A. Full Fine-Tuning of Language Models

Fully fine-tuning language models is a widely used approach for task-specific optimization. For instance, In [17], the authors explore the use of the BERT model for sentiment analysis, employing a full fine-tuning approach. In this method, the pre-trained BERT model, initially trained on a large text corpus, is further fine-tuned on a specific sentiment analysis dataset. This involves updating all the parameters of the BERT model to adapt it to the task at hand. The full fine-tuning process takes advantage of BERT's capability to understand the intricate dependencies between words in a sentence and to leverage various contextual cues. As a result, the fine-tuned model demonstrates significantly improved performance in sentiment analysis tasks.

Similarly, In another study [18], a new model called LUKE is introduced, which is specifically designed to enhance the representation of entities within text. LUKE incorporates an entity-aware self-attention mechanism, allowing the model to pay special attention to entities mentioned in the text. By fully fine-tuning the BERT model on several tasks related to entity recognition, LUKE is able to accurately identify and represent the semantic relationships between entities and the surrounding text, resulting in better overall performance.

In the work of Nogueira et al. [19], BERT is utilized to enhance the performance of passage re-ranking tasks. The approach involves a two-step process: first, an initial retrieval system provides a list of candidate passages for a given query. Then, the BERT model, which has been fully fine-tuned, is used to re-rank these passages. The BERT model processes the input queries and candidate passages, encoding the sentences and analysing the semantic relationships between the words. This allows the model to more accurately determine the relevance of each passage to the query, improving the final ranking.

Khashabi et al. [20] introduce the UnifiedQA model, which is built on transformer architectures and trained using a transfer learning approach. This model is capable of answering questions across various formats. The key innovation is the unification of all question-answering tasks into a single text format, which is then used to fully fine-tune the language model. This unification allows the model to leverage a larger and more diverse dataset during training, leading to increased accuracy and flexibility in handling different types of question-answering tasks.

Full fine-tuning allows models to adapt comprehensively to specific tasks, maximizing their performance. However, this approach requires updating all model parameters, leading to high computational costs and memory requirements, making it less feasible for resource-constrained environments or scenarios requiring frequent model updates.

B. Parameter-Efficient Fine-Tuning (PEFT)

To mitigate the resource demands of full fine-tuning, parameter-efficient fine-tuning (PEFT) methods have been proposed. Pfeiffer et al. [21] introduced a novel approach for adapting transformer models using adapter modules. Rather than fully fine-tuning all the parameters of the model, this method involves adding a small set of new parameters, known as adapters, for each specific task. These adapter modules are connected to different layers of the transformer model, allowing the model to be optimized for various tasks without altering the primary parameters of the model. This approach significantly reduces the time and computational resources required for fine-tuning, while also facilitating the sharing of base models across multiple tasks. The results demonstrate that using adapter modules can achieve performance comparable to full fine-tuning methods, and in some cases, even surpass them.

Li and Liang [22] introduced Prefix-Tuning, another PEFT method, where instead of fine-tuning all model parameters, a continuous prefix of text is optimized. This prefix is added to the input of a transformer model, guiding it during text generation. This method offers a cost-effective and flexible alternative to traditional full fine-tuning approaches. The prefix is directly appended to the input sequence, and only a small number of parameters related to this prefix are optimized, leaving the rest of the model unchanged. The models used in their experiments, primarily GPT-2 and other generative transformer models, demonstrated that Prefix-Tuning could achieve performance similar to, or even better than, full fine-tuning while only optimizing a small portion of the model's parameters.

Stickland and Murray [23] proposed the PALs (Projected Attention Layers) method, another PEFT approach that adapts the BERT model to various tasks by replacing standard attention layers with projected attention layers. This method allows the model to be effectively adapted to various tasks by adding additional parameters to the attention layers, which are fine-tuned for each specific task. This approach enables efficient adaptation without the need to fine-tune all of the model's parameters, significantly reducing the computational load while maintaining high performance across multiple tasks.

Zhang et al. [24] introduced LoRA-FA, a novel method designed to improve the efficiency of fine-tuning large language models. The primary goal of LoRA-FA is to reduce the memory overhead associated with fine-tuning LLMs by minimizing the need for activation memory without sacrificing performance or incurring heavy recomputation costs. LoRA-FA works by keeping the low-rank weight matrix A fixed while only updating the higher-rank weight matrix B. This effectively reduces the activation memory required during the fine-tuning process. The method has been extensively tested across various models and scales, with empirical results showing that LoRA-FA offers comparable accuracy to full parameter fine-tuning and LoRA, while significantly reducing memory costs.

PEFT methods significantly reduce computational and memory requirements, enabling fine-tuning on resource-limited devices. However, these methods may not achieve the same level of performance improvement as full fine-tuning, especially for tasks requiring deep model adaptations.

C. Hybrid Fine-Tuning Approaches

Hybrid approaches aim to balance the comprehensive adaptation of full fine-tuning with the efficiency of PEFT. These approaches combine the strengths of both full fine-tuning and parameter-efficient fine-tuning methods, aiming to optimize model performance while minimizing computational resources. These approaches leverage the comprehensive capabilities of full fine-tuning by updating a significant portion of the model's parameters while simultaneously employing parameter-efficient techniques to reduce the overall computational cost.

Sar et al. [25] introduced the USE model, designed to generate high-quality representations of sentences and textual phrases. This model is particularly geared towards improving performance across a range of natural language processing (NLP) tasks and machine learning applications. The USE model employs two main architectures for generating vector representations of sentences:

Transformer-based Encoder: This version of the USE model is built upon the transformer architecture and utilizes the multi-head attention mechanism to understand semantic dependencies within the text. Due to its ability to model complex relationships between words, this architecture is particularly well-suited for applications requiring high precision and detail.

Deep Averaging Network (DAN): This version is lighter and faster, focusing on the simplicity of averaging word embeddings rather than employing the full transformer architecture.

In [26], the authors propose an approach that integrates Low-Rank Adaptation (LoRA) with traditional fine-tuning techniques. This hybrid method allows for the efficient fine-tuning of large language models by updating both a small, low-rank set of parameters and a larger set of fully fine-tuned parameters. This combination enables the model to achieve high performance with reduced computational costs compared to traditional full fine-tuning methods.

Similarly, the work by Kim et al. [27] explores a hybrid fine-tuning approach that combines Prefix-Tuning with full parameter fine-tuning. By optimizing a prefix of continuous text in conjunction with full fine-tuning, the model benefits from the flexibility and efficiency of parameter-efficient methods while retaining the comprehensive improvements provided by full fine-tuning. The results demonstrate that this hybrid approach can achieve performance levels comparable to full fine-tuning while requiring fewer computational resources.

Hybrid approaches effectively combine the strengths of full and parameter-efficient fine-tuning, achieving a good trade-off between performance and resource efficiency. However, their complexity can increase implementation overhead and may require careful tuning to balance the contributions of different components.

By critically analysing these methods, this paper identifies the potential of hybrid approaches to leverage the advantages of knowledge sharing and efficiency, forming the foundation for the proposed hybrid fine-tuning method.

Methodology

In this section, we introduce a novel hybrid model designed to enhance the fine-tuning process of large language models (LLMs) using the Low-Rank Adaptation (LoRA) technique. Our proposed method is structured to combine the strengths of multiple fine-tuned models to improve overall accuracy and efficiency in various text classification tasks. Each component model in our hybrid architecture shares the same pre-trained weights W^*, but has its own learnable parameters ΔW_m, which correspond to the matrices A and B in the LoRA method. After fine-tuning for each task, the weights W_m are defined as:

$$W_m = W^* + \Delta W_m \tag{5}$$

During the fine-tuning process on the data of task m, the objective is to optimize ΔW_m. For this optimization is formulated as follows:

$$\mathcal{L}(\Delta W_m) = \min_{\Delta W} \sum_{n=1}^{N} -\log p(y_n \,|X_n;\, W^* + \Delta W_m) \tag{6}$$

where N denotes the number of training samples for task m, X_n represents the input samples, and y_n are their corresponding labels.

Fig. 2 illustrates the proposed framework and the integration of these fine-tuned models.

Each of these fine-tuned models functions as an independent module within the proposed framework, as depicted on the left side of Fig. 2. These modules are

specifically tailored to handle individual tasks by employing the LoRA (Low-Rank Adaptation) fine-tuning technique. This approach allows for efficient adaptation of large language models to specific tasks without requiring complete retraining of the model, thus optimizing computational resources. For example, as shown in the left section of Fig. 2, the i-th module is designed for the task of sentiment analysis. The input to this module is a sentence, which could represent any textual content, such as a review, a social media post, or a customer feedback comment. The module processes the input and classifies it into one of the predefined categories like here: positive or negative sentiment.

To ensure accurate classification, a task-specific head is integrated into the module. This head maps the model's internal representations to the class space relevant to the sentiment analysis task. This mapping step is crucial for transforming the abstract feature representations learned by the model into interpretable outputs aligned with the specific requirements of the task.

The training process is guided by the computation of an error signal, which quantifies the difference between the predicted class and the true class label. This error is calculated using the cross-entropy loss function (7), a widely used metric in classification problems that effectively handles the probabilistic nature of model outputs. Once the error is computed, it is backpropagated through the network, enabling the model to adjust its parameters and improve its performance on the sentiment analysis task during subsequent iterations.

$$Loss_{cls} = \mathcal{L}(Head(LM(X)), y) \quad (7)$$

This modular design not only enhances the flexibility of the overall framework but also ensures that each module is highly specialized and optimized for its respective task, ultimately contributing to the robustness and adaptability of the proposed system.

In this way, it is possible to fine-tune k distinct modules on k different tasks. Each module is trained to learn task-specific knowledge, ensuring optimal performance for its assigned task. This approach ensures that the expertise gained by each module is highly specialized, tailored to the unique requirements of the respective task.

The main proposed model, depicted on the right side of Fig. 2, is essentially a combination of these individual modules. By integrating the knowledge acquired by each module, the overall model leverages the specialized expertise of all modules to perform complex or multifaceted tasks effectively.

This design allows the system to benefit from the modularized training of diverse tasks, enabling a flexible and scalable architecture. As a result, the proposed framework can tackle multiple tasks simultaneously or sequentially by drawing on the task-specific knowledge embedded within its constituent modules. This modular combination enhances the adaptability and efficiency of the model, particularly in multi-task learning scenarios.

Based on the model depicted in Fig. 2, the outputs of the fine-tuned models are initially aggregated using the aggregation function Agg1 and converted into a vector. It is important to note that these model outputs represent the values computed by the language model before passing through the softmax layer and being transformed into probability values. Subsequently, these aggregated values pass through an attention layer. The attention layer computes a new vector for each of these input vectors, determining how much attention each task's output should receive and adjusting the values accordingly. All outputs from the attention layer are then passed to a second aggregation function, Agg2, and finally combined with the output of the neural network layer using the Agg3 function, forming the final output of the model.

Aggregation functions are versatile tools designed to combine multiple inputs into a singular, cohesive output. The most widely used aggregation functions include:

Mean Function: Computes the average of the input values, providing a balanced representation of the data.

Max/Min Function: Identifies the highest or lowest value among the inputs, highlighting the most extreme values.

Sum Function: Adds up all input values, offering a cumulative measure of the inputs.

These are just a few examples, and a variety of other functions can also be utilized depending on the specific needs of the task at hand.

Attention Layer: This layer utilizes the self-attention mechanism, where the outputs from the Agg1 function serve simultaneously as queries (Q), keys (K), and values (V). After applying Agg1, each output from the fine-tuned models functions as its own query, key, and value. This allows the model to not only focus on its own output but also to dynamically attend to the outputs of other models fine-tuned on different tasks. This interaction enables the model to weigh the relevance of each output in the context of the others, leading to a more informed and refined final result.

Normalization: To enhance the model's performance, the aggregated data is normalized. Normalizing the data ensures that the model inputs fall within a specified range, which aids in accelerating the learning process and improving prediction accuracy. Moreover, this process prevents the model from disproportionately focusing on features with larger scales, which could lead to imbalances in learning. Overall, data normalization not only reduces the model's training time but also helps in improving its generalization ability.

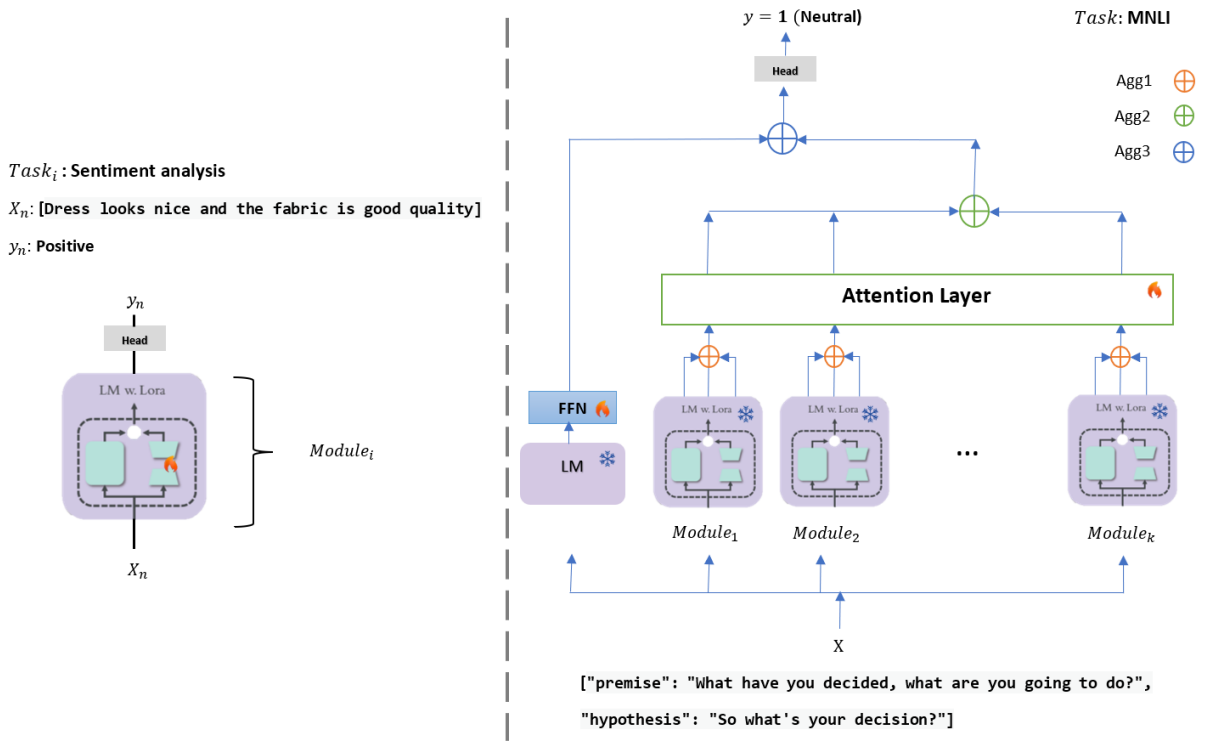


Fig. 2: The left image depicts a single module undergoing fine-tuning of a language model on a specific task (sentiment analysis) using the LoRA method. The right image illustrates the proposed framework, which combines these fine-tuned modules for the target task (here, MNLI).

For normalization, after each aggregation function, a normalization layer is added. In this layer, after calculating the mean μ and variance σ^2, the data is normalized according to (8):

$$\hat{X}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{8}$$

where ϵ is a small value added to prevent division by zero. Following normalization, by adding two learnable parameters, the output of the normalization layer is given by

$$y_i = \gamma \hat{x}_i + \beta \tag{8}$$

During the training phase of the proposed model, only the parameters of the Feed-Forward Network (FFN) and the attention layers are updated. The language model and all k modules remain frozen, with only their learned knowledge being shared across tasks. This design allows the model to focus on learning the new task (e.g., the MNLI task shown in Fig. 2) without altering the parameters of the pre-trained language model or the task-specific modules.

By leveraging this knowledge-sharing property, the framework can achieve high performance and accuracy on previously unseen tasks without the need to fine-tune the entire language model for each new task. This approach ensures that the expertise gained from prior tasks is effectively utilized to generalize to new scenarios, significantly reducing the need for extensive retraining.

Additionally, since the proposed framework employs parameter-efficient methods during the module training phase, such as LoRA (Low-Rank Adaptation), the computational cost of training is kept minimal. Unlike full fine-tuning approaches, which require updating the entire language model, this method modifies only a small subset of parameters. Consequently, the training and inference processes are computationally efficient, allowing the model to run effectively on a single GPU.

This efficiency makes the framework practical and scalable, particularly in resource-constrained settings, while maintaining high performance across both seen and unseen tasks. It demonstrates the capability of leveraging modular and efficient design principles to achieve robust task adaptation with minimal computational overhead.

Experiments and Results

To evaluate the effectiveness of the proposed method, several experiments were conducted. This section details the experiments and the results obtained from them.

A. Datasets

To fine-tune and evaluate the model, the GLUE dataset [28] was used, which encompasses multiple tasks. The details of the dataset are provided in Table 1.

To assess the generalization ability of the model, four additional datasets—STS-B, IMDB, AG News, and TREC—were employed, none of which were used in the training process of the hybrid method.

Table 1: Details of the datasets

	Task	Type	#Class	#Train	#Test	Labels	Metric
Single-sentence	SST-2	Sentiment	2	6920	872	positive, negative	Accuracy
	CoLA	acceptability	2	8551	1042	grammatical, not_grammatical	Matthews_Correlation
Sentence-pair	MNLI	NLI	3	392702	9815	entailment, neutral, contradiction	Accuracy
	QNLI	NLI	2	104743	5463	entailment, not_entailment	Accuracy
	RTE	NLI	2	2490	277	entailment, not_entailment	Accuracy
	WNLI	NLI	2	635	72	entailment, contradiction	Accuracy
	MRPC	Paraphrase	2	3668	408	equivalent, not_equivalent	Accuracy
	QQP	Paraphrase	2	363846	40431	equivalent, not_equivalent	Accuracy
Single-sentence	IMDB	Sentiment	2	25000	25000	positive, negative	Accuracy
	AG_NEWS	News Categorization	4	120000	7600	World, Sports, Business, Science/Technology	Accuracy
	TREC	Question Classification	6	5452	500	Abbreviation, Entity, Description, Human, Location, Numeric	Accuracy
Sentence-pair	STS-B	Sentiment. similarity	Regression	5749	1500	-	Pearson

The STS-B dataset, part of the GLUE benchmark [28], is designed for evaluating semantic similarity between sentence pairs. Each pair is annotated with a similarity score ranging from 0 to 5, where higher scores indicate greater semantic similarity.

The IMDB dataset [29], widely used for sentiment analysis, contains 50,000 movie reviews evenly split into 25,000 training samples and 25,000 test samples. Each review is labelled as either positive or negative, and the dataset is balanced, ensuring equal representation of both sentiment classes.

The AG News dataset [30] is a benchmark dataset used for news categorization. It consists of 120,000 training samples and 7,600 test samples, categorized into four classes: World, Sports, Business, and Science/Technology. Each sample includes a title and a brief description of the news article, making it ideal for evaluating text classification methods.

The TREC dataset [31], widely utilized for question classification, contains 5,452 training questions and 500 test questions categorized into six main types: Abbreviation, Entity, Description, Human, Location, and Numeric. These classes are further divided into finer subcategories, offering a hierarchical structure that is useful for question classification and intent detection tasks.

B. Evaluation Metrics

In this subsection, we describe the evaluation metrics used to assess model performance across different datasets, as summarized in Table 1.

Each metric is selected based on the specific nature and objectives of the corresponding task.

Accuracy is a standard evaluation metric for classification tasks. It measures the proportion of correctly predicted samples relative to the total number of samples. This metric is widely used in datasets such as SST-2, IMDB (sentiment analysis), AG_NEWS, TREC, MNLI, QNLI, RTE, WNLI (natural language inference), and MRPC, QQP (paraphrase detection) [28]. The formula for accuracy is given as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

where:

TP (True Positives): The number of correctly predicted positive samples.

TN (True Negatives): The number of correctly predicted negative samples.

FP (False Positives): The number of incorrectly predicted positive samples.

FN (False Negatives): The number of incorrectly predicted negative samples.

Matthews Correlation Coefficient (MCC) [28] is a robust evaluation metric specifically suited for binary classification tasks, particularly in scenarios with imbalanced datasets. It considers all four categories of the confusion matrix (true positives, true negatives, false

positives, and false negatives) to provide a balanced assessment of model performance. In this study, MCC is employed for the CoLA dataset, which focuses on grammatical acceptability. The formula for MCC is:

$$MCC = \frac{(FP*FN)-(TP*TN)}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \quad (10)$$

MCC ranges from −1 to +1, where +1 indicates perfect prediction, 0 indicates no better than random prediction, and −1 indicates total disagreement between prediction and observation.

Pearson Correlation Coefficient [28] is used to measure the strength and direction of the linear relationship between predicted and actual values in regression tasks. It is particularly relevant for the STS-B dataset, where the objective is to evaluate semantic similarity scores between sentence pairs. A higher correlation indicates a stronger agreement between predicted and true scores. The formula for the Pearson correlation is:

$$Pearson\ Correlation = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (11)$$

where:

x_i and y_i: Predicted and actual values, respectively.

$\bar{x}$ and $\bar{y}$: Mean of predicted and actual values, respectively.

n: Number of samples.

A higher correlation value (closer to +1) indicates a stronger agreement between predicted and true scores.

C. Experimental Settings

For the experiments, the language models DistilBERT [32] with approximately 66 million parameters, BERT-base [3] with approximately 110 million parameters, and the ELECTRA [33] model in both small and base versions with approximately 14 million and 110 million parameters, respectively, were used. Each of the tasks was fine-tuned individually on these models using the LoRA method. The implementations were carried out in Python using the Transformers [34] and Huggingface PEFT [35] libraries.

In the experiments, the number of fine-tuned models was set to k=9, except for the experiment examining the number of tasks, where this variable is adjusted. During the LoRA fine-tuning process, the ΔW_m matrices were applied to all three matrices—key, value, and query—in the language model's attention module with a rank of r=4, chosen based on preliminary experiments where this value provided a balance between computational efficiency and model accuracy.

Lower values of r were found to reduce the model's expressive power, while higher values increased computational costs without significant gains in performance. The AdamW optimizer was used in all experiments, and the models were trained for 20 epochs. The initial learning rate was set to $2e-5$ with weight decay=0.01, and the batch size was 16 for all datasets. We employed the NVIDIA GeForce RTX 3090 24GB. The mean function was used for both aggregation functions Agg1 and Agg2, while the sum function was employed for Agg3. The number of heads in the model is 4. The code related to this paper is available at this link[1].

D. Evaluation of the Proposed Method's Performance

In this section, we present the results of training the proposed method on all datasets after 20 epochs. The results, as shown in Table 2 for the two language models utilized, indicate the validation accuracy achieved.

Table 2: The effectiveness of the proposed method (AttEns) and other methods

LLM	Method	WNLI	QQP	MRPC	RTE	MNLI-mm	MNLI	QNLI	CoLA	SST2	Average
DistilBERT	Model	0.5493	0.3771	0.3161	0.4729	0.3267	0.3178	0.4794	0.0000	0.4908	0.3700
	FFN	0.4366	0.7648	0.7107	0.5523	0.5288	0.5238	0.6908	0.1458	0.8337	0.5763
	FT_LoRA	0.4366	0.8432	0.6838	0.4440	0.7389	0.7426	0.7426	0.0000	**0.9162**	0.6164
	AttEns	**0.6447**	**0.8544**	**0.8014**	**0.6642**	**0.7610**	**0.7523**	**0.8473**	**0.3952**	0.8933	**0.7348**
BERT-base	Model	0.5633	0.6066	0.6789	0.5090	0.3302	0.3226	0.4880	0.0950	0.5149	0.4565
	FFN	0.3802	0.7600	0.6985	0.5956	0.5372	0.5294	0.6847	0.3084	0.8589	0.5947
	FT_LoRA	**0.5774**	0.8615	0.7009	0.5848	0.8087	0.7940	**0.8813**	**0.4993**	0.9082	0.7351
	AttEns	0.5633	**0.8654**	**0.8186**	**0.6714**	**0.8102**	**0.8007**	0.8795	0.4988	**0.9105**	**0.7576**
ELECTRA-Small	Model	0.5915	0.4742	0.3137	0.5270	0.3303	0.3216	0.5249	0.0104-	0.4988	0.4477
	FFN	0.5774	0.7542	0.7132	0.5523	0.4869	0.4787	0.7100	0.3268	0.6938	0.5881
	FT_LoRA	0.5633	0.8362	0.7377	0.5631	0.7823	0.7671	0.8504	0.3463	0.8772	0.7026
	AttEns	**0.5774**	**0.8396**	**0.8235**	**0.6787**	**0.7870**	**0.7708**	**0.8533**	**0.5312**	**0.8853**	**0.7496**
ELECTRA-Base	Model	0.3943	0.3604	0.6838	0.5018	0.3585	0.3532	0.4946	0.020-	0.5080	0.4568
	FFN	0.4788	0.8000	0.7436	0.5956	0.6472	0.6320	0.7810	0.5208	0.8337	0.6703
	FT_LoRA	0.4084	0.8822	0.8431	0.7003	0.8647	0.8607	0.9203	0.6012	0.9288	0.7788
	AttEns	**0.5774**	**0.8825**	**0.8627**	**0.7689**	**0.8651**	**0.8632**	**0.9211**	**0.6362**	**0.9369**	**0.8126**

[1] https://github.com/Azadeh297/Attention-hybrid-method

The models evaluated are as follows:

Model: The base language model without any additional training or fine-tuning was used to infer the data, and its accuracy was calculated.

FFN: A single neural network layer was added on top of the base language model and fine-tuned. In this configuration, the parameters of the base model remain fixed, and only the parameters of the added neural network layer are updated.

FT_LoRA: The language model was fine-tuned separately on the data using the LoRA method.

AttEns: The proposed method.

The values listed in Table 2 for all datasets represent accuracy, except for the CoLA dataset, where the evaluation metric is the Matthews correlation coefficient. Based on the results in Table 2, the proposed method has achieved the highest accuracy in most tasks. This improvement is attributed to the use of the attention mechanism and the integration of information from multiple tasks, allowing the model to identify more complex patterns and thereby achieve higher accuracy.

To better explain this improvement, consider the QNLI task, which is related to natural language inference. In this task, the AttEns method achieved an accuracy of over 92%. One reason for this improvement could be the method's ability to recognize complex relationships between sentences. For instance, imagine that the model needs to compare two sentences to determine whether the second sentence is a logical conclusion of the first. In traditional methods like FFN or FT_LoRA, this process is carried out directly without utilizing information from other tasks. However, in the AttEns method, the model also leverages information from other tasks, such as recognizing contradictions in MNLI and semantic sentence matching in QQP. This combination of information allows the model to perform better in identifying complex relationships, particularly in tasks related to natural language inference.

In tasks like RTE, MRPC, and WNLI, there is a significant difference in the accuracy obtained from the proposed method compared to the single fine-tuned model. One possible reason for this is the smaller number of samples in these datasets compared to others. The proposed hybrid model has demonstrated higher accuracy in situations where limited data is available, potentially because the single fine-tuned model might have overfitted due to the small dataset size, resulting in poorer performance. This finding suggests that the proposed method can offer better generalizability even in data-limited scenarios.

By comparing the results obtained from different language models, in many cases, the larger language model has yielded better results, indicating that using a larger language model with more parameters can be effective in improving model performance.

To further evaluate the proposed method, we applied it to four additional datasets, IMDB, STS-B, AG_NEWS and TREC where the fine-tuned models for these datasets were not used in the combination of the proposed method. The results are shown in Table 3. According to the results, the AttEns method outperformed the FFN and FT_LoRA methods on all four datasets. For example, on the IMDB dataset, the AttEns method achieved an accuracy of 0.9397, which is clearly better than the FFN (0.8405) and FT_LoRA (0.8789) methods. This improvement in accuracy demonstrates the proposed method's high generalization ability to unseen data and indicates better performance in real-world scenarios. Imagine that the model needs to detect the sentiment of a movie review in the IMDB dataset. In traditional methods like FFN or FT_LoRA, the model only uses the training data from the same dataset, limiting its ability to generalize to new data. However, in the AttEns method, the model uses the attention mechanism to also leverage information from other related tasks. For example, if the model learned how to identify similar sentences in the QQP dataset, it could enhance this knowledge and apply it to better understand the sentiment in movie reviews.

Table 3: Results of the proposed method on two datasets not used in the combination of the proposed method (BERT-base language model)

Method	STS-B	IMDB	AG_NEWS	TREC
Model	-0.0608	0.4956	0.2505	0.0180
FFN	0.2247	0.8405	0.8942	0.7666
AttEns	**0.4158**	**0.9397**	**0.9107**	**0.8220**

On the STS-B dataset, the AttEns method also achieved an accuracy of 0.4158, which is an improvement compared to other methods. This improved performance indicates that the proposed model has high generalization ability even in scenarios where data is limited or heterogeneous.

The results show that using the attention mechanism and integrating information from various tasks enables the model to better identify and analyse complex features in new data, which is particularly important in real-world scenarios and unseen data.

E. Evaluating the Impact of k

In this section, the impact of the number of components, k, in the proposed hybrid model is examined. The results of this experiment are presented in Fig. 3. This chart demonstrates that increasing the number of components generally improves the accuracy of the hybrid model, although the extent of this improvement varies at different points. As observed in Fig. 3, the accuracy of the model significantly improves when the number of components increases from 3 to 5.

This enhancement is due to the increased capacity of the model to learn and integrate diverse information from various tasks.

When the number of components reaches 7, accuracy continues to improve in some datasets, but the improvement is not as pronounced. As the number of components increases, the model can examine each part of the data with greater detail, resulting in better performance. However, while the model continues to improve in accuracy, this improvement becomes slower compared to earlier stages. This slowdown occurs because, with more components, the model needs to process more information, requiring more resources for complete and optimized processing. This observation highlights the importance of selecting an appropriate number of components to optimize the model's performance.

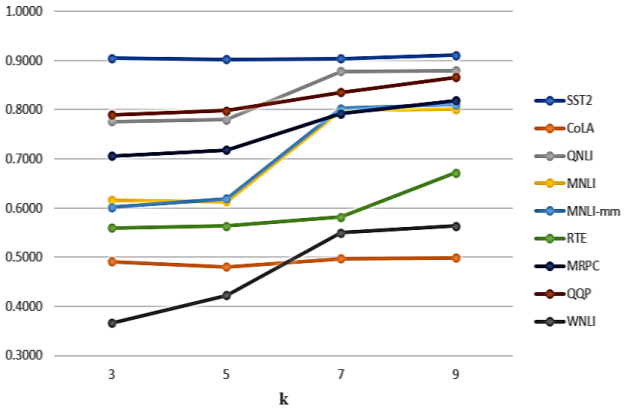


Fig. 3: Results of the proposed method for different values of k (BERT-base language model).

Limitations and Future Work

While the proposed hybrid fine-tuning method demonstrates notable improvements in accuracy and computational efficiency, several limitations must be acknowledged. Firstly, the method's effectiveness has been primarily validated on text classification tasks. Its applicability to other NLP domains, such as sequence generation or machine translation, remains unexplored and requires further investigation. Secondly, the reliance on specific datasets like GLUE may constrain the generalizability of the findings to other domains or languages. Future research should aim to extend this approach to a broader range of datasets and languages while examining its compatibility with other parameter-efficient fine-tuning techniques. Additionally, exploring the trade-offs between various aggregation functions and attention mechanisms could provide valuable insights for further optimizing model performance.

Results and Discussion

The results obtained from various datasets indicate that the proposed AttEns method consistently outperforms traditional fine-tuning approaches such as FFN and FT_LoRA across multiple NLP tasks. As shown in Table 2, AttEns achieves the highest accuracy in most datasets, particularly excelling in QNLI (92.11%), RTE (86.51%), and IMDB (93.97%). This improvement is primarily due to the model's ability to integrate task-specific information through an attention-based ensemble mechanism, which enhances its ability to identify complex patterns and generalize across tasks. Additionally, the method exhibits superior performance on datasets with limited samples, such as RTE and MRPC, where single-task fine-tuning often leads to overfitting.

Furthermore, the generalization ability of AttEns is evident in its strong performance on IMDB, STS-B, AG_NEWS, and TREC, despite these datasets not being explicitly incorporated into the model's training. The improvement in STS-B (41.58%) suggests that the attention mechanism enables the model to leverage knowledge from related tasks, leading to better sentence similarity evaluation. Moreover, the analysis of k, the number of task components, reveals that increasing k enhances performance up to a certain point, after which the improvement plateaus due to computational constraints. Overall, these results highlight the effectiveness of AttEns in improving language model accuracy, particularly in multi-task learning and low-data scenarios.

Conclusion

This paper introduced a hybrid approach for fine-tuning large language models using the LoRA method, which is capable of improving model accuracy by learning multiple tasks simultaneously. The results from the experiments showed that this method outperformed traditional fine-tuning methods, especially on the GLUE dataset. The use of the attention mechanism to integrate and influence different tasks was one of the main factors contributing to the success of this method. Additionally, the method demonstrated good generalizability on unseen data. Ultimately, this research marks a significant step towards reducing computational costs and enhancing the efficiency of large language models in various natural language processing tasks.

Author Contributions

All authors contributed equally to the conception, design, methodology, data analysis, manuscript preparation, and revision of this research. All authors have reviewed and approved the final version of the manuscript.

Acknowledgment

We sincerely thank the authors of previous studies whose work has informed and inspired this research. We

also extend our heartfelt appreciation to the respected referees for their thorough review of this paper.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

Abbreviations

NLP	Natural Language Processing
NLI	Natural Language Inference
LLM	Large Language Model
LoRA	Low Rank Adaptation
PEFT	Parameter-Efficient Fine-Tuning
MCC	Matthews Correlation Coefficient
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
FT	Fine Tuning
FFN	Feed Forward Network
GLUE	General Language Understanding Evaluation
CoLA	Corpus of Linguistic Acceptability
SST-2	Stanford Sentiment Treebank
MRPC	Microsoft Research Paraphrase Corpus
QQP	Quora Question Pairs
STS-B	Semantic Textual Similarity Benchmark
MNLI	Multi-Genre NLI
QNLI	Question NLI
RTE	Recognizing Textual Entailment
WNLI	Winograd NLI
IMDB	Internet Movie Database
TREC	Text Retrieval Conference
AttEns	Attention Ensemble

References

[1] K. I. Roumeliotis, N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," Future Internet, 15(6): 192, 2023.

[2] T. Brown et al., "Language models are few-shot learners," in Proc. Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 33: 1877-1901, 2020.

[3] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[4] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, X. Qiu, "Full parameter fine-tuning for large language models with limited resources," arXiv preprint arXiv:2306.09782, 2023.

[5] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[6] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[7] Hugging Face. https://huggingface.co/, 2023.

[8] Eric Wang. Alpaca-lora. https://github.com/tloen/alpaca-lora, 2023.

[9] A. Vaswani et al., "Attention is all you need," in Proc. Advances in neural information processing systems 30 (NIPS 2017), 2017.

[10] J. Achiam et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[11] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., 21(140): 1-67, 2020.

[12] D. Narayanan et al., "Efficient large-scale language model training on gpu clusters using megatron-lm," in Proc. the International Conference for High Performance Computing, Networking, Storage and Analysis: 1-15, 2021.

[13] O. Sharir, B. Peleg, Y. Shoham, "The cost of training nlp models: A concise overview," arXiv preprint arXiv:2004.08900, 2020.

[14] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, "Peft: State-of-the-art parameter-efficient fine-tuning methods," 2022.

[15] A. Hernández, J. M. Amigó, "Attention mechanisms and their applications to complex systems," Entropy, 23(3): 283, 2021.

[16] S. Dathathri et al., "Plug and play language models: A simple approach to controlled text generation," arXiv preprint arXiv:1912.02164, 2019.

[17] C. Sun, X. Qiu, Y. Xu, X. Huang, "How to fine-tune bert for text classification?," in Proc. Chinese computational linguistics: 18th China National Conference (CCL 2019): 194-206, 2019.

[18] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," arXiv preprint arXiv:2010.01057, 2020.

[19] R. Nogueira, K. Cho, "Passage Re-ranking with BERT," arXiv preprint arXiv:1901.04085, 2019.

[20] D. Khashabi et al., "Unifiedqa: Crossing format boundaries with a single qa system," arXiv preprint arXiv:2005.00700, 2020.

[21] J. Pfeiffer et al., "Adapterhub: A framework for adapting transformers," arXiv preprint arXiv:2007.07779, 2020.

[22] X. L. Li, P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," arXiv preprint arXiv:2101.00190, 2021.

[23] A. C. Stickland, I. Murray, "Bert and pals: Projected attention layers for efficient adaptation in multi-task learning," in Proc. International Conference on Machine Learning, PMLR: 5986-5995, 2019.

[24] L. Zhang, L. Zhang, S. Shi, X. Chu, B. Li, "Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning," arXiv preprint arXiv:2308.03303, 2023.

[25] D. Cer et al., "Universal sentence encoder," arXiv preprint arXiv:1803.11175, 2018.

[26] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," arXiv preprint arXiv:1701.06538, 2017.

[27] X. Wang, L. Aitchison, M. Rudolph, "LoRA ensembles for large language model fine-tuning," arXiv preprint arXiv:2310.00035, 2023.

[28] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," arXiv preprint arXiv:1804.07461, 2018.

[29] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, "Learning word vectors for sentiment analysis," in Proc. the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: 142-150, 2011.

[30] X. Zhang, J. Zhao, Y. LeCun, "Character-level convolutional networks for text classification," in Proc. Advances in neural information processing systems 28 (NIPS 2015), 2015.

[31] X. Li, D. Roth, "Learning question classifiers," in Proc. COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[32] V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[33] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.

[34] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations: 38-45, 2020.

Biographies

Azadeh Beiranvand Borjele completed Bachelor's degree in Software Engineering in 2005 and Master's degree in Artificial Intelligence in 2012 at Shahid Chamran University, Ahvaz, Iran. Currently, She is currently a doctoral student in Artificial Intelligence at University of Kashan, Kashan, Iran. Her research interests include graph representation learning, graph neural networks, large language models and dynamic complex networks.

- Email: a.Beiranvand@grad.kashanu.ac.ir
- ORCID: 0009-0007-7077-8896
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://scholar.google.com/citations?user=81IV9sEAAAAJ&hl=en

Mahdiye Sarhadi Dadiyan completed Bachelor's degree in Information Technology Engineering in 2010 at Zahedan PNU, Zahedan, Iran and Master's degree in Computer Engineering (Artificial Intelligence) in 2016 at Kharazmi University, Tehran, Iran. Right now, She is a doctoral student in Artificial Intelligence at University of Kashan, Kashan, Iran. Her research interests include reinforcement learning, deep learning, large language models.

- Email: mahdiye.sarhadi@grad.kashanu.ac.ir
- ORCID: 0009-0006-5351-1885
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Javad Salimi Sartakhti is an Assistant Professor of Artificial Intelligence in the department of Computer Engineering at the University of Kashan, Iran. He obtained his B.Sc. degree in computer engineering from the University of Kashan and his M.Sc. degree in Software Engineering from the Tarbiat Modares University, Tehran, Iran, in 2008 and 2013, respectively. In January 2017, he obtained his Ph.D. degree in Artificial Intelligence at the Isfahan University of Technology. He ranked first among students of computer engineering in all three degrees. His main research interests are LLM, NLP, and Deep learning.

- Email: salimi@kashanu.ac.ir
- ORCID: 0000-0003-1183-1232
- Web of Science Researcher ID: HJY-2812-2023
- Scopus Author ID: 51864592100
- Homepage: https://faculty.kashanu.ac.ir/salimi/en

How to cite this paper:
A. Beiranvand, M. Sarhadi, J. Salimi Sartakhti, “Hybrid fine-tuning of large language models using lora: enhancing multi-task text classification through knowledge sharing,” J. Electr. Comput. Eng. Innovations, 13(2): 417-430, 2025.

DOI: 10.22061/jecei.2025.11314.794

URL: https://jecei.sru.ac.ir/article_2303.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Comparative Evaluation of Model Predictive Current Controlled Matrix Converter versus AC-DC-AC Converter

M. Nabizadeh [1], ***P. Hamedani*** [2,*], ***B. Mirzaeian Dehkordi*** [1]

[1] *Department of Electrical Engineering, University of Isfahan, Isfahan, Iran.*

[2] *Department of Railway Engineering and Transportation Planning, University of Isfahan, Isfahan, Iran.*

Article Info

Article History:
Received 11 November 2024
Reviewed 27 January 2025
Revised 14 February 2025
Accepted 02 March 2025

Keywords:
AC-DC-AC converter
Current control
Matrix converter
Model Predictive Control (MPC)
Weighting factor

*Corresponding Author's Email Address:
p.hamedani@eng.ui.ac.ir

Abstract

Background and Objectives: Due to the disadvantages of the traditional AC-DC-AC converters, especially in electric drive applications, Matrix Converters (MCs) have been widely researched. MCs are well-known structures that remove the DC-Link capacitor and provide bidirectional power flow, while also giving the ability to control reactive power flow, which the AC-DC-AC converter lacks.
Methods: In this work, Model Predictive Current Control (MPCC) is utilized in conjunction with the MC to provide more versatility and controllability than traditional control methods. The work endeavors to investigate the current control of the MC utilizing the finite control set Model Predictive Control (MPC) approach.
Results: Current tracking performance, reactive power control, and switching frequency minimization have been included in the objective function of the controller. Moreover, the results have been compared to the traditional AC-DC-AC converters under similar circumstances. The MC can reduce the switching frequency by 40% compared to the AC-DC-AC converter while maintaining the same current THD value. Additionally, it achieves a 58% reduction in current THD compared to the AC-DC-AC converter at the same average switching frequency. However, in the MC, the mitigation of reactive power and the reduction in switching frequency have opposing effects on the current tracking performance.
Conclusion: This work proposes an MPCC method for the MC with an RL load, effectively controlling load current and reactive power. The reduction of switching commutations was also evaluated using different weighting factors in the prediction strategy for both the MC and AC-DC-AC converters. Simulation results demonstrate that the MC outperforms the AC-DC-AC converter in dynamic response and reactive power control.

Introduction

Nowadays, in multifarious industrial applications, Matrix Converters (MCs) are preferred as a replacement of the traditional AC-DC-AC converters including DC-link capacitors [1]-[4]. Eliminating the DC-link capacitors, results in a more compact, lower weight, and more reliable converter structure [4], [5]. Moreover, the MC has a relatively lower number of switching elements with bidirectional power flow ability [4], [5]. The ability to control the grid-side power factor and produce low distorted input and output waveforms makes MCs attractive for different industrial applications [5].

Since the MC topology was introduced, different methods have been applied to control it [4]. The pulse-width modulation [6], [7] approach and space vector modulation [8], [9] strategies are the two most prevalent approaches used to control the MCs. Moreover, in the control of electrical motor drives fed by MCs, Direct

Torque Control (DTC) has been utilized [10]-[12]. But, in combination with the MC, the mentioned methods are complicated. Other control methods have been proposed for MCs [13]-[16]. However, the main drawback of these control methods is their complexity which makes them unsuitable for industrial applications.

To provide a superior dynamic response and to overcome the complexity in the control of MCs, Model Predictive Control (MPC) was suggested. In addition, MPC is compatible with the system's nonlinear nature and restrictions [17]-[18]. Several works have reviewed the MPC strategy for MCs [17]-[21]. Moreover, in the case of induction motor drives supplied with an MC, different MPC methods have been suggested. The common MPC strategies that have been developed for MCs are Predictive Current Control (PCC) [22]-[24], Predictive Torque Control [25]-[27], Predictive Voltage Control [28]-[29], and Predictive Power Control [30]-[31]. Among them, PCC is the most popular. Different objectives have been considered in the cost function of the PCC, including reactive power [32]-[34], Common-Mode (CM) Voltage [35], switching Losses [36], and efficiency [36]. In [37], the standard weighting factor selection in the objective function is replaced by a fuzzy decision-making strategy, demonstrated through a case study on controlling load and supply currents in a direct matrix converter (DMC). This approach eliminates the need for weighting factors and introduces a simplified selection scheme. However, fuzzy control strategies rely heavily on expert knowledge, making them prone to errors if the rules or membership functions are poorly designed. They can also be computationally demanding in complex systems and lack predictive capabilities, limiting their effectiveness in applications requiring anticipation of future system behavior.

In [38], to mitigate the effects of unbalanced grid voltages, an extended instantaneous power theory generates source current references, ensuring sinusoidal source and balanced output currents. An extended state observer (ESO) eliminates the need for grid voltage sensors by estimating grid voltages and providing delayed voltage information for current reference calculations. However, the ESO is sensitive to model inaccuracies, noise, and parameter variations, which can degrade performance. It also requires precise tuning and adds computational complexity, making real-time implementation challenging in some systems.

Due to the efficacy of the PCC method, this paper concentrates on the MC with the Model Predictive Current Control (MPCC) strategy. Moreover, although many works have studied the MPC of MCs, a comparison between the MC and AC-DC-AC converter with MPCC has not been presented. As a result, this paper investigates the performance of the MPCC strategy for the MC and AC-DC-AC converter. The remainder of this manuscript is structured along these lines: first, the mathematical model of the MC is explained. Then, the MPCC of the MC is described. After that, the simulation results are given for an MC with MPCC and an AC-DC-AC converter with MPCC. Finally, the conclusion is given.

Mathematical Model of an MC

Fig. 1 presents the topology of the three-phase MC. The MC consists of nine bidirectional power electronic switches that supply a three-phase R-L load. The MC is connected to the three-phase power grid using an L-C-R filter (as shown in Fig. 1). The switching condition of each bidirectional power electronic switch is represented by S_{xy}, where x is the source phase ($x \in \{u, v, w\}$) and y is the load phase ($y \in \{a, b, c\}$) (as shown in Fig. 1).

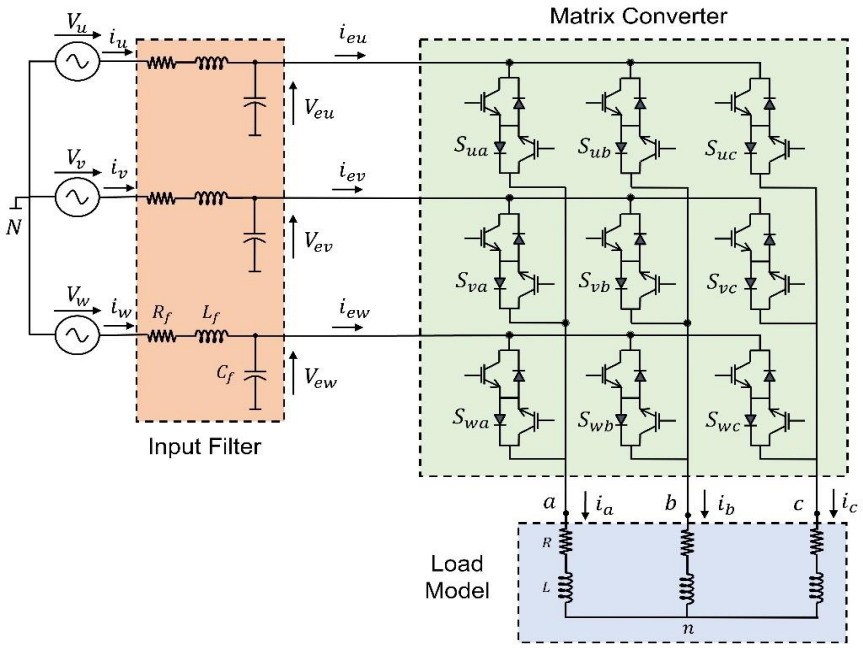


Fig. 1: Topology of a three-phase MC [19].

Note that in the operation of the MC, the short-circuiting of the power source must be avoided by the following restriction [19]:

$$S_{uy} + S_{vy} + S_{wy} = 1 \quad \forall \quad y \in \{a, b, c\} \tag{1}$$

Moreover, in the case of inductive loads, the load current must be continuous to avoid overvoltage in the power electronic elements.

The output voltage of the MC can be derived from the input voltage as follows [21], [22]:

$$\begin{bmatrix} v_{aN}(t) \\ v_{bN}(t) \\ v_{cN}(t) \end{bmatrix} = \underbrace{\begin{bmatrix} S_{ua} & S_{va} & S_{wa} \\ S_{ub} & S_{vb} & S_{wb} \\ S_{uc} & S_{vc} & S_{wc} \end{bmatrix}}_{\boldsymbol{T}} \cdot \begin{bmatrix} v_{eu}(t) \\ v_{ev}(t) \\ v_{ew}(t) \end{bmatrix} \tag{2}$$

or

$$\boldsymbol{v}_o(t) = \boldsymbol{T} \cdot \boldsymbol{v}_e(t) \tag{3}$$

in which $\boldsymbol{T}$ is the transformation matrix and $\boldsymbol{v}_o(k)$ and $\boldsymbol{v}_e(k)$ are the output voltage and input voltage of the MC, respectively:

$$\boldsymbol{v}_o(t) = \begin{bmatrix} v_{aN} \\ v_{bN} \\ v_{cN} \end{bmatrix} \quad and \quad \boldsymbol{v}_e = \begin{bmatrix} v_{eu}(t) \\ v_{ev}(t) \\ v_{ew}(t) \end{bmatrix} \tag{4}$$

Note that the above voltages are written relative to source mid-point N.

The phase voltage relative to the load neutral can be written as [37]:

$$\boldsymbol{v}_{on}(t) = \begin{bmatrix} v_{an} \\ v_{bn} \\ v_{cn} \end{bmatrix} = \begin{bmatrix} v_{aN} - v_{nN} \\ v_{bN} - v_{nN} \\ v_{cN} - v_{nN} \end{bmatrix} \tag{5}$$

v_{nN} can be extracted as [19]:

$$v_{nN} = \frac{v_{aN} + v_{bN} + v_{cN}}{3} \tag{6}$$

where *n* is the neutral point of the load.

The input current can be derived from the output current of the MC as [22]:

$$\begin{bmatrix} i_{eu}(t) \\ i_{ev}(t) \\ i_{ew}(t) \end{bmatrix} = \underbrace{\begin{bmatrix} S_{ua} & S_{ub} & S_{uc} \\ S_{va} & S_{vb} & S_{vc} \\ S_{wa} & S_{wb} & S_{wc} \end{bmatrix}}_{\boldsymbol{T}^T} \cdot \begin{bmatrix} i_a(t) \\ i_b(t) \\ i_c(t) \end{bmatrix} \tag{7}$$

or

$$\boldsymbol{i}_e(t) = \boldsymbol{T}^T \cdot \boldsymbol{i}_o(t) \tag{8}$$

$\boldsymbol{i}_e(k)$ and $\boldsymbol{i}_o(k)$ are the input current of the MC and load current, respectively:

$$\boldsymbol{i}_e(t) = \begin{bmatrix} i_{eu}(t) \\ i_{ev}(t) \\ i_{ew}(t) \end{bmatrix} \quad and \quad \boldsymbol{i}_o(t) = \begin{bmatrix} i_a(t) \\ i_b(t) \\ i_c(t) \end{bmatrix} \tag{9}$$

Using the circuit theory, the input filter state-space model can be derived as [21], [22]:

$$\dot{x}(t) = \underbrace{\begin{bmatrix} 0 & \frac{1}{C_f} \\ -\frac{1}{L_f} & -\frac{R_f}{L_f} \end{bmatrix}}_{\boldsymbol{A_c}} x(t) + \underbrace{\begin{bmatrix} 0 & -\frac{1}{C_f} \\ \frac{1}{L_f} & 0 \end{bmatrix}}_{\boldsymbol{B_c}} u(t) \tag{10}$$

R_f, L_f, and C_f are the filter resistor, inductor, and capacitor, respectively. The state and input vectors are as:

$$\boldsymbol{x}(t) = \begin{bmatrix} \boldsymbol{V}_e(t) \\ \boldsymbol{i}_s(t) \end{bmatrix} \quad and \quad \boldsymbol{u}(t) = \begin{bmatrix} \boldsymbol{V}_s(t) \\ \boldsymbol{i}_e(t) \end{bmatrix} \tag{11}$$

$\boldsymbol{V}_s(t)$ and $\boldsymbol{i}_s(t)$ are the voltage space vector and current space vector at the power supply side, respectively, and can be written as:

$$\boldsymbol{V}_s(t) = 2/3(v_u + \boldsymbol{a}v_v + \boldsymbol{a}^2 v_w) \tag{12}$$

$$\boldsymbol{i}_s(t) = 2/3(i_u + \boldsymbol{a}i_v + \boldsymbol{a}^2 i_w) \tag{13}$$

Using (10), the filter discrete model can be obtained as [39]:

$$\boldsymbol{i}_s(k+1) = \boldsymbol{A}_q(2,1)\,\boldsymbol{V}_e(k) + \boldsymbol{A}_q(2,2)\,\boldsymbol{i}_s(k) + \boldsymbol{B}_q(2,1)\,\boldsymbol{V}_s(k) + \boldsymbol{B}_q(2,2)\,\boldsymbol{i}_e(k) \tag{14}$$

with:

$$\boldsymbol{A}_q = e^{\boldsymbol{A}_c T_s} \quad and \quad \boldsymbol{B}_q = \int_0^{T_s} e^{\boldsymbol{A}_c (T_s - \tau)} \boldsymbol{B}_c d\tau \tag{15}$$

$\boldsymbol{A}_c$ and $\boldsymbol{B}_c$ matrixes can be calculated from (4). T_s is the sampling time.

MPCC of the MC

Fig. 2 illustrates the block diagram of the MPCC for an MC. For all 27 switching conditions of the MC, the load current must be calculated in the next sampling instant. The optimum switching condition for minimizing the objective function is chosen and utilized in the converter in the next sampling instant.

The resistive-inductive load discrete-time model can be written as [40], [41]:

$$\boldsymbol{i}_o(k+1) = \left(1 - \frac{RT_s}{L}\right)\boldsymbol{i}_o(k) + \frac{T_s}{L}(\boldsymbol{v}_{on}(k)) \tag{16}$$

where *k* is the sampling instant. *R* and *L* are the load resistor and inductor, respectively.

$\boldsymbol{v}_{on}(k)$ and $\boldsymbol{i}_o(k)$ are the load phase voltage and load current and can be calculated from (2)-(8).

The general cost function can be written as:

$$g(k+1) = g_i(k+1) + \lambda_Q\, g_Q(k+1) + \lambda_s\, n_{sw}(k+1) \tag{17}$$

where g_i, Q, and n_{sw} are the objective terms regarding the load current, reactive power, and number of switching commutations, respectively. λ_Q *and* λ_S are the weighting factors that adjust the reactive power, and switching commutations, respectively.

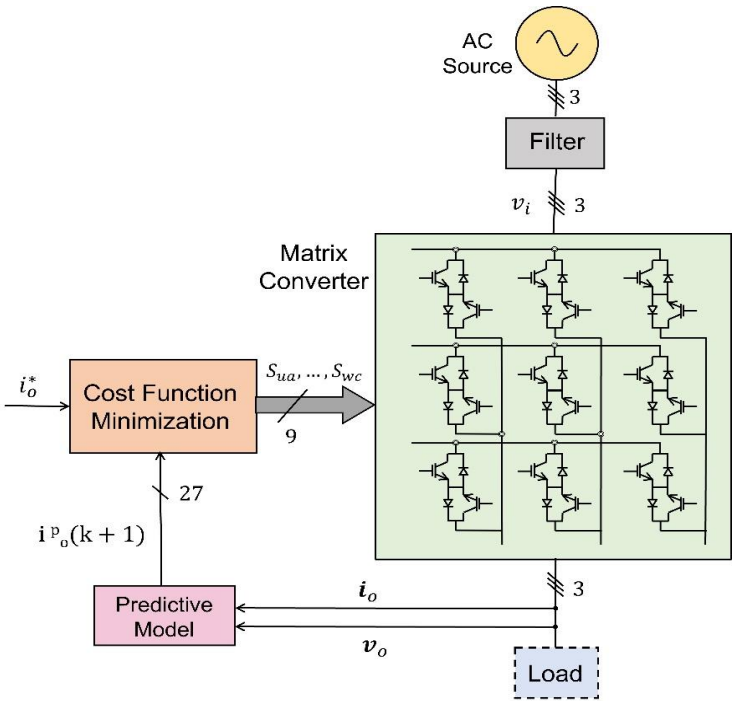


Fig. 2: Diagram of the MPCC for the MC [20].

g_i can be written as [40], [41]:

$$g_i = \left|i_{o\alpha}^*(k+1) - i_{o\alpha}^p(k+1)\right| + \left|i_{o\beta}^*(k+1) - i_{o\beta}^p(k+1)\right| \tag{18}$$

with

$$i_{o\alpha} = \frac{2}{3}\left(i_{oa} - \frac{1}{2}\, i_{ob} - \frac{1}{2}\, i_{oc}\right) \tag{19}$$

$$i_{o\beta} = \frac{2}{3}\left(\frac{\sqrt{3}}{2}\, i_{ob} - \frac{\sqrt{3}}{2}\, i_{oc}\right) \tag{20}$$

Moreover, g_Q can be written as [39]:

$$g_Q = |Q^* - Q(k+1)| \tag{21}$$

in which Q^* is the desired reactive power value. $Q(k+1)$ is the grid-side reactive power and can be computed as [37]:

$$\begin{aligned} Q(k+1) &= Im\{\boldsymbol{V}_s(k+1)\cdot \bar{\boldsymbol{\imath}}_s(k+1)\} \\ &= v_{s\beta}(k+1)\, i_{s\alpha}(k+1) \\ &\quad - v_{s\alpha}(k+1)\, i_{s\beta}(k+1) \end{aligned} \tag{22}$$

In addition, n_{sw} can be defined as [39]:

$$n_{sw}(k+2) = \sum_{x=u,v,w}\sum_{y=a,b,c}\left|S_{xy}(k+2) - S_{xy}(k+1)\right| \tag{23}$$

In Fig. 3 the flowchart of the MPCC for an MC is given.

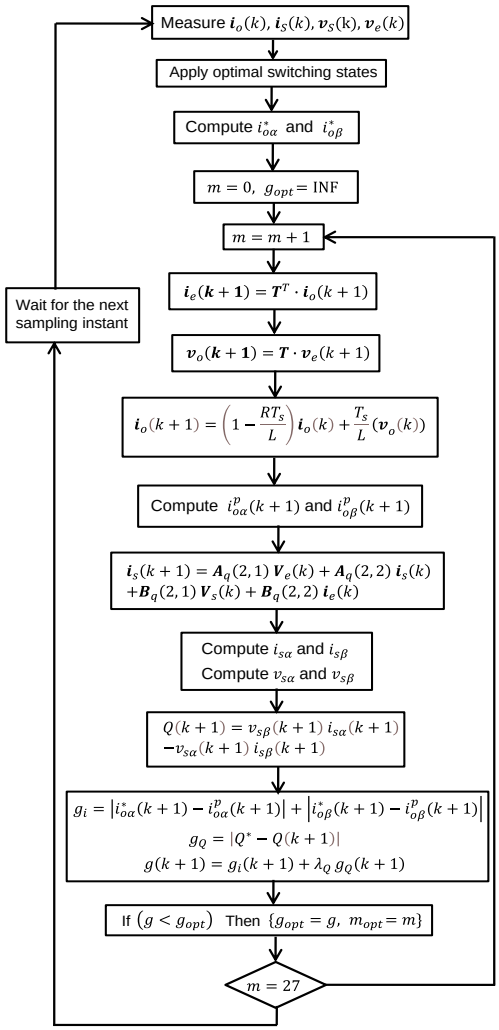


Fig. 3: Flowchart of the MPCC of the MC.

Results and Discussion

The proposed MPCC strategy has been verified by simulating an MC using MATLAB/Simulink. An AC source with a 180 V amplitude and a 50 Hz frequency has been used. The resistance, inductance, and capacitance of the input filter are 0.5 Ω, 400 µH, and 21 µF, respectively. A three-phase load with resistance of 10 Ω and inductance of 30 mH has been used. The sampling time T_s and simulation time-step T_{sim} are 20 µsec and 1 µsec, respectively.

Fig. 4 presents the simulation results of the MC with the MPCC. The weighting factors λ_Q and λ_s are set to zero. The reference currents amplitude is set to 8 A at t=0 sec and is changed from 8 A to 4 A at t=0.1 sec. Fig. 4(a) demonstrates the load current in the MC with the MPCC strategy. The converter has an excellent current tracking performance. Figs. 4(b)-(c) present the phase voltage (V_{an}), and line-to-line voltage (V_{ab}) of the MC with the MPCC method. Fig. 4(d) shows the current of the power supply i_u and the input current of the MC i_{eu}. It is visible that the quality of the power supply current is improved due to the existence of the L-C-R filter. Fig. 4(e) reveals the voltage and current at the power supply side. For λ_Q=0, the input power factor is not controlled. Fig. 4(f) demonstrates the instantaneous reactive power at the power supply side. The reactive power is not controlled.

However, the MC with the MPCC procedure can control the grid-side reactive power. The reactive power at the power supply side can easily be adjusted by increasing the weighting factor λ_Q in the objective function of MPCC.

Fig. 5 shows the simulation results of the MC with MPCC for λ_Q=0.008, indicating an increased weighting factor for the instantaneous grid-side reactive power control parameter in the MPC, compared to Fig. 4.

Fig. 6 presents the simulation results of the MC with MPCC for λ_Q=0.02, the maximum allowable weighting factor for the instantaneous grid-side reactive power control parameter in the MPC, while maintaining the proper current tracking. The parameters used in the simulation are like Fig. 4.

The simulation results demonstrate a significant reduction in the instantaneous grid-side reactive power between the three cases. In Figs. 5-6, where the weighting factor for reactive power control was increased, the system exhibits a marked improvement in minimizing reactive power flow to the grid. This reduction is evident in the smoother and more stable reactive power waveform, which is closer to zero compared to the first simulation. The enhanced control strategy effectively mitigates the fluctuations in reactive power, leading to better overall grid-side power quality and improved system efficiency.

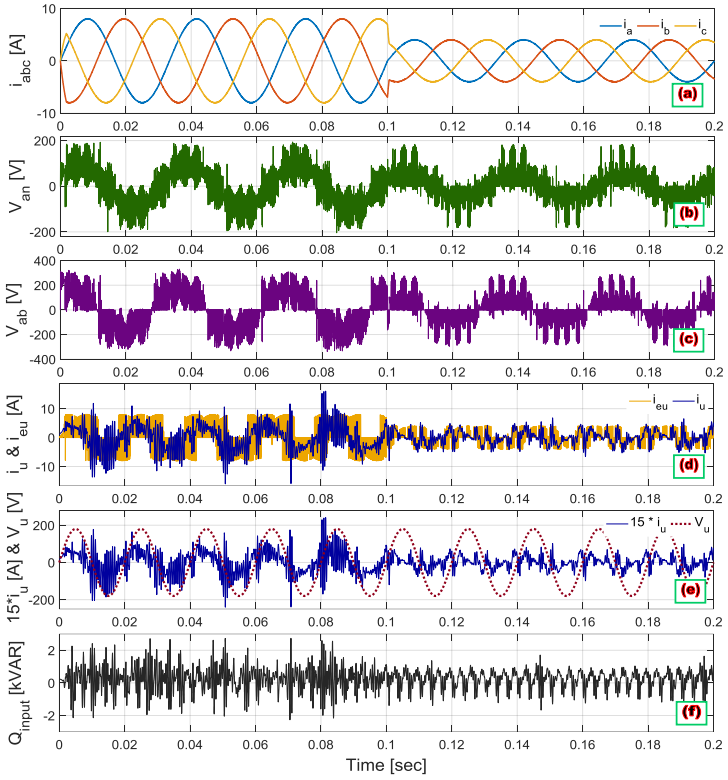


Fig. 4: Simulation results of the MC with MPCC for λ_Q=0: (a) load currents; (b) phase voltage V_{an}; (c) line voltage V_{ab}; (d) i_u and i_{eu}; (e) 15×i_u and V_u voltage; (f) reactive power.

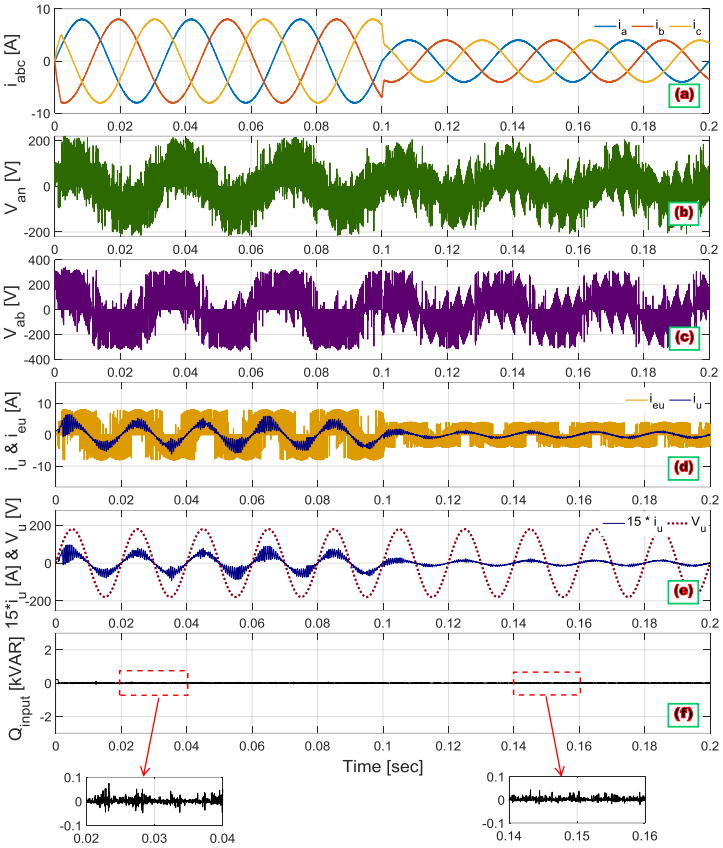


Fig. 5: Simulation results of the MC with MPCC for λ_Q=0.008: (a) load currents; (b) phase voltage V_{an}; (c) line voltage V_{ab}; (d) i_u and i_{eu}; (e) 15×i_u and V_u voltage; (f) reactive power.

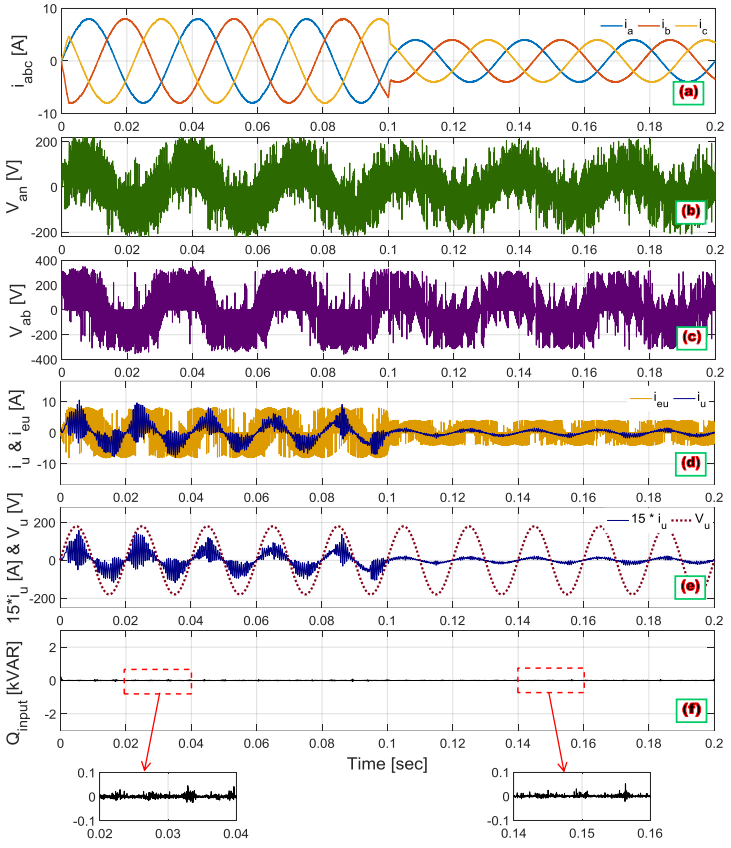


Fig. 6: Simulation results of the MC with MPCC for λ_Q=0.02: (a) load currents; (b) phase voltage V_{an}; (c) line voltage V_{ab}; (d) i_u and i_{eu}; (e) 15×i_u and V_u voltage; (f) reactive power.

In the next part, the behavior of the MC and the AC-DC-AC converter using the MPCC is compared. Fig. 7 illustrates the structure of the AC-DC-AC converter. The simulation parameters of the AC-DC-AC converter are the same as the MC. The DC-link capacitor is C=2 mF in the AC-DC-AC converter. Fig. 8 presents the simulation outcomes in the AC-DC-AC converter with MPCC.

Fig. 8(a) shows the load current in the AC-DC-AC converter with the MPCC method. Figs. 8(b)-(c) present the phase voltage (V_{an}), and line-to-line voltage (V_{ab}) of the AC-DC-AC converter with MPCC method. Fig. 8(d) represents the voltage and current at the power supply side. Fig. 8(e) illustrates the instantaneous power supply side reactive power.

The reactive power is high because the AC-DC-AC converter cannot control it. Moreover, it is visible that the phase and line voltage waveforms as well as the power supply side voltage and current waveforms are more distorted in the AC-DC-AC converter in comparison with the MC.

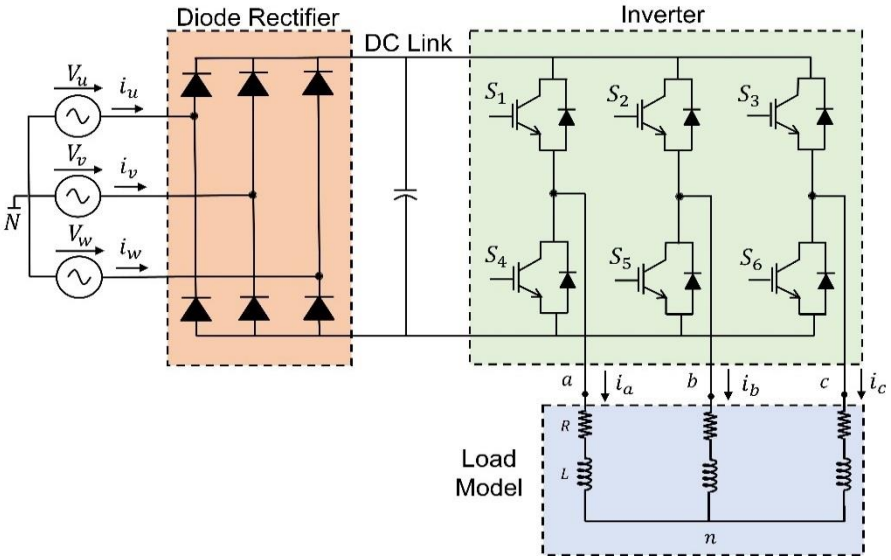


Fig. 7: Structure of the AC-DC-AC converter.

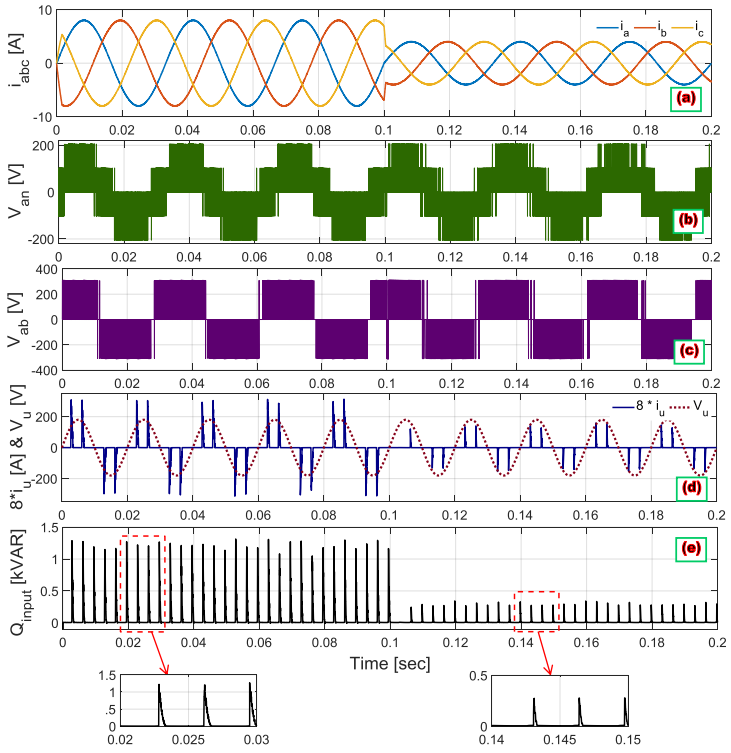


Fig. 8: Simulation results of the AC-DC-AC converter with MPCC: (a) load currents; (b) phase voltage V_{an}; (c) line voltage V_{ab}; (d) 8×i_u and V_u voltage; (e) reactive power.

Table 1 compares the load current THD in the AC-DC-AC converter and MC, both using MPCC. The weighting factor is λ_S=0. The current THD is lower in the MC with λ_Q=0. By increasing λ_Q, the grid-side reactive power of the MC is controlled. Consequently, the tracking performance and THD index of the load current deteriorate. If the reference current amplitude is set to 4 A, the current THD is reduced from 0.983% in the AC-DC-AC converter to 0.594% in the MC, which is a 39.58% reduction. If the reference current amplitude is set to 8 A, the current THD is reduced from 0.458% in the AC-DC-AC converter to 0.318% in the MC, which is a 30.57% decrease.

In the next part, an extra term related to the number of switching commutations is added to the objective function of the MPCC to mitigate the switching frequency of the MC. The system performance and the average switching frequency in the MC have been evaluated with various values of the weighting factor λ_S.

Fig. 9 illustrates the effect of two different values of λ_S on the performance of the MC. The weighting factor λ_Q is set to zero.

The weighting factor λ_S is changed from zero to 0.06 at t=0.05 sec. Figs. 9(a)-(b) show the load current and gate pulse S_{ua} in the MC with MPCC. As can be seen, increasing λ_s mitigates the average switching frequency of the MC from 11.132 kHz to 7.227 kHz. However, it is evident that in higher values of λ_s, there is more current distortion. Note that the switching frequency in Fig. 11 is determined based on the number of switching commutations in a period T for different switches. The average switching frequency of the MC can be calculated as:

$$f_{s-avg} = \frac{1}{9T}\sum_{x=u,v,w}\sum_{y=a,b,c} N_{xy} \quad (24)$$

where N_{xy} is the number of switching commutations for a specific switch, x is the source phase ($x \in \{u, v, w\}$) and y is the load phase ($y \in \{a, b, c\}$).

Fig. 10 presents the effect of two different values of λ_S on the performance of the AC-DC-AC converter. The simulation parameters are the same as Fig. 9. In the AC-DC-AC converter, by increasing λ_s from zero to 0.06, the average switching frequency reduces from 11.205 kHz to 8.39 kHz.

Table 1: Comparison of the load current THD [%] in AC-DC-AC converter and MC with MPCC for λ_S=0

		AC-DC-AC Converter	Matrix Converter		
			λ_Q=0	λ_Q=0.008	λ_Q=0.02
I_{ref} =4 A	Min	0.976	0.578	0.841	0.975
	Max	1.031	0.611	0.911	1.090
	Avg.	0.983	0.594	0.900	1.034
I_{ref} =8 A	Min	0.452	0.300	0.475	0.688
	Max	0.471	0.324	0.526	1.237
	Avg.	0.458	0.318	0.488	0.878

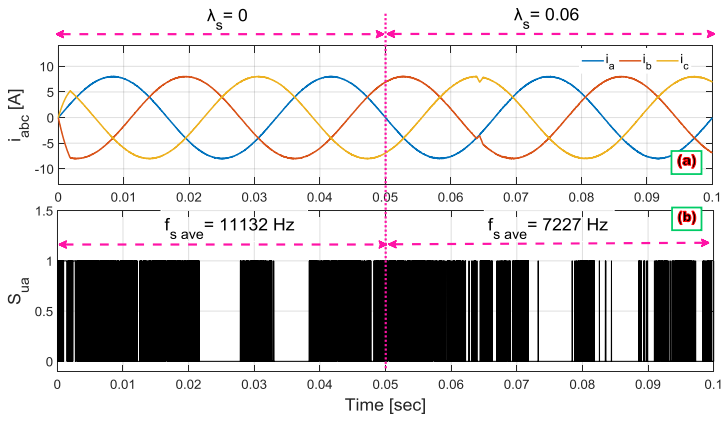


Fig. 9: Impact of λ_s on the performance of the MC: (a) load currents; (b) gate pulse S_{ua}.

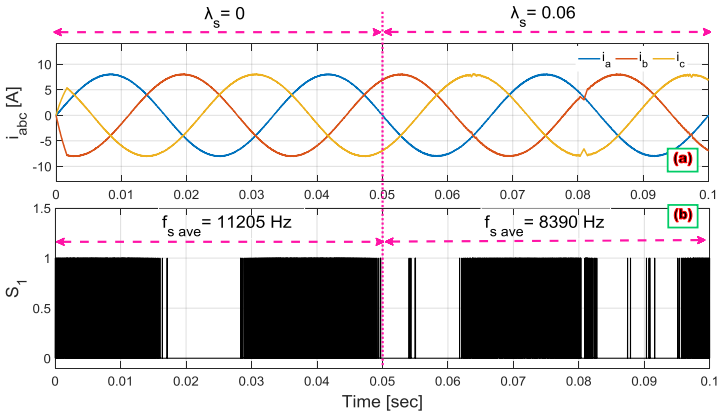


Fig. 10: Impact of λ_s on the behavior of the AC-DC-DC converter: (a) load currents; (b) gate pulse S_1.

Table 2 presents the load current THD of the AC-DC-AC converter and MC with MPCC for and λ_S=0.05.

Fig. 11 compares the current THD versus the average switching frequency in the matrix and AC-DC-AC converters with MPCC. It is visible that for a similar switching frequency, the current THD is lower in the MC compared to the AC-DC-AC converter and vice versa. As an example, at THD=0.48%, the average switching frequency is reduced from 10.4 kHz in the AC-DC-AC converter to 6.5 kHz in the MC. At the average switching frequency of 7 kHz, the THD is reduced from 1.01% in the AC-DC-AC converter to 0.435% in the MC.

Table 2: Comparison of the load current THD [%] in AC-DC-AC converter and MC with MPCC for λ_S=0.05

		AC-DC-AC Converter	Matrix Converter		
			λ_Q=0	λ_Q=0.008	λ_Q=0.02
I_{ref} =4 A	Min	1.246	0.872	1.015	1.133
	Max	1.299	3.082	1.175	1.352
	Avg.	1.277	1.111	1.056	1.187
I_{ref} =8 A	Min	0.589	0.398	0.576	0.800
	Max	0.627	0.455	0.708	1.352
	Avg.	0.617	0.419	0.661	1.187

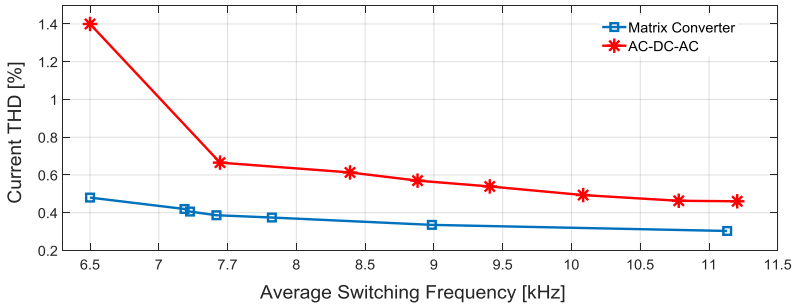


Fig. 11: Current THD versus average switching frequency in MC and AC-DC-AC.

Conclusion

This work has proposed an MPCC method for the MC with an RL load. The suggested method has succeeded in controlling the load current and power supply side reactive power, while other objectives were easily considered in the predictive controller. Additionally, this paper has evaluated the mitigation of the number of switching commutations in the prediction strategy. In this

regard, the prediction control with different weighting factors has been applied to the matrix and the AC-DC-AC converters. Finally, the results of the AC-DC-AC converter and MC were compared. Simulation results have revealed the excellent dynamic response and control of the reactive power at the power supply side in the MC using the MPC approach compared to the traditional AC-DC-AC converter. The AC-DC-AC converter is unable to control the reactive power at the power supply side, but the MC can almost completely mitigate it. The MC can decrease the switching frequency by 40% compared to the AC-DC-AC converter in the same current THD value, and the current THD by 58% compared to the AC-DC-AC converter in the same average switching frequency. However, mitigation of the reactive power and the switching frequency has the opposite effect on the current tracking quality in the MC.

The future work will concentrate on the MPC of motor drives fed by MCs including different objectives in the predictive control process.

Author Contributions

M. Nabizadeh carried out the simulations. Dr. Hamedani was the supervisor and Prof. Mirzaeian Dehkordi was the adviser of the current research paper. Dr. Hamedani wrote the manuscript. All authors revised and discussed the results and approved the final manuscript.

Acknowledgment

The author gratefully acknowledges the respected reviewers and the editor of JECEI for their helpful comments and accurate reviewing of this paper.

Conflict of Interest

The author declares no potential conflict of interest regarding the publication of this work.

Abbreviations

CM	Common-Mode
DTC	Direct Torque Control
MC	Matrix Converter
MPCC	Model Predictive Current Control
MPC	Model Predictive Control
PCC	Predictive Current Control

References

[1] L. Empringham, J. W. Kolar, J. Rodriguez et al., "Technological issues and industrial application of matrix converters: a review," IEEE Trans. Ind. Electron., 60(10): 4260-4271, 2013.

[2] J. Rodriguez, M. Rivera, J. W. Kolar, et al., "A review of control and modulation methods for matrix converters," IEEE Trans. Ind. Electron., 59(1): 58-70, 2012.

[3] S. Muller, U. Ammann, S. Rees, "New time-discrete modulation scheme for matrix converters," IEEE Trans. Ind. Electron., 52(6): 1607-1615, 2005.

[4] P. W. Wheeler, J. Rodriguez, J. C. Clare, L. Empringham, A. Weinstein, "Matrix converters: a technology review," IEEE Trans. Ind. Electron., 49(2): 276-288, 2002.

[5] K. B. Lee, F. Blaabjerg, "Improved sensorless vector control for induction motor drives fed by a matrix converter using nonlinear modeling and disturbance observer," IEEE Trans. Energy Convers., 21(1): 52-59, 2006.

[6] P. G. Potamianos, E. D. Mitronikas, A. N. Safacas, "Open-circuit fault diagnosis for matrix converter drives and remedial operation using carrier based modulation methods," IEEE Trans. Ind. Electron., 61(1): 531-545, 2014.

[7] T. D. Nguyen, H. H. Lee, "Generalized carrier-based PWM method for indirect matrix converters," in Proc. 2012 IEEE Third International Conference on Sustainable Energy Technologies (ICSET): 223-228, 2012.

[8] H. M. Nguyen, H. H. Lee, T. W. Chun, "Input power factor compensation algorithms using a new direct-SVM method for matrix converter," IEEE Trans. Ind. Electron., 58(1): 232-243, 2011.

[9] T. D. Nguyen, H. H. Lee, "A new SVM method for an indirect matrix converter with common-mode voltage reduction," IEEE Trans. Ind. Inf., 10(1): 61-72, 2014.

[10] D. Casadei, G. Serra, A. Tani, "The use of matrix converters in direct torque control of induction machines," IEEE Trans. Ind. Electron., 48(6): 1057-1064, 2002.

[11] H. H. Lee, H. M. Nguyen, T. W. Chun, W. H. Choi, "Implementation of direct torque control method using matrix converter fed induction motor," in Proc. IEEE International Forum on Strategic Technology: 51-55, 2007.

[12] H. Dan, P. Zeng, W. Xiong, M. Wen, M. Su, M. Rivera, "Model predictive control-based direct torque control for matrix converter-fed induction motor with reduced torque ripple," CES Trans. Electr. Mach. Syst., 5(2): 90-99, 2021.

[13] J. Monteiro, J. F. Silva, S. F. Pinto et al., "Matrix converter-based unified power-flow controllers: advanced direct power control method," IEEE Trans. Power Deliv., 26(1): 420-430, 2011.

[14] X. Wang, Q. Guan, L. Tian, "A novel adaptive fuzzy control for output voltage of matrix converter," in Proc. IEEE Power Electronics and Motion Control Conference (IPEMC): 53-58, 2012.

[15] C. F. Calvillo, F. Martell, J. L. Elizondo et al., "Rotor current fuzzy control of a DFIG with an indirect matrix converter," in Proc. IEEE Industrial Electronics Society (IECON): 4296-4301, 2011.

[16] T. S. Sivarani, S. J. Jawhar, C. A. Kumar, "Novel intelligent hybrid techniques for speed control of electric drives fed by matrix converter," in Proc. IEEE Computing, Electronics and Electrical Technologies (ICCEET): 466-471, 2012.

[17] M. Rivera, P. Wheeler, A. Olloqui, D. Khaburi, "A review of predictive control techniques for matrix converters—Part I," in Proc. IEEE Power Electronics and Drive Systems Technologies Conference (PEDSTC): 582-588, 2016.

[18] M. Rivera, P. Wheeler, A. Olloqui, "Predictive control in matrix converters—Part II: Control strategies, weaknesses and trends," in Proc. IEEE International Conference on Industrial Technology (ICIT): 1098-1104, 2016.

[19] S. Toledo, D. Caballero, E. Maqueda, J. J. Cáceres, M. Rivera, R. Gregor, P. Wheeler, "Predictive control applied to matrix converters: A systematic literature review," Energies, 15(20): 7801, 2022.

[20] M. Khosravi, M. Amirbande, D.A. Khaburi, M. Rivera, J. Riveros, J. Rodriguez, A. Vahedi, P. Wheeler, "Review of model predictive control strategies for matrix converters", IET power Electron., 12(12): 3021-3032, 2019.

[21] R. Vargas, J. Rodriguez, C. A. Rojas, M. Rivera, "Predictive control of an induction machine fed by a matrix converter with increased efficiency and reduced common-mode voltage," IEEE Trans. Energy Convers., 29(2): 473-485, 2014.

[22] R. Vargas, J. Rodriguez, U. Ammann, P. W. Wheeler, "Predictive current control of an induction machine fed by a matrix converter

with reactive power control," IEEE Trans. Ind. Electron., 55(12): 4362-4371, 2008.

[23] L. Tarisciotti et al., "Modulated predictive control for indirect matrix converter," IEEE Trans. Ind. Appl., 53(5): 4644-4654, 2017.

[24] Y. Liu, Y. Liu, B. Ge, H. Abu-Rub, "Interactive grid interfacing system by matrix-converter-based solid state transformer with model predictive control," IEEE Trans. Ind. Inf., 16(4): 2533-2541, 2020.

[25] J. Rodriguez, R. Pontt, R. Vargas et al., "Predictive direct torque control of an induction motor fed by a matrix converter," in Proc. IEEE European Conference on Power Electronics and Applications: 1-10, 2007.

[26] M. López, J. Rodriguez, C. Silva, M. Rivera, "Predictive torque control of a multidrive system fed by a dual indirect matrix converter," IEEE Trans. Ind. Electron., 62(5): 2731-2741, 2015.

[27] M. Uddin, S. Mekhilef, M. Mubin, M. Rivera, J. Rodriguez, "Model predictive torque ripple reduction with weighting factor optimization fed by an indirect matrix converter," Electr. Power Compon. Syst., 42(10): 1059-1069, 2014.

[28] S. Toledo et al., "Active and reactive power control based on an inner predictive voltage control loop for AC generation systems with direct matrix converter," in Proc. IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC): 1-6, 2019.

[29] S. Toledo et al., "Active and reactive power control based on predictive voltage control in a six-phase generation system using modular matrix converters," in Proc. IEEE International Conference on Industrial Technology (ICIT): 1059-1065, 2020.

[30] M. Ortega, F. Jurado, J. Carpio, "Control of indirect matrix converter with bidirectional output stage for micro-turbine," IET Power Electron., 5(6): 659-668, 2012.

[31] S. Yusoff, L. De Lillo, P. Zanchetta, P. Wheeler, "Predictive control of a direct AC/AC matrix converter power supply under non-linear load conditions," in Proc. IEEE International Power Electronics and Motion Control Conference (EPE/PEMC): DS3c.4-1-DS3c.4-6, 2012.

[32] M. Rivera, J. Rodriguez, J. R. Espinoza, H. Abu-Rub, "Instantaneous reactive power minimization and current control for an indirect matrix converter under a distorted AC supply," IEEE Trans. Ind. Inf., 8(3): 482-490, 2012.

[33] C. F. Garcia, M. E. Rivera, J. R. Rodríguez, P. W. Wheeler, R. S. Peña, "Predictive current control with instantaneous reactive power minimization for a four-leg indirect matrix converter," IEEE Trans. Ind. Electron., 64(2): 922-929, 2017.

[34] M. Roostaee, B. Eskandari, M. R. Azizi, "Predictive current control with modification of instantaneous reactive power minimization for direct matrix converter," in Proc. IEEE Power Electronics, Drives Systems and Technologies Conference (PEDSTC): 199-205, 2018.

[35] M. Rivera, J. Rodriguez, J. Espinoza, B. Wu, "Reduction of common-mode voltage in an indirect matrix converter with imposed sinusoidal input/output waveforms," in Proc. IEEE Industrial Electronics Society Conference (IECON 2012): 6105-6110, 2012.

[36] R. Vargas, U. Ammann and J. Rodriguez, "Predictive Approach to Increase Efficiency and Reduce Switching Losses on Matrix Converters," IEEE Transactions on Power Electronics, 24(4): 894-902, 2009.

[37] F. Villarroel, J. Espinoza, C. Rojas, C.; Molina, E. A. Espinosa, "multiobjective ranking based finite states model predictive control scheme applied to a direct matrix converter," in Proc. IECON Proceedings (Industrial Electronics Conference): 2941-2946, 2010.

[38] W. Xiong, Y. Sun, J. Lin, M. Su, H. Dan, M. Rivera, J. M. Guerrero, "A cost-effective and low-complexity predictive control for matrix converters under unbalanced grid voltage conditions," IEEE Access, 7: 43895-43905, 2019.

[39] J. Rodriguez, P. Cortes, Predictive control of power converters and electrical drives, John Wiley & Sons, 2012.

[40] P. Hamedani, M. Changizian, "A new hybrid predictive-pwm control for flying capacitor multilevel inverter," J. Electr. Comput. Eng. Innovations (JECEI), 12(2): 353-362, 2024.

[41] P. Hamedani, "Multistep model predictive control of diode-clamped multilevel inverter," J. Electr. Comput. Eng. Innovations (JECEI), 13(1): 117-128, 2025.

Biography

Matin Nabizadeh was born in Kerman, Iran, in 2000. He received the B.Sc. degree in Electrical Engineering from University of Isfahan, Iran, in 2022. He is currently a M.Sc. student in Electrical Engineering in University of Isfahan. His research interests include power electronics and motor drives, intelligent systems, and renewable energy

- Email: m.nabizadeh@eng.ui.ac.ir
- ORCID: 0009-0009-8641-7336
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Pegah Hamedani was born in Isfahan, Iran, in 1985. She received B.Sc. and M.Sc. degrees from University of Isfahan, Iran, in 2007 and 2009, respectively, and the Ph.D. degree from Iran University of Science and Technology, Tehran, in 2016, all in Electrical Engineering. Her research interests include power electronics, control of electrical motor drives, supply system of the electric railway (AC and DC), linear motors & MAGLEVs, and analysis of overhead contact systems. She is currently an Assistant Professor with the Department of Railway Engineering and Transportation Planning, University of Isfahan, Isfahan, Iran. Dr. Hamedani was the recipient of the IEEE 11th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC'20) best paper award in 2020.

- Email: p.hamedani@eng.ui.ac.ir
- ORCID: 0000-0002-5456-1255
- Web of Science Researcher ID: AAN-2662-2021
- Scopus Author ID: 37118674000
- Homepage: https://engold.ui.ac.ir/~p.hamedani/

Behzad Mirzaeian Dehkordi was born in Shahrekord, Iran, in 1966. He received the B.Sc. degree in Electronics Engineering from Shiraz University, Iran, in 1985, and the M.Sc. and Ph.D. degrees in Power Engineering from the Isfahan University of Technology (IUT), in 1994 and 2000, respectively. In September 2002, he joined the Department of Electrical Engineering, University of Isfahan, where he is currently a Professor of Electrical Engineering. He was a Visiting Professor with the Power Electronic Laboratory, Seoul National University (SNU), South Korea, from March 2008 to August 2016. His research interests include power electronics and motor drives, intelligent systems, and renewable energy

- Email: mirzaeian@eng.ui.ac.ir
- ORCID: 0000-0002-1124-8138
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://engold.ui.ac.ir/~mirzaeian/

How to cite this paper:
M. Nabizadeh, P. Hamedani, B. Mirzaeian Dehkordi, "A comparative evaluation of model predictive current controlled matrix converter versus AC-DC-AC converter," J. Electr. Comput. Eng. Innovations, 13(2): 431-442, 2025.

DOI: 10.22061/jecei.2025.11435.804

URL: https://jecei.sru.ac.ir/article_2304.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Hybrid Three-Layered Approach for Intrusion Detection Using Machine Learning Methods

A. Beigi *

School of Computer Engineering, Shadid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 08 December 2024
Reviewed 05 February 2025
Revised 18 February 2025
Accepted 09 March 2025

Keywords:
Intrusion detection systems
Network security
Machine learning
NSL-KDD data set

*Corresponding Author's Email Address: *akrambeigi@sru.ac.ir*

Abstract

Background and Objectives: Intrusion Detection Systems (IDS) are crucial for safeguarding computer networks. However, they face challenges such as detecting subtle intrusions and novel attack patterns. While signature-based and anomaly-based IDS have been widely used, hybrid approaches offer a promising solution by combining their strengths. This study aims to develop a robust hybrid IDS that effectively addresses these challenges.

Methods: We propose a three-layered hybrid IDS that leverages machine learning techniques. The first layer utilizes a signature-based approach to identify known intrusions. The second layer employs an anomaly-based approach with unsupervised learning to detect unknown intrusions. The third layer utilizes supervised learning to classify intrusions based on training data. We evaluated the proposed system on the NSL-KDD dataset.

Results: Experimental results demonstrate the effectiveness of our proposed hybrid IDS in accurately detecting intrusions. Comparisons with recent studies using the same dataset show that our system outperforms existing approaches in terms of detection accuracy and robustness.

Conclusion: Our research presents a novel hybrid IDS that effectively addresses the limitations of traditional IDS methods. By combining signature-based, anomaly-based, and supervised learning techniques, our system can accurately detect both known and unknown intrusions. The promising results obtained from our experiments highlight the potential of this approach in enhancing network security.

Introduction

The widespread adoption of the internet has led to a significant increase in data exchange between diverse devices. Ensuring secure communication among these devices is paramount, making network security a critical research area in today's interconnected world. Intrusion Detection Systems (IDSs) play a vital role in bolstering network security, often employed in conjunction with other protective measures such as firewalls and access control mechanisms [1]. Intrusion Detection Systems are designed to safeguard critical resources by actively monitoring network traffic and system events for any signs of unauthorized access. Intruders employ a diverse range of tactics, including [2]:

- Denial of Service (DoS) attacks: These attacks aim to overwhelm a system with excessive traffic, making it unavailable to legitimate users.
- Remote-to-Local (R2L) attacks: These attacks exploit vulnerabilities to gain unauthorized access to a system from a remote location.
- User-to-Root (U2R) attacks: These attacks enable attackers to gain root-level privileges on a system by exploiting vulnerabilities in user accounts.
- Probing attacks: These attacks involve scanning

systems for vulnerabilities that can be later exploited for malicious purposes.

Furthermore, intruders constantly develop new and sophisticated techniques to breach system defenses, making it crucial for IDSs to adapt and evolve.

The performance of an IDS is analyzed by creating a specialized dataset comprising network traffic features to capture and learn attack patterns. Intrusion detection is framed as a classification problem, where various Machine Learning and Data Mining techniques are applied to categorize network data into normal and malicious traffic. The dataset includes both normal and anomalous network traffic, providing the classifier with sufficient examples to identify and differentiate between these patterns effectively [3].

However, some intrusion samples are almost identical to normal samples, leading to false positives where IDS mistakenly classify normal traffic as attacks [4]. To address this issue, researchers often incorporate signature-based methods into their system design. The signature-based method relies on a database of known attack signatures, comparing these signatures with incoming samples to identify intrusions [5].

Another significant challenge in designing an IDS is detecting intrusions that have no prior examples in the training data. To tackle this, researchers employ anomaly-based methods. This approach involves detecting abnormal patterns by monitoring data for deviations from expected behavior. Any significant deviation is flagged as an intrusion [6]. Algorithms utilizing anomaly-based methods can be further divided into supervised and unsupervised learning techniques [7]. Supervised learning algorithms require labeled training data, while unsupervised learning algorithms do not, allowing them to identify new types of intrusions without prior examples.

These challenges can be simultaneously addressed using hybrid approaches that combine signature-based and anomaly-based methods. To maximize the effectiveness of hybrid IDS, attention should be paid to two critical points:

Comprehensive Detection: Hybrid systems should be capable of detecting all types of attacks, including those with training samples (known attacks), those without training samples (unknown attacks), and those that closely resemble normal samples.

Sequence of Methods: The order in which signature-based and anomaly-based methods are applied is crucial and can significantly affect the system's accuracy. For instance, if an anomaly-based algorithm is applied first, it might fail to detect intrusions that are very similar to normal traffic, leading to false negatives. Conversely, applying a signature-based method first can help filter out known attacks, allowing the anomaly-based method to focus on detecting unknown and subtle deviations.

While implementing hybrid methods increases temporal complexity, this approach effectively addresses the outlined challenges and enhances overall system performance.

This study proposes a hybrid intrusion detection system with three detection layers to address the mentioned challenges effectively. Machine learning methods are applied throughout the system. The first detection layer employs heuristic rules, a signature-based approach, to tackle the challenge of intrusions that closely resemble normal samples. The second layer uses a genetic algorithm-based clustering method, an anomaly-based approach, to detect unknown attacks. In the third detection layer, a Back Propagation Neural Network (BPNN) classification is utilized to distinguish known attacks from normal samples [8].

The innovation of the proposed IDS lies in the integration of both supervised and unsupervised learning algorithms alongside a signature-based approach, enhancing its capability to detect a wide range of intrusions. Experiments conducted on the NSL-KDD dataset demonstrate the efficiency of the proposed IDS, showing superior performance compared to other methods using the same dataset.

The remainder of this paper is organized as follows: First, we provide a brief review of recent hybrid approaches for IDS similar to our research. Next, we describe the proposed method, outlining the main idea and the proposed algorithms. Then, we introduce the dataset used and present experimental results demonstrating the efficiency of our method. Finally, we conclude the paper.

Related Studies

Many researchers have developed intrusion detection systems using signature-based, anomaly-based, or hybrid approaches, each aiming to improve the efficiency of their system by leveraging the strengths of these methods. This section reviews some recent studies in this area.

In [7], scholars presented a fuzzy semi-supervised learning (FSSL) method. This approach categorizes an unlabeled training dataset into three fuzzy groups: low, medium, and high, using a neural network classifier. The "low" and "high" fuzzy groups are then combined with a labeled training dataset to form a new training dataset, which is subsequently used to train a neural network. This method enhances the classification efficiency in IDS. The use of a semi-supervised learning method makes it suitable for detecting unknown intrusions; however, it is less effective in detecting attacks that are very similar to normal samples.

In [9], researchers proposed a two-layered hybrid approach. The first layer employs a novel intrusion detection method based on changing cluster centers of samples. The second layer uses the K-Nearest Neighbors

(KNN) algorithm to implement two different detection modules for anomaly and signature-based detection. The second layer's modules reduce the rates of false positives and false negatives from the first detection module. Their experiments show that this method effectively identifies both known and unknown attacks with a high detection rate and low false positive rate, although it struggles with detecting intrusions similar to normal samples.

In [10], a hybrid approach called Hierarchical Filtration of Anomalies (HFA) was introduced. This method first uses a decision tree to separate normal and attack samples. Next, the detected normal samples are re-evaluated by a random forest algorithm, while the detected attack samples are rechecked by the KNN algorithm. Samples classified as normal by KNN are sent back to the random forest for final consideration. This method is effective for detecting known attacks but not for unknown intrusions due to its reliance on supervised algorithms.

Authors in [11] introduced a hybrid IDS based on Artificial Bee Colony (ABC) and Artificial Fish Swarm (AFS) algorithms. Their method preprocesses the training data and divides them into clusters using Fuzzy C-Means Clustering. An initial population is generated for each training subset, and irrelevant features are omitted using Correlation-based Feature Selection (CFS). Trustworthy and straightforward rules are then produced to detect normal and abnormal activities for each subset, which are combined to create the final rules. The ABC-AFS hybrid algorithm is trained on these rules, and a test dataset is used to evaluate system efficiency. However, this method's effectiveness depends heavily on the training dataset, making it less effective for detecting attacks without training samples.

In [12], a multi-level hybrid IDS using Support Vector Machine (SVM) and extreme learning machine was proposed. The method applies the K-means algorithm to create a modified training dataset and employs an anomaly-based approach. Their results indicate limited success in detecting attacks that closely resemble normal samples. Another study, [13], proposed a deep learning-based IDS using Recurrent Neural Networks (RNN-IDS). RNNs can retain previous information and apply it to the current output, making them effective for supervised classification. This method outperforms traditional learning methods like decision trees, SVM, and neural networks but lacks a solution for detecting unknown intrusions.

In [14], a fuzziness-based semi-supervised learning approach via ensemble learning (FSSL-EL) was applied in a framework for IDS. This framework learns from labeled data and analyzes unlabeled and noisy data using a fuzziness-based method. The results of the supervised and unsupervised parts are then combined via an ensemble system. A recently published study, [15], proposed an IDS (TSE-IDS) based on hybrid feature selection and two-level classifier ensembles. This method uses particle swarm optimization, the ant colony algorithm, and a genetic algorithm in the feature selection phase, and rotation forest and bagging in the classifier ensembles stage. The intrusion detection method is anomaly-based, but no specific solution is provided for similar normal intrusions.

In [16], a Deep Convolutional Neural Network (DCNN) model was used for anomaly detection. The model includes an input layer, three convolution and sub-sampling pairs, three fully connected layers, and an output layer with a single sigmoid unit. This research explores the suitability of deep learning approaches for IDS but does not address detecting intrusions similar to normal samples. Another research combines multiple learners to create an ensemble learner called Decision Tree Bagging Ensemble (DTBE) [17]. DTBE first identifies anomalies and then classifies attacks. This method also lacks a specified solution for detecting similar normal intrusions.

In ICVAE-DNN [18], an intrusion detection model is proposed by integrating an Improved Conditional Variational Autoencoder (ICVAE) with a Deep Neural Network (DNN). This model aims to learn and uncover sparse representations between network data features and their corresponding classes. The trained ICVAE decoder generates new attack samples based on specified intrusion categories, effectively balancing the training dataset and enhancing the diversity of training samples. This process improves the detection rate of unbalanced attacks.

In a recent work [19], a hybrid intrusion detection system is proposed. This system leverages the CFS-DE feature selection algorithm for dimensionality reduction and selects an optimal subset of features for improved classification. This system achieves good accuracy by leveraging a diverse set of feature categories during the selection process. It utilizes the CFS-DE algorithm for feature selection.

In [20], an intrusion detection system called IGAN-IDS has been developed to address the challenge of unbalanced class intrusion detection using samples generated by IGAN (Improved Generative Adversarial Network). The system is composed of three main modules: feature extraction, IGAN, and a deep neural network (DNN). Initially, a feed-forward neural network is employed to transform raw network features into feature vectors, which serve as the input for the IGAN module. IGAN then generates synthetic samples to balance the dataset. Finally, a DNN, incorporating convolutional layers and fully connected layers, is utilized for the final intrusion detection, leveraging both the original and the IGAN-generated samples to improve detection accuracy and robustness against class imbalance.

The I-SiamIDS is a two-layer IDS designed for detecting intrusions in unbalanced datasets. This model addresses class imbalance by identifying majority and minority classes at the algorithm level [21]. The first layer employs an ensemble of binary extreme Gradient Boosting, Siamese Neural Network and Deep Neural Network to hierarchically filter input samples and detect potential attacks. Detected attacks are then passed to the second layer, where a multi-class extreme Gradient Boosting classifier further classifies these attacks into specific categories.

CNN-BiLSTM [22] introduces an IDS designed to address class imbalance by combining One-Side Selection with the Synthetic Minority Over-sampling Technique. This hybrid sampling approach enhances model performance by reducing noise in the majority class and increasing the representation of minority class samples. The proposed model leverages Convolutional Neural Networks for spatial feature extraction and Bidirectional Long Short-Term Memory networks for temporal feature extraction. Experimental results on the KDDTest+ dataset demonstrate an accuracy of 83.58%.

Research [23] introduces an intrusion detection method that addresses class imbalance by employing the Adaptive Synthetic Sampling algorithm. To enhance feature extraction and mitigate the impact of redundant information, the model incorporates an improved Convolutional Neural Network based on the Split Convolution Module. This combined approach, termed AS-CNN, is evaluated using the standard NSL-KDD dataset to assess its effectiveness in detecting network intrusions.

BAT-MC [24] presents a two-stage deep learning anomaly detection model that integrates Bidirectional Long Short-Term Memory (BLSTM) networks with an attention mechanism. The attention mechanism effectively prioritizes critical network flow vectors generated by the BLSTM, enabling the model to focus on essential features for precise classification. BAT-MC incorporates convolutional layers alongside BLSTM to capture both spatial and temporal characteristics in network traffic. This end-to-end approach eliminates the need for manual feature engineering, allowing the model to automatically learn hierarchical features. Experimental results on the KDDTest+ dataset demonstrate an accuracy of 84.25%.

Another study [25] proposed a multi-layered approach utilizing k-nearest neighbors (KNN), hyper-learning machines, and hierarchical hyper-learning machines. This work explores the application of Software-Defined Networking (SDN) to mitigate the false positive rate in DoS attack detection systems. The proposed approach combines flow-based and packet-based analysis techniques. Experimental evaluation on the NSL-KDD dataset achieved accuracy of 84.29% on the KDDTest+ subset.

In [26], an Intrusion Detection System (IDS) named DSN was proposed, utilizing a Deep Stacking Network that integrates multiple base classifiers, including decision trees, k-nearest neighbors, deep neural networks, and random forests. The study focuses on the NSL-KDD dataset, with experimental results demonstrating an accuracy of 86.8%.

Recent research [27] explores network intrusion detection by evaluating the performance of various classification techniques, including Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbor, on the NSL-KDD dataset. Adopting the CRISP-DM methodology, the study aims to identify and analyze anomalous patterns in network traffic. Among the approaches, the Decision Tree demonstrated the highest performance, achieving complete accuracy on the KDDTrain+ dataset and 80% accuracy on the KDDTest+ dataset.

[28], a recent study, investigates the effectiveness of various shallow machine learning algorithms, including Random Forest, Decision Tree, Gaussian Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting, AdaBoost, and Linear Discriminant Analysis, for intrusion detection. This research leverages the NSL-KDD dataset and applies feature selection techniques, such as SelectKBest and Correlation Feature Selection, to enhance model performance. Among the evaluated algorithms, Gradient Boosting achieved the highest accuracy, with 86% in binary classification tasks.

As mentioned earlier, it is crucial that the algorithms are effective in real-world applications and exhibit high performance [17]. A key limitation of many IDSs is their difficulty in detecting unknown attacks and those that closely resemble normal traffic. These limitations significantly reduce both the efficiency and accuracy of such systems. Our proposed method addresses these challenges to improve detection accuracy, and is therefore evaluated using the widely recognized and realistic NSL-KDD dataset. Our system uniquely integrates signature-based detection, genetic algorithm-based clustering, and backpropagation neural networks within a novel three-layered architecture. Experimental results demonstrate that this approach achieves higher detection accuracy compared to other methods evaluated on the same dataset.

Proposed Algorithm

The proposed algorithm comprises two main processes: (1) Clustering and Classification Learning Process Using the Training Dataset, (2) Diagnosis and Separation of Attacks from Normal Samples in the Experimental Dataset. Fig. 1 illustrates the mechanism of the proposed intrusion detection system.

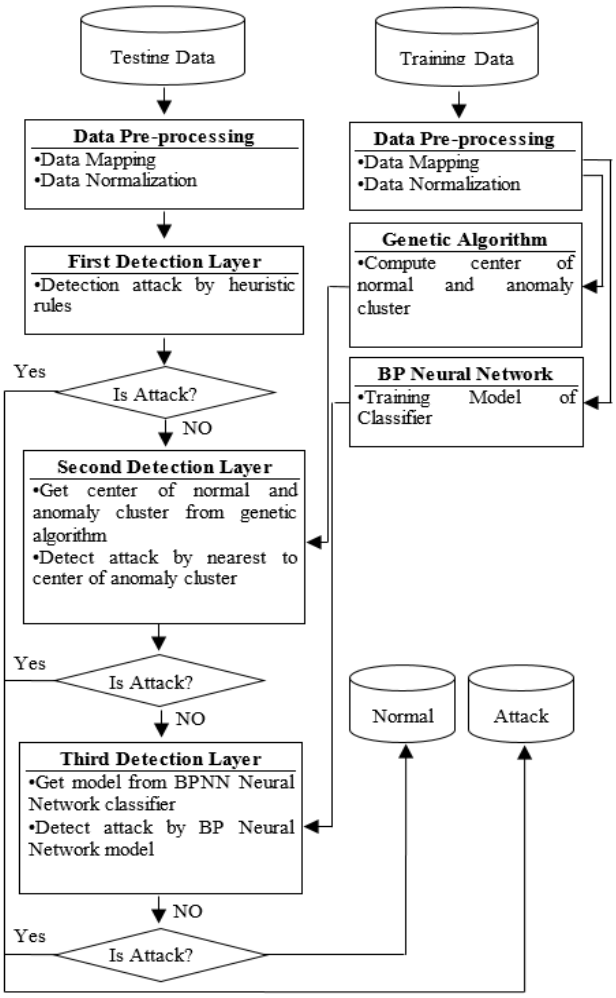


Fig. 1: Proposed intrusion detection system.

The algorithm begins by performing mapping and pre-processing operations on the training and testing datasets. The mapping algorithm involves two processes:

1. Converting discrete feature data to continuous values.
2. Normalizing all values to numbers between zero and one.

For more details on mapping and pre-processing operations, refer to [12]. Next, the training set data is sent to both the genetic-based clustering algorithm and the Back Propagation Neural Network (BPNN) classification algorithm. The genetic-based clustering algorithm divides the data into abnormal and normal subsets and sends these cluster centers to the second diagnostic layer. The BPNN classification algorithm is also trained using this data, and the resulting model is sent to the third diagnostic layer.

In the subsequent phase, each sample in the testing set is processed through all three detection layers for labeling. These three layers utilize detection methods based on heuristic rules, genetic-based clustering, and BPNN classification. Each layer examines the testing samples, and if an anomaly is detected, the sample is labeled as an attack. If none of the layers detect the sample as abnormal, it is labeled as normal. This process continues until all the testing data are labeled. The three detection layers are described as follows:

First Detection Layer: Heuristic Rules

This layer uses a signature-based detection approach with heuristic rules to identify intrusions that closely resemble normal samples.

Second Detection Layer: Genetic-Based Clustering

This layer employs a genetic algorithm-based clustering method to detect unknown attacks by identifying anomalies.

Third Detection Layer: BPNN Classification

This layer uses a Back Propagation Neural Network (BPNN) to classify known attacks and normal samples, providing a final decision on the sample's status.

By integrating these layers, the proposed algorithm addresses the challenges of detecting both known and unknown attacks, as well as intrusions that closely resemble normal behavior. This layered approach enhances the overall accuracy and efficiency of the intrusion detection system.

A. First Detection Layer Based on Heuristic Laws

Many attacks of the Remote to Local (R2L) type exhibit significant similarities with normal samples, making them difficult for classifiers to accurately diagnose. To address this challenge, we implement a detection layer based on heuristic rules. In an intrusion detection system, the expertise of a professional on signature attacks can be translated into heuristic rules in the form of "if-then" statements [4]. By examining the behavior and signature of an attack, we can formulate these rules.

For instance, heuristic rules can effectively detect Guess-password and Warezmaster attacks, which are types of R2L attacks [29]. In Guess-password attacks, the intruder repeatedly attempts different passwords to gain access. If a user fails multiple login attempts in a network event and either succeeds or fails in logging in, such patterns can indicate an attack. Therefore, in datasets where the number of failed logins is recorded in the num-failed-logins attribute and the login status in the logged-in attribute, we can create a heuristic rule based on these attributes.

Warezmaster attacks occur when a file server mistakenly grants write permissions to guest users. During such an attack, an intruder accesses the server using a guest account, creates a hidden folder, and uploads files that can later be downloaded by other users [30]. Heuristic rules can be formulated to detect such activity. The algorithm for this detection layer is illustrated in Fig. 2.

This detection layer uses two heuristic rules:

Heuristic Rule 1: If the number of login failures exceeds one and the user fails to log in, the input sample is detected as an attack.

Heuristic Rule 2: If the protocol is TCP and the service type is FTP or FTP-DATA, the input sample is detected as an attack if either of the following conditions is true:

- The connection time length and the number of bytes sent are greater than zero, and the number of bytes received is zero.
- The guest user has created one or more folders or files in the source.

```
Algorithm. First Detection Layer - heuristic rules
Input:
        Ts_i : Instance Data of Testing Dataset
Output:
        isattack : 0 is not attack and 1 is attack
-----------------------------------------------------------
1:  //rule 1: guess_passwd detection
2:  if (num_failed_logins >0) AND (is_guest_login==0)
3:      isattack=1;
4:      Return;
5:  //rule 2: warezmaster detection
6:  if ((protocol_type == tcp) AND ((service ==ftp)
        OR  (service ==ftp_data )))
7:      if (((duration > 0) AND  (src_bytes >0)
        AND (dst_bytes ==0))
        OR ((hot >0 AND) (is_guest_login ==1)))
8:          isattack=1;
9:          Return;
```

Fig. 2: The algorithm of the first detection layer based on heuristic rules.

These heuristic rules leverage specific attributes and behaviors to distinguish between normal and malicious activities effectively. This layer's ability to detect subtle patterns characteristic of R2L attacks enhances the overall accuracy of the intrusion detection system.

B. Second Detection Layer Based on Genetic Algorithm Clustering

The first detection layer uses heuristic rules to identify several known intrusions, but some intrusions are not detected at this stage. These include both known and unknown intrusions. One of the major challenges of Intrusion Detection Systems (IDSs) is detecting unknown intrusions for which no training examples exist. To address this challenge, a clustering technique can be used to identify unknown attacks. Various clustering algorithms, such as the K-means algorithm and genetic algorithms, have been developed for this purpose. Researchers in [31] utilized a genetic algorithm in their IDS, demonstrating its efficiency over the K-means algorithm based on their results. We have also employed a genetic algorithm in our proposed IDS.

The genetic algorithm is an adaptive and meta-heuristic optimization technique based on the principle of survival of the fittest [32]. In such algorithms, all individuals compete within a generation to acquire resources, and the most successful ones are allowed to produce offspring. Iterations of this approach yield progressively better results with each generation. In this population-based search method, each sample is represented as a chromosome. Chromosomes can be encoded in various forms, including binary, tree structures, permutations, and real values [31]. In our proposed IDS, we use real values for representation. Each generation consists of ten chromosomes, each containing two genes, with each gene comprising 41 alleles. In the proposed algorithm, genes represent the centers of clusters, and alleles denote the characteristics of each center. The initial population is randomly selected from the training dataset, and the quality of each chromosome is evaluated using a proposed fitness function, as shown in Fig. 3. This function helps select the most appropriate chromosomes to produce offspring for the next generation.

```
Algorithm. Fitness of genetic clustering
Input:
        ch: chromosome;
        Tr_d: Data of Training Dataset
Output:
        F_v: Value of Fitness
-----------------------------------------------------------------
1:  set label for all instance of tr_d by nearest to Gen1 OR Gen2;
2:  split tr_d to cluster1 and cluster2 by labels;
3:  compute intra and extra quality of cluster1 AND  cluster2;
4:  set Fitness value(F_v) for qualities;
5:  Return F_v;
```

Fig. 3: The algorithm of the fitness function of genetic clustering,

In this algorithm, one gene on the input chromosome is randomly selected as the center of the normal cluster, and another gene is selected as the center of the abnormal cluster. Then, all samples in the training dataset are labeled based on their Euclidean distance to one of the cluster centers, resulting in two clusters: normal and abnormal. After clustering, new cluster centers are determined using the mean values of the samples within each cluster. Intra-cluster and inter-cluster distances are then calculated as follows:

Intra-Cluster Distance: For each cluster, the Euclidean distance of all samples from the new cluster center is calculated. This value represents the distance within the cluster.

Inter-Cluster Distance: The Euclidean distance between the two new cluster centers is calculated, and a new point is defined at the midpoint of this distance. The Euclidean distance of all samples from this midpoint is then calculated. This value represents the inter-cluster distance.

The fitness function output is higher when the intra-cluster distance is minimized and the inter-cluster distance is maximized. After determining the fitness of all chromosomes, five chromosomes are selected using the Tournament Selection Method. Recombination (at a rate

of 0.8) and mutation (at a rate of 0.01) are performed on them to produce new chromosomes. These new chromosomes are combined with half of the previous generation to form a new generation, and the evaluation process begins again using the fitness function. After 15 generations, the algorithm converges, and in the final generation, the genes of the best chromosomes are considered the centers of the normal and abnormal clusters. These cluster centers are then sent to the second diagnostic layer, where the detection layer labels the experimental input samples based on these clusters. The algorithm for the second recognition layer is shown in Fig. 4.

```
Algorithm. Second Detection Layer - clustering
Input:
        Ts_i : Instance Data of Testing Dataset
        Ts_c1 : center of clusters anomaly
        Ts_c2 : center of clusters normal
Output:
        isattack : 0 is not attack and 1 is attack
-----------------------------------------------------------------------
1:  if (distacnce(Ts_i ,Ts_c1) < distacnce(Ts_i ,Ts_c2) )
2:          isattack=1;
3:  Return;
```

Fig. 4: The algorithm of the second detection layer based on clustering.

In this detection layer, if the input test sample is closer to the center of the abnormal cluster than to the center of the normal cluster, it is labeled as abnormal. This clustering-based approach allows for the identification of unknown intrusions by leveraging the patterns and characteristics inherent in the data.

C. The Third Detection Layer Based on Back Propagation Neural Network

While the first and second detection layers address known and clustered attacks, some attacks may still evade detection due to their similarity to normal samples. To address this challenge and enhance the Intrusion Detection System's (IDS) accuracy, a Back Propagation Neural Network (BPNN) classifier is employed in the third detection layer. The BPNN is a feed-forward neural network used for supervised learning in various applications such as pattern recognition and image processing [8].

The BPNN consists of multiple layers: an input layer, one or more hidden layers, and an output layer. Each layer comprises nodes connected by activation functions, such as hyperbolic tangent or sigmoid functions. During the learning process, the network's settings and connection weights are initially established. For each training sample, a forward calculation is performed from the input layer through the hidden layers to the output layer. A backward calculation follows to correct errors and adjust connection weights. This iterative process enables the network to learn the training samples and recognize unknown patterns effectively. BPNNs have demonstrated high detection rates compared to other neural network techniques [33].

In the proposed IDS, the BPNN classifier is trained using the training dataset. The trained model is then deployed in the third detection layer, where it examines the input experimental samples. If an intrusion is detected, the BPNN classifier labels it accordingly. The algorithm of the third detection layer based on the BPNN is illustrated in Fig. 5.

```
Algorithm. Third Detection Layer - Classifier
Input:
        Ts_i : Instance Data of Testing Dataset
        Net_structure : structure of Back Propagation Neural Network
Output:
        isattack : 0 is not attack and 1 is attack
-----------------------------------------------------------------------
1:  if (distacnce(Ts_i ,Ts_c1) < distacnce(Ts_i ,Ts_c2) )
2:          isattack=1;
3:  Return;
```

Fig. 5: The algorithm of the third detection layer based on BPNN classification.

The third detection layer's algorithm includes the following steps:

Training the BPNN Classifier:

- The classifier is trained using the training dataset.
- Network settings and connection weights are established.
- Forward and backward calculations are performed iteratively to minimize errors and adjust weights.

Deploying the Trained Model:

- The trained BPNN model is sent to the third detection layer.
- The layer uses the trained model to classify input experimental samples.

Intrusion Detection:

- The BPNN classifier examines each input sample.
- If the sample is classified as an intrusion, it is labeled as such.

By incorporating the BPNN classifier in the third detection layer, the IDS can effectively identify intrusions with feature values similar to normal samples, thereby increasing overall detection accuracy. This comprehensive multi-layer approach ensures robust detection of both known and unknown intrusions, leveraging the strengths of heuristic rules, genetic algorithm clustering, and supervised learning.

Experiments and Results

Researchers utilize various datasets to demonstrate the efficiency of their intrusion detection systems. Evaluating an algorithm with real data is crucial as it reflects the algorithm's practical applicability and robustness [17]. In this study, we evaluated the proposed

algorithm using the real-world and widely utilized NSL-KDD dataset. This section discusses the employed dataset, the evaluation methodology of the proposed method, and the experimental results. The proposed algorithms were implemented in MATLAB and executed on a computer equipped with an Intel Core i5 CPU, featuring 4 cores at 2.2 GHz, and 4 GB of RAM.

D. Data Set

The initial event for developing an intrusion detection system took place in 1998, supported by the DARPA organization [3]. During this event, a cyber-attack scenario was simulated at the Air Force Base. This simulation was repeated in 1999 with enhancements by the Computer Security Association [2]. Over seven weeks, raw network TCP/IP data was collected. However, for this data to be useful in learning algorithms, feature extraction was necessary. A team of researchers [3] won the KDD International Knowledge Discovery and Data Mining Tools Competition by presenting a feature extraction method for this raw data set, resulting in the creation of the KDDCUP99 data set. This data set has since become a widely used benchmark in intrusion detection research [7]. In [30], researchers identified several shortcomings in the KDDCUP99 dataset using statistical methods, which significantly impacted system evaluation performance. To address these issues, they proposed an improved dataset known as NSL-KDD, which allows for better comparison of different intrusion detection models.

The NSL-KDD data set maintains the 41 features of the KDDCUP99 data set, and the class labels remain the same. The characteristics of each instance and their details are listed in Table 1 [34].

Most features in the data set have continuous values between 0 and 1, but some are symbolic. Attributes such as land, logged_in, is_host_login, and is_guest_login have values of 0 or 1 and can be treated as continuous properties. The protocol attribute has 3 values, the service attribute has 66 values, and the flag attribute has 11 distinct values that can be converted to continuous values using various methods [12]. Each instance's class determines whether it is normal or an attack. There are several types of attacks in the NSL-KDD database, detailed and classified in Table 2 [7].

The NSL-KDD dataset comprises four subsets: KDDTrain+, KDDTrain+_20%, KDDTest+, and KDDTest-21. The KDDTrain+ and KDDTrain+_20% datasets are used to train learning algorithms, while the KDDTest+ and KDDTest-21 datasets are used to evaluate the performance of intrusion detection algorithms. The training dataset samples are categorized as either normal or known attacks. In contrast, the test dataset includes both known attacks and types of attacks absent in the training dataset, considered unknown intrusions in IDSs. The number of samples for each type in each dataset is shown in Table 3.

Table 1: Description of input sample features [34]

No.	Feature Category	Feature Name	Data Type
1	Basic Features	duration	Continuous
2		protocol_type	Symbolic
3		service	Symbolic
4		flag	Symbolic
5		src_bytes	Continuous
6		dst_bytes	Continuous
7		land	Symbolic
8		wrong_fragment	Continuous
9		urgent	Continuous
10	Content Features	hot	Continuous
11		num_failed_logins	Continuous
12		logged_in	Symbolic
13		num_compromised	Continuous
14		root_shell	Continuous
15		su_attempted	Continuous
16		num_root	Continuous
17		num_file_creations	Continuous
18		num_shells	Continuous
19		num_access_files	Continuous
20		num_outbound_cmds	Continuous
21		is_host_login	Symbolic
22		is_guest_login	Symbolic
23	Traffic Features	count	Continuous
24		srv_count	Continuous
25		serror_rate	Continuous
26		srv_serror_rate	Continuous
27		rerror_rate	Continuous
28		srv_rerror_rate	Continuous
29		same_srv_rate	Continuous
30		diff_srv_rate	Continuous
31		srv_diff_host_rate	Continuous
32	Host-based Features	dst_host_count	Continuous
33		dst_host_srv_count	Continuous
34		dst_host_same_srv_rate	Continuous
35		dst_host_diff_srv_rate	Continuous
36		dst_host_same_src_port_rate	Continuous
37		dst_host_srv_diff_host_rate	Continuous
38		dst_host_serror_rate	Continuous
39		dst_host_srv_serror_rate	Continuous
40		dst_host_rerror_rate	Continuous
41		dst_host_srv_rerror_rate	Continuous

Table 2: Details of attacks type [7]

DoS	U2R	R2L	PROBE
Back	Perl	FTP write	IP sweep
Ping of Death	Buffer Overflow	Guess password	NMAP
Smurf	Load module	IMAP	Port sweep
Land	Rootkit	Multi HOP	Satan
Teardrop		Phf	
		SPY	
		Wareclient	
		Warezmaster	

Table 3: Details of NSL-KDD subsets

Subsets	#Normal data	#Known attack	#Unknown attack	Total
KDDTrain+	67343 53%	58630 47%	0	125973
KDDTrain+_20%	13449 53%	11743 47%	0	25192
KDDTest+	9711 43%	9083 40%	3750 17%	22544
KDDTest-21	2152 18%	5958 50%	3740 32%	11850

According to Table 3, the total number of training samples in KDDTrain+ is approximately five times that of KDDTrain+_20%. Utilizing the KDDTrain+ dataset increases the accuracy of learning algorithms, whereas using the KDDTrain+_20% dataset increases the speed of learning algorithms. Additionally, the total number of samples in the KDDTest-21 collection is about half of those in the KDDTest+ collection, but the number of unknown attacks in both is equal. Therefore, achieving high detection accuracy in an IDS that uses the KDDTest-21 dataset for evaluation is more challenging.

E. Evaluation Criteria

Four states of detection occur for an event in the process of intrusion detection:

True Positive (TP): The input sample is correctly identified as an intrusion at the end of the detection process.

True Negative (TN): The input sample is correctly identified as normal at the end of the detection process.

False Positive (FP): The input sample is actually normal but is incorrectly identified as an intrusion at the end of the detection process.

False Negative (FN): The input sample is actually an intrusion but is incorrectly identified as normal at the end of the detection process.

These four measures (TP, TN, FP, FN) are crucial for evaluating and calculating the accuracy and efficiency of an Intrusion Detection System (IDS). Important criteria for evaluating intrusion detection systems based on these detection modes are defined as follows [12]:

Accuracy (ACC): This metric represents the percentage of correctly labeled samples out of the total samples, as shown in (1):

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

Detection Rate (DR): This metric represents the percentage of correctly identified abnormal samples out of the total number of abnormal samples, as shown in (2):

$$DR = \frac{TP}{TP+FN} \quad (2)$$

False Positive Rate (FPR): This metric represents the percentage of incorrectly identified normal samples (false positives) out of the total number of normal samples, as shown in (3):

$$FPR = \frac{FP}{FP+TN} \quad (3)$$

According to these criteria, an IDS performs better when it has higher accuracy (ACC) and detection rate (DR), and a lower false positive rate (FPR) [13], [16].

Experimental Results and Discussion

In this section, we present the results of experiments conducted on the proposed intrusion detection system. We used the KDDTrain+ dataset for training and both KDDTest+ and KDDTest-21 datasets for testing. For evaluating the proposed IDS, all data in the dataset are divided into two classes: normal and attack, and all 41 features of data points have been used in the algorithm's training phase. Table 4 indicates the result of testing performed on the two datasets, KDDTest+ and KDDTest-21, using the proposed IDS.

Table 4 shows the number and ratio of the four parameters: True Positive, True Negative, False Positive, and False Negative. It also shows the False Positive Rate, Detection Rate, and Accuracy of the algorithm.

Table 4: Test results of proposed method on KDDtest+ and KDDtest-21

Subsets	KDDTrain+	KDDTest-21
#TP	10404 46%	7269 61%
#TN	9275 41%	1746 15%
#FP	436 2%	406 3%
#FN	2429 11%	2429 21%
FPR	4.49	18.87
DR	81.07	74.95
ACC	87.29	76.08

As can be seen, the proposed method has achieved significant results. Although many researchers use the NSL-KDD dataset to demonstrate the performance of their intrusion detection systems, few have provided the total accuracy of their method. As mentioned in 2th Section, the following algorithms have good results in the NSL-KDD data set:

- FSSL [7]
- RNN-IDS [13]
- FSSL-EL [14]
- TSE-IDS [15]
- DCNN [16]
- DTBE [17]
- ICVAE-DNN [18]
- CFS-DE_based [19]
- IGAN-IDS [20]
- I-SiamIDS [21]
- SVM [35]

- em J48 [36]
- Two-tier classifier [37]
- Ensemble J48 PART [38]
- CNN-BiLSTM [22]
- AS-CNN [23]
- BAT-MC [24]
- HSDN [25]
- DSN [26]
- CRISP-DM [27]
- Pardeshi [28]

In this section, we compare the total accuracy of the proposed algorithm with that of methods mentioned earlier. We focus on recent works with remarkable results, avoiding comparisons with older studies. Fig. 6 and Fig. 7 display the total accuracy obtained by using the KDDTest+ and KDDTest-21 datasets, respectively.

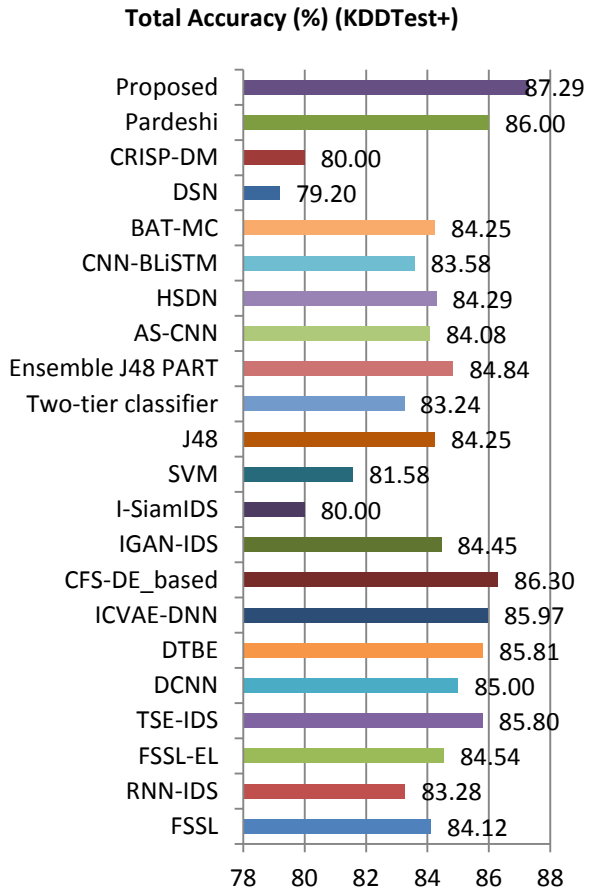


Fig. 6: Comparison of the accuracy of solutions using KDDTest+ data set for binary classification.

It's important to note that the reported results in the CFS-DE_based method, a commonly used approach, are based on testing with all features in dataset. Also, it is worth noting that CFS-DE_based [19], IGAN-IDS [20], I-SiamIDS [21], SVM [35], J48 [36], Two-tier classifier [37], Ensemble J48 PART [38], CNN-BiLSTM [32], HSDN [25], DSN [26], CRISP-DM [27] and Pardeshi [28] have not reported results on KDDTest-21.

The results show that the total accuracy of the proposed method in both experimental datasets is superior to all other mentioned algorithms. As described in 3th Section, this method employs a three-layer hybrid system, with each layer responsible for detecting different types of attacks. Additionally, the sequence of execution of these three layers significantly contributes to the overall success rate and detection of intrusions. As illustrated in Fig. 6 and Fig. 7, the proposed method achieves a higher degree of accuracy. Despite the smaller number of samples and the similar types of attacks in the KDDTest-21 set compared to KDDTest+, the proposed method still maintains high intrusion detection accuracy. Notably, more studies report results on the KDDTest+ database than on KDDTest-21.

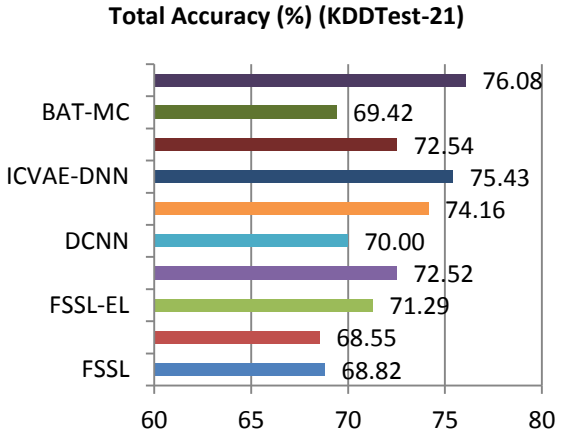


Fig. 7: Comparison of the accuracy of solutions using KDDTest-21 dataset for binary classification.

Evaluating intrusion detection systems requires considering metrics beyond accuracy. As previously mentioned, DR measures the proportion of actual attacks correctly identified, reflecting the system's ability to detect threats, while FPR quantifies the proportion of normal instances incorrectly classified as attacks, directly impacting the number of false alarms. A desirable IDS aims for a high DR and a low FPR. Our proposed method achieved a DR of 81.07% on KDDTrain+ and 74.95% on KDDTest-21, demonstrating strong performance. Our FPR was 4.49% on KDDTest+ and 18.87% on KDDTest-21, representing acceptable results.

It is important to note that many related studies do not report both FPR and DR directly, limiting direct comparison. Among the studies that do, ICVAE-DNN [18] reports DRs of 77.43% on KDDTrain+ and 72.86% on KDDTest-21, while RNN-IDS [13] achieved a DR of 72.95% on KDDTrain+. For FPR, FSSL-EL [14] reported 5.31% on KDDTest+ and 20.35% on KDDTest-21, and RNN-IDS [13] reported 3.6% on KDDTest+, which is lower than our method's FPR on the same dataset. It should be noted that other reviewed studies either used different evaluation metrics or calculated DR and FPR separately for each attack type, making direct comparison in this section inappropriate.

The difference in FPR between KDDTest+ (4.49%) and KDDTest-21 (18.87%) for our method highlights the significant impact of dataset characteristics. The KDDTest-21 dataset, with its higher proportion of unknown attacks, presents a greater challenge for accurate classification. This suggests that while our method effectively handles known attack patterns, further refinement is necessary to improve its ability to generalize to unseen attacks and consequently reduce the FPR in such challenging scenarios.

One crucial aspect to consider when deploying intrusion detection systems in real-world environments is the speed of detection execution. To address this, we have examined the time taken to process an input sample at each of the detection layers individually.

Table 5: The spent time of each detection layer

Detection layer	Detection method	Type of learning	Spent time (S)
Based on heuristic rules	Signature	-	$1.0753*10^{-4}$
Based on clustering	Anomaly	Unsupervised	$3.3016*10^{-4}$
Based on neural networks	Anomaly	Semi-supervised	$614.5862*10^{-4}$

Table 5 indicates that the first and second detection layer does not require much time to diagnose, and their time consuming is close to real-time. The third detection layer is more time-consuming due to the BPNN classifier. It can be said that by using of powerful processors it is possible to implement intrusion detection systems in real-time environments.

Conclusion and Future Work

In this paper, we proposed an intrusion detection system (IDS) employing a hybrid approach to address the challenges of detecting intrusions without training samples and distinguishing between normal samples and those with similar patterns. The system incorporates three detection layers, each employing distinct methodologies to differentiate intrusions from normal samples. The first Detection Layer utilizes signature detection, leveraging heuristic rules to identify known intrusions similar to normal samples. The second Detection Layer is an anomaly-based approach using a clustering method to detect unknown intrusions. The third Detection Layer is another anomaly-based approach using a classification method, specifically a back propagation neural network (BPNN), to detect known intrusions with available training samples.

For evaluation, we utilized the NSL-KDD dataset. By providing solutions for handling similar normal intrusions, intrusions without training samples, and intrusions with training samples, the proposed system demonstrated an increased detection rate and overall accuracy. A comparative analysis of the proposed method's total accuracy against several recently proposed methods evaluated using the NSL-KDD dataset highlights the effectiveness and success of our approach. The hybrid nature of our system, combining signature-based and anomaly-based methods, ensures robust performance across various intrusion scenarios.

Future research will focus on exploring alternative supervised learning algorithms to replace the back propagation neural network in the third detection layer. We anticipate that finding a more efficient algorithm could further enhance the detection rate and accuracy of the proposed IDS, thus continuing to improve its effectiveness in real-world applications.

Author Contributions

A. Beigi designed the experiments, analyzed the data, interpreted the results, and authored the manuscript.

Acknowledgment

Akram Beigi acknowledges the financial support received from Shahid Rajaee Teacher Training University under grant number 5973/15.

Conflict of Interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

IDS	Intrusion Detection System
DoS	Denial of Service
R2L	Remote-to-Local
U2R	User-to-Root
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ACC	Accuracy
DR	Detection Rate
FPR	False Positive Rate
BPNN	Back Propagation Neural Network
KNN	K-Nearest Neighbors
SVM	Support Vector Machine

References

[1] A. Thakkar, R. Lohiya, "A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions," Artif. Intell. Rev., 55(1): 453-563, 2022.

[2] S. Venkatesan, "Design an intrusion detection system based on feature selection using ML algorithms," Math. Stat. Eng. Appl., 72(1): 702-710, 2023.

[3] A. Thakkar, R. Lohiya, "A review of the advancement in intrusion detection datasets," Procedia Comput. Sci., 167: 636-645, 2020.

[4] M. Sabhnani, G. Serpen, "KDD feature set complaint heuristic rules for R2L attack detection," in Security and Management, 310-316, 2003.

[5] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," Cybersecurity, 2(1): 20, 2019.

[6] S. Aljawarneh, M. Aldwairi, M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," J. Comput. Sci., 25: 152-160, 2018.

[7] R. A. R. Ashfaq, X. Z. Wang, J. Z. Huang, H. Abbas, Y. L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," Inf. Sci., 378: 484-497, 2017.

[8] I. Goodfellow, Y. Bengio, A. Courville, "6.5 Back-Propagation and Other Differentiation Algorithms," in Deep Learning, MIT Press, 200-220, 2016. ISBN 9780262035613.

[9] C. Guo, Y. Ping, N. Liu, S. S. Luo, "A two-level hybrid approach for intrusion detection," Neurocomputing, 214: 391-400, 2016.

[10] P. Kar, S. Banerjee, K. C. Mondal, G. Mahapatra, S. Chattopadhyay, "A hybrid intrusion detection system for hierarchical filtration of anomalies," in Inf. Commun. Technol. Intell. Syst., 417-426, Springer, Singapore, 2019.

[11] V. Hajisalem, S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," Comput. Netw., 136: 37-50, 2018.

[12] W. L. Al-Yaseen, Z. A. Othman, M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Syst. Appl., 67: 296-303, 2017.

[13] C. Yin, Y. Zhu, J. Fei, X. He, "A deep learning approach for intrusion detection using recurrent neural networks," IEEE Access, 5: 21954-21961, 2017.

[14] Y. Gao, Y. Liu, Y. Jin, J. Chen, H. Wu, "A novel semi-supervised learning approach for network intrusion detection on cloud-based robotic system," IEEE Access, 6: 50927-50938, 2018.

[15] B. A. Tama, M. Comuzzi, K. H. Rhee, "TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system," IEEE Access, 7: 94497-94507, 2019.

[16] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, K. Han, "Enhanced network anomaly detection based on deep neural networks," IEEE Access, 6: 48231-48246, 2018.

[17] P. Illy, G. Kaddoum, C. M. Moreira, K. Kaur, S. Garg, "Securing fog-to-things environment using intrusion detection system based on ensemble learning," in Proc. 2019 IEEE Wireless Commun. Netw. Conf. (WCNC), 1-7, 2019.

[18] Y. Yang, K. Zheng, C. Wu, Y. Yang, "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network," Sensors, 19(11): 2528, 2019.

[19] R. Zhao, Y. Mu, L. Zou, X. Wen, "A hybrid intrusion detection system based on feature selection and weighted stacking classifier," IEEE Access, 10: 71414-71426, 2022.

[20] S. Huang, K. Lei, "IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks," Ad Hoc Netw., 105: 102177, 2020.

[21] P. Bedi, N. Gupta, V. Jindal, "I-SiamIDS: an improved Siam-IDS for handling class imbalance in network-based intrusion detection systems," Appl. Intell., 51(2): 1133-1151, 2021.

[22] K. Jiang, W. Wang, A. Wang, H. Wu, "Network intrusion detection combined hybrid sampling with deep hierarchical network," IEEE Access, 8: 32464-32476, 2020.

[23] Z. Hu, L. Wang, L. Qi, Y. Li, W. Yang, "A novel wireless network intrusion detection method based on adaptive synthetic sampling and an improved convolutional neural network," IEEE Access, 8: 195741-195751, 2020.

[24] T. Su, H. Sun, J. Zhu, S. Wang, Y. Li, "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset," IEEE Access, 8: 29575-29585, 2020.

[25] M. Latah, L. Toker, "Minimizing false positive rate for DoS attack detection: A hybrid SDN-based approach," ICT Express, 6(2): 125-127, 2020.

[26] Y. Tang, L. Gu, L. Wang, "Deep Stacking Network for Intrusion Detection," Sensors, 22(1): 25, 2022.

[27] Y. Yuliana, D. H. Supriyadi, M. R. Fahlevi, M. R. Arisagas, "Analysis of NSL-KDD for the Implementation of Machine Learning in Network Intrusion Detection System," J. Inform. Inf. Syst. Softw. Eng. Appl. (INISTA), 6(2): 80-89, 2024.

[28] N. G. Pardeshi, D. V. Patil, "Binary and Multiclass Classification Intrusion Detection System using Benchmark NSL-KDD and Machine Learning Models," in Proc. 2024 Int. Conf. Data Sci. Netw. Secur. (ICDSNS), 1-7, 2024.

[29] D. Gümüşbaş, T. Yıldırım, A. Genovese, F. Scotti, "A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems," IEEE Syst. J., 15(2): 1717-1731, 2020.

[30] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE Symp. Comput. Intell. Secur. Def. Appl., 1-6, 2009.

[31] N. B. Aissa, M. Guerroumi, "A genetic clustering technique for anomaly-based intrusion detection systems," in Proc. 2015 IEEE/ACIS Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distrib. Comput. (SNPD), 1-6, 2015.

[32] D. Greiner, J. Periaux, D. Quagliarella, J. Magalhaes-Mendes, B. Galvan, "Evolutionary algorithms and metaheuristics: applications in engineering design and optimization," Math. Probl. Eng., 2018.

[33] F. Salo, A. B. Nassif, A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," Comput. Netw., 148: 164-175, 2019.

[34] P. Mishra, V. Varadharajan, U. Tupakula, E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," IEEE Commun. Surv. Tutorials, 21(1): 686-728, 2018.

[35] Q. M. Alzubi, M. Anbar, Z. N. Alqattan, M. A. Al-Betar, R. Abdullah, "Intrusion detection system based on a modified binary grey wolf optimization," Neural Comput. Appl., 2019.

[36] N. T. Pham, E. Foo, S. Suriadi, H. Jeffrey, H. F. M. Lahza, "Improving performance of intrusion detection system using ensemble methods and feature selection," in Proc. Australas. Comput. Sci. Week Multiconf., 1-6, 2018.

[37] H. H. Pajouh, G. Dastghaibyfard, S. Hashemi, "Two-tier network anomaly detection model: a machine learning approach," J. Intell. Inf. Syst., 48: 61-74, 2017.

[38] N. Paulauskas, J. Auskalnis, "Analysis of data pre-processing influence on intrusion detection using NSL-KDD dataset," in 2017 Open Conf. Electr. Electron. Inf. Sci. (eStream), 1-5, 2017.

Biographies

Akram Beigi is currently an Assistant Professor in the Computer Engineering Department at Shahid Rajaee Teacher Training University, Tehran, Iran. She earned her M.Sc. and Ph.D. in Computer Engineering – Artificial Intelligence from Iran University of Science and Technology. Her research interests include machine learning, deep learning, cybersecurity, and multi-agent systems. Her expertise spans various AI-driven applications, particularly in anomaly detection, optimization, biometric authentication, and intelligent decision-making systems.

- Email: akrambeigi@sru.ac.ir
- ORCID: 0000-0003-2268-8734
- Web of Science Researcher ID: ADA-5427-2022
- Scopus Author ID: 36661958200
- Homepage: https://www.sru.ac.ir/en/school-of-computer/akram-beigi/

How to cite this paper:
A. Beigi, "A hybrid three-layered approach for intrusion detection using machine learning methods," J. Electr. Comput. Eng. Innovations, 13(2): 443-454, 2025.

DOI: 10.22061/jecei.2025.11530.811

URL: https://jecei.sru.ac.ir/article_2308.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Second Generation Current Conveyor Employing a Flipped Voltage Follower and Improved DC Voltage Gain Operational Transconductance Amplifier

E. Tavassoli [1], *S. M. Anisheh* [1,*], *M. Radmehr* [1]

Department of Electrical Engineering, Sari Branch, Islamic Azad University, Sari, Iran.

Article Info

Article History:
Received 04 November 2024
Reviewed 20 January 2025
Revised 26 February 2025
Accepted 09 March 2025

Keywords:
Current-mode circuit
Second generation current conveyor
Flipped voltage follower
Operational amplifier

*Corresponding Author's Email Address:
s.m.anisheh@ee.kntu.ac.ir

Abstract

Background and Objectives: The background of this research is the significance of current conveyors as essential building blocks in current-mode circuits. The objective is to design and simulate a second generation current conveyor (CCII) in a 180-nm CMOS process, aiming to achieve low impedance, accurate voltage copying, and high DC voltage gain.

Methods: The proposed CCII design utilizes a flipped voltage follower (FVF) to provide low impedance. A novel operational transconductance amplifier (OTA) is introduced to accurately copy the voltage within the circuit. This OTA employs a positive feedback technique to increase its output resistance, thereby enhancing DC voltage gain and reducing input impedance. The performance of the presented CCII is evaluated through simulations in a 180-nm CMOS technology using Cadence software.

Results: The simulation results show the successful operation of the CCII circuit. Key performance metrics include voltage and current tracking errors of 0.3% and 0.1%, respectively, and a bandwidth of 1.4 GHz.

Conclusion: The research concludes that a new OTA and CCII have been successfully simulated in a 180-nm CMOS process. The proposed CCII design, based on FVF and a novel OTA with positive feedback, achieves improved DC voltage gain without compromising other specifications like power consumption, UGBW, and stability. The tracking errors in the proposed method are lower compared to existing approaches.

Introduction

Analog signal processing can be achieved through either voltage-mode or current-mode techniques. Voltage-mode circuits typically experience a fixed bandwidth product, which results in a decrease in amplifier bandwidth as voltage gain increases [1]. To address this issue, various approaches have been suggested to surmount the gain-dependent bandwidth constraint. One promising approach involves utilizing current-mode devices, which are capable of operating at high frequencies. While voltage-mode circuits have been widely used, current-mode circuits have emerged as a promising alternative due to their inherent advantages. These include a higher slew rate (SR), a wider operating frequency range, and a bandwidth that remains constant across various gain levels [3].

Current conveyors are regarded as fundamental components in current-mode circuits. These versatile analog components have garnered significant attention from researchers [3], [4]. They are characterized as three-port structures. When the ports are labeled X, Y, and Z, the following equation describes their behavior:

$$\begin{bmatrix} I_Y \\ V_X \\ I_Z \end{bmatrix} = \begin{bmatrix} 0 & M & 0 \\ 1 & 0 & 0 \\ 0 & N & 0 \end{bmatrix} \begin{bmatrix} V_Y \\ I_X \\ V_Z \end{bmatrix} \quad (1)$$

where in (1), I and V represent the currents and voltages

at the respective nodes. This discussion focuses on the second generation current conveyor (CCII), where the value of M is zero. An ideal CCII exhibits the following characteristics:

- High impedance at node Y, low impedance at node X, and high output impedance at node Z.
- Precise voltage transfer from node Y to node X and current transfer from node X to node Z.
- High speed performance for a specified bias current.
- Low operating supply voltage.

Several CCII implementations have been proposed in the literature [3]-[10]. In [3], a current-mode instrumentation amplifier using a fully differential operational floating conveyor (FD-OFC) is presented. The FD-OFC enhances design flexibility and noise rejection, requiring only one circuit for simplicity and low-voltage operation (1.2 V). The design is analyzed analytically and simulated in 130-nm technology using Cadence software.

References [4] and [5] introduce the operational floating current conveyor (OFCC) and discuss its applications. The OFCC operates efficiently with a 1.2V supply voltage and offers a wider bandwidth. It consists of two CCII blocks, a non-inverting trans-impedance amplifier, and a current steering circuit. The CCII includes a unity-gain amplifier followed by a common source amplifier [6]. In the first CCII, voltage tracking occurs between nodes Y and X, while current tracking is performed between nodes W and Z in the second CCII. This approach provides high performance using a simple circuit topology, but the resistance at terminal X of the CCII can be relatively high.

Reference [7] explores three realizations of the OFC. The first realization employs two CCII blocks and a non-inverting trans-impedance amplifier. The second OFC uses a CCII block, a non-inverting trans-impedance amplifier, and a positive current follower. The third OFC configuration includes a CCII block, an inverting trans-impedance amplifier, and a positive current follower. However, these realizations require a lower impedance at terminal X and a voltage gain closer to unity.

Reference [8] presents a digitally controlled OFCC-based filter, incorporating a trans-conductance amplifier and a bandpass filter to reduce power consumption. However, this approach has a limited bandwidth. A CMOS OFCC structure suitable for instrumentation amplifiers is proposed in [9], designed in 90-nm process. While it reduces power consumption, the bandwidth improvement is not significant. Reference [10] introduces a logarithmic amplifier based on the OFCC, comprising an OFCC, a diode, and a grounded resistor. The CCII CMOS circuit uses a cascade current mirror to increase output impedance, but the low impedance condition at node X is not fully achieved.

This paper proposes a new CMOS CCII with improved specifications, including low voltage and current tracking errors and high bandwidth. The proposed CCII utilizes a FVF to achieve low impedance at node X. A novel OTA is introduced to accurately copy the voltage from node Y to node X. By employing positive feedback, the OTA's output resistance and DC voltage gain are increased, further reducing the impedance at node X and increasing the input impedance at node Y.

The subsequent sections of this article are organized as follows. Initially, the circuit structure of the presented OTA and CCII will be discussed in detail. Next section presents the simulation results and compares them to existing works. To wrap up, the key findings of the research presented will be summarized.

Proposed Circuits

In this section, the proposed circuits of OTA and CCII are explained and the theoretical analysis are presented.

Proposed OTA

OTAs are fundamental components in numerous analog and mixed-mode circuits [11]-[16]. The conventional OTA circuit was first introduced in [17]. The presented OTA is illustrated in Fig. 1. Nodes V_O and V_{out} represent the outputs of the first and second stages, respectively. The input differential pair comprises transistors M_1 and M_2, biased by the FVF circuit. The FVF is capable of operating with low supply voltages [18]-[20]. M_{2a}, M_{2b}, M_{3a}, and M_{3b} devices, in conjunction with M_{1a} and M_{1b} transistors acting as current sources, constitute the FVF. This structure incorporates two additional signal amplification paths.

In the first path, the source of M_2 device is connected to the gate of M_9 transistor using the FVF. Additionally, there is a path connecting the source of M_1 to the gate of M_{10}. Consequently, the input signal is present at the gate-source of M_9 and M_{10}, increasing the first-stage transconductance from $g_{m1,2}$ to $g_{meff1} = g_{m1,2} + g_{m9,10}$. This enhancement leads to an improvement in DC voltage gain.

FVF-based nonlinear current mirrors (NLCM), consisting of M_{11}, M_{15}, M_{17}, M_{19} transistors and M_{12}, M_{14}, M_{16}, M_{18} devices, are used in the second stage to improve the SR. As depicted in Fig. 1, the gate of M_{11} is connected to V_{O+}, and similarly, the gate of M_{12} is connected to V_{O-}. This connection provides an additional signal amplification path by increasing the second-stage transconductance.

This paper modifies the OTA structure described in [17] to boost DC voltage gain without impacting power consumption. The output resistance of the second stage is improved by employing positive feedback, which leads to an increased DC voltage gain. Transistor pairs (M_{23}, M_{24}) and (M_{25}, M_{26}) are added to the conventional structure as current sources.

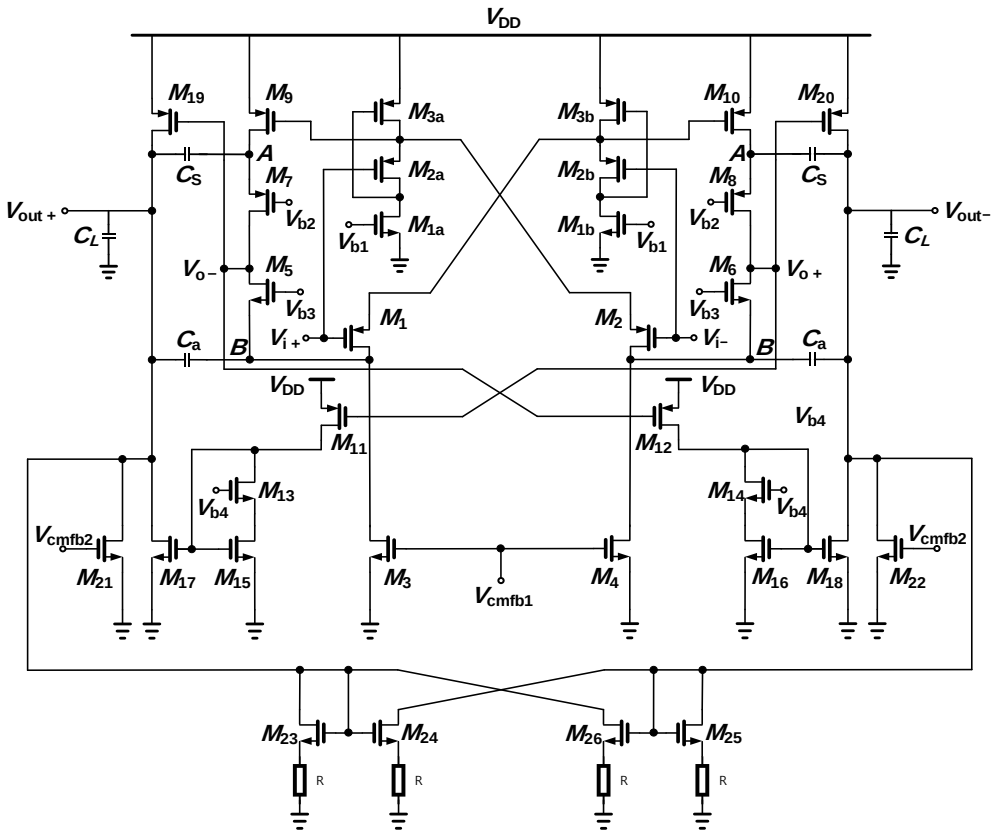


Fig. 1: The proposed OTA.

Positive feedback is established by connecting these current sources to the outputs [21]. The DC voltage gain of the suggested OTA is determined as follows:

$$\begin{bmatrix} I_Y \\ V_X \\ I_Z \end{bmatrix} = \begin{bmatrix} 0 & M & 0 \\ 1 & 0 & 0 \\ 0 & N & 0 \end{bmatrix} \begin{bmatrix} V_Y \\ I_X \\ V_Z \end{bmatrix}$$

$$A_d = A_1 \times A_2 \quad (2)$$

where A_1 is the DC voltage gain of the first stage and it can be obtained as below:

$$A_1 = G_{meff1} R_{out1} \quad (3)$$

and,

$$R_{OUT1} = g_{m7} r_{ds7} r_{ds9} \| \left(g_{m5} r_{ds5} (r_{ds1} \| r_{ds3}) \right) \quad (4)$$

A_2 represents the DC voltage gain of the second stage and is calculated as follows:

$$A_2 = G_{meff2} R_{out2} \quad (5)$$

where,

$$G_{meff2} = g_{m19+} g_{m11} g_{m17} (r_{ds11} \| r_{ds13}) \quad (6)$$

$$R_{out2} = \frac{1}{g_{ds17} + g_{ds19} + g_{ds21} + G_{cs}} \quad (7)$$

$$G_{cs} = \frac{1}{\frac{g_{m26} r_{ds26} R + r_{ds26} + R}{1 - g_{m26} r_{ds26}}} + \frac{g_{m23}}{1 + g_{m23} R} \quad (8)$$

In the above equations, R_{out} refers to the output resistance of each stage. g_m is the transconductance of a NMOS/PMOS device. Moreover, r_{ds} is the resistance of the drain-source of the utilized transistors, $g_{ds}=1/r_{ds}$, and $G_{cs}=1/R_{cs}$. R_{cs} is the resistance of the current source seen from the node V_{out+} and it can be calculated according to Fig. 2. If the denominator of (7) is close to zero but its value is positive, then the differential gain increases significantly and the system is stable.

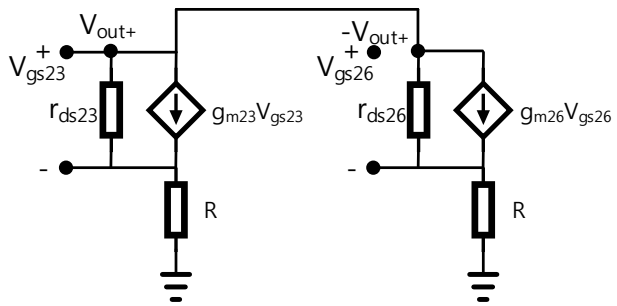


Fig. 2: Equivalent small-signal model for calculating the output resistance of the current source.

Proposed CCII

The proposed CCII structure is depicted in Fig. 3. The set of NMOS devices (M_1, M_2, and M_3), and PMOS devices (M_4, M_5, M_6) form the FVF. Transistor M_7 works as a reference current generator. (M_1, M_4, M_8) is a FVF circuit.

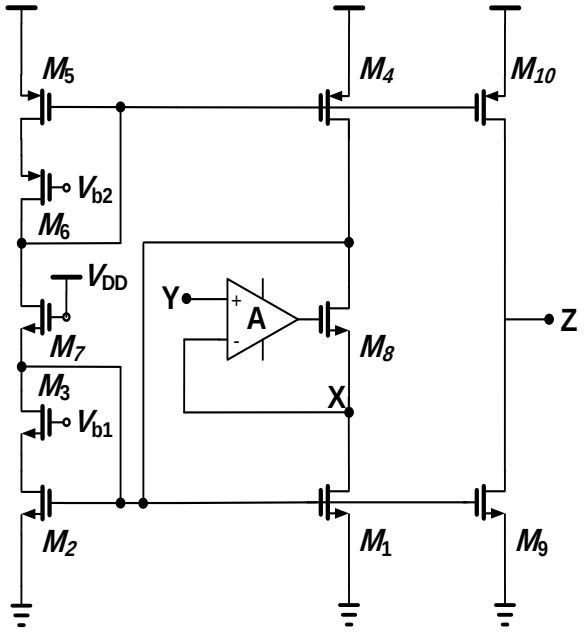


Fig. 3: The proposed CCII based on the OTA.

The choice of using a FVF in the proposed CCII offers several key advantages over traditional configurations.

1- Low Impedance: The FVF structure inherently provides low output impedance, which is crucial for ensuring proper current transfer and minimizing signal distortion in current-mode circuits. This low impedance directly contributes to the improved performance of the CCII, especially in driving high-frequency loads.

2- Enhanced Voltage Tracking: The FVF configuration helps in accurate voltage tracking by reducing the mismatch between the input and output voltages. This is particularly beneficial for achieving the precise replication of the input voltage in the CCII structure.

3- Impact on Performance: By using FVF, we can significantly enhance the performance in terms of bandwidth and DC gain. The low impedance provided by the FVF ensures that the circuit can operate at higher speeds and with greater precision, resulting in an overall improvement in the CCII's voltage and current tracking capabilities.

The operational amplifier with gain A is the amplifier shown in Fig. 1, which is used for three purposes. First, the voltage of nodes X and Y should be close to each other. Second, the resistance at the X node should be reduced. Finally, the resistance at the Y is high. The transistors (M_9, M_{10}) are used in CCII output.

From Fig. 3, it can be seen that the input impedance of the Y node is infinite. The output impedance of node X is equal to:

$$R_x = \frac{1}{A_d g_{m1} g_{m8} (r_{ds4} \| r_{ds8})} \quad (9)$$

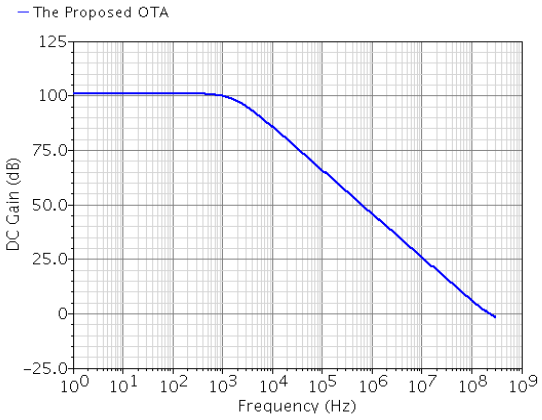
The above equation shows that as the Ad increases, the resistance at node X decreases. The output resistance in node Z is obtained from the following equation.

$$R_z = r_{ds9} \| r_{ds10} \quad (10)$$

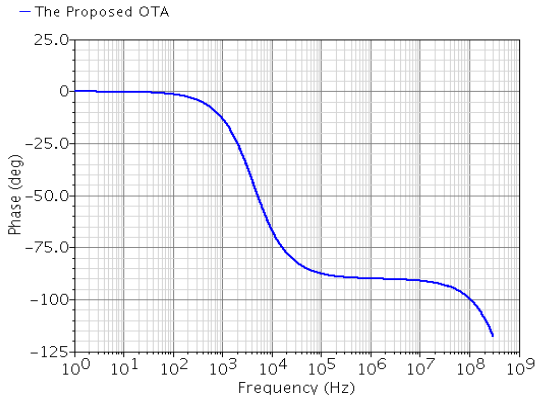
Simulation Results

Both the presented OTA and CCII are simulated in a standard 180-nm CMOS process with a supply voltage of 1.8 V. The component values are similar to those used in the OTA presented in [17]. For the CCII, the dimensions of the transistors are as follows: the NMOS has dimensions of 1.5µm / 0.18µm, and the PMOS has dimensions of 3.6µm / 0.18µm. Fig. 4 illustrates the open-loop frequency responses of the OTA. The DC voltage gain is measured at 101 dB, representing an enhancement of 7 dB over the traditional structure. The OTA's unity-gain bandwidth (UGBW) is 230 MHz, and its phase margin is 61°, indicating circuit stability.

Monte-Carlo (MC) simulations of the OTA were conducted to assess process and mismatch variations. Fig. 5 shows the MC histograms of the designed amplifier based on 1000 simulation runs. The mean and standard deviation values for the DC voltage gain are 104.2 dB and 7.4 dB, respectively. Similarly, the mean and standard deviation values for the phase margin are obtained 66.6 and 1.1 degrees, respectively.

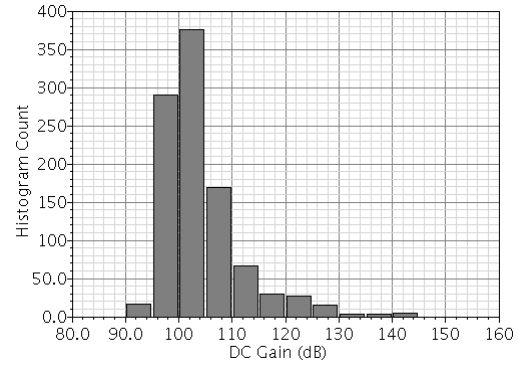


(a)

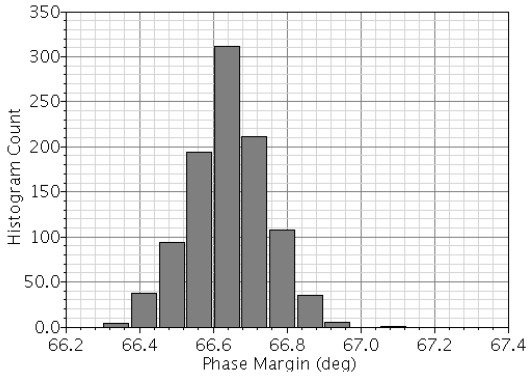


(b)

Fig. 4: The open-loop frequency response of the designed OTA: (a) DC voltage gain, (b) phase response across the frequency range.



(a)



(b)

Fig. 5: Histogram of Monte Carlo (MC) simulation results for the proposed OTA: (a) DC voltage gain, (b) phase margin.

Table 1: A comparative analysis of the developed amplifier's performance versus existing design, highlighting the improvement in voltage gain

Parameter	This work	[17]	[19]
Technology (nm)	180	180	180
Supply Voltage (V)	1.8	1.8	1.8
DC Voltage Gain (dB)	101	93	84
Input-Referred Noise@100kHz $(\mu v/\sqrt{Hz})$	0.32	0.31	0.34
Silicon Area (mm²)	0.022	0.021	0.07
Differential Output Swing (V_{pp})	2.8	2.8	2.8
Phase Margin (^{o})	62	65	77
Power Dissipation (mW)	2.8	2.8	3.1
SR (V/µs)	650	494	63
UGBW (MHz)	230	216	91
Load Capacitor (pF)	1	1	100
Operating Region	Strong Inversion	Strong Inversion	Strong Inversion
FOM_S (MHz×pF/mA)	147	138	5290
FOM_L (V×pF/µs×mA)	417	316	3660

Table 1 presents the post-layout simulation results for the designed OTA and its competitor. The proposed OTA exhibits a higher DC voltage gain compared to the other solution, while maintaining comparable parameters such as UGBW, power consumption, and stability. To evaluate the relative performance of the two competitors, the figures of merit described in (11) and (12) are utilized [22]-[30].

$$FOM_S = \frac{UGBW\ C_L}{I_T} \quad (11)$$

$$FOM_L = \frac{SR\ C_L}{I_T} \quad (12)$$

where I_T is the total circuit current and C_L is the load capacitor. From Table 1, it can be concluded that the figure of merit of the presented OTA is better.

The total power consumption of the proposed CCII is 3.1 mW.

Fig. 6 illustrates the input voltage tracking of the CCII, with an error of 0.3%. Fig. 7 depicts the output current tracking I_z/I_x, with an error of 0.1% across process and temperature corners. The bandwidth in TT (27°C), FF (-40°C), and SS (90°C) is 1.1 GHz, 1.7 GHz, and 1.1 GHz, respectively.

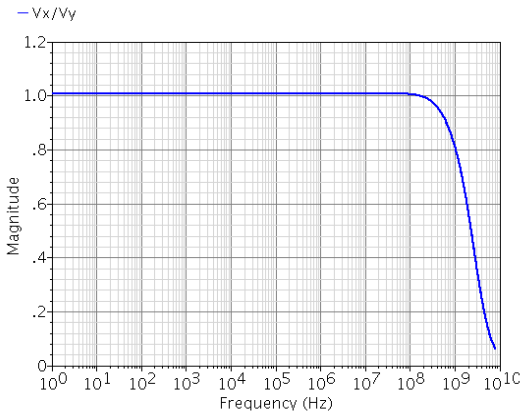


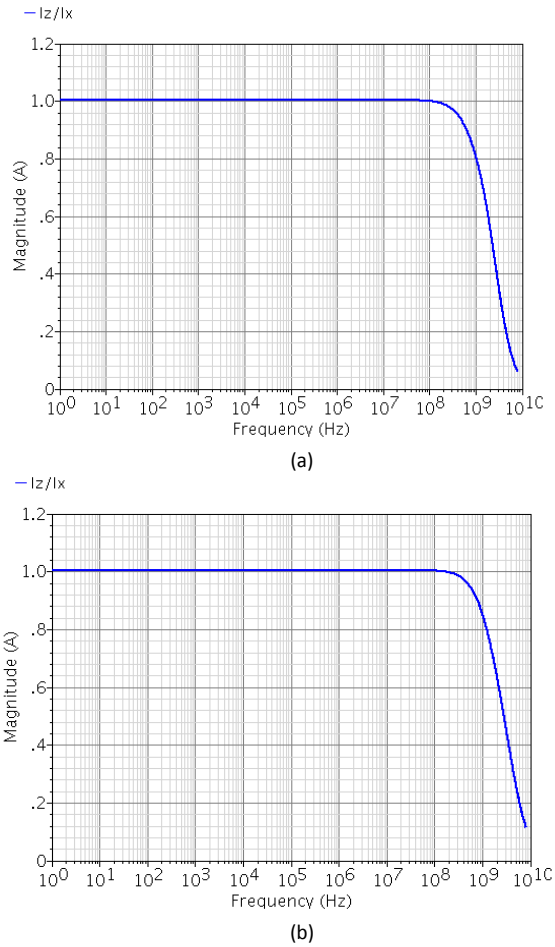
Fig. 6: Input terminals voltage tracking V_x/V_y. This figure illustrates the accuracy of voltage transfer between the terminal V_x and the terminal V_y.

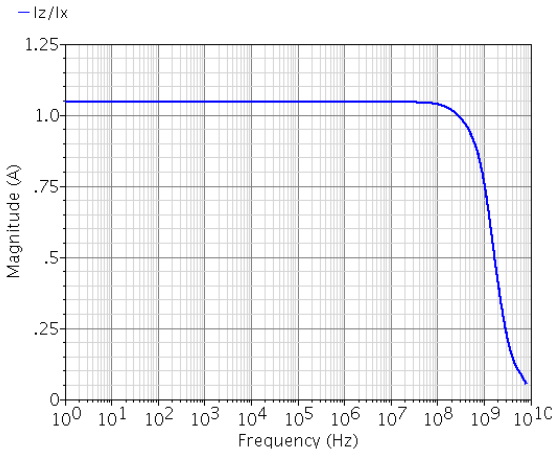
Table 2 presents a performance comparison between the designed CCII and one existing method. The results clearly demonstrate that the proposed method exhibits lower current and voltage tracking errors compared to existing approaches. Additionally, the proposed method offers a wider bandwidth than the existing methods. In [5] and [9], conventional methods are used in the CCII design to achieve certain performance metrics. However, the absence of positive feedback and innovative approaches in the OTA design leads to limitations in DC gain and other performance characteristics. In [18], a FVF is utilized, which affects the operating voltage. Compared to our design, this approach may not provide the optimal DC performance.

Table 2: Comparison of the performance between the developed design and existing similar works

References	This work	[5]	[9]	[18]
Process (nm)	180	130	90	500
Supply Voltage (V)	1.8	1.2	1.2	1.5
Input voltage tracking error (%)	0.3	0.%	-	-
Output current tracking error (%)	0.1	0.5	-	-
Bandwidth (GHz)	1.5	1.2	0.104	0.1
Power Consumption (mW)	3.1	1.5	3	0.6

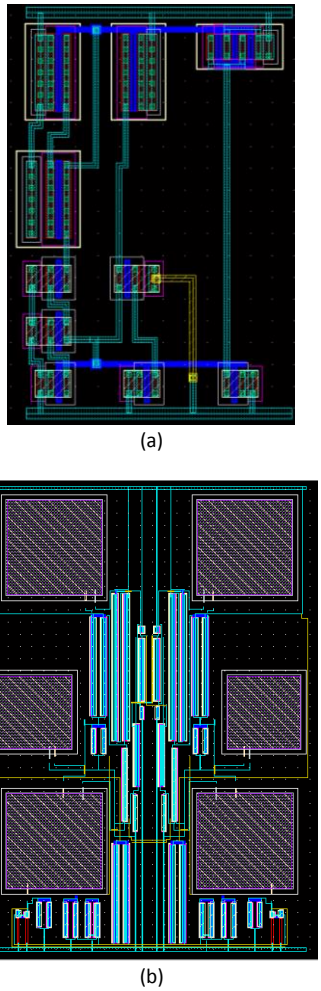
The layout view of the designed circuit is shown in Fig. 8. The physical dimension of the designed CCII is 12 µm x 17 µm, and the dimension of the designed OTA is 120 µm x 180 µm. These details provide a clearer understanding of the physical implementation and area utilization of the proposed design.





(c)

Fig. 7: Outputs current tracking I_z/I_x in different process corners and temperatures: (a) TT (27°C), (b) FF (-40°C), (c) SS (90°C). This figure shows the accuracy of current transfer in different operating scenarios.



(a)

(b)

Fig. 8: Layout view of the proposed design: (a) Proposed CCII, (b) Proposed OTA.

Conclusion

Current conveyors are essential components in many current-mode circuits. This paper presents a novel OTA and CCII designed in a 180-nm CMOS process. The

presented circuits operate at a supply voltage of 1.8 V. The designed CCII employs a FVF to achieve low impedance at node X. A new OTA is introduced to accurately copy the voltage within the circuit. The proposed OTA utilizes positive feedback to increase its output resistance, resulting in improved DC voltage gain. Simulations were conducted to evaluate the performance of the designed CCII. The DC voltage gain of the OTA was increased by approximately 7 dB compared to its competitor without affecting specifications such as power consumption, UGBW, and stability.

The current and voltage tracking errors in the proposed method are lower than those of existing methods. The proposed CCII design, with its low impedance and accurate voltage copying features, can be utilized in filter circuits, analog and digital signal processing, and signal amplification in telecommunications systems.

These features ensure that signals are transmitted with minimal attenuation and distortion, improving overall system performance.

Author Contributions

Conceptualization and design, E. Tavassoli; formal analysis, E. Tavassoli; software, E. Tavassoli; investigation, S. M. Anisheh.; writing—original draft preparation, E. Tavasolli; writing—review and editing, S. M. Anisheh. supervision, S. M. Anisheh, and M. Radmehr.

Conflict of Interest

The authors have disclosed that there are no potential conflicts of interest related to the publication of this work.

Acknowledgment

The authors would like to also express their gratitude to Alireza Ghorbani for his technical assistance and support, and for useful discussions and hints.

Abbreviations

MC	Monte-Carlo
OTA	Operational transconductance amplifier
CCII	Second generation current conveyor
FVF	Flipped voltage follower
SR	Slew rate
FD-OFC	Fully differential operational floating conveyor
OFCC	Operational floating current conveyor
NLCM	Nonlinear current mirrors
UGBW	Unity-gain bandwidth

References

[1] K. C. Smith, A. Sedra, "The current conveyor—A new circuit building block," Proc. IEEE, 56(8): 1368-1369, 1968.

[2] G. Yun et al., "An ultrasound receiver with bandwidth-enhanced current conveyor and element-level ultrasound transmitter for ultrasound imaging systems," IEEE Solid-State Circuits Lett., 7: 98-101, 2024.

[3] M. Kumngern, F. Khateb, T. Kulej, "Low-voltage low-power differential difference current conveyor transconductance amplifier and its application to a versatile analog filter," IEEE Access, 12: 92523-92535, 2024.

[4] N. M. Edward, Y. H. Ghallab, H. Mostafa, Y. I. Ismail, "A CMOS based operational floating current conveyor and its applications," in Proc. IEEE International Conference on Electronics, Circuits, and Systems (ICECS): 494-495, 2015.

[5] N. M. Edward, Y. H. Ghallab, H. Mostafa, Y. I. Ismail, "A CMOS based operational floating current conveyor," in Proc. IEEE International Conference on Electronics, Circuits, and Systems (ICECS), 2015.

[6] G. Palmisano, G. Palumbo, "A simple cmos CCII+," Int. J. Circuit Theory Appl., 23(6): 599-603, 1995.

[7] H. M. Hassan, A. M. Soliman, "Novel cmos realizations of the operational floating conveyor and applications," J. Circuits Syst. Comput., 14(6): 1113-1143, 2005.

[8] D. Nand, N. Pandey, R. Pandey, S. Jindal, "Operational floating current conveyor based digitally controlled hearing aid," in Proc. 5th International Conference on Signal Processing and Integrated Networks (SPIN): 647-651, 2018.

[9] F. Elsayed, M. Rashdan, M. Salman "CMOS operational floating current conveyor circuit for instrumentation amplifier application," Int. J. Electr. Electron. Eng. Telecommun., 9(5): 317-323, 2020.

[10] N. Pandey, P. Tripathi, R. Pandey, R. Batra, "OFCC based logarithmic amplifier," in Proc. International Conference on Signal Processing and Integrated Networks (SPIN): 522-525, 2014.

[11] R. S. Assaad, J. Silva-Martinez, "The recycling folded cascode: A general enhancement of the folded cascode amplifier," IEEE J. Solid-State Circuits, 44(9): 2535-2542, 2009.

[12] N. Ning, F. Yang, S. Zhiling, L. Rui, W. Shuangyi, "A low-sensitivity negative resistance load fully differential OTA under low voltage 40nm CMOS logic process," Chinese scientific papers: 1-7, 2010.

[13] A. Dabas, S. Kumari, M. Gupta, R. Yadav, "Design and analysis of class AB RFC OTA with improved performance," in Proc. 8th International Conference on Signal Processing and Communication (ICSC): 582-586, 2022.

[14] M. Anand, Anushree, J. Dhanoa, "Gain and gain-bandwidth enhancement of recyclic folded cascode OTA using floating voltage source," in Proc. IEEE Delhi Section Conference (DELCON): 1-5, 2022.

[15] C. Taol, L. Lei, Z. Chen, Z. Hong, Y. Huang, "A 50MHz bandwidth TIA based on two stage pseudo-differential OTA with cascode negative resistance and R-C compensation technique," in Proc. IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS): 1-4, 2022.

[16] S. Adibi, R. Rubino, P. Toledo, and P. Crovetti, "Design of an analog and of a digital-based OTA in flexible integrated circuit technology," in Proc. 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS): 1-4, 2022.

[17] S. M. Anisheh, H. Shamsi, "Two-stage class-AB OTA with enhanced DC gain and slew rate," Int. J. Electron. Lett., 5(4): 438-448, 2017.

[18] R. G. Carvajal et al., "The flipped voltage follower: a useful cell for low-voltage low-power circuit design," IEEE Trans. Circuits Syst. I Regul. Pap., 52(7): 1276-1291, 2005.

[19] S. M. Anisheh, H. Abbasizadeh, H. Shamsi, C. Dadkhah, K. Y. Lee, "84 dB DC-gain two-stage class-AB OTA," IET Circuits Devices Syst., 13(5): 614-621, 2019.

[20] S. M. Anisheh, H. Abbasizadeh, H. Shamsi, C. Dadkhah, K. Y. Lee, "98-dB gain class-AB OTA with 100 pF load capacitor in 180-nm digital CMOS process," IEEE Access,7: 17772-17779, 2019.

[21] S. M. Anisheh, H. Shamsi, M. Mirhassani, "Positive feedback technique and split-length transistors for DC-gain enhancement of two-stage op-amps," IET Circuits Devices Syst., 11(6): 605-612, 2017.

[22] K. P. Ho, C. F. Chan, C. S. Choy, K. P. Pun, "Reversed nested Miller compensation with voltage buffer and nulling resistor," IEEE J. Solid-State Circuits, 38(10): 1735-1738, 2003.

[23] A. D. Grasso, G. Palumbo, S. Pennisi, "Advances in reversed nested miller compensation," IEEE Trans. Circuits Syst. I, Regul. Pap., 54(7): 1459-1470, 2007.

[24] V. Michal, "OTA slew-rate and bandwidth enhancement based on dynamic input-overdriven current mirror," in Proc. International Conference on Applied Electronics (AE): 1-4, 2023.

[25] P. Li, L. Luo, Q. Wei, "A CLASS AB operational transconductance amplifier (OTA) with a slew rate enhancement (SRE) auxiliary circuit," in Proc. 3rd International Conference on Communication Technology and Information Technology (ICCTIT): 20-23, 2023.

[26] S. L. Tripathi, B. Raj, S. Verma, V. Narula, S. Saxena, "High gain multistage CMOS amplifier design at 45nm technology node," in Proc. IEEE Devices for Integrated Circuit (DevIC): 237-242, 2023.

[27] S. Choudhary, R. Chowdhury, T. Sharma, "Comparative analysis of operational transconductance amplifiers (OTA) for biomedical applications," in Proc. International Conference on Circuit Power and Computing Technologies (ICCPCT): 1330-1334, 2023.

[28] R. Jayachandran, E. B. Kumaradhas, A. S, "Design of second order low pass filter using inverter-based CMOS operational transconductance amplifier (OTA) through gm/ID methodology," in Proc. IEEE International Conference of Electron Devices Society Kolkata Chapter (EDKCON): 519-524, 2024.

[29] S. Nasabi, S. S. Kotbagi, S. Javalagi, M. M, R. R. Patgar, B. A. Gunhalkar, "Design of operational transconductance amplifier (OTA) for 1.2V bandgap reference (BGR)," in Proc. 4th International Conference on Intelligent Technologies (CONIT): 1-6, 2024.

[30] S. Dutta, R. Ranjan, P. Kumari, P. Sinha, R. K. Ranjan, D. K. Singh, "Operational transconductance amplifier (OTA) based dual mode Multiphase Oscillator," in Proc. 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST): 332-337, 2024.

Biographies

Emad Tavassoli received the M.S degree in Electrical Engineering (Electronic) from Islamic Azad University Science and Research Branch Tehran, Iran in 2013. He is currently a student of Electronic Engineering at Islamic Azad University, Sari branch. His research interests include bilayer graphene, differential operational floating conveyors.

- Email: emadtavassoli@gmail.com
- ORCID: 0009-0008-2515-4731
- Web of Science Researcher ID: MFK-2999-2025
- Scopus Author ID: NA
- Homepage: NA

Seyed Mahmoud Anisheh received the Ph.D. degree in Electrical Engineering (Electronics) from K.N. Toosi University of Technology, Tehran, Iran in 2017. His research interests include high performance Op-Amps, automatic analog circuit layout generation, RFIC, and signal processing.

- Email: s.m.anisheh@ee.kntu.ac.ir
- ORCID: 0000-0003-0982-8608
- Web of Science Researcher ID: MFK-3141-2025
- Scopus Author ID: NA
- Homepage: NA

Mehdi Radmehr was born in 1974 and received the B.Sc., M.Sc., and Ph.D. degrees in Electrical Engineering from University of Tehran, Tarbiat Modares, and Islamic Azad University, Science and Research campus, Tehran, Iran, in 1996, 1998, and 2006 respectively. He is a specializing in power electronics, motor drives and power quality. He has worked for Mazandaran Wood and Paper Industries as an advisor since 1997 before starting his Ph.D. study. He has joined the scientific staff of Islamic Azad University, Sari branch since 1998.

- Email: maradmehr@gmail.com
- ORCID: 0000-0003-1678-9758
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
E. Tavassoli, S. M. Anisheh, M. Radmehr, "A second generation current conveyor employing a flipped voltage follower and improved DC voltage gain operational transconductance amplifier," J. Electr. Comput. Eng. Innovations, 13(2): 455-462, 2025.

DOI: 10.22061/jecei.2025.11326.795

URL: https://jecei.sru.ac.ir/article_2307.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Implementing Yosys & OpenROAD for Physical Design (PD) of an IoT Device for Vehicle Detection via ASAP7 PDK

S. H. Rakib [1,*], ***S. N. Biswas*** [2]

[1]*Physical Design Engineer, Neural Semiconductor Limited, Dhaka, Bangladesh.*

[2]*Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.*

Article Info

Article History:
Received 12 December 2024
Reviewed 10 February 2025
Revised 07 March 2025
Accepted 16 March 2025

Keywords:
VLSI
IoT
Synthesis
Physical Implementation
Setup and Hold Timing

*Corresponding Author's Email Address:
s.h.rakib.153@gmail.com

Abstract

Background and Objectives: The automobile industry is becoming more technologically advanced. Modern vehicles are expensive, but they have cutting-edge security features. As a result, the average individual who can afford low-end vehicles must forego the latest improvements, such as greater safety. Therefore, the main goal was to create a small Internet of Things device that could be used on a mobile device to notify the user when a car comes from the opposite direction. It will promote human safety by alerting users to that vehicle. The preparation, integration, and deployment of a modern IoT-based vehicle detection device have been described in this work. From there, it goes through the OpenROAD toolchain, and OpenSTA is used for static timing analysis (STA) and the ASAP7 PDK is used for the design. In this paper, provide a performance evaluation study across all three metrics (power, performance and area), as well as the entire design flow from hardware description to final implementation.
Methods: The entire behavioral level code goes through several stages before reaching a physical perspective, where various tools are used for multiple tasks. To obtain the desired physical-level architecture, first, use a tool to obtain the netlist file, including every G-cell map with a PDK-specific gate-level representation. Then, several stages will be followed to get the device's physical view.
Results: Throughout the entire experiment, the transition from RTL to GDSII was successfully achieved. Once the complete design is finished, the area, power, and timings all appear fine. Another unique characteristic is that the chip employed 7nm technology. The 5 GHz frequency was attained when the chip functioned flawlessly without DRC or any connection problems, timing, or DRV violations. Less than 1 percent is the maximum allowable IR loss maintained. Over 80% of the total space was utilized effectively.
Conclusion: To build an IoT gadget manufacturable with the best PPA, the general experiment was to write RTL code and proceed to the tap-out stage. The experiment achieved the best result by utilizing open-source chip design tools. Additionally, there are no DRC violations, timing problems, or power loss.

Introduction

It was Kevin Ashton who first introduced the Internet of Things or IoT. All of the physical objects in the environment that communicates with one another via the internet is referred to as IoT systems [1].

According to several industry projections, there will be approximately 50 billion smart devices linked to the Internet of Things (IoT) by 2030. These devices will aid in the development of novel solutions for issues affecting society as a whole, including telemetry, healthcare, home automation, energy conservation, security, wearable computing, asset tracking, public infrastructure maintenance, etc. [2].

The Association for Safe International Road Travel (ASIRT) estimates that 20–50 million people are injured or incapacitated and that approximately 1.3 million people die in traffic accidents annually. Worldwide, traffic accidents cost $518 billion, or 1% to 2% of each nation's yearly GDP [3]. As one of the biggest causes of death, road accidents claim 1.3 million lives annually. Therefore, vehicle detection technology may be a useful way to lower traffic accidents and improve public safety [4].

This IoT device was introduced to assist in decreasing accidents by warning drivers in advance. However, this gadget will be upgraded in the future to include the capability of providing a prompt alert to the nearby rescue team and ensuring prompt treatment to save the life of the passenger in the event of an accident.

The Internet of Things (IoT) is a computer process in which every physical thing has sensors, microcontrollers, and transceivers for enabling communication. It also has proper protocol stacks built in to enable the objects to communicate with one another and with users [5].

Various techniques can be employed for vehicle detection. When it comes to reducing accidents or promoting automated vehicles, this detecting procedure is essential. The statistical approach was one of the strategies. Statistical techniques effectively integrate activity data with sophisticated classify arithmetic knowledge about vehicle flight patterns. This method effectively locates and tracks the cars [6].

On numerous fronts, open-source EDA is quickly facilitating new waves of innovation. It expedites the scientific process and makes study findings applicable to contemporary business practices, according to academic scholars. Open-source EDA is a supplement and enhancer of commercial EDA for EDA experts and the industry ecosystem [7].

This device is based on real-time data capturing and is MCU-powered. OpenROAD flow and the Yosys, OpenROAD, and OpenSTA tools are utilized in the design of this [8]. A fully autonomous RTL-GDSII flow for quick architectural and design space exploration, early QoR prediction, and thorough physical design implementation is called OpenROAD-flow-scripts (ORFS) [9].

Select open-source tools because they are user-friendly, widely available, and compliant with industry standards. This project's complete ASIC flow is shown next.

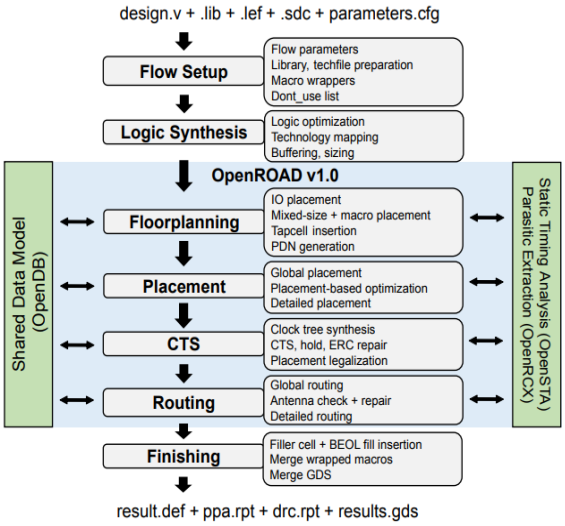


Fig. 1: An overview of the OpenROAD tool's implementation of the OpenROAD design flow.

The design flow from RTL to the final physical design using OpenROAD, help understand the whole design process covered in the study.

In the summer of 2020, OpenROAD fulfilled many "proof points" enabling the automated construction of a manufacturing layout in TSMC 65LP and GLOBALFOUNDRIES 12LP technologies, including a 12nm SOC tape-in. These "proof points" included passing all physical verification checks as well as electrical and timing correctness checks [10].

The selection of Yosys was based on its wide range of features, broad support for Verilog-2005, and ability to map to any standard cell library used in ASICs. The goal of OpenROAD usage in this design is to lower the obstacles that now prevent designers from implementing innovative technologies on hardware, including those related to cost, skill, and unpredictability [11].

A detailed analysis of power, time, and area wraps up the whole flow. Make use of 7nm PDK. To ensure that the design operates flawlessly, choose the best achievable frequency. And appropriately set all the restraints. The Objectives Provide a comprehensive system design and support the deployment of high-performing IoT endpoints for traffic safety.

Technology is getting smaller day by day. Additionally, since smaller technology uses less space, it can accommodate a larger, more complicated design in a smaller space. And because of its tiny size, the design uses less power overall. That's why 7nm PDK was selected for the design, as it was the smallest technology available on OpenRoad Flow [12].

ASAP7 is an Advanced-Node Research PDK that is open-source. OpenROAD has also made the ASAP7 advanced-node research PDK from Arizona State University publicly available to supplement the SKY130 open-source manufacturable PDK [13]. Advanced

patterning technologies and scaling boosters (single diffusion break, contact-over-active-gate, dense crossovers, etc.) are reflected in the design principles of this incredibly realistic PDK [10].

First, strive to reach the optimal frequency that can accept the design without any violations. Power consumption is also given significant consideration, as it represents a major challenge for Internet of Things (IoT) devices. Since most IoT devices are portable charging systems that do not connect directly to a charging system, the primary goal when designing IoT devices is to minimize power consumption and maximize speed while maintaining functionality.

The flow's default units are Time: 1ps; capacitance: 1fF; voltage: 1v; power: 1pW; distance: 1um.

Device Structural Overview

Based on information from the neighboring mobile network about whether or not any cars are approaching me, the MCU in this device manages network connectivity, sensor data, and the generation of the clock, reset, and data input signals for the IoT device. An LED display cannot display unwanted data until a reset signal is received. A clock signal is required for the synchronous operation of an Internet of Things device. Data from the MCU is saved by an IoT device, which also manages system operation by displaying it on an LED display. Manage the display_enable switch as well.

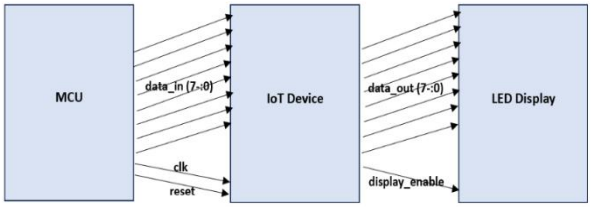


Fig. 2: Structural overview of the device, highlighting its key components and their interconnections.

The architecture is described before the detailed design begins which shows the structure of the IoT device also how the components are connected.

The entire design flow is divided into three sections:

- RTL Design
- RTL to Gate Level Netlist Conversion
- Physical Implementation

The SDC file contains design constraints that are specified following the design's operating frequency of 5GHz. As an example, the clock uncertainty is set to 20% of the clock period, the IO delay is set to 30%, and the maximum transition for both the data path and the clock path is set to 10%. As opposed to the typical setting of 5 to 15% of the clock duration. It's probably set at 5 to 10 for the clock path and 10 to 15% for the data paths [14]. The transition time of a CMOS gate is known to have a significant impact on its performance, including its propagation delay time and short circuit power dissipation [15].

Verilog Module

For the IoT gadget, here is the code for functional work.

```
module iot_device (
    input wire      clk,          // Clock signal
    input wire      reset,        // Reset signal
    input wire [7:0] data_in,     // Input data from network/microcontroller
    output reg [7:0] data_out,    // Output data to display
    output reg      display_enable // Enable signal for display
);

    // State register
    reg [7:0] internal_data;

    // Control logic
    always @(posedge clk or posedge reset) begin
        if (reset) begin
            internal_data   <= 8'b0;
            data_out        <= 8'b0;
            display_enable  <= 1'b0;
        end else begin
            internal_data   <= data_in;
            data_out        <= internal_data;
            display_enable  <= 1'b1;  // Enable display when not in reset
        end
    end

endmodule
```

Fig. 3: Verilog code for implemented IoT device.

Physical design starts with the Verilog code for the IoT device, which represents the logic of the device.

Synthesis with Yosys

Yosys, a platform for managing and storing data, was first used to synthesize the Verilog of the Hardware device (IoT) into a gate-level netlist [16]. Additionally, Yosys provides us with a glimpse of the device schematic.

Following synthesis using the ASAP7 PDK and this is the Yosys report.

```
Number of wires:                 32
Number of wire bits:             46
Number of public wires:          13
Number of public wire bits:      27
Number of ports:                  5
Number of port bits:             19
Number of memories:               0
Number of memory bits:            0
Number of processes:              0
Number of cells:                 36
  DFFASRHQNx1_ASAP7_75t_R        17
  INVx3_ASAP7_75t_R              18
  TIEHIx1_ASAP7_75t_R             1

Chip area for module '\iot_device': 7.800300
  of which used for sequential elements: 6.444360 (82.62%)
```

Fig. 4: Report of gate level netlist.

Physical Design with OpenROAD

For getting a concrete view of the design; several physical implementation phases need to be run each having different target and tasks corresponding the gate-level netlist and generated SDC from Yosys output.

A. Floor Planning

The overall goal of floor planning is to reduce the design's length and overall area [17]. The chip's area and shape must be determined at this point. Therefore, there

needs to be a few basic activities carried out, such as setting the block's core area, die area, utilization, and aspect ratio as well as positioning the physical cell, macros, and IO port on the appropriate edge.

According to this design, the floorplan should begin at 50% utilization and terminate at 55%. However, a 75% utilization rate is initially adopted in real projects [18].

For this design, the die area and core areas are set at (0.0 0.0 4.5 4.5) and (0.108 0.27 4.374 4.32). Port placement is done using Metals 4 and 5, respectively, for the vertical and horizontal layers.

B. Power Planning

An integrated circuit system's physical PDN extends its hierarchy across multiple tiers. The voltage regulator module (VRM), which is essentially a DC-DC converter that raises input DC voltage to the nominal supply voltage level as needed by the IC chip, provides power to the network [19].

In order to allocate power to every component in the design—including standard cells and macros—power planning is required.

```
add_pdn_stripe -grid {top} -layer {M1} -width {0.018} -pitch {0.54} -offset {0} -followpins
add_pdn_stripe -grid {top} -layer {M2} -width {0.018} -pitch {0.54} -offset {0}
add_pdn_stripe -grid {top} -layer {M5} -width {0.12} -spacing {0.072} -pitch {0.75} -offset {0.13}
```

The top layer in this block is made of metal 5, the intermediate layer is made of metal 2, and the special route is made of Metal1.

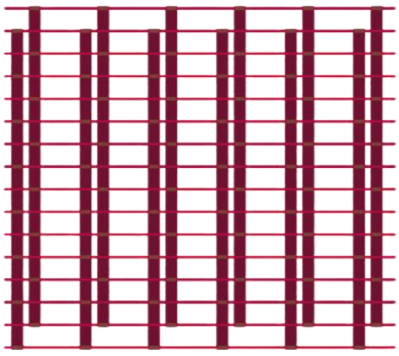


Fig. 5: Illustration of the correct power grid distribution to ensure proper power delivery to the cell.

The power grid distribution sends adequate power to all regions of the design, which is an important factor to take into account for the device's reliability.

C. Placement

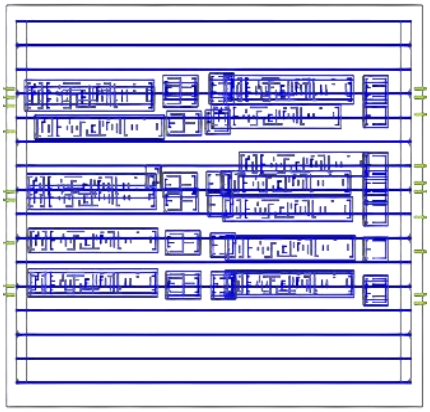
The goal of placement is to handle optimization goals like HPWL, routed wire length, time, power, routing, etc. while figuring out where to put instances (such as standard cells and macros) [20].

Some problems can be resolved during the placement phase; for example, hard blockage can be used to reduce notch congestion, partial blockage can be used to address issues with cell density, and padding can be used to address issues with pin density [21]. Additionally, it is necessary to verify utilization; if it deviates significantly from the floorplan utilization, debug it to determine the cause. After the placement stage, the DRV and setup time must also be checked.

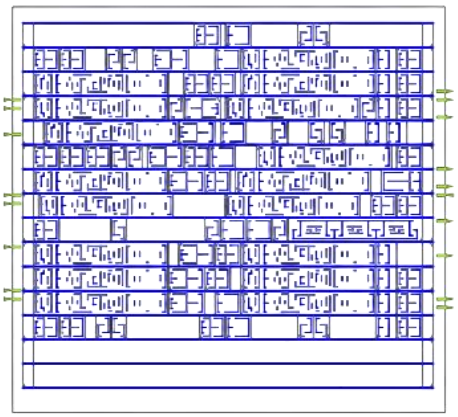
Proper placement of the standard cell will reduce the likelihood of timing violations, power outages, or physical violations. Proper placement in the core area will facilitate easier connections between ports and cells, and if the cell is placed in a scattered configuration, power outages will not arise during design. Also, in the absence of congestion, there won't be any problems with a lack of tracks, which was one of the causes of the short violation. Finished the placement stage since it has a big impact on the design and doesn't throw off the sign-off process.

The placement process consists of two basic steps: global placement and detailed placement [22].

The first tool uses global placement, which places cells that are genuinely existent in the netlist without adhering to any legality rules when doing so. Not positioned correctly in the row either. There are also overlaps between them. Subsequently, the tool places each cell precisely, adhering to the placement guidelines, so that they don't overlap, and placing them in between rows.



(a)

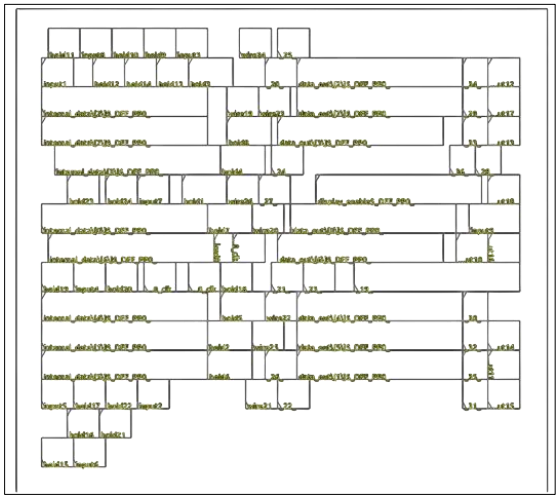


(b)

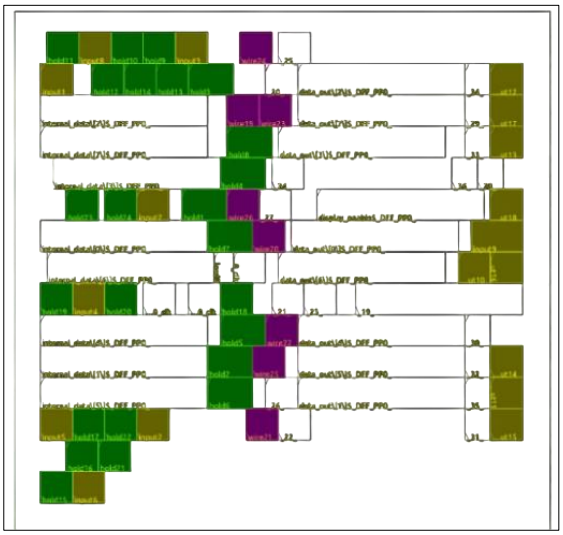
Fig. 6: (a) Incorrect cell placement after global placement, showing overlapping cells and misalignment. (b) Corrected detailed placement, ensuring proper cell alignment within designated rows.

A comparison of incorrect and corrected cell placements highlights the importance of proper cell placement for efficient routing and performance.

Place every cell closer together at this placement stage in order to fulfill the timing requirement as well. Then, by resizing the cell, optimization is carried out to decrease the slack.



(a)



(b)

Fig. 7: a) place all cells in the design listed in the netlist file; b) highlighting the cells that were optimized to meet the timing constraints. (Shown in violet, yellow, and green).

The placement of all cells and optimized cells for timing shows how cell placement affects timing performance.

Table 1: The report following placement details are included in the table

Parameter	Value
Instances	117
Nets	97
Core Area	17.277 um^2
Place Instances Area	11.474 um^2
Utilization	69.96%

D. Clock Tree Synthesis

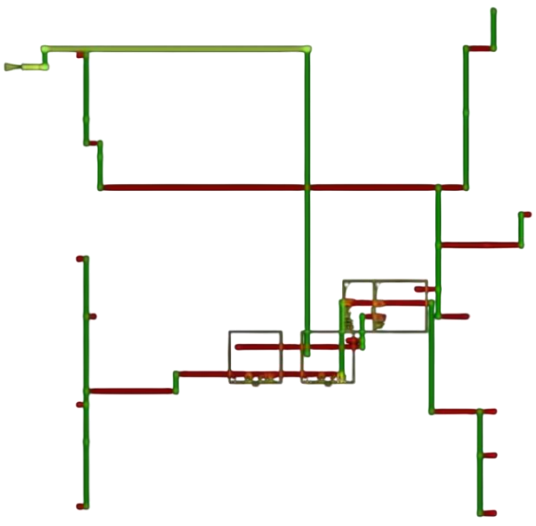
The movement and processing of data inside a chip are coordinated and controlled by a clock signal [23]. The clock port to every clock pin on the flop is connected by a clock tree.

A clock signal on the flop is required for design operation. Therefore, the tree must be balanced because all of the flops may run simultaneously. Minimize the clock skew to balance the tree. The time disparity between the arrival times of the clock signal at each different flip is known as clock skew. Use an inverter or buffer at this point to balance the clock tree.

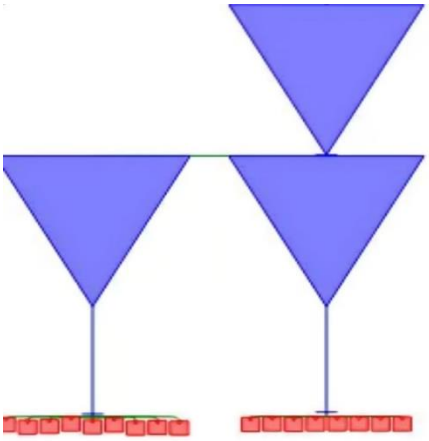
Since the clock is the most important component of a synchronous system design, all flops must be synchronized for the design to operate flawlessly and shift and store any data [24]. CTS buffers or inverters are used to balance clock trees since they require a clock that is always on. If the clock path's wire width is not increased, cross-talk will result. Consequently, use NDR on the clock net. When creating clock trees, use the CTS buffer and inverter cell since their rise and fall times are the same. In addition, the wire length is affected by the location of the Flop during the installation step. It also impacts the setup and hold times.

The addition of positive clock skew helps with design setup violations by increasing the time it takes for data to get from the launch flop to the capture flop [25]. On the other hand, skew added to the data would impact the difficult need for a more stable time and might result in design metastability, which will complicate the hold analysis. A cluster group is made up of virtually identical sorts of insertion delay flops.

The software also makes an effort to create cluster groups based on the given skew and utilizes clustering to aid in tree balancing.



(a)



(b)

Fig. 8: Illustrations of: a) A clock port on the left boundary connecting all sequential cells b) A debugger window identifying problematic sequential cell groups with similar skew for timing issue resolution.

It is essential to fix timing issues, and the port clocks connecting sequential cells, and the debugger indicating timing concerns, play a crucial role in this scenario.

17 DFF flip-flops are linked to the clock port, as seen in the above image. Additionally, it can be seen that the design includes 17 DFF from the synthesis report.

E. Routing

Routing is the technique of creating a physical link between each instance by utilizing the metal layer to connect all of the instances with ports or instance to instance.

Prior to estimating the parasitic value, the tool completed global routing. In order to maintain the continuity of the N-well, a filler cell is also placed on this step. Once placed DEF is inserted, the tool reads LEF.

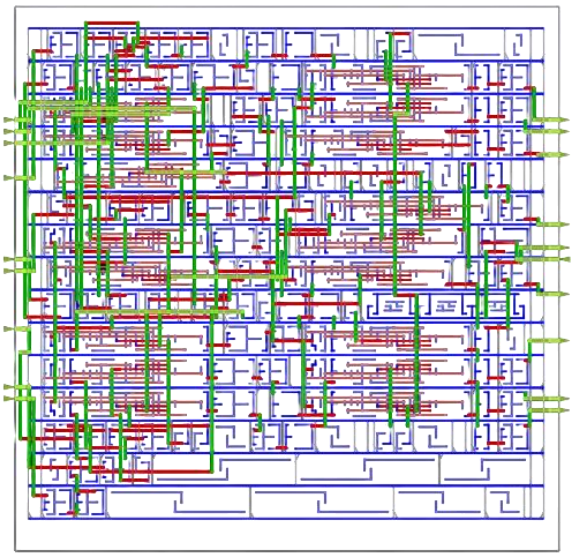


Fig. 9: Signal routing indicating physical connections between all the communicating objects within the design.

Signal routing also has the critical task of routing the signals between all the object in the design such that they are connected correctly in the end layout.

In order to accomplish global routing, it defines global routing cells, or gcells, and minimizes wire length and vias while limiting congestion and overflow within the cells [9].

Afterward, the tool finishes the detailed routing as efficiently as feasible, making use of the routing truck and being aware of the basic DRC rule of the design.

Result and Discussion

There are three primary (PPA) focuses in this study. First is performance, often known as timing, in a VLSI digital design. Next are power and area. The battery lifetime and manufacturing cost will be impacted by increased power and area. However, if the timing is right, the device will function as intended.

Moreover, the design will cease to function if the timing is violated. Primarily focused on using this concept to solve the time issue. Furthermore, this design must not have any time violations or negative slack. Contrasted with industry-level comparisons. With less than 1% IR loss, area utilization is also about 80% without filler, according to the industry level.

The tool checked the design rules after finishing the detailed route, attempted to resolve any violations, and produced a design with the fewest possible timing and physical violations. Next comes the sign-off check, which includes a final, physical verification check, RC parasitic check, sign-off time check, and power analysis. The design may then be tapped out for manufacturing.

In the final design stage, there are 476 single-cut vias and no DRC, connectivity, or DRV violations are present.

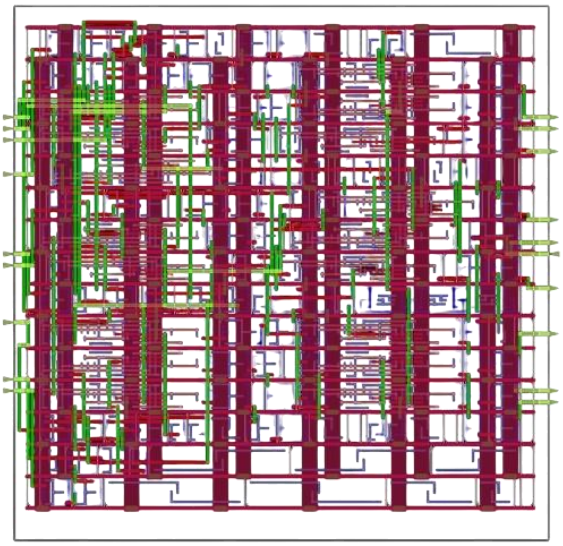


Fig. 10: Design final physical layout — completed place and route after optimizations.

The final physical layout, following optimization, is

given to show that the design process was accomplished effectively. It displays the results of the placement, routing, and optimization processes, indicating that the design fits the requirements and is ready for manufacture. This value is important because it demonstrates the efficacy of the design process and the overall operation of the IoT device.

The overall design area is 14 um2 with an aspect ratio of 1, of which 80 percent is eventually used. The overall locations of all 193 cells. There are 476 single-cut vias in the final design stage, and there are no DRC, connection, or DRV violations.

Table 2: Total number of cells used in the final designed layout

Cell Type	Count
Filler	56
Tap	30
Tie	18
Clock Buffer	3
Timing repair buffer	50
Inverter	18
Clock Inverter	1
Sequential cell	17
Total	193

Timing is a more important component of design as it affects functionality if it is done incorrectly. Thus, there shouldn't be any timing violation in the design. Timing violations mostly fall into two categories: setup and hold. The clock period mostly determines setup. The equations for Setup and Hold violation check:

$$Tc2q + Tcomb + Tsetup \leq Tclk + Tskew \quad (1)$$

$$Tc2q + Tcomb \geq Thold + Tskew \quad (2)$$

The setup analysis uses the worst data path delay if the path delay is maximal since the chip will pass other setup analyses if it operates with the largest delay

Timing analysis is done with OpenSTA, a tool that is also integrated with OpenROAD flow scripts. Also, KLayout is utilized for Physical Verification Checks.

Tool performed hold analysis using best case as in best scenario there is minimal data path delay and if chip functions here flawlessly, it will function for other analysis views [26].

Table 3: Final Design Timing Report

	Required Time	Arrival Time	Slack
Setup	100	92.78	7.22
Hold	84.6	94.22	9.62

The final complete timing report for the setup and hold design, including cell and net delay, is provided [26]. WNS and TNS are both greater than 0 because this design does not have any negative slack; for setup and hold analysis, these are 7.32 ps and 9.62 ps, respectively. Negative slack causes functional mismatches, whereas positive slack affects overall performance with regard to power but has no influence on device functionality. More positive slack will cause the device to consume more power.

For this design, an IEEE 1481-1999 SPEF file was also created. By supplying the parasitic values of each net, the SPEF file aids the STA tool in accurately calculating the delay [27].

```
=============================================================================
finish report_checks -path_delay max
-----------------------------------------------------------------------------
Startpoint: data_out[0]$_DFF_PP0_
          (rising edge-triggered flip-flop clocked by clk)
Endpoint: data_out[0] (output port clocked by clk)
Path Group: clk
Path Type: max

Fanout     Cap    Slew   Delay    Time   Description
---------------------------------------------------------------------------
                          0.00    0.00   clock clk (rise edge)
                          0.00    0.00   clock source latency
     1    1.01    0.00    0.00    0.00 ^ clk (in)
                                         clk (net)
                0.29    0.09    0.09 ^ clkbuf_0_clk/A (BUFx2_ASAP7_75t_R)
     2    1.28    7.04   11.10   11.19 ^ clkbuf_0_clk/Y (BUFx2_ASAP7_75t_R)
                                         clknet_0_clk (net)
                7.05    0.07   11.26 ^ clkbuf_1_1_0_clk/A (BUFx2_ASAP7_75t_R)
     9    5.32   19.29   19.03   30.30 ^ clkbuf_1_1_0_clk/Y (BUFx2_ASAP7_75t_R)
                                         clknet_1_1_0_clk (net)
               19.30    0.28   30.57 ^ data_out[0]$_DFF_PP0_/CLK (DFFASRHQNx1_ASAP7_75t_R)
     1    1.39   18.68   44.24   74.81 v data_out[0]$_DFF_PP0_/QN (DFFASRHQNx1_ASAP7_75t_R)
                                         _16_ (net)
               18.68    0.11   74.92 v _36_/A (INVx2_ASAP7_75t_R)
     1    0.68    7.16    6.45   81.37 ^ _36_/Y (INVx2_ASAP7_75t_R)
                                         net18 (net)
                7.16    0.06   81.43 ^ output10/A (BUFx2_ASAP7_75t_R)
     1    0.10    3.82   11.34   92.78 ^ output10/Y (BUFx2_ASAP7_75t_R)
                                         data_out[0] (net)
                3.82    0.00   92.78 ^ data_out[0] (out)
                                92.78   data arrival time

                          200.00  200.00   clock clk (rise edge)
                            0.00  200.00   clock network delay (propagated)
                          -40.00  160.00   clock uncertainty
                            0.00  160.00   clock reconvergence pessimism
                          -60.00  100.00   output external delay
                                  100.00   data required time
---------------------------------------------------------------------------
                                  100.00   data required time
                                  -92.78   data arrival time
---------------------------------------------------------------------------
                                    7.22   slack (MET)
```

(a)

```
=============================================================================
finish report_checks -path_delay min
-----------------------------------------------------------------------------
Startpoint: data_in[5] (input port clocked by clk)
Endpoint: internal_data[5]$_DFF_PP0_
          (rising edge-triggered flip-flop clocked by clk)
Path Group: clk
Path Type: min

Fanout     Cap    Slew   Delay    Time   Description
---------------------------------------------------------------------------
                          0.00    0.00   clock clk (rise edge)
                          0.00    0.00   clock network delay (propagated)
                         60.00   60.00 v input external delay
     1    0.59    0.00    0.00   60.00 v data_in[5] (in)
                                         data_in[5] (net)
                0.09    0.03   60.03 v hold15/A (BUFx2_ASAP7_75t_R)
     1    0.46    4.59   10.28   70.31 v hold15/Y (BUFx2_ASAP7_75t_R)
                                         net59 (net)
                4.59    0.03   70.34 v input6/A (BUFx2_ASAP7_75t_R)
     1    0.49    4.66   11.79   82.14 v input6/Y (BUFx2_ASAP7_75t_R)
                                         net14 (net)
                4.66    0.04   82.17 v hold16/A (BUFx2_ASAP7_75t_R)
     1    0.61    4.92   12.00   94.18 v hold16/Y (BUFx2_ASAP7_75t_R)
                                         net60 (net)
                4.93    0.05   94.22 v internal_data[5]$_DFF_PP0_/D (DFFASRHQNx1_ASAP7_75t_R)
                                94.22   data arrival time


                          0.00    0.00   clock clk (rise edge)
                          0.00    0.00   clock source latency
     1    1.01    0.00    0.00    0.00 ^ clk (in)
                                         clk (net)
                0.29    0.09    0.09 ^ clkbuf_0_clk/A (BUFx2_ASAP7_75t_R)
     2    1.28    7.04   11.10   11.19 ^ clkbuf_0_clk/Y (BUFx2_ASAP7_75t_R)
                                         clknet_0_clk (net)
                7.05    0.08   11.27 ^ clkbuf_1_0_0_clk/A (BUFx2_ASAP7_75t_R)
     9    5.19   18.89   18.89   30.16 ^ clkbuf_1_0_0_clk/Y (BUFx2_ASAP7_75t_R)
                                         clknet_1_0_0_clk (net)
               18.91    0.36   30.52 ^ internal_data[5]$_DFF_PP0_/CLK (DFFASRHQNx1_ASAP7_75t_R)
                         40.00   70.52   clock uncertainty
                          0.00   70.52   clock reconvergence pessimism
                         14.09   84.60   library hold time
                                 84.60   data required time
---------------------------------------------------------------------------
                                 84.60   data required time
                                -94.22   data arrival time
---------------------------------------------------------------------------
                                  9.62   slack (MET)
```

(b)

Fig. 11: GBA report for: (a) Setup analysis and (b) Hold analysis, showing timing checks for signal synchronization.

The GBA report for setup and hold analysis is key for ensuring the design meets timing requirements.

Usually, the sum of the circuit's dynamic and static power results in the overall power dissipation.

$$P_{total}=P_{dynamic}+P_{leakage} \quad (3)$$

$$P_{dynamic}=P_{switching}+P_{short\_circuit\ (internal)} \quad (4)$$

Dynamic and static power loss are the two basic forms that occur in designs.

Static power loss, often referred to as leakage power loss, and dynamic loss due to switching and short circuit loss. The design's switching activity is the primary determinant of switching power.

In reality, logic transitions from 0 to 1 and 1 to 0 in a design resulting in dynamic power [28]. Furthermore, short circuit power is mostly impacted by a cell's transition time. Here is the equation for Dynamic Power loss in a design.

$$P_{switch}=a\times f\times C_{load}\times V_{dd}^2 \quad (5)$$

$$P_{sc}=t_{sc}\times V_{dd}\times I_{peak} \quad (6)$$

Leakage power, however, is the most crucial component. Considering that loss of data happens while a chip or gadget is not in use. Therefore, this loss must be reduced [29]. A device's lifespan will be shortened otherwise.

Table 4: Breakdown of the total power in terms of switching, leakage, and internal power for each individual type of cell

	Internal Power (mW)	Switching Power (mW)	Leakage Power (mW)	Total Power (mW)
Sequential	0.0938	0.00258	1.68E-06	0.0964
Combinational	0.016	0.00728	2.92E-06	0.0232
Clock	0.0212	0.0349	1.41E-07	0.0562

The design's overall power dissipation is 0.1758 mW, while most IoT devices consume 0.1 to 1 mW power [30].

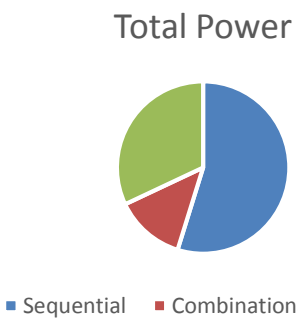


Fig. 12: A graphic representation of the percentages of total power used in a certain type of cell.

A maximum IR loss of less than 1% can be attained with wider stripes. When the layer becomes wider, it will reduce the total IR loss since R will also get smaller. If this design encounters an IR drop because of cell congestion in a specific area, the cell must be divided via placement blockage.

The following table contains the power analysis (IR) report for this design, with IR drop percentages equal to the ratio of the worst-case voltage to the IR drop.

Table 5: Table on IR droop analysis for VDD and VSS

Net	VDD	VSS
Supply Voltage	770 mV	0 mV
Worst-case voltage	769 mV	0.473 mV
Average voltage	770 mV	0.0719 mV
Average IR drop	0.0732 mV	0.0719 mV
Worst-case IR drop	0.604 mV	0.473 mV
Percentage drop	0.08%	0.06%

Due to the tiny size of the design, there isn't much IR drop; nevertheless, if it does exceed the limit, it varies from design to design. For example, in a moderate instance count design, the dynamic allowable IR drop is less than 10% and the static drop is 2% [31]. If the limit is crossed, the problem can be resolved by distributing the cell by putting cell blockages where IR problems arise, or by decreasing the R by widening the stripe layer or adding more stripes to the IR congested area.

Conclusion

The development, assembly, and implementation of a physical prototype for a cutting-edge Internet of Things device that can recognize cars are all covered in this study. Combining the ASAP7 PDK with Yosys for synthesis, OpenROAD for physical design, and OpenSTA timing analysis, a focused high-performance Internet of Things device was created. These attest to the design's true operation at maximum efficiency. The project primarily examines a chip's RTL-GDSII flow and the steps a designer takes to bring a chip from RTL to production. Additionally, using an open-source program, I completed the task and saw that the device's power, performance, and area (PPA) were all satisfactory.

Future Work

Starting at 10 GHz and satisfying the criteria with 1 mW of power, less area, and sufficient operating without timing violations were the main challenges. A significant setup and hold violation occurred on the 10 GHz frequency. Additionally, there was a brief infraction when an additional buffer and inverter were inserted to balance the clock tree to match the design at a 10GHz frequency. Thus, the area also grows. In addition, restricting IR drop to less than 1% and selecting power stripe width, which may supply power to the cell, were the other obstacles. This design can use a frequency of 5GHz to operate the device within the parameters.

But there are further issues as well; just 80% of the available space is used. Future studies could attempt to employ fewer filler cells and more device space. In order to improve system performance overall, future work will focus on developing vehicle identification algorithms,

investigating power-aware design, and investigating advanced design methodologies. Will also continue to work on constructing the entire gadget, including the MCU and LED sections. For design, the industry-standard tool can be utilized to get precise results.

Author Contributions

S. H. Rakib was responsible for planning and structuring the research, conducting the experiments, collecting the data, analyzing the results, and writing the manuscript.

S. N. Biswas reviewed the manuscript and provided necessary feedback for improvements.

Acknowledgment

The author gratefully acknowledges Neural Semiconductor Limited for the knowledge and support that contributed to this work.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

ASAP7	A design kit for 7-nm FinFET predictive processes
HPWL	Half Perimeter Wire Length
GDSII	Graphic Design System

References

[1] N. B. Soni, J. Saraswat, "A review of IoT devices for traffic management system," in Proc. International Conference on Intelligent Sustainable Systems, 2017.

[2] H. Jayakumar, A. Raha, Y. Kim, S. Sutar, W. S. Lee, V. Raghunathan, "Energy-efficient system design for IoT devices," in Proc. 2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC), 2016.

[3] E. Nasr, E. Kfoury, D. Khoury, "An IoT approach to vehicle accident detection, reporting, and navigation," in Proc. IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), 2016.

[4] C. V. S. Babu et al., "IoT-based smart accident detection and alert system," in Handbook of Research on Deep Learning Techniques for Cloud-Based Industrial IoT, 2023.

[5] R. K. Kodali, G. Swamy, B. Lakshmi, "An implementation of IoT for healthcare," in Proc. IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2015.

[6] G. Punyavathi, M. Neeladri, M. K. Singh, "Vehicle tracking and detection techniques using IoT," Mater. Today Proc., 51(1): 909-913, 2021.

[7] A. Hosny, A. B. Kahng, "Open-source EDA and machine learning for IC," in Proc. International Conference on VLSI Design and Embedded Systems, 2020.

[8] "GitHub-The OpenROAD Project," [Online]. Available: https://github.com/TheOpenROAD-Project.

[9] T. Ajayi et al., "INVITED: Toward an open-source digital flow: First learnings from the openroad project," in Proc. 2019 56th ACM/IEEE Design Automation Conference (DAC), 2019.

[10] A. B. Kahng et al., "The OpenROAD Project: Unleashing Hardware Innovation,".

[11] "OpenROAD Flow Scripts documentation!," [Online]. Available: https://openroad-flow-scripts.readthedocs.io/en/latest/index2.html.

[12] V. Vashishtha, L. T. Clark, "ASAP5: A predictive PDK for the 5 Nm node," Microelectron. J., 126: 105481-105481, 2022.

[13] L. T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy, G. Yeric, "ASAP7: A 7-nm finFET predictive process design kit," Microelectron. J., 53: 105-115, 2016.

[14] D. Velenis, M. C. Papaefthymiou, E. G. Friedman, "Reduced delay uncertainty in high performance clock distribution networks," in Proc. 2003 Design, Automation and Test in Europe Conference and Exhibition. 2003.

[15] P. Maurine, M. Rezzoug, N. Azemard, D. Auvergne, "Transition time modeling in deep submicron CMOS," IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 21(11): 1352-1363, 2002.

[16] "GitHub-Yosys," [Online]. Available: https://github.com/YosysHQ/yosys.

[17] S. N. Adya, I. L. Markov, "Fixed-outline floorplanning: enabling hierarchical design," IEEE Trans. Very Large Scale Integr. VLSI Syst., 11(6): 1120-1135, 2004.

[18] P. Kadarkarai et al., "Implementation of a PnR flow at block level to achieve the high utilisation rate," in Proc. International Conference on Inventive Computation Technologies (ICICT), 2024.

[19] M. Chakraborty, D. Saha, A. Chakrabarti, S. Bindai, "A CAD approach for pre-layout optimal PDN design and its post-layout verification," Microprocess. Microsyst., 65: 158- 168, 2019.

[20] C. K. Cheng, A. B. Kahng, I. Kang, L. Wang, "RePlAce: Advancing solution quality and routability validation in global placement," IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 38(9): 1717-1730, 2018.

[21] S. M. Das, S. M. Rafi, "Reducing cell density congestion issue in chip design," J. Microelectron. Solid State Dev., 6(3): 11-17, 2020.

[22] Sarrafzadeh, M. Wang, "NRG: Global and detailed placement," in Proc. International Conference on Computer Aided Design, 1997.

[23] N. Patel, "A novel clock distribution technology multisource clock tree system (MCTS)," Int. J. Adv. Res. Electr. Electron. Instrum. Eng., 2(6): 2234-2239, 2013.

[24] E. G. Friedman, "Clock distribution networks in synchronous digital integrated circuits," Proc. IEEE, 89(5): 665-692, 2001.

[25] D. Harris, M. Horowitz, D. Liu, "Timing analysis including clock skew," IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 18(11): 1608-1618, 1999.

[26] B. Rebaud, M. Belleville, C. Bernard, Z. Wu, M. Robert, P. Maurine, "Setup and hold timing violations induced by process variations, in a digital multiplier," in Proc. 2008 IEEE Computer Society Annual Symposium on VLSI, 2008.

[27] J. H. Kim, W. Kim, Y. H. Kim, "Efficient statistical timing analysis using deterministic cell delay models," IEEE Trans. Very Large Scale Integr. VLSI Syst., 23(11): 2709-2713, 2015.

[28] M. Mostafa, M. Watheq El-Kharashi, M. Dessouky, A. M. Zaki, "A novel flow for reducing dynamic power and conditional performance improvement," IEEE Trans. Circuits Syst. I Regul. Pap., 68(5): 2003-2016, 2021.

[29] N. P. Bose, N. Santhi, "Efficient leakage reduction approach for low power VLSI design using modified feedback sleeper stack technique," Int. J. Electron. Commun. Eng., 11(3), 2024.

[30] P. Mayer, M. Magno, L. Benini, "Smart power unit—mW-to-nW power management and control for self-sustainable IoT devices," IEEE Trans. Power Electron., 36(5): 5700-5710, 2020.

[31] V. G. Menon, S. Jacob, S. Joseph, P. Sehdev, M. R. Khosravi, F. Al-Turjman, "An IoT-enabled intelligent automobile system for smart cities," Internet Things, 18, 2022.

Biographies

Saeed Hossen Rakib received his bachelor's degree in Electrical and Electronic Engineering from Ahsanullah University of Science and Technology in Dhaka in June 2022. After that, he began working as a physical design engineer in Neural Semiconductor Limited. ASIC Design, Low Power IC Design, VLSI Circuit and System Design, Microelectronics and Nanotechnology, Electronics, SoC Design are among his areas of interest in study.

- Email: s.h.rakib.153@gmail.com
- ORCID: 0009-0008-3727-6699
- Web of Science Researcher ID: N/A
- Scopus Author ID 58920493500
- Homepage: NA

Satyendra Nath Biswas (M'98) received his B.Sc. from BUET and M.Sc. & Ph.D. from Yamaguchi University, Japan. He has worked as an R&D Engineer in Dhaka and Canada and was a Research Assistant at the University of Ottawa. Currently, he is an Assistant Professor at Georgia Southern University. His research focuses on VLSI design, built-in self-testing and reconfigurable computing. He is a Professional Engineer (P.Eng.) and a member of IEICE.

- Email: sbiswas.eee@aust.edu
- ORCID: 0000-0002-5334-0784
- Web of Science Researcher ID: N/A
- Scopus Author ID 14048019400
- Homepage: NA

How to cite this paper:
S. H. Rakib, S. N. Biswas, "Implementing yosys & openroad for physical design (PD) of an IoT device for vehicle detection via ASAP7 PDK," J. Electr. Comput. Eng. Innovations, 13(2): 463-472, 2025.

DOI: 10.22061/jecei.2025.11258.785

URL: https://jecei.sru.ac.ir/article_2315.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Usability of Iranian Math Apps for Kids

N. Zanjani *

Computer Engineering Department, Refah University College, Tehran, Iran.

Article Info

Article History:
Received 31 December 2024
Reviewed 03 February 2025
Revised 13 March 2025
Accepted 26 March 2025

Keywords:
Usability
Math apps
Kids' education
Kids' apps
HCI

*Corresponding Author's Email Address: *Zanjani@refah.ac.ir*

Abstract

Background and Objectives: The widespread use of mobile apps among children has introduced both opportunities and challenges, particularly in the realm of educational tools. Usability is critical for these apps, as it ensures that young users can easily engage with and benefit from educational content. The objective of this study is to evaluate the usability of Iranian Android math apps designed for children.
Methods: This study is the expert review research and focuses specifically on Android applications designed to teach mathematics to children aged 6-9 years (preschool to grade 3). The apps were selected from two popular Iranian app stores: Bazar and Myket. A total of 100 math apps were intentionally chosen. Each app was tested for 15 minutes by the researcher to evaluate usability based on 39 usability factors derived from the literature on human-computer interaction. Non-functional, non-interactive, paywalled, or text-only apps were excluded, leaving 44 apps for detailed analysis.
Results: 98% of the apps showed consistency in navigation and visual elements. 77% of the apps provided feedback to users, indicating when a mistake was made or when a task was completed. However, only 9% offered positive feedback. 86% of the apps had appropriately sized icons and text, making them accessible to children. However, about 40% of the apps needed improvement in terms of simplifying the language and instructions to suit young children's comprehension levels. 89% of the apps offered little to no personalization options. Most apps (56%) relied heavily on text prompts rather than audio or visual cues, making navigating difficult for younger children without adult assistance. 75% of the apps did not encourage children to engage in online transactions and 73% were free of advertisements, creating a safer and less distracting learning environment.
Conclusion: While many Iranian math apps for children adhered to basic usability principles, there was a gap between research recommendations and their practical application, particularly in areas related to engagement, feedback, and personalization. Developers could partner with schools and education organizations to create apps that align with specific curriculums, have more personalized features, engage children using cartoon characters, and include interactive educational tools. Educational tools and platforms should provide environments that allow students to interact more with content, teachers, and classmates. This can be achieved through live chats, group discussions, and increased interactions with digital content such as quizzes and interactive assignments. Further, using gamification elements such as scoring, badges, and challenging levels can make learning process more engaging.

Introduction

Mathematics is a fundamental subject for science and technology, which is necessary for the growth of countries and progress in fields such as medicine and engineering. However, many students struggle with mathematics. Research shows that a strong foundation in

elementary school is critical to success at higher levels of mathematics. In addition, students' academic success in mathematics positively and significantly predicts the level of their computational thinking skill [1]. Traditional teaching methods focus too much on memorization and can hinder students' deep understanding and interest. To solve these problems, the use of digital teaching methods is suggested to strengthen students' mathematical and problem-solving skills [2].

On the other hand, usability plays a key role in the design and development of digital products and interfaces, as it directly impacts user satisfaction, efficiency, and overall experience [3]. The main aspect of usability is ensuring that a system is intuitive and easy to use, allowing users to navigate and interact with it seamlessly. By prioritizing usability, designers can enhance user engagement and retention, reduce errors or frustration during interactions, and increase productivity and efficiency. Overall, the importance of usability lies in its ability to create positive user experiences, ultimately contributing to the success and effectiveness of digital products and services[4], [5].

The ever-increasing popularity of touchscreen devices and virtual apps has caused a significant increase in the utilization of touchscreen interfaces by children both for gaming and educational purposes [6]. Studies show that globally, over half of children under the age of 3 regularly use touchscreen devices [7]. Fig. 1 shows the result of a survey on U.S children's engagement with digital devices [8]. However, the usability of software plays a crucial role in children's engagement with these applications. Research in Human-Computer Interaction (HCI) and Interaction Design and Children (IDC) has demonstrated that the design of interfaces significantly shapes children's interactions with touchscreen apps [9]. However, several existing mathematics applications focus more on the content, and the usability of applications is ignored [10].

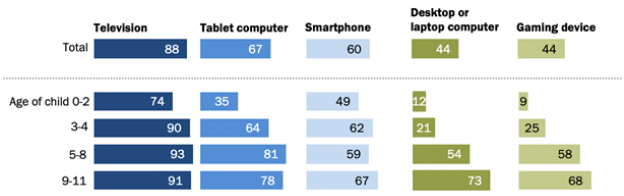


Fig. 1: Children's engagement with digital devices [8].

Therefore, the usability of kids' apps is of vital importance, due to the unique characteristics and needs of young users. Children, especially those in the early stages of development, require interfaces that are intuitive, engaging, and easy to navigate to effectively interact with digital content [11]. The usability of educational apps is essential to ensure they effectively facilitate learning and engagement among children [12]. Usability considerations such as clear navigation, age-appropriate language, and interactive elements tailored to children's cognitive abilities are crucial in ensuring that kids can fully benefit from and enjoy educational or entertainment apps. By prioritizing usability in the design and development of kids' apps, developers can create safe, engaging, and enriching digital experiences that cater to the specific needs and preferences of young users, ultimately contributing to their overall growth and development [13].

However, there is a lack of thorough evaluations on the usability of children's apps that are promoted with different educational goals [14]. Without a clear evaluation framework, parents and educators struggle to determine which applications provide both a user-friendly experience and effective educational support. A comprehensive review is necessary to establish usability benchmarks, ensuring that digital learning tools meet the developmental needs of children. Additionally, as many children's apps seem to prioritize generating advertising revenue or collecting user data over usability, this raises concerns regarding ethical design practices and children's digital safety [15].

Given the abundance of Iranian apps designed for children, examining the usability of these apps and how they align with the principles of Human-Computer Interaction is paramount. Iranian children grow up in an environment where their language, culture, and learning styles differ from those of children in other countries. Therefore, they should be designed in simple Persian and use familiar symbols for Iranian children. Moreover, in many countries, there are strict regulations to protect children in the digital space. However, this is not the case in Iran. Examining this issue as a usability parameter, can help policymakers and developers establish better standards for children's digital safety.

Hence, the purpose of this research is to see if there is a gap between research and practice in Iranian educational apps developed for children. To narrow the scope of this research we focus on those Iranian Android apps that teach mathematics to children. Therefore, the question of this research is: to what extent do Iranian educational android math apps adhere to scientific usability principles. We should clarify that this study does not assess the educational effectiveness of these apps but solely their design principles. The result of this research can highly be useful for designers and developers of educational apps in Iran. Besides it can give parents and educators a comprehensive insight to select proper educational apps for children.

Review of the Related Literature

The first step in evaluating applications is finding a suitable framework for analyzing them [16]. Researchers [17], identify a scarcity of reliable evaluation tools for

educational apps, with existing tools often being either too complex, outdated, or lacking scientific backing. Overall, the literature review underscores the necessity for effective evaluation tools to navigate the vast array of educational apps and ensure they provide genuine educational benefits. In this part, a review of evaluation tools is presented as well as research that analyzed interface design practices based on design recommendations in the literature.

In the user experience design process for children, the designer should first attract the child by appealing to their imagination, interests, and motivation [18]. Then, the designer should provide a high degree of freedom for the child to explore and experiment in the new world [19]. Finally, the designer should plan how the experience in the virtual world can be integrated into the real world for a holistic and engaging experience [20]. The best practices in user experience design for kids include showing respect by considering child thinking [21], using clear and consistent interfaces, sticking to plain talk with simple language and design elements, and gaining trust by providing safe and secure products [22], encouraging interaction through interactive features, and rewarding loyalty with virtual goods and incentives. These practices aim to create a positive and engaging online experience for children while also promoting repeat purchases and customer retention [23].

Soni and her colleagues developed a framework called TIDRC [9]. The TIDRC framework consists of 57 design suggestions that are divided into 19 interface dimensions and seven overarching categories depending on the impacted interface features. It provides practical design suggestions based on research for meeting the cognitive, physical, and socio-emotional needs of children in touchscreen interaction design. It includes advice on visual design, audio features, interactive elements, application responsiveness, informational features, physical gestures, target features, and socio-emotional contextual features. These recommendations are tailored to different age groups, from 2-7 years old to 7-11 years old, and aim to help create engaging and effective interaction.

In another study, a collection of 23 usability guidelines was established for designing mobile-based Augmented Reality (AR) apps intended for kindergarten-aged children. Researchers carried out a thorough literature review to pinpoint existing usability principles from various sources. They then organized expert meetings to evaluate and enhance these principles. Lastly, they utilized factor analysis to group the refined principles into four categories: cognition, orientation, design, and support. Cognition features focused on cognitive and intellectual elements like learnability, efficiency, and minimizing memory load. Orientation emphasized user comprehension and interaction, such as enjoyment and customizability. Design principles centered on application usage, for instance, interactivity and simplicity, and finally support Geared towards user assistance, including error management and early testing. These guidelines aim to direct the development of more engaging, easy-to-learn, and user-friendly AR apps made for kindergarten children [24].

There is another instrument called the E.T.E.A. (Evaluation Tool for Educational Apps) [17]. It is designed to assess the quality of educational apps targeted at children aged 3 to 6 years. The instrument was developed based on a review of existing rubrics and checklists for evaluating educational apps, and it underwent Exploratory Factor Analysis (EFA) to confirm its validity and reliability. The E.T.E.A. aims to provide a simple, valid, and reliable tool for parents, custodians, and educators to make informed decisions when selecting educational apps for young children. The E.T.E.A. consists of a thirteen-item assessment questionnaire that evaluates four key dimensions of educational apps including usability, efficiency, parental control, and security. Usability assesses how easy the app is to use for children, including the clarity of instructions, consistency of visual elements, and the overall ease of navigation. Efficiency evaluates how effectively the app facilitates learning and engagement for children. Parental Control examines the features that allow parents to monitor and control their child's use of the app, such as providing feedback on the child's progress and ensuring no disruptive advertisements are present. Security focuses on the app's privacy policies and how it manages personal data, ensuring that children's information is protected.

Besides papers that discuss dimensions for evaluating apps, other researchers recommend some points to enhance the usability of apps for children. For example, Anthony and her colleagues recommend providing visual feedback, especially for children, wherever possible during surface gesture interaction on mobile devices [25]. Another key issue that Meyer and his colleagues raise is the frequent use of annoying video ads and the fact that children are often tempted to watch these ads in exchange for rewards [26]. These distractions are annoying and can cause users' dissatisfaction. Besides they can pull young users away from the real learning objectives of the app, making it harder for them to focus on what they should be gaining from the experience [27].

To assess the usability of Iranian math applications created for children, the researcher combined and categorized various dimensions and aligned them with HCI guidelines and principles, as suggested by Shneiderman and Plaisant [4], who advocate for designing interfaces for children as if they are new or inexperienced users. This led to 39 parameters including criteria such as

consistency [9], [12], [17], [23], [24], [28], informative feedback [12], [17], [28], design adapted to children's skills and interaction styles [9], [17], [24], [28]. Table 1 presents these criteria in detail.

Besides considering general usability guidelines for novice or first-time users, limiting the vocabulary to a few commonly used concept terms is crucial [23], [28]. It is also important to keep the number of actions minimal to ensure that new and inexperienced users can complete basic tasks successfully [9], thereby reducing anxiety, boosting confidence, and offering positive reinforcement. Providing informative feedback on task completion is beneficial, and offering clear, specific error messages when users make mistakes is essential [12], [17], [28]. Thoughtfully crafted video demonstrations and online tutorials can also be effective [4].

Method

The method of this research is expert review. In this method, the researcher is an expert in usability studies and evaluates objects systematically based on standard parameters. This method doesn't require direct testing on users. A natural way to evaluate interfaces is to show them to users and gather their feedback. While informal demos with test subjects can offer some insight, formal expert reviews are generally more effective. Expert review enables faster and more standardized evaluation. Experts can analyze applications using design principles and scientific criteria without being influenced by children's behavioral variables. Young children may not be able to provide precise or logical feedback. Utilizing experts in this field ensures that analyses are based on well-established and scientifically valid criteria. Many usability evaluation studies have employed expert review, as it offers more efficient methods for systematically analyzing the design features of applications. [4].

In this study, the usability of Iranian Android math apps designed for children aged 6 to 9 (preschool to grade 3) was evaluated. Given the restrictions on iOS usage in Iran, the apps were selected from two major Iranian Android app stores: Bazar[1] and Myket[2]. These platforms are widely used due to their accessibility. To ensure to have a representative sample, we started by manually searching for 100 apps using keywords like "teaching math to kids," "math learning apps," and "kids math education" in Persian. It helped to find a diverse range of apps that directly focused on teaching math to children. Since it was aimed to explore a broad range of apps, no limits were set on the number of downloads, which varied from just 10 to over 500,000. Between July 22 and August 5, 2024, we selected 100 apps, intentionally excluding any older versions of apps already on the list. Bazar and Myket don't offer an option to filter app search results based on what users specifically need. On top of that, the "number of hits" shown includes not just the apps themselves but also every time the search term appears in the app titles and descriptions. This means that searching for something like "teaching math to kids" often brings up a lot of irrelevant results.

The researcher, applied strict criteria to decide which apps would move forward for testing. Only apps that worked properly and offered interactive features were included for detailed usability analysis. Interactive applications are programs that allow users, especially children, to actively engage with them rather than merely viewing static content such as text or videos. Therefore, Apps that didn't function correctly, required purchases before offering any meaningful content or lacked interactive elements (such as text-only apps) were excluded. Additionally, apps that were simply videos or non-interactive presentations, were excluded, as these didn't meet the goal of helping children actively engage with math learning. After this filtering process, it ended up with a final sample of 44 apps, which were then put through a detailed usability review.

Each of the 44 apps was tested for around 15 minutes. The decision to use a 15-minute window came from the average time children tend to spend on educational apps in a single session. During the testing phase, the apps were assessed based on 39 usability factors drawn from existing research on Human-Computer Interaction (HCI) and educational app design. These factors included how easy the app was to navigate, the clarity of its instructions, how well it used visual and auditory feedback, and whether the content was appropriate for the target age group.

For each app, the researcher tried to complete as many tasks as possible within the 15-minute timeframe. If an app was more complex, extra time was given to make sure all its features were thoroughly assessed. Throughout the process, we noted any design issues or problems with responsiveness and made observations about how intuitive the app was for a child to use. The researcher made notes explaining how the gameplay aligned with the criteria for each score.

To evaluate the apps, each usability factor was rated from 0 to 4. A score of 0 indicated that the app failed to meet usability expectations, while a score of 4 showed that it fully complied with usability guidelines. For each program, each of the 39 items shown in Table 1 was investigated and a number between 0 and 4 was assigned to the App for that parameter. Finally, the percentage of programs that scored 3 or higher on each parameter was determined, as reported in Fig 2.

Each app was assessed on several key areas:

[1] https://cafebazaar.ir/?l=en
[2] https://myket.ir/

Table 1: HCI Guidelines for designing interfaces for kids

HCI Guidelines	Details
Strive for consistency	Metaphors, navigation, content, and visual elements [9], [12], [17], [23], [24], [28].
Providing informative feedback	Children know if they make a mistake [28] Visual/audio [9] Appropriate, and clear [12], [17], [28]]
Considering kids skills	The use of animation and images matches with children's skills [28]. Instructions in apps are presented in a manner suitable for children [17] The app is user-friendly for children, allowing for easy scrolling and navigation [17], [24], [28] Icon sizes are appropriate and manageable for children [28] Interactive widgets are intentionally designed to be visually larger [9] The language used is simple and suitable for the target age group [23], [28] Abstract signs and symbols are eliminated [9] Avoidance of visually complex backgrounds in applications is recommended [9] Use a minimum of 14-point font size and appropriate spacing [9], [28]
Reduce user frustration	Avoid App crash, hanging, or freeze [24], [28] Fast load [12]
Attract users	An engaging design with clear visuals [24], [29] Bright colors that attract children [28] Stimulation of children's imagination, interests, and motivation [23] Incorporation of interactive cartoon characters on the screen [23]
Getting the users' attention	Use sound effects and voice to get users' attention [9]
Consider Learnability and Retention overtime	Once they receive help from adults, children should be able to use it on their own. The app should be easy to use without requiring special training. [12], [24], [28] Reduce short-term memory load [5], [23]
Legal and security concerns	The app does not encourage children to engage in any online transactions [23] The app is free from advertisements, such as pop-up messages [17]
Personalization	The application should be flexible enough to get customized based on user requirements [9], [24], [28] Easily users can start or stop any activity at any time [28] The application should allow users to bypass instructions or content that isn't part of the gameplay [24] The app should offer children a significant level of freedom to explore and experiment within the new environment [23]
Interaction Styles	Clearly distinguish clickable items from other elements on the screen [9]. Restrict the functionality of clickable items to their intended purpose [9]. Avoid using extensive menus in apps designed for children [9]. Ensure the menu is suitable for touchscreen use [28]. Minimize the use of text prompts [9]. Include audio prompts alongside visual cues [9]. Use animated prompts to illustrate gestures [9]. Provide audio support for text labels and instructions [9]. Avoid implementing rotation gestures [28]. Do not utilize pinch-to-zoom gestures [28]. Avoid drag-and-drop gestures [28].

Navigation and Interface Consistency: how smoothly the app's design and navigation flowed, helping children move easily through different sections without getting confused.

Feedback Mechanisms: how well the app informed users about mistakes or task completion, such as through visual cues or sounds.

Engagement and Attractiveness: how visually appealing and engaging the app was for children. This included elements like animations, vibrant colors, and fun tasks to hold their attention.

Personalization Options: whether the app allowed users to adjust settings like difficulty levels or themes to cater to individual learning preferences.

Cognitive Design: whether the app's language and instructions were simple and clear enough for children to understand without adult assistance, and if it used helpful audio or visual cues to support learning.

After completing the usability testing of all 44 apps, the scores were compiled and analyzed. The frequency of scores for each usability factor was calculated to help identify patterns. This analysis helped to highlight where most apps were doing well and where improvements were needed. For example, apps that scored below 39 points (out of a possible 156) were flagged as having significant usability problems. Based on Table 1, 39 parameters were examined for each App, when an App is

scored below 39, it means that in most of the questions it had gained one or less.

Before starting the full usability evaluation, a pilot test was conducted with five randomly selected apps. This pilot phase helped fine-tune the evaluation process, making sure that the chosen usability factors were appropriate for the apps being tested. It confirmed that 15 minutes was an appropriate amount of time for testing most apps, as children typically use educational apps in short bursts.

Since no children were directly involved in the study, there were no ethical concerns related to user participation. However, each app was carefully reviewed to ensure that it was age-appropriate and that no harmful content or in-app purchases were present that could exploit children. Apps that included excessive advertisements or encouraged online transactions were flagged and excluded from the final sample to protect children from potential risks.

Results and Discussion

A quick overview of math apps for kids on Bazar and Myket reveals that most of them are free, with full versions accessible, and only a small number require payment. About 96% of the apps, largely released or updated between 2019 and 2024, don't cost anything. The developers of most of these apps are freelancers or private companies.

When it comes to math apps for kids, nearly 12% of them don't work at all! 11% just display text to teach kids math!! 18% aren't interactive and only provide videos for teaching math to kids. Some videos are simply recordings of a classroom lesson, while others are more creative and designed as animations. However, none of them create any interactive tools for the child. Besides 15% asked kids in-app purchases or register and provide some information such as their mobile phone number without even letting the user examine a demo of the application to decide whether he wants to use it or not. All these apps were excluded and the remaining 44 ones were analyzed.

From the remaining 44 teaching math to kids' apps that we explored, 18% of them only have one main function mostly counting or clock. About 28% offer two features such as adding and subtraction, and 40% provide three or four functionalities mostly adding, subtraction, multiplication, and division. Apps with five or more functions make up just 14% and add functionalities such as clock, shapes, and pattern recognition to the functionalities mentioned above. These apps generally aim at two groups: children and parents or teachers. Most are designed for kids, while only 8% are specifically created for parents or teachers. Besides, user ratings vary from 2.2 to 5, with an average of around 4.

Results show that nearly 98% of apps "strive for consistency" and follow the same navigation and visual elements around the app. This finding aligns with other research emphasizing the importance of a consistent interface for usability in educational apps. For instance, Soni and her colleagues developed a framework (TIDRC) that includes design suggestions for cognitive and physical needs, which supports the need for consistency in app design [9].

Regarding providing feedback to users, the majority of the apps (77%) in our sample "provide feedback" and let the kids know about their mistakes. This finding is consistent with recommendations from other studies, such as those by Anthony and colleagues, who advocate for visual feedback during interactions to enhance usability for children [9]. However, only 9% of apps use positive phrases such as "My dear try again" to inform the user of his mistake. Others, just create a text, sound, vibration, or a changed color, among them, 4% make harsh sounds and that may cause kids to experience stress. This indicates a significant gap in the quality of user interaction. This discrepancy suggests that developers are more focused on visual consistency than on increasing user motivation through positive feedback. The impact of positive on children's learning and motivation is significant, influencing their engagement and persistence in educational tasks. Positive feedback has been shown to enhance motivation, self-efficacy, and skill acquisition, while negative feedback can lead to decreased motivation and increased anxiety [30], [31].

Concerning the parameter of "considering children's skill levels", more than half of the apps meet this criterion. Specifically, in regards to icon and text sizes, about 86% of apps designed icons and texts large enough to be suitable for kids. It is also notable that almost none of the studied apps used abstract signs and symbols that are vague for kids and avoid complex backgrounds. All the apps we explored were easy to use and children could learn to start working with them without the help of adults. However, instructions and language used in nearly 40% of apps need to be edited to be more understandable for kids. This difference suggests that more attention was paid to visual design than to the simplicity of language.

The Average score for the recommendations to help attract users is 2.1 which shows apps are not attractive enough and don't stimulate children's imagination, interest, and motivation. Since visual appeal and effective interaction play an important role in children's learning, this deficiency can negatively affect children's concentration, motivation, and information retention. Research has shown that the use of interactive elements such as cartoon characters [23] , and gamification [32] can increase children's engagement with educational content, which ultimately improves understanding of concepts and increases the duration of interaction with educational apps [33]. However, Only 16% of apps use interactive cartoon characters which is recommended to

engage. This low rate suggests that most developers have not paid enough attention to the psychology of children's learning and the role of engagement in enhancing their cognitive performance. As mentioned earlier, 15% of the 100 apps we selected to analyze, asked kids in-app purchases before allowing them to do any interaction and we excluded these apps. However, fortunately, 75% of 44 apps studied in this research do not encourage children to engage in any online transaction and 73% are free from advertisements. This is a positive finding compared to other studies [26] that have highlighted the negative impact of intrusive ads on children's focus and learning objectives. However, while these numbers represent a positive trend in creating a safe environment for children, 25% of apps still have in-app features and 27% have ads, which can be distracting for younger users.

In regards to personalization, 89% of apps don't provide customization options to users. Others allow very limited personalization features. Just 2% of apps let the user change difficulty level. Further, in 2% of apps, the user can choose between the automatic process of the app and the user-based selection process.

Only 7 % of apps let the user connect or disconnect the background music and in 2% of apps, the user can ask for the repetition of instructions. This statistic shows that developers have paid less attention to the individual needs of users, while personalization options can help improve the user experience and increase children's engagement with the app. Finally, analyzing interaction style parameters, shows that nearly in all apps, clickable items are clearly distinguishable and their functionality is restricted to the intended purposes. Extensive menus are not used and menus are suitable for touchscreen use. However, more than half of the apps (56%) use text prompts extensively instead of audio or visual cues and 82% don't provide audio support for text labels and instruction. Moreover, 73% don't include audio cues with visual prompts, and 91% don't employ animated prompts to illustrate gestures. This shows that many apps are not designed optimally for interaction with younger users, who may not have reading skills. This shortcoming can reduce accessibility and ease of understanding of content for children.

Fig. 2 gives a visual view of the results explained. While most apps meet basic usability standards, there are still some significant shortcomings in areas that could impact how well children can learn and stay engaged with these tools.

A. Adherence to Basic Usability Principles

Most of the apps (98%) followed the key usability principles, particularly when it came to consistent navigation and visual elements. This consistency helps children use the apps more easily, reducing confusion and making it simpler to complete tasks. The apps also generally did well in terms of physical accessibility, like having buttons and text that were the right size for children's small hands and developing motor skills. These design elements are crucial because they allow children to interact with the apps without getting frustrated, helping them focus more on learning rather than figuring out how to use the app.

B. Feedback and Engagement Issues

While many apps (77%) provided feedback when children made mistakes or completed tasks, only 9% offered positive, encouraging feedback. Most apps relied on simple cues like text or color changes, and a few even used harsh sounds to indicate errors, which can cause stress for young users. The lack of supportive feedback is a missed opportunity to motivate children and encourage learning. Positive reinforcement is especially important in educational settings, as it can boost a child's confidence and willingness to keep trying.

The apps also scored low on how engaging they were, with an average score of 2.1 out of 4. Most apps lack the interactive features that make learning fun for kids, like cartoon characters or exciting visuals. Only 16% of the apps used such engaging elements. Without these, children are less likely to stay focused on the math content, which reduces the app's overall effectiveness.

C. Challenges with Cognitive Design

Another major issue was the cognitive design of many apps. While over half of the apps did take into account the skill levels of young children, about 40% still used complex language and instructions that would be difficult for a 6- to 9-year-old to understand. This presents a significant barrier to effective learning since kids may struggle to comprehend what they're supposed to do.

Additionally, more than half of the apps (56%) relied too much on text prompts and didn't offer enough visual or audio cues to help non-readers navigate the content. Effective app design should consider children's developmental stages, and many of these apps failed to offer the necessary supports for younger users. Apps that don't offer clear guidance or break things down into manageable steps are harder for children to use independently, limiting their learning potential.

D. Limited Personalization

A key shortcoming in many of the apps was the lack of personalization. Only 11% of the apps allowed users to tailor their experience, such as adjusting the difficulty level or changing the feedback style. Personalization is essential because it allows the app to cater to each child's unique learning pace and needs. Without this flexibility, the apps are less likely to meet the diverse learning styles of children effectively. This is especially important for kids who need extra support or, conversely, those who may require more challenging tasks.

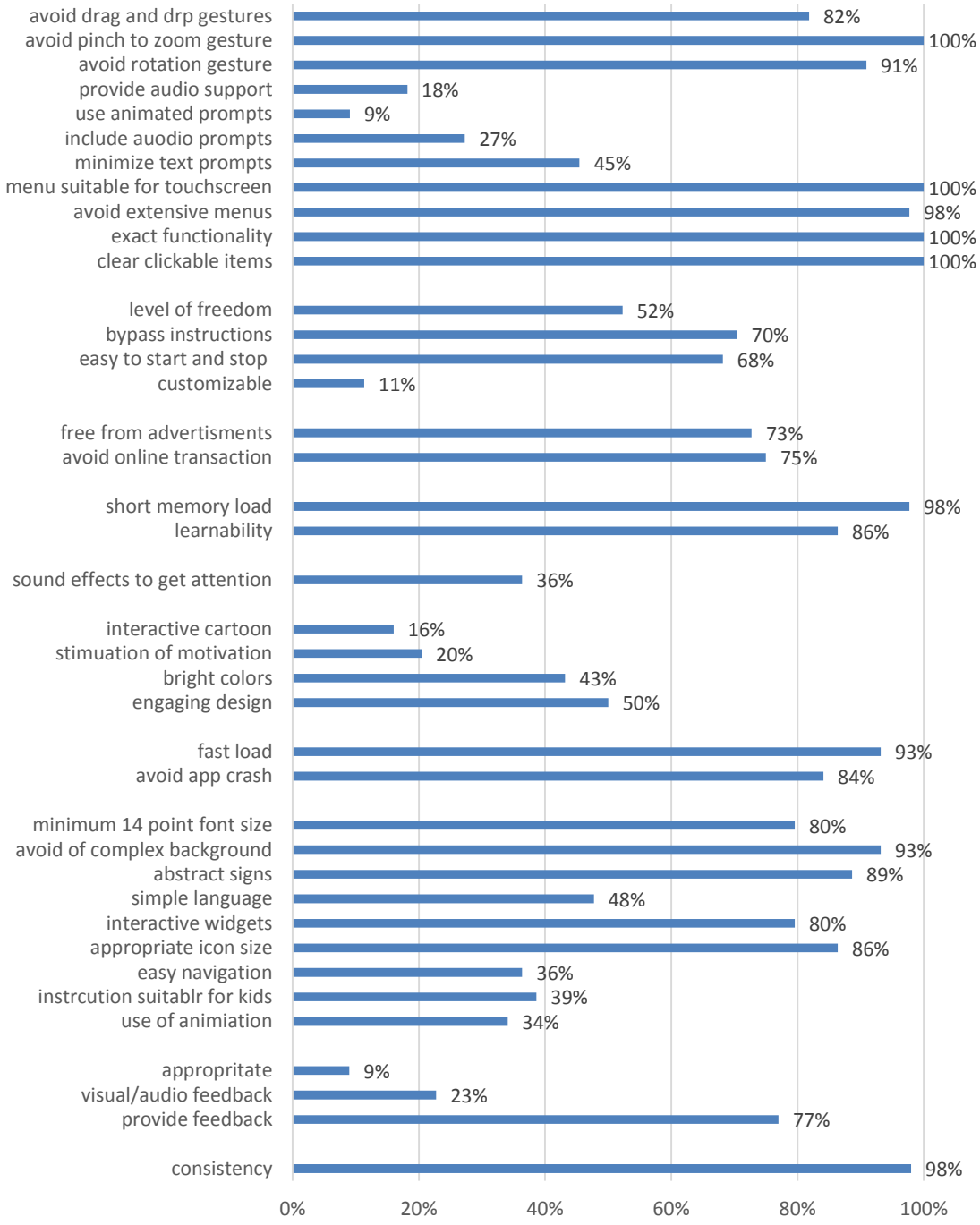


Fig. 2: The rate of apps' adherence to HCI factors.

E. Safe and Distraction-Free Learning

On a more positive note, 75% of the apps did not push for online transactions, and 73% were free of ads. This is a huge plus for parents and teachers, as it means the apps provide a safer, less distracting environment for learning. Ads and in-app purchases can take away from the educational experience and introduce risks for children, so their absence in many of the apps is a strong point.

F. Gaps between Research and Practice

One of the key findings from this study is the gap between what research suggests for educational app design and what's being implemented. While many apps followed basic physical usability guidelines, they often didn't perform well in areas like cognitive design, engagement, and feedback—all of which are essential for effective learning.

In terms of cognitive design, about 40% of apps used complex language and instructions that were difficult for children to understand. In addition, more than 56% of apps relied heavily on text messages and did not use audio or visual prompts appropriate for illiterate or low-literate children. Good cognitive design helps children understand new concepts with minimal mental effort and not get discouraged from using the app [34], [35]. Developers should use simple language, meaningful icons, audio prompts, and educational animations to help children.

Regarding engagement, the average user engagement score was 2.1 out of 4, indicating low engagement for most apps. Furthermore, only 16% of apps used interactive cartoon characters, even though research has shown that interactive characters help increase attention, motivation, and enjoyment of learning. Effective learning happens when children are actively involved. The lack of engaging features can make children lose interest in the app [33]. Developers can make the learning experience more engaging by adding gamification elements (scoring, badges, and challenges) and using animated characters.

Concerning providing positive feedback, although 77% of apps provide some form of feedback, only 9% provide positive, motivating feedback. This means that most apps use neutral or even negative methods when providing feedback on user mistakes, which can make the learning experience stressful. Research has shown that positive, motivating feedback increases children's self-confidence and improves the learning process [30], [31]. Developers should use positive statements and engaging visual feedback to motivate children to learn more.

Developers need to understand the importance of adding more interactive features, offering positive feedback, and simplifying language for younger users. By following research-based recommendations more closely, developers can create apps that not only meet usability standards but also enhance the learning experience for children.

Conclusion

This research sheds light on how Iranian Android math apps for children are performing in terms of usability, uncovering both strengths and areas for growth. The results are relevant for various groups—app developers, educators, parents, and policymakers—who all have a stake in improving educational technology for young learners.

This study found significant gaps in how the apps engage with children, provide feedback, and support their cognitive development. Developers can use these findings to create better, more user-friendly educational apps. Many of the apps were too text-heavy or complicated for young children. Developers should focus on making apps simpler and more intuitive by using age-appropriate language, visuals, and sounds. Only 9% of the apps provided positive feedback, which is crucial for motivating kids. Positive feedback helps build confidence and encourages children to keep learning, making the experience more rewarding [36]. With just 11% of the apps offering any customization, there's a big opportunity for improvement. Developers should consider adding features like adjustable difficulty levels or options for different learning styles. This way, each child can have a personalized learning experience that fits their unique needs, helping them learn more effectively. Developers could partner with schools and education organizations to create apps that align with specific curriculums, have more personalized features, use cartoon characters to engage children, and include interactive educational tools. By doing so, they can ensure that their apps meet both educational and usability standards, increasing their appeal and credibility within the educational community.

Educational tools and platforms should provide environments that allow students to interact more with content, teachers, and classmates. This can be achieved through live chats, group discussions, and increased interactions with digital content such as quizzes and interactive assignments. Further, using gamification elements such as scoring, badges, and challenging levels can make the learning process more engaging.

Teachers and parents are key decision-makers when it comes to choosing educational tools for children. This research provides them with insights on how to pick the best apps for learning. The study shows which app features make a real difference in usability and learning. By choosing apps with positive feedback, simple designs, and engaging content, educators and parents can ensure that children are using tools that help them learn. Since many of these apps still require adult help, educators and parents need to be aware of this. They might need to step in to provide additional support or guidance, especially when apps don't have enough visual or audio cues.

Policymakers and educational organizations can also use this research to set better standards for educational apps, ensuring they meet both usability and learning requirements. Based on the gaps found in this research, education authorities could develop clear guidelines that app developers must follow. These guidelines should focus on keeping the design simple, the navigation intuitive, and the content engaging and suitable for young learners. Further, policymakers could introduce a certification process that ensures only high-quality educational apps are recommended for use in schools. This would give teachers and parents a trusted way to know which apps are the most effective and safe for children to use.

In summary, the findings of this research have practical implications for how educational apps are developed,

selected, and used. By applying these insights, developers can create more engaging and effective tools paying more attention to interactive design and gamification. While educators and parents can make informed choices about which apps to use, choosing apps that provide positive feedback and voice guidance while committing to security settings. This ultimately leads to better learning experiences for children, helping them succeed in both the digital and classroom environments.

While this study offers valuable insights into how usable Iranian math apps are for children, there are a few limitations that should be considered to better understand the findings and their relevance. By recognizing these limitations, we can better understand the study's context and scope.

The research focused only on math apps from two Iranian Android app stores—Bazar and Myket. This means the study might not fully capture the broader range of educational apps available in Iran, especially those on global platforms like the Google Play Store or the iOS App Store. Additionally, since only math apps were studied, the findings may not apply to apps for other subjects, like science or reading.

Each app was tested for only 15 minutes, which might not be enough time to fully understand its usability. Some issues, like whether children get tired of the app, how easy it is to learn to use, or how engaging it remains over time, may not have been noticeable in such a short test. A longer testing period could reveal more about how children interact with these apps in the long run.

This study focused on apps designed for kids aged 6 to 9. As a result, the findings may not apply to apps meant for older children. Usability needs vary by age, so future studies could explore how apps perform for children of different ages to get a broader picture.

Since this study focused on Iranian apps, the findings are shaped by local cultural, educational, and technological factors. While some usability principles are universal, the way children learn and interact with apps can vary across different cultures, which limits how well these findings apply to educational apps in other regions.

Another limitation is that the study did not involve direct user testing with children. Instead, the apps were evaluated by a single researcher. Watching how children interact with the apps in real-life settings could have provided richer insights into how usable they are and might have uncovered more usability challenges that weren't evident through researcher evaluation alone.

Future research could expand by including a broader range of apps, and testing the apps directly with children to gather more comprehensive findings.

Since this study focused on Iranian apps, future research could look at how educational apps from other countries perform. A global comparison could uncover best practices that developers worldwide can use to improve their apps.

Educational bodies have an important role to play in ensuring these improvements are made. By setting clear standards for educational apps and promoting best practices, they can help bridge the gap between research and practical application, leading to better-designed, more effective learning tools. With ongoing research and collaboration between educators, developers, and researchers, we can improve the quality of educational apps for children, giving them the tools they need to succeed in today's digital world.

Author Contributions

N. Zanjani, designed the research, collected the data, carried out the data analysis, interpreted the results and wrote the manuscript.

Acknowledgment

The author would like to express sincere gratitude to the developers and publishers of the educational apps evaluated in this study, whose work provided the foundation for this analysis.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

HCI	Human-Computer Interaction
IDC	Interaction Design and Children
AR	Augmented Reality
TIDRC	Touchscreen Interaction Design Recommendations for Children
E.T.E.A.	Evaluation Tool for Educational Apps
EFA	Exploratory Factor Analysis

References

[1] H. Özgür, "Relationships between computational thinking skills, ways of thinking and demographic variables: A structural equation modeling," Int. J. Res. Edu. Sci., 6(2): 299-314, 2020.

[2] S. Annamalai, A. C. Omar, S. N. A. Salam, " Rory's Mathematics Adventure's (ROMAAD) mobile game-based learning applicatioN: AN evaluation of usabilitY," Int. J. Edu. Psychol. Couns., 7(48): 575-85, 2022.

[3] B. S. Nugraha, M. Suyanto, E. Utami, "Innovative approaches in child-friendly user interfaces: A systematic literature review on technologies, motoric skill and evaluation," Presented at 7th International Conference of Computer and Informatics Engineering (IC2IE), 2024.

[4] B. Shneiderman, C. Plaisant, Designing the User Interface: Strategies for Effective Human-Computer Interaction: Pearson Education India, 2010.

[5] B. Al-Haimi, "Usability guidelines of mobile learning application," J. Inf. Syst. Res. Innovation, 5(9), 2013.

[6] S. Yağmur, M. P. Çakır, "Usability evaluation of a dynamic geometry software mobile interface through eye tracking" Presented at the third International Conference on Learning and Collaboration Technologies held as Part of HCI International, 2016.

[7] M. L. Courage, L. M. Frizzell, C. S. Walsh, M. Smith, "Toddlers using tablets: They engage, play, and learn", Front. Psychol., 12: 1-18, 2021.

[8] B. Auxier, M. Anderson, A. Perrin, E. Turner, "Children's engagement with digital devices, screen time", Pew Research Center, 2020.

[9] N. Soni, A. Aloba, K. S. Morga, P. J. Wisniewski, L. Anthony, "A framework of touchscreen interaction design recommendations for children (tidrc) characterizing the gap between research evidence and design practice," in Proc. the 18th ACM International Conference on Interaction Design and Children, 2019.

[10] D. F. Murad, B. D. Wijanarko, R. Leandros, F. I. Ramadhan, Y. A. P. Siregar, "Interaction design of mathematics learning applications for elementary school students," Presented at 3rd International Symposium on Material and Electrical Engineering Conference (ISMEE), 2021.

[11] M. Martens, "Issues of access and usability in designing digital resources for children," Lib. Inf. Sci. Res., 34(3): 159-68, 2012.

[12] A. Ibrahim, M. Al-Rajab, K. Hamid, M. Aqeel, S. Muneer, M. Parveen et al., "Usability evaluation of kids' learning apps," Presented at International Conference on Business Analytics for Technology and Security (ICBATS), 2023.

[13] P. Markopoulos, J. C. Read, S. MacFarlane, J. Hoysniemi, Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers: Elsevier, 2008.

[14] M. Meyer, J. M. Zosh, C. McLaren, M. Robb, H. McCaffery, R. M. Golinkoff et al., "How educational are "educational" apps for young children? App store content analysis using the Four Pillars of Learning framework," J. Children Media, 15(4): 526-48, 2021.

[15] R. Binns, U. Lyngs, M. Van Kleek, J. Zhao, T. Libert, N. Shadbolt, "Third-party tracking in the mobile ecosystem," Proceedings of the 10th ACM Conference on Web Science; 2018.

[16] C. A. C. Domínguez, D. O. Mina, V. Agredo-Delgado, P. H. Ruiz, D, M, AlSekait, "Towards to usability guidelines construction for the design of interactive mobile applications for learning mathematics," Iberoamerican Workshop on Human-Computer Interaction, 2020.

[17] S. Papadakis, J. Vaiopoulou, M. Kalogiannakis, D. Stamovlasis "Developing and exploring an evaluation tool for educational apps (ETEA) targeting kindergarten children," Sustainability, 12(10): 4201, 2020.

[18] J. H. D. Doong, Loneliness among Chinese Emerging Adults in America and the Role of the Church: A Practical Theology Inquiry: Fuller Theological Seminary, Center for Advanced Theological Study, Dissertations & Theses, 3733234, 2015.

[19] N. Vanderschantz, A. Hinze, "Designing an internet search interface for children," Presented at 32nd BCS Human Computer Interaction Conference (BCS HCI 2018), 2018.

[20] B. H. Chandana, N. Shaik, P. Chitralingappa, "Exploring the frontiers of user experience design: VR, AR, and the future of interaction," Presented at International Conference on Computer Science and Emerging Technologies (CSET), 2023.

[21] A. Banke, C. Lauff, "Usability testing with children: History of best practices, comparison of methods & gaps in literature," DRS Biennial Conference Series, DRS2022: Bilbao, 2022.

[22] G. Ragone, P. Buono, R. Lanzilotti, "Designing safe and engaging ai experiences for children: Towards the definition of best practices in UI/UX Design," arXiv preprint arXiv:240414218, 2024.

[23] A. K. Sandhu, K. Bhardwaj, "Interfaces for kids," in Proc. the 11th Asia Pacific Conference on Computer Human Interaction, 2013.

[24] N. Tuli, A. Mantri, "Evaluating usability of mobile-based augmented reality learning environments for early childhood," Int. J. Hum. Comput. Interac., 37(9): 815-27, 2021.

[25] L. Anthony, Q. Brown, J. Nias, B. Tate, "Examining the need for visual feedback during gesture interaction on mobile touchscreen devices for kids," in Proc. the 12th International Conference on Interaction Design and Children, 2013.

[26] M. Meyer, V. Adkins, N. Yuan, H. M. Weeks, Y. J Chang, J. Radesky, "Advertising in young children's apps: A content analysis," J. dev. Behav. Pediatr., 40(1): 32-39, 2019.

[27] N. Holmberg, Effects of online advertising on children's visual attention and task performance during free and goal-directed internet use: A media psychology approach to children's website interaction and advert distraction, Lund University, 2016.

[28] M. M. S. Missen, A. Javed, H. Asmat, M. Nosheen, M. Coustaty, N. Salamat et al., "Systematic review and usability evaluation of writing mobile apps for children," New Rev. Hypermedia Multimedia, 25(3): 137-60 2019.

[29] Y. Cakan, A. Kaya, C. A. Gumussoy, Usability Analysis of a Mobile Learning Application. Handbook of Research on Modern Educational Technologies, Applications, and Management: IGI Global, p. 198-212, 2021.

[30] A. Câmpean, M. Bocoș, A. Roman, D. Rad, C. Crișan, M. Maier et al., "Examining teachers' perception on the impact of positive feedback on school students," Edu. Sci., 14(3): 257, 2024.

[31] M. Merrick, E. Fyfe, "Should I stay or should I go? Children's motivation in response to feedback and its association with math self-concept, math self-efficacy, and math anxiety," 2024.

[32] M. Maryana, C. Halim, H. Rahmi, "The impact of gamification on student engagement and learning outcomes in mathematics education," Int. J. Bus. Law Edu., 5(2): 1697-1708, 2024.

[33] L. Y. Fei, M. R. Norfariza, E. H. Kazi, "Student engagement as a mediator between online classroom management and learning outcomes in Beijing universities," Int. J. Eval. Res. Edu., 14(1): 94, 2024.

[34] Q. Yu, J. Wang, "A framework design of children's educational app based on metacognitive theory," Presented at International Conference on Human-Computer Interaction, 2022.

[35] N. Blom, "Design cognition in design and technology classrooms," Debates in Design and Technology Education, Routledge, p. 209-220, 2022.

[36] A. Ani, "Positive feedback improves students' psychological and physical learning outcomes," Indones. J. Educ. Stud., 22(2), 2019.

Biographies

Nastaran Zanjani is a faculty member of the Computer Engineering Department at Refah College University. She holds a Bachelor's degree in Electrical Engineering with a specialization in Electronics from the Shahid Beheshti University. She also has a Master's degree in Telecommunications Engineering from Khaje Nasir Toosi University and a Ph.D. in Information Technology from the Queensland University of Technology in Australia. Her areas of expertise include information technology and HCI.

- Email: Zanjani@refah.ac.ir
- ORCID: 0000-0002-5307-683X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://refah.ac.ir/cv/81/zanjani

How to cite this paper:
N. Zanjani, "Usability of Iranian math apps for kids," J. Electr. Comput. Eng. Innovations, 13(2): 473-484, 2025.

DOI: 10.22061/jecei.2025.11596.818

URL: https://jecei.sru.ac.ir/article_2316.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Designing Multiband, Reconfigurable Printed Antenna for Modern Communication Systems

M. Zahiry [1], ***S. M. Hashemi*** [1,*], ***J. Ghalibafan*** [2]

[1]*Department of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

[2]*Department of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran.*

Article Info

Article History:
Received 03 February 2025
Reviewed 10 March 2025
Revised 15 April 2025
Accepted 22 April 2025

Keywords:
CPW-fed monopole antenna
Impedance matching stubs
Multiband
Reconfigurable

*Corresponding Author's Email Address: *sm.hashemi@sru.ac.ir*

Abstract

Background and Objectives: Printed monopole antenna has an omnidirectional radiation pattern but a narrow impedance bandwidth. To achieve multiband behavior, modification of the antenna shape is necessary. There is no systematic approach for the shape modification and commonly it is done by parametric studies, adjusting, tuning or optimization of an initial shape.

Methods: The method for designing a multiband antenna is based on transmission line theory and lumped element model. It applies to all desired multiband operations without needing tuning or optimizing a complex configuration. Every length and dimension are computed from the mathematical formula or Smith chart as a graphical tool. The proposed matching method in the design of a multiband and reconfigurable CPW-fed monopole antenna employed. Both series and parallel impedance matching stubs are investigated and compared with each other.

Results: To explain the challenges, the proposed method was applied to a desired antenna. It showed that using matching stubs in the CPW line can design a multiband and reconfigurable antenna. Also, the measurements have been done and compared with the simulations and show a good agreement.

Conclusion: Compared to the microstrip line, the CPW feeding of the monopole antenna has the advantage that both parallel and series stubs can be implemented for the matching of the antenna. In this case, the required space for these stubs placed inside the antenna and no extra space needed. So, the printed size of the proposed antenna does not change. The impedance matching method for the integrated stubs with the antenna has been proposed. The high-pass and low-pass properties of each matching network were considered. The authors showed that this method can successfully design multiband, reconfigurable CPW-fed monopole antennas.

Introduction

Today, the use of different frequency bands in a single device for access to various telecommunication services is very common. An Antenna as a part of the wireless systems, play an important role in achieving this goal. The appropriate antenna should have a multiband operation and an omnidirectional pattern, where full spatial coverage is required. One of prevalent and simple way to get an omnidirectional pattern is the monopole antenna. But the monopole antenna due to the resonance characteristic is very sensitive to frequency and does not have multiband properties. So, a lot of changes have been made to the monopole antenna shape by researchers to overcome this problem [1]-[16]. In order to attain

multiband properties, in reference [1] two connected strip monopoles with different lengths are tuned, in [2] two stacked T-shaped monopoles with different sizes are adjusted, in [3] the dimensions of G-shaped monopole are optimized, in [4] the geometry of shorted parasitic inverted-L wire closely placed to the inverted L-shaped monopole is founded by tuning, in [5] appropriate slits into the CPW feeding line are optimized by particle swarm optimization, in [6] dimensions of modified T-shaped monopole are obtained by parametric studies, in [7] the lengths of the monopole antenna and an n-shaped slot on the radiating element properly adjusted, in [8] the parameters of two inverted-L slots etched on the radiator element optimized and tuned. In [9]-[24] by using a more complicated structure and shape for the monopole antenna, the multiband operation has been achieved.

Coplanar waveguide (CPW) implementation of microwave and antenna devices is widely considered in research. This is due to its attractive features such as a single metallic layer, low dispersion and loss, more design parameters for impedance matching, low radiation losses, and easy connection of series and shunt components. Some of the antenna mentioned in the past are fed by the CPW line. Furthermore, the CPW line also used in microwave devices such as filter [25], phase shifters [26], power splitters [27], amplifier [28], coupler [29] and etc. The matching stubs described here can be used for a better design of them.

The multiband antenna design method described here is very simple and completely applicable to all of the desired multiband operations without the need to tuning or optimizing a complex configuration. Every length and dimension are computed from the mathematical formula or Smith chart as a graphical tool. Because the shape of the monopole antenna does not change and the matching network integrated with the antenna structure, no need to increase in the monopole size.

In this paper, first in Section II the theory of impedance matching for CPW line based on the single-stub tuning developed and customized for CPW-fed monopole antenna. In this way, the effects of both series and parallel impedance matching stubs modeled with transmission line theory. Because the monopole antenna can be matched in the first operational band through the antenna length and proper feeding line, the matching network used for the second operational band and should not have any adverse effect on the first one. This can be explained through the concept of high-pass and low-pass properties of each matching network. To have a better view, each stub before the first resonance frequency modeled with the lumped elements.

In Section III, the proposed matching method in the design of a multiband and reconfigurable CPW-fed monopole antenna employed. Both series and parallel impedance matching stubs are investigated and compared with each other. Also, the effect of the difference between the upper and lower frequency bands on the success of the proposed method investigated. The tuning of the second band with the length of the stubs has been done and the independence of the two frequency bands is studied. For all case, the measurement results are compared with the Feko software simulations and show good agreement.

Matching Stubs in CPW Line

Single stub matching network is based on a short or open- ended transmission line (stub) with a specified length which connected to the feeder line at a certain position in the form of serial or parallel. Due to fabrication constraints, the parallel stub can be implemented in the microstrip line but the serial one not. For the CPW transmission line, both serial and parallel connections can be considered, due to presence of ground and center conductor in the same plane. The theory of single stub matching is well known [30]. But for using this theory in the design of the CPW-fed monopole antenna, we need to clarify some of the points. One of them is about the rotation direction in the Smith chart (toward the generator or load) that will be different here with usual impedance matching methods.

The other one is about stub termination (short or open) which must be selected in serial and parallel connection of stubs. The termination choosing for serial or parallel connection is related to the high-pass and low-pass properties of matching network. To design a multiband monopole antenna, the first operational band was matched (with a length of about quarter-wavelength) and for the second band, the matching stub must be designed so that the first one not disturbed by the matching network.

A. Modeling of Serial and Parallel Stubs

Let's consider the Fig. 1 (a) which part of center strip conductor expanded into surrounding ground and ended without connecting to it. Hence, this part can be modeled as an open-end CPW parallel stub (see Fig. 1 (b)). For circuit analysis based on the Smith chart, moving from the open circuit gives rise to a capacitive reactance that must be connected in parallel with the main CPW line (see Fig. 1 (c)).

The input impedance of this stub can be calculated from transmission line theory as:

$$Z_{in} = -jZ_0\, cot(\beta l_p) \tag{1}$$

where Z_0 is the open-end CPW parallel stub characteristic impedance (here we assumed that it is equal to the main line), l_p is the physical length of it and β is the phase constant. The values of Z_0 and β dependent on the physical dimensions can be found in [31].

If the stub, before the quarter-wavelength, modeled as a capacitance, then from (1) we have:

$$C = \frac{tan(\beta l_P)}{\omega Z_0} \quad (2)$$

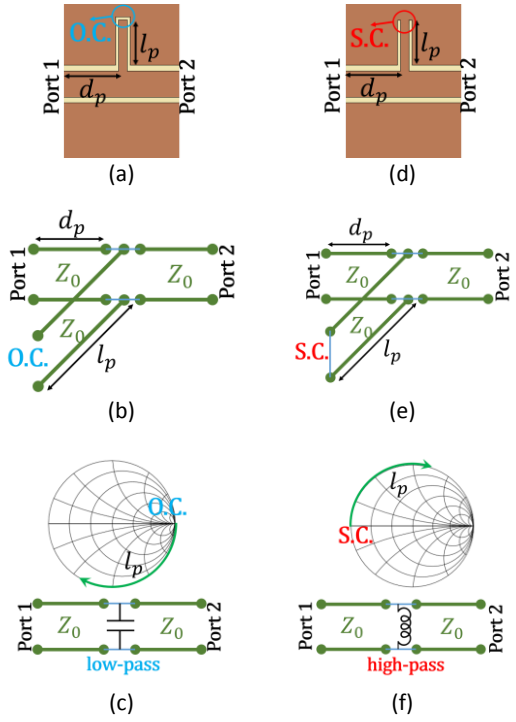


Fig. 1: Parallel stubs in CPW transmission line. Open-end CPW parallel stub (a) board layout, (b) equivalent transmission line and (c) equivalent circuit element with Smith chart. Short-end CPW parallel stub (d) board layout, (e) equivalent transmission line and (f) equivalent circuit element with Smith chart.

For Fig. 1 (d) the center strip conductor expanded into surrounding ground and connected to it to form a short circuit. Similarly, this part can be modeled as a short-end CPW parallel stub (see Fig. 1 (e)). In this case, moving from the short circuit gives rise to an inductive reactance that must be connected in parallel with the main CPW line (see Fig. 1 (f)). To explain the inductive properties of this stub, we must consider the current following around the end of termination like a loop. The input impedance of this short parallel stub can be calculated from transmission line theory as:

$$Z_{in} = jZ_0\, tan(\beta l_p) \quad (3)$$

where the parameters are defined similar to (1). The equivalent inductance value, before the quarter-wavelength, determined by:

$$L = \frac{Z_0\, tan(\beta l_p)}{\omega} \quad (4)$$

The values of C and L in (2) and (4) are a function of l_p (stub length), Z_0 and β.

In Fig. 2 (a), the center strip changed so that to form an interdigital structure. In circuit analysis, this structure can be modeled obviously as a series capacitance (see Fig. 2 (c)) and also based on transmission line theory it can be considered as a series open circuit CPW stub (see Fig. 2 (b)). So, the input impedance of the stub and the equivalent capacitance value, before the quarter-wavelength, can be computed with (1) and (2) respectively. The difference is that this open circuit stub, connected in series with the main CPW line (compare Fig. 1 (a) and Fig. 2 (a)).

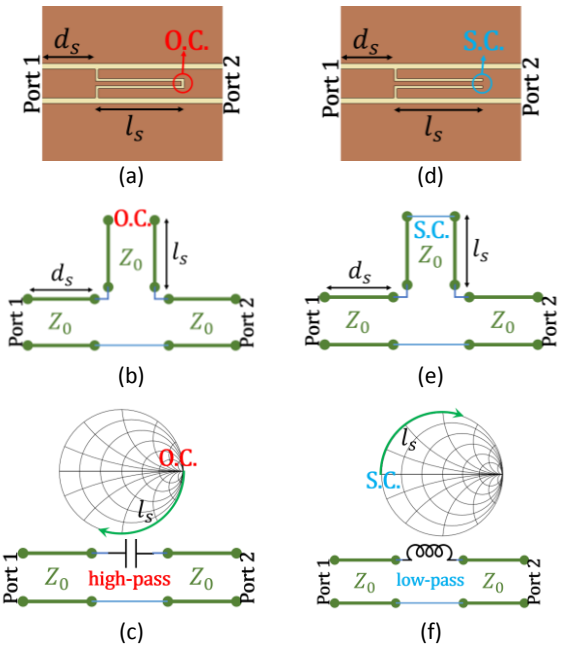


Fig. 2: Series stubs in CPW transmission line. Open-end CPW series stub (a) board layout, (b) equivalent transmission line and (c) equivalent circuit element with Smith chart. Short-end CPW series stub (d) board layout, (e) equivalent transmission line and (f) equivalent circuit element with Smith chart.

In Fig. 2 (d), the interdigital structure shorted in the end and form a short-end CPW series stub (see Fig. 2 (e)). Based on the Smith chart, moving from the short circuit gives rise to an inductive reactance that must be connected in series with the main CPW line (see Fig. 2 (f)). So, the input impedance of the stub and the equivalent inductance value, before the quarter-wavelength, can be computed with (3) and (4) respectively.

To achieve a multiband antenna, the matching network used for the second operational band, should not have any adverse effect on the first one. This simple lumped element model is useful when we must consider the effect of different stubs on the first operational bandwidth based on the high-pass and the low-pass properties of each matching network. Thus, we can conclude that the stubs shown in Fig. 1 (a) and Fig. 2 (d) are low-pass and suitable for matching of the second operational band, but the others are not useful for our purpose.

The simple lumped element equivalent circuit model is

accurate before the first resonance frequency (the quarter-wavelength) and does not predict the resonance. But the transmission line model does it (see Fig. 3). In Fig. 3 (a) and (b) the $S_{21}$of low-pass CPW stubs simulated in full wave (FW) and compared with the transmission line (TL) model. The values of capacitance for open-end CPW parallel stub computed with (2) and illustrated in Fig.3 (a). The FW simulation for parallel stub agrees with TL model properly.

But for CPW series stub, the resonance frequency is shifted (see Fig. 3 (b)). It is because of the low width of surrounding ground for CPW series stub in comparison with the common CPW line. The values of inductance for short-end CPW series stub computed with (4) and illustrated in Fig. 3 (b). In Fig. 3 (c) the structure of low-pass CPW stubs for multi-band monopole antenna illustrated.

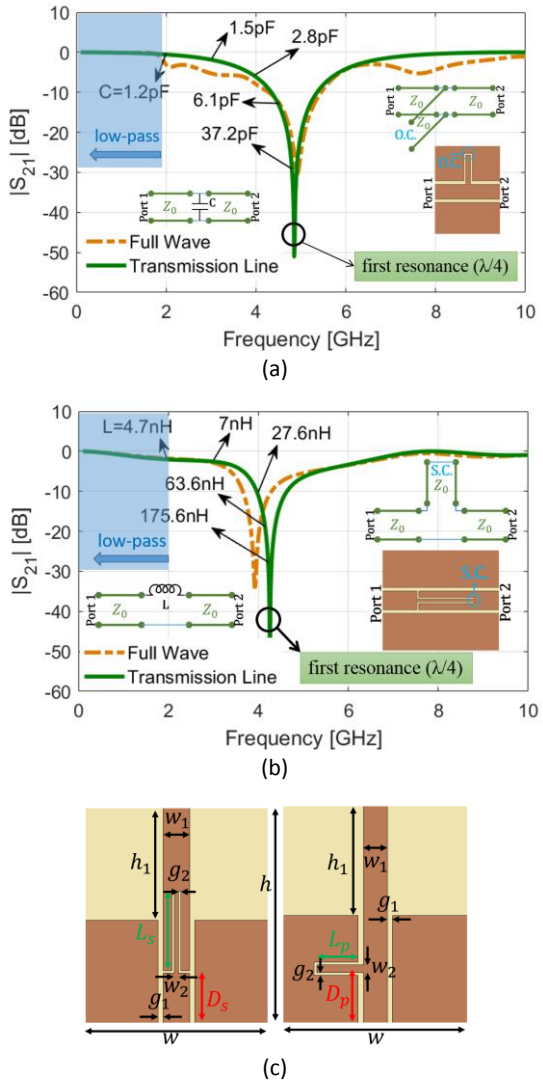


Fig. 3: The low-pass CPW stubs. The S_{21} magnitude of (a) open-end CPW parallel stub and (b) short-end CPW series stub. The FW and TL simulations compared and equivalent circuit element values before first resonance computed. (c) The low-pass CPW stubs solution for multi-band monopole antenna.

The D_S (D_P) and L_S (L_P) dimensions as the series (parallel) stub position and length must be designed to achieve multiband antenna. The following is an explanation of the design process with help of some examples.

B. CPW Stubs Design for Multi-Band Monopole Antenna

At the first, a CPW fed monopole antenna designed for the first frequency band based of the common method (with a length of about quarter-wavelength). Then the input impedance of the antenna gets through simulation at all frequencies. In this step, for more accurate results, the subminiature version A (SMA) connector is also considered in simulations. Now the input impedance of the antenna in the second desired band has been achieved and the matching problem is specified. To match this load to a 50Ω line the Smith chart can be used as a graphical tool. Since the goal of this work is the integration of matching circuit with the antenna, rotation direction in the Smith chart chosen toward the load (inside the antenna).

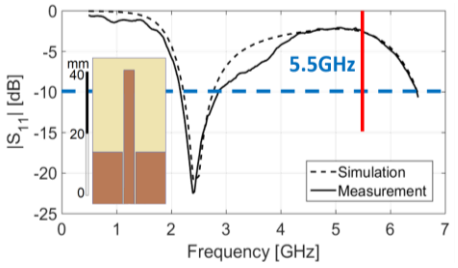
For parallel stubs case, the admittance Smith chart could be used and the normalized load must be plotted. The D_P calculated from the SWR circle intersection with the $1 + jb$ circle. The length of the open-end parallel stub (L_P) that gives a susceptance of $-jb$ can be found on the Smith chart. The design process for the CPW series stubs is similar to the above.

Let's assume the first frequency band of the required antenna is 2.4 GHz and it is desirable to be matched in 5.5 GHz. The monopole antenna with such properties is illustrated in Fig. 4 (a). As can be seen, the antenna in 5.5 GHz poorly matched and it is required to a matching network. Now the input impedance of the antenna in the 5.5 GHz can be normalized and plotted on the Smith chart.

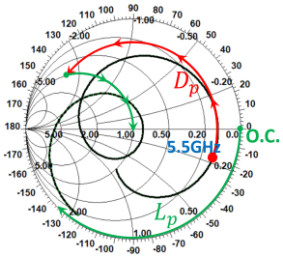
Because the SMA connector modeled when simulating the antenna, the input impedance computed in the place of the SMA input as a wave port. Thus, the de-embedding result (removing the influence of the SMA connector) leads to 5.5 GHz point on the Smith chart rotate counterclockwise (CCW). The rotated 5.5 GHz point plotted on the Smith chart as illustrated in Fig. 4 (b).

First parallel CPW stubs case is explained. The rotation along a constant radius SWR circle chosen CCW to embed the matching network into the antenna board. This brings the solution to a point on the $1 + jb$ circle. Here, due to limited space available in antenna board, just the first intersection point which leads to the shortest distance (D_P) considered. The D_P calculated from the wavelengths toward load (WTL) scale. In this desired antenna, the normalized admittance at the intersection point is $1 + jb$ ($b < 0$) as illustrated in Fig. 4 (b). Thus, the solution requires a parallel stub with a susceptance of $-jb$. The length of the open-end parallel CPW stub (L_P) that lead to a susceptance of $-jb$ can be found through starting at

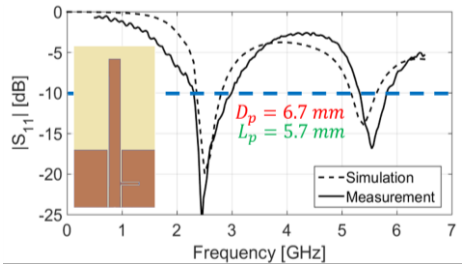
$y = 0$ (open circuit), and moving along the outer edge of the Smith chart ($|\Gamma| = 1$) toward the generator to the $-jb$ point. So, the both of D_P and L_P have been achieved and the design of matching network completed. The printed circuit board and the radiation patterns of parallel stub solution illustrated in Fig. 5.



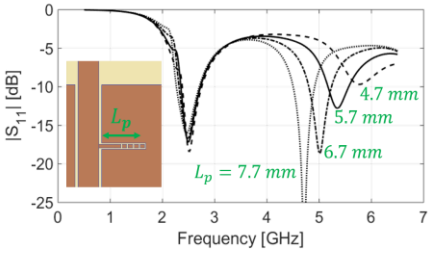
(a)



(b)



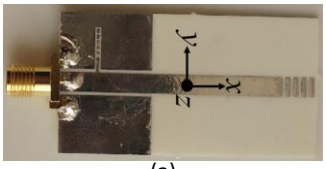
(c)



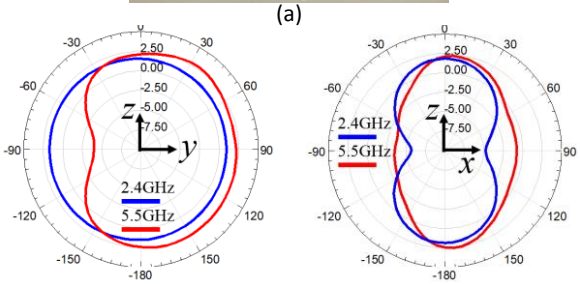
(d)

Fig. 4: The open-end CPW parallel stub design process and results for 5.5 GHz. (a) Simulation and measurement of primary 2.4 GHz CPW-fed monopole antenna. (b) The Smith chart graphical solution to find D_p and L_p for the open-end CPW parallel stub. (c) Simulation and measurement of parallel stub solution in CPW transmission line for impedance matching in 5.5 GHz. (d). The effect of different values of stub length (L_p) on the second band to have reconfigurable antenna.

For the series CPW stub case, a same matching problem assumed (both Fig. 4 (a) and Fig. 6 (a) are the same). In this case the impedance Smith chart could be used (see Fig. 6 (b)).

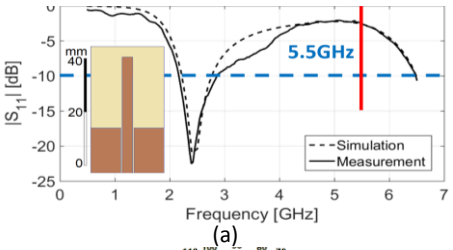


(a)

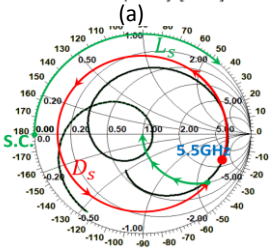


(b) (c)

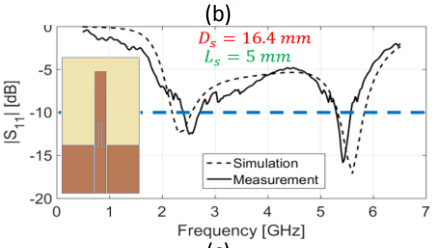
Fig. 5: Printed circuit board (Ro4003 ε_r=3.55, thickness=0.8mm), (a) and simulated radiation patterns of parallel stub solution for impedance matching in 2.4 GHz and 5.5 GHz, (b) *yz* (c) *xz* plane. ($h_1 = 30mm, w_1 = 3.5mm, g_1 = 0.5mm$)



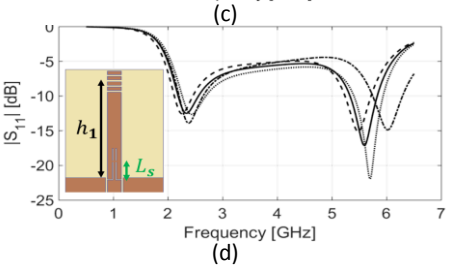
(a)



(b)



(c)



(d)

Fig. 6: The short-end CPW series stub design process and results for 5.5 GHz. (a) Simulation and measurement of primary 2.4 GHz CPW-fed monopole antenna. (b) The Smith chart graphical solution to find D_p and L_p for the short-end CPW series stub. (c) Simulation and measurement of series stub solution in CPW transmission line for impedance matching in 5.5 GHz. (d) The effect of different values of stub length (L_s) and monopole antenna length (h_1) on the second band to have reconfigurable antenna.

Again, the rotation along a constant radius SWR circle chosen CCW and this leads to intersection with a point on the $1 + jx$ circle. In our example, the first intersection point leads to a small length for D_S (see Fig. 3 (c)) and the assembly of the SMA connector is difficult. Thus, the rotation continues along a constant radius SWR circle to intersect with second point on the $1 + jx$ circle (see Fig. 6 (b)). In this desired antenna, the normalized impedance at the intersection point is $1 + jx$ $(x < 0)$ as illustrated in Fig. 6 (b). Thus, the solution requires a series stub with a reactance of $-jx$. The length of the short-end series CPW stub (L_S) that lead to a reactance of $-jx$ can be found through starting at $z = 0$ (short circuit), and moving along the outer edge of the Smith chart ($|\Gamma| = 1$) toward the generator to the $-jx$ point. So, the both of D_S and L_S have been achieved and the design of matching network with series CPW stub completed.

In the next section, this method implemented and the simulation and measurement result presented.

Result and Discussions of Reconfigurable Multiband Monopole Antenna

This section illustrates that the implementation of matching stubs in the CPW line with the proposed method can be used for design of a multiband and reconfigurable antenna. Here are a few examples to explain the challenges and to compare the both series and parallel methods.

In the first case, assume the frequency band of the required antenna is 2.4 GHz and it is desirable to be matched in 5.5 GHz (see Fig. 4 (a)). The usage of the Smith chart as a graphical tool for matching the antenna illustrated in Fig. 4 (b). The designed D_P and L_P dimensions as the parallel stub position and length are 6.7 mm and 5.7 mm respectively. The results of the simulation and measurement of the designed antenna have been shown in Fig. 4 (c) and as can be seen, there is a good agreement between them. Thus, the validation of the method described in the previous section approved. Due to manufacturing constraints, the parallel stub position cannot be changed, so the D_P has a fixed value. But the parallel stub length (L_P) can be changed. The effect of L_P variation on the input impedance matching of the antenna illustrated in Fig. 4 (d). The variable L_P implemented here by connecting of the pads through the copper ribbon. As can be seen in Fig. 4 (d), by adjusting the L_P value the second band of the antenna tuned as a reconfigurable antenna. The advantage of this implementation is that the first frequency band remains unchanged and the two frequency bands are independent (see Fig. 4 (d)). It is evident that the first frequency band is dependent on monopole length. The printed circuit board and the radiation patterns of parallel stub solution illustrated in Fig. 5.

For the series CPW stub case, again assume the frequency band of the required antenna is 2.4 GHz and it is desirable to be matched in 5.5 GHz (see Fig. 6 (a)). By using the Smith chart (see Fig. 6 (b)), the D_S and L_S dimensions as the series stub position and length, attain 16.4 mm and 5 mm respectively. Since the D_S has a fixed value, only the series stub length (L_S) can be changed by shortening the stub in appropriate length. Unlike the parallel stub, the first and second frequency band for series stub are not independent. Thus, to freeze the first frequency band, both of L_S and h_1 need to be tuned and it is a difficult task (see Fig. 6 (c)). The printed circuit board of series stub solution and the radiation patterns illustrated in Fig. 7.

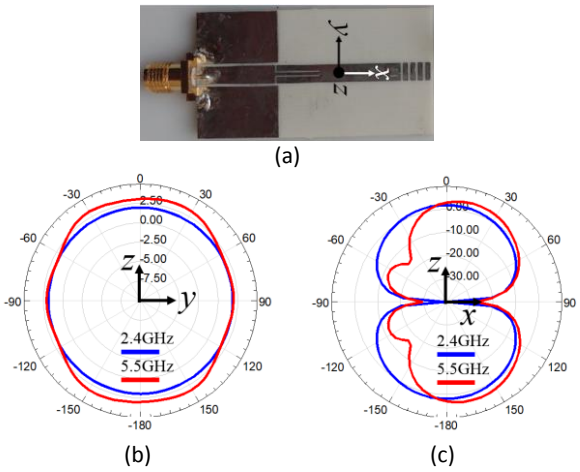


Fig. 7: Printed circuit board (a) and simulated radiation patterns of series stub solution for impedance matching in 2.4 GHz and 5.5 GHz, (b) *yz* (c) *xz* plane. ($h_1 = 18mm, w_1 = 3.5mm, g_1 = 0.5mm$)

To investigate the effect of the difference between the upper and lower frequency bands on the success of the proposed method, let's consider another problem and assume the frequency band of the required antenna is 3.5 GHz and it is desirable to be matched in 5.5 GHz. In this case, the low pass properties of the matching network are more important to have no adverse effect on the first band of the antenna (see Fig. 8). The matching stubs for both series and parallel cases designed with the help of the Smith chart (see Fig. 8. (b) and (c)). The designed $D_P(D_S)$ and $L_P(L_S)$ dimensions as the parallel (series) stub position and length are 3 (12) mm and 5.5 (3.7) mm respectively.

For parallel stub, in addition to the 5.5 GHz, another frequency band around 2.4 GHz has been matched by chance (see Fig. 9 (a)). But the authors check and found that the additional matched band is not due to the parasitic effect and the parallel stub length and position are also suitable for matching in 2.4 GHz. In the series stub case, the 3.5 GHz and 5.5 GHz bands are near to each other and merged. On the other hand, a wideband antenna instead of multiband antenna has been achieved

(see Fig. 9 (b)). The printed circuit board of parallel and series stub solution in CPW-fed monopole antenna for impedance matching in 5.5 GHz illustrated in Fig. 9.

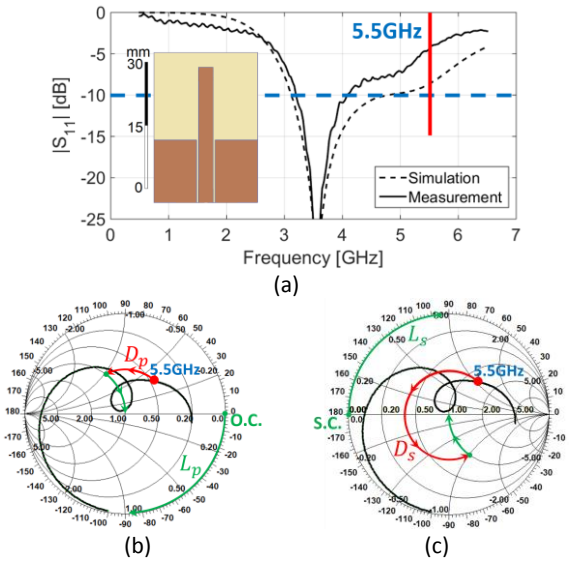


Fig. 8: (a) Simulation and measurement of antenna for 3.5 GHz and it is desirable to be matched in 5.5 GHz, (b), (c) The matching stubs for both series and parallel cases designed with the help of the Smith chart.

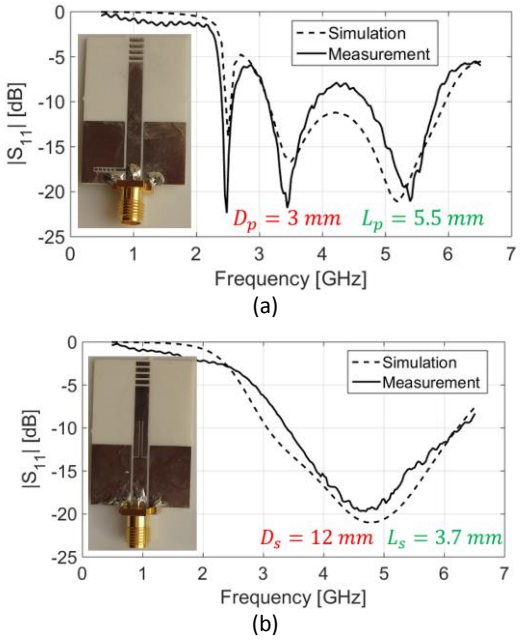


Fig. 9: Simulation and measurement of (a) parallel stub solution, (b) series stub solution.

Since, this work is about the antenna design method, a comparison has been made with previous works in this field, which illustrated in Table 1.

Table 1: Comparison with other works

Ref	Antenna Shape	Design Approach	Bandwidth	Reconfigurable
[1]	Two monopoles with different lengths	Tuning the antenna parameters	Dual-Band	NA
[2]	Two stacked T-shaped monopoles of different sizes	Adjusting the antenna parameters	Dual-Band	NA
[3]	G-shaped profile	The antenna dimensions are optimized	Dual-Band	NA
[4]	Shorted parasitic inverted-L	Parametric studies	Triple-Band	NA
[5]	Embedding appropriate slits into the 50 Ω CPW feeding line	Particle swarm optimization approach	Multiband	NA
[6]	Modified T-shaped antenna	Parametric studies	Dual-Band	NA
[7]	n-shaped slot	Adjusting the lengths of the element and slot	Dual-Band	NA
[8]	Two inverted-L slots	Adjusting the lengths of the element and slot	Triple-Band	NA
This work	Parallel and series stubs	Serial and parallel stubs Using Smith chart	Dual-Band Triple-Band Wide-Band	Yes

The design methods in previous works are based on changing the antenna parameters to reach the desired solution, and the ability to change the frequency by changing the dimensions is limited. Therefore, the design method is more complicated and the reconfigurable capability of the antenna is reduced.

Also, in some previous works, the size of the antenna increases due to the change in the shape of the antenna which in some cases leads to a decrease in the omnidirectionality of the antennas. The method presented in this work allows the implementation of dual-band, triple-band and wide-band antennas without

changing the shape of the antenna and with a simple design method.

Conclusion

Compared to the microstrip line, the CPW feeding of the monopole antenna has the advantage that both parallel and series stubs can be implemented for the matching of the antenna. In this case, the required space for these stubs placed inside the antenna and no extra space needed. So, the printed size of the proposed antenna does not change. The impedance matching method for the integrated stubs with the antenna has been proposed. The concept of high-pass and low-pass properties of each matching network considered and the authors showed that this method can be successfully used for designing of multiband, reconfigurable CPW-fed monopole antenna.

Author Contributions

M. Zahiry carried out the simulation, S. M. Hashemi has presented the idea of the research and J. Ghalibafan contributed to the analysis of results and implementation. All authors contributed to the writing of the article.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Acknowledgment

We appreciate the editorial team and reviewers for their time and constructive feedback.

Abbreviations

CPW	Coplanar Waveguide
SMA	Subminiature Version A
SWR	Spectral Angle Mapper
CCW	Counterclockwise
TL	Transmission Line
FW	Full Wave
WTL	Wavelengths Toward Load

References

[1] H. M. Chen, Y. F. Lin, C. C. Kuo, K. C. Huang, “A compact dual-band microstrip-fed monopole antenna,” in Proc. IEEE Antennas and Propagation Society International Symposium. 2001 Digest. Held in conjunction with: USNC/URSI National Radio Science Meeting (Cat. No.01CH37229), 2: 124-127, 2021.

[2] Y. L. Kuo, K. L. Wong, “Printed double-T monopole antenna for 2.4/5.2 GHz dual-band WLAN operations,” IEEE Trans. Antennas Propag., 51(9): 2187-2192, 2003.

[3] C. Y. Pan, C. H. Huang, T. S. Horng, “A new printed G-shaped monopole antenna for dual-band WLAN applications,” Microw. Opt. Technol. Lett., 45: 295-297. 2005.

[4] J. Y. Jan, L. C. Tseng, “Small planar monopole antenna with a shorted parasitic inverted-L wire for wireless communications in the 2.4-, 5.2-, and 5.8-GHz bands,” IEEE Trans. Antennas Propag., 52(7): 1903-1905, 2004.

[5] W. C. Liu, “Design of a multiband CPW-fed monopole antenna using a particle swarm optimization approach,” IEEE Trans. Antennas Propag., 53(10): 3273-3279, 2005.

[6] S. B. Chen, Y. C. Jiao, W. Wang, F. S. Zhang, “Modified T-shaped planar monopole antennas for multiband operation,” IEEE Trans. Microwave Theory Tech., 54(8): 3267-3270, 2006.

[7] S. T. Fan, Y. Z. Yin, W. Hu, K. Song, B. Li, “Novel CPW-FED printed monopole antenna with an n-shaped slot for dual-band operations,” Microw. Opt. Technol. Lett., 54: 240-242. 2012.

[8] H. Chen, X. Yang, Y. Z. Yin, S. T. Fan, J. J. Wu, “Triband planar monopole antenna with compact radiator for WLAN/WiMAX applications,” IEEE Antennas Wireless Propag. Lett., 12: 1440-1443, 2013.

[9] X. L. Sun, S. W. Cheung, T. I. Yuk, “Dual-Band monopole antenna with frequency-tunable feature for WiMAX applications,” IEEE Antennas Wireless Propag. Lett., 12: 100-103, 2013.

[10] X. Ren, S. Gao, Y. Yin, “Compact tri-band monopole antenna with hybrid strips for WLAN/WiMAX applications,” Microw. Opt. Technol. Lett., 57: 94-99. 2015.

[11] P. Beigi, J. Nourinia, Y. Zehforoosh, B. Mohammadi, “A compact novel CPW-fed antenna with square spiral-patch for multiband applications,” Microw. Opt. Technol. Lett., 57: 111-115. 2015.

[12] X. Wang, Y. Yu, W. Che, “A novel dual-band printed monopole antenna based on planar inverted-cone antenna (PICA),” IEEE Antennas Wireless Propag. Lett., 13: 217-220, 2014.

[13] Z. Tang, K. Liu, Y. Yin, R. Lian, “Design of a compact triband monopole antenna for WLAN and WiMAX applications,” Microw. Opt. Technol. Lett., 57: 2298-2303. 2015.

[14] H. Huang, Y. Liu, S. Zhang, S. Gong, “Multiband metamaterial-loaded monopole antenna for WLAN/WiMAX applications,” IEEE Antennas Wireless Propag. Lett., 14: 662-665, 2015.

[15] R. Pandeeswari, S. Raghavan, “A CPW-fed triple band OCSRR embedded monopole antenna with modified ground for WLAN and WIMAX applications,” Microw. Opt. Technol. Lett., 57: 2413-2418. 2015.

[16] H. Liu, P. Wen, S. Zhu, B. Ren, X. Guan, H. Yu, “Quad-Band CPW-Fed monopole antenna based on flexible pentangle-loop radiator,” IEEE Antennas Wireless Propag. Lett., 14: 1373-1376, 2015.

[17] X. Fang, G. Wen, D. Inserra, Y. Huang, J. Li, “Compact wideband CPW-Fed meandered-slot antenna with slotted Y-Shaped central element for Wi-Fi, WiMAX, and 5G applications,” IEEE Trans. Antennas Propag., 66(12): 7395-7399, 2018.

[18] R. Wu, P. Wang, Q. Zheng, R. Li, “Compact CPW-fed triple-band antenna for diversity applications,” Electron. Lett., 51(10): 735-736, 2015.

[19] H. Li, G. Wang, X. Gao, L. Zhu, “CPW-Fed multiband monopole antenna loaded with DCRLH-TL unit cell,” IEEE Antennas Wireless Propag. Lett., 14: 1243-1246, 2015.

[20] M. Wu, M. Chuang, “Multibroadband slotted bow-tie monopole antenna,” IEEE Antennas Wireless Propag. Lett., 14: 887-890, 2015.

[21] S. Ahmad et al., "A metasurface-based single-layered compact AMC-Backed dual-band antenna for off-body IoT devices," IEEE Access, 9: 159598-159615, 2021.

[22] D. Chen, C. Zhang, H. Zhang, C. Zhao, "A novel printed monopole antenna with a tapered-line resonator loading," IEEE Access, 2025.

[23] M. Borhani, P. Rezaei, A. Valizade, “Design of a reconfigurable miniaturized microstrip antenna for switchable multiband systems,” IEEE Antennas Wireless Propag. Lett., 15: 822-825, 2015.

[24] E. Nasrabadi, P. Rezaei, “A novel design of reconfigurable monopole antenna with switchable triple band-rejection for UWB applications,” Int. J. Microwave Wireless Tech., 8: 1223-1229, 2015.

[25] A. Franc, E. Pistono, D. Gloria, P. Ferrari, "High-performance shielded coplanar waveguides for the design of CMOS 60-GHz bandpass filters," IEEE Trans. Electron Devices, 59(5): 1219-1226, 2012.

[26] J. H. Park, H. T. Kim, W. Choi, Y. Kwon, Y. K. Kim, "V-band reflection-type phase shifters using micromachined CPW coupler and RF switches," J. Microelectromech. Syst., 11(6): 808-814, 2002.

[27] S. G. Mao, Y. Z. Chueh, "Broadband composite right/left-handed coplanar waveguide power splitters with arbitrary phase responses and balun and antenna applications," IEEE Trans. Antennas Propag., 54(1): 243-250, 2006.

[28] D. Parveg, M. Varonen, D. Karaca, A. Vahdati, M. Kantanen, K. A. I. Halonen, "Design of a D-Band CMOS amplifier utilizing coupled slow-wave coplanar waveguides," IEEE Trans. Microwave Theory Tech., 66(3): 1359-1373, 2018.

[29] M. Nedil, T. A. Denidni, L. Talbi, "Novel butler matrix using CPW multilayer technology," IEEE Trans. Microwave Theory Tech., 54(1): 499-507, 2006.

[30] D. M. Pozar, "Microwave Engineering," 4th Edition, John Wiley & Sons, 2011.

[31] R. N. Simons, "Coplanar Waveguide Circuits, Components, and Systems," John Wiley & Sons, 2004.

Biographies

Mojtaba Zahiry received the B.Sc. degree in Communications Engineering from Arak University (Arak-Iran) in 2015, M. Sc. Degree from Shahid Rajaee Teacher Training University (Tehran-Iran) in 2017. His research interests include antenna theory, antenna design and simulation, microstrip antenna, multi-band antenna and wide-band antenna.

- Email: mojtaba.zahiryy@gmail.com
- ORCID: 0000-0001-6574-7429
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Seyed Mohammad Hashemi was born in Tehran, Iran, in 1983. He received the B.Sc., M.Sc. and Ph.D. degrees in Electrical Engineering from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2006, 2008 and 2013 respectively. From 2012 to 2013 he joined Aalto University, Finland, as a Visiting Scholar. Since 2015, he joined Communications Engineering at the Department of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran, where he is now Associate Professor. His research interests include Applied Electromagnetics, Optimization Methods, Antenna and Microwave Engineering.

- Email: sm.hashemi@sru.ac.ir
- ORCID: 0000-0003-1484-9008
- Web of Science Researcher ID: MWY-3892-2025
- Scopus Author ID: 57192714769
- Homepage: https://www.sru.ac.ir/en/faculty/school-of-electrical-engineering/seyed-mohamad-hashemi/

Javad Ghalibafan received the B.S. degree from the Ferdowsi University of Mashhad in 2007 and the M.S. and Ph.D. degrees from Iran University of Science Technology in 2009 and 2013, respectively. In 2014 he joined Department of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran, where he is now Associate Professor, and the head of Antenna & Microwave Lab. His research interests include the analysis, design, and measurement of artificial electromagnetic materials; antenna and microwave devices; metamaterial; and magnetic material.

- Email: jghalibafan@shahroodut.ac.ir
- ORCID: 0000-0001-7113-0951
- Web of Science Researcher ID: EVH-9176-2022
- Scopus Author ID: 35931876800
- Homepage: https://shahroodut.ac.ir/en/as/index.php?id=S455

How to cite this paper:
M. Zahiry, S. M. Hashemi, J. Ghalibafan, "Designing multiband, reconfigurable printed antenna for modern communication systems," J. Electr. Comput. Eng. Innovations, 13(2): 485-493, 2025.

DOI: 10.22061/jecei.2025.9666.653

URL: https://jecei.sru.ac.ir/article_2318.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

PAPER TYPE? (Research paper, short paper, Review paper *et al.*)

Instructions and Formatting Rules for Authors of Journal of Electrical and Computer Engineering Innovations, JECEI

F. Author, S. Author, T. Author

Affiliations of the Authors: (Department, Faculty, University (Institution), City, Country)

Article Info

Article History:
Received
Reviewed
Revised
Accepted

Keywords:
The author(s) shall provide up to 6 keywords to help identify the major topics of the paper

*Corresponding Author's Email Address:

Abstract

Background and Objectives: This section should be the shortest part of the abstract and should very briefly outline the following information: 1-What is already known about the subject, related to the paper in question. 2- What is not known about the subject and hence what the study intended to examine (or what the paper seeks to present). In most cases, the background can be framed in just 2–3 sentences, with each sentence describing a different aspect of the information referred to above; sometimes, even a single sentence may suffice. The purpose of the background, as the word itself indicates, is to provide the reader with a background to the study, and hence to smoothly lead into a description of the methods employed in the investigation.
Methods: The methods section is usually the second-longest section in the abstract. It should contain enough information to enable the reader to understand what was done, and how.
Results: The results section is the most important part of the abstract and nothing should compromise its range and quality. This is because readers who peruse an abstract do so to learn about the findings of the study. The results section should therefore be the longest part of the abstract and should contain as much detail about the findings as the journal word count permits.
Conclusion: This section should contain the most important take-home message of the study, expressed in a few precisely worded sentences. Usually, the finding highlighted here relates to the primary outcome measure; however, other important or unexpected findings should also be mentioned. It is also customary, but not essential, for the authors to express an opinion about the theoretical or practical implications of the findings, or the importance of their findings for the field. Thus, the conclusions may contain three elements: 1- The primary take-home message 2-The additional findings of importance 3-The perspective.

Introduction

This document provides an example of the desired layout for JECEI paper and can be used as a template for Microsoft Word versions 2003 and later. It contains information regarding desktop publishing format, type sizes, and typefaces. Style rules are provided to explain how to handle equations, units, figures, tables, abbreviations, and acronyms. Sections are also devoted to the preparation of appendixes, acknowledgments, references, and authors' biographies. For additional information including electronic file requirements for text and graphics, please refer to www.autjournal.com.

Technical Work Preparation

Please use automatic hyphenation and check your spelling. Additionally, be sure your sentences are complete and that there is continuity within your paragraphs. Check the numbering of your graphics and make sure that all appropriate references are included.

Doi:

A. Template

This document may be used as a template for preparing your technical work.

B. Format

If you choose not to use this document as a template, prepare your technical work in single-spaced, double-column format, on paper 21.6×27.9 centimeters (8.5×11 inches or 51×66 picas). Set top and bottom margins to 25 millimeters (0.98 inch) and left and right margins to about 20 millimeters (0.79 inch). Do not violate margins (i.e., text, tables, figures, and equations may not extend into the margins). The column width is 82 millimeters (3.2 inches). The space between the two columns is 6 millimeters (0.24 inch). Paragraph indentation is 4.2 millimeters (0.17 inch). Use full justification. Use either one or two spaces between sections, and between text and tables or figures, to adjust the column length.

C. Typefaces and Sizes

Please use a proportional serif typeface such as Calibri and embed all fonts. Table 1 provides samples of the appropriate type sizes and styles to use.

D. Section Headings

A primary section heading is enumerated by a Roman numeral followed by a period and is centered above the text. A primary heading should be in capital letters.

A secondary section heading is enumerated by a capital letter followed by a period and is flush left above the section. The first letter of each important word is capitalized and the heading is italicized.

A tertiary section heading is enumerated by an Arabic numeral followed by a parenthesis. It is indented and is followed by a colon. The first letter of each important word is capitalized and the heading is italicized.

A quaternary section heading is rarely necessary, but is perfectly acceptable if required. It is enumerated by a lowercase letter followed by a parenthesis. It is indented and is followed by a colon. Only the first letter of the heading is capitalized and the heading is italicized.

E. Figures and Tables

Figure axis labels are often a source of confusion. Try to use words rather than symbols. As an example, write the quantity "Magnetization," or "Magnetization, *M*," not just "*M*." Put units in parentheses. Do not label axes only with units. As in Fig. 1, write "Magnetization (kA/m)" or "Magnetization (kA·m^{-1})," not just "kA/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K." Figure labels should be legible, approximately 8- to 10-point type.

Large figures and tables may span both columns, but may not extend into the page margins. Arrange these one column figures and tables at either top or end of a page, or at the end of the paper right before the references. Figure captions should be below the figures; table captions should be above the tables. Do not put captions in "text boxes" linked to the figures. Do not put borders around your figures. Use Insert | Reference | Caption to number your tables and figures, and use Insert | Reference | Cross- reference to refer to their numbers.

Table 1: Samples of Calibri sizes and styles used for formatting a pes technical work

Point Size	Purpose in Paper	Special Appearance
9	Table text, figure text footnotes, subscripts, superscripts, references, bio, Figure caption, keywords	Table Title
10	Body text, equations, author affiliation, abstract	*Subheadings*
11		**Section Titles**
12	***Author Name***	

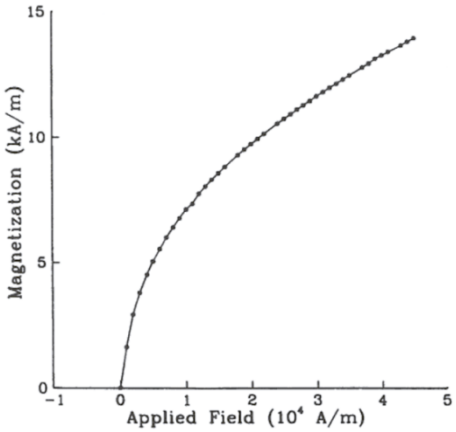


Fig. 1: Magnetization as a function of applied field. (Note that there is a colon after the figure number followed by two spaces.)

All figures and tables must appear near, but not before, their first mention in the text. Use the abbreviation "Fig. 1," even at the beginning of a sentence.

To insert images in Word, use Insert | Picture | From File.

F. Numbering

Number reference citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [2], [3] are each numbered with separate brackets [1][1]-[2]. Refer simply to the reference number, as in [2]. Do not use "Ref. [2]" or "reference [2]" except at the beginning of a sentence: "Reference [2] shows…."

Number footnotes separately with superscripts

(Insert | Footnote). Place the actual footnote at the bottom of the column in which it is cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Use Arabic numerals for figures and Roman numerals for tables. Appendix figures and tables should be numbered consecutively with the figures and tables appearing in the rest of the paper. They should not have their own numbering system.

G. Units

Metric units are preferred for use in IEEE publications in light of their global readership and the inherent convenience of these units in many fields. In particular, the use of the International System of Units is advocated. This system includes a subsystem of units based on the meter, kilogram, second, and ampere (MKSA). British units may be used as secondary units (in parentheses). An exception is when British units are used as identifiers in trade, such as 3.5-inch disk drive.

H. Math and Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the "Equation" markup style. Write the equation number in parentheses using Insert | Caption.

Use the Microsoft Equation Editor for all math objects in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

To make your equations more compact, you may use the slash (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Use parentheses to avoid ambiguities in denominators. Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit Tesla).

Use Insert | Reference | Caption to number equations. Refer to "(1)," not "Eq. 1" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ...". Punctuate equations when they are part of a sentence, as in

$$\int_0^{r_2} F(r,\varphi)\, dr\, d\varphi = [\sigma r_2 / (2\mu_0)] \cdot \int_0^{\infty} \exp(-\lambda\,|\,z_j - z_i\,|)\,\lambda^{-1}\, J_1(\lambda r_2)\, J_0(\lambda r_i)\, d\lambda \quad (1)$$

Use two column tables to locate equations and their numbers properly in one line, as follows:

$$I_F = I_B = -I_C = A^2 I_{A1} + A I_{A2} + I_{A0} = \frac{-J\sqrt{3}E_A}{Z_1 + Z_2} \quad (2)$$

where I_F is the fault current. Be sure that the border is off.

Results and Discussion

The Results section should briefly present the experimental data in text, tables or figures. Tables and figures should not be described extensively in the text.

The Discussion should focus on the interpretation and the significance of the findings with concise objective comments that describe their relation to other work in the area. It should not repeat information in the results. The final paragraph should highlight the main conclusion(s), and provide some indication of the direction future research should take.

Conclusion

As the Conclusion section is the most important element of a manuscript, so it must be more expanded scientifically and contently at least half a page length.

Example:

In this study, a forecast model was developed to determine the generation of MSW in the municipalities of the CCS, Chiapas State, Mexico. A MLR was used to obtain the forecast model with social and demographic explanatory variables. Two forecast models were presented and analyzed, with variables that met the multicollinearity test. The most important variables to predict the rate of MSW generation in the study area were the population of each municipality (XPop), the population born in another municipality (XPbam) and the population density (XPd). XPop is the most influential explanatory variable of waste generation, particularly it is related in a positive way. XPbam is less related to waste generation. XPd is the variable that least influences waste generation prediction; in addition, it can present problems of correlation with other explanatory variables. Although other variables, such as daily per capita income (XDpi) and average schooling (XAs), are very important, they do not seem to have an effect on the response variable in this study. The user of this forecast model should use model 2, since it is the one with the highest parsimony (it uses fewer variables); R^2_{adj}, MAPE, MAD and RMSE values indicated high influence on the explained phenomenon and high forecasting capacity. Additionally, it is important to mention that when using the models proposed for forecasting purposes, it is necessary to make a transformation in the explanatory and response variables (use inverse of natural logarithm). The inferences made on the municipalities of the study area showed that, except in some municipalities, the MSW generation rate usually presented a gradual increase

with respect to population growth and with respect to the number of inhabitants that were born in another entity (migration). Finally, this study can be a solid basis for comparison for future research in the area of study. It is possible to use different mathematical models such as artificial neural network, principal component analysis, time-series analysis, etc., and compare the response variable or the predictors.

Author Contributions

Each author role in the research participation must be mentioned clearly.

Example:

A. Mahboobi, B. Bagheri, and C. Ahmdi designed the experiments. A. Mahboobi collected the data. A. Mahboobi carried out the data analysis. A. Mahboobi, B. Bagheri, and C. Ahmdi interpreted the results and wrote the manuscript.

Acknowledgment

The following is an example of an acknowledgment. (Please note that financial support should be acknowledged in the unnumbered footnote on the title page.)

The author gratefully acknowledges the IEEE I. X. Austan, A. H. Burgmeyer, C. J. Essel, and S. H. Gold for their work on the original version of this document.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

Define less common abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title unless they are unavoidable.

Example:

MS	Multispectral
SMF	Spectral Matched Filter
SAM	Spectral Angle Mapper
MSD	Matched Subspace Detector
OSP	Orthogonal Subspace Projection
CEM	Constrained Energy Minimization
ASD	Adaptive Subspace Detector
STD	Sparsity Based Target Detector
KSAM	Kernel Based SAM
DTD	Difference Based Target Detection
AP-CR	Attribute Profile Based Collaborative Representation
ROC	Receiver Operating Characteristic

MS	Multispectral
SMF	Spectral Matched Filter
SAM	Spectral Angle Mapper
MSD	Matched Subspace Detector
OSP	Orthogonal Subspace Projection
CEM	Constrained Energy Minimization
ASD	Adaptive Subspace Detector
STD	Sparsity Based Target Detector
KSAM	Kernel Based SAM
DTD	Difference Based Target Detection
AP-CR	Attribute Profile Based Collaborative Representation
ROC	Receiver Operating Characteristic

References

References are important to the reader; therefore, each citation must be complete and correct. There is no editorial check on references; therefore, an incomplete or wrong reference will be published unless caught by a reviewer or discusser and will detract from the authority and value of the paper. References should be readily available publications. List only one reference per reference number. If a reference is available from two sources, each should be listed as a separate reference. Give all authors' names; do not use *et al.*

Samples of the correct formats for various types of references are given below.

Periodicals:

[1] J. F. Fuller, E. F. Fuchs, K. J. Roesler, "Influence of harmonics on power distribution system protection," IEEE Trans. Power Deliv., 3(2): 549-557, 1988.

Books:

[2] E. Clarke, Circuit Analysis of AC Power Systems, vol. I. New York: Wiley: 81, 1950.

Technical Reports:

[3] E. E. Reber, R. L. Mitchell, C. J. Carter, "Oxygen absorption in the Earth's atmosphere," Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1968.

[4] S. L. Talleen. (1996, Apr.). The Intranet Architecture: Managing information in the new paradigm. Amdahl Corp., Sunnyvale, CA.

Papers Presented at Conferences (Unpublished):

[5] D. Ebehard, E. Voges, "Digital single sideband detection for interferometric sensors," presented at the 2nd Int. Conf. Optical Fiber Sensors, Stuttgart, Germany, 1984.

[6] Process Corp., Framingham, MA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annu. Meeting.

Papers from Conference Proceedings (Published):

[7] J. L. Alqueres, J. C. Praca, "The Brazilian power system and the challenge of the Amazon transmission," in Proc. IEEE Power Engineering Society Transmission and Distribution: 315-320, 1991.

Dissertations:

[8] S. Hwang, "Frequency domain system identification of helicopter rotor dynamics incorporating models with time periodic coefficients," Ph.D. dissertation, Dept. Aerosp. Eng., Univ. Maryland, College Park, 1997.

Standards:

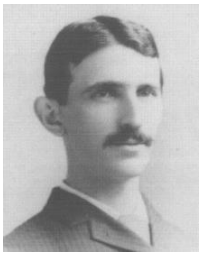
[9] IEEE Guide for Application of Power Apparatus Bushings, IEEE Standard C57.19.100-1995, Aug. 1995.

Patents:

[10] G. Brandli and M. Dick, "Alternating current fed power supply," U.S. Patent 4 084 217, Nov. 4, 1978.

Biographies

A technical biography for each author may be included, but without any title, as it is seen herein. It should begin with the author's name (as it appears in the byline). A photograph and an electronic file of the photo should also be included for each author. The photo should be black and white, glossy, and 3.0 centimeters (1.18 inches) wide by 3.8 centimeters (1.5 inches) high. The head and shoulders should be centered, and the photo should be flush with the left margin. The following is an example of the text of a technical biography:



Nikola Tesla (M'1888, F'17) was born in Smiljan in the Austro-Hungarian Empire, on July 9, 1856. He graduated from the Austrian Polytechnic School, Graz, and studied at the University of Prague. His employment experience included the American Telephone Company, Budapest, the Edison Machine Works, Westinghouse Electric Company, and Nikola Tesla Laboratories. His special fields of interest included high frequency. Tesla received honorary degrees from institutions of higher learning including Columbia University, Yale University, University of Belgrade, and the University of Zagreb. He received the Elliott Cresson Medal of the Franklin Institute and the Edison Medal of the IEEE. In 1956, the term "tesla" (T) was adopted as the unit of magnetic flux density in the MKSA system. In 1975, the Power Engineering Society established the Nikola Tesla Award in his honor. Tesla died on January 7, 1943.

- Email:
- ORCID:
- Web of Science Researcher ID:
- Scopus Author ID
- Homepage:

How to cite this paper:
F. Author, S. Author, T. Author," Instructions and Formatting Rules for Authors of Journal of Electrical and Computer Engineering Innovations, JECEI," J. Electr. Comput. Eng. Innovations, ×(×): ×××-×××, ××××.

DOI:

URL: http://jecei.srttu.edu/journal/authors.note