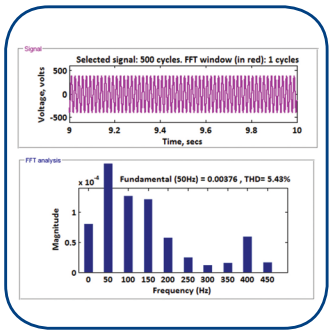
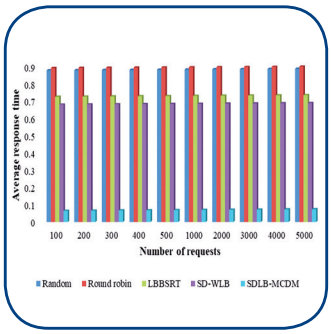
Volume 14, Number 4, 2023 ISSN 1847-6996

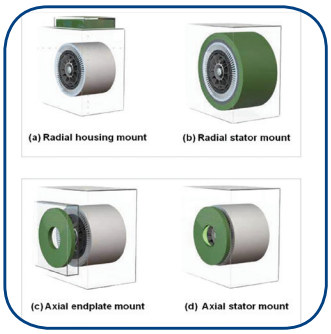
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

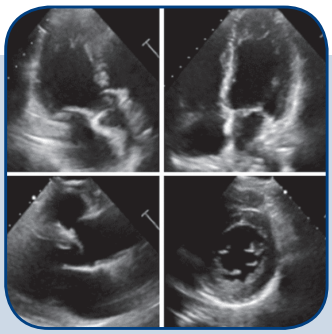
IJECES International Journal of Electrical and Computer Engineering Systems

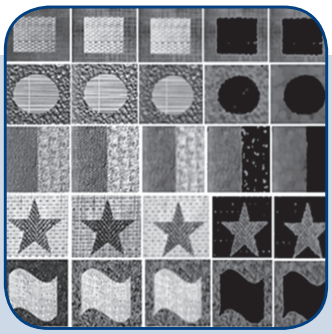
International Journal of Electrical and Computer Engineering Systems

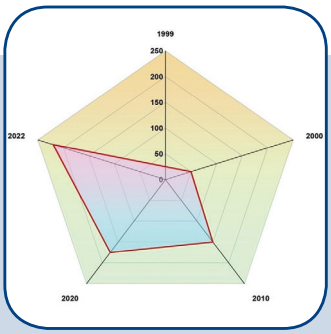












ISSN: 1847-6996

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

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

Osijek, Croatia | Volume 14, Number 4, 2023 | Pages 371 - 495

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

CONTACT

International Journal of Electrical and Computer Engineering Systems (IJECES)

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

Subscription Information

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

Journal is referred in:

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

Bibliographic Information

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

IJECES online

https://ijeces.ferit.hr

TABLE OF CONTENTS

Multi-Resolution Feature Embedded Level Set Model for Crosshatched Texture Segmentation

Original Scientific Paper

Prabhakar K.
Ballari Institute of Technology and Management,
Department of Electronics and Communication Engineering,
Near Allipur, Ballari-583104, India
prabhakar@bitm.edu.in

Sadyojatha K.M.
Ballari Institute of Technology and Management,
Department of Electronics and Communication Engineering,
Near Allipur, Ballari-583104, India
sadyojatha@bitm.edu.in

Abstract *– In image processing applications, texture is the most important element utilized by human visual systems for distinguishing dissimilar objects in a scene. In this research article, a variational model based on the level set is implemented for crosshatched texture segmentation. In this study, the proposed model's performance is validated on the Brodatz texture dataset. The cross-hatched texture segmentation in the lower resolution texture images is difficult, due to the computational and memory requirements. The aforementioned issue has been resolved by implementing a variational model based on the level set that enables efficient segmentation in both low and high-resolution images with automatic selection of the filter size. In the proposed model, the multi-resolution feature obtained from the frequency domain filters enhances the dissimilarity between the regions of crosshatched textures that have low-intensity variations. Then, the resultant images are integrated with a level set-based active contour model that addresses the segmentation of crosshatched texture images. The noise added during the segmentation process is eliminated by morphological processing. The experiments conducted on the Brodatz texture dataset demonstrated the effectiveness of the proposed model, and the obtained results are validated in terms of Intersection over the Union (IoU) index, accuracy, precision, f1-score and recall. The extensive experimental investigation shows that the proposed model effectively segments the region of interest in close correspondence with the original image. The proposed segmentation model with a multi-support vector machine has achieved a classification accuracy of 99.82%, which is superior to the comparative model (modified convolutional neural network with whale optimization algorithm). The proposed model almost showed a 0.11% improvement in classification accuracy related to the existing model*

Keywords*: crosshatched texture, level set, morphological processing, multiresolution, texture segmentation*

1. INTRODUCTION

In image processing applications, texture is a fundamental property of object surfaces and is extensively present in natural images [1-2]. It has an extensive range of applications like texture classification [3-4], spectral shape retrieval video recognition [5], and retrieval [6-7]. Among these available topics, texture classification is an active research area, where it has gained more attention among the researcher's communities in the field of pattern recognition [8] and computer vision [9]. The typical applications of texture classification comprise object recognition, content-based image retrieval, fabric inspection, remote sensing, and medical image analysis [10-12]. On the other hand, the purpose of texture segmentation is to discriminate between the regions, which have dissimilar textures [13]. The texture segmentation is done by using two major methods such as filter bank methods and statistical-based methods [14-15]. The existing methods are implemented by applying the bank of filters to the image, and the filter response is studied to set the image's local behavior. The existing methods' performance completely depends on the texture, which has distinctive statistical properties [16-17].

This research article aims to segment the crosshatched textures and to automate the filter size in order to enhance the dissimilarity between regions with low-intensity variations. Hence, the filter size must be selected cautiously, because when the filter size is large, it causes uncertainties across the boundary re-

gions and if small, then perhaps, it fails to confine certain variations in a few textures. The accurate descriptor and exact filter size are the necessary attributes for good texture segmentation. The main contributions are listed as follows:

- Used histogram equalization technique that improves the contrast of the collected images, which are acquired from the Brodatz texture dataset.
- Developed a precise model for the automatic selection of the filter size. Further, the texture segmentation for crosshatched texture images is done in an unsupervised way. The multi-resolution feature embedded level set model is proposed for texture segmentation.
- The experimental results are validated by using different evaluation measures known as the IoU index, precision, f1-score, recall and accuracy. The segmentation results of the proposed model are compared with the Mix-Normalized-Cut (MixNCut) model, and the classification results are compared with the modified Convolutional Neural Network (CNN) with Whale Optimization Algorithm (WOA). The proposed segmentation model obtains near-perfect accuracy on all crosshatched texture images acquired from Brodatz texture dataset. The proposed model effectively sorts image data into interpretable information and it is utilized in an extensive range of applications like remote sensing, and medical imaging. Industrial application, image retrieval, etc.

This research article is organized as follows: research papers related to texture image segmentation and classification are surveyed in Section 2. The mathematical and theoretical explanation of the proposed model are specified in Section 3. The experimental outcomes of the proposed model is stated in Section 4, and the conclusion of this work is mentioned in Section 5.

2. LITERATURE SURVEY

Maskey and Newman [18] developed a novel texture directionality measure, wherein both global and local directionality aspects were considered. In this literature study, the developed texture directionality measure was employed in different applications that included a circuit board image classification task, striped shirt classification, striped fabric classification and a CNN initialization task. The extensive experimental analysis stated that the suggested measure was superior than the then-existing measures. However, a higher-end graphics processing unit system was required to perform the classification tasks, which proved to be computationally expensive. Ranganath et al. [19] implemented a new image texture classification model named pixel range calculation. As per the derived results, the developed model not only provided superior classification results but also consumed limited computational time related to the conventional models. However, the developed model was ineffective in multi-class texture classification on larger datasets. Dixit et al. [20] integrated a modified CNN model with WOA for effective texture classification. The inclusion of WOA made CNN more effective and robust in texture classification by the selection of optimal parameters. As specified earlier, the implementation of the deep learning models was computationally expensive compared to other machine learning models.

Hilal et al. [21] implemented a new bi-dimensional entropy-based measure for texture classification that included multi-channel methods (FuzEnM2D and FuzEnV2D), and single channel method (FuzEnC2D). The extensive experimental results demonstrated that the developed measure outperformed the well-known texture analysis measures, but the computational time was higher, which needed to be reduced as part of a future extension. Raja et al. [22] developed a new descriptor called Optimized Local Ternary Pattern (OLTP) for effective texture classification. In this literature study, the developed descriptor's effectiveness was validated on two standard datasets namely, Usptex and Brodatz. The obtained simulation outcomes demonstrated that the use of OLTP descriptors effectively improved the texture classification accuracy. However, the developed OLTP descriptor applied only to the gray-scale images, which was considered a major concern in this study. Soares et al. [23] introduced a new class-independent method for segmenting the texture regions from the images. However, the developed method was restricted to a limited number of dissimilar texture classes in the images.

Khan et al. [24] implemented a new descriptor, Overlapped Multi-oriented Triscale Local Binary Patterns (OMTLBP) that effectively retains image classification accuracy under different conditions like illumination, scale, and orientation. In addition to this, Pan et al. [25] developed a new descriptor: Scale-Adaptive LBP (SALBP) for effective texture classification. The effectiveness of the developed OMTLBP and SALBP descriptors was validated on different online datasets like Outex, Brodatz, etc. The obtained experimental outcomes demonstrated the superiority of the developed OMTLBP and SALBP descriptors against traditional texture descriptors using classification accuracy. However, the developed OMTLBP and SALBP descriptors included the concern of high computational time. Feng et al. [26] implemented a new objective measure based on the smallest univalue segment assimilating nucleus method for multi-focus image fusion. By inspecting experimental results, the high computational time was a major issue in this literature. To address the abovementioned problems, a new multi-resolution feature embedded level set model is implemented for crosshatched texture segmentation.

3. METHODOLOGY

In recent decades, the possibility of inaccurate segmentation is high, when the parameters are selected at the stage of feature extraction or the stage of segmen-

tation voluntarily. Additionally, the supervised texture segmentation schemes require a priori knowledge, and it may not be viable to provide information regarding the number of texture regions and the type of textures to be segmented. Particularly, if the volume of images to be handled is large viz thousands of images in art galleries, large numbers of satellite imagery, etc. [27-30]. The supervised texture segmentation techniques are difficult to provide the prior knowledge manually, and this has motivated the present researchers to look for fully automated schemes. In this research article, an automated model named Multi-Resolution Feature Embedded Level Set Model is implemented for effective crosshatched texture segmentation.

3.1. DATASET DESCRIPTION

In this research article, the proposed multi-resolution feature embedded level set model's performance is evaluated on the Brodatz texture dataset. In text classification, the Brodatz texture dataset is one of the popular datasets, which is recorded from the University of Southern California. The original dataset has rotated images, which are generated by using simple computer-graphics methods. The statistical contribution of the Brodatz texture dataset is denoted in Table 1. The sample images from Brodatz texture dataset are indicated in Fig. 1.

Table 1. Statistical contribution of the Brodatz texture dataset

Features	Values
Size	1.02 GB
Image format	8-bit gray-scale images
Texture patch size	640×640 pixels
Total number of samples	4480
Number of classes with unique samples	40
Number of classes	112

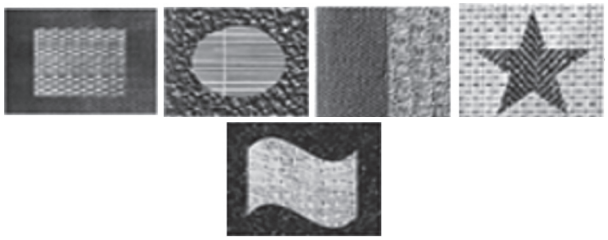


Fig. 1. Sample images of the Brodatz texture dataset

3.2. PRE-PROCESSING

After acquiring the images from Brodatz texture dataset, the image pre-processing is carried out by using a histogram equalization technique that adjusts the crosshatch image intensities for improving the contrast of the images. Let f be considered as the crosshatch image, which is denoted by a matrix integer pixel intensities m, that ranges between 0 to L-1. The histogram equalized crosshatch image g is mathematically determined in equation (1).

$$g_{i,j} = floor\,(L-1)\sum_{n=0}^{f(i,j)} p_n \tag{1}$$

Where, p represents the normalized histogram value of crosshatch image f with a bin-possible intensity, L indicates the intensity value of range 256, and the term floor() rounds off the nearest integers, which are equivalent in transforming the pixel intensity value k, and it is mathematically denoted in equation (2).

$$T(k) = floor(L-1)\sum_{n=0}^{k} p_n \tag{2}$$

The pixel intensities f and g are considered as the continuous random values of X=Y on [0, L-1] in the transformation section, which is mathematically defined in equation (3).

$$Y = T(X) = (L-1)\int_0^x p_X\,(x)dx \tag{3}$$

Where T indicates the cumulative distributive function of X multiplied by L-1 and p_X represents the probability density function of crosshatch image f. In addition to this, Y is represented by $T(X)$, which is uniformly distributed on [0 ,L-1] namely $p_Y\,(y)$=1/(L-1), and it is mathematically defined in equations (4), (5), and (6).

$$\int_0^y p_Y(z)dz = \frac{1}{L-1} = \int_0^{T^{-1}(y)} p_X(w)dw \tag{4}$$

$$\frac{d}{dy}\left(\int_0^y p_Y(z)dz\right) = p_Y(y) = p_X(T^{-1}(y))\frac{d}{dy}(T^{-1}(y)) \tag{5}$$

$$\begin{aligned} &\frac{dT}{dx_{|x=T^{-1}(y)}}\frac{d}{dy}\left(T^{-1}(y)\right) \\ &= (L-1)p_X\left(T^{-1}(y)\right)\frac{d}{dy}\left(T^{-1}(y)\right) = 1 \end{aligned} \tag{6}$$

where, $p_Y\,(y)$=1/(L-1).

3.3. AUTOMATIC COMPUTATION OF THE FILTER SIZE

In image processing applications, the textures are complicated visual patterns poised on sub-patterns, which show signs of orientation, color, slope, size, etc. For significant texture segmentation, the image features and the selection of the filter size play an important role. Hence, the filter size has to be chosen carefully, since it is capable of capturing the pattern fully, and yields identical values when repeated over the entire homogeneous region.

In this article, an algorithm is developed for the automatic computation of the filter size, and the steps involved are listed below:

- Divide the test crosshatch image into n number of partitions of equal size (size is normally chosen as 21×21, which is usually found to be sufficiently large for the study made on all the textures in Brodatz dataset) and compute the mean of each partition.
- Every partition starts with a smaller filter size of 3×3. Increase the size of the filter until the mean of the filter equals the mean of the partition.

- Group the obtained filter sizes into two clusters.
- Compute mean value of the lower cluster and use it as the filter size.

The method used here involves the computation of a minimum filter size. It is a simple technique in which the filter size is progressively increased until the average within the filter, and average of the partition, are both similar. It ensures that for every region, a maximum filter size is estimated which covers the maximum homogeneous area within that region. The minimum average size among all such regions is selected as the filter size. The algorithm presented above computes the size of the filter involuntarily and is then segmented using a level set. The flowchart of the automated filter selection is indicated in Fig. 2.

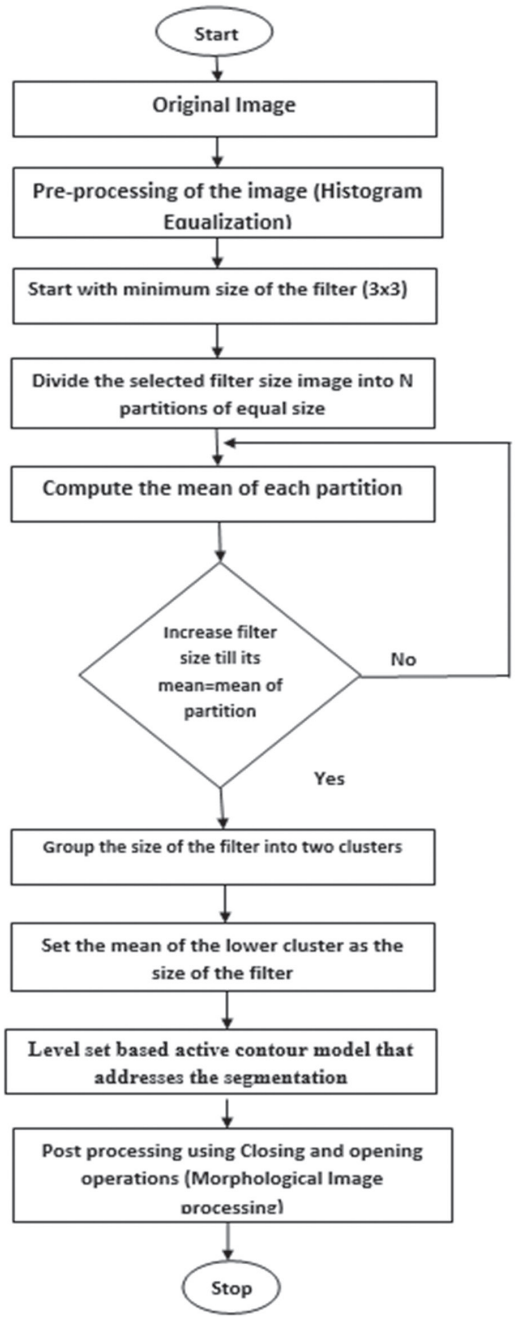


Fig. 2. Flowchart of the automated filter selection

In this article, an algorithm is developed for the automatic computation of the filter size, and the steps involved are listed below:

- Divide the test crosshatch image into n number of partitions of equal size (size is normally chosen as 21×21, which is usually found to be sufficiently large for the study made on all the textures in Brodatz dataset) and compute the mean of each partition.
- Every partition starts with a smaller filter size of 3×3. Increase the size of the filter until the mean of the filter equals the mean of the partition.
- Group the obtained filter sizes into two clusters.
- Compute mean value of the lower cluster and use it as the filter size.

The method used here involves the computation of a minimum filter size. It is a simple technique in which the filter size is progressively increased until the average within the filter, and average of the partition, are both similar. It ensures that for every region, a maximum filter size is estimated which covers the maximum homogeneous area within that region. The minimum average size among all such regions is selected as the filter size. The algorithm presented above computes the size of the filter involuntarily and is then segmented using a level set. The flowchart of the automated filter selection is indicated in Fig. 2.

3.4. MULTI-RESOLUTION FEATURE EMBEDDED LEVEL SET MODEL

In the proposed model, the preprocessed crosshatch image is constructed with two different crosshatched textures that have subjective boundaries. After preprocessing, the next step is selecting the correct size of the frequency domain filter, which is automated. The results obtained are integrated with a level set based active contour model that addresses the segmentation of crosshatched texture images. Any noise incurred during histogram equalization is eliminated by a post-processing step, using morphological processing. The automation process of the frequency domain filters is represented below. Initially it starts with the mean calculation of the filter, which is graphically shown in Fig. 3.

5	3	6
2	1	9
8	4	7

(a)

*	*	*
*	5	*
*	*	*

(b)

Fig. 3. (a) 3×3 Image, and (b) result of the mean filter applied to 3×3 Image

Fig. 3(a) shows a 3×3 image with a center pixel value of 1. Fig. 3(b) shows the mean filter with a center pixel value of 5. The mean filter is a simple sliding-window spatial filter that replaces the center value in a window

with the average (mean) of all the pixel values. An example of mean filtering of a single 3×3 window of values is shown in Fig. 3. The filter size is decided interactively by starting with a minimum possible size of 3×3, gradually increasing the size, and continuing the size of the filter until the mean of the pixels within the area is similar. It is important to select the correct filter size for proper segmentation.

The original image is divided into n number of partitions of equal size (the size of the image is normally chosen as 21×21, which is found to be sufficiently large on all the textures in Brodatz dataset), and computes the mean of each partition. Every partition starts with a smaller filter size of 3×3, which is placed at the center of the partition, and then increases the size of the filter until the mean of the filter equals the mean of the partition. The size of the filter obtained is to be grouped into two clusters. The mean value of the lower cluster is used as the filter size. Thus, the obtained automated filtered image is considered for segmentation using a level set.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this research article, the proposed multi-resolution feature embedded level set model is implemented utilizing a Matlab software environment. An extensive experimental analysis is performed on a computer with an Intel core i5-6200U computer processing unit, an 8 GB Random Access Memory, and a Linux operating system. The segmentation performance of the proposed multi-resolution feature embedded level set model is analyzed using an evaluation measure by name of IoU index. It is a statistical measure used for gauging the similarity and diversity of sample sets, where A and B indicate ground truth and segmented regions. It is mathematically stated in equation (7).

$$I(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \quad (7)$$

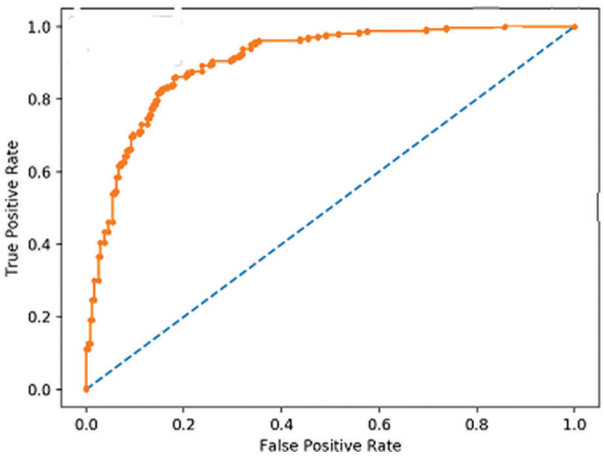


Fig 4. ROC curve of the proposed model

On the other hand, the classification performance of the proposed multi-resolution feature embedded level set model is validated using evaluation measures like accuracy, recall, precision and f1-score, which are stated in equations (8), (9), (10) and (11). True Positive is denoted as (TP), False Positive as (FP), True Negative as (TN), and False Negative as (FN). Meanwhile, the Receiver Operating Characteristic (ROC) curve of the proposed model is stated in Fig 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (10)$$

$$F1-score = \frac{2 \times Precision \times Recall}{Precision+Recall} \times 100 \quad (11)$$

4.2. QUANTITATIVE EVALUATION IN TERMS OF SEGMENTATION

The proposed multi-resolution feature embedded level set model is tested and the obtained results are verified in terms of the IoU index. The segmentation results of the proposed multi-resolution feature embedded level set model are mentioned in Table 2 and Fig. 5. The original images with two crosshatched textures are acquired from Brodatz texture dataset, and their subjective boundaries are represented in Fig. 5(a). The histogram results of the original image are shown in Fig. 5(b), and the automated filter size selection of the developed algorithm is shown in Fig. 5(c). The segmentation results of the level set algorithm are shown in Fig. 5(d). Any noise encountered during the histogram process is eliminated by the operation of opening and closing the morphological image processing, as shown in Fig. 5(e). The selection of the filter size is very vital to perform proper segmentation in the original texture image. The algorithm initiates with a 3×3 filter size and then, increases the filter size until the size of the filter equals the mean of the image partition. If the above condition is satisfied, the value of the lower cluster is used as the filter size.

The segmented results are validated by using the IoU index value, which is presented in Table 2, wherein the reference IoU value is one. The IoU index of the image under study is greater than 0.9, which indicates that the proximity with the initial image is excellent, as shown in Table 2. The experimental outcome indicates that the proposed model is capable of segmenting the region of interest in close correspondence with the texture image. The proposed multi-resolution feature embedded level set model is compared with MixNCut [26]. The proposed segmentation model achieves precise accuracy on all the crosshatched texture images. It is measured by means of "raw" pixels that identify optimum segmentation. The proposed segmentation model significantly outperforms the existing segmentation model - MixNcut [26]. The experimental results of the proposed and the existing segmentation models, in terms of running time and IoU index, are shown in Table 2.

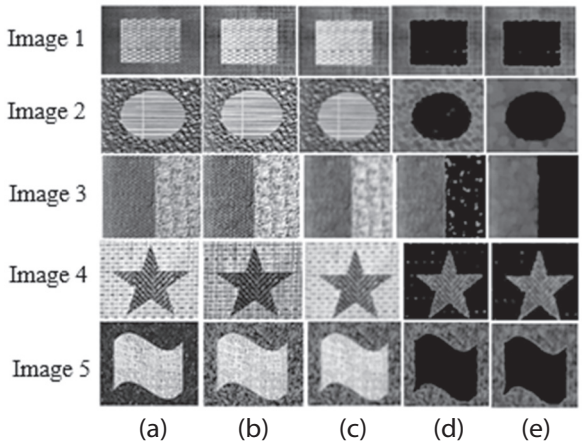


Fig. 5. Segmented output image (a) original image (Crosshatched) constructed from Brodatz texture dataset, b) histogram pre-processed image, (c) automated filter size selected by the proposed algorithm, (d) Segmented image through level set Algorithm, and (e) opening and closing results (Morphological Image Processing)

Table 2. Segmentation results in terms of IoU index value and running time

Index	Image 1		Image 2		Image 3		Image 4		Image 5	
Model	MixNcut [26]	Proposed	MixNcut [26]	Proposed	MixNcut [26]	Proposed	MixNcut [26]	Proposed	MixNcut [26]	Proposed
IoU	0.96	0.98	0.93	0.95	0.92	0.95	0.93	0.97	0.94	0.95
Time (sec)	9.09	6.30	6.61	5.42	7.15	5.35	8.05	5.72	7.89	5.81

4.2. QUANTITATIVE EVALUATION IN TERMS OF CLASSIFICATION

In the classification section, the proposed multi-resolution feature embedded level set model is tested with different classification techniques namely: random forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Multi-SVM (MSVM) by means of f1-score, accuracy, precision, and recall with different cross fold validations: 5 and 10 folds. By inspecting Tables 3 and 4, it is seen that the combination of multi-resolution feature embedded level set model with histogram equalization and MSVM has obtained higher classification performance in five-fold cross-validation with f1-score of 98.90%, accuracy of 99.82%, precision of 99.12%, and recall of 98.88% on the Brodatz texture dataset. The graphical depiction of the proposed multi-resolution feature embedded level set model with different classifiers and testing percentages is shown in Fig. 6 and 7.

As seen in the comparative analysis in Table 5, the proposed multi-resolution feature embedded level set model with MSVM achieved comparatively higher classification results, when related to a model named modified CNN with WOA. The proposed model gained an f1-score of 98.90%, accuracy of 99.82%, precision of 99.12%, and recall of 98.88% on the Brodatz texture dataset. However, the modified CNN with WOA has obtained an accuracy of 99.71%, precision of 96.70%, recall of 95.80%, and f1-score of 96.20% on the Brodatz texture dataset. As depicted in the literature survey section, the proposed model effectively resolves the problems of higher computational time, and achieves better segmentation and classification performance.

Table 3. Experimental results of the proposed model with five-fold cross-validation (80:20% training and testing)

Classifiers	Without histogram equalization				With histogram equalization			
	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
Random forest	90.34	91.76	93.23	89.28	93.24	95.30	96.66	92.20
KNN	92.95	92.56	92.45	93.33	95.90	98.24	96.32	97
SVM	93.73	92.34	92.37	92.28	97.74	99.22	98.32	97.22
MSVM	94.34	93.43	92.10	92.90	98.90	99.82	99.12	98.88

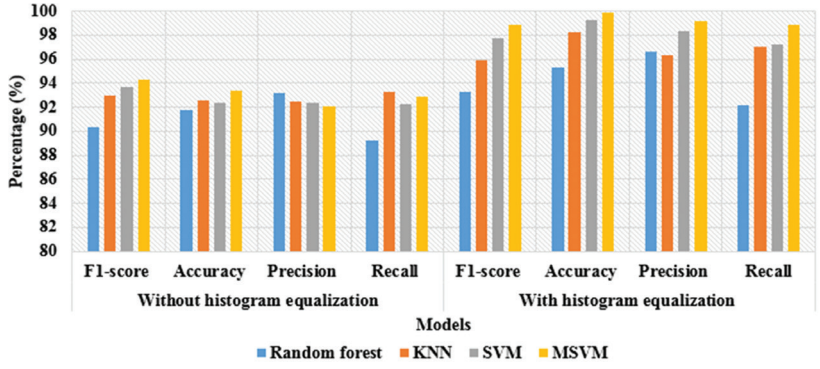


Fig. 6. Graphical depiction of the proposed model with five-fold cross-validation (80:20% training and testing)

Table 4. Experimental results of the proposed model with ten-fold cross-validation (90:10% training and testing)

Classifiers	Without histogram equalization				With histogram equalization			
	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
Random forest	88.12	90.43	91.13	84.24	92.20	94.38	93.62	90.28
KNN	90.44	90.55	91.15	90.34	92.90	94.28	93.34	94.07
SVM	92.66	91.55	90.36	87.99	93.79	95.20	95.35	94.25
MSVM	92.38	91.49	90.18	88.95	94.98	95.83	95.17	94.87

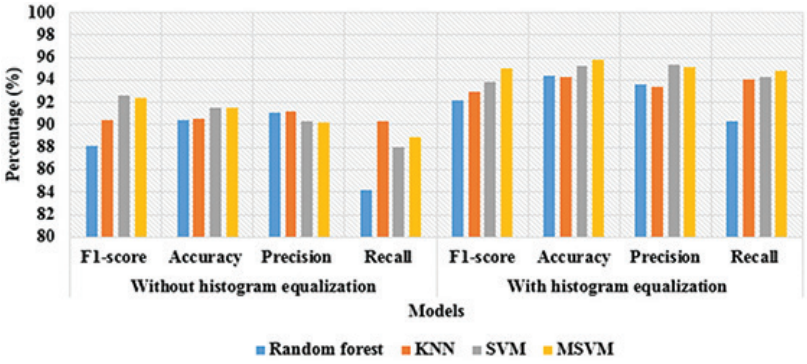


Fig. 7. Graphical depiction of the proposed model with ten-fold cross-validation (90:10% training and testing)

Table 5. Comparative results between the proposed and the existing models

Models	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
Modified CNN with WOA [20]	96.20	99.71	96.70	95.80
Multi-resolution feature embedded level set with MSVM	98.90	99.82	99.12	98.88

5. CONCLUSION

In image processing applications, the main pre-processing steps for object detection are image segmentation and shape detection. In this research article, an effective model is proposed for computing the appropriate features and automatic selection of filter size in context of unsupervised texture segmentation. The proposed model determines the minimum size of the filter for texture feature extraction in order to enhance discrimination and segmentation capabilities. In this study, a multi-resolution feature embedded level set model is introduced, that segments challenging images like cross-hatched texture images, which are acquired from the Brodatz texture dataset. The experimental results show that the proposed model provides longer-range interactions and captures the complex region appearances. The proposed segmentation model with MSVM classifier has achieved a classification accuracy of 99.82%, which is superior compared to other models. The proposed model is practical and robust, and it is employed for other dissimilar types of texture images in future work.

6. REFERENCES

[1] B. Pascal, N. Pustelnik, P. Abry, "Strongly convex optimization for joint fractal feature estimation and texture segmentation", Applied and Computational Harmonic Analysis, Vol. 54, 2021, pp. 303-322.

[2] S. Kinge, B. S. Rani, M. Sutaone, "Markov Random Field Based Color Texture Segmentation", International Journal of Recent Technology and Engineering, Vol. 8, No. 5, 2020, pp. 1055-1060.

[3] P.-P. Niu, L. Wang, X. Shen, Q. Wang, X.-Y. Wang, "Texture image segmentation using Vonn mixtures-based hidden Markov tree model and relative phase", Multimedia Tools and Applications, Vol. 79, No. 39-40, 2020, pp. 29799-29824.

[4] U. Barman, R. D. Choudhury, "Soil texture classification using multi class support vector machine", Information Processing in Agriculture, Vol. 7, No. 2, 2020, pp. 318-332.

[5] B. Wang, X. Yuan, X. Gao, X. Li, D. Tao, "A hybrid level set with semantic shape constraint for object segmentation", IEEE Transactions on Cybernetics, Vol. 49, No. 5, 2019, pp. 1558-1569.

[6] S. Aouat, I. Ait-hammi, I. Hamouchene, "A new approach for texture segmentation based on the Gray Level Co-occurrence Matrix", Multimedia Tools and Applications, Vol. 80, No. 16, 2021, pp. 24027-24052.

[7] B. Hui, Y. Liu, J. Qiu, L. Cao, L. Ji, Z. He, "Study of texture segmentation and classification for grading small hepatocellular carcinoma based on CT images", Tsinghua Science and Technology, Vol. 26, No. 2, 2021, pp. 199-207.

[8] T. Kavzoglu, H. Tonbul, "An experimental comparison of multi-resolution segmentation, SLIC and K-means clustering for object-based classification of VHR imagery", International Journal of Remote Sensing, Vol. 39, No. 18, 2018, pp. 6020-6036.

[9] D. Hamill, D. Buscombe, J. M. Wheaton, "Alluvial substrate mapping by automated texture segmentation of recreational-grade side scan sonar imagery", PloS one, Vol. 13, No. 3, 2018, p. e0194373.

[10] Y. Almakady, S. M. Mahmoodi, M. Bennett, "Gaussian Markov random fields-based features for volumetric texture segmentation", Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, San Jose, CA, USA, 28-30 March 2019, pp. 212-215.

[11] D. Chen, Q. Chen, F. Zhu, "Pixel-level texture segmentation based AV1 video compression", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12-17 May 2019, pp. 1622-1626.

[12] N. Borodinov, W.-Y. Tsai, V. V. Korolkov, N. Balke, S. V. Kalinin, O. S. Ovchinnikova, "Machine learning-based multidomain processing for texture-based image segmentation and analysis", Applied Physics Letters, Vol. 116, No. 4, 2020, p. 044103.

[13] B. Pascal, S. Vaiter, N. Pustelnik, P. Abry, "Automated data-driven selection of the hyperparameters for total-variation-based texture segmentation", Journal of Mathematical Imaging and Vision, Vol. 63, No. 7, 2021, pp. 923-952.

[14] S. Y. Idrissi, "An active contour model for texture image segmentation using Rényi divergence measure", Mathematical Modelling and Analysis, Vol. 27, No. 3, 2022, pp. 429-451.

[15] P. Pawlus, R. Reizer, M. Wieczorowski, "Functional importance of surface texture parameters", Materials, Vol. 14, No. 18, 2021, p. 5326.

[16] L. Davy, N. Pustelnik, P. Abry, "Combining dual-tree wavelet analysis and proximal optimization for anisotropic scalefree texture segmentation", https://hal.science/hal-03864647 (accessed: 2022).

[17] Y. Kang, "Retinex Algorithm and Mathematical Methods Based Texture Detail Enhancement Method for Panoramic Images", Mathematical Problems in Engineering, Vol. 2022, 2022, p. 6490393.

[18] M. Maskey, T. S. Newman, "On measuring and employing texture directionality for image classification", Pattern Analysis and Applications, Vol. 24, No. 4, 2021, pp. 1649-1665.

[19] A. Ranganath, M. R. Senapati, P. K. Sahu, "Classification of textures using pixel range calculation method", Proceedings of the International Conference on Computing, Communication, and Intelligent Systems, Greater Noida, India, 19-20 February 2021, pp. 652-657.

[20] U. Dixit, A. Mishra, A. Shukla, R. Tiwari, "Texture classification using convolutional neural network optimized with whale optimization algorithm", SN Applied Sciences, Vol. 1, No. 6, 2019, p. 655.

[21] M. Hilal, A. S. Gaudêncio, P. G. Vaz, J. Cardoso, A. Humeau-Heurtier, "Colored texture analysis fuzzy entropy methods with a dermoscopic application", Entropy, Vol. 24, No. 6, 2022, p. 831.

[22] G. M. Raja, M. Thaha, R. Latha, A. Karthikeyan, "Texture classification using optimized local ternary patterns with nonlinear diffusion as pre-processing", Multimedia Tools and Applications, Vol. 79, No. 5-6, 2020, pp. 3831-3846.

[23] L. A. Soares, K. F. Côco, P. M. Ciarelli, E. O. T. Salles, "A class-independent texture-separation method based on a pixel-wise binary classification", Sensors, Vol. 20, No. 18, 2020, p. 5432.

[24] M. J. Khan, M. A. Riaz, H. Shahid, M. S. Khan, Y. Amin, J. Loo, H. Tenhunen, "Texture representation through overlapped multi-oriented tri-scale local binary pattern", IEEE Access, Vol. 7, 2019, pp. 66668-66679.

[25] Z. Pan, X. Wu, Z. Li, "Scale-adaptive local binary pattern for texture classification", Multimedia Tools and Applications, Vol. 79, No. 9, 2020, pp. 5477-5500.

[26] Y. Feng, R. Guo, X. Shen, X. Zhang, "A measure for the evaluation of multi-focus image fusion at feature level", Multimedia Tools and Applications, Vol. 81, No.13, 2022, pp. 18053-18071.

[27] A. Ranganath, M. R. Senapati, P. K. Sahu, "Estimating the fractal dimension of images using pixel range calculation technique", The Visual Computer, Vol. 37, No. 3, 2021, pp. 635-650.

[28] A. Ranganath, M. R. Senapati, P. K. Sahu, "A novel pixel range calculation technique for texture classification", Multimedia Tools and Applications, Vol. 81, No. 13, 2022, pp. 17639-17667.

[29] A. Ranganath, P. K. Sahu, M. R. Senapati, "A novel approach for detection of coronavirus disease from computed tomography scan images using the pivot distribution count method", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 10, No. 2, 2022, pp. 145-156.

[30] A. Ranganath, P. K. Sahu, M. R. Senapati, "Detection of COVID from Chest X-Ray Images using Pivot Distribution Count Method", Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, Noida, India, 26-27 August 2021, pp. 373-378.

Segmentation of Medical Images with Adaptable Multifunctional Discretization Bayesian Neural Networks and Gaussian Operations

Original Scientific Paper

1. **Gomathi Ramalingam**
Department of Electronics and Communication Engineering,
University College of Engineering-Dindigul,
Tamilnadu, India
gomathiaudece@gmail.com

2. **Selvakumaran Selvaraj**
Department of Electrical and Electronics Engineering
PSNA College of Engineering and Technology,
Dindigul, Tamil Nadu, India

3. **Visumathi James**
Department of Computer Science and Engineering
Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
Chennai, India

4. **Senthil Kumar Saravanaperumal**
Computer Science and Engineering Department,
Sethu Institute of Technology,
Virudhunagar, India

5. **Buvaneswari Mohanram**
Department of CSE,
Paavai Engineering College,
Namakkal. India

Abstract *– Bayesian statistics is incorporated into a neural network to create a Bayesian neural network (BNN) that adds posterior inference aims at preventing overfitting. BNNs are frequently used in medical image segmentation because they provide a stochastic viewpoint of segmentation approaches by producing a posterior probability with conventional limitations and allowing the depiction of uncertainty over following distributions. However, the actual efficacy of BNNs is constrained by the difficulty in selecting expressive discretization and accepting suitable following disseminations in a higher-order domain. Functional discretization BNN using Gaussian processes (GPs) that analyze medical image segmentation is proposed in this paper. Here, a discretization inference has been assumed in the functional domain by considering the former and dynamic consequent distributions to be GPs. An upsampling operator that utilizes a content-based feature extraction has been proposed. This is an adaptive method for extracting features after feature mapping is used in conjunction with the functional evidence lower bound and weights. This results in a loss-aware segmentation network that achieves an F1-score of 91.54%, accuracy of 90.24%, specificity of 88.54%, and precision of 80.24%.*

Keywords*: medical image processing, segmentation, Gaussian filtering, discretization*

1. INTRODUCTION

To classify each pixel in an image, image segmentation is an important and difficult topic. Segregation of covering noise using scan images to detect diseases, lung segmentation using computed tomography (CT) visuals to help distinguish between cancerous and benign lung lesions, edge detection of internal organs using eye scan images to detect disease premature, and image segmentation for surgical preparation are just a few of its many uses [1]. It is highly difficult to accomplish precise and consistent separation of defects in medical imaging that are substantial variations in the outline or dimensions of the pathophysiology across people, although there may be a poor distinction between the classification target and adjacent tissues. It has been a few years since deep neural networks (DNNs) surpassed human specialists in several medical image segmentation tasks.

Decisions on how to segment an image are decided solely by the model, and this might lead to inaccurate predictions. Reliability is critical in areas like medical

diagnostics and self-driving car technology, where mistakes can have devastating effects if they are made in life-threatening conditions. The lack of interpretability of DNN predictions has led many to dismiss these systems as black boxes. Finally, the deep architectures used by DNNs may lead to overfitting as a result [2]. As a result, huge datasets and regularisation procedures are needed to thwart this trend. As an alternative, Bayesian neural networks (BNNs) might do away with the shortcomings of deep neural networks (DNNs) altogether.

For a BNN to be accurate, it must consider the uncertainty in its predictions, hence it treats weights and biases as random variables. Using a scalable methodology, these models help to minimize overfitting and to provide a sense of uncertainty. A disciplined approach to predicting and interpreting model outputs using BNNs is made possible by this method. Medical image segmentation relies heavily on uncertainty estimates since it may shed light on the level of confidence in a segmentation's conclusions and so assist doctors to make better diagnoses [3]. As a result of the requirement to use such approaches in real-world systems, researchers have become increasingly interested in deep learning and learning with Bayesian networks. BNNs are often trained using variational inference (VI), which transforms normal Bayesian learning into an optimization issue. Steps in the Medical image processing framework are shown in Fig. 1.

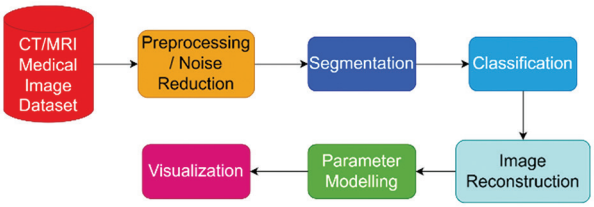


Fig. 1. Steps in Medical image processing framework

Approximating the posterior is easy using VI since it allows you to choose a vaguely resembling probability distribution and then optimizes it. Kullback–Leibler (KL) divergence may be used to quantify how close an estimate and the genuine subsequent models are, and reducing this discrepancy is comparable to normalizing the adverse indication lower limit. Based on VI, it's easy to build BNNs and they're resistant to overfitting. Graves presented a method to evaluate a factorized subsequent or a partial calculator using data subsampling techniques. This made it possible for the method to handle vast volumes of data [4].

An effective stochastic gradient VI technique, known as Bayes by Backprop (BBB), was devised by Blundell et al. based on this to measure weight uncertainty. The use of Gaussian approximation distributions indeed results in a larger number of model parameters, which makes this variational technique less suitable for usage with big complicated models than prior variational approaches [5]. The model quality and uncertainty estimate of probabilistic backpropagation and approximation inference approaches that rely upon deviation reduction take greatly enhanced over these VI techniques. However, the enormous processing costs, scale, and nonlinearity make these VI techniques computationally infeasible.

GPs provide exact estimates of uncertainty for trustworthy forecasting, also the association among GPs were already been widely investigated. Neal claims that when the BNN width grows, GP is the limiting distribution. The system having dropping levels is analogous to approximation for dynamic modeling with Monte Carlo (MC) simulations in the forecast at runtime. Using the Monte Carlo dropping (MC-Dropout) technique to deep GPs may be seen as a Bayesian calculation. They share a commonality of deep structures, which explains their close association. According to Ma et al., BNNs provide a contained randomized development, and for applying variational methods for training and reasoning throughout a functional domain, BNN models are established.

An investigation of the VI between GPs led Sun et al. to create functional variational BNNs (FBNNS) that are used to specify an aim and to estimate the role at limited inputs. Carvalho et al. introduced a GP variational family technique based on the fELBO goal, which allows functional VI to be used for high-dimensional feature extraction problems. For every network design, this approach delivers well-calibrated uncertainty estimates and only one forward pass is required for predictive inference [6].

In the case of limited hyperparameters and a big dataset, the query optimizer prefers to optimize the aggregate log-likelihood component rather than the negative log-likelihood. As a result, determining the optimal hyperparameters will be difficult [7]. It is also computationally difficult when utilizing powerful BNNs to compute the model posterior, culminating in systems that are costly to run. Many people have worked to develop BNNs that can both scale and estimate. When it comes to segmenting medical images, the model's performance is hampered by intricate border interactions, greater visual variability, and low extracellular matrix contrast [8].

Consequently, the identification of the most significant attributes is critical to the model's effectiveness. Here, we introduce a reliable FBNN that does medical image segmentation to overcome these concerns. Variational inference is performed on the posterior and variational prior probabilities, which we refer to as GPs. The value obtained from the KL deviation unit is analogous to the adverse log probability when using a threshold with KL deviation in the variation. Rather than immediately increasing in proportion to the number of variational parameters, the KL divergence grows as the quality of the discretization parameters improves. As a bonus, we present a medical image segmentation that is cognizant of content [9].

Feature upsampling uses content-aware reassembly of features to retrieve semantic data from input feature maps, which improves model performance while decreasing computational costs. With one forward pass, the suggested BNN may perform predictive inference in high-dimensional problems. This technique outperforms several current Bayesian models for segmentation evaluation criteria and confidence approximations on open medical image data sources.

The research work proposed in this article possesses the following key contributions: To assist the model in discovering the most optimum variational parameters, we present a weighted version of the loss function parameter model in function space. An adaptive framework is presented for the segmentation of medical images. When building this model, we took advantage of a computationally efficient up-sampling component. Experiments were conducted with the openly available dataset that shows that the proposed model outmatches several current approaches in terms of segmentation performance measures and uncertainty estimations.

2. LITERATURE SURVEY

In the field of medical image segmentation, UNet has achieved greater success. FCNs are used to create a novel architecture in this network. With this UNet design, numerous better segmentation processes, like as residual as well as dense mechanisms, have been presented. An algorithm created by Isensee et al., called nnU-Net, can autonomously arrange itself for any biomedical job, encompassing pre-processing, network design and training, and post-processing. nnU-Net is a free, open-source technology that outperforms most existing techniques [10]. For the COVID-19 lesion region, a content-aware residual UNet was proposed by Xu et al.

Segmentation results are more accurate and computational costs are reduced with this approach. There is no probability calculation for the segmentation findings of these vanilla DNNs, thus they make judgments only based on point prediction. In image segmentation, BNNs generate predictions for each pixel as well as estimations of pixel-wise uncertainty. When it comes to making decisions, it's critical to consider the principles of uncertainty. Deep ensemble methods and Bayesian approximation are the two most used approaches to quantifying segmentation ambiguity in biomedical imaging using BNNs [11]. As a strong approach for assessing uncertainty, the deep ensemble assesses the model's uncertainty by training numerous models and collecting their output predictions' variances, which are then averaged together.

On two publicly accessible echocardiography datasets, Dahal et al. examined various ensemble-based indecision approaches. Four metrics were used to measure uncertainty and show how uncertainty estimates are done automatically and enhance the outcomes. They were shown to be effective [12]. Since BBB and other typical Bayesian approaches cannot accurately approximate the average Bayesian model, Wilson and others have shown that deep ensemble models can. Dropout is a regularisation term in the MC-Dropout technique of Bayesian approximation, which is commonly used in medical imaging applications to quantify the prediction uncertainty [13]. This approach is widely used since it is simple to implement.

As a result, weight perturbation-based methods suffer from large variations in gradient estimation since they use the weight space's intractable prior and parameterized a GP as the discretization posterior [14]. CNN was used for the GP prior kernel and the GP posterior kernel. There are issues with their approach, however, such as the inability of fELBO loss to get the optimum solution in function space in all cases and the enormous computing cost. CARAFE is used as an optimized feature average pooling operator in the CAUNet, unlike the other Bayesian approaches discussed [15]. Competingly efficient, the suggested segmentation approach enhances model performance indicators and uncertainty estimations.

It is vital to segment important items in medical images and extracts information from segmented sections to aid physicians in making correct diagnoses. Feature extraction, clustering approaches, bayesian hierarchical frameworks, dynamic patterns, computer vision, etc. are common in early medical image segmentation approaches. An updated edge identification approach based on mathematical morphology was proposed in [16] for CT scans of the lungs. Disc inspection using Hausdorff-based template matching was performed by [17], while ventricular fragmentation in brain CT images was accomplished using a similar method by [18].

To segment cardiac MRI images in 2D and prostate MRI images in 3D, [19] suggested a shape-based technique using horizontal sets. Liver tumors in abdominal CT images were segmented using the activation profile model by [19], while [20] designed a methodology for medical anatomy data segmentation using level sets and SVM classifiers (Support Vector Machines). Brain MRI images were segmented using Markov random fields (MRF) by Held et al. Image segmentation is still one of the most demanding subjects in computer vision owing to the difficulty of feature representation, even though several algorithms have been described and are successful in specific conditions.

Because of issues like a blur, noise, low contrast, etc., medical images make it more difficult to extract discriminating characteristics than standard RGB images [23]. Convolutional neural networks (CNN) accomplish hierarchical visual features of images, making medical image segmentation the hottest research area in image processing and computer vision thanks to the rapid growth of deep learning techniques. Thanks to the robustness of CNNs for feature learning, medical image segmentation is not negatively impacted by common

image artifacts such as clutter, distortion, sharpness, etc. Semantic segmentation and instance segmentation are the two main types of image segmentation tasks now in existence [24].

Image segmentation is used as a probabilistic classifier that labels each pixel in an image with a certain category. Instance segmentation, in contrast to semantic segmentation, requires not just pixel-level categorization but also the ability to differentiate instances based on specified categories. Since every organ and tissue is unique, there have been surprisingly few papers on detection and segmentation in healthcare image segmentation [25]. We take a look back at the progress made in applying deep learning to the task of segmenting medical images. Machine learning is generally broken down into subfields called supervised learning, semi-supervised learning, and unsupervised learning based on the availability of labels for training data. It is challenging to get a significant amount of labeled data for medical imaging, despite supervised learning's benefit of training models based on correctly tagged data.

Unsupervised learning, on the other hand, can be done without labeled data, although it is more challenging. Weakly supervised learning bridges the gap between supervised and unsupervised learning by simply requiring a subset of data to be labelled [26]. Medical image segmentation had previously been the subject of several model-driven techniques reported by researchers before deep learning's mainstream adoption. Image clustering, region analysis, and random forest are only a few examples of model-driven approaches in medical image analysis that were summarised in detail by [27]. Using a variety of mathematical models, the authors of [21] summarised the many techniques used to do segmentation on medical images.

For medical image segmentation, only a small number of research using model-driven approaches have been reported recently, whereas an increasing number of studies are using data-driven techniques. The growth and improvement of deep learning models for medical image segmentation are the primary topics of this research. Shen et al. gave a comprehensive study of deep learning's use in medical image processing [28]. Here, we take a look back at how far we've come with computer-aided illness diagnosis and prognosis, tissue segmentation, machine learning, and deep learning in the field of medicine. An overview of deep learning approaches was recently given by [29], which discusses the application of deep learning to various tasks such as image classification, object identification, segmentation, registration, and more [22].

The recent growth of semantic and medical image segmentation was discussed by [30], who divided deep learning-based image segmentation alternatives into six categories: deep modern architecture, data biosynthetic pathway, loss function-based, sequential models, weakly carefully monitored, and multi-task approaches. In [31], the authors evaluated artificial neural networks (ANNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) as well as classical machine learning algorithms like Markov random fields, k-means clustering, and random forest to create a more comprehensive survey on medical image segmentation (RNNs). Solutions for medical image segmentation with incomplete data sets were evaluated by the authors of [32], with the authors highlighting the constraints of both sparse and weak annotations as important obstacles. All of these surveys are crucial to improving medical image segmentation methods.

The methodologies such as network architecture, training methods, and difficulties were all covered in [33]. This article breaks down the most well-known and widely-used network architectures for image segmentation. The J Digit imaging method for training deep neural network models is covered in the section devoted to training methods. The problems of employing deep learning algorithms for medical image segmentation are described in detail in the next section. In a recent study, the researchers [34] examined the progress made in applying deep learning to chemotherapy and the possibilities for its future use. In their recent article, the authors of [35] summarised the state-of-the-art methods for quantitative brain MRI image segmentation using deep learning. Incomplete supervision, inexact supervision, and incorrect supervision were the primary areas of concern [36]. Optimizing approaches for medical image semantic segmentation are evaluated and summarised by Eelbode et al. [37], with a particular emphasis on Dice scores as well as Jaccard values.

2.1 RESEARCH GAP

Unfortunately, there is a critical flaw in the formulation that prevents the loss from always leading to appropriate variational parameters when adopting existing approaches. Two factors, the inverse log-likelihood as well as the deviation among the Bayesian approximation that the prior distribution, contribute to the optimization of the specified loss, and their relative importance is determined by the size of the collection and the model's parameters [16]. When training a model with few parameters using a large dataset, the continuous random variable likelihood will initially be very high, whereas another term will be extremely low. It is expected that the optimizer will give priority to the log probability term. Consequently, identifying the best values for the model parameters will be challenging [17].

In addition, the inference of the model posterior is computationally intractable, and powerful BNNs often have a high number of parameters, leading to computationally costly models. Consequently, a lot of effort has gone into creating approximation BNNs that can scale well [18]. Last but not least, the medical image segmentation job has its unique challenges due to the complicated border interactions, increased appearance variance, and poor surrounding tissue contrast.

That's why it's crucial to accurately extract the most relevant features for the model to work.

3. SYSTEM MODEL

Here, we quickly examine a generic explanation and then present the probabilistic VI design, and operational VI. Let us now turn our attention to what we refer to as content-aware upsampling.

In a chosen dataset that comprising of pairs of data denoted by $D=\langle a_x, b_{ox} \rangle$, where x is ranging from 1 to P. The objective variable a_x is a scalar, and we have a total of P observation data pairings where D is the feature vector and $b_x \in Q$ are the feature matrix, and these are represented by a_x and b_x, respectively. Neural networks are used in parametric regression, where a collection of parameters (a) is used to define an appropriate function for the inputs and outputs (b) of the neural network. A prior probability over through the domain of system parameters was inserted in BNNs, where weights and biases are considered random variables, and the weight matrixes for each layer are treated as random variables. This prior distribution is a prediction of which parameters may have created the results before any data was observed. The probability over parameter space was calculated by applying Bayes' theorem to a set of statistically independent observations, i.e., D. $P(A \mid B)$ defines the conditional probability, and $P(A)$ and $P(B)$ are the associated probability of A and B.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1)$$

where $P(B)$ has been estimated as,

$$P(B) = \int P(B|A)\ P(A)\ dA, \quad (2)$$

$$\text{and} P(B|A) = \prod_{x=1}^{P} P(a_x|b_p, A) \quad (3)$$

which represents the normalization component and the log-likelihood value respectively. The parameters of the network that are most likely to be affected by the observed data are represented by the probability density. Assuming a SoftMax probability for the classification problem, we may conclude that,

$$P(a_x = d|a_x, B) = softmax(g(a_x{:}B)) \quad (4)$$

$$\text{as·well·as·} P(b_x|a_x, B) = \mathcal{G}(b_x; g(a_x{:}B), \mu^2 J) \quad (5)$$

for the assumed data pairs.

Image segmentation and tracking have received a lot of emphasis from deep learning applications, but the recognition system hasn't received the same level of attention thus far. Though several promising works have been published, this does not mean there aren't a lot of possibilities. In point-based registration, finding good features extracted that allow for the proper identification of associated points is a common difficulty. The difficulty of finding a specific constituent in a medical image is addressed by image detection and recognition. The images are often volumetric. Consequently, fast parsing is essential.

The most common method is marginal space learning, which is both efficient and resilient in detecting organs. To make it even more efficient, its deep learning counterpart uses a computer vision boost cascading instead of a probabilistic tree. Still, the complete volume must be processed to accurately identify anatomical features. Using deep reinforcement learning, the search process may be replaced by an artificial entity that explores anatomy to identify anatomical structures. In only a few seconds, the approach can identify hundreds of landmarks throughout a complete dataset.

Because it lacks a closed-form equation, the real posterior distribution could be computed analytically in practice; an approximation must be employed instead. The posterior is frequently approximated using VI. Free variational parameters are used to construct an approximate variational distribution. In addition, the parameters of the approximation distribution and the real posterior are modified to minimize the dispersion.

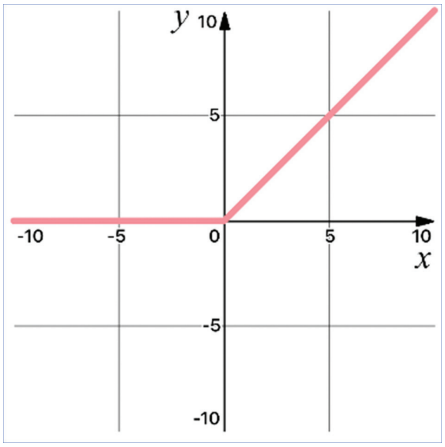


Fig. 2. Activation Function

$$\hat{\rho} = argmin_{\rho}(\mathcal{G}(b_x; g(a_x{:}B)) \quad (6)$$

Decreasing the negative parameters is the same as minimizing the KL divergence, as shown by the equation. Due to BNNs being highly nonlinear, this equation cannot be used to compute h analytically. It's possible to lower the volatility of the baseline predictor for the discretization of BNNs by using the re-parametrization approach, which has just been developed in VI. This technique, known as re-parametrization, was used to choose samples from the discretization distribution in which a point-wise combination and the standard deviation are nonnegative and may be represented. The loss function may be estimated by combining the re-parametrization method with an MC approximation.

The up-sampling approach has three advantages. It has a wide field of view, is inexpensive and fast to calculate, and is content-aware. A parameterization activator with content-aware kernels enhances the effectiveness in object identification, classification techniques, and

texture features compared to traditional benchmarks. We use the following steps to accomplish upsampling. When compressing the input feature channel, we first used a convolution layer. The reassembly kernels were then encoded depending on the content of input features using a second higher-order convolution layer. SoftMax function was performed on each reassembly kernel geographically before upsampling to ensure that all kernel values were equal. Upsampling does not rescale or modify the average scores of the convolution layer because of normalization. It is possible, to sum up, the kernel forecasting component in this way:

$$Y_g = V(C_{3\times3}(C_{1\times1}(a_x))) \cdot \tag{7}$$

where the convolutional with the kernel is denoted by the symbol C. One pixel in a kernel has the location in the feature map, which is the result of the content encoder operation. V is the algorithm that works the view of the underlying Kernel, and a_x is the feature map after the information encoder. As a final step, we reconstruct the characteristics into anticipated kernels, which we may represent as kernels.

3.2 PROPOSED MODEL

Autoencoder and linear interpolation blocks comprise the Bayesian antecedent system. Extension of the model that takes advantage of the encoding-decoding structure. We use the model as a feature upsampling operator instead of a separate feature extraction tool. Using convolutions instead of pooling processes result in connections with a smaller memory footprint when training because no switches are required to map the output of the pools back to their inputs. Additionally, networks with smaller memory footprints can be effectively understood and analyzed by using only deconvolutions instead of unspooling operations. For the following network layers, we can use a larger receptive field with less signal complexity by downsampling input data.

During each step of the network's left side, two times as many features are calculated as the preceding layer. Feature extraction and spatial expansion of lower-resolution feature maps are performed in the right section of the network so that the essential information may be assembled to produce a two-channel volumetric segmentation. We use soft-max voxelwise mostly on feature space maps obtained by the final fully connected layer, each with a 1×1 kernel size, to produce randomized segmentation of their original image regions. There are one to three convolutional layers, each with half the average of 5 × 5 × 5 kernels, that follow each phase of the CNN's efficient distribution channels. This increases the input data size. It's just like on the left side of the network, where we use residual functions to learn in the convolutional phases.

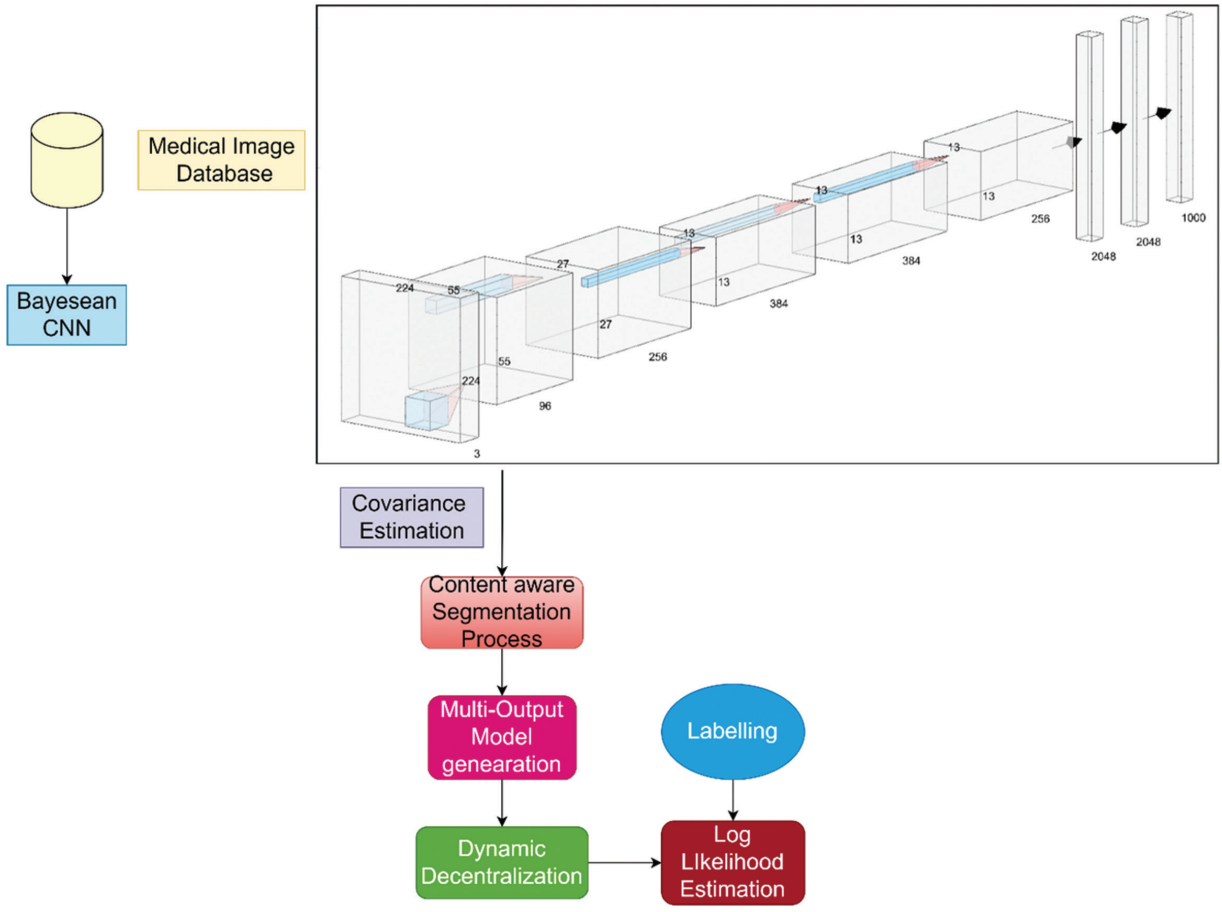


Fig. 3. Proposed image segmentation model architecture

Of the identical accuracy as that of the initial input data, the system recommendations in two dimensions are processed by a softmax layer. To determine the possibility of each voxel being in the foreground or background, each layer provides the probability. In medical datasets like the types we're dealing with, the architecture of interest is often only a small percentage of the scan. An untrained network will be significantly biased towards the background because of this, and learning will become stuck at the global optimum of the gradient descent. Consequently, the foreground is sometimes absent or just partially discernible. Sample reweighting was used in several earlier ways to reduce loss functions where foreground areas were given greater weight than ambient ones during learning.

Our goal is to maximize the coefficient that takes a number between Zero and one in this research, which we define as an objective function. It is unnecessary to weigh a variety of various classes when using this formulation, and the results obtained are superior to those obtained through another network trained to optimize a logistic regression loss with sample reweighting, as we have experimentally observed. This is because we don't have to weigh samples of different classes. A ReLU activation function and the CNN precondition model are also employed. The correlation coefficient H can be calculated accurately by a Bayesian CNN with arbitrarily numerous convolutional filters, which are identical.

$$\mathcal{M} = \rho LM[r(g^{y})|| \, q(g^{y})] - \sum_{j=1}^{M} \wp_{r0}(g(y_j)) \left[\log q \, (z_j[g(y_j)]\right] \tag{8}$$

Additionally, it is possible to get the inverse correlation matrix and eigenvalue of K. Thus, the kernel for segmentation tasks was included in the design. This KL divergence component may be computed using mappings that have been parametrized. A kernel matrix represents a square matrix for massive Bayesian CNNs because of the zero pixel-pixel correlations in Bayesian CNNs without pooling layers. Images with dimensions H and W that have C layers can be used as inputs to the algorithm.

$$\sum_{mn} \frac{1}{M} \sum_{l=1}^{M} f_l(a_m) \oplus f_l(a_n) + G(a_m, a_n)\rho(a_m, a_n) \tag{9}$$

We utilized a loss function that is inverse to the BNN's true loss value to train it. KL term size and log-likelihood term expectations are both dependent on the dataset and model parameters. KL divergence can be substantially larger than low log-likelihood when the dataset is small but the model has many parameters. Optimizers will always favor high-value terms over low-value ones in these two scenarios. This means that it will be hard to get the best dynamic features for CNN. Consequently, a b-weighted loss function has been included to overcome this issue.

The extra expanding route, multi-dimensional identity process, and dissimilar multi-scale convolutional blocks are the three key additional advanced components in the suggested model compared to the standard U-Net. To enhance the model's learning power through dual supervision, a new, expanding path is developed to introduce a new learning loss, namely, auxiliary loss. Since the suggested approach incorporates both the image features from the convolution layer and the intermediary classification values from the extra expanding path, it can produce more precise segmentation results. In addition, the suggested model suggests a strategy to deal with the issue of irrelevant information by employing two successive self-attention components, dense space orientation focus, and connection recognition, to capture the significance of characteristics in both the positional and multichannel aspects.

Applying a dilated convolution block in the model, the input feature map is transformed into a dense feature matrix of reduced size, which is then utilized to estimate the spatial dependencies, allowing for efficient estimation of the weights of features in the spatial structure. Furthermore, the proposed model's blocks are put to use to address the semantic gap through the usage of multiscale convolution kernels that are coupled in a variety of ways (series and parallel) to accommodate for the differences in convolutional size. Thus, the created multi-scale feature maps may be used more effectively, and the combined feature maps can better minimize semantic discrepancies since they maintain more extensive semantic information with varied scales. The model was designed to segment 2D medical images, but it can be readily expanded to a 3D model to analyze 3D medical images in a manner analogous to how U-Net can be extended to a 3D model.

The initial upgrade to the model is the inclusion of an extended path to U-Net so that dual supervision may provide more precise results when segmenting medical images. Specifically, the suggested model contains just a single contracting route, like U-Net, but unlike U-Net, after the most speculative characteristic mappings have been collected at the innermost part of the convolution layer, they are transmitted to two expanding pathways with identical topologies. Both the previous expansive path, which was already present in U-Net and the new expansive path, which was just introduced, have names.

Similar to the skip connection operation in U-Net, we combine the feature maps produced by the corresponding layer of the convolution layer with the feature maps produced by the transfer function fully - connected up-sampling operations in the final layer to create the feature maps used in the additional expansive path. The new expansive path's combined feature maps are then given to the layer that follows it, as well as the layer that the initial expansive path maps to. As a result, the unsampled feature maps from the new expansive path are added to the feature maps again from the convolution layer and the original expansive path

in each layer. Last but not least, the supplementary expansive path is used to train the deep model using a pair of segmentation losses, one from the primary extensive channel and one from the supplementary extensive channel.

There are two types of uncertainty in BNN predictions: perceptual and aleatoric. Model uncertainty, another name for ambiguity, is a way of quantifying the unknown. This makes evaluating Bayesian uncertainty estimation difficult because there is no integral gain for the estimations. To assess the models' ability to estimate uncertainty, we combined the ground truth label with predictions from the target model and predictive entropy. The consistency map and the confidence map are necessary to compute these metrics. Accordingly, the correctness and uncertainty maps may be calculated by matching the regression coefficients labels and model predictions, respectively. The four types of evaluations are False Positive (TP), True Negative (TN), False Positive and False Negative (FP), and True Negative (TN) wrong and certain (False Negative, FN).

4. RESULTS AND DISCUSSION

The lung segmentation dataset was the subject of several investigations. The Kaggle Data Science Bowl presented the lung segmentation dataset in 2017 [38]. There are two- and three-dimensional CT images with lung segmentation labels labeled. There is a total of 512×512 pixels in each image, so it's rather detailed. The lung region was isolated from the surrounding areas and the images were divided at random into three sets: training, validation, and testing. We utilized 450 training images, 120 validation images, and 400 testing images in this work. There was an 8-node Bayesian CNN GP preceding network with weight and bias basis functions with variances of 0.2 and 0.08. Content-aware upsampling used the same hyperparameters.

To verify that the transverse members of the covariance matrix are not zero, we set L to 0.30 and included a threshold value of 1.03 to the transverse. We had an initial learning rate of 0.01 a weight decay rate of 0.04, and an annealing factor of 0.998, the optimization technique used stochastic gradient descent (SGD). We trained in batches of four, and for the DRIVE dataset, we trained in batches of twenty patches. Systems were accomplished for 200 epochs for segmentation experiments. Here, the proposed models were trained for 125 epochs. A Linux computer with a Geforce P5000 GPU was used for the tests, which were carried out using PyTorch for all of the models. There are several different baselines against which we might measure the suggested approach's efficacy and superiority. BNN's GP prior in feature space is the inspiration for our concept.

As a result, this method was the most natural starting point. All of these deep learning inference methods use an approximation of Bayesian inference and express uncertainty naturally. It is possible to assess uncertainty in a model using a deep ensemble. Class-likelihood estimates are produced by the UNet design with a softmax output layer, and this architecture may express the network's uncertainty. Non-Bayesian uncertainty quantification approaches such as Softmax and Deep Ensemble are frequently employed. Moreover, to properly test the influence of various parameters of b upon segmentation efficiency, we applied the described segmentation method.

Table 1. Performance analysis with existing models

	F1-Score	Accuracy(%)	Specificity	Precision
Softmax	**0.8798**	**85.78**	**0.8154**	**0.7854**
FBNN	**0.8547**	**85.78**	**0.8124**	**0.7754**
Proposed	0.9154	90.24	0.8854	0.8024

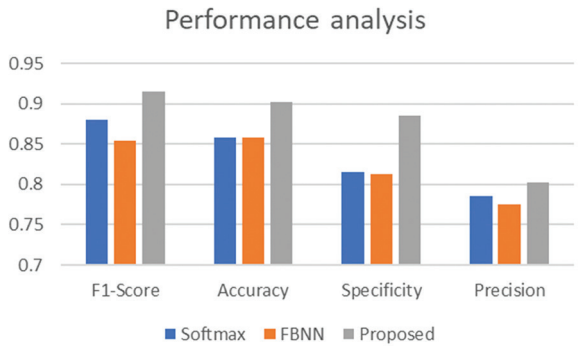


Fig. 4. Performance analysis

So that all models could be compared fairly, we chose the same UNet design, which consists of five encoder and decoder blocks. The dropout frequency was adjusted to 0.2 for MC-Dropout in the ensemble of four identical UNets. Mostly from the dataset of lung segmentation, the various approaches are evaluated. For this test set, the results were obtained by utilizing the suggested and baseline procedures. In this dataset, our technique surpassed all of the baseline methods, notably precision, our method's outcome is somewhat better than that of both b1 and b2. MC-Dropout was the poorest of the test outcomes. The superiority of our strategy in comparison to this dataset shows that the suggested method is useful in increasing the model's performance and generalization capacity.

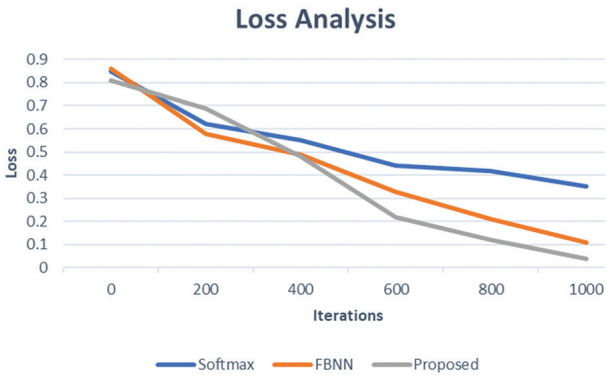


Fig. 5. Loss Analysis

Estimated ROC curves are used to demonstrate the overall efficiency of the suggested and baseline approaches on the four datasets. The true positive rate (TPR) is shown alongside the false positive rate (FPR). The more accurate a model is, the closer the ROC curve gets to the top-left boundary of the coordinates of ROC. In terms of techniques, the b2 slope is perhaps the most left-leaning among them. Figures also demonstrate that our technique has the biggest area under the ROC curve compared to the other methods (AUC). That our approach works better than others is further supported by these findings. Using a variety of Bayesian frameworks and datasets, a qualitative analysis may make use of the outcomes of segmentation and uncertainty in the models.

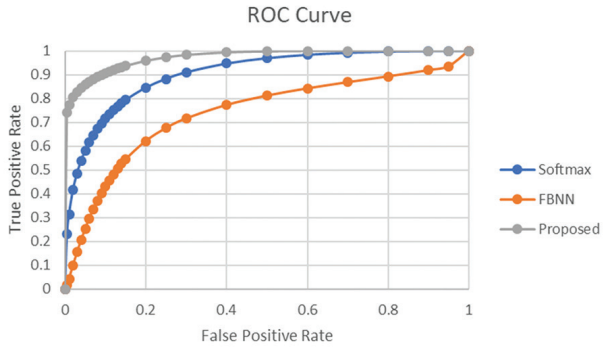


Fig. 6. ROC curve Analysis

This target's contours aren't well-segmented when using Prob. UNet, MCDropout, and FBNN predictions against the respective ground truth labels. Our approach has a segmentation boundary that is closer to the real world than the baseline methods do. When used on the lung segmentation dataset, MC-Dropout and SWAG both under-segment and under-represent the overall shape. In the lung segmentation test, our solution outperformed the standard methods by a wide margin. Aside from this, the suggested approach was able to properly detect and segment the existence of vessels in blood and the images of the retina. Small blood arteries, on the other hand, did not segment effectively using the baseline approaches.

Table 2. ROC Formulation

FPF	TPF	Lower	Upper
0.005	**0.2301**	**0.0169**	**0.7407**
0.01	**0.3135**	**0.043**	**0.7718**
0.02	**0.4168**	**0.0996**	**0.8061**
0.03	**0.486**	**0.1545**	**0.8282**
0.04	0.5384	0.2056	0.8449
0.05	0.5807	0.2523	0.8587
0.06	0.6159	0.2949	0.8705
0.07	0.6461	0.3337	0.8808
0.08	0.6723	0.369	0.8901
0.09	0.6955	0.4012	0.8985

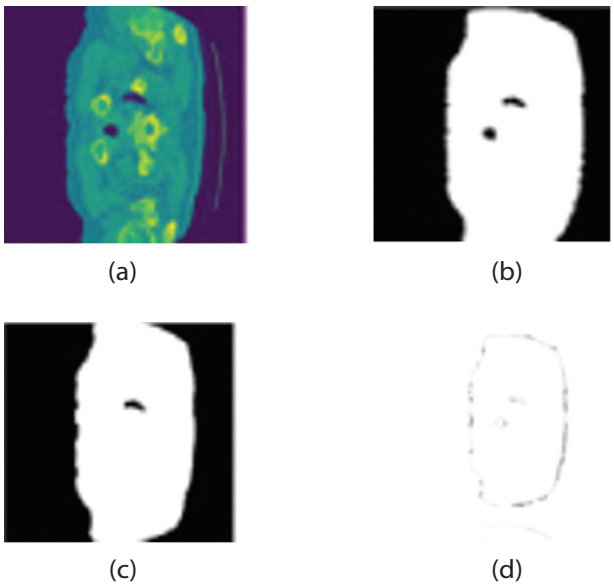
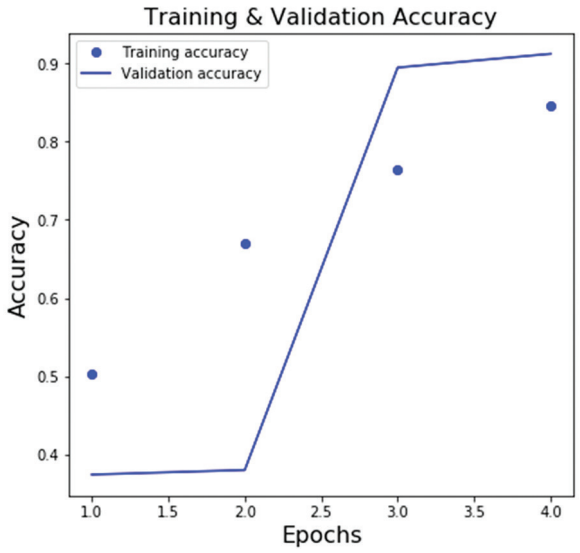
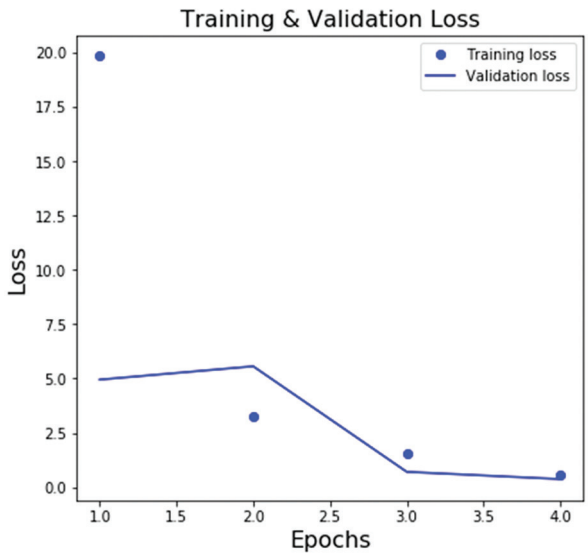


Fig. 7. Segmentation Outcomes: (a) Original (b) Edge detection (c) Region identification (d) Segmented Features

The incorrect and uncertain pixels are highlighted in the correctness and confidence maps, respectively. The confidence maps for various approaches in the image show that models usually have considerably larger ambiguity for the boundary of the classes, which indicates that the model underperforms over these classes. A high level of confidence is also found in pixels that are distant from the edge of the screen. It can be observed from the figure's correctness maps that wrong areas tend to be located along the boundaries of classes. These findings show that models typically produce large uncertainty estimations when the forecast is incorrect. When compared to the baseline approaches, all segmented regions, even accurate regions, are subject to considerable uncertainty. Overall, our technique outmatches the baseline methods in terms of segmentation and estimations of uncertainty.



(a)



(b)

Figure 8. (a) Training and validation Accuracy (b) Training and validation Loss

5. CONCLUSION

A reliable method for segmenting medical images has been discussed in this research paper. This can be defined as the distribution function and the variational posterior of the variational goal. It was possible to train BNNs with GPs by using a model variation that included a b-weight mostly on KL deviation unit for the functional VI goal, which we presented. A CNN that takes advantage of the downsampling operator to enhance model performance while lowering computing costs was also presented. The proposed approach outperforms current methods in terms of reliability, ambiguity prediction, and accuracy, according to the findings of the assessment studies. The performance is restricted by the function before and the parameterization of the likelihood function. The Bayesian CNN prior network must be properly designed, and the finite difference posterior must be precisely parameterized.

As shown by the experimental findings, the suggested technique achieves better performance, uncertainty estimates, and inference time than the state-of-the-art methods. This results in a loss-aware segmentation network that achieves an F1-score of 91.54%, accuracy of 90.24%, specificity of 88.54%, and precision of 80.24%. We need to carefully build the Bayesian CNN GP prior network and parametrize the variational posterior to maximize the suggested method's performance, which is constrained by the stated function before and the parameterization of the posterior distribution. We hope to improve upon our current technique of parameterizing the covariance of the posterior distribution and investigate the potential of applying it to the segmentation of 3D medical images in the future. The automated configuration of a BNN like nnU-Net, including pre-processing, network design, and training for medical image segmentation, presents a significant challenge.

6. REFERENCES

[1] X. Chen, Y. Zhao, C. Liu, "Medical image segmentation using scalable functional variational Bayesian neural networks with Gaussian processes", Neurocomputing, Vol. 500, 2022, pp. 58-72.

[2] S. D. Halle, D. N. Dunn, A. H. Gabor, M. O. Bloomfield, M. Shephard, "Bayesian dropout approximation in deep learning neural networks: analysis of self-aligned quadruple patterning", Journal of Micro/Nanopatterning, Materials, and Metrology, Vol. 21, No. 04, 2022.

[3] J. Steinbrener, K. Posch, J. Pilz, "Measuring the Uncertainty of Predictions in Deep Neural Networks with Variational Inference", Sensors, Vol. 20, No. 21, 2020, p. 6011.

[4] T. Hussain, "ViPS: A novel visual processing system architecture for medical imaging", Biomedical Signal Processing and Control, Vol. 38, 2017, pp. 293-301.

[5] C. Villacampa-Calvo, D. Hernández-Lobato, "Alpha divergence minimization in multi-class Gaussian process classification", Neurocomputing, Vol. 378, 2020, pp. 210-227.

[6] Y. Li, Q. Zhang, S. W. Yoon, "Gaussian process regression-based learning rate optimization in convolutional neural networks for medical images classification", Expert Systems with Applications, Vol. 184, 2021, p. 115357.

[7] A. Maier, C. Syben, T. Lasser, C. Riess, "A gentle introduction to deep learning in medical image processing", Zeitschrift für Medizinische Physik, Vol. 29, No. 2, 2019, pp. 86-101.

[8] S. Bansal, "Determining Disease Using Machine Learning Algorithm in Medical Image Processing: A Gentle Review", Biomedical Statistics and Informatics, Vol. 6, No. 4, 2021, p. 84.

[9] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A. K. Nandi, "Medical image segmentation using deep learning: A survey", IET Image Processing, Vol. 16, No. 5, 2022, pp. 1243-1267.

[10] B. Wiestler, B. Menze, "Deep learning for medical image analysis: a brief introduction", Neuro-Oncology Advances, Vol. 2, Supplement_4, 2020, pp. iv35-iv41.

[11] A. Mobiny, P. Yuan, S. K. Moulik, N. Garg, C. C. Wu, H. Van Nguyen, "DropConnect is effective in modeling uncertainty of Bayesian deep networks", Scientific Reports, Vol. 11, No. 1, 2021.

[12] B. Pandey, D. Kumar Pandey, B. Pratap Mishra, W. Rhmann, "A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions", Journal of King Saud University - Computer and Information Sciences, Vol. 34, No. 8, 2022, pp. 5083-5099.

[13] S. Atasever, N. Azgınoglu, D. S. Terzı, R. Terzı, "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning", Clinical Imaging, 2022.

[14] Z. Qiu, Y. Hu, J. Zhang, X. Chen, J. Liu, "FGAM: A pluggable light-weight attention module for medical image segmentation", Computers in Biology and Medicine, Vol. 146, 2022, p. 105628.

[15] L. Zhang, B. Zhu, C. Ma, "Utsn-Net: Medical Image Semantic Segmentation Model Based on Skip Non-Local Attention Module", SSRN Electronic Journal, 2022.

[16] P. Tang, P. Yang, D. Nie, X. Wu, J. Zhou, Y. Wang, "Unified medical image segmentation by learning from uncertainty in an end-to-end manner", Knowledge-Based Systems, Vol. 241, 2022, p. 108215.

[17] R. Wang, S. Chen, C. Ji, J. Fan, Y. Li, "Boundary-aware context neural network for medical image segmentation", Medical Image Analysis, Vol. 78, 2022, p. 102395.

[18] J. Wan, S. Yue, J. Ma, X. Ma, "A coarse-to-fine full attention guided capsule network for medical image segmentation", Biomedical Signal Processing and Control, Vol. 76, 2022, p. 103682.

[19] Z. Wu, J. Wei, J. Wang, R. Li, "Slice imputation: Multiple intermediate slices interpolation for anisotropic 3D medical image segmentation", Computers in Biology and Medicine, Vol. 147, 2022, p. 105667.

[20] X. Xie, X. Pan, W. Zhang, J. An, "A context hierarchical integrated network for medical image segmentation", Computers and Electrical Engineering, Vol. 101, 2022, p. 108029.

[21] X. Guo, C. Yang, Y. Yuan, "Dynamic-weighting hierarchical segmentation network for medical images", Medical Image Analysis, Vol. 73, 2021, p. 102196.

[22] S. Rani, "A novel mathematical morphology-based edge detection method for medical images", CSI Transactions on ICT, Vol. 4, No. 2-4, 2016, pp. 217-225.

[23] M. Lalonde, M. Beaulieu, L. Gagnon, "Fast and robust optic disc detection using pyramidal decomposition and Hausdorff-based template matching", IEEE Transactions on Medical Imaging, Vol. 20, No. 11, 2001, pp. 1193-1200.

[24] W. Chen, R. Smith, S.-Y. Ji, K. R. Ward, K. Najarian, "Automated ventricular systems segmentation in brain CT images by combining low-level segmentation and high-level template matching", BMC Medical Informatics and Decision Making, Vol. 9, No. S1, 2009.

[25] A. Tsai et al. "A shape-based approach to the segmentation of medical imagery using level sets", IEEE Transactions on Medical Imaging, Vol. 22, No. 2, 2003, pp. 137-154.

[26] J. Song, Z. Zhang, "Magnetic Resonance Imaging Segmentation via Weighted Level Set Model Based on Local Kernel Metric and Spatial Constraint", Entropy, Vol. 23, No. 9, 2021, p. 1196.

[27] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, Vol. 60, No. 6, 2017, pp. 84-90.

[28] M. Florkowski, "Classification of Partial Discharge Images Using Deep Convolutional Neural Networks", Energies, Vol. 13, No. 20, 2020, p. 5496.

[29] S. Budd, E. C. Robinson, B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis", Medical Image Analysis, Vol. 71, 2021, p. 102062.

[30] S. Budd, E. C. Robinson, B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis", Medical Image Analysis, Vol. 71, 2021, p. 102062.

[31] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, G. Hamarneh, "Deep semantic seg-

mentation of natural and medical images: a review", Artificial Intelligence Review, Vol. 54, No. 1, 2020, pp. 137-178.

[32] H. Seo et al. "Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications", Medical Physics, Vol. 47, No. 5, 2020.

[33] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation", Medical Image Analysis, Vol. 63, 2020, p. 101693.

[34] M. H. Hesamian, W. Jia, X. He, P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges", Journal of Digital Imaging, Vol. 32, No. 4, 2019, pp. 582-596.

[35] P. Meyer, V. Noblet, C. Mazzara, A. Lallement, "Survey on deep learning for radiotherapy", Computers in Biology and Medicine, Vol. 98, 2018, pp. 126-146.

[36] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, B. J. Erickson, "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions", Journal of Digital Imaging, Vol. 30, No. 4, 2017, pp. 449-459.

[37] T. Eelbode et al. "Optimization for Medical Image Segmentation: Theory and Practice When Evaluating with Dice Score or Jaccard Index", IEEE Transactions on Medical Imaging, Vol. 39, No. 11, 2020, pp. 3679-3690.

[38] "Data Science Bowl 2017 | Kaggle", Data Science Bowl 2017 | Kaggle. https://www.kaggle.com/c/data-science-bowl-2017. (accessed: 2022)

Feature Extraction Based on ORB- AKAZE for Echocardiogram View Classification

Original Scientific Paper

1. **Shamla Beevi A.**
Department of Computer Science and Engineering
National Institute of Technology Calicut,
Kerala, India
beevishamla471@gmail.com

2. **Ratheesha S.**
SOTI India Private Limited,
Kochi, Kerala, India

3. **Saidalavi Kalady**
Department of Computer Science and Engineering
National Institute of Technology Calicut,
Kozhikode, Kerala, India

4. **Jenu James Chakola**
Department of Cardiology
Aster MIMS Hospital Kottakkal
Malappuram Kerala, India

Abstract *– In computer vision, the extraction of robust features from images to construct models that automate image recognition and classification tasks is a prominent field of research. Handcrafted feature extraction and representation techniques become critical when dealing with limited hardware resource settings, low-quality images, and larger datasets. We propose two state-of-the-art handcrafted feature extraction techniques, Oriented FAST and Rotated BRIEF (ORB) and Accelerated KAZE (AKAZE), in combination with Bag of Visual Word (BOVW), to classify standard echocardiogram views using Machine learning (ML) algorithms. These novel approaches, ORB and AKAZE, which are rotation, scale, illumination, and noise invariant methods, outperform traditional methods. The despeckling algorithm Speckle Reduction Anisotropic Diffusion (SRAD), which is based on the Partial Differential Equation (PDE), was applied to echocardiogram images before feature extraction. Support Vector Machine (SVM), decision tree, and random forest algorithms correctly classified the feature vectors obtained from the ORB with accuracy rates of 96.5%, 76%, and 97.7%, respectively. Additionally, AKAZE's SVM, decision tree, and random forest algorithms outperformed state-of-the-art techniques with accuracy rates of 97.7%, 90%, and 99%, respectively.*

Keywords*: Ultrasound, Echocardiography, SRAD, ORB, AKAZE*

1. INTRODUCTION

1. INTRODUCTION

Cardiac echocardiography produces images that assist experts in determining the function and diseases related to the human heart. Unlike magnetic resonance imaging (MRI), X-Ray, and computed tomography (CT) scans, ultrasound (US) imaging is free from radiation and is highly portable [1]. Echocardiography is an ultrasound medical imaging modality for obtaining cross-sectional views of the human heart. They support doctors in the visualization of valve failure, blood clots, changes in the velocity of blood inflow and outflow, chamber enlargement, damaged tissues, and muscles [2,3]. Doctors frequently recommend a 2D transthoracic echocardiogram (TTE), in which high-frequency US waves from a probe or transducer are placed over standard locations on the anterior chest wall to obtain different heart views. Different anatomical sections or views are obtained by adjusting or tilting the plane of waves that pass through the body. Standard views include parasternal long-axis (PLAX) view, parasternal short-axis (PSAX) views, apical 2-chamber (A2C) view, apical 4-chamber (A4C) view, apical 5-chamber (A5C) view, subcostal view, the suprasternal view that provide clear anatomy of heart [4-6]. A transesophageal echocardiogram inserts a tube through the oesophagus to create a close-up view of the heart. Complicated cases, including infants and children, are studied and analyzed with 3D echocardiography to create detailed 3D images before surgeries [7]. Echocardiography is an essential diagnostic tool in cardiology. Cardiac-related diseases like cardiomyopathies, ventricular dysfunction, coronary artery disease (CAD), congenital diseases, left ventricular hypertrophy (LVH), pulmonary hypertension, and stenosis is identified by skilled sonographers and experts by analysis of echocardiogram(echo) images [1]. However, the need for skilled experts in the interpretation of echo images has hindered the public from obtaining comprehensive benefits. We use the potentials of image processing and computer vision to address these shortcomings in echocardiogram analysis.

Deriving information from images automates various decision-making processes in medicine, industry, auto-

mobiles, surveillance, defense, and many more fields. In recent years, handcrafted features extracted from data and fed into Machine learning (ML) models have outperformed human abilities in similar tasks. Sophisticated advancements, including feature extraction and representation with deep networks, have exponentially increased the capabilities of computer vision for building novel solutions in various fields. Application of computer vision in echocardiogram images for automated view identification and disease diagnosis can considerably impact rural areas that lack human expertise and other resources. Ultrasound images are corrupted with speckle noise, which shows multiplicative and granular behavior [8]. Speckle removal in echocardiogram medical images helps easily interpret the diseased tissue [9]. Extracting relevant features such as local patches, textures, color information, and edges from medical images and categorizing these images using a machine learning model continues to bring a massive leap in health care. In this paper, we propose an automated view classification model for echocardiogram images based on feature vectors obtained from Oriented Fast and Rotated Brief (ORB) and Accelerated KAZE (AKAZE), followed by the feature representation approach bag of visual words (BOVW). Four primary views of echo images were classified using machine learning classifiers such as support vector machine (SVM), decision tree, and random Forest classifier, and their performance was assessed.

The primary contributions of this work are:

- A pipeline for classifying echocardiographic images using machine learning techniques has been developed.
- Speckle was removed from echocardiographic images using the SRAD method to improve classification accuracy.
- ORB and AKAZE were used to perform cost-effective manual feature extraction from echocardiographic images, and features represented using the BoVW method.
- Our results demonstrated superior results compared to various machine learning models for classifying echo images.

The structure of this paper is as follows: The related work in handcrafted feature extraction, feature representation, and ML models for medical image classification is discussed in section 2, along with despeckling of ultrasound images. The materials and techniques used for our study are the focus of section 3. Section 4 contains the results, and Section 5 contains the conclusion and an explanation of the future direction of our work.

2. RELATED WORKS

This study focused on the noise reduction and feature extraction of echocardiographic images. We first extracted features using the AKAZE and ORB methods, and then classified the data using different machine-learning models. This section discusses previously written literature on the previously stated modules.

2.1. DESPECKLING OF ULTRASOUND MEDICAL IMAGES

Benzarti et al. proposed an integrated method for denoising medical images using logarithmic transformation and a nonlinear diffusion tensor [9]. Speckle noise is multiplicative, and logarithmic transformation converts multiplicative noise to additive [10]. performed a comparative study on spatial and frequency domain denoising filters on ultrasonic B-mode images. They quantitatively analyzed the performance of the filters in terms of Peak Signal Noise Ratio (PSNR) value. Durte-Salazar et al. [8] explained 27 different methods that eliminate speckle noise in medical ultrasound images, which extensively covers conventional methods like spatial, diffusion, wavelet filtering, and recent techniques based on deep learning [11]. Evaluated the performance of different filtering methods like frost, mean, Kuan, median, and speckle- reducing anisotropic diffusion filter (SRAD) on liver US image data. SRAD filter showed better results on denoising medical ultrasound images from their experiments.

2.2. FEATURE EXTRACTION TECHNIQUES

It is possible to reduce image dimensions and, as a result, processing costs by selecting significant features that highlight the images' inherent content. Tareen et al. presented a comparative study on the performance of various feature extraction algorithms, SIFT, SURF, KAZE, AKAZE, ORB, and BRISK, that can be extensively applied for im-age registration [12]. Quantitative comparisons between these methods were mainly made based on several key points and corners, feature descriptors identified, and computational cost. Wei Li et al. proposed AKAZE for extracting salient features from echocardiogram videos and compared the performance with SIFT extraction technique [13]. Feature representation methods like a bag of words (BOW), sparse coding, and fisher vector (FV) are utilized to classify eight viewpoints. ORB feature matching was suggested by Rublee et al. for significant applications, including object recognition and patch-tracking on a smartphone [14]. Their study confirmed that ORB is a substitute to SIFT or SURF. Chhabra et al. developed content-based image retrieval (CBIR) system with the descriptors obtained from SIFT and ORB [15]. K-means clustering is applied to descriptors of every image to form 32 clusters, and the mean of these clusters constitutes the 32D feature vector. They also utilized locality-preserving projection (LPP) for dimensionality reduction [16]. Examined traditional approaches for extracting remarkable object recognition features such as Bag of Words, HOG-SVM (Histogram of Oriented Gradients-Support Vector Machine), and deep learning-based methods CNN and pre-trained Alexnet CNN.

2.3. FEATURE REPRESENTATION TECHNIQUES

Representing extracted features from images that are representative and discriminative is essential to develop classification models [17]. Presented an optimal correlation-based BOVW model and utilized the modest visual dictionary to implement image classification. Caleanu et al. presented a bag of features (BoF) approach to develop a histogram of visual words for binary classification of liver lesions in the CEUS dataset [18]. Tiang et al. conducted a comprehensive survey on the latest image feature extraction and representation techniques focusing on the fusion of global and local features for CBIR and automatic image annotation [19]. They also looked into generating visual-word image representations using vector-quantized region features [20]. Proposed a feature representation for microscopy image classification. A feature representation for microscopy image classification was proposed by [20]. They created feature vector (FV) descriptors from different local features and a separation-guided dimension reduction (SDR) model to transform the FV descriptors to low dimensionality [21]. Developed a sparse coding-based key point detector for low-dimension mapping of descriptors retaining complete discriminative features.

2.4. MACHINE LEARNING FOR MEDICAL IMAGE CLASSIFICATION

Image classification requires subtle features to be extracted from each image so that discriminating attributes can make efficient categorization or recognition. K. S. Jothi et al. [17] proposed a heart disease prediction model based on the Decision Tree and k-nearest neighbor (KNN) algorithms, two popular data mining algorithms, and obtained promising results in terms of accuracy. Presented eight machine learning algorithms for classifying major stroke types, ischemic and hemorrhage. u Random forest classifier performed better than other algorithms with an accuracy of 95.97%. Introduced a random forest algorithm with a correlation-based feature selection approach for early diagnosis of heart disease on the UCI heart database. Designed a classification model combining particle swarm optimization (PSO) and SVM for brain tumour prediction. Intensity, shapes, and texture-based features were derived from segmented MRI images to build a subset of relevant features. Thepage and Jadhav [19] investigated on covid-19 chest X-Ray database for automatic identification of virus infection. The feature set obtained via local binary patterns (LBP) was used to a train random tree - random forest - KNN ensemble model, which showed convincing results.

3. MATERIALS AND METHODS

This section describes the dataset, tools and libraries, and methods used in this study.

3.1. DATASETS

The dataset for this study was obtained from Aster MIMS Hospital Kottakkal in Kerala, India, with the approval of the Scientific Research Committee (SRC) and the institutional ethics committee (IEC). The dataset includes 112 echocardiogram videos collected from 56 patients (including 31 regional wall motion abnormalities and 25 normal cases). Each frame obtained is of size 600 x 800 pixels. All images were acquired using the Philips Epiq 7C cardiology US system. Echocardiogram videos are stored in digital imaging and communications in medicine (DICOM) format and sampled at a rate of 15 frames per second. Cropping was used to remove from each image information related to image acquisition, identifying information, and other information outside the image sector. These images were resized to 200 X 200 pixels. Figure 1 shows sample images from the dataset.

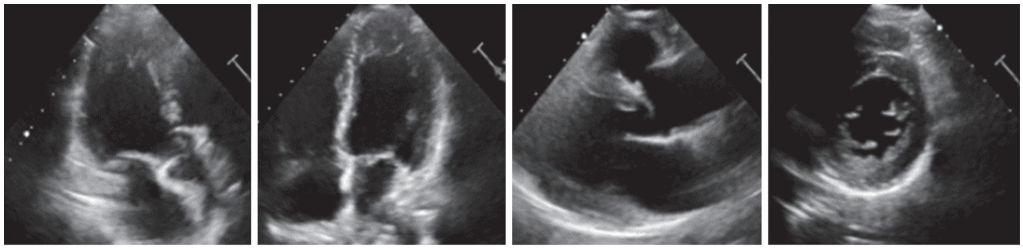


Fig. 1. Four views of echocardiogram image (a)A2C (b)A4C (c)PLAX (d)PSAX

The four categories of views are present in the dataset, namely: A2C, A4C, PLAX, and PSAX. Table 1 shows the distribution of data by views. Table 1. Distribution of data in the collected dataset.

Table 1. Distribution of data in the collected dataset

View	A2C	A4C	PLAX	PSAX	Total
Images	401	401	404	401	1604

3.2. LIBRARIES AND TOOLS

We employed the Python programming language and the Spyder integrated development environment (IDE) for our experiment. Several Python libraries were used to create the models, including numpy, pandas, matplotlib, sklearn, seaborn, and scipy. Additionally, we prepared the dataset, edited images, produced visual representations, and plotted the outcomes in Spyder.

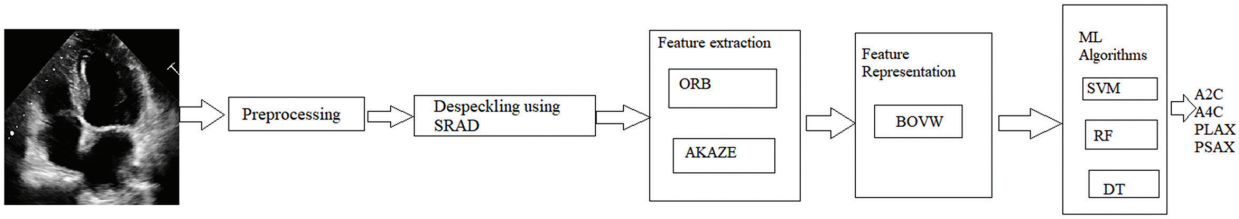


Fig. 2. Architecture of the proposed system

Figure 2 shows the pipeline architecture of our proposed system. It includes mainly four modules: Denoising echo images using speckle reducing anisotropic diffusion filter (SRAD) method, feature extraction using ORB (Oriented FAST and Rotated BRIEF) and AKAZE(Accelerated-KAZE), feature representation using a bag of words (BoW) and classification using different machine learning algorithms SVM, Decision tree and random forest classifier.

3.3. DESPECKLING USING SRAD

This study used the SRAD method to remove speckle noise from the echo image. It is a partial differentiation technique proposed by Yu and Acton in 2002 [10] to reduce speckle noise in ultrasound images. It is a useful method for maintaining edge and detail while lowering noise. The square speckle scale function is computed using this method by considering the image's mean and variance. The diffusion coefficient is calculated using the normalized discrete Laplacian and the normalized discrete gradient magnitude, as well as the gradient direction. It is presented in equation (3). Equation (4) provides the formula for the instantaneous coefficient of variation (ICOV) (4). Equation (1) and (2) contains the expression for partial derivatives.

$$\frac{\partial I(x,y,t)}{\partial t} = \mathrm{div}[c(q)\nabla I(x,y,t)] \tag{1}$$

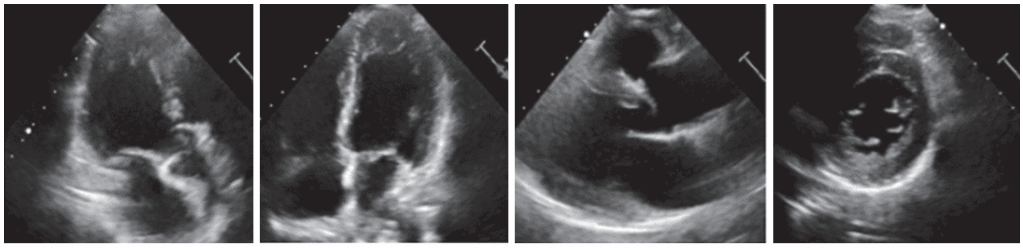
$$I(x,y,t) = I_0(x,y,t)\left.\frac{\partial I(x,y,t)}{\partial \vec{n}}\right|_{\partial\Omega} \tag{2}$$

c(q): coefficient of diffusion; ∇: gradient operator; div: divergence operator; $I_0(p, q)$: image intensity.

$$c(q) = \frac{1}{1+\frac{[q^2(x,y)-q_0{}^2(t)]}{[q_0{}^2(t)-q_0{}^2 9t]}} \tag{3}$$

q (x, y, t): instantaneous variation coefficient

$$q = \sqrt{\frac{\frac{1}{2}(\nabla I\, I)^2-\left(\frac{1}{4}\right)^2\left(\frac{\nabla^2 I}{I}\right)^2}{\left[1+\left(\frac{1}{4}\right)\left(\frac{\nabla^2 I}{I}\right)\right]^2}} \tag{4}$$



(a) A2C (b) A4C (c) PLAX (d) PSAX

Fig. 3. Denoised Images using SRAD (a)A2C (b)A4C (c)PLAX (d)PSAX

Figure 3 shows the denoised echo images after applying the SRAD method. Denoised images will help to improve the accuracy of image view classification.

3.4. FEATURE EXTRACTION USING ORB AND A-KAZE

Scale-invariant feature transform (SIFT), speeded-up robust features (SURF), features from accelerated segment test (FAST), binary robust independent elementary features (BRIEF), ORB, KAZE, and AKAZE are key point-based 2D feature detection algorithms. ORB was proposed [11]. This sophisticated method combines the BRIEF descriptor and the FAST keypoint detector and replaces SIFT and SURF, a scale and rotation invariant feature extraction method. The magnitude is ten times and 100 times faster when comparing ORB to SURF and SIFT. [12-13]. KAZE is a nonlinear diffusion filtering method based on partial differential equations (PDEs). AKAZE reduces the feature extraction complexity by fast explicit diffusion (FED). Compared to KAZE and AKAZE, the main drawback of other feature extraction techniques is their high computational expense. In our experiments, we used the ORB method to extract 500 key points, each of which had a 32-element descriptor, from the denoised echo image, and the AKAZE method discovered roughly 450 key points, each of which had a 61-element descriptor. View prediction requires extracting features that distinguish between the different echocardiogram views.

The main disadvantage of hand engineering techniques is that they are highly data-dependent.

Feature extraction plays a critical role in building superior machine learning models by avoiding the inconvenience of training and developing extensive data-driven deep networks in low-resource settings. Accurate representation of features is critical for producing accurate results.

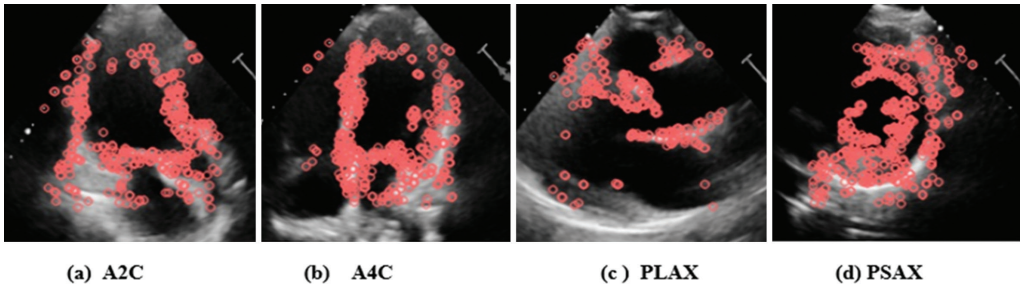
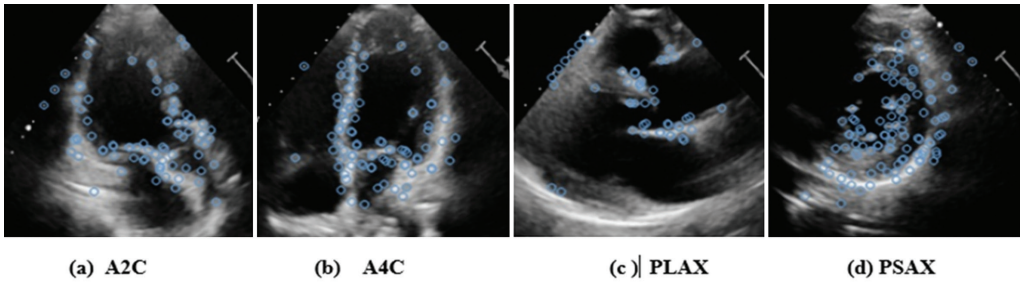


Fig. 4. Key points extracted by ORB (a)A2C (b)A4C (c)PLAX (d)PSAX



Figures 4 and 5 show features extracted using the ORB and AKAZE methods, respectively. Even though the ORB method can extract more features than the AKAZE method, the latter can extract more relevant features.

3.5. FEATURE REPRESENTATION USING BOVW

BOVW, the feature representation concept used in computer vision, is borrowed from the bag of words (BOW) method in natural language processing (NLP) [14-16]. Keypoints and descriptors used to construct a visual dictionary after clustering and frequency histogram of features define each image's feature vector. Code words in the vocabulary are the most relevant features, and the histogram shows the count of occurrence of these features in the entire image set. Figure 6 shows the histogram of a bag of visual words using ORB and AKAZE feature extraction techniques.

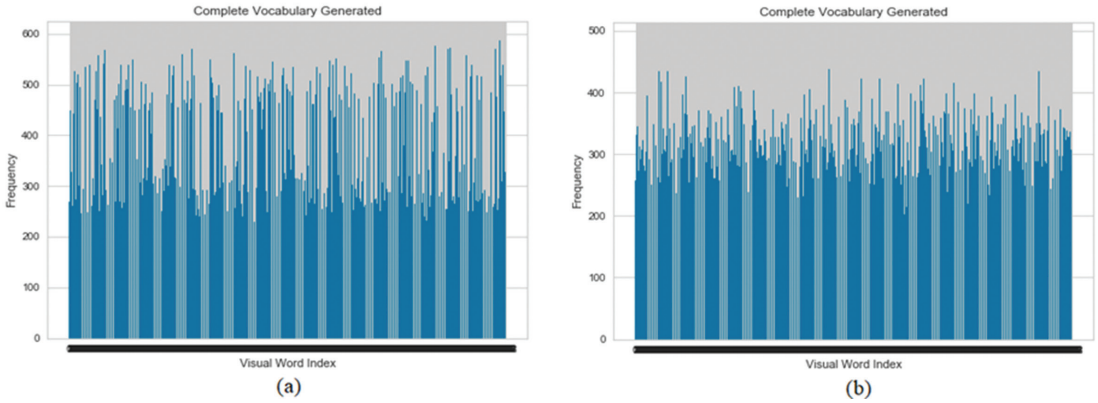


Fig. 6. Histogram of Visual words obtained from (a) ORB and (b) AKAZE

3.6. ECHOCARDIOGRAM VIEW CLASSIFICATION USING ML MODELS

Fig. 1 depicts the proposed system's architecture. Initially, the input image is preprocessed, and then the SRAD algorithm is used to eliminate speckle noise. The input image is first preprocessed, and then the SRAD algorithm is used to remove speckle noise. ORB and AKAZE were used to extract features from these denoised images. Key point descriptors from ORB and AKAZE were clustered into 1024 visual words using the k-means algorithm. We fed the feature vectors generated by ORB and AKAZE to three different ML algorithms for the classification of 4 primary echocardiogram views. Supervised models SVM, decision tree, and random forest classifier [17-20] trained using feature vectors labelled with corresponding views.

Table 2. Performance comparison between ML models

Feature extraction Method	ML Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)
ORB	SVM	96.5	96.57	97	97
ORB	Decision Tree	76	75.5	76	76
ORB	Random Forest	97.7	97.7	95	98
AKAZE	SVM	97.7	97.7	98	98
AKAZE	Decision Tree	90	89.9	90	90
AKAZE	Random Forest	99	99	99	99

Four standard views labeled on the images are A2C, A4C, PLAX, and PSAX. The models were trained over 1205 images and tested with 402 images. For both feature extraction techniques, confusion matrix and classification reports were found for each model to compare their performance.

3.7. PERFORMANCE MEASURES

Various performance assessment metrics have been applied to the performance evaluation of the classifiers. Mathematical expressions for accuracy, specificity, sensitivity, and precision are shown in equations (5), (6), (7), and (8). - These metrics can be calculated from the confusion matrix. The confusion matrix assists practitioners in determining whether the results are of high performance.

The model's accuracy refers to the total number of correct predictions over a total number of predictions. It is given by equation (5).

$$\text{Accuracy} = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{5}$$

Specificity indicates the proportion of actual false samples, which the model predicted as a false sample itself. The formula for specificity is given in equation (6).

$$\text{Specificity} = \frac{T_N}{T_N+F_P} \tag{6}$$

Sensitivity or recall tells the ratio of samples predicted true over the actual true samples. It is given by equation (7).

$$\text{Sensitivity} = \frac{T_P}{T_P+F_N} \tag{7}$$

Precision is a metric that calculates the proportion of true positives to the sum of true positives and false positives. Equation (8) provides it.

$$\text{Precision} = \frac{T_P}{T_P+F_P} \tag{8}$$

The area under the curve (AUC) of the Receiver operator characteristic (ROC) curve is a graphical representation of the performance of our machine learning classifier. The higher the AUC, the better the model performance. It depicts the trade-off between false positive rates plotted along the X-axis and true positive rates along the Y- axis.

4. RESULTS

Figures 1, 3, 4, and 5 display the four main views of an echocardiogram, denoised image using SRAD, key points detected using ORB, and key points detected using AKAZE, respectively.

Performance comparisons in terms of accuracy, specificity, sensitivity, and precision among ML models used for the experiment based on ORB and AKAZE have been presented in Table 2. AKAZE and ORB coupled with Random Forest showed excellent performance with an overall accuracy of 99 % and 97.7%.

The ROC curve for the view classification of three ML models with ORB and AKAZE feature extraction methods used for the study can see in figs 7 and 8. The ROC curve offers a graphic representation of a classifier's effectiveness. The area under the ROC curve, or AUC, provides a scalar metric that sums up the classifier's effectiveness. An improved classifier will have a higher AUC value, with 1 denoting the ideal classifier. The normalized confusion matrix exhibiting the performance of the Random Forest classifier with AKAZE is depicted in Figure 9. From the confusion matrix, we can demonstrate that our findings are sound.

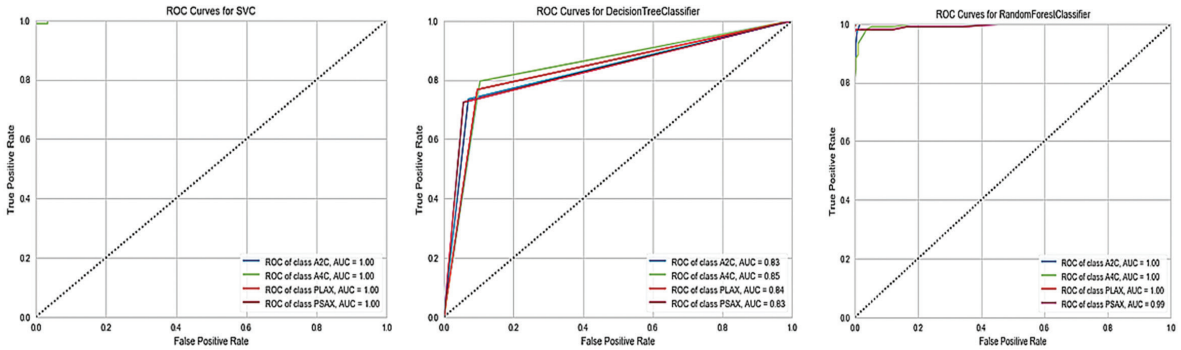


Fig. 7. ROC curve of ORB with SVM, Decision Tree and Random Forest classifier

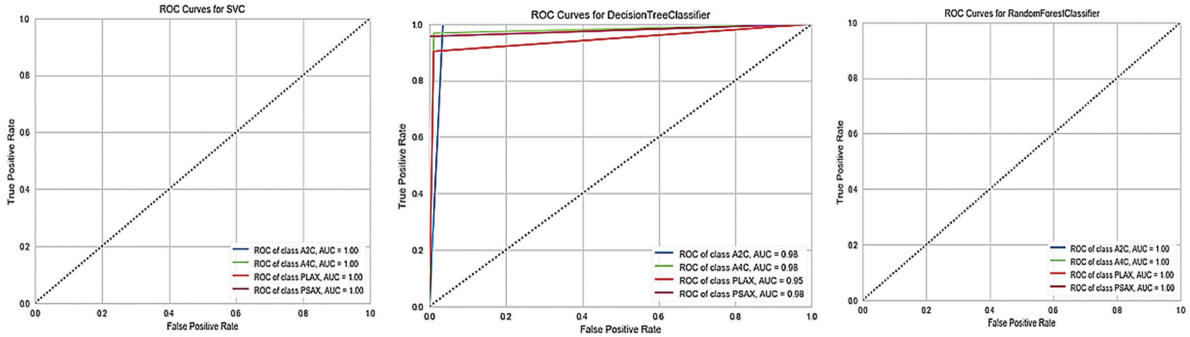


Fig. 8. ROC curve of AKAZE with SVM, Decision Tree and Random Forest classifier

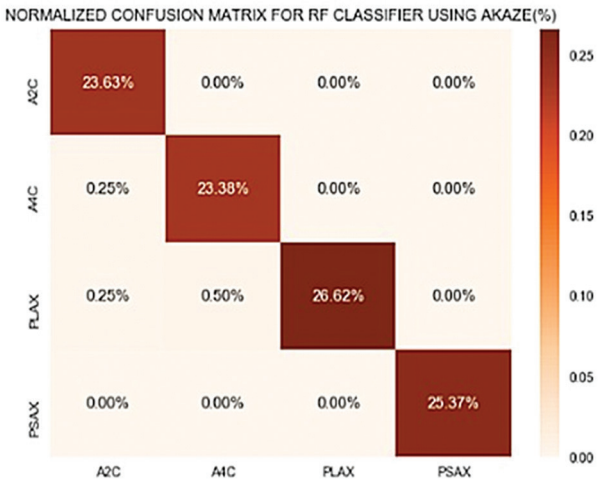


Fig. 9. Confusion matrix of Random Forest classifier using AKAZE

The confusion matrix allows a more in-depth examination of the model's behavior. Here, we only presented the confusion matrix for the model that employed the AKAZE method for feature extraction and a random forest classifier to produce the best results compared to other models. The features of echocardiographic images are incredibly complex. AKAZE employs a more complex descriptor extraction method than ORB. It results in more accurate, robust, and computationally less expensive echocardiographic image classification results.

5. CONCLUSION AND FUTURE WORK

This study combined ORB-AKAZE-based feature extraction with BOVW to classify echocardiogram images into four primary views. Automated view classification will make it easier to complete subsequent cardiac echo tasks, such as disease prediction and segmenting the region of interest. The ensembling technique Random Forest classifier outperformed both SVM and decision tree in predicting the views of echo images. Handcrafted features are typically not robust for large datasets and are computationally intensive. Future research will need to incorporate additional echocardiographic views, such as the apical five-chamber (A5C), apical three-chamber (A3C), and others, into our model. Our current research focuses on automated deep-learning feature extraction for cardiac echocardiographic images.

DATA AVAILABILITY

Due to reasonable privacy and security concerns, the datasets used in this study are not publicly available. The datasets were collected from Aster MIMS Hospital Kottakal, Kerala, India. The authors make de-identified data available upon reasonable request and with permission from the Institutional Ethics Committee (EC/13/2021 dated 25th March 2021). Experiments carried out following applicable laws and regulations. We want to extend thanks to the members of Aster MIMS (doctors, echo technicians, and staff) for their cooperation.

CODE AVAILABILITY:

https://github.com/shamlabeevia/2D-Echocardiographic-View-classification

6. REFERENCES:

[1] D. Ghorbani et al. "Deep learning interpretation of echocardiograms", NPJ Digital Medicine, Vol. 3, No. 1, 2020, pp. 1-10.

[2] F. Pereira et al. "Automated detection of coarctation of the aorta in neonates from two-dimensional echocardiograms", Journal of Medical Imaging, Vol. 4, No. 1, 2017, pp. 014502-014502.

[3] H. Moghaddasi, S. Nourian, "Automatic assessment of mitral regurgitation severity based on extensive textural features on 2d echocardiography videos", Computers in Biology and Medicine, Vol. 73, 2016, pp. 47-55.

[4] A. Madani, R. Arnaout, M. Mofrad, R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning", NPJ Digital Medicine, Vol. 1, No. 1, 2018, pp. 1-8.

[5] H. Khamis, G. Zurakhov, V. Azar, A. Raz, Z. Friedman, D. Adam, "Automatic apical view classification of echocardiograms using a discriminative learning dictionary", Medical Image Analysis, Vol. 36, 2017, pp. 15-21.

[6] H. Vaseli et al. "Designing lightweight deep learning models for echocardiography view classification", Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 10951, 2019, pp. 93-99.

[7] A. A Roest, A. De Roos, "Imaging of patients with congenital heart disease", Nature Reviews Cardiology, Vol. 9, No. 2, 2012, pp. 101-115.

[8] C. A. Duarte-Salazar, A. E. Castro-Ospina, M. A. Becerra, E. D. Trejos, "Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: an overview", IEEE Access, Vol. 8, 2020, pp.15983-15999.

[9] F. Benzarti, H. Amiri, "Speckle noise reduction in medical ultrasound images", arXiv:1305.1344, 2013.

[10] Y. Yu, S. T. Acton, "Speckle reducing anisotropic diffusion", IEEE Transactions on Image Processing, Vol. 11, No. 11, 2002, pp. 1260-1270.

[11] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "Orb: An efficient alternative to sift or surf", Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6-13 November 2011, pp. 2564-2571.

[12] P. Chhabra, N. K. Garg, M. Kumar, "Content-based image retrieval system using orb and sift features", Neural Computing and Applications Vol. 32, No. 7, 2020, pp. 2725-2733.

[13] S. B. Kibria, M. S. Hasan, "An analysis of feature extraction and classification algorithms for dangerous object detection", Proceedings of the 2nd International Conference on Electrical & Electronic Engineering, Rajshahi, Bangladesh, 27-29 December 2017, pp. 1-4.

[14] J. Jiang, D. Wu, Z. Jiang, "A correlation-based bag of a visual word for image classification", Proceedings of the IEEE 3rd Information Technology and Mechatronics Engineering Conference, Chongqing, China, 3-5 October 2017, pp. 891-894.

[15] C. D. Caleanu, G. Simion, "A bag of features approach for cause liver lesions investigation", Proceedings of the 42nd International Conference on Telecommunications and Signal Processing, Budapest, Hungary, 1-3 July 2019, pp. 323-326.

[16] D. P. Tian, "A review on image feature extraction and representation techniques", International Journal of Multimedia and Ubiquitous Engineering, Vol. 8, No. 4, 2013, pp. 385-396.

[17] K. A. Jothi, S. Subburam, V. Umadevi, K. Hemavathy, "Heart disease prediction system using machine learning", Materials Today: Proceedings, 2021.

[18] M. A. Alim, S. Habib, Y. Farooq, A. Rafay, "Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model", 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies, Sukkur, Pakistan, 29-30 January 2020, pp. 1-5.

[19] S. D. Thepade, K. Jadhav, "Covid-19 identification from chest x-ray images using local binary patterns with assorted machine learning classifiers", Proceedings of the IEEE Bombay Section Signature Conference, Mumbai, India, 2020, pp. 46-51.

[20] A. Naseri, A. Rezaei Nasab, "Automatic identification of minerals in thin sections using image processing", Journal of Ambient Intelligence and Humanized Computing, 2021, pp. 1-13.

[21] Yang, Z., Xu, Q., Bao, S., Cao, X., Huang, Q., "Learning with Multiclass AUC: theory and algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, Vol. 44, No. 11, pp. 7747-7763.

Implementation and evaluation of EMAES – A hybrid encryption algorithm for sharing multimedia files with more security and speed

Original Scientific Paper

Riddhi Somaiya
Saurashtra University, Department of computer science
Rajkot, India
riddhisomaiya@gmail.com

Atul Gonsai
Saurashtra University, Department of Computer Science
Rajkot, India
atul.gosai@gmail.com

Rashmin Tanna
Gujarat technological university, AVPTI, Electronics and Communication department
Rajkot, India
rashminstanna@gmail.com

Abstract *– In this era of smartphones, a huge amount of multimedia files like audio, video, images, animation, and plain text are shared. And with this comes the threat of data being stolen and misused. Most people don't think about the security of data before uploading it to any platform. Most apps used on smartphones upload our data to their server. Not only this, but other third-party apps can also read that data while it is being transmitted. One solution to this problem is encrypting the data before sharing it and decrypting it back at the other end so that even if it is intercepted in between the transmission, it would be impossible to decrypt it. In this paper, a newly designed hybrid encryption algorithm EMAES that includes the efficiency of MAES (Modified Advanced Encryption Standard) and security of ECC (Elliptic Curve Cryptography) was implemented in MATLAB as well as in android studio 4.0. using a mobile messaging application. Also, it was tested for different speeds and security parameters. Further, it was compared with standard algorithms like the RC4, RC6 and Blowfish as well as with other hybrid algorithms like RC4+ECC, RC6+ECC and Blowfish+ECC. The EMAES was found 30% more efficient in terms of encryption and decryption time. The security of EMAES also showed improvement when compared with other hybrid algorithms for parameters like SSIM (structural similarity index measure), SNR (Signal to Noise Ratio), PSNR(Peak Signal to Noise Ratio), MSE (Mean Squared Error) and RMSE (Root Mean Squared Error). And finally, no significant improvement was found in the CPU and RAM usage.*

Keywords*: Cryptography, AES, ECC, EMAES, MAES, RC4, RC6, hybrid encryption algorithm*

1. INTRODUCTION

The most commonly used security encryption algorithm is Rijndael, which is also known as AES (Advanced Encryption Standard) in the standardized form. It is used in the WPA2 security standard for Wi-Fi networking. In our previous research work, we modified the original algorithm and found that its efficiency improved by 68%. The implementation of this Modified AES (MAES) algorithm in MATLAB software was done in [1].

Current research is being done to make MAES more secure; for that, dual-layer security with the combination of another algorithm is proposed. MAES was extended with dual-layer security with the combination of ECC where ECC (Elliptic Curve Cryptography) is used to generate a random key every time for MAES. The hybrid algorithm was then implemented in MATLAB as well as in the Android app for comparison with other standard algorithms like RSA, ECC, AES, RSA+ECC, etc... in terms of efficiency and security. The resultant hybrid algorithm EMAES (ECC and MAES) proved to be more efficient and secure for sharing multimedia files as compared to other algorithms.

The original name of AES (Advanced Encryption Standard) is Rijndael and was selected by NIST during

the AES selection process [2]. It is the first and only algorithm that is a publicly accessible cipher approved by the National Security Agency (NSA). It is based on substitution–permutation network design principle and is efficient for both software and hardware. AES performs well on a large variety of hardware, from 8-bit smart cards to high-performance computers. It has a fixed block size of 128 and three categories of key sizes 128, 192, or 256 bits. It operates on a 4×4 column array of bytes. The key size used for an AES cipher specifies the number of transformations rounds to produce cipher from plaintext and vice versa, it moves through 4 major functions in each round i.e.

- SubBytes – based on a lookup table each byte is substituted with another by a non-linear substitution step.
- ShiftRows – last three rows of the state are cyclically shifted several times in this transposition step.
- MixColumns – it operates on the columns by combining the 4 bytes in each column of the state by a linear mixing operation.
- AddRoundKey – using bitwise xor each byte of the round key is combined with each byte of the state.

MAES (Modified Advanced Encryption Standard) is the faster version of AES. After reviewing the encryption algorithms, AES was found to be more secure and compatible with both hardware and software [3]. So, we decided to improve its efficiency as per today's requirements. Generating the same sbox and inverse sbox every time was requiring more CPU time. Also, in the mixcolumns part, a large number of multiplication processes were consuming more CPU time.

AES generates an SBOX having 256 entries and an Inverse SBOX by calculating inverse GF (28) of all the 256 entries every time it is initiated. This research eliminates all these calculations by adding a fixed SBOX as well as Inverse SBOX. At the time of the mixcolumns() procedure, AES multiplies all the substituted data which will be one element from SBOX with a poly matrix as shown in Fig. 1.

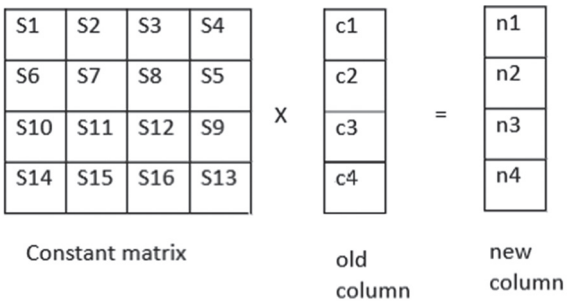


Fig. 1. Mix columns process in AES

To multiply just four bytes from data, the CPU will go through the below-given calculation:

$$SBOX'_{0,c} = (\{02\} \bullet sbox_{0,c}) + (\{03\} \bullet sbox_{1,c}) + sbox_{2,c} + sbox_{3,c}$$

$$SBOX'_{1,c} = sbox_{0,c} + (\{02\} \bullet sbox_{1,c}) + (\{03\} \bullet sbox_{2,c}) + sbox_{3,c}$$

$$SBOX'_{2,c} = sbox_{0,c} + sbox_{1,c} + (\{02\} \bullet sbox_{2,c}) + (\{03\} \bullet sbox_{3,c})$$

$$SBOX'_{3,c} = (\{03\} \bullet sbox_{0,c}) + sbox_{1,c} + sbox_{2,c} + (\{02\} \bullet sbox_{3,c})$$

MAES solves this calculation problem by taking two fixed matrices i.e., SBOX2 and SBOX3. Where in

$$SBOX2_{(R,C)} = SBOX_{(R,C)} \times 2$$

$$SBOX3_{(R,C)} = SBOX_{(R,C)} \times 3$$

This will eliminate a lot of excessive computational load on the CPU and increase the speed of operation. Results in [4] show that the percentage improvement in the encryption process is 65.386% as described in Table 1.

Table 1. Comparison of AES and MAES algorithms

Input Data Type	Execution Time		Improvement in efficiency due to Modifications
	AES (second)	Modified AES (second)	
Text (1024 bytes)	7.548	4.506	40.30%
Audio (40000 bytes)	166.633	38.575	76.99%
Image (777845 bytes)	1019.369	215.308	78.87%
Average Percentage improvement in efficiency		**65.386%**	

ECC (Elliptic Curve Cryptography) algorithm is based on the algebraic structure of elliptic curves over finite fields, public key cryptography is done. Fig. 2 shows examples of such elliptic curves. Elliptic curve-based algorithms use slightly smaller key sizes than the variants of the non-elliptic curve. The disparity in the corresponding key sizes increases significantly with rising key sizes. ECC is a public key cryptography (PKC) that has authentication keys, both public and private over finite fields which are based on elliptic curves [5].

In this paper, the EMAES i.e., ECC + Modified AES is implemented in ANDROID STUDIO 4.0 for encrypting and decrypting data in a Wi-Fi Direct chat application for smartphones. Because it is the only practical way to test the algorithm physically on the network with all its aspects. The application is tested on 5 different android phones having different configurations. Also, EMAES is compared with standard encryption algorithms like Blowfish, RC4, and RC6 in the same scenario as in [6-8]. Similarly, it is also tested and compared with the latest hybrid algorithms. Finally, we could conclude from the results that EMAES is approximately 30% more efficient (speedy), uses 25% fewer resources, and is secure as compared to another standard as well as hybrid algorithms.

1.1. RELATED WORK

An improved hybrid cryptographic framework is presented in [9] for an efficient cancellable biometric authentication system that is more secure against hackers. The main contribution of this work is the incorporation of Rubik's Cube encryption into a hybrid framework containing AES, RC6 and Chaos encryption algorithms.

Experimental simulation results confirm the promising results of the proposed Hybrid Encryption framework for efficiently encrypting stored biometric images. Therefore, it is more suitable for protecting biometric templates compared to traditional methods.

An investigation on secure communications based on hybrid encryption algorithms to improve encryption algorithms for wireless sensor networks was done in [10]. The study proposed an encryption scheme that combines the advantages of AES and ECC. This document uses hybrid encryption technology and selects the AES symmetric encryption algorithm to encrypt the data while the ECC algorithm encrypts the key and the HMAC algorithm to authenticate the message and ensure message integrity. Through simulation verification, it is found that this process significantly improves performance.

The hybrid approach described in [11] combines AES, ECC, and SHA256. Referring to existing methods, the proposed hybrid solution is similar to encrypting both text and images using the AES algorithm. The proposed method is more efficient than the previously considered methods because it is more efficient in encrypting text. The proposed method is less efficient for image encryption than the current method.

Hybrid encryption of cross-border e-commerce information is implemented in [12] through the steps of key and private key generation, key management and distribution, and key exchange in hybrid encryption. Experimental Results Compared to the existing encryption methods, the experimental results show that the hybrid encryption method developed in this paper has a longer decryption time and a reduced data error rate of 2.44 MB, resulting in higher security.

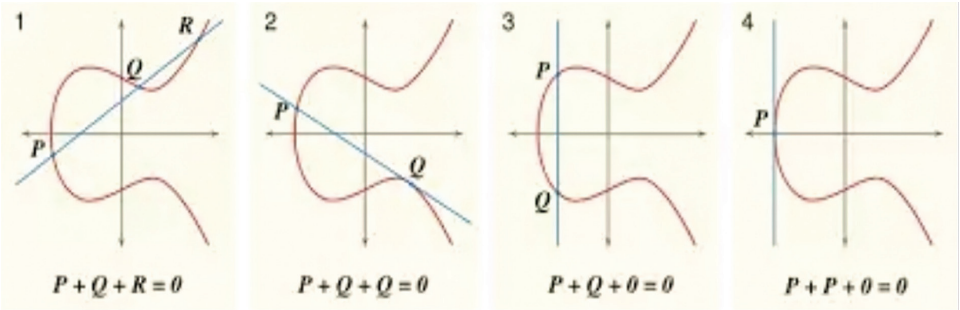


Fig. 2. Examples of Symmetric elliptic curves

In [13] a hybrid encryption method for quantum secure videoconferencing combined with blockchain, and adopt two "one-time pad" and AES quantum encryption methods to solve the problem of the low-key ratio of quantum keys was developed. A cache-efficient query method based on a B+ tree was developed, which was found to be 3.15 times more efficient than the original blockchain query.

As per the authors of [14], the hybrid ECC-AES model was found to take less time than the AES model and other existing models. Current algorithms have certain security issues, such as vulnerability to plaintext attacks, brute-force attacks, side-channel attacks, and computational complexity. The proposed algorithm was able to solve the key exchange problem experienced by AES.

The proposed HAC-based security authentication method [15] achieves a minimum communication cost of 0.017 seconds, a calculation time of 0.060 seconds, and minimum memory usage of 2.502MB, respectively. Hybrid cryptography functions in two ways. One relies on Advanced Encryption Standard (AES) and Elliptic Curve Cryptography (ECC), and the other on Rivest Shamir Adleman (RSA) and AES.

In [16] proposed the idea to use Blowfish for encryption, Message Digest 5 (MD5) for integrity, and Elliptic Curve Diffie Hellman (ECDH) for authentication. The proposed algorithm gives the best results when using two computers (A and B) compared to many other algorithms in terms of ciphertext size, encryption time, decryption time, and throughput.

AES, ECC, and Serpent were used to design an encryption scheme to secure data in an IOT-enabled system. The proposed scheme [17] improves security measures using both symmetric and asymmetric cryptosystems. A two-dimensional classification of existing studies on hybrid cryptography models based on processing phase and scope is presented in [18]. As a result, we can compare the study with other current models that help improve the performance of hybrid models after this COVID-19 pandemic.

Article [19] presents a general model of various hybrid encryption schemes that improve data security. This white paper also presents a comparative study of various traditional and hybrid models actively used for data security. The hybrid scheme can provide a higher level of security than AES and should be chosen if maximum security is required.

Security issues of information transmission and methods of hybrid encryption algorithms were described in [20]. It also considers and analyzes the different characteristics of algorithms on different systems and some common cases of hybrid cryptography, demonstrating the advantages of combining them. A hybrid encryption algorithm enhances transmission security without

causing additional problems. It also explains how, for example, cryptographic algorithms can be combined to increase security.

Lightweight hybrid cryptography techniques were explored in [21], primarily using a set of rules based entirely on AES for plaintext encryption and the Elliptic Curve Diffie-Hellman (ECDH) protocol for key encryption. The simplicity of the AES implementation makes it easy, and the complexity of ECDH makes it secure. The design is simulated in Spyder Tool, and Modelsim and implemented in Xilinx Vivado. The effect shows that the proposed lightweight model offers a normal level of security with reduced computational power. Along with the realization of a project to implement multimedia input on his FPGA, a key authentication system for enhanced security was proposed.

2. IMPLEMENTATION ENVIRONMENT

2.1.1. ANDROID

Android is a mobile operating system designed for smart devices such as smartphones, smart tv, smartwatch, etc. It is developed by Open Handset Alliance and commercially sponsored by Google. It is an open-source and free [22] operating system. It has been the best-selling operating system in the world since 2011. As of March 2020, the app store i.e., Google play store features more than 2.9 million applications.

2.1.2. ANDROID STUDIO

Designed specifically for Android development and built on JetBrains' IntelliJ IDEA software, Android Studio is the Android operating system's official IDE (Integrated development environment). It supports developers to design, code, test and launch the application easily and fast. It contains various tools for learning android applications, designing the user interface, coding, compiling and debugging environments along with various testing features.

Developers can create virtual android devices to test the application. Android Studio also supports the installation of the application on real android devices and the logging of performance statistics.

2.1.3. ANDROID APPLICATION

A software application developed to run on android supported devices. It is distributed as a .apk file that contains all the resources of that application. Android apps could be coded in various languages such as java, c++, kotlin, etc., using an android software development kit and JVM i.e., Java Machine. The official development environment is called Android Studio.

2.1.4. WIFI DIRECT

Wi-Fi CERTIFIED Wi-Fi Direct® is a Wi-Fi connection without the requirement of a wireless router or an internet connection. Like Bluetooth, Wi-Fi is a way of communicating wirelessly. The concept of "ad-hoc" Wi-Fi mode has similarities with Wi-Fi Direct [23]. However, Wi-Fi Direct has an easier way to automatically discover and connect to nearby devices like cameras, Mobile phones, PCs, printers and gaming devices compared to an ad-hoc Wi-Fi connection. Using the latest Wi-Fi security i.e., Wi-Fi Protected Setup™ supported devices, one can make a point-to-point connection or a group of several devices can connect simultaneously and exchange files, play media, print documents or display screens between them [24].

3. PROPOSED WORK

3.1.1. EMAES

The EMAES is an improved version of MAES where MAES improves the efficiency of AES and EMAES provides better security to it. AES is almost impossible to crack without the knowledge of its KEY. As we used MAES in a mobile chat application, it was necessary to share the key with the receiver so that the data could be decrypted. When we share the key wirelessly, it becomes vulnerable to attacks from third-party intruders. i.e., hackers/crackers who can read the key and then decrypt the data easily. As a solution to this, we used the ECC algorithm that helped to generate random private and public keys. Here, both devices (same application in two different smartphones) share their public key and create a shared key with the help of their private key and others' public key. Now this shared key is used as KEY to encrypt and decrypt the data in MAES.

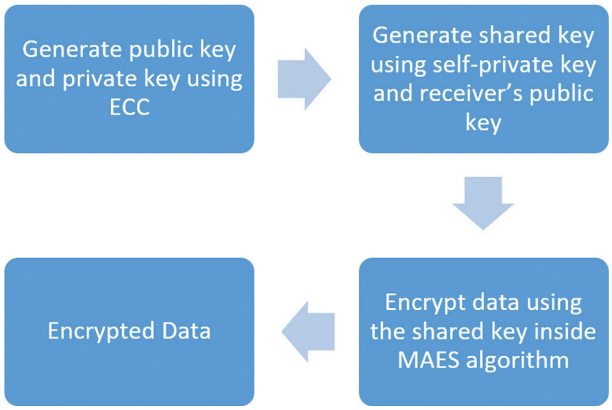


Fig. 3. EMAES algorithm Encryption process

The EMAES algorithm encryption process is described in Fig. 3. At first, 256 bits of public key and private key will be generated. The sender and the receiver will have to exchange their public keys to share the data. A shared key will be generated using its own private key and the receiver`s public key. This shared key will be used as the KEY in the MAES algorithm to encrypt the data.

The decryption process of the EMAES algorithm will be the same as the Encryption process. As shown in

Fig. 4, the decryptor end will generate a public key and a private key. Also, a shared key using a self-private key and the sender's public key will be generated. Then the shared key will be used in the MAES algorithm as a secure key and the encrypted data will get decrypted.

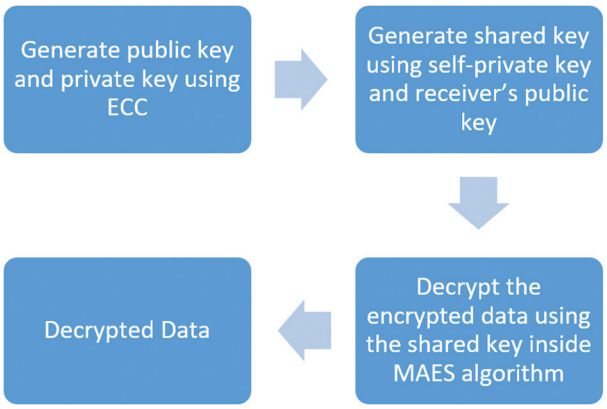


Fig. 4. EMAES algorithm decryption process

Here, the shared key in ECC can be generated in any size i.e., 128, 198, or 256 bits according to the requirement of MAES.

This way EMAES algorithm gives us all the advantages of AES along with improved speed through MAES and better security through ECC. We get dual layer security where data and key both are hidden from third parties.

3.1.2. EMAES chat app

EMAES chat app is a testing tool in the form of a chat application on the android platform. One can select any encryption algorithm provided in it and test them all one by one. This application requires Wi-Fi and location service to be kept ON. The user can connect to another phone and then test the selected encryption algorithm by sending and receiving multimedia messages. The messages sent and received through the application are encrypted as well as decrypted using the selected encryption algorithm. The test parameters recorded in the google firebase real-time database are later used for the comparison of algorithms.

3.1.3. Application Flow

Fig. 5 shows the connection screen of the app encryption algorithm for testing purposes. Currently available algorithms are EMAES (proposed in this article), MAES, AES, AES+ECC, BLOWFISH, BLOWFISH+ECC, RC4, RC4+ECC, RC6, RC6+ECC. The option of NO ENCRYPTION is also available which sends data without encryption. On the next screen, the user can connect with another android phone with the EMAESChat app installed and opened.

3.1.4. Chat room screen

After starting the chat room, the application instantly generates the private and public keys and sends them to the other user only if one of the algorithms with ECC is selected. Similarly, the app receives the public key of the opposite user and generates the shared key. In other cases, a simple 32-bit constant key is selected as the public key. Users can send text messages as well as multimedia files like drawings, images, audio, video, and other file formats on the chatroom screen. A screenshot of the chat room screen is shown in Fig. 6. When the user presses the send button, the application converts the data selected to be sent in string format and encrypts that string with the selected encryption algorithm. At the receiver end, the application receives the encrypted message and decrypts it with the selected algorithm using a generated shared secret key. Its decryption results in a string that is converted into the exact original message sent by the sender. The same process is repeated every time the user sends and receives the messages. The shared key will expire when the user exits the chat room if it is previously generated by the system.

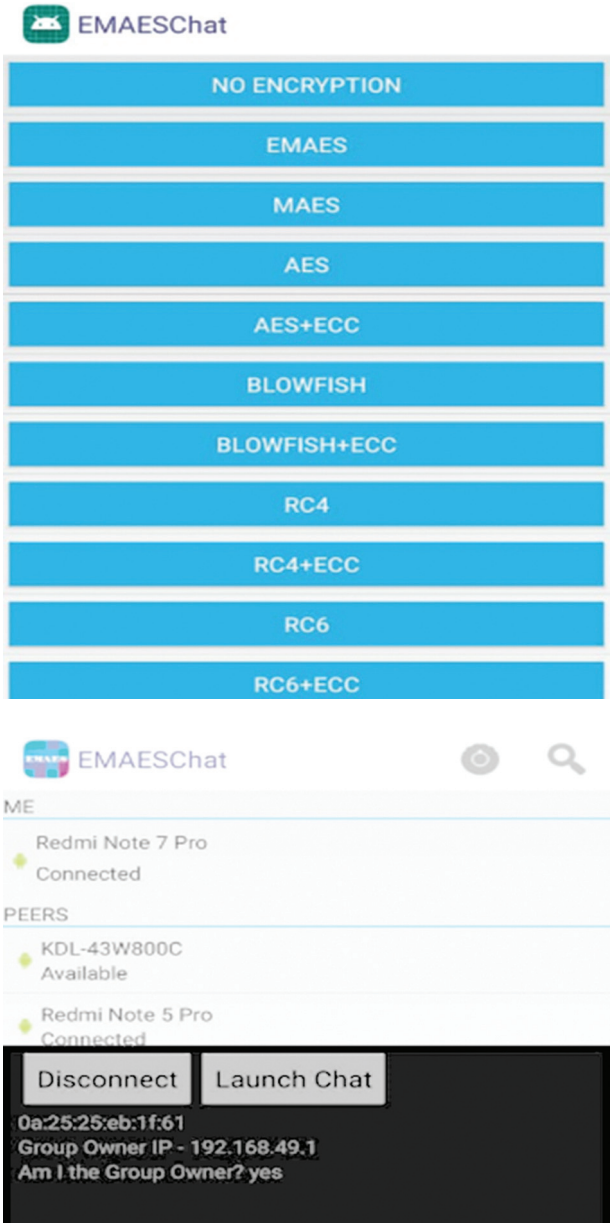


Fig. 5. Screenshot of connection screen

Fig. 6. Screenshot of chatroom screen sending and receiving messages

4. RESULTS

In Table 2 multimedia files like text, image, audio, and video files are used to encrypt and decrypt with EMAES. This table shows the size of each file and the time taken to encrypt and decrypt the data. The original file is compared with the decrypted file to search for errors or noise using the parameters like SSIM (structural similarity index measure), SNR (Signal to Noise Ratio), PSNR(Peak Signal to Noise Ratio), MSE (Mean Squared Error) and RMSE (Root Mean Squared Error). The data after decryption was found exactly the same as the original. Also, table 2 shows the utilization of resources like CPU, RAM, and NETWORK while encrypting, sending, receiving and decrypting the data, similar to the work done in [25]. Here, 56kb of text took 0.049 seconds to encrypt and 0.035 seconds to decrypt with 0 noise and errors. Same way, 1.82 MB of audio took 1.1 seconds to encrypt and 0.52 seconds to decrypt. Many other file types like .wav, .pdf, .doc, etc., were also tested successfully during the research.

Table 2. Proposed work implemented on multi-Media files

PARAMETERS	TEXT	AUDIO	IMAGE	VIDEO
Size	52 kb	1.82MB	4MB	27MB
Encryption Time (sec)	0.049	1.1	36.40	93.16
Decryption Time (sec)	0.035	0.52	17.99	54.21
SSIM	1	1	1	1
PSNR	INF	INF	INF	INF
SNR	0	0	0	0
MSE	0	0	0	0
RMSE	0	0	0	0
CPU (%)	12	24	29	37
RAM (MB)	121	131	261	315
NETWORK (bps)	39.23	49.90	59.10	65.32

Table 3 depicts the comparison of EMAES with standard algorithms like Blowfish, RC4, and RC6. All the algorithms were implemented in the EMAESChat app and compared based on various parameters while sending and receiving messages. 10kb, 100kb, and 1 MB data were considered for small, medium, and large sizes respectively. The parameters considered for comparison were Encryption time, Decryption time, SSIM, SNR, and MSE. The table also shows the utilization of resources like CPU, RAM, and NETWORK while sending and receiving multimedia files. A comparison of EMAES with hybrid algorithms like Blowfish+ECC, RC4+ECC, and RC6+ECC is depicted in table 4. The comparison parameters used were the same as that in table 3.

A comparison of EMAES with both standard and hybrid algorithms shows that EMAES provides better security as it is completely based on AES. Also, it has the advantages of ECC. Moreover, its strength can be seen by comparing SSIM (Structure Similarity Index), SNR (Signal to Noise Ratio), and MSE(Mean Squared Error) values, as they are more reliable than that of any other algorithms.

A chart based on table 3 is shown in Fig. 7. This will allow us to visualize the numerical differences between EMAES and other standard algorithms with the parameters considered for testing. It could be seen from the chart that the average execution time (encryption and decryption) of EMAES is 0.90 sec which is the fastest. The second fastest is Blowfish with 1.29 sec. Hence, we can say that EMAES is at least 30% faster than all the standard algorithms. Also, the resource utilization is 25% less than other algorithms. But for this security is not compromised.

Fig. 8. represents a chart based on Table 4. With the help of this, comparison parameter values used to compare the EMAES with other hybrid algorithms could be analyzed. It was seen that the average execution time of EMAES was 0.90 sec and that of the fastest hybrid algorithm i.e., Blowfish+ECC was 0.98 sec which is still 8.1% faster. Also, resource utilization and security were not compromised.

Table 3. Comparison of EMAES with standard algorithms

Comparison Parameters/ Algorithms	Data size (Small / Medium / Large)	Encryption Time (seconds)	Decryption Time (seconds)	SSIM (Structured Similarity Index)	SNR (Signal-to-Noise Ratio)	MSE (Mean Squared Error)	CPU usage (%)	RAM usage (kb)	Network throughput (bps)
EMAES	S	0.05	0.5	0.9	0	0	12	121	39.23
	M	0.51	0.74	0.9	0	11	24	131	49.9
	L	1.78	1.86	0.9	-1.3	5.3	29	261	59.65
Blowfish	S	0.07	0.07	1	0	0	12	122	10.1
	M	0.95	0.71	0.96	65	78	34	101	35.6
	L	1.13	1.1	0.99	-1.8	4.2	21	168	65.3
RC4	S	0.03	0.04	1	0	0	18	141	23.1
	M	0.31	0.38	0.96	0	70	22	137	61.1
	L	3.32	2.7	0.99	0	2.4	29	161	78.8
RC6	S	0.59	0.91	0.99	0	2.2	16	204	20.6
	M	0.58	0.59	0.9	0	121	19	246	22.3
	L	2.64	1.95	0.99	-2.2	4.8	21	303	32.8

Table 4. Comparison of EMAES with hybrid algorithms

Comparison Parameters / Algorithms	Data size (Small / Medium / Large)	Encryptio Time (seconds)	Decryption Time (seconds)	SSIM (Structured Similarity Index)	SNR (Signal-to-Noise Ratio)	MSE (Mean Squared Error)	CPU usage (%)	RAM usage (kb)	Network throughput (bps)
EMAES	S	0.05	0.5	0.9	0	0	12	121	39.23
	M	0.51	0.74	0.9	0	11	24	131	49.9
	L	1.78	1.86	0.9	-1.3	5.3	29	261	59.65
Blowfish+ECC	S	0.45	0.07	1	0	0	12	122	10.9
	M	0.75	0.71	0.96	65	78	34	123	36.1
	L	1.93	1.97	0.99	-1.8	4.2	21	231	67.3
RC4+ECC	S	0.04	0.04	1	0	0	18	146	26.1
	M	0.32	0.38	0.96	0	70	22	154	61.9
	L	3.21	3.1	0.99	0	2.4	29	211	79.4
RC6+ECC	S	0.6	0.98	0.99	0	2.2	16	211	23.2
	M	0.72	0.66	0.9	0	121	19	255	25.3
	L	2.89	2.1	0.99	-2.2	4.8	21	335	34.8

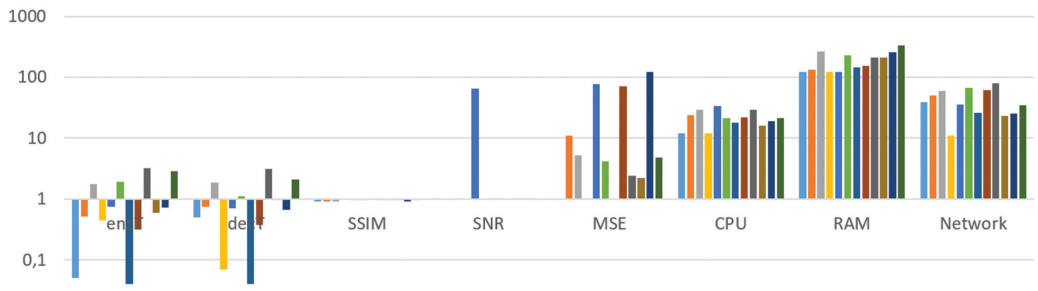


Fig. 7. Visual representation of comparison parameters of EMAES and other standard algorithms in logarithmic scale to the base 10

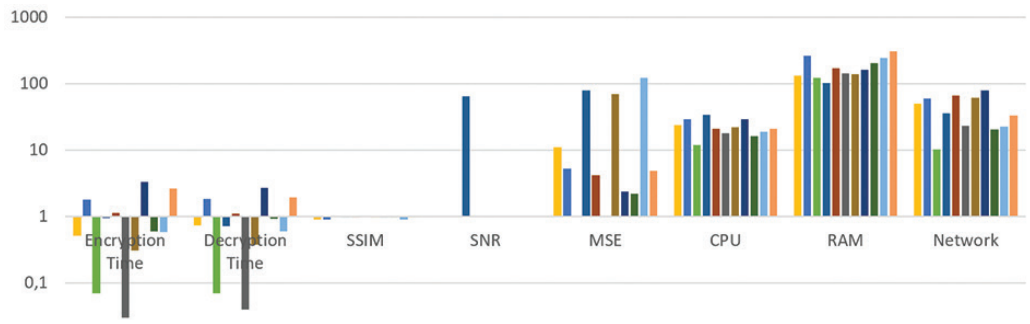


Fig. 8. Visual representation of comparison parameters of EMAES and other hybrid algorithms in logarithmic scale to the base 10

5. CONCLUSION

This research proposes a new hybrid algorithm that gives multiple benefits of speed, accuracy, and security with minimum resource utilization.

The algorithm was previously designed as MAES i.e., modified AES to increase efficiency. But AES is secure enough only until its key is hidden. When we need to share the data with another device, we also need to share the key to decrypt the data. This was a big risk to data security. Then the combination of MAES with ECC improved the security by making it a dual layer. Here, AES is well known for data security while ECC gives a strong public key technique that is next to impossible to hack.

The algorithm was previously tested in MATLAB to check its security and efficiency against other encryption algorithms, the positive outcomes encouraged us to test it in a live environment. Hence the EMAES was implemented in an android chat application i.e., the EMAESChat app which works on Wi-Fi-direct and was tested in multiple android smartphones sharing different kinds of multimedia files.

Finally, by comparing the results with standard as well as hybrid algorithms, it could be concluded that EMAES is on average at least 30% faster than the fastest algorithm i.e., blowfish, and 8% faster in execution time when compared with the fastest hybrid algorithm i.e., blowfish+ECC. In terms of resource utilization like CPU, RAM, and network also EMAES is at least 25% better. The most significant feature of EMAES is that it does not compromise security while achieving efficiency.

Future work could include the implementation of EMAES on FPGA and test with a large number of devices on the Internet. It could also be tested in cloud computing and IoT environments. There is also the scope for comparing it with other live streaming and video calling algorithms.

6. REFERENCES

[1] R. Tanna, R. Tanna, J. Bhadeshiya, "Improvement in the execution time of AES algorithm by modifications in sbox and mix columns for multimedia applications", International Journal of Research in Engineering, IT and Social Sciences, Vol. 8, No. 4, 2018, pp. 65-68.

[2] J. Nechvatal, E. Barker, L. Bassham, W. Burr, M. Dworkin, J. Foti, E. Roback, "Report on the Development of the Advanced Encryption Standard (AES)", Journal of research of the National Institute of Standards and Technology, Vol. 106, No. 3, 2001, pp. 511–577.

[3] R. C. Somaiya, A. M. Gonsai, R. S. Tanna, "WLAN security and efficiency issues based on encryption techniques", International Journal of Research in Engineering, IT and Social Sciences, Vol. 6, No. 9, 2016, pp. 27-32.

[4] R. Somaiya, A. Gonsai, R. Tanna, "Design and Implementation of a New Encryption Algorithm in MATLAB for Multimedia Files", Vidhyayana, Vol. 6, No. 6, 2021, pp. 985-999.

[5] X. Li, J. Chen, D. Qin, W. Wan, "Research and realization based on hybrid encryption algorithm of improved AES and ECC", Proceedings of the International Conference on Audio, Language and Image Processing, Shanghai, China, 23-25 November 2010, pp. 396-400.

[6] H. K. Verma, R. K. Singh, "Enhancement of RC6 block cipher algorithm and comparison with RC5 & RC6," Proceedings of 3rd IEEE International Advance Computing Conference, Ghaziabad, India, 22-23 Feb. 2013, pp. 556-561.

[7] B. Schneier, "Description of a new variable-length key, 64-bit block cipher (Blowfish)", International workshop on fast software encryption, Springer, Vol. 809, No. 6, 1994, pp. 191-204.

[8] R. Singh, "An overview of android operating system and its security features", International Journal of Engineering Research and Applications, Vol. 4, No. 2, 2014, pp. 519-521.

[9] M. Helmy et al. "A hybrid encryption framework based on Rubik's cube for cancelable biometric cyber security applications", Optik, Vol. 258, No. 168773, 2022, pp. 30-42.

[10] Y. Huiwei, "Application of hybrid encryption algorithm in hardware encryption interface card", Security and communication networks, Vol. 2022, No. 7794209, 2022, pp. 1-11.

[11] P. William, A. Choubey, G. S. Chhabra, R. Bhattacharya, K. Vengatesan, S. Choubey, "Assessment of Hybrid Cryptographic Algorithm for Secure Sharing of Textual and Pictorial Content", Proceedings of the International Conference on Electronics and Renewable Systems, Tuticorin, India, 16-18 March 2022, pp. 918-922.

[12] J. Li, S. Wang, Z. Zhang, Y. Xu, "Research on Hybrid Encryption of Cross Border E-commerce Transaction Information Based on B+ Search Tree Algo-

rithm", Proceedings of the 2nd EAI International Conference on IoT and Big Data Technologies for Health Care, Leicester, UK, 18-19 October 2021, pp. 290-307.

[13] D. Zhu, J. Zheng, H. Zhou, J. Wu, N. Li, L. Song, "A Hybrid Encryption Scheme for Quantum Secure Video Conferencing Combined with Blockchain", Mathematics, Vol. 10, No. 17, 2022, pp. 30-37.

[14] S. Ahmad, S. Mehfuz, J. Beg, "Hybrid cryptographic approach to enhance the mode of key management system in cloud environment", Journal of Supercomputing, Vol. 78, No. 17, 2022, pp.1-37.

[15] S. P. Kavitha, I. Mandal, C. Rangaswamy, "Hybrid and Adaptive Cryptographic-based secure authentication approach in IoT based applications using hybrid encryption", Pervasive and Mobile Computing, Vol. 82, No. 101552, 2022, pp. 8-25.

[16] H. A. Abdulhameed, A. A. Abdulhameed, M. F. Mosleh, A. T. Mohammad, "Lightweight security protocol for WSNs using hybrid cryptography algorithm", AIP Conference Proceedings, Vol. 2547, No. 1, 2022, p. 060006.

[17] S. Das, S. Namasudra, "A Novel Hybrid Encryption Method to Secure Healthcare Data in IoT-Enabled Healthcare Infrastructure", Computers & Electrical Engineering, Vol. 101, No. 107991, 2022, pp 107-115.

[18] A. A.-R. El-Douh, S. F. Lu, A. Elkony, A. S. Amein., "A Systematic Literature Review: The Taxonomy of Hybrid Cryptography Models", Lecture Notes in Networks and Systems, Springer, Vol. 439, 2022, pp. 714-21.

[19] P. Soni, R. Malik., "A Comparative Study of Various Traditional and Hybrid Cryptography Algorithm Models for Data Security", Modeling, Simulation, and Optimization, Springer Nature, Vol. 292, No.28, 2022, pp. 31-47.

[20] Q. Zhang, "An Overview and Analysis of Hybrid Encryption: The Combination of Symmetric Encryption and Asymmetric Encryption", Proceedings of the 2nd International Conference on Computing and Data Science, Stanford, CA, USA, 28-29 January 2021, pp. 616-622.

[21] P. Kitsos, G. Kostopoulos, N. Sklavos O. Koufopavlou, "Hardware implementation of the RC4 stream cipher", Proceedings of the 46th Midwest Symposium on Circuits and Systems, Cairo, Egypt, 27-30 Dec. 2003, pp. 1363-1366.

[22] D. Camps-Mur, A. Garcia-Saavedra, P. Serrano, "Device-to-device communications with Wi-Fi Direct: overview and experimentation", IEEE Wireless Communications, Vol. 20, No. 3, 2013, pp. 96-104.

[23] T. Alam, M. Aljohani, "An approach to secure communication in mobile ad-hoc networks of Android devices", Proceedings of International Conference on Intelligent Informatics and Biomedical Sciences, Okinawa, Japan, 28-30 November 2015, pp. 371-375.

[24] S. K. Ghosh, S. Rana, A. Pansari, J. Hazra, S. Biswas, "Hybrid Cryptography Algorithm for Secure and Low-Cost Communication", Proceedings of International Conference on Computer Science, Engineering and Applications, Gunupur, India, 13-14 March 2020, pp. 1-5.

[25] A. Mammenp, S. KN, R. Bhakthavatchalu, "Implementation of Efficient Hybrid Encryption Technique", Proceedings of the 2nd International Conference on Intelligent Technologies, Hubli, India, 24-26 June 2022, pp. 1-4.

Deep learning approach and topic modelling for forecasting tourist arrivals

Original Scientific Paper

Houria Laaroussi
Mohammed V University in Rabat,
Mohammadia School of Engineering, EST SALE, Department of Engineering Science
RABAT, Morocco
hourialaaroussi@research.emi.ac.ma

Fatima Guerouate
Mohammed V University in Rabat,
Mohammadia School of Engineering, EST SALE, Department of Engineering Science
RABAT, Morocco
fatima.guerouate@est.um5.ac.ma

Mohamed Sbihi
Mohammed V University in Rabat,
Mohammadia School of Engineering, EST SALE, Department of Engineering Science
RABAT, Morocco
mohamed.sbihi@yahoo.fr

Abstract *– Online review data attracts the attention of researchers and practitioners in various fields, but its application in tourism is still limited. The social media data can finely reflect tourist arrivals forecasting. Accurate prediction of tourist arrivals is essential for tourism decision-makers. Although current studies have exploited deep learning and internet data (especially search engine data) to anticipate tourism demand more precisely, few have examined the viability of using social media data and deep learning algorithms to predict tourism demand. This study aims to find the key topics extracted from online reviews and integrate them into the deep learning model to forecast tourism demand. We present a novel forecasting model based on TripAdvisor reviews. Latent topics and their associated keywords are captured from reviews through Latent Dirichlet Allocation (LDA), These generated features are then employed as an additional feature into the deep learning (DL) algorithm to forecast the monthly tourist arrivals to Hong Kong from USA. We used machine learning models, artificial neural networks (ANNs), support vector regression (SVR), and random forest (RF) as benchmark models. The empirical results show that the proposed forecasting model is more accurate than other models, which rely only on historical data. Furthermore, our findings indicate that integration of the topics extracted from social media reviews can enhance the prediction.*

Keywords*: Online review data, Tourist arrivals forecasting , Latent Dirichlet Allocation, Deep Learning*

1. INTRODUCTION

The tourism sector has expanded greatly over the last several years. Tourism demand forecasting has become a striking issue in the area of prediction research because of the important economic impact of the rapid-growing tourism industry [1]. Forecasting is vital to the tourism planning process. Because of the perishable nature of tourism [2]. Tourist arrivals forecasting provides valuable information to decision makers in order to make crucial decisions and planning [3]. Accurate forecasts can serve to allocate resources, and contribute to reducing the risks of decision failures and the costs of attracting [4].

Overall, tourism demand forecasting techniques can be categorized into three classes: time series models, econometric models, and artificial learning techniques [5,6]. Econometric and time series models fail to learn the nonlinear dependencies in the data. Machine learning techniques usually need to have manual features [7]. Deep learning (DL) is a neural network with many layers. DL methodologies have gained the interest of scholars due to their successful applications. Many hidden layers can often capture the non-linear characteristics of data, it can handle the nonlinearity that exists in tourism demand data and it aims to build important features automatically [8].

In earlier studies, the most crucial data were historical data and search engine data [8,9]. On the other hand, social media data can make a better contribution to predicting tourism demand. Therefore, Social media data can supplement traditional data. Furthermore, search engine data can enhance the prediction performance of tourism demand [9], but it is not easy for users to choose the most relevant information in search engines with a such large number of information available over search engines and it cannot completely describe tourists' preferences. Then, to increase forecasting accuracy, incomplete data must be reinforced with relevant information. Social media can be used as a solution to address this problem.

The latest approaches for forecasting tourism demand are seen to have two significant limitations. First, the majority of studies employed deep learning methods that perform well in tourism demand prediction, but they didn't take into account the influence of social media data [5,8,10,11]. In the deep learning techniques, they only utilized the motor engine data or traditional data. For instance, [8] employed a deep learning approach and search engine data to predict the monthly Macau tourist arrival volumes. [10] used a deep learning model called the Bayesian Bidirectional Long Short-Term Memory (BBiLSTM) network in conjunction with past tourist volumes, tourism prices, and income to forecast quarterly tourist arrivals to Singapore. [11] used LSTM networks with multisource time series data to predict daily tourism demand. [5] Incorporate multivariate time series data into a deep learning model to forecast the daily tourism volume. This study tries to incorporate social media data into the deep learning model.

Second, the studies that take social media into account focus on quantitative data. For example, [12] utilized likes on DMO Facebook pages as a novel explanatory variable of tourism demand. [13] Employed the average review rating and the volume of the review as indicators from online review data to forecast tourist arrivals. [14] Used review volume and ratings collected from internet review sources to forecast tourism demand. .There are few studies that use qualitative data in the prediction of tourism demand. For example, [15] adopted news topics to forecasting tourism demand. In this paper, we seek to fill this gap. We discuss the impact of textual data as qualitative data that can be obtained from social media to enhance the prediction of tourism demand. We use Topic modeling to discover coherent and interesting topics of reviews extracted from social media.

The purpose of this paper is to develop a new model using social media data and strong features for tourism demand prediction. In particular, we predict tourist arrivals to Hong Kong from USA in the period from January 2013 to January 2020. First, we collected the reviews shared on Trip Advisor. Then, we extract the important topics and the key keywords from posts using a topic modeling technique, latent Dirichlet allocation (LDA), which focuses on identifying topics within a collection of reviews. It can present the distribution of keywords by topic and the distribution of topics by review. All the keywords are used as new features in the prediction model. Furthermore, we predict tourist arrivals in Hon Kong from USA based on the long short-term memory (LSTM) algorithm. To do this, text mining frequency inverse document frequency (TF-IDF) is applied for vector representation. TF-IDF converts each keyword into numeric vectors to construct an LSTM model to predict tourism demand. In addition, an empirical analysis was performed to measure the accuracy of our model. Random forest, support vector regression, and artificial neural network are used for experiments.

This article will present the related work in Section 2. We will describe the proposed methods in section 3. Section 4 will provide the results and discussion. In Section 5, we will present the conclusions of the proposed study and future work.

2. RELATED WORK

2.1. TOURISM DEMAND FORECASTING MODELS

Several researchers have used Time series, econometric, and AI methods for predicting tourism demand [16]. Time series methods include naïve, moving average, exponential smoothing, and BoxeJenkins models. The autoregressive moving average (ARIMA) and Seasonal ARIMA are widely employed models and give better performance [17]. Time series models are limited by their incapacity to forecast fluctuations that are not used the past observations [18].

Causal econometric models seek to analyze the causal relationships between tourist arrivals and their determinants (such as Exchange rates, income; relative prices, and expenditure [19]. The insufficient information on the causal structure is the main limitation of econometric models [20]. Popular econometric forecasting models comprise the autoregressive distributed lag model [19], the error correction model [21], the vector autoregressive model [22], and the time-varying parameter model [23]. The most important limitations of econometric models are the existence of multicollinearity among the independent variables, difficulties in the data collection [24] and the model may depend on some predictor variables which are not available at the moment of prediction. They are limited also by the large amount of time and substantial skills required to establish correct relationships. Time series forecasting models and econometric models are often incapable of simulating complicated nonlinear properties of the destination demands [25] when non-linearity and noise exist in tourism demand data.

Artificial Intelligence (AI) forecasting methods, including Artificial Neural Networks, rough sets theory, fuzzy

time series method, support vector regression, grey theory, and modern deep learning. AI models have higher adaptability and can explore non-linear relationships. Artificial NNs (ANNs) are one of the most popular nonlinear modeling methods used in tourism demand studies [26]. For example, Kon and Turner [26] showed that ANNs often had best accurate than time-series models. Cho [27] examined three time-series methods and the ANN model for predicting visitor arrivals. It was found that ANN is more accurate than other models.

Overfitting, slow convergence, and unpredictable solutions during training, NN redundancy, and getting stuck in local minima are the most drawbacks of ANNs. Deep learning is a neural network with many layers. Deep learning can explore more complex non-linear patterns in the data, can handle complex data with various structures, prevents over fitting problems for a large number of inputs, and can Improves model performance. In the last few years, a few publications used the deep learning to model tourism demand [28]. For example, Law et al. [8] compared the forecasting performances of LSTM and other methods (naive method, SVR, ANN), to predict Monthly Macau tourist arrival volumes. The experiments demonstrated that the DL technique (LSTM) performs better than other models. Li and Cao [29] forecast tourism flow based on the LSTM technique. The results showed that the LSTM technique is more accurate than the benchmark models.

2.2. TOURISM DEMAND FORECASTING WITH ONLINE DATA

A different kind of data has been used for tourism demand predicting. Based on the data type, previous literature related to tourism demand can be decomposed into two categories: traditional data and online data. Traditional data is statistical data. The majority of statistical data are extracted daily [30], monthly [31], quarterly [32], and annually [33] from tourism industry organizations; considering the major lack of data about tourist arrivals, it is difficult to apply these data to forecasting tourism demand. Moreover, the data size is limited [34]. Due to these disadvantages, the application of traditional data in tourist demand studies is limited.

Several studies have considered internet data as a complement to traditional data and as a new feature; hence, they employed big data to understand tourist satisfaction [35]. One of the major advantages of online data provided by search engines and social media is that it is real-time [36]. Two kinds of online data have emerged in the tourist arrivals prediction literature: search engine data and social media data [13]. Several studies have focused on the impact of search engine data on tourism demand forecasting performance [37]. For example, Law et al. [8] used search engine data as features in the LSTM method to forecast the monthly Macau tourist arrival volumes; likewise, Wen et al. [9] used search queries for forecasting tourism demand. The results demonstrate that models that contain a composite search index show good prediction performance; Bangwayo-Skeete and Skeete [31] try to define that Google data can be a factor for tourism prediction. And found that Google search data can improve forecasting performance.

Social media plays an important role in information search. It has been used by tourists to share their experiences. Limited researchers have attempted to predict tourist arrivals using online reviews data. For instance, Önder et al. [12] indicate that Facebook likes data incorporated into econometric models can be exploited as a new variable to explain tourism demand at different destinations. Park et al. [15] apply the news as data for predicting tourist arrivals in Hong Kong. Peng et al. [38] implemented social network data, sentiment analysis, and Gradient Boosting Regression Trees to forecast Huang Shan tourism demand, which has always resulted in good forecasting performance. Fronzetti et al. [39] employed Factor Augmented Autoregressive and Bridge models with social network and semantic variables which have the highest performance than other algorithms based on GoogleTrend data. In most of these studies, the researchers attempted to capture only the likes or the sentiments to build a predictive model. However, topics in the reviews can be used as factors in the predictive model.

3. METHODS

As schematically illustrated in Fig. 1, this framework includes six stages. In the first step, we collect online reviews about tourism demand from the Trip Advisor travel forum and the number of tourist arrivals. Then, the unstructured data were pre-processed. In Step 3, the LDA topic model is used to extract topics and their keywords. The keywords that have impacted tourism demand can be defined as features using TF-IDF representation and Pearson correlation. These features are used as input in the LSTM method to generate a predictive model. Finally, we evaluated the forecasting performances of the proposed model. .

3.1. DATA EXTRACTION

Historical data: Tourist arrivals to Hong Kong from the USA are used to measure tourism demand. We collected the monthly number of tourists From January 2013 to January 2020 from the partnernet website (https://partnernet.hktb.com/).

Online post data: The online review data used in this study were collected from the TripAdvisor travel forum. TripAdvisor is one of the most popular travel platforms and contains several topics or posts, submitted by users. We collected posts using WebHarvy software From January 2013 to January 2020.

3.2. DATA PREPROCESSING

All the data collected were preprocessed using three essential steps:

Tokenization: The first thing in text was breaking the social media posts into words, phrases, or other significant pieces named tokens. We used NLTK word tokenizer to tokenize the posts.

Stop words removal: are words which do not contribute to meaning and are dropped from the text to provide a simpler analysis of the text. For our study, the stop words were removed using the nltk corpus of stopwords.

Lemmatization: tried to get the base form of a word, known a lemma. In our study, lemmatisation is implemented using the WordNet Lemmatizer.

3.3. FEATURE CONSTRUCTION

Our paper uses topic modeling to determine important topics impacting the tourism demand from the posts. Topic modeling is adopted for discovering a set of topics. The most frequently employed topic modeling is the Latent Dirichlet Allocation (LDA) model introduced by Blei [17] in 2003. LDA is very useful and effective to find topics; it is a probabilistic generative model that was used to solve the latent semantic analysis; LDA supposes that each post can be represented as a probabilistic distribution over topics, and each topic in the LDA model is also represented as a probability distribution over words [18], as shown in Fig. 2.

The graphical representation of LDA is illustrated in Fig. 3. Each topic has a corresponding probability distribution for different words. Each post $m \in \{1,...,M\}$ is considered as a multinomial distribution $\theta(d)$ over K topics, and each topic Z is supposed to have a multinomial distribution ϕ over the collection of words W. In concept, LDA detect the topics and the distribution of these topics in each post from a corpus of posts M.

As illustrated in Fig. 4., the LDA process takes all the posts as input. This data must be preprocessed. The result of LDA is K topics, which contain N words.

Selecting the right number of topics K is a crucial component of topic modeling. The topic coherence is utilized in this study to evaluate the topic model findings, which detects semantic similarity between high-ranking words in the topic. In particular, the coherence measures the frequencies of occurrence in documents with which the high-ranking and lower-ranking terms that are related to the same topic.

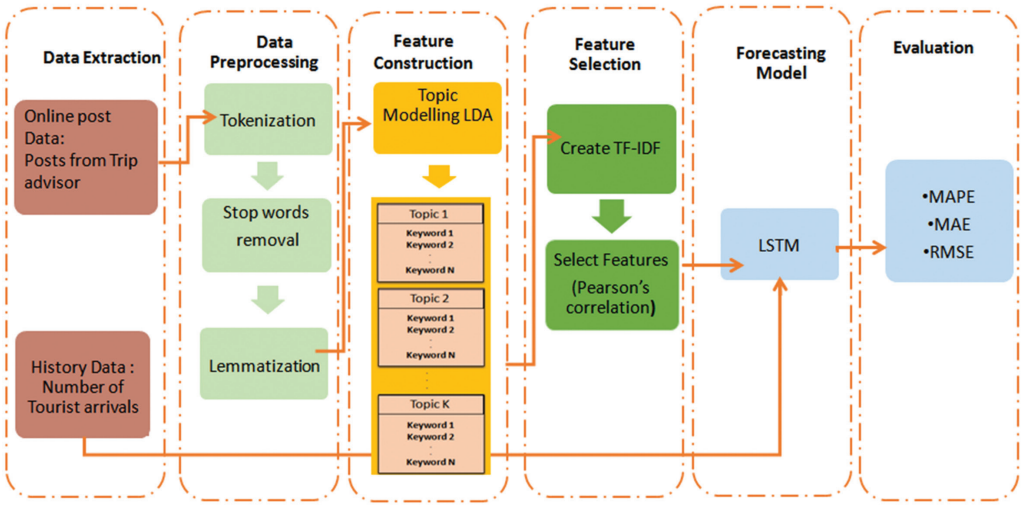


Fig. 1. Framework of proposed model

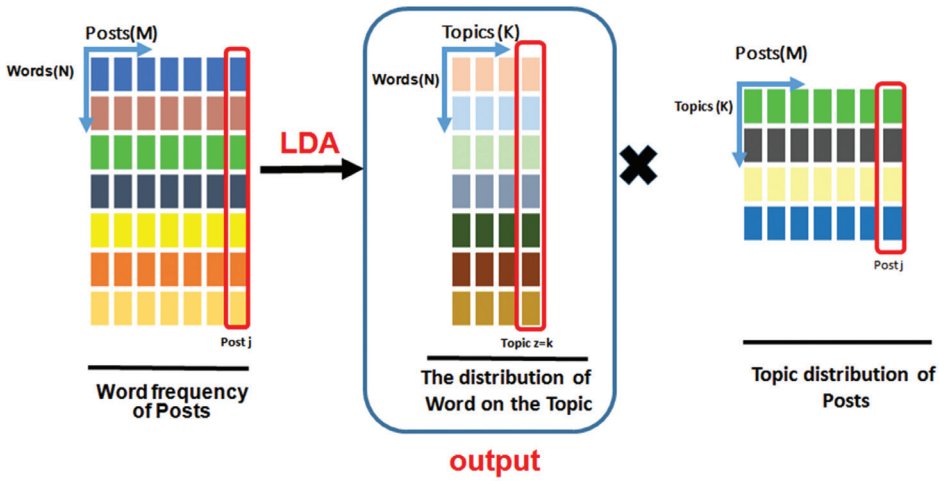


Fig. 2. The implementation of the LDA model

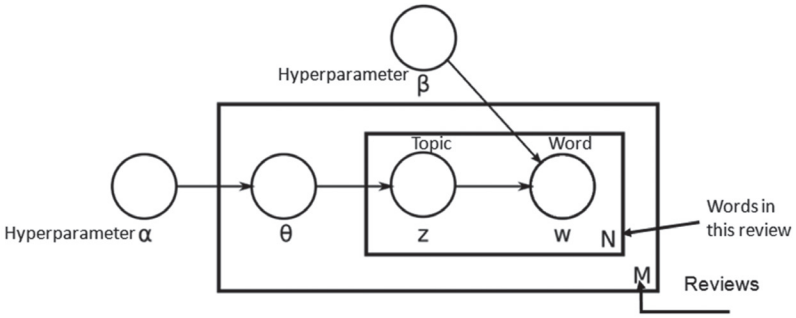


Fig. 3. The graphical model of LDA.

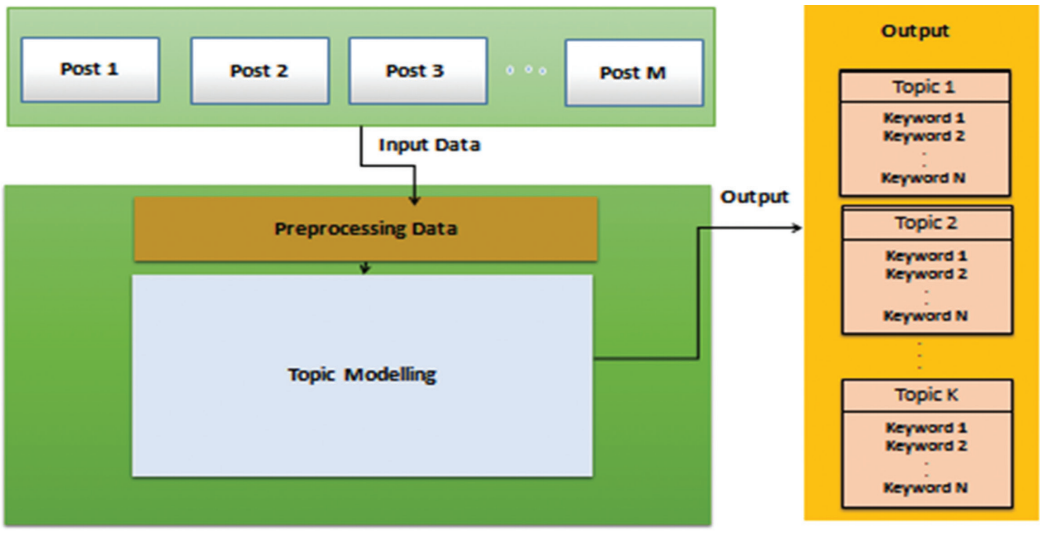


Fig. 4. Output of LDA model

3. 4. FEATURE SELECTION

- Numerical representation of the word: The TF-IDF for Term frequency- inverse Document Frequency, is employed to convert a review into structured format. It measures the importance of a given word in a given review, and the meaning of a word is defined by its frequency.
- Pearson's correlation: Pearson's correlation is the procedure of choosing an influential factor from a large collection of processed data.

3.5. TOURISM DEMAND PREDICTION BASED ON DEEP LEARNING

After identifying the appropriate features, the prediction models may be employed to predict the tourism demand in Hong Kong. In this paper, one deep learning model is chosen as the principal method, and three machine learning models are chosen as benchmark methods, i.e. LSTM, SVR, RF, and ANN. They are recurrent prediction technique, kernel prediction technique, ensemble prediction technique, and nonlinear prediction technique, respectively.

LSTM is a kind of Recurrent Neural Network (RNN) with additional variables to remember the sequence of data. LSTM is developed by Hochreiter to resolve the vanishing gradient problem of RNN. Each LSTM unit is a cell that saves information which is updated by three gates: the input gate, the forget gate, and the output gate. The input gate determines what significant information can be added from the current step and the output gate decides the part of the cell state being outputted. An illustration of LSTM unit structure is shown in Fig. 5.

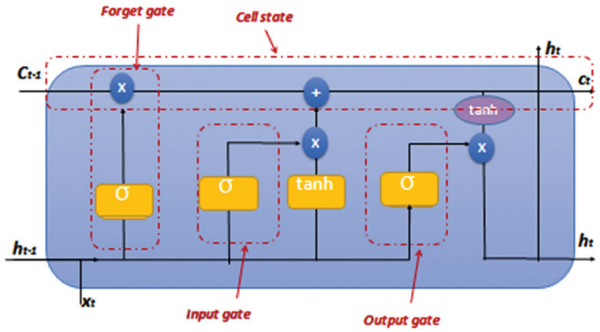


Fig. 5. The LSTM unit structure

Fig. 6. displays the model implemented for tourism demand forecasting with LSTM. The dataset used in this model is the association between the data that arrived from the TF-IDF model and the number of tourists monthly. This model consists of one hidden layer of LSTM units followed by an output layer for forecasting tourism demand.

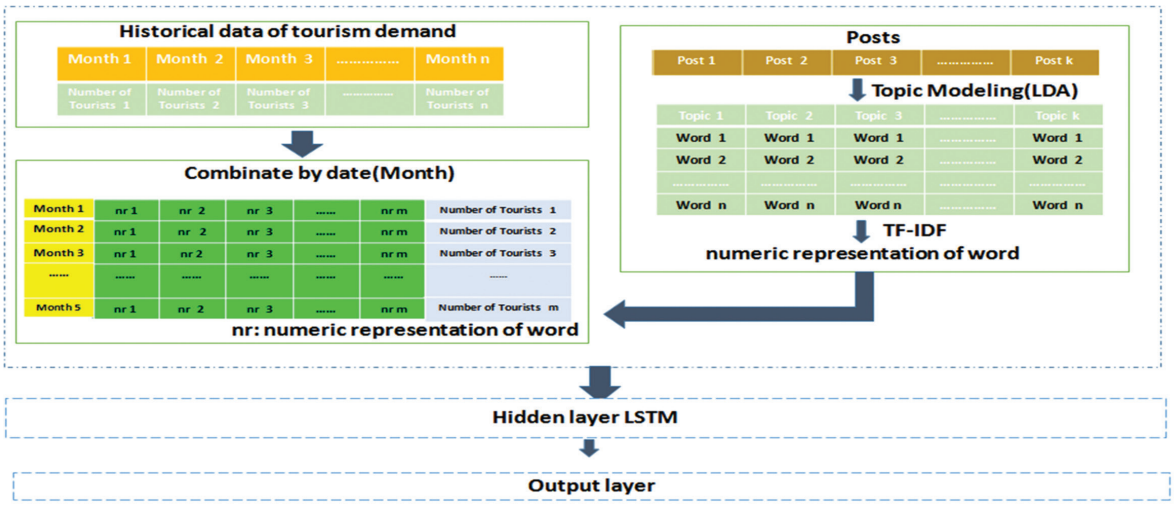


Fig. 6. Lstm in our framework

3. 6. EVALUATION METRICS

For the evaluation of all models, we use Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root-Mean-Square Error (RMSE) measures. The smaller values of RMSE, MAE, and MAPE conduct to higher accuracy of the best model.

4. RESULTS AND DISCUSSION

The monthly number of tourists is illustrated in Fig. 7. In this study, first, WebHarvy software is used to collect reviews from TripAdvisor, which are registered in the excel document. Various kinds of data were crawled, comprising the origins of the posters (just from USA), the year and the month of the post, the title of the post, and the contents of the post. In total, 4987 posts and 17532 words were obtained. Right away after the collection of the posts. We undertake data cleaning or data preprocessing. Approximately 5484 words were left.

4.1. FEATURE CONSTRUCTION

To construct the LDA model, it is essential to define the optimal number of topics K. K ranges from 1 to 20. Figure 8 displays the coherence score values. The model that provided the highest coherence score is chosen. In our case, the top coherence score was reached at K = 10. Ten topics are extracted for analysis and discussion because they have the best coherence score among the other topics.

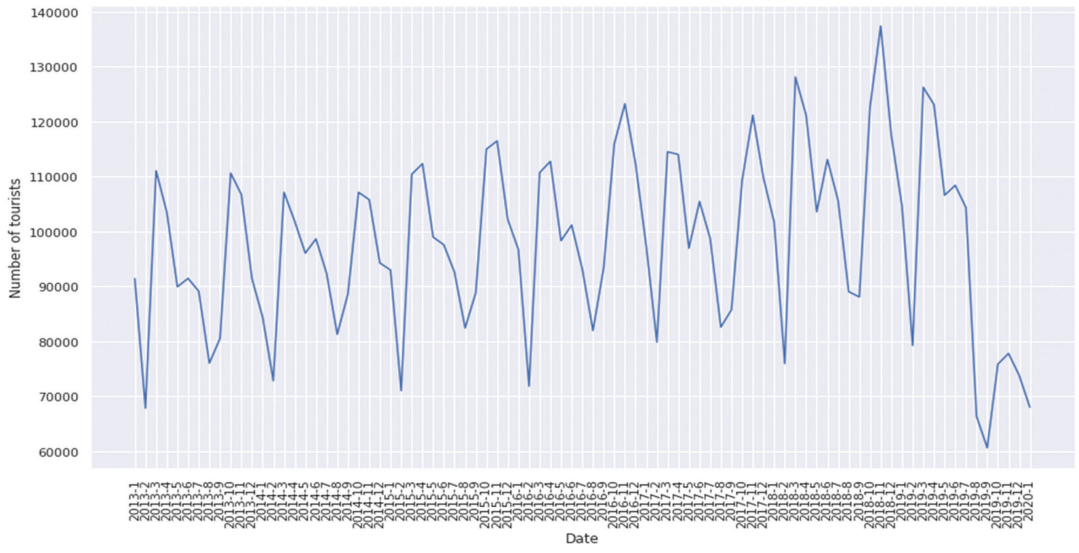


Fig. 7. Actual tourist arrivals to HongKong from USA

After trying several values for the number of topic K, we set it to 10. LDA does not generate a label for each topic; we manually assign a label to a topic word. Table 1 shows the top five words from the top ten topics identified from the posts using the Latent Dirichlet Allocation (LDA) topic modeling method. LDA can distill essential information regarding tourism demand from social media data. The ten topics are: Transport, lodging, Dining, Weather, Visit, Experience, Currency, Shopping, Busy place, Guide.

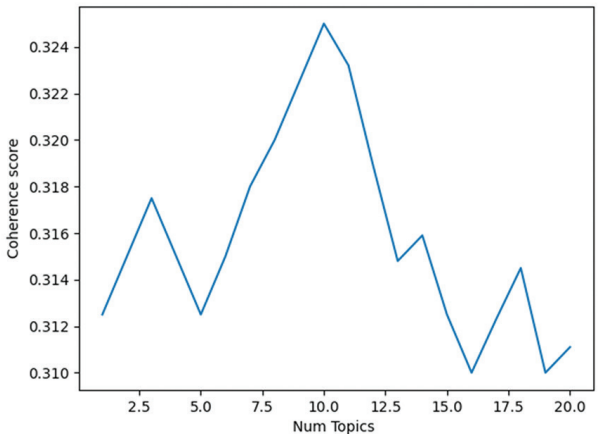


Fig. 8. The Coherence score

Table 1. Important topics identified using Latent DirechletAllocation.

Topics	Words
transport	Airport, metro, tickets, reservations, train.
lodging	Hotel, accommodation, street, Apartment, views
Dining	Bar, restaurant, food, eat, casino
Weather	Weather, degrees, down, heat, rain
Visit	Visit, trip, Disneyland, Oceanpark, museum.
Experience	Good, Bad, very interesting, Amazing, fun
Currency	Exchange rate, credit card, People, want, Money
Shopping	Shop, Price, Gift, Fashion, Clothing
Busy place	Busy, Place, People, crowd, Holiday
Guide	Guide, Sightseeing,Location, Hong Kong , Ticket

4.2. PEARSON CORRELATION

The data is randomly divided in certain percentages, 80% of our data is selected as the training data and the rest is considered as testing data.

We calculated the correlation between the numbers of tourists and all other features and selected the words that have a correlation of more than 0.01. It reduced the number of variables by more than 40%.

4.3. HYPERPARAMETER OF LSTM MODEL

For training our proposed model, it should discover the optimal combination of hyper parameters to get the best performance. After many experiments, we obtain the following combination: The model is trained for 50 epochs with a batch size of 64 and learning rate was 0.002. The optimizer used is the Adam optimizer.

4.3. MODEL EVALUATION

The error indices of the proposed model are calculated and illustrated in Table 3. The LSTM method outperforms the SVR, RF and the ANN in terms of RMSE, MAE and MAPE. The RMSE value of the LSTM method is 840.87 The MAE value of the LSTM method is 743.32, while the MAPE value is 5.145.

To test the effectiveness of integrating reviews data and topics in the predictive model, we calculate the error indices for the LSTM, RF, SVR and ANN methods with only historical data (the number of tourists only). The results are shown in Table 3. In terms of RMSE, MAE and MAPE, we observe that models based only on the number of tourists as a feature show unsatisfactory results and achieve higher values compared to models that use topics as features (Table 2). Furthermore, our approach considering LSTM model outperforms all other models with topics and with only traditional data.

Fig. 9. demonstrate the comparison between actual tourist arrivals and predictions of all models that use topics as features. The actual tourist arrival and prediction of the proposed model are shown as a blue line and an orange line, respectively. In contrast, the predictions of other benchmark models are denoted by a green dashed line for RF, a red dashed line for SVR and a purple dashed line for ANN. The forecasting accuracy of the LSTM forecasting model is higher than that of the machine learning forecasting models.

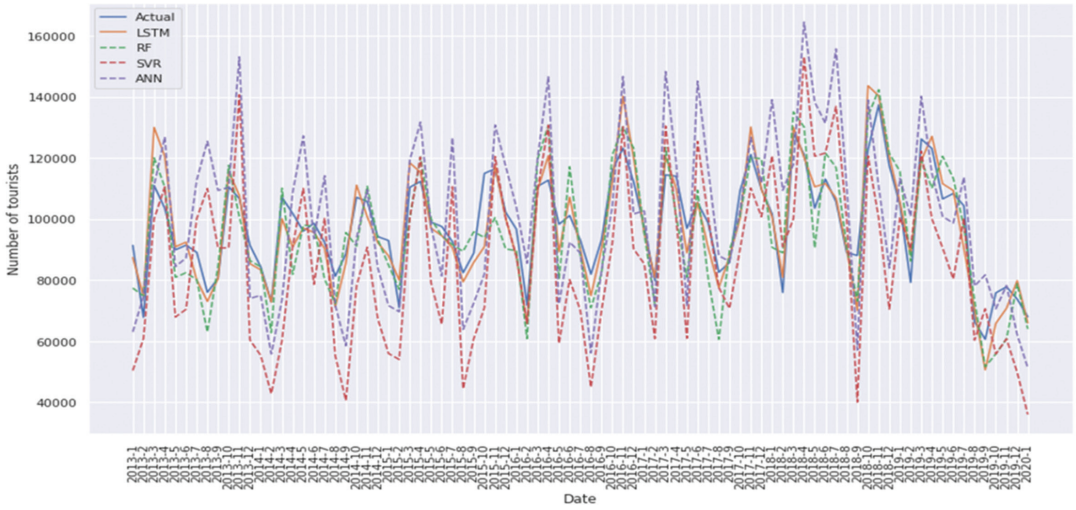


Fig. 9. Actual tourist arrivals and predictions of all models

Table 2. Prediction performance for each model with social media data

	Training data			Test data		
	RMSE	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
LSTM	**840.87**	**743.32**	**5.145**	**960.54**	**855.46**	**5.475**
SVR	1653.98	1399.42	8.835	1721.39	1585.74	8.934
RF	1139.48	932.62	6.483	1183.73	1024.83	6.947
ANN	1463.82	1256.39	7.719	1594.35	1594.36	8.475

Table 3. Prediction performance for each model with only historical data

	Training data			Test data		
	RMSE	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
LSTM	**1680.93**	**1655.68**	**8.879**	**1764.46**	**1710.32**	**9.121**
SVR	1826.29	1802.90	9.209	1884.35	1865.46	9.532
RF	1734.73	1729.52	8.971	1799.64	1784.63	9.249
ANN	1772.38	1768.31	9.138	1862.35	1855.63	9.438

5. CONCLUSIONS& FUTURE WORK

In this article, we attempt to construct an accurate model for tourism demand prediction, which is a very challenging task. Researchers have used search engine data and traditional data in a forecasting model. Although search engine data can complement traditional data, it produces less information than social media data. We presented a novel model that uses social media data by determining the main topics present in the online reviews and adding them as new factors in the deep learning model to predict tourism demand. We extracted data from the TripAdvisor review platform; we discovered a list of influential variables with topic modeling. We used all features in the deep learning LSTM method to predict tourism demand. Experiments indicated that our model achieved better accuracy than the other machine learning methods with social media data on the one hand and with only the number of tourists on the other hand. This study showed that forecasting performance can be greatly enhanced by using online review data to predict tourism demand; In future work, we will take sentiment analysis of social media reviews into consideration to predict tourism demand. Moreover, search query data and other types of data (such as weather data, and temperature data) can be combined to supplement the social media data to provide accurate forecasts.

6. REFERENCES:

[1] X. Xu, R. Law, W. Chen, L. Tang, "Forecasting tourism demand by extracting fuzzy Takagi–Sugeno rules from trained SVMs", CAAI Transactions on Intelligence Technology, Vol. 1, No. 1, 2016, pp. 30-42.

[2] F. L. Chu, "Forecasting tourism: A combined approach", Tourism Management, Vol. 19, No. 6, 1998, pp. 515-520.

[3] M. E. Nor, A. I. M. Nurul, M. S. Rusiman, "A Hybrid Approach on Tourism Demand Forecasting", Journal of Physics: Conference Series, Vol. 995, No. 1, 2018.

[4] A. G. Assaf, G. Li, H. Song, M. G. Tsionas, "Modeling and Forecasting Regional Tourism Demand Using the Bayesian Global Vector Autoregressive (BGVAR) Model", Journal of Travel Research, Vol. 58, No. 3, 2019, pp. 383-397.

[5] J. W. Bi, Y. Liu, H. Li, "Daily tourism volume forecasting for tourist attractions", Annals of Tourism Research, Vol. 83, 2020, p. 102923.

[6] H. Song, G. Li, "Tourism demand modelling and forecasting-A review of recent research", Tourism Management, Vol. 29, No. 2, 2008, pp. 203-220.

[7] J. W. Bi, H. Li, Z. P. Fan, "Tourism demand forecasting with time series imaging: A deep learning model", Annals of Tourism Research, Vol. 90, 2021.

[8] R. Law, G. Li, D. K. C. Fong, X. Han, "Tourism demand forecasting: A deep learning approach", Annals of Tourism Research, Vol. 75, 2019, pp. 410-423.

[9] L. Wen, C. Liu, H. Song, "Forecasting tourism demand using search query data: A hybrid modelling approach", Tourism Economics, Vol. 25, No. 3, 2019, pp. 309-329.

[10] A. Kulshrestha, V. Krishnaswamy, M. Sharma, "Bayesian BILSTM approach for tourism demand forecasting", Annals of Tourism Research, Vol. 83, 2020.

[11] J. W. Bi, C. Li, H. Xu, H. Li, "Forecasting Daily Tourism Demand for Tourist Attractions with Big Data: An Ensemble Deep Learning Method", Journal of Travel Research, Vol. 61, No. 8, 2022, pp. 1719-1737.

[12] I. Önder, U. Gunter, A. Scharl, "Forecasting tourist arrivals with the help of web sentiment: A mixed-frequency modeling approach for big data", Tourism Analysis, Vol. 24, No. 4, 2019, pp. 437-452.

[13] H. Li, M. Hu, G. Li, "Forecasting tourism demand with multisource big data", Annals of Tourism Research, Vol. 83, 2020, p. 102912.

[14] M. Hu, H. Li, H. Song, X. Li, R. Law, "Tourism demand forecasting using tourist-generated online review data", Tourism Management, Vol. 90, 2022, p. 104490.

[15] E. Park, J. Park, M. Hu, "Tourism demand forecasting with online news data mining", Annals of Tourism Research, Vol. 90, 2021.

[16] H. Song, R. T. R. Qiu, J. Park, "A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting", Annals of Tourism Research, Vol. 75, 2019, pp. 338-362.

[17] Gerard S. Forecasting tourist arrivals in Barbados. Annals of Tourism Research, 1995, Vol. 22, No. 4, pp. 804-818.

[18] R. Law, "Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting", Tourism Management, Vol. 21, No. 4, 2000, pp. 331-340.

[19] V. S. Lin, A. Liu, H. Song, "Modeling and Forecasting Chinese Outbound Tourism: An Econometric Approach", Journal of Travel and Tourism Marketing, Vol. 32, No. 1-2, 2015, pp. 34-49.

[20] X. Sun, W. Sun, J. Wang, Y. Zhang, Y. Gao, "Using a Grey-Markov model optimized by Cuckoo search algorithm to forecast the annual foreign tourist arrivals to China", Tourism Management, Vol. 52, 2016, pp. 369-379.

[21] H. Choyakh, "A Model of Tourism Demand for Tunisia: Inclusion of the Tourism Investment Variable", Tourism Economics, Vol. 14, No. 4, 2008, pp. 819-838.

[22] H. Song, S. F. Witt, "Forecasting international tourist flows to Macau", Tourism Management, Vol. 27, No. 2, 2006, pp. 214-224.

[23] S. Page, H. Song, D. C. Wu, "Assessing the Impacts of the Global Economic Crisis and Swine Flu on Inbound Tourism Demand in the United Kingdom", Journal of Travel Research, Vol. 51, No. 2, 2012, pp. 142-153.

[24] P. J. Sheldon, T. Var, "Tourism forecasting: A review of empirical research", Journal of Forecasting, Vol. 4, No. 2, 1985, pp. 183-195.

[25] B. Zhang, Y. Pu, Y. Wang, J. Li, "Forecasting hotel accommodation demand based on LSTM model incorporating internet search index", Sustainability, Vol. 11, No. 17, 2019, pp. 1-14.

[26] S. C. Kon and L. W. Turner, "Neural Network Forecasting of Tourism Demand", Tourism Economics, Vol. 11, No. 3, 2016, pp. 301-328.

[27] C. Vincent, "A comparison of three different approaches to tourist arrival forecasting", Tourism management, Vol. 24, No. 3, 2003, pp. 323-330.

[28] K. He, L. Ji, C. W. D. Wu, K. F. G. Tso, "Using SARIMA-CNN-LSTM approach to forecast daily tourism demand", Journal of Hospitality and Tourism Management, Vol. 49, 2021, pp. 25-33.

[29] Y. Li , H. Cao, "Prediction for Tourism Flow based on LSTM Neural Network", Procedia Computer Science, Vol. 129, 2018, pp. 277-283.

[30] L. Shao-Jiang, C. Jia-Ying, L. Zhi-Xue, "A EMD-BP integrated model to forecast tourist number and applied to Jiuzhaigou", Journal of Intelligent and Fuzzy Systems, Vol. 34, No. 2, 2018, pp. 1045-1052.

[31] P. F. Bangwayo-Skeete, R. W. Skeete, "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach", Tourism Management, Vol. 46, 2015, pp. 454-464.

[32] C. Goh, R. Law, H. M. K. Mok, "Analyzing and Forecasting Tourism Demand: A Rough Sets Approach", Journal of Travel Research, Vol. 46, No. 3, 2008, pp. 327-338.

[33] C. H. Wang, L. C. Hsu, "Constructing and applying an improved fuzzy time series model: Taking the tourism industry for example", Expert Systems with Applications, Vol. 34, No. 4, 2008, pp. 2732-2738.

[34] H. Song, H. Liu, "Predicting Tourist Demand Using Big Data", Analytics in Smart Tourism Design, 2017, pp. 13-29.

[35] S. Sun, Y. Wei, K. L. Tsui, S. Wang, "Forecasting tourist arrivals with machine learning and internet search index", Tourism Management, Vol. 70, 2019, pp. 1-10.

[36] Y. Yang, B. Pan, H. Song, "Predicting Hotel Demand Using Destination Marketing Organization's Web Traffic Data", Journal of Travel Research, Vol. 53, No. 4, 2014, pp. 433-447,.

[37] B. Zhang, X. Huang, N. Li, R. Law, "A novel hybrid model for tourist volume forecasting incorporating search engine data", Asia Pacific Journal of Tourism Research, Vol. 22, No. 3, 2017, pp. 245-254.

[38] T. Peng, J. Chen, C. Wang, Y. Cao, "A Forecast Model of Tourism Demand Driven by Social Network Data", IEEE Access, Vol. 9, 2021, pp. 109488-109496.

[39] A. F. Colladon, B. Guardabascio, R. Innarella, "Using social network and semantic analysis to analyze online travel forums and forecast tourism demand", Decision Support Systems, Vol. 123, 2019.

An Effective Deep Learning Based Multi-Class Classification of DoS and DDoS Attack Detection

Original Scientific Paper

Arun Kumar Silivery
Department of Computer Science & Engineering,
University College of Engineering, Osmania University, Hyderabad- 500 007
arunsilivery@osmania.ac.in

Kovvur Ram Mohan Rao
Department of Information Technology
Vasavi College of Engineering, Osmania University, Hyderabad- 500 031
krmrao@staff.vce.ac.in

L. K. Suresh Kumar
Department of Computer Science & Engineering,
University College of Engineering, Osmania University, Hyderabad- 500 007
suresh.l@uceou.edu

Abstract *– In the past few years, cybersecurity is becoming very important due to the rise in internet users. The internet attacks such as Denial of service (DoS) and Distributed Denial of Service (DDoS) attacks severely harm a website or server and make them unavailable to other users. Network Monitoring and control systems have found it challenging to identify the many classes of DoS and DDoS attacks since each operates uniquely. Hence a powerful technique is required for attack detection. Traditional machine learning techniques are inefficient in handling extensive network data and cannot extract high-level features for attack detection. Therefore, an effective deep learning-based intrusion detection system is developed in this paper for DoS and DDoS attack classification. This model includes various phases and starts with the Deep Convolutional Generative Adversarial Networks (DCGAN) based technique to address the class imbalance issue in the dataset. Then a deep learning algorithm based on ResNet-50 extracts the critical features for each class in the dataset. After that, an optimized AlexNet-based classifier is implemented for detecting the attacks separately, and the essential parameters of the classifier are optimized using the Atom search optimization algorithm. The proposed approach was evaluated on benchmark datasets, CCIDS2019 and UNSW-NB15, using key classification metrics and achieved 99.37% accuracy for the UNSW-NB15 dataset and 99.33% for the CICIDS2019 dataset. The investigational results demonstrate that the suggested approach performs superior to other competitive techniques in identifying DoS and DDoS attacks.*

Keywords*: DDoS, deep learning, Alexnet, Resnet-50, DCGAN, Atom search optimization algorithm*

1. INTRODUCTION

In today's world, systems for Information and Communication Technology (ICT) significantly influence every element of society and business. At the same time, cyberattacks on ICT systems are becoming more sophisticated and more frequent [1-3]. This significantly impacts network performance, resulting in instability that would render the network unusable. Thus, ICT systems require a very effective network security solution. One of the most popular methods for spotting different kinds of malicious network attacks is Intrusion Detection System (IDS).

The two primary strategies for detecting intrusion are signatures-based IDS and anomalies-based IDS [4-6]. The signature-based IDS is also known as Knowledge-based Detection or Misuse Detection. It is as effective as upgrading the database at a specified time since it focuses on finding a "signature" or patterns of intrusion event. The anomaly-based IDS is also known as Behavior-based Detection. It is based on comparing reliable behavioral patterns with unexpected behaviors while observing routine activities [7-9]. An administrator uses an Intrusion Prevention System (IPS) to stop threats like denial of service (DoS), Distributed DoS (DDoS) attacks, Trojan horses, etc., when the IDS system detects them.

The DoS and DDoS attacks are significant traits in anomaly-based IDS. As a result of the modification of various services, there has been a growth in these attacks over the past ten years, establishing them as a

severe threat to the security of networks [10-12]. Additionally, tracking security attacks has become a significant hurdle for most organizations despite spending excessive money to secure the system. However, large-scale cyberattacks are still happening, attackers are becoming more sophisticated, and tools to defend against them have become outdated.

Every DDoS assault is based on the same basic idea. Using a network protocol, the attacker floods the server with spoof request packets. Since the target server cannot distinguish between them, it begins serving every packet. The server becomes overloaded and crashes while attempting to fulfill all the requests [13-15]. Because of this, an attacker can exhaust all server resources, resulting in a denial of service. It is referred to as a distributed denial of service attack. In this, multiple devices send packets instead of a single source machine.

In the past few years, many IDS techniques have been presented based on various approaches, such as mathematical formulations and data mining techniques like machine learning. Poor performances are caused by the difficulty in managing the high-dimensional network traffic data by these statistical formulations and conventional machine learning models [16-18]. Furthermore, most existing techniques used only binary classification, such as whether it is an attack. Therefore, better approaches are required for IDS, such as deep learning-based techniques. Due to its powerful learning and feature extraction capabilities, particularly in scenarios involving large datasets, deep learning has been widely recommended for IDS in recent years. Multiple layers are used in deep learning approaches to gradually extract essential features from the raw input without domain knowledge [19, 20].

Therefore, an Alexnet-Resnet-50-based deep learning model is presented in this paper for multi-class classification of DoS and DDoS attack detection. The Atom search algorithm-based hyperparameter optimization and Deep Convolutional Generative Adversarial Network (DCGAN) based Data augmentation is implemented to increase the classifier's efficiency. Our framework makes use of ResNet's superior capacity for learning. From the input data, it automatically extracts the essential elements. This produces better outcomes and avoids the trouble of manually selecting features. The deep network structure of Alexnet leads to a faster training process and avoids the vanishing gradient problem. Therefore, combining these techniques achieves effective DoS and DDoS attack classification results.

In this paper, the primary contributions are listed as follows,

- An effective deep learning technique Alexnet has been proposed to create the intrusion detection system. This model effectively detects various cyber threats like DoS and DDoS attacks.
- To deal with the imbalanced data issue on CCIDS 2019 and UNSW-NB15 Dataset, an effective Deep Convolutional Generative Adversarial Network (DCGAN) is implemented.
- To improve the classification process and achieve high accuracy, the significant features are extracted by the ResNet-50-based technique.
- To decrease the learning complexity of the classifier, the parameters are estimated by the Atom search optimization algorithm, which is used to achieve fast and high-accuracy classification.
- Finally, extensive performance assessments of several existing techniques have been carried out using recall, f1-score, accuracy, and precision metrics.

The remaining portion of this paper is organized as follows: Section 2 provides the relevant studies in this area. Section 3 explains the steps of the suggested strategy. In Section 4, the experimental analysis is entirely detailed. The paper is finally finished with Section 5.

2. LITERATURE REVIEW

Convolutional Neural Network (CNN) based hybrid deep learning model was used by Alghazzawi et al. [21] for DDoS attack detection. Initially, they conducted the preprocessing to prepare the raw data for processing. Afterward, the feature selection process based on the chi-squared test was implemented. Then, the features were extracted using CNN; the Bidirectional long/short-term memory (BiLSTM) system was used to detect DDoS attacks. By employing standard performance criteria, including f1-score, recall, precision, and accuracy, the authors evaluated the findings on the CIC-DDoS2019 dataset.

To identify the data representing network traffic patterns, including both regular and DDoS traffic, Aamir and Zaidi [22] devised a clustering-based technique. To extract the features from the dataset, Principal Component Analysis (PCA) was used with two clustering techniques such as k-means and agglomerative techniques. Then a voting technique was implemented to provide the labels for the data to distinguish between assaults and legitimate traffic. After labeling, trained models for future classification are obtained using the supervised machine learning algorithms and Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Random Forest (RF).

Panigrahi et al. [23] suggested the IDS based on the Consolidated Tree Construction (CTC) method to address the issue of class imbalance. A Supervised Relative Random Sampling (SRRS) technique has been developed for the preprocessing stage to overcome the imbalanced data issue. An Improved Infinite Feature Selection for Multi-class Classification (IIFS-MC) has been implemented to choose the best features of the sample. Finally, J48Consolidated, equipped with CTC, was used to identify potential threats. NSL-KDD, ISCX-IDS2012, and the CICIDS2017 dataset were used for performance assessment.

Khan [24] developed the hybrid deep learning framework to anticipate and categorize harmful cyberattacks in the network using a Convolutional Recurrent Neural Network (CRNN). In this system, CNN conducts convolution to capture local information, while recurrent neural networks (RNNs) capture temporal features to enhance the performance and prediction of the ID system. Experiments were conducted on a publicly accessible CSE-CIC-DS2018 dataset to assess the detection capacity of the suggested approach.

Bi-LSTM with an attention mechanism-based deep learning technique was presented by Fu et al. [25] for traffic anomaly detection. This system initially uses a CNN network to extract sequence information from the traffic data, after which it uses the attention method to reassign the weights of each channel. Finally, Bi-LSTM was used to learn the extracted features. The authors used Adaptive Synthetic Sampling (ADASYN) method to expand minority class samples to address data imbalance issues. Moreover, to decrease the dimensionality of the data, a modified stacked autoencoder was utilized to improve information fusion. For performance assessment, the NSL-KDD dataset was used.

Mighan and Kahani [26] created a hybrid SAE-SVM approach for a cyber-security IDS. The suggested approach employed stacked autoencoder networks for feature extraction and SVM for classification. In the Autoencoder network, multilayer perceptrons were employed for feature extraction. Binary classification and multi-class classification were both done by the authors. The trials were based on accuracy, training time, prediction time, and other performance parameters using the UNB ISCX 2012 IDS dataset.

Based on supervised machine learning, Moualla et al. [27] developed the IDS that contains many Phases. The dataset's unbalanced class problem was first solved by the Synthetic Minority Oversampling Technique (SMOTE) approach. The Extremely Randomized Trees Classifier was then used to choose the key features for each class in the dataset according to the Gini Impurity criterion. Finally, the attacks were classified using the pre-trained extreme learning machine (ELM) model, and the experiments were conducted on the UNSW-NB15 dataset.

3. PROPOSED METHODOLOGY

The proposed technique contains four major conceptual components. They are preprocessing, data augmentation, feature extraction, and classification. Initially, the raw data are preprocessed using several techniques. Afterward, Deep Convolutional Generative Adversarial Network (DCGAN) based data augmentation technique is implemented to increase the samples of the minority classes to rectify the imbalanced data issue. Then, the ResNet-50-based deep learning technique is applied to perform the feature extraction. Finally, an optimized Alexnet-based technique is implemented to perform multi-class attack detection. Fig. 1 displays the overall system architecture.

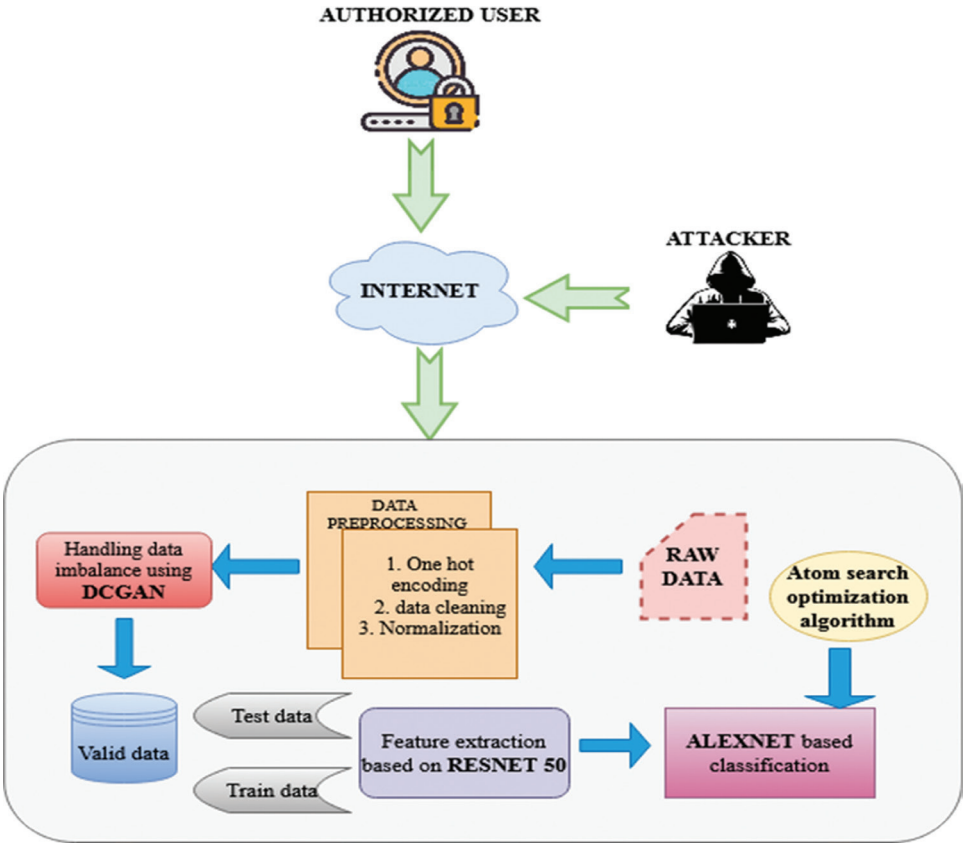


Fig. 1. System Framework

3.1. PREPROCESSING

Analyzing and cleaning the raw datasets before implementing the proposed technique is necessary because they are typically inaccurate, uniform, or comprehensive.

Removing socket data: The CICIDS2019 and UNSW NB15 datasets are in.csv file format, and all socket features—including server and client IP addresses, time stamps, flow ID sources, and destination ports—have been eliminated from the datasets. Because both regular users and intruders have the same IP address; In addition, characteristics like Bwd Bulk Rate (Avg), Bwd Packet/Bulk (Avg), Bwd PSH (Flags), Fwd Packet/Bulk (Avg), Fwd Bytes/Bulk (Avg), Bwd URG, Fwd URG (Flags), and Fwd Bulk Rate (Avg) that have the same value across all rows are eliminated.

Encoding: One Hot Encoder is employed in this paper. This encoder creates a new column for each label represented in the dataset and assigns a value of 1 or 0 depending on whether the record falls into that category.

Normalization: In this study, Min-Max normalization is used. The primary goal of this normalization is to equalize the values of the features within the range 0 and 1.

$$x' = \frac{x - x_{mn}}{x_{mx} - x_{mn}} \tag{1}$$

Here, x_{mn} and x_{mx} denote minimum and maximum eigenvalue, respectively, the normalized eigenvalue denoted by x' and the original eigenvalue denoted by x.

3.2. DCGAN-BASED DATA AUGMENTATION

To avoid the imbalance class problem, Deep Convolutional Generative Adversarial Network (DCGAN) is applied for oversampling rare classes. In the UNSW-NB15 dataset, the 'analysis,' 'Shellcode,' 'Worms,' and 'backdoor' classes are increased by the DCGAN technique. Convolutional neural layers are utilized for this network's generator and discriminator models. The following objective function organizes this network:

$$V(D,G) = E_x \sim p_{data(x)}[\log D(x)] E_z \sim p_{z(z)}[\log(1 - D(G(z)))] \tag{2}$$

Here, the actual sample is denoted by x, $D(x)$ denotes that the discriminator networks would correctly identify x as an actual sample, $G(z)$ denotes the actual sample from the noise z created by the generator network G, and a probability $D(G(z))$ denotes that the discriminator network D will recognize $G(z)$ as an actual sample.

In both the generating and discriminating networks, stride convolution is replaced by the pooling operation in DCGAN. Additionally, global pooling is used in place of the fully connected layer to increase model stability. Following that, Equations (3) and (4) determine the discriminator loss $Ls(D)$ and generator loss $Ls(G)$.

$$Ls(G) = \frac{1}{N} \sum_{i=1}^{N} -\log(D(G(z_i))) \tag{3}$$

$$Ls(D) = \frac{1}{N} \sum_{i=1}^{N} -\log(D(x_i)) - \log(1 - D(G(z_i))) \tag{4}$$

In network training, SGD updates the parameters of discriminator and generator networks based on the loss functions mentioned above. The generator's initial layer is the entire connection layer, followed by the convolution, batch normalization, and leakyReLu activation function, and finally, tanh activates the final layer of convolution. In the discriminator, all the layers follow convolution, batch normalization, and leakyReLu activation, except the first and output layers. Batch normalization is not used in these layers.

3.3. RESNET-50-BASED FEATURE EXTRACTION

After the preprocessing, the data is transferred to the ResNet-50 network for feature extraction. The primary goal of this section is to extract meaningful representations from the data to increase the recognition accuracy of the proposed classifier. In this network, the input data is processed through the 1D convolution layers and 16 residual blocks, which are utilized to extract deep features from the data. These blocks are used to address the decomposition and gradient disappearance problems that are typically present in general CNNs. The residual block improves a system's performance without depending on network depth. The following provides the residual function.

$$y = F(x, W) + x \tag{5}$$

Here, the weight is denoted by W, the residual input block is denoted by x, and the output is denoted by y.

The residual block in this network comprises two ReLU activation levels, three 1D convolution layers, and three 1D Batch Normalization layers. Convolution and batch normalization layers match dimensions and skip connections, respectively. Moreover, ReLU layers perform the nonlinear activation, and batch normalization layers are utilized to speed up and stabilize the model. The features are extracted and sent to the average pooling layer in the residual block. The features are pooled, and the pooling results are transferred to the output layer to produce the final features. These features are most intricate and distinctive. To decrease the overfitting, the dropout was set to 0.2.

3.4. ALEXNET-BASED INTRUSION CLASSIFICATION

After the feature extraction, the extracted features are transferred to the Alexnet technique for final intrusion classification. The traditional AlexNet contains three fully connected layers, five convolutional layers, three max-pooling layers, and two Local Response Normalization (LRN) layers. To reduce the complication of the network, we used three 1D convolutional layers, two fully connected layers, three pooling layers, and

two LRN layers. The convolutional layer is connected to each pooling layer. The random inactivation neuron operation is added to the previous two fully linked layers to prevent the proposed model from overfitting. At last, a softmax layer is the last layer for intrusion classification.

To lower the suggested deep learning framework's computational cost, the pooling layers in the proposed architecture are employed to reduce the size of the feature map. After the first two sessions, response normalization is carried out to lower the test error rate of the suggested network. Network input layers and network input as a whole are both normalized in this layer. The normalization procedure of this layer is explained in the following equation.

$$N_e^x = \frac{b_e^x}{(z+\alpha\sum_{j=\max(0,x-c/2)}^{\min(T-1,x+c/2)}(b_e^x)^2)^\gamma} \quad (6)$$

Here, the normalization of b_e^x neurons activity is denoted by N_e^x, calculated at point e using kernel *k*. *z*, *c*, α, and γ are constants, and the whole kernel's range inside the layer is denoted as *T*. Following that, the output's learned representation is sent into the Softmax layer for multi-class classification aids in calculating classification probabilities. The final classification layer uses these probabilities to categorize the various types of attacks. Then to enhance the classifier's performance, the Atom search optimization algorithm is used to optimize Alexnet's network parameters, including momentum, epoch, initial learning rate, and mini-batch size.

3.4.1. Atom Search optimization algorithm

This section implements the combination of the swarm and physics-based algorithm named the Atom Search optimization (ASO) algorithm to optimize the classifier's parameters. The ASO is created using an analysis of the dynamics of molecules and a heuristic algorithm that relies on the kept population. In other words, it may be assumed that the functioning of the suggested ASO relies on the search for the global optima while simulating the mobility of the atoms, which is governed by interactivity and reservation forces. The ASO approach is incredibly straightforward to build and performed exceptionally well. The location of i^{th} atom in a population of *N* atoms is now computed as,

$$x_i = [x_i^1,.....,x_i^d,.....x_i^D], \quad i = 1,.....N \quad (7)$$

Here, in a D-dimensional space, the i^{th} atom's d^{th} position component is ($d = 1\ldots D$). The fitness assessment of the existing population of atoms is used to determine the masses of atoms. That is described as,

$$Ms_i(nt) = e^{-\frac{Fit_i(nt)-Fit_{best}(nt)}{Fit_{worst}(nt)-Fit_{best}(nt)}} \quad (8)$$

$$ms_i(t) = \frac{Ms_i(t)}{\sum_{j=1}^{N} Ms_j(t)} \quad (9)$$

At the nth iteration, all atoms' worst and best fitness is denoted as Fitworst(nt) and Fitbest(nt) correspondingly. An ith atom's and normalized mass are denoted as Msi(nt) and msi(nt), respectively. One way to express the overall force exerted by the other atoms on the i^{th} atom is as follows:

$$F_i^d(nt) = \sum_{j\in Kbest} rand_j F_{ij}^d(nt) \quad (10)$$

$$F_{ij}^{'}(nt) = -\eta(nt)[2(h_{ij}(nt))^{13} - (h_{ij}(nt))^7] \quad (11)$$

Here, the depth function η (*t*) is expressed as,

$$\eta(nt) = \alpha(1-\frac{nt-1}{mT})^3 e^{-\frac{20nt}{mT}} \quad (12)$$

Here, the maximum number of iterations is denoted by '*mT*,' and the depth weight is denoted as '*a*'. A definition of h_{ij} is expressed as,

$$h_{ij}(nt) = \begin{cases} h_{\min} & \frac{r_{ij}(nt)}{\sigma(nt)} < h_{\min} \\ \frac{r_{ij}(nt)}{\sigma(nt)} & h_{\min} \le \frac{r_{ij}(nt)}{\sigma(nt)} \le h_{\max} \\ h_{\max} & \frac{r_{ij}(nt)}{\sigma(nt)} > h_{\max} \end{cases} \quad (13)$$

Here, the upper and lower range of *h* is denoted as h_{max} and h_{min} correspondingly. The length scale σ (*nt*) is described as

$$\sigma(nt) = \left\| x_{ij}(nt), \frac{\sum_{j\in K\,best} x_{ij}(nt)}{K(nt)} \right\|_2 \quad (14)$$

and

$$\begin{cases} h_{\min} = g_O + g(nt) \\ h_{\max} = u \end{cases} \quad (15)$$

Here, the drift factor '*df*' has the following definition.

$$d_f(nt) = 0.1 \times \sin(\frac{\pi}{2} \times \frac{nt}{mT}) \quad (16)$$

The definition of the constraint force is,

$$G_i^d(nt) = \lambda(nt)(x_{best}^d(nt) - x_i^d(nt)) \quad (17)$$

The Lagrangian multiplier is described as,

$$\lambda(nt) = \beta e^{-\frac{20nt}{mT}} \quad (18)$$

In this case, the multiplier weight is β. The i^{th} atom's force at time *nt* for each repetition is expressed as,

$$F_r = F_i + G_i \quad (19)$$

$$a_i^d(nt) = \frac{F_r^d(nt)}{m_i(nt)} = -\alpha\left(1 - \frac{nt-1}{mT}\right)^3 e^{-\frac{20\,nt}{mT}} \times \sum_{j \in Kbest} \frac{rand_j\left[2 \times (h_{ij}(nt))^{13} - (h_{ij})^7\right]}{m_i(nt)} \frac{(x_i^d(nt) - x_i^d(nt))}{\left\|x_i(nt), x_j(nt)\right\|2} + \beta e^{-\frac{20\,nt}{mT}} \frac{x_{best}^d(nt) - x_i^d(nt)}{m_i(nt)} \quad (20)$$

In this case, the multiplier weight is denoted as β, and the depth weight is denoted as 'a'. The first K atoms with the highest fitness values make up the subset of an atom population known as K_{best}; K is denoted by

$$K(nt) = N - (N-2) \times \sqrt{\frac{nt}{mT}} \quad (21)$$

The i^{th} atom's speed and position at time $nt + 1$ are updated accordingly.

$$v_i^d(nt+1) = rand_i^d v_i^d(nt) + a_i^d(nt) \quad (22)$$

$$x_i^d(nt+1) = x_i^d(nt) + v_i^d(nt+1) \quad (23)$$

Here, the atom's velocity is denoted by v_i^d, the acceleration of the atom is denoted by $a_i^d(nt)$, and i^{th} atom's position is denoted by $x_i^d(nt)$. Finally, the algorithm's best optimal value initializes the classifier's hyperparameters. The optimized values are momentum=0.9, weight decay=0.005, epoch=100, initial learning rate=0.001, and mini-batch size=32.()

4. RESULTS AND DISCUSSION

The performance and efficiency of the proposed IDS are assessed by several experiments on the UNSW-NB15 and CICIDS2019 datasets using typical performance metrics, which are discussed in this section. The experiments were carried out with Python programming language, and Keras was used to run all of the simulations with Tensorflow as the backend, using an Intel Core i7-7700 CPU and 32 GB RAM.

4.1. DATASET DESCRIPTION

4.1.1. UNSW-NB15dataset

The Australian Centre for Cyber Security research team developed the UNSW-NB15 dataset. More than 2.5 million network packets are used to replicate this data set. This data set includes non-anomalous packets and nine other types of attacks (Exploits, Reconnaissance, DoS, Generic, Shellcode, Fuzzers, Backdoors, Worms, and Analysis). The data set is highly unbalanced because more than 87% of the packets are non-anomalous.

4.1.2. CICIDS2019 dataset

This dataset includes a variety of DDoS assaults that can be conducted via TCP/UDP application layer protocols. The taxonomy of attacks in the dataset is carried out in terms of exploitation-based and reflection-based assaults. More than 80 flow features are included in the dataset. The dataset was gathered over two various days for testing and training analysis. The assaults in the dataset include DDoS attacks using DNS, NTP, NetBIOS, SYN, MSSQL, UDP-Lag, LDAP, and SNMP.

4.2. PERFORMANCE METRICS

The proposed intrusion model's performance was assessed utilizing its accuracy, precision, recall, and F-score criteria. An overview of the classification metrics is given in this section. The proportion of accurately classified data out of all classified data is how accurately something is classified. The accuracy of an optimistic prediction is estimated by precision. A low rate of false positives indicates high precision. Recall counts the instances that are classified correctly as positive. F-score integrates recall and precision. It is possible to define it as the average of precision and recall. The measures mentioned above can be expressed mathematically as,

$$Accuracy = \frac{TruPsv + TruNeg}{TruPsv + TruNeg + FlsPsv + FlsNeg} \quad (24)$$

$$\mathrm{Pr}ecision = \frac{TruPsv}{TruPsv + FlsPsv} \quad (25)$$

$$\mathrm{Re}\,call = \frac{TruPsv}{TruPsv + FlsNeg} \quad (26)$$

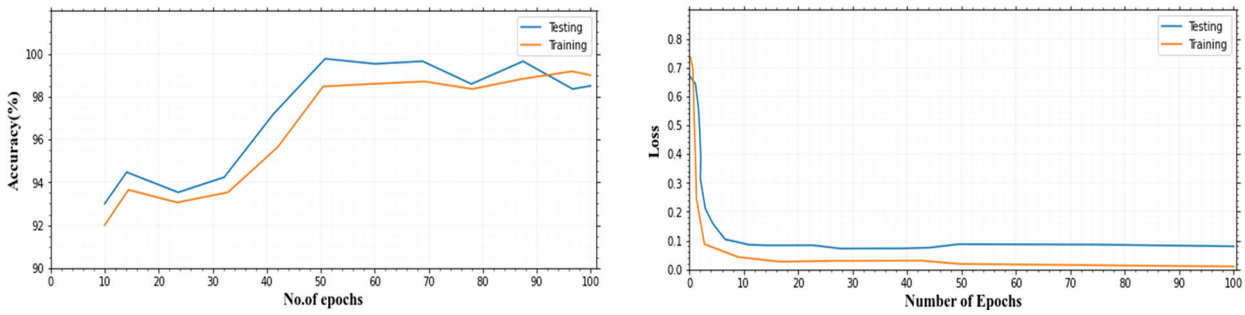
$$F1 - Score = \frac{2 * precision * recall}{precision + recall} \quad (27)$$

Here, true positives (TruPsv) are the class or value of occurrence that was accurately predicted. False positives (FlsPsv) occur when the actual class differs from the anticipated one, and false negatives (FlsNeg) are values for no events that were mistakenly predicted to occur. Correctly estimated no event values are referred to as true negative (TruNeg) values.

4.3. TRAINING AND TESTING

In this method, the entire dataset is split into two groups: one is used to train the network (70% of data), and the other is used to test it (30%). The training set is used to train the proposed model for 100 epochs to decrease the error in the model to the fullest degree possible, and a 0.001 learning rate allows the model to train faster. Moreover, the proposed ASO algorithm optimizes momentum=0.9, weight decay=0.005, mini-batch size=32, and the bias of each layer in the classifier. The training and testing accuracy and loss of the proposed approach for both datasets are shown in Figs. 2 and 3

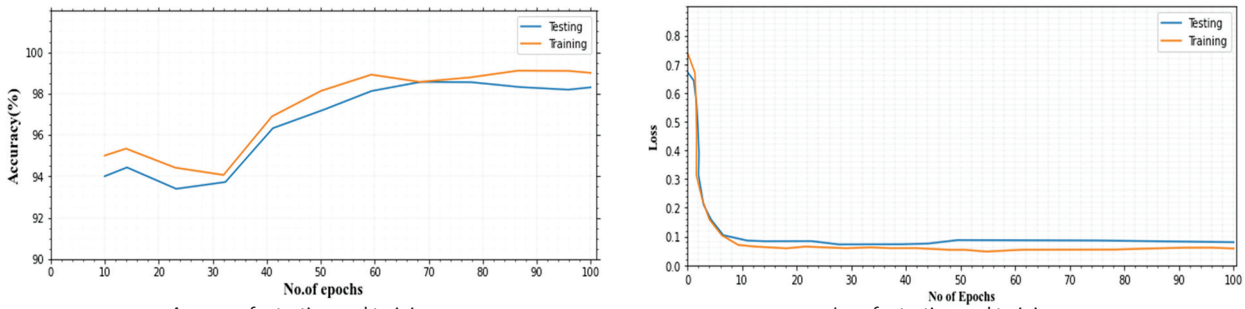
The testing and training accuracy of the presented technique for both datasets are shown in Figs. 2 and 3, and the loss values range from 0.001 to 0.004. When evaluated using benchmark datasets, the suggested model's performance displays a similar pattern, demonstrating the model's ability to predict attacks from other categories besides those stated. Additionally, as the image shows, the training loss is relatively high initially but gradually decreases as the training progresses. Once the training epoch reaches 20, the error usually decreases more slowly.



Accuracy for testing and training

Loss for testing and training

Fig. 2. (a) Testing and training accuracy, (b) testing and training loss for the CICIDS2019 dataset



Accuracy for testing and training

Loss for testing and training

Fig. 3. (a) Testing and training accuracy, (b) Testing and training loss for UNSW-NB15 dataset

4.4. PERFORMANCE EVALUATION ON THE CICIDS2019 DATASET

Various tests have been performed on the CICIDS2019 dataset to assess the effectiveness of the suggested approach. The multi-class classification result of the proposed approach is given in Table 1.

Table 1. Multi-class classification of the proposed approach on the CICIDS2019 dataset

Attack types	F1-score	Recall	Precision	Accuracy
Normal	99.88	99.89	99.87	99.92
DNS	99.25	99.27	99.24	99.29
NTP	99.33	99.4	99.38	99.41
NetBIOS	99.75	99.81	99.78	99.89
SYN	99.76	99.77	99.74	99.78
MSSQL	98.72	98.77	98.77	98.79
UDP	99	99.04	99	99.09
LDAP	99.18	99.24	99.2	99.25
SNMP	98.98	98.92	98.95	98.98
UDP-LAG	98.97	98.94	98.96	98.99

Table 1 shows that the proposed approach attains superior outcomes for all attacks on the CICIDS2019 dataset regarding f1-score, recall, precision, and accuracy. Particularly, Normal, NetBIOS, and SYN classes attain superior results with correspondingly 99.92%, 99.89%, and 99.78% accuracy. Moreover, the classification performance on UDP and LDAP provides the best performance. MSSQL detection performance is average compared to all the classes, with 98.79% accuracy, 98.77% precision, 98.77% recall, and 98.72% f1-score.

A graphical representation of Table 1 is shown in Fig. 4.

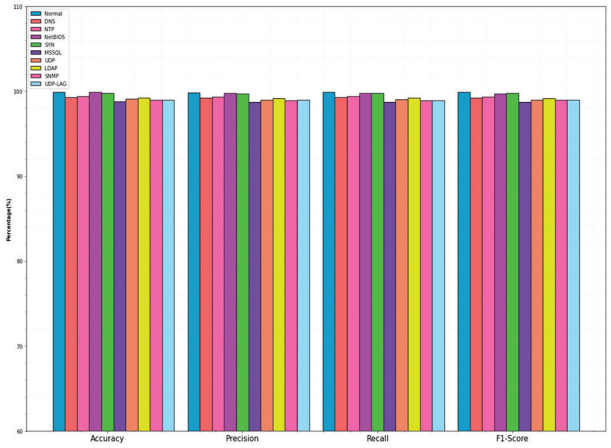


Fig. 4. Multi-class classification of the CICIDS2019 dataset

Table 2. Comparison of the proposed approach on the CICIDS2019 dataset

Technique	F1-score	Recall	Precision	Accuracy
Adaboost Regression [28]	-	96.74	95.93	95.87
XNN[29]	99	99.2	99	99.3
LSTM [30]	97.8	98	98.1	98
KNN [31]	97	97	96	98
Proposed	99.28	99.30	99.28	99.33

Table 1 shows the differentiation of the performance of the suggested work with other standard techniques tested under the CICIDS2019 dataset. The table shows

that the IDS model based on the proposed approach incurs the best results in terms of recall, accuracy and precision, and f1-score. Compared to all other methods, the performance of the Adaboost technique is poor (95.87% accuracy), and the XNN technique provides the best performance with 99% f1-score, 99.2% recall, 99% precision, and 99.3% accuracy. However, these are just as good as our proposed approach. The Long short-term memory (LSTM) and K-nearest Neighbor (KNN) perform similarly with 98% accuracy. A graphical representation of Table 1 is shown in Fig. 5.

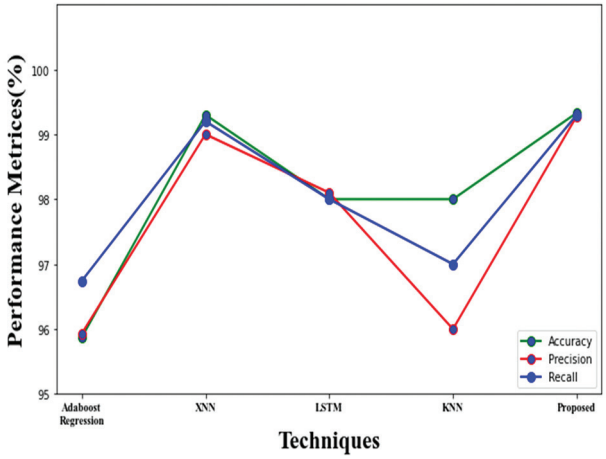


Fig. 5. Comparison of the proposed approach on the CICIDS2019 dataset

4.5. PERFORMANCE EVALUATION ON UNSW-NB15 DATASET

On the UNSW-NB15 dataset, numerous experiments are performed to analyze the effectiveness of the presented strategy. The multi-class classification result of the proposed approach is given in Table 3.

Table 3. Multi-class classification of the proposed approach on the UNSW-NB15 dataset

Attack types	F1-score	Recall	Precision	Accuracy
Normal	99.88	99.88	99.87	99.89
Analysis	99.79	99.76	99.78	99.81
Backdoors	99.64	99.63	99.61	99.69
DoS	99.18	99.19	99.16	99.21
Exploits	98.93	98.94	98.91	99
Fuzzers	99.15	99.11	99.08	99.12
Generic	99	99.02	99	99.04
Reconnaissance	99.04	99.06	99	99.1
Shellcode	99.73	99.74	99.71	99.79
Worms	99	99.09	99.04	99.12

Table 3 shows that the proposed approach's multi-classification performance is superior and achieves better values for all the attack classes. All the classes attain above 99% accuracy for all the classes. Specifically, the proposed approach classifies normal, Analysis, and Shellcode with 99.89%, 99.81%, and 99.79% accuracy. Compared to all attacks, the classification performance on the proposed approach of Reconnaissance and Exploits is average, with 99% accuracy. These values are the best. However, compared to all other classes, these values are shallow. The graphical representation of this table is presented in Fig. 6.

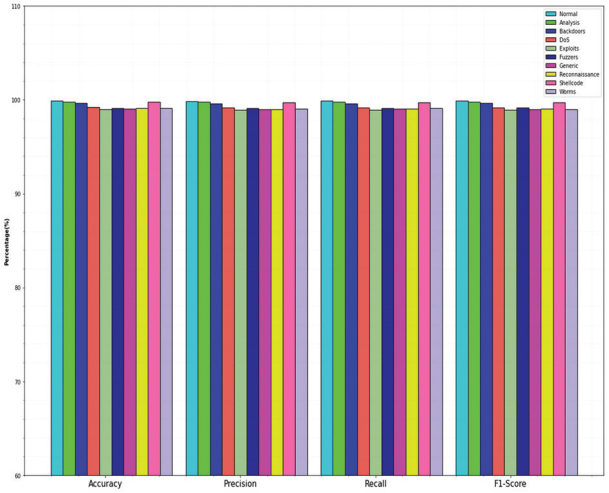


Fig. 6. Multi-class classification of the UNSW-NB15 dataset

Table 4. Comparison of the proposed approach on the UNSW-NB15 dataset

Technique	F1-score	Recall	Precision	Accuracy
SVM-ANN [32]	87.01	-	96.89	97.98
RLF-CNN [33]	-	89.3	-	88.7
ELM [34]	96.08	-	-	98.19
ANN[35]	-	-	-	97.89
Proposed	99.33	99.34	99.31	99.37

According to the residual blocks, a deep learning algorithm allows the construction of deeper networks to identify more critical network traffic characteristics. Our model outperforms existing deep learning techniques, as seen in Table 5 and Fig. 6. CNNs can perform better in network intrusion detection with residual learning. Due to issues with a class imbalance in the training set, none of the other models perform well in terms of minor classes. To overcome the abovementioned problem, our model uses a class imbalance strategy based on DCGAN. In all other existing techniques, the performance of RLF-CNN is mildly decreased due to its decreased weights in the loss function. Moreover, the performance of ANN (Artificial Neural Network) and SVM-ANN (Support vector machine-ANN) are similar. However, it is not more than the proposed approach. A graphical representation of the accuracy comparison is displayed in Fig. 7.

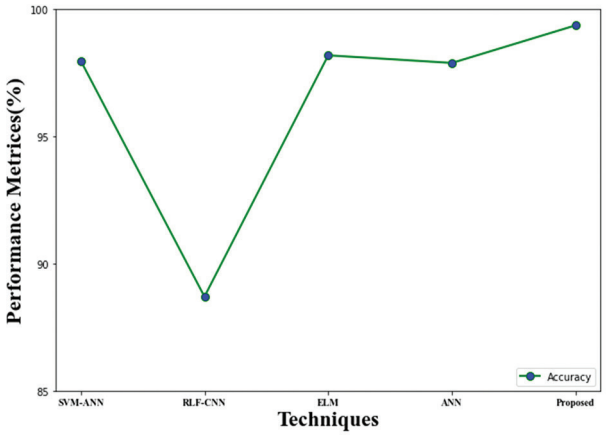


Fig. 7. Accuracy comparison of the proposed approach on the UNSW-NB15 dataset

4.6. COMPARISON OF DATA AUGMENTATION

The impact of data augmentation on the classification criteria for both datasets is illustrated in Fig. 8. The data augmentation technique enhances the classifier performance, as is shown in this graph. Using a DCGAN-based data augmentation strategy, the classifier in the CICIDS2019 dataset achieves 99.33% accuracy. It only accomplishes 99.12% without DCGAN. Using data augmentation approaches, the classifier for the UNSW-NB15 dataset achieves 99.37% accuracy. Compared to training samples without data augmentation, the number of training samples produced by this method is much higher.

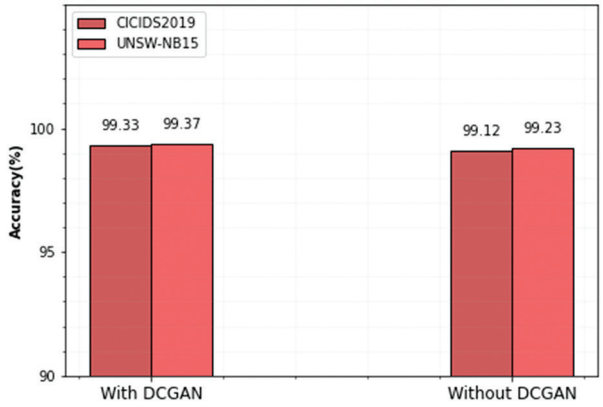


Fig. 8. Analysis of Data Augmentation Technique

In the absence of data augmentation, the same group of training instances is utilized for each epoch; however, when data augmentation is present, different training instances are generated for each epoch. Because of this, the proposed algorithm employs the DCGAN data augmentation technique to perform better and attain higher accuracy in both datasets.

5. CONCLUSION

This research presented the Deep learning-based intrusion detection system, which utilized a Resnet-50-based effective technique to extract features from the network data. The suggested IDS is validated using the UNSW-NB15 and CICIDS2019 datasets—a deep learning system built on the optimized Alexnet to identify the attacks effectively. The UNSW-NB15 and CICIDS2019 datasets had the highest accuracy, each at 99.37%. The suggested model performs admirably in a multi-class setting regarding f1-score, recall, precision, and accuracy measures. The research conducted here also aimed to offer guidance in selecting the optimum dataset for the model. The optimal dataset for the suggested model has been determined to be the UNSW-NB15 dataset. In the future, more assessment metrics will be conducted to analyze the system's efficacy with less time and resource usage.

Conflict of interests: 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.

Acknowledgments

We declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

Availability of data and material: Not applicable

Code availability: Not applicable

Authors' contributions: The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Ethics approval: This material is the author's original work, which has yet to be previously published elsewhere. The paper reflects the author's research and analysis truthfully and completely.

6. REFERENCES

[1] Z. Tang, H. Hu, C. Xu, "A federated learning method for network intrusion detection", Concurrency and Computation: Practice and Experience, Vol. 34, No. 10, 2022, p. e6812.

[2] B. Cao, C. Li, Y. Song, Y. Qin, C. Chen, "Network Intrusion Detection Model Based on CNN and GRU", Applied Sciences, Vol. 12, No. 9, 2022, p. 4184.

[3] M. Chalé, N. D. Bastian, "Generating realistic cyber data for training and evaluating machine learning classifiers for network intrusion detection systems", Expert Systems with Applications, Vol. 207, No.1, 2022, pp. 117936.

[4] J. Toldinas, A. Venčkauskas, R. Damaševičius, Š. Grigaliūnas, N. Morkevičius, E. Baranauskas, "A novel approach for network intrusion detection

using multistage deep learning image recognition", Electronics, Vol. 10, No. 15, 2021, p. 1854.

[5] G. D. C. Bertoli, L. A. P. Júnior, O. Saotome, A. L. Dos Santos, F. A. N. Verri, C. A. C. Marcondes, J. M. P. De Oliveira, "An end-to-end framework for machine learning-based network intrusion detection system", IEEE Access, Vol. 9, No. 1, 2021, pp. 106790-106805.

[6] H. Alavizadeh, H. Alavizadeh, J. Jang-Jaccard, "Deep Q-Learning based Reinforcement Learning Approach for Network Intrusion Detection", Computers, Vol. 11, No. 3, 2022, p. 41.

[7] H. Zhang, J. L. Li, X. M. Liu, C. Dong, "Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection", Future Generation Computer Systems, Vol. 122, No. 1, 2021, pp. 130-143.

[8] T. Kim, W. Pak, "Real-time network intrusion detection using deferred decision and hybrid classifier", Future Generation Computer Systems, Vol. 132, No. 1, 2022, pp. 51-66.

[9] J. Liu, Y. Gao, F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM", Computers & Security, Vol. 106, No. 1, 2021, p. 102289.

[10] A. Rosay, K. Riou, F. Carlier, P. Leroux, "Multilayer perceptron for network intrusion detection", Annals of Telecommunications, Vol. 77, No. 5, 2022, pp. 371-394.

[11] Z. Wang, Y. Zeng, Y. Liu, D. Li, "Deep belief network integrating improved kernel-based extreme learning machine for network intrusion detection", IEEE Access, Vol. 9, 2021, pp. 16062-16091.

[12] A. Ghorbani, S. M. Fakhrahmad, "A Deep Learning Approach to Network Intrusion Detection Using a Proposed Supervised Sparse Auto-encoder and SVM", Iranian Journal of Science and Technology, Transactions of Electrical Engineering, Vol.1, No. 1, 2022, pp. 1-18.

[13] R. Atefinia, M. Ahmadi, "Network intrusion detection using multi-architectural modular deep neural network", The Journal of Supercomputing, Vol. 77, No. 4, 2021, pp. 3571-3593.

[14] E. A. Shams, A. Rizaner, A. H. Ulusoy, "A novel context-aware feature extraction method for convolutional neural network-based intrusion detection systems", Neural Computing and Applications, Vol. 33, No. 20, 2021, pp. 13647-13665.

[15] S. Rawat, A. Srinivasan, V. Ravi, U. Ghosh, "Intrusion detection systems using classical machine learning techniques vs. integrated unsupervised feature learning and deep neural network", Internet Technology Letters, Vol. 5, No. 1, 2022, p. e232.

[16] H. Han, H. Kim, Y. Kim, "An Efficient Hyperparameter Control Method for a Network Intrusion Detection System Based on Proximal Policy Optimization", Symmetry, Vol. 14, No. 1, 2022, p. 161.

[17] X. Huang, "Network intrusion detection based on an improved long-short-term memory model in combination with multiple spatiotemporal structures", Wireless Communications and Mobile Computing, Vol. 10, No. 1, 2021, pp. 12-34.

[18] C. M. Hsu, M. Z. Azhari, H. Y. Hsieh, S. W. Prakosa, J. S. Leu, "Robust network intrusion detection scheme using long-short term memory based convolutional neural networks", Mobile Networks and Applications, Vol. 26, No. 3, 2021, pp. 1137-1144.

[19] Y. N. Kunang, S. Nurmaini, D. Stiawan, B. Y. Suprapto, "Attack classification of an intrusion detection system using deep learning and hyperparameter optimization", Journal of Information Security and Applications, Vol. 58, No. 1, 2021, p. 102804.

[20] L. Yu, J. Dong, L. Chen, M. Li, B. Xu, Z. Li, C. Zhang, "PBCNN: packet bytes-based convolutional neural network for network intrusion detection", Computer Networks, Vol. 194, No. 1, 2021, p. 108117.

[21] D. Alghazzawi, O. Bamasag, H. Ullah, M. Z. Asghar, "Efficient detection of DDoS attacks using a hybrid deep learning model with improved feature selection", Applied Sciences, Vol. 11, No. 24, 2021, p. 11634.

[22] M. Aamir, S. M. A. Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification", Journal of King Saud University-Computer and Information Sciences, Vol. 33, No. 4, 2021, pp. 436-446.

[23] R. Panigrahi, S. Borah, A. K. Bhoi, M. F. Ijaz, M. Pramanik, Y. Kumar, R. H. Jhaveri, "A consolidated decision tree-based intrusion detection system

for binary and multi-class imbalanced datasets", Mathematics, Vol. 9, No. 7, 2021, p. 751.

[24] M. A. Khan, "HCRNNIDS: hybrid convolutional recurrent neural network-based network intrusion detection system", Processes, Vol. 9, No. 5, 2021, p. 834.

[25] Y. Fu, Y. Du, Z. Cao, Q. Li, W. Xiang, "A Deep Learning Model for Network Intrusion Detection with Imbalanced Data", Electronics, Vol. 11, No. 6, 2022, p. 898.

[26] S. N. Mighan, M. Kahani, "A novel scalable intrusion detection system based on deep learning", International Journal of Information Security, Vol. 20, No. 3, 2021, pp. 387-403.

[27] S. Moualla, K. Khorzom, A. Jafar, "Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset", Computational Intelligence and Neuroscience, Vol.11, No.1, 2021, pp. 23-56.

[28] B. M. M. AlShahrani, "Classification of cyber-attack using Adaboost regression classifier and securing the network", Turkish Journal of Computer and Mathematics Education, Vol. 12, No. 10, 2021, pp. 1215-1223.

[29] S. Aziz, M. T. Faiz, A. M. Adeniyi, K. H. Loo, K. N. Hasan, L. Xu, M. Irshad, "Anomaly Detection in the Internet of Vehicular Networks Using Explainable Neural Networks (xNN)", Mathematics, Vol. 10, No. 8, 2022, p. 1267.

[30] V. Q. Ta, M. Park, "MAN-EDoS: A Multihead Attention Network for the Detection of Economic Denial of Sustainability Attacks", Electronics, Vol. 10, No. 20, 2021, p. 2500.

[31] S. Mohammed, "A Machine Learning-Based Intrusion Detection of DDoS Attack on IoT Devices", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 10, No. 4, 2021, pp. 12-45.

[32] P. Rana, I. Batra, A. Malik, A. L. Imoize, Y. Kim, S. K. Pani, S. Rho, "Intrusion Detection Systems in Cloud Computing Paradigm: Analysis and Overview", Complexity, Vol. 11, No. 2, 2022, pp. 34-67.

[33] J. Man, G. Sun, "A residual learning-based network intrusion detection system", Security and Communication Networks, Vol. 18, No. 1, 2021, pp. 56-89.

[34] S. Moualla, K. Khorzom, A. Jafar, "Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset", Computational Intelligence and Neuroscience, Vol. 2, No. 4, 2021, pp. 67-89.

[35] A. M. Aleesa, M. Younis, A. A. Mohammed, N. M. Sahar, "Deep-intrusion detection system with enhanced UNSW-NB15 dataset based on deep learning techniques", Journal of Engineering Science and Technology, Vol. 16, No. 1, 2021, pp. 711-727.

An Intelligent Server load balancing based on Multi-criteria decision-making in SDN

Original Scientific Paper

Vani K. A.
Department of Information Science and Engineering,
Dayananda Sagar College of Engineering, Visveshwaraya Technological University, Bangalore, India
vanika-ise@dayanandasagar.edu

RamaMohanBabu K. N.
Department of Information Science and Engineering,
Dayananda Sagar College of Engineering, Visveshwaraya Technological University, Bangalore, India
ramamohanbabu-ise@dayanandasagar.edu

Abstract *–In an environment of rising internet usage, it is difficult to manage network traffic while maintaining a high quality of service. In highly trafficked networks, load balancers are crucial for ensuring the quality of service. Although different approaches to load-balancing have been proposed in traditional networks, some of them require manual reconfiguration of the device to accommodate new services due to a lack of programmability. These problems can be solved through the use of software-defined networks. This research paper presents a dynamic load-balancing algorithm for software-defined networks based on server response time and content mapping. The proposed technique dispatches requests to servers based on real-time server loads. This technique comprises three different modules, such as a request classification module, a server monitoring module, and an optimized dynamic load-balancing module using content-based routing. There are a variety of robust mathematical tools to address complex problems that have multiple objectives. Multi-Criteria Decision-Making is one of them. The performance of the proposed scheme has been validated by applying the Weighted Sum Method of the multi-criteria decision-making technique. The proposed method Server load balancing based on Multi-criteria Decision Making[SDLB-MCDM] is compared with different load-balancing schemes such as round robin, random, load-balancing scheme based on server response time [LBBSRT], and An SDN-aided mechanism for web load-balancing based on server statistics [SD-WLB]. The experimental results of SDLB-MCDM show a significant improvement of 58% when weights are equal and 50% when unequal weights are assigned to various QoS parameters in comparison with the ROUND ROBIN, RANDOM, LBBSRT and SD-WLB techniques.*

Keywords*: Quality of Service. Software-Defined Networks, Load-Balancing, Open Flow*

1. INTRODUCTION

Over the past few years, there has been a remarkable increase in services residing in modern data centers. Some of the critical components of data centres include different types of servers, storage systems, switches, routers, and application delivery controllers. The applications of data centres range from social networking, video streaming, web search, data storage, data processing and many more. With these growing applications, the frequency of communication between the nodes has increased to a greater extent.

Further, the users who access these applications expect greater QoS with a minimum response time from the application servers. The response time is the amount of time a service provider takes to respond to a request.

However, data centre operators must deal with the complexity of managing traffic both within and across data centers. This includes providing the necessary resources and establishing a connection, regardless of how they are hosted. On the other hand, network management and dynamic configuration using traditional networks impose a challenging task. The configuration of the network components in traditional networks is very laborious and time-consuming for the network operators. This is due to fixed functionalities of network components, vendor dependency and structural complexities and many more [1]. This architecture consists of a control plane, a data plane, and a management plane, as shown in Fig. 1. The control plane and the data plane are decoupled. The entire global view will be present in the controller, which acts as the brain of the network. The data plane is regarded as the forwarding plane that governs the flow rules laid out by the controller. The communication between SDN controllers and data plane elements is carried out via the Open Flow

protocol. This protocol enables flow-level programmability in software-defined networking, which may be used to program the network according to application QoS needs as well as network traffic conditions [2,3].

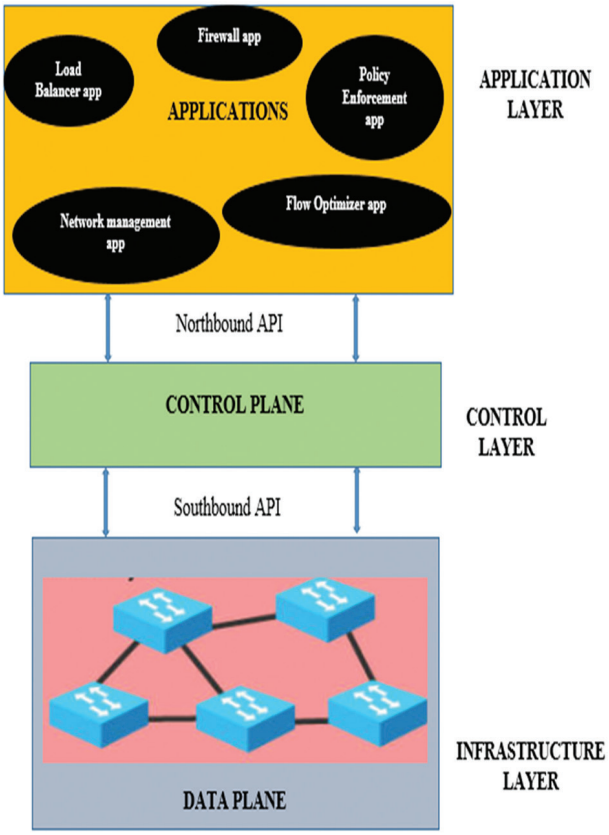


Fig. 1. SDN Architecture

During heavy traffic scenarios, deploying a dynamic load-balancing technique can aid in managing the network traffic more effectively. However, optimizing the response time while handling heavy network traffic and mapping the content is another challenging task. Though much research is carried out to address load-balancing in SDN, as discussed in [4], most of them perform load-balancing either based on the server's response time or content mapping. But this proposed research work takes both response time and content mapping into account while performing load-balancing in the server pool.

The major contribution of this research work is server load-balancing based on response time and content mapping, as well as optimization of routing rules and mathematical analysis using Weighted Sum Method [WSM] of the Multi-Criteria Decision-Making [MCDM] technique to select the best server in the server pool.

The paper is organized as follows: Section 2 discusses the related work and introduces an overview of software-defined networks along with strategies for load-balancing. Section 3 reviews the proposed model. Section 4 covers experimentation Section 5 covers evaluation and results. Section 6 concludes the work with a future scope.

2. RELATED WORK

There have been a number of studies on load- balancing in software-defined networks. Nevertheless, this research work focuses on providing a dynamic load-balancing solution in data center networks that is based on response time in SDN. The response time is one of the crucial aspects when we are evaluating the QoS of any network. Some of the research related to controller response times and server response times is discussed in this section.

2.1. LOAD BALANCING BASED ON THE CONTROLLER'S RESPONSE TIME

This section provides some of the research work related to load-balancing based on a controller's response time.

The authors in [5] proposed an SDN framework for load-balancing based on the controller's response time that makes use of network heterogeneity and context-aware vertical mobility concepts. This scheme designed a mechanism for load dissemination between controllers called reducing the overhead. The scheme studies the bandwidth requirement based on ongoing traffic, not the type of service requirement. The study by Senthil et al. aims to compare the performance of two load-balancing algorithms, flow-based load-balancing and traffic pattern-based load-balancing, using distributed controller architecture [6]. Authors in [7] provided a mathematical analysis of existing techniques in SDN and proposed the Response Surface Methodology to reduce the response time of a controller. While adding a new QoS policy to this scheme requires repetition and analysis to determine the QoS-related outcome.

To reduce the response time during load-balancing among the controllers, a two-phase dynamic controller clustering is proposed in [8]. According to the scheme, the optimal cluster size was not taken into consideration. The majority of research studies achieved load-balancing during heavy loads but did not achieve continuous load-balancing among the controllers. To address this issue, a new scheme named multiple threshold load-balancing (MTLB) switch migration scheme is proposed in [9]. Most of the research focused on the static assignment of controllers and switches. Due to this, some of the controller's response time was high. In order to reduce the controller's response time, a two-phase algorithm is proposed in [10].

2.2. LOAD-BALANCING BASED ON SERVERS RESPONSE TIME

In conventional networks, it was extremely challenging to take advantage of server reaction time due to hardware restrictions. Many academics have suggested load-balancing plans based on server response time in SDN to fill this need. In this section, several of these methods are covered. The authors of [11] recom-

mend load-balancing based on server response time. This scheme supports the same kind of data traffic as the other scheme.. The research work discussed in [12] performs server load-balancing based on switch port statistics. [13] Investigated how to maximize server utilization while minimizing response time in a cloud environment using SDN-based load- balancing. This scheme makes use of an application module and server pool. The type of service provided is classified as compute request or data request in this case. The authors of [14] discussed multiple server tests in demonstrating the quality of service with a limited number of servers to demonstrate the benefits of SDN in accessing servers. To assess performance, the scheme compared round-robin, random, and least-bandwidth algorithms. In order to exploit the dynamic performance of servers using SDN and to showcase the limitations of traditional networks, the authors in [15] have designed server load-balancing based on round-robin and weighted round-robin techniques using POX controller [16]. However, this technique attempts to address server load-balancing using the POX Controller.

For the efficient distribution of load among multiple servers based on bandwidth and round-robin fashion, the authors in [17] have proposed server load-balancing using SDN. This scheme compared the results of bandwidth-based and round-robin-based load- balancing and proved that the former yields better results in comparison with the round-robin technique. Based on the concept of server clustering that is widely used to provide availability and achieve high performance and scalability, the study in [18] proposed a novel dynamic weighted random selection load- balancing algorithm. This technique considers real-time server loads when assigning requests among the servers. This method works well in a single-controller architecture. The authors in [19] proposed a multiple regression-based search algorithm for selecting an optimal server with an optimal routing path. The scheme distributes the traffic to the server with the fewest connections and the lowest path cost from the floodlight controller. Further utilizing the concept of correlation analysis, this scheme predicts the response time based on the load and bandwidth.

This proposed method considers diverting the incoming requests to the appropriate server based on the type of traffic with optimized routing rules.

3. PROPOSED METHOD

The proposed system model is depicted in Fig. 2. The system is composed of clients and servers connected to a Ryu controller, along with a load balancer module.

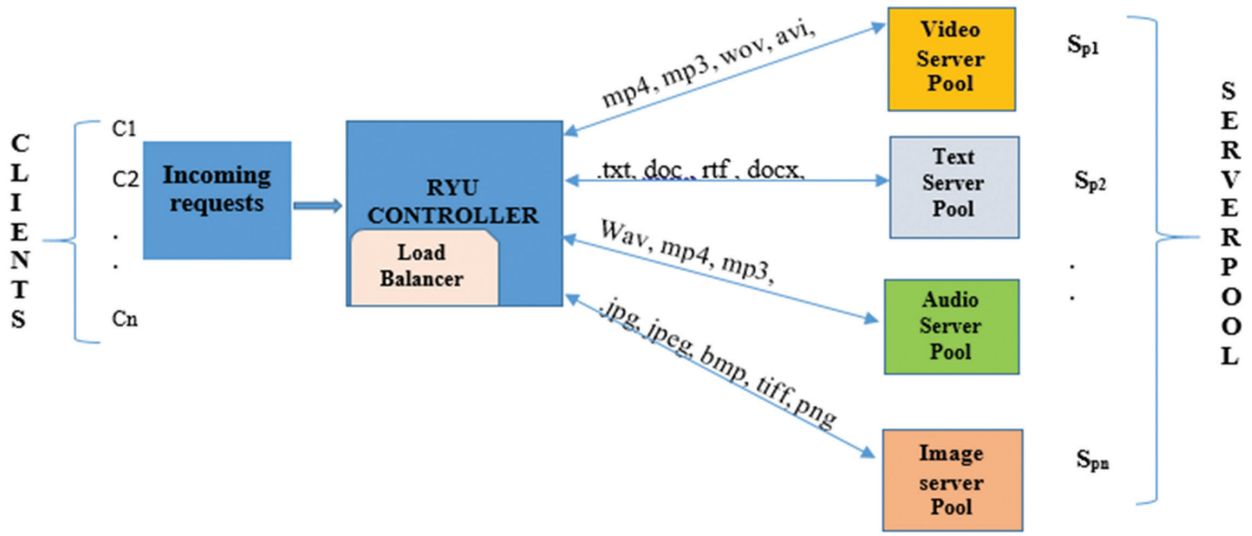


Fig. 2. System Architecture

The proposed model works on the principle of optimised routing rules laid out by the controller. This model is designed to support web services at different server pools. At each level, the controller directs the requests to the respective servers in the server pool based on the load balancer result for the required content type and response time. The different server pools are classified as video server pool, audio server pool, image server pool, and text server pool, respectively, as depicted in Fig. 2. This architecture consists of mainly three modules, namely: the request classification module, the server monitoring module, and the optimized dynamic load-balancing module. These three modules are discussed in detail below.

3.1 REQUEST CLASSIFICATION MODULE

The main idea behind creating this module is to classify the type of request based on its content. The model makes use of URL mapping instead of regular IP map-

ping. The request classification module is depicted in Fig. 3 below. Let us consider a scenario where the client requests a video by specifying it in the URL. e.g., myapp/switch/app/video. The request is sent to the controller via the OpenFlow Zodiac switch. The classification module determines the type of request, such as video, image, text, or audio. Once this information is extracted, it is sent to the load-balancing module.

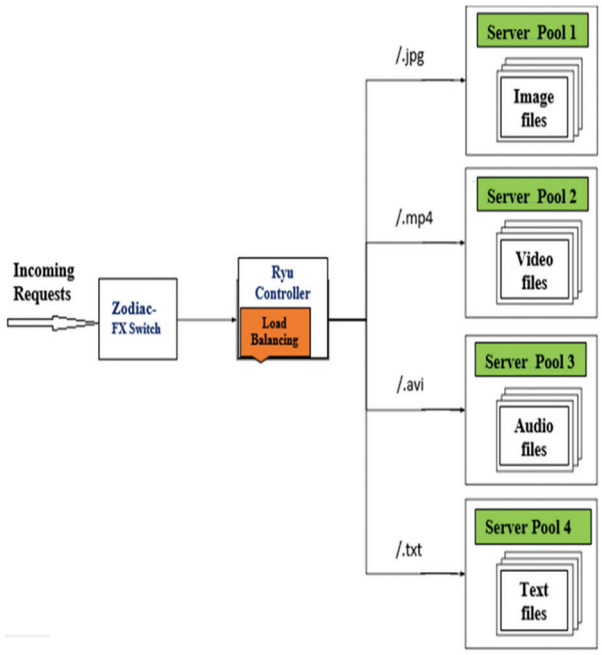


Fig. 3. Request classification module

The algorithm for request classification module is described in algorithm 1.

Algorithm 1: Request classification module

Input : Server Metrics

Output: Request classification

While*(true)*

Read (*Content-Type ='img')*

If *(Content-Type ='img')*

Send image data

If *(Content-Type ='video')*

Send video data

If *(Content-Type ='txt')*

Send text data

If *(Content-Type ='audio')*

Send audio data

End

3.2 SERVER MONITORING MODULE

This module is implemented in such a way that the servers in the server pool keep sending the load information to the controller every 5 milliseconds [ms]. The algorithm for this module is described in Algorithm 2. The severity metrics, like CPU utilization, memory utilization, requests per second, time per request, transfer rate, waiting time, and many more, are sent to the controller. The most interesting part of this module is the response time of the server. The response time of each server in different server pools is collected via this module based on real statistics.

Algorithm 2: Server Monitoring

Input: Server metrics

Output: Server monitoring

Start

While *(true)*

start the servers

if*(time='T' ms)*

start the server script for sending metrics

for each*(T=5 ms)*

Send metrics like CPU utilization, Memory Utilization, requests_per_second, time_per_request, transfer_rate, waiting_time to the controller

time.sleep (*INTERVAL_SECONDS*)

end

3.3 OPTIMIZED DYNAMIC LOAD-BALANCING MODULE USING CONTENT-BASED ROUTING

This module implements dynamic load-balancing using content-based routing. Upon the arrival of the client's request, the content is parsed by the load-balancing module in the controller, which runs algorithm 3 to find the server with the least response time in each server pool, and the controller installs the flow based on the requested content and the server with the least response time. Based on the content, for example, if the request pertains to images, it will be forwarded to the image server pool; similarly, if the request is to retrieve video, it will be forwarded to the server that handles video; the same holds true for text and audio files.

Algorithm 3: Optimized Load-Balancing module based on the requested content

Input: server metrics

Output: Best server $[B_S]$ with fast response time

Start the RYU controller

while *(true)*

Initially B_s*=null*

if (*time='T' ms*)

Collect server metrics and run the optimized load balancer module

Initialize Load-balancing module to Read the content of the request

if*(Content Type= 'Img")*

send the request to image server queue

if*(Content Type= 'video")*
send the request to video server queue
if*(Content Type= 'text")*
send the request to text server queue
if*(Content Type= 'audio")*
Send the request to audio server queue
Calculate the server with least response time[Rt]
Forward the requested content to the server with minimum response time [Rt], according to equation 2
end

The response time [Rt_s] and the average response time [$ARTs$] of each server are calculated as given in equations [1] and [2], respectively.

$$Rt_s = \sum_{i=1}^{n} X_i \quad (1)$$

Where

$$Xi = (N_{s1} * R_{s1}) + (N_{s2} * R_{s2}) + \cdots + (N_{sn} * R_{sn})$$
$$ART_S = \sum_{i=1}^{n} \left(\frac{X_i}{T_{nr}}\right) \quad (2)$$

Here, 'X_i' represents the response time of each server serving 'n' number of requests. 'T_{nr}' represents the total number of requests. 'N_{s1}' is the number of requests served by server 1, and 'R_{s1}' is the response time of server 1 serving the required content. Similarly, 'N_{s2}' is the number of requests served by server 2, and 'R_{s2}' is the response time of server 2. The requests served by the nth server are represented by Nsn, and the response time of the nth server is represented by R_{sn}.

4. EXPERIMENT AND RESULTS

The experiment setup consists of a controller, an OpenFlow switch, a pool of web servers, and various client machines. The experimental testbed is as shown in Fig. 4, the experiment is carried out in data centre network where a number of clients and various web servers, such as Apache 2, Ngnix, and SimpleHTTPServer, are connected to the RYU controller via a real-time Zodiac-fx switch. The load balancer module is placed within the RYU controller.

Initially, the experiment was carried out to perform load-balancing based on various techniques such as round robin, random, LBSSRT, SD-WLB, and SDLB-MCDM. The single-objective optimization and analysis approach is no longer widely used due to the increasing complexity and multiplicity of the load-balancing problem. Due to the fact that perfect load-balancing is driven by multiple dimensions, a good decision-maker may look into various parameters, such as non-economical or economical, that can be compromised in certain situations. The experiment is formulated using the multi-criteria decision-making [MCDM] mathematical model to find a suitable solution for the load-balancing problems involving multiple and conflicting objectives. This model works on the basic principle of the weighted sum method [WSM], i.e., the rank of the best load-balancing technique is evaluated based on the WSM of the MCDM technique [20-25].

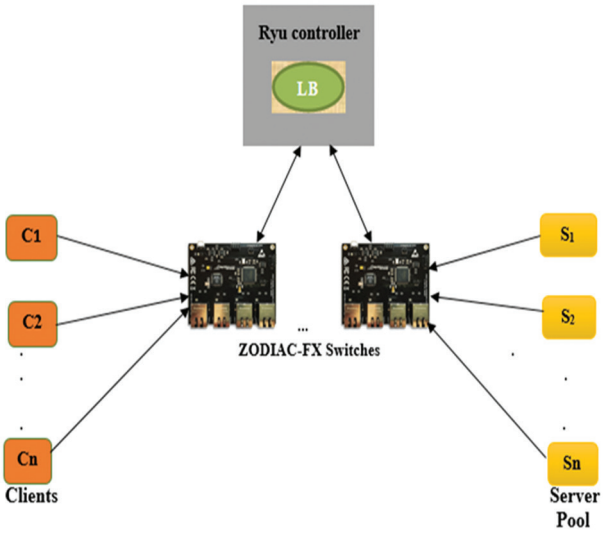


Fig. 4. Experimental setup

This technique takes into account various parameters and values, along with criteria.The criteria column represents the various methods used for evaluation, such as round robin, random, LBBSRT, SD-WLB and the proposed method SDLB-MCDM. The parameters to be considered are outlined in Table 1 below:

Table 1. Parameters used

Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time
Random	0.875 s	2496.68 kbps	3.357ms	297.85/s	4
Round robin	0.888 s	2246.57 kbps	3.223 ms	290.06/s	3
LBBSRT	0.723 s	3445.85 kbps	2.452 ms	312.14/s	2
SD-WLB	0.678 s	3876.45 kbps	2.126 ms	366.31/s	2
SDLB-MCOM	0.065 s	6687.23 kbps	0.157 ms	543.67/s	1

When we look at the measuring units of each of these parameters, they are different. In order to resolve this issue, the weighted sum method is used. The steps of experimentation using the WSM-MCDM technique are as follows:

Step 1: Construct a conversion scale that ranges from low to excellent as shown below in Table 2.

Table 2. Conversion scale

Low	1
Below average	2
Average	3
Good	4
Excellent	5

Step 2: To obtain the decision matrix, assume that the decision maker has determined the importance (or measure of performance) of alternative Ai based on criterion Cj (for i = 1,2,3,..., M and j = 1,2,3,...,N), and W_i represents the weights assigned, as shown in Table 3.

Table 3. Decision matrix

Criteria

Alternative	C_1	C_2	C_3	...	C_n
	W_1	W_2	W_3	...	W_n
a_1	a_{11}	a_{12}	a_{13}	...	a_{1n}
a_2	a_{21}	a_{22}	a_{23}	...	a_{2n}
a_3	a_{31}	a_{32}	a_{33}	...	a_{3n}
⋮	⋮	⋮	⋮	⋮	⋮
a_m	a_{m1}	a_{m2}	a_{m3}	...	a_{mn}

Here, a decision-makers primary objective is to select the best alternative or rank all possible alternatives. After considering all of the decision criteria, Pi (for i = 1, 2, 3 …M) represents the final preference for alternative Ai. We can calculate the preference Pi for alternative A_i (i = 1, 2, 3…M) using the formula below [26-29].

$$Pi = \sum_{j=1}^{N} A_{ij} Wj \quad (3)$$

(for i=1, 2,3… M)

Step 3: The next step is to categorize the parameters as beneficial or costly. The beneficial parameters are the ones whose higher values are preferred, and the costly parameters are the ones whose lower values are preferred [30]. Accordingly, the table is categorized by parameters as shown below in Table 4.

Table 4. Parameter categorizing table

	Costly	**Beneficial**	**Beneficial**	**Beneficial**	**Costly**
Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time
Random	0.875 s	2496.68 kbps	3.357 ms	297.85 /s	4
Round robin	0888s	2246.57 kbps	3.223 ms	290.06 /s	3
LBBSRT	0.723 s	3445.85 kbps	2.452 ms	312.14 /s	2
SD-WLB	0.678 s	3876.45 kbps	2.126 ms	366.31 /s	2
SDLB-MCDM	0.065 s	6687.23 kbps	0.157 ms	543.67 /s	1

Step 4: Further, the table needs normalization. In order to normalize the following expressions are used.

$$Costly = \max(a_{ij}) / a_{ijj} \quad (4)$$

$$Costly = \max(a_{ij}) / a_{ijj} \quad (5)$$

Step 5: Applying the expression in equations (4) and (5), the table is normalized as shown below in Table 5 below.

Table 5. Normalized values

	Costly	**Beneficial**	**Beneficial**	**Beneficial**	**Costly**
Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time
Random	0.875	2496.68	3.357	297.85	4
Round robin	0.888	2246.57	3.223	290.06	3
LI313SRT	0.723	3445.85	2 452	312.14	2
SD-WLB	0.678	387645	2.126	366.31	2
SDLB-MCDM	0.065	6687.23	0.157	543.67	1

Step 6: The next step is to obtain a weighted normalized matrix by adding weights to all these criteria. Here the proposed technique is evaluated for both equal and unequal weights for all the criteria, as shown below in Table 6 below.

Weightage	**20%**	**20%**	**20%**	**20%**	**20%**
Normalization	Costly	Beneficial	Beneficial	Beneficial	Costly
Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time
Random	0.875	2496.68	3.357	297.85	4
Round robin	0.888	2246.57	3.223	290.06	3
LBBSRT	0.723	3445.85	2.452	312.14	2
SD-WLB	0.678	3876.45	2.126	366.31	2
SOLB-MCDM	0.065	6687.23	0.157	543.67	1

Step 7: The next step is to obtain the performance matrix to select the best among the given alternatives, as shown in Table 7 below.

Table 7. Performance Matrix for equal weights

Weightage	**20%**	**20%**	**20%**	**20%**	**20%**
Normalization	Costly	Beneficial	Beneficial	Beneficial	Costly
Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time
Random	0.014857	0.05	0.07467	0.2	0.10957
Round robin	0.01464	0.066667	0.06719	0.192017	0.106704
LBBSRT	0.017981	0.1	0.103058	0.146083	0.114827
SD-WLB	0.019174	0.1	0.115936	0.126661	0.134755
SDLB-MCDM	0.2	0.2	0.2	0.009354	0.2

Step 8: Obtain the performance ranking matrix as shown in Table 8.

From the final performance table, it is seen that the proposed method SDLB-MCDM stands out best among all the other techniques such as random, round robin, LBBSRT, and SD-WLB. The results are discussed in the next section.

Table 8. Ranking Matrix for equal weights

Weightage	20%	20%	20%	20%	20%	
Normalization	Costly	Beneficial	Beneficial	Beneficial	Costly	RANK
Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time	
Random			0.449097			4
Round robin			0.447217			5
LBBSRT			0.481948			3
SD-WLB			0.496525			2
SDLB-MCDM			0.809354			1

Step 9: The performance matrix for unequal weights are shown in Table 9 below.

Table 9. Performance Matrix for unequal weights

Weightage	25%	20%	10%	20%	25%
Normalization	Costly	Beneficial	Beneficial	Beneficial	Costly
Criteria	Average Response time	Transfer rate	Time per renuest	Request per second	Waiting time
Random	0.018571	0.0625	0.07467	0.1	0.10957
Round robin	0.0183	0.083333	0.06719	0.096008	0.106704
LBBSRT	0.022476	0.125	0.103058	0.073041	0.114827
SD-WLB	0.023968	0.125	0.115936	0.06333	0.134755
SDLB-MCDM	0.25	0.25	0.2	0.004677	0.2

Step 10: Obtain the final performance ranking matrix as shown below in Table 10

Table 10. Performance ranking

Weightage	20%	20%	20%	20%	20%	
Normalization	Costly	Beneficial	Beneficial	Beneficial	Costly	RANK
Criteria	Average Response time	Transfer rate	Time per request	Request per second	Waiting time	
Random			0.365312			4
Round robin			0.371536			5
LBBSRT			0.438402			3
SD-WLB			0.462988			2
SDLB-MCDM			0.904677			1

From the final performance table, it is seen that the proposed method SDLB-MCDM stands out best among all the other techniques such as random, round robin, LBBSRT, and SD-WLB. The results are discussed in the next section.

5. RESULTS

An analysis of the results obtained using the real experimental setup implemented using an OpenFlow environment is presented in this section. The experimental setup included a Ryu controller and Zodiac-FX switch, as well as web servers such as Apache 2, Ngnix, and SimpleHTTPServer, and a set of client machines installed with Ubuntu 20.0. The steps are configured as follows:

The hosts are configured to use services such as image data, video data, audio data, and text data. Apache Bench is used to generate the traffic. Here different metrics such as average response time, transfer rate, time-per-request, requests-per-second, and waiting time are considered for the performance evaluation of SDLB-MCDM. The comparison of SDLB-MCDM with different techniques like random, round robin, LBBSRT, and SD-WLB is considered. In this experiment, the SDLB-MCDM module runs on a Ryu controller that runs three different algorithms: the request classification module, the server monitoring module, and the optimized dynamic load-balancing module using content-based routing. Averaging ten experiments yielded the reported results. The proposed mechanism, SDLB-MCDM, shows better performance in comparison with other techniques, and this mechanism can be used in many data center environments.

The graphs shown in Fig. 5 illustrate the average response time of different schemes like round robin, random, LBBSRT, SD-WLB, and SDLB-MCDM. The proposed scheme (SDLB-MCDM) shows better performance in comparison with other techniques.

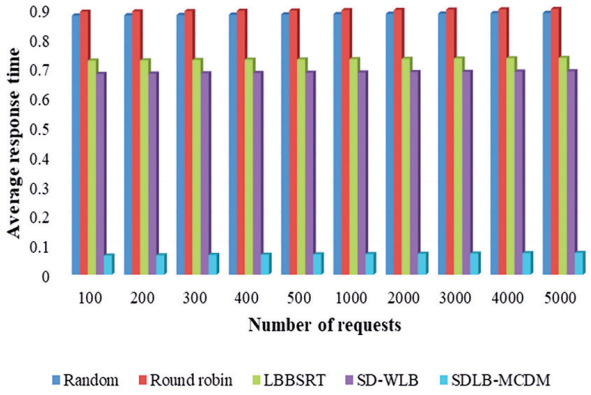


Fig. 5. Average Response Time

Fig. 6 depicts the transfer rate, which indicates that the proposed technique performs better at transferring more data in comparison with other techniques.

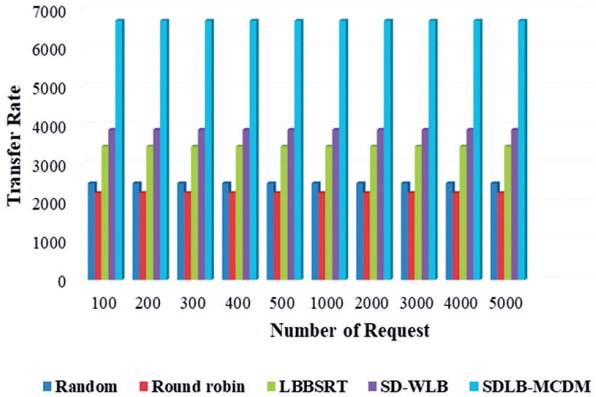


Fig. 6. Transfer Rate

The time-per-request is depicted in Fig. 7, which indicates the proposed [SDLB-MCDM] technique takes very little time to serve the request.

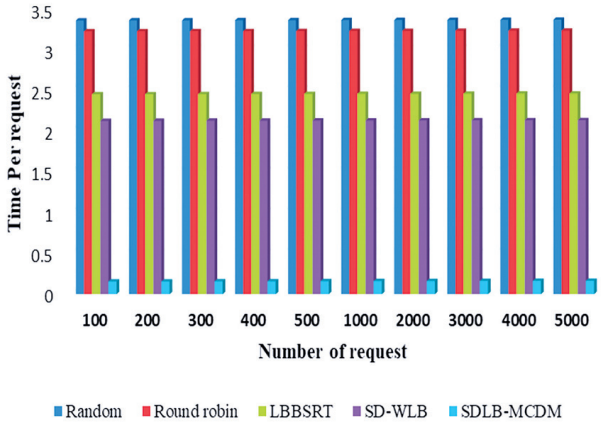


Fig. 7. Time per request

The results in Fig. 8 clearly indicate that the proposed method [SDLB-MCDM] serves a greater number of requests per second in comparison with other techniques.

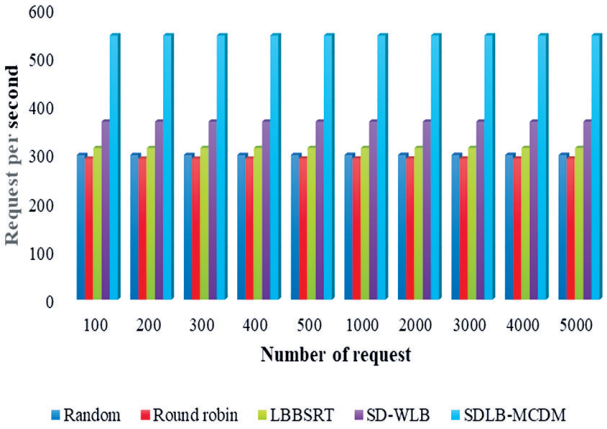


Fig. 8. Request per second

It is very important for any method to have a short waiting period that indicates a very small number of requests are waiting in the queue.

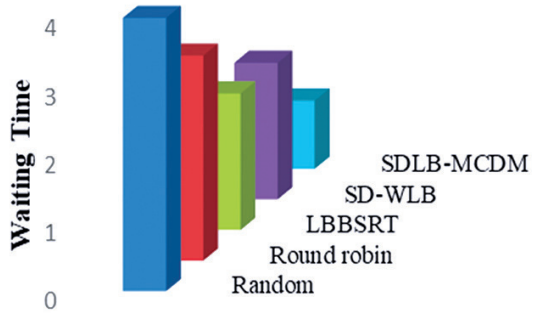


Fig. 9. Waiting Time

The results shown in Fig. 9 clearly indicate that the proposed method has a very low number of outstanding requests.

6. CONCLUSION

This proposed research work presents an optimized load-balancing in a software-defined network based on a multi-criteria decision-making technique [SDLB-MCDM]. The SDLB-MCDM method proposes three algorithms based on response time and content mapping to choose the best server among the pool of servers. In order to appreciate the efficacy and feasibility of the proposed technique, different parameters are considered for decision-making rather than a single parameter, which makes it more efficient in comparison with other techniques. The proposed technique makes use of WSM and the MCDM method to determine the load-balancing technique. The experimental results show a 58% improvement in the performance of the proposed method when equal weights are assigned. The results show a 50% improvement in the performance of SDLB-MCDM when unequal weights are assigned. The performance results under both equal and unequal weights show better performance in comparison with round robin, random, LBBSRT, and SD-WLB techniques.

The proposed SDLB-MCDM method can be adopted in data centre networks where load-balancing among many virtual machines is a major challenge. The future scope of this research work can be tested in a heterogeneous environment with different servers.

7. REFERENCES

[1] C. H. Benet, "Traffic Management in Software-Defined Data Center Networks", Faculty of Health, Science and Technology, Department of Mathematics and Computer Science, Karlstad University, PhD Thesis ,2021.

[2] M. M. Tajiki, B. Akbari, N. Mokari, "Optimal Qos-aware network reconfiguration in software defined cloud data centers", Computer. Networks, Vol. 120, 2021 pp. 71-86.

[3] K. Benzekki, A. El Fergougui, A. Elbelrhiti Elalaoui, "Software-defined networking (SDN): a survey", Secure Communication Networks, Vol. 9, No. 18, 2016, pp. 5803-5833.

[4] L. Li, Q. Xu, "'Load balancing researches in SDN: A survey", Proceedings of the 7th IEEE International Conference on Electronics Information and Emergency Communication, Shenzhen, China, 21-23 July 2017, pp. 403-408.

[5] M. Alotaibi, A. Nayak, "Linking handover delay to load balancing in SDN-based heterogeneous networks", Computer. Communication., Vol. 173, 2021, pp. 170-182.

[6] S. Prabakaran, R. Ramar, Software-defined network: Load balancing algorithm design and analysis", The International Arab Journal of Information Technology, Vol. 18, No. 3, 2021, pp. 312-318.

[7] A. Moravejosharieh, S. Ahmad, K. Ahmadi, "Shortening the response time in software-defined networking: A sensitivity analysis approach", Proceedings of the IEEE 7th International Conference on Communications and Electronics, Hue City, Vietnam 18-20 July 2018, pp. 90-95.

[8] H. Sufiev, Y. Haddad, L. Barenboim, J. Soler, "Dynamic SDN controller load balancing", Future Internet, Vol. 11, No. 3, 2019, pp. 1-21.

[9] H. Mokhtar, X. Di, Y. Zhou, A. Hassan, Z. Ma, S. Musa, "Multiple-level threshold load balancing in distributed SDN controllers", Computer Networks, Vol. 198, 2021, pp. 108-369.

[10] T. Wang, F. Liu, J. Guo and H. Xu, "Dynamic SDN controller assignment in data center networks: Stable matching with transfers", Proceedings of the 35th Annual IEEE International Conference on Computer Communications, San Francisco, CA, USA, 2016, pp. 1-9.

[11] H. Zhong, Y. Fang, J. Cui, "Reprint of 'LBBSRT: An efficient SDN load balancing scheme based on server response time", Future Generation Computer Systems., Vol. 80, 2018, pp. 409-416.

[12] K. Soleimanzadeh, M. Ahmadi, M. Nassiri, "SD-WLB: An SDN-aided mechanism for web load balancing based on server statistics", ETRI Journal, Vol. 41, No. 2, 2019, pp. 197-206.

[13] A. A. Abdelaziz, E. Ahmed, A. T. Fong, A. Gani, M. Imran, "SDN-Based load balancing service for cloud servers", IEEE Communication Magazine, Vol. 56, No. 8, 2018, pp. 106-111.

[14] J. V. O. Farias, E. F. Coutinho, C. I. M. Bezerra, "Applying load balancing algorithms for multiple access management on software-defined networking servers", Proceedings of the 10th Euro-American Conference on Telematics and Information Systems, New York, NY, USA, November 2020 ,pp. 1-5.

[15] A. S. Linn, S. H. Win, S. T. Win, "Server Load Balancing in Software Defined Networking", National Journal of Parallel and Soft Computing, Vol. 1, No. 1, 2019, pp. 261-265.

[16] S. Kaur, K. Kumar, J. Singh, N. S. Ghumman, "Round-robin based load balancing in Software Defined Networking," Proceedings of the 2nd International Conference on Computing for Sustainable Global Development, New Delhi, India, 2015, pp. 2136-2139.

[17] A. K. Arahunashi, G. G. Vaidya, K. V. Reddy, "Implementation of Server Load Balancing Techniques Using Software-Defined Networking", Proceedings of the 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, Bengaluru, India, 2018, pp. 87-90.

[18] M. L. Chiang, H. S. Cheng, H. Y. Liu, C. Y. Chiang, "SDN-based server clusters with dynamic load balancing and performance improvement", Cluster Computing, Vol. 24, No. 1, 2021, pp. 537-558.

[19] G. S. Begam, M. Sangeetha, N. R. Shanker, "Load Balancing in DCN Servers through SDN Machine Learning Algorithm", Arabian Journal for Science and Engineering, Vol. 47, No. 2, 2022, pp. 1423-1434.

[20] M. Cinelli, M. Kadziński, G. Miebs, M. Gonzalez, R. Słowiński, "Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system", European Journal of Operational Research, Vol. 302, No. 2, 2022, pp. 633-651.

[21] B. Sarkar, "Fuzzy decision making and its applications in cotton fibre grading", Soft Computing in Textile Engineering, 2011, pp. 353-383..

[22] S. G. Ozcan, M. Sayit, "Improving the QoE of DASH over SDN: A MCDM Method with an Intelligent Approach", Proceedings of the 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops, Paris, France, 2019, pp. 100-105.

[23] S. Chakraborty, T. K. Jana, S. Paul, "On the application of multi criteria decision making technique for multi-response optimization of metal cutting process", Intelligent Decision Technologies, Vol. 13, No. 1, 2019, pp. 101-115.

[24] J. Gyani, A. Ahmed, M. A. Haq, "MCDM and Various Prioritization Methods in AHP for CSS: A Comprehensive Review", IEEE Access, Vol. 10, 2022, pp. 33492-33511.

[25] E. Triantaphyllou, A. Sánchez, "A sensitivity analysis approach for some deterministic multi-criteria

decision-making methods", Decision Sciences., Vol. 28, No. 1, 1997, pp. 151-194.

[26] P. C. Fishburn, "Letter to the Editor—Additive Utilities with Incomplete Product Sets: Application to Priorities and Assignments", Operations Research, Vol. 15, No. 3, 1967, pp. 537-542.

[27] S. Das, A. Ghosh, "A Fuzzy Multi-Criteria Decision-Making Approach for Grading of Raw Jute", Journal of Natural Fibers, Vol. 18, No. 5, 2021, pp. 685-693.

[28] A. Hussain, J. Chun, M. Khan, "A novel multicriteria decision making (MCDM) approach for precise decision making under a fuzzy environment", Soft Computing, Vol. 25, No. 7, 2021, pp. 5645-5661.

[29] J. Ali and B. H. Roh, "A Novel Scheme for Controller Selection in Software-Defined Internet-of-Things (SD-IoT)", Sensors, Vol. 22, No. 9, 2022, p. 3591.

[30] J. Ali, B. H. Roh, S. Lee, "QoS improvement with an optimum controller selection for software-defined networks", PLoS ONE, Vol. 14, No. 5, 2019.

An Efficient Switch Migration Scheme for Load Balancing in Software Defined Networking

Original Scientific Paper

Thangaraj Ethilu
Annamalai University,
Department of Computer Science and Engineering, Tamilnadu, India.
ethilthangaraj@yahoo.co.in

AbiramiSathappan
Annamalai University
Department of Computer Science and Engineering, Tamilnadu, India.
reachabisv@gmail.com

Paul Rodrigues
King Khalid University
Department of Computer Engineering, Abha, Saudi Arabia
drpaulprof@gmail.com

Abstract *–Software-defined networking (SDN) provides increased flexibility to network management through distributed SDN control, and it has been a great breakthrough in network innovation. Switch migration is extensively used for workload balancing among distributed controllers. The time-sharing switch migration (TSSM) scheme proposes a strategy in which more than one controller is allowed to share the workload of a switch via time sharing during overloaded conditions, resulting in the mitigation of ping-pong controller difficulty, a reduced number of overload occurrences, and better controller efficiency. However, it has increased migration costs and higher controller resource consumption during the TSSM operation period because it requires more than one controller to perform. Therefore, we have proposed a strategy that optimizes the controller selection during the TSSM period based on flow characteristics through a greedy set coverage algorithm. The improved TSSM scheme provides reduced migration costs and lower controller resource consumption, as well as TSSM benefits. For its feasibility, the implementation of the proposed scheme is accomplished through an open network operating system. The experimental results show that the proposed improved TSSM scheme reduces the migration cost and lowers the controller resource consumption by about 36% and 34%, respectively, as compared with the conventional TSSM scheme.*

Keywords*: Quality of Service. Software-Defined Networks, Load-Balancing, Open Flow*

1. INTRODUCTION

The challenges in network management have tremendously increased due to the rapid deployment of cloud computing, big data applications, the internet of multimedia things, and increased data traffic. The traditional network architecture system combines a data plane and a control plane in each switch, with the former handling packet processing and the latter handling decision making and management. Therefore, updating the latest algorithms and new policies on the switches is very complex and time-consuming because all the switches involved in the given network need to be reconfigured one after another by system administrators or workers [1].

Currently, software-defined networking techniques create a unique view of network management in network applications where the control plane in the switches is shifted to a central unit known as the controller. Therefore, the controller can manage multiple switches in the network. In this modern approach, monitoring, and control of network switches are much simpler as compared with conventional network management techniques because the controller unit can provide such information about the switches. Furthermore, the latest algorithms and new control policies are easily updated to the switches via a set of rules in the controller [2]. Apart from this, SDN can support a wide range of applications, including (i) resisting cyber-attacks; (ii) identifying malicious access points; and (iii) providing anonymous authentication, etc. [3-7].

A lone controller in a large network will be a tough option since it creates a bottleneck in network management; therefore, distributed SDN control (DSC) is demanded in the network applications, and it acts as a promising solution in large network management with the several numbers of switches [8]. The DSC allows multiple controllers to coordinate with each other to manage the entire network. Where each controller is managing a subset of switches (i.e., a subnet), as well as workflow, these can be exchanged among controllers for the use of teamwork. Each controller involves distributing the workload for the subnets and reassigning its switches' workloads through the regular check-up of each subnet, called "controller placement" [9]. The placement of controllers is based mainly on load balancing, and it is applied through several techniques, including the workgroup control technique [10], the deep reinforcement learning technique [11], and so on. The outcome of such control techniques may widely alter the switches in the subnet and lead to instability in the subnet via ping-pong operation. Furthermore, controller placement techniques are not considered effective during short-lasting flows such as distributed denial of service and impulses [12].

Switch migration provides a smooth alteration of subnets with a lesser period and overcomes the above-mentioned issues. In each time frame (or time interval or period), a switch migration method is examining the workload status of each controller in the network to determine whether they are overloaded (busy) or lightly loaded (available to share other works). If it is overloaded, the migration method in a network relocates a switch from the busy controller subnet to a lightly loaded controller subnet. Most of the existing switch migration methods follow the smallest slice of the migration: one single switch, which is migrated at the beginning of the period. Once the switch is migrated, it remains in the latest subnet until the switch is selected for the next period. Most importantly, these migration methods always ask a controller to oversee one switch for a complete period. Therefore, the controller in these methods gets into the ping-pong difficulty of an "elephant flow situation (i.e., flow carries many packets) and goes into the serious trouble of a subnet that is unstable [13].

2. LITERATURE REVIEW

Over the years, several studies have detailed the various issues in the DSC network. Conventionally, controller load balancing is achieved through dynamic controller placement methods. Chan et al. [14] proposed a method that could minimize the service interruption time by smoothly transferring the workflow from the compromised controller to another controller. The leader controller redundancy is detailed in [15], where a lightly loaded controller can act as a leader in case of failure in the regular leader controller unit. Controller placement methods and challenges are reviewed in [9]. It has insisted that the controllers maintain fairness during the sharing of their workloads. Ref [16] proposed a reliable deployment method because of reducing packet loss and improving network stability, and it has achieved its objectives compared to other controller placement methods. Kim et al. [17] have proposed a method that improves the output of dispersed data stores in an Open Daylight controller cluster by consistently spreading the shared leaders to the cluster members. Ref [18] proposed a method in which controllers collaborate to reroute traffic to avoid congestion during a switch's busy or overloaded period. A software-defined cyber-seek framework is proposed in [19], where a hybrid controller is used for cloudlets and local networks. Prediction-based controllers are proposed in [20], and they predict the network load and perform the device transfer based on the prediction. The controller placement studies like the workgroup control technique and the deep reinforcement learning technique proposed in [10] and [11], respectively, show that these techniques are not effective during impulses, distributed denial of service, etc. Apart from the dynamic controller placement approach, methods for workload balancing for DSC are grouped into three categories: (i) switch migration, (ii) flow migration, and (iii) flow splitting.

Switch Migration: switch control can be transferred from overloaded controllers to lightly loaded controllers, considering workload reduction. The study [21] has discussed the switch migration because of CPU and memory allocation exceeding a controller's threshold level, but it does not define the way of choosing the targeted controllers. Switch migration using the Q-learning technique is discussed in [22], and it has reduced the standard deviation of the controller's workload. Cui et al. [23] have used the response time of the controller for switch migration. By using this technique, the switch is transferred with the largest load on the controller and the quickest reaction time. Ref. [24] proposed a method that targeted controller selection for switch migration based on CPU utilization, memory size, bandwidth, etc. Hu et al. [25] proposed a simulated annealing algorithm for selecting the targeted controller to reduce the switch migration cost.

Flow Migration: The flow migration method only transfers the hardness (i.e., flow beyond the threshold level) of the flow instead of migrating a whole switch. Hu et al. proposed a technique in which a "super controller administers every controller in the system and regulates the flow managed by them [26]. Ref. [27] proposed a game theory approach that managed the flow of each controller through workload exchange between them. Maity et al. [28] proposed a traffic-aware consistent approach for reducing the flow migration duration, and they achieved about a 15% reduction in flow migration time when compared with the conventional flow migration methods. Also, with the use of a traffic-aware flow migration approach, ref [29] has proposed a method to reduce the data plane load and achieved a 13% reduction when compared with the two-phase update approach.

Flow Splitting: This method allows a switch to be managed by more than one controller at the same time. Gorkemli et al. [30] discussed a method in which switches are required to negotiate with their controllers for flow splitting using a virtual overlay on the data plane. Ref. [31] proposed an approach based on convex quadratic programming for load balancing as well as reducing new switch-controller appointments through modeling the mapping between controllers and switches.

The control relation graph-based controller placement method for software-defined networking (SDN) is presented in [32]. It demonstrates that the proposed approach reduces management costs through load balancing and response time in LEO satellite networks. Zhang et al. proposed an SDN-based space-terrestrial integrated network architecture. In addition, it has presented an efficient dynamic controller placement and adjustment algorithm for better load balancing and response time [33]. Chen et al. proposed a dynamical control domain division problem to reduce the management cost. In addition, it has presented a heuristic algorithm to choose the best controller for better load balancing [34].

However, considering the practical viability of Open Flow, a switch cannot be controlled by more than one controller simultaneously considering synchronization and complex design. Therefore, flow migration and flow splitting methods are non-compliant to the OpenFlow protocol and cannot be implemented in the real-time controller platform.

2.1. PROBLEM DESCRIPTION AND CONTRIBUTION

As discussed in the literature section, most of the switch migration methods are having issues with ping-pong difficulty. The ping-pong difficulty of the controller is explained in the following example. Let us consider two controllers [C_p and C_q] and three switches [S_a, S_b, S_c] in the network. The maximum manageable workload for each controller is 100 PIMS per second. The switches S_a, $S_{b,}$ and S_c produce 60, 80, and 60 PIMS per period, respectively. In time t, C_p handles switches S_a & S_b then controller C_q manages to switch S_c. Since $\alpha_{c\_p} = \delta_a(t) + \delta_b(t) = 60 + 80 > \beta_{c\_p}$(100 PIMS), C_p is overloaded and requires switch migration. In most of the switch migration methods, an overloaded controller will request and takeover a switch for a whole period from other controllers. Therefore, Switch S_a is transferred to controller C_q`s subnet at time t+1. Though at period t+1, $\alpha_{c\_q} = \delta_c(t) + \delta_a(t) = 60 + 60 > \beta_{c\_q}$ (100 PIMS), controller C_q will be overloaded. So, controller C_q *a*sks C_p to take over a switch again in time t+2, which makes ping-pong difficult.

Recently, W.K. Lai et al. [35] proposed a time-sharing switch migration scheme (TSSM) that mitigates the ping-pong difficulties in the controllers by sharing the workload of a switch that is supervised by two controllers at the same time during overloaded conditions. It proposes a strategy whereby switch migration is performed in a time-sharing manner, where the workload of the switch is divided between two controllers within a given period. Considering the previous example, at time t+1, controller C_p manages 20 PIMs of S_a, and the remaining 40 PIMs are handled by C_q through migration. During this time, both controllers C_p and C_q are managing the workload of switch S_a. Hence, C_p's workload becomes $\alpha_{c\_p} = \delta_a(t) + \delta_b(t) = 20 + 80 \leq \beta_{c\_p}$(100 PIMS) and, C_q's workload turns out to be $\alpha_{c\_q} = \delta_c(t) + \delta_a(t) = 60 + 40 \leq \beta_{c\_q}$(100 PIMS). Therefore, both controllers are not overloaded (busy) in period t+1. Similarly, at ti*me* t+2, C_q initially processed 40 PIMS, and the remaining 20 PIMS have been sent to the C_p controller subnet. In this approach, The TSSM scheme can successfully conquer the ping-pong difficulty of the controller.

Specifically, it proposes a strategy where two controllers, namely an overload controller (one) and a lightly loaded controller (it can be many, but this paper utilizes one), are combined, and the switch from an overloaded to a lightly loaded controller subnet is made at an adequate point in time. The outcome of this technique shows that it has considerably reduced overload occurrences of the controllers and effectively balanced the workload of all the controllers with improved controller efficiency as compared with the existing switch migration methods such as group-based dynamic controller placement [10], churn-triggered migration [30], and the "best-fit migration [32] method. Nevertheless, it is observed that more than one lightly loaded controller operation in the TSSM provides better controller efficacy than the original one (i.e., discussed in the paper) with the increased switch migration cost. In addition, this method has higher controller resource consumption during TSSM operation since the migration switch is managed (i.e., controlled) by more than one controller in the network.

Therefore, we proposed a strategy that optimizes the selection of lightly loaded controllers during the TSSM period and allows more than one lightly loaded controller for switch migration during the TSSM period without increasing migration costs. The controller is selected based on flow characteristics through a greedy set coverage algorithm, which reduces the controller's resource consumption by reducing the number of controllers participating in the flow processing. The improved TSSM scheme provides reduced migration costs and lower controller resource consumption, as well as TSSM benefits. The implementation of the proposed scheme is accomplished through an open network operating system (ONOS) for its feasibility, and it can respond to about one million flow processing requests per second.

In summary, software-defined networking (SDN) leads to an efficient administration process in network management through easy updating of network poli-

cies and the latest algorithms. Typically, distributed SDN is adopted in network management, considering bottleneck issues. Load balancing is a critical factor in the SDN, and it can be managed through (i) the dynamic controller placement method, (ii) switch migration, (iii) the flow splitting method, and (iv) the flow migration method. Considering the practical viability of Open Flow, a switch cannot be controlled by more than one controller simultaneously, considering synchronization and complex design. Therefore, flow migration and flow splitting methods are non-compliant with the OpenFlow protocol and cannot be implemented on the real-time controller platform. Considering the OpenFlow protocol and its implementation in the real-time controller platform, the dynamic controller placement method with switch migration is a better solution for load balancing.

The conventional switch migration methods suffer from ping-pong difficulty during the switch migration process because the whole single switch is migrated in the beginning period. It causes instability issues in the switch migration. The ping-pong difficulty is rectified by a time-sharing switch migration scheme. This method significantly reduces the overload occurrences of the controller, which leads to better load balancing. However, the selection of controllers during the TSSM period is random. So that it could increase the switch migration cost and higher controller resource consumption during TSSM operation since the migration switch is managed by more than one controller in the network. Therefore, our paper has proposed an improved TSSM scheme, and it has the following merits: (i) It contains all the merits of a conventional TSSM scheme, including the removal of ping-pong controller action during the switch migration process, a reduction in controller overload occurrences, and better controller efficiency. (ii) The selection of controllers during TSSM is specified and optimized through the greedy set algorithm, which reduces the switch migration cost and controller resource consumption. (iii) It provides better controller efficiency and load balancing compared with the conventional TSSM scheme. The structure of the paper is shown in Fig. 1.

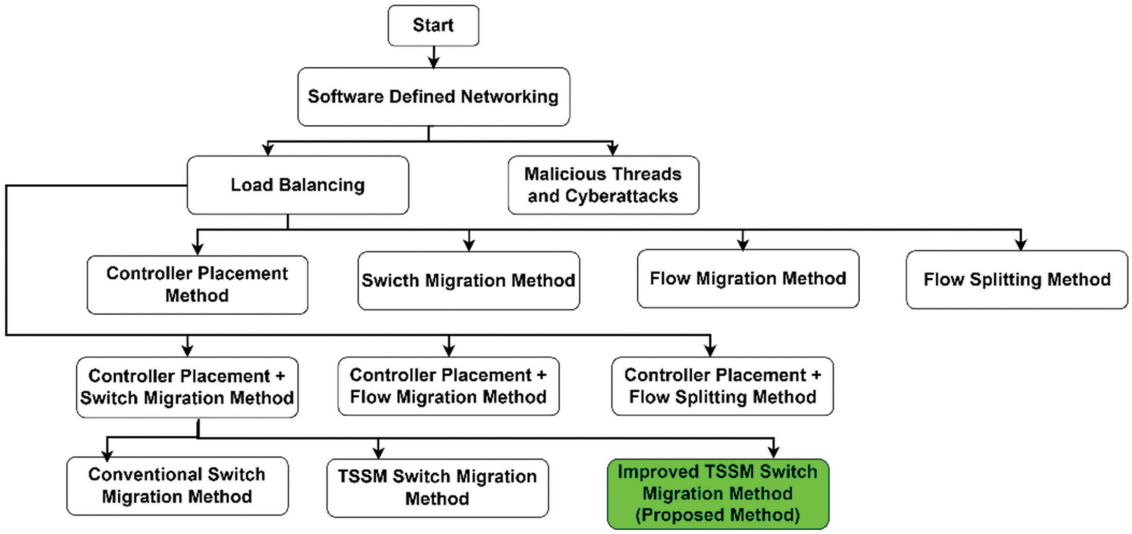


Fig. 1. Structure of the paper

2.2. ORGANIZATION OF THE PAPER

The paper is structured as follows: The literature review and problem description are covered in Section II of this paper. The background knowledge of the distributed SDN control network, OpenFlow protocol rules, and network model is detailed in Section III. The proposed improved TSSM scheme and matching algorithms are discussed in Section IV, and the performance evaluation of the proposed method is presented in Section V. Finally, the concluding statement is summarized in Section VI.

3. DISTRIBUTED SDN CONTROLLER

The architecture of the distributed SDN control network, the switch transfer procedure in the OpenFlow protocol, and network models are discussed in this section.

3.1. DISTRIBUTED SDN CONTROL NETWORK ARCHITECTURE

Two common control methods are typically followed in the distributed SDN control network, namely, (i) the hierarchical method and (ii) the flat control method, also called circular chain control [8]. In the hierarchical method, the central distributed controller (called the leader) has the idea of a network global view and is updating the network policies and latest algorithms to the sub-controllers, as shown in Fig. 2(a). The sub-controller takes control (is in charge) of the subnet of its switches, as well as reports its status to the leader. It is noted that the new leader will be selected if the original leader is broken down in the hierarchical method [15]. In the case of circular chain control, controllers have information about the local view of the network

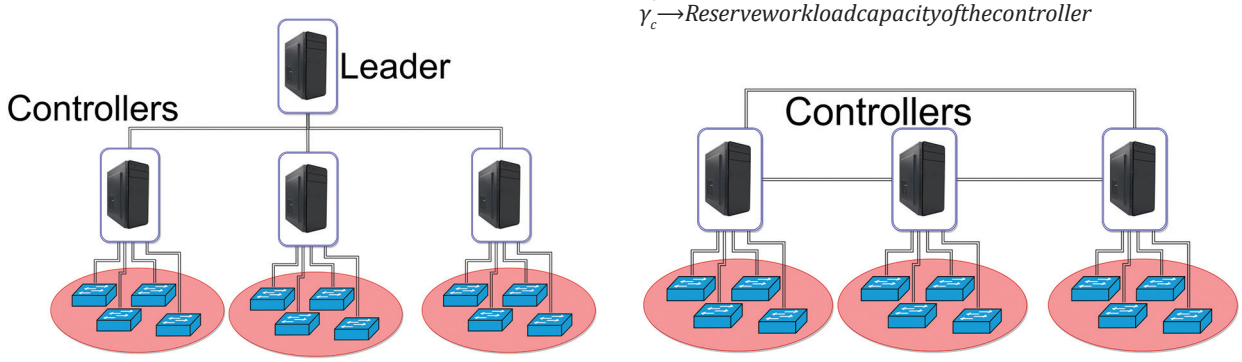
and authority over their own subnet. The involved controllers are swapping information among themselves in a distributed manner, as shown in Fig. 2(b).

The hierarchical method is considered in this paper to apply the proposed switch migration scheme. The leader is responsible for monitoring the status of each sub-controller as well as performing the TSSM scheme to select the lightly loaded controller for the overloaded controller during flow fluctuations, flow traffic, impulses, distributed denial of service, and so on. Afterward, two sub-controllers (overloaded and lightly loaded) are committed to sharing their workloads and migrating the switch where it is required.

Specifically, the threshold level of the sub controller is also defined in the leader to avoid unwanted switch migrations. When the workload of the controller is more than the threshold level, it is considered overloaded, and it is selected based on the maximum capacity and reserve capacity of the controller. Generally, the threshold level is selected between 90 and 95% of the maximum capacity, as recommended by network administrators. The threshold level of the controller is also noted as the maximum workload of the controller, and it is defined in eq. (1).

$$\varphi_c=\beta_c-\gamma_c \tag{1}$$

$\varphi_c \longrightarrow Thersholdworkloadlevelofthecontroller$
$\beta_c \longrightarrow Maximumworkloadcapacityofthecontroller$
$\gamma_c \longrightarrow Reserveworkloadcapacityofthecontroller$



(a) (b)

Fig. 2. Control methods for the DSC architecture: (a) Hierarchical method, (b) Flat method

3.2. TRANSFERRING PROCESS FOR SWITCHES IN OPENFLOW PROTOCOL

OpenFlow permits a switch transfer among various subnets and creates a connection with several controllers. Based on switch S_n's point of view.

The following roles are determined by each associated controller C_p.

- OFPCR_ROLE_EQUAL (Equal): This default role makes controller C_p to have full authority to switch S_n, and C_p can send commands to S_n and receive the status. Similarly, all the controllers have full access to S_n when it is acting in this role.
- OFPCR_ROLE_SLAVE (Slave): If the controller C_p role is changed to slave, then C_p can only read the status from switch S_n.
- OFPCR_ROLE_MASTER (Master): It is like as equal role and controller C_p has complete authority to S_n. Though, it is insisted that only one controller (e.g., C_p) is considered as a master controller for a switch Sn and other controllers are regarded as slaves to switch S_n.

Transferring process for the switches is defined in the OpenFlow protocol is shown in Fig. 3. Switch transferring process is initiated by the master controller since it has full authority over the switch. For example, controller's C_p and C_q are the master and targeted (slave) controllers respectively, for the switch S_n. It is insisted that overloaded controllers are transferring a switch to other controllers for workload balancing with the help of the leader (controller). Once the master controller (C_p) gets a command from the leader, it will then send a transfer request to switch S_n to targeted controller C_q.

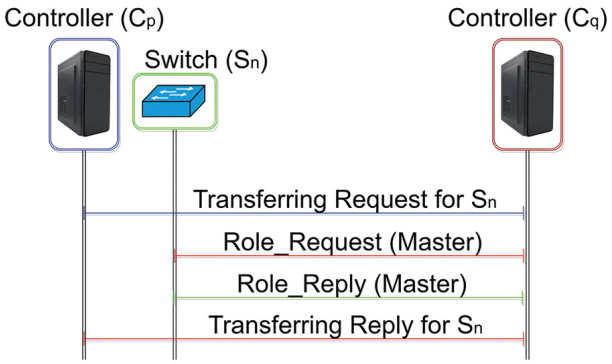


Fig. 3. Switch transferring process in OpenFlow Protocol

After that, controller C_q asks switch Sn to change the role of S_n control to master instead of slave through the Role_Request (Master) message, and switch S_n will provide a confirmation message to C_q via Role_Reply (Master). After all, C_q provides notification message to C_p for the victorious migration of switch S_n and then controller C_p acts as a slave controller for switch S_n.

The switch migration is supported by the OpenFlow protocol in versions 1.2, 1.3, 1.4, and 1.5 (most recent version). It is observed that OpenFlow regulation only instructs about how to alter (migrate) the switches between controllers for their roles and exchange messages between controllers. However, deciding target controllers and switches for migration is not defined by OpenFlow. The proposed improved TSSM scheme achieves optimized controller selection and determines when to execute switch migration during the TSSM period.

3.3. NETWORK MODEL

Let us predict an SDN-based network comprised of a collection S_n of switches and a group C_n of controllers. A switch (e.g., S_a) in S_n is controllable by a controller in C_n (e.g., C_p) with a model of one switch is controlled by a controller simultaneously recommended by OpenFlow, i.e., Cp is acting as a master controller for Sa, and it can be changed after the switch migration.

The workload of each controller is determined through Packet_In messages (PIMs) sent from the switches. Particularly, switches workload (δ(t)) are determined through the number of PIMs generated by a switch in each period ´*t*´. Subsequently, controller workload capacity is defined as the maximum amount of PIMS that can be handled in each period. For example, if switches S_a to S_z are administered by controller C_p then the workload of the controller C_p is calculated as,

$$WL_{pt} = \alpha_{c_p} = \sum_{S_a}^{S_z} \delta(t) \quad (2)$$

Generally, the maximum workload (αc) of the controller shall be less than the maximum capacity of the controller (βc) considering the requirement of reserve load during unwanted situations such as flow fluctuation, abrupt demand, etc. In this paper, hierarchical control of DSC architecture is considered; therefore, the leader collects workload from all the controllers at every period and directs the switch migration between controllers when required.

4. PROPOSED SWITCH MIGRATION SCHEME

During the initial stage, controller placement methods or network operators are used to configure the network switches, where each switch is controlled by a master controller. As discussed in the previous section, conventional switch migration methods include migrating a switch at the beginning of the period as well as a complete part of a switch even though it is not required. Thus, connections between controllers and switches are static for the whole period. In the case of TSSM, switch migration is allowed through time-sharing, and switches in the network can dynamically alter their connections with the controller in each period. In addition, the TSSM scheme effectively overcomes the controller ping-pong difficulty, as discussed in Section 2.1. Nevertheless, controller resource consumption is higher during the TSSM period, which could increase the migration cost of the method compared to other migration methods since it allows more than one controller to share their switch loads during the TSSM period. It is observed that migration costs are estimated based on the utilization of controllers and switches. Therefore, this paper has proposed an algorithm that significantly reduces the number of controllers associated with the switches based on flow characteristics during time-sharing migration. We have introduced a greedy set coverage algorithm to achieve the optimal association between the switches and controllers during the time-sharing migration period, such that the number of controllers associated with the switch is reduced, which subsequently reduces controller resource consumption and lowers the migration cost. The following algorithms are designed for the successful

Algorithm 1: Locating Overloaded and Lightly Loaded Controllers

1. $C_{over} \leftarrow$ ø and $C_{light} \leftarrow$ ø ;
2. **foreach** $C_p \in C$ **do**
3. $\alpha_{c\_p} \leftarrow 0$;
4. **foreach** $S_a \in S_p$ *do*
5. $\alpha_{c\_p} \leftarrow \alpha_{c\_p} + \delta_{a,t}^{(p)}$;
6. if $\alpha_{c\_p} > \varphi_{c\_p}$ then
7. $C_{over} \leftarrow C_{over}$ U {Cp} ;
8. else if $\alpha_{c\_p} < \lambda \times \varphi_{c\_p}$ **then**
9. $C_{light} \leftarrow C_{light}$ U {Cp} ;
10. If $C_{over} \neq$ ø and $C_{light} \neq$ ø then
11. Use Algorithm 2 for load balancing between C_{over} and C_{light} ;

Algorithm 1: Locating Overloaded and Lightly Loaded Controllers

This algorithm is ensured to find all the overloaded (called busy) and lightly loaded controllers (called assistant or target controllers) in the given network, symbolized by *Cover* and *Clight*, respectively. The workload of each controller (e.g., α_{c_p}) is estimated based on Eq. (2) through adding the loads of each switch (e.g., $\delta_{a,t}^{(p)} + \delta_{b,t}^{(p)} + \cdots$) in the subnet, it is described in the algorithm code between 3 and 5 lines. Afterward, controller workload (e.g., α_{c_p}) is compared with the threshold level (φ_{c_p}) and if it is more than the threshold level then it is considered as an overloaded controller and included in the overload controllers (characterized in lines 6 -7) unit in the leader. Then lightly loaded controllers are determined based on a lightly loaded coefficient 'λ', value between 0.9 and 0.95 (selected by network administrators) and it is included in line 8. Afterwards, lightly loaded coefficient is multiplied with the threshold value, and if the workload of the controllers is less than the multiply value, then it is considered a lightly

loaded controller, and it is added to the lightly loaded controller unit in the leader. It is insisted that switch migration be carried out when both C_{over} and C_{light} controllers are non-empty, checked in line 10.

Lemma 1: Let assume overloaded and lightly controllers are subset of main controller $\ddot{C}$ (i.e., $\xi Cover \& \xi Clight \in \ddot{C}$) and all the switches are included within this domain is represented as $\ddot{S}$ (i.e, $\xi S \in \ddot{S}$), then complexity of the time computation for algorithm 1 is estimated as O ($\xi C_{over} + \xi C_{light} + \xi S$) + T_2, T_2 is the computation time of algorithm 2.

Proof: In algorithm 1, Line 1 requires a constant amount of time to initialize both C_{over} and C_{light}. Then, in lines 2-9, the outer for-loop has iterations similar to the number of controllers positioned in this domain, but lines 3, 6, 7, 8, and 9 all need O (1) time.

Lines 4-5's inner for-loop (together with the outer for-loop) examines every switch in overloaded controller S_p. As a result, the outer for-loop takes ($\xi C_{over} + \xi C_{light}$)× O (1) + O ($\xi S$) = O ($\xi C_{over} + \xi C_{light} + \xi S$). Line 11 then performs Algorithm 2 and uses T_2 time. To summarize, Algo. time complexity is O ($\xi C_{over} + \xi C_{light} + \xi S$) + T_2.

Algorithm 2: Switch Migration Segment for Load Balancing

1. SORT (C_{over}, α_{c_p} - φ_{c_p});
2. SORT (C_{light}, φ_{c_q} - α_{c_q});
3. **foreach** $C_p \in C_{over}$ do
4. SORT (Sp , $\delta_{a,t}^{(p)}$);
5. **while** $\alpha_{c\_p} > \varphi_{c\_p}$ do
6. **if** $C_{light} =$ ø then
7. Cease this module ;
8. Pick the optimized controllers C_{q_1}, C_{q_2},… from C_{light} ;
9. (Controller-Switch Association Matrix) ← Algorithm 3 (Request PIM´s of Switch, Switches from Cover)
10. (S_a, [τ_1 , τ_2 , …] , [n_1 , n_2 , …]) ← Algorithm 4 (C_p , [C_{q1}, C_{q2}, …]) ;
11. Transfer S_a to [C_{q1}, C_{q2}, …]'s subnet after [τ_1 , τ_2 , …] units of time ;
12. $\alpha_{c\_p} \leftarrow \alpha_{c\_p}$ - [n_1, n_2, …];
13. $\alpha_{c\_q1} \leftarrow \alpha_{c\_q}$ + [n_1, n_2, …] ;
    $\alpha_{c\_q2} \leftarrow \alpha_{c\_q}$ + [n_1, n_2, …] ;
14. **if** $\alpha_{c_{q[1,2,\ldots]}} \geq \lambda \times \varphi_{c_{q[1,2,\ldots]}}$ **then**
15. $C_{light} \leftarrow C_{light} \setminus \{C_q [1,2, \ldots]\}$;
16. **else**
17. SORT ($\boldsymbol{C_{light}}$, φ_{c_q} - α_{c_q}) ;

Algorithm 2: Ordering the pair of overloaded and assistant controllers and switch migration.

The aim of this algorithm is to share the workload between controllers by locating the pair of overloaded and lightly loaded controllers. The SORT function helps sort the overloaded and lightly loaded controllers in decreasing workload order. The overload controllers are sorted in code line 1, whereas line 2 provides the sorted information about the lightly loaded controller. Hence, a controller with extremely leftover capacity will be considered the first to contribute to the workload of an overloaded (busy) controller. The code in lines between 3 and 17 handles each controller in the network through for-loop by most overloaded controller to the lowest overloaded one. Line 4 sorts of the switches under C_p administration based on their workload in conjunction with decreasing order. The while loop in line 5-16 keep on decreasing the workload of the C_p by migrating a switch until it gets below threshold workload. However, if there is no assistant controller to help (i.e., C_{light} is empty), and more overload controllers are still in the domain then algorithm 2 terminates as given in line 6 -7. Otherwise, if we want to select again a lightly loaded controller C_q for sharing workload then time sharing switch migration scheme is to be activated. For that initially, Algorithm 3 is executed to find the optimum controllers [C_{q1} , C_{q2} , …] for TSSM in view of reduced controller resources consumption and lower migration cost. Afterward, once the optimized controllers are discovered then TSSM scheme is executed based on Algorithm. 4. The output of Algorithm. 4 provides three output parameters as noticed in line 10. In which, ´τ´ gives the information about what time switch S_a should migrate to other controllers, whereas ´n´ provides the information of how much of PIMs to be migrated to each controller. Afterward, workload updates of C_p and [C_{q1} , C_{q2}, …] is performed in line 11 to 13 and if [C_{q1}, C_{q2}, …] is exceeded the threshold level then these controllers are removed from the lightly loaded controllers as given in line 14, otherwise these controllers are again going for the sorted function in the lightly loaded controller unit as given in line 17 and line 2.

Lemma 2: This property proves that algorithm 2 must be converge and it does not run forever due to the finite number of overloaded controllers. Let consider sum of lightly loaded controllers, and switches are represented as $|C_{light}| = \xi_{light}$, $|S| = \xi_S$, respectively. In the worst scenario, algorithm 2 takes ξ_S ($T_3 + T_4$ + O (ξ_{light} + $\log_2 \xi_S$)) time, where T_3 and T_4 is the calculation time of Algorithm 3 and 4.

Proof: Lines 1 and 2 of algorithm 2 take time required for the sorting of overload O ($\xi_{over} + log_2\, \xi_{over}$) and lightly loaded O ($\xi_{light} + log_2\, \xi_{light}$) controllers. We choose an overloaded controller C_p(i.e., line 3), an assistance controller C_q (i.e., line 8), and shift the load of a switch Sp from C_p to C_q (i.e., lines 9-12) in the for-loop. Except for lines 4, 9, 10, and 16, each of the residual statements

inside the for-loop takes O (1) time. Then, line 4 takes the time to sort the switch Sp and it is estimated as O ($|S_p|$ $log2$ $|S_p|$). Line 9 detects a switch and the time (T_3) for finding the lightly loaded controller optimization by algorithm 3 and line 10 takes the time (T_4) required for the switch migration by algorithm 4. Therefore, considering all the time taken by each line then the total time complexity of the algorithm 2 is estimated as, ξ_S (T_3+T_3+O(ξ_{light}+log_2 ξ_S)).

Algorithm 3: Selection of Optimised Controller for TSSM Scheme

1. **Initialization**: controller-switch association{};
   set switches ={};
2. SORT (S_p, $\delta_{a,t}^{(p)}$; SORT (S_q, $\delta_{a,t}^{(q)}$);
3. **end-to-end traffic distribution**:
   Flow_pair = Flow_sort (flow)
4. **while** C_p in the Flow_pair:
   Traversing traffic on the network
5. Path_swicth = Dijkstra (Network Topology, C_p);
   Calculate the flow path
6. **while** Constantly Traversing Controller and Switch Path Sets
7. S is a set, {S_a, S_b, … S_z} is a subset of S, and $US_a = S$
8. **if** $S_a = S$, S_a is selected as the optimal coverage set of S.
9. **if** an element x satisfies $x \in S$, and $x \in S_a$, then S_a is the part of the optimal coverage set of S.
10. **if** $S_a \subset S_b$ exists, S_a is removed from { S_a, S_b, … S_z}
    Let S_n (x) denote a set from{S_a, S_b, … S_z} satisfying $x \in$ Sa, and
11. **if** S_n (x) $\subset$ S_n (y), element y is removed from S.
12. Perform the Corresponding Action
13. **else**:
14. Implement a Greedy Strategy to Select the Controller that Covers the Most Switches
15. **end if**
16. **end while**
17. **end while**
18. This Algorithm ceases until all switches are associated.
19. Use Algorithm 4 for TSSM scheme;

Algorithm 3: Selection of optimized controller for TSSM scheme in view of reduced migration cost

The objective of this algorithm is to provide optimized controllers for the lightly loaded controllers in the TSSM operation. The optimized controller is selected based on flow characteristics such that it reduces the controller's resource consumption and, subsequently, the switch migration cost. The greedy set coverage algorithm [36] is utilized for the optimized controller selection and is presented in Algorithm 3. This algorithm requires the PIMs of each switch in the overloaded controller C_p, as well as the controller's threshold level, network topology map, and so on. The Flow_sort function in line between 3 and 12 estimates the total amount of flow in each path and sorts it in descending order. In lines 4-6, execute and select a controller that covers most of the switches in the path. The sorted path set will continue to be covered until all switches are associated. Subsequently, the optimized controllers are selected, and then Algorithm 4 is executed for the TSSM strategy.

Lemma 3: Let assume all the lightly loaded controllers are sorted in the lightly loaded controller domain, it is represented as $|C_{light}|$=ξ_{light}. Let $C_{q1} > C_{q2}$, then in any optimal solution exists based on optimal flow that $X_1 <$ 1, or $X_2 = 0$. then complexity of the time computation for algorithm 3 is estimated as O ((C_{light})log^2(C_{light})).

Proof: In algorithm 1, Line 1 requires a constant amount of time to initialize controller-switch association. Line 2 takes the time required to sort the overload and lightly loaded controllers, represented as O (ξ_{over} + log^2 ξ_{over}), and O (ξ_{light} + log^2 ξ_{light}) respectively. Line 5 requires a time to estimate the flow path between switches and lightly loaded controllers, represented as O (log C_{light}). Then the total complexity of the time is computed as O ((C_{light})log^2(C_{light})).

Algorithm 4: Time to Switch Migration Estimating Segment

1. $\Delta_{over} \leftarrow min\ (\alpha_{c\_p} - \varphi_{c\_p})$ & $\Delta_{light} \leftarrow$ $max\ (\varphi_{c\_q} - \alpha_{c\_q})$;
2. $S_p^{\mu} \leftarrow$ ø and $S_p^{v} \leftarrow$ ø;
3. **foreach** $S_a \in S_p$ **do**
4. **if** $\delta_{a,t}^{(p)} \geq \Delta_{over}$ **then**
5. $S_p^{\mu} \leftarrow S_p^{\mu} \cup \{S_a\}$;
6. **else**
7. $S_p^{v} \leftarrow S_p^{v} \cup \{S_a\}$;
8. **if** $S_p^{\mu} \neq$ ø **then**
9. $S_a \leftarrow$ the last switch of S_p^{μ};
10. τ=[% of Δ_{light} *with respect to* Δ_{over}]×(L_t) ;
11. **else**

12. $S_a \leftarrow$the first switch of S_p^{ν};

13. $\tau \leftarrow 0$ then $n \leftarrow \delta_{a,t}^{(p)}$;

14. $\delta_{a,t}^{(p)} \leftarrow \delta_{a,t}^{(p)} - n$ and $\delta_{a,t}^{(p)} \leftarrow$ n ;

15. **return** (S_a, τ, n);

Algorithm 4: Time to Switch Migration Estimating Segment

After the optimum lightly loaded controllers (C_q [1,2, …]) are defined from Algorithm 3, they are combined with overloaded controllers to perform three tasks via the execution of Algorithm 4. The responsibilities are, (i) select a switch (from an overloaded controller) for sharing their workload with lightly loaded controllers, (ii) determine the switch migration time (τ), and (iii) calculate the number of PIMS (n) that lightly loaded controllers will process. Based on Algorithm 4, line 1, execute and consider 'Δ_{light}' be the remaining capacity of the lightly loaded controllers, and 'Δ_{over}' is believed to be the minimum amount of overload in the overloaded controllers. After that, switches in the overloaded controllers are split into two subnets namely S_p^{μ} and S_p^{ν}, respectively, where switch load is more than 'Δ' then it is sorted in S_p^{μ} with decreasing load order and S_p^{ν} includes remaining switches in the overloaded controllers, respecting codes are given in line 2-7. At first, switches near to 'Δ' (might be the very last switch in Spμ based on load soring order) is selected in the S_p^{μ} subnet for migration in view of reducing number of migrations (executed in line 8 - 9), because minimum amount of overload in the overloaded controllers can be easily getting placed in the lightly loaded controllers. The estimation of switch migration time depends on amount of PIM's generation in the switch, Δ_{light} in the optimum lightly loaded controllers, Δ_{over} in the switch. For example, if Δ_{light} is half of the Δ_{over} value and PIMs produced rate is constant then switch migration time is estimated as half of the period length as given in eq. (3). If τ=0, then switch migration occurs at beginning of the period as executed in line 13. Furthermore, once the switches in S_p^{μ} subnet is empty then S_p^{ν} subnet is considered for the better load balancing even though it is not overloaded, it is executed in line 11 - 12. This process will be repeated until all the controllers are load balanced via optimal controller finding (Algorithm 3) for each switch in the time-sharing scheme and then finally back to Algorithm 2.

Lemma 4: In the worst scenario, given ξS switches in S, Algo. 3 takes O(ξS) time.

Proof: The first two lines of Algo. 3 require a consistent amount of time to initialize. Because $S_p \subseteq S$, the for-loop in lines 3-7 repeats at most ξS times, and each statement takes O (1) time. Each statement in lines 8-15 obviously requires O (1) time. To summarize, Algo. 3's temporal complexity is O (1) + ξS O (1) + O (1) = O(ξS).

The relationship between all the algorithms is summarized in Fig. 4, and Table. 1. It is shown that the Algorithm 1 is used to locate the overload and lightly loaded controllers in the SDN domain. The whole switch migration is performed through an algorithm. 2. The optimization of controller selection for the TSSM scheme is achieved through Algorithm 3, and Algorithm 4 handles the TSSM process.

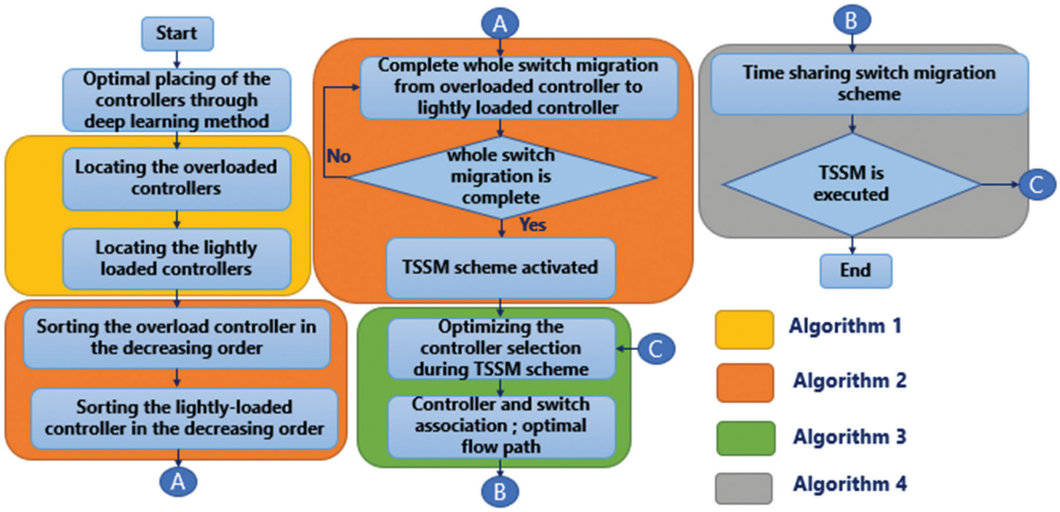


Fig. 4. Relationship between algorithms used in the improved TSSM scheme

Table. 1. Functions of Algorithms used in the improved TSSM scheme

Algorithms	Process
1	It is used to locate the overloaded and lightly loaded controllers in the SDN domain.
2	Initially, it is sorting the overloaded and lightly loaded controllers based on their overloading and PIMS accessibility. After that, it performs the whole switch migration from overloaded controllers to lightly loaded controllers.
3	It achieves optimized controller selection based on flow path through a greedy set algorithm for the TSSM operation.
4	It performs the TSSM operation and achieves better load balancing.

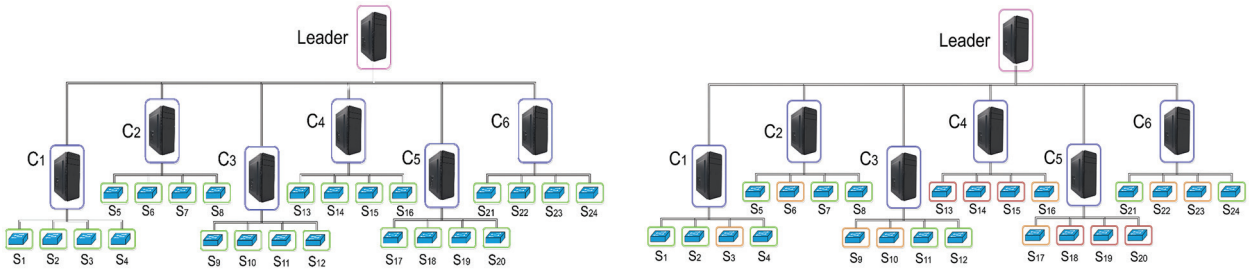


Fig. 5. Network topology used in the simulation test platform: (a) at '0' second, (b) at 21st second

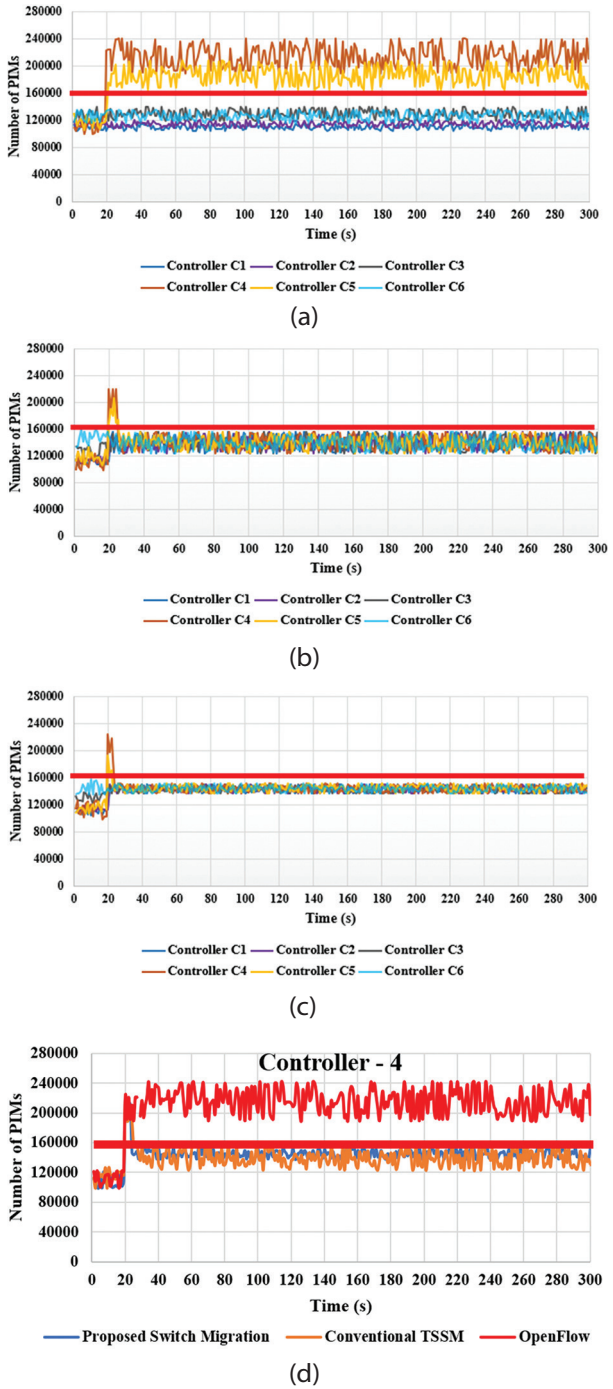


Fig. 6. Comparison of workload of controllers: (a) OpenFlow Method, (b)Conventional TSSM method, (c) Proposed Switch migration method, (d) Controller-4 comparsion results.

5. EVALUATION AND ANALYSIS

The performance of the proposed switch migration scheme is evaluated through time domain simulation analysis. The ONOS platform is considered the test platform, and hierarchical DSC architecture is adopted for the experimental network, it has seven controllers and 24 switches, as shown in Fig. 5. Therefore, one controller is acting as a leader, and its primary role is managing the other six controllers in the network; it does not involve itself in switch management; six secondary controllers are controlling their switches in its subnet. In this test platform, simulation time is considered 300 seconds and is divided into 60 periods. Each secondary controller has a PIMs handling capacity of 10^6 PIMs per period length of 5 seconds. In addition, the threshold for each controller is set at 800,000 PIMs per period. As a result, the total affordable load for the controller is estimated to be 4.8×10^6 PIMs per period. The loads of the switches are classified into three levels, namely, (i) light load, (ii) medium load, and (iii) large load. During light load, each switch generates around 21000 PIMs per second, whereas if a switch produces 42,000 PIMs per second, it is called medium load. However, if a switch is generating more than 63,500 PIMs per second, then it is called a large load. It is observed that if all the switches are lightly loaded, then the total affordable load for the controller is 2.52 × 106 PIMs per period, which is about 52.5 % of the total affordable load. But if all switches are considered as large loads, then the total load will be 7.62×10^6 PIMs per period, which is much higher than the total affordable load for the controller. Therefore, in this simulation study, the simulation starts with a small load in all switches, and the load will be randomly increased in the switches by cbench tool as simulation time increases, for evaluating the performance of the switch migration method. For example, at 21st seconds, 10 switches (S1, S2, S4, S5, S7, S8, S11, S12, S21, S24,) are generating about 21,500 PIMS/s, and 8 switches (S3, S6, S9, S10, S16, S17, S22, S23,) are generating 43,000 PIMs/s, and the remaining switches (S13, S14, S15, S18, S19, S20,) are producing 64,000 PIMs/s. Therefore, total controller workload is 4.715×10^6 PIMs per period, and there must be switch migration by both conventional (complete switch) and TSSM scheme. For evaluating the performance of the proposed method, three cases are considered: (i) work loads of occurrences, (ii) occurrences of overload, (iii) controller resource consumption, and (iv) migration cost.

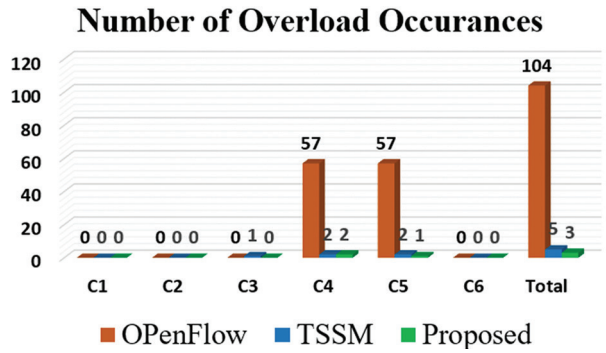


Fig. 7. Comparison of number of overload occurrences in conventional and proposed method

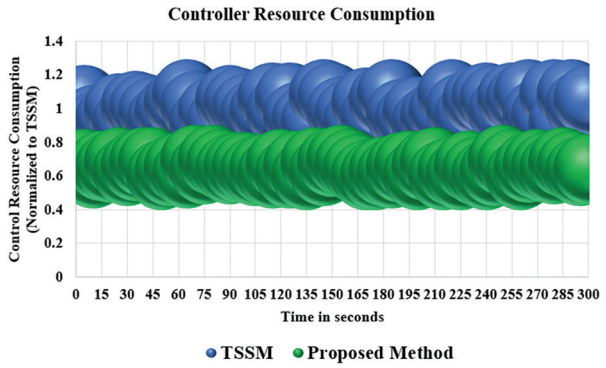


Fig. 8. Comparison of controller resource consumption between TSSM and proposed switch migration method

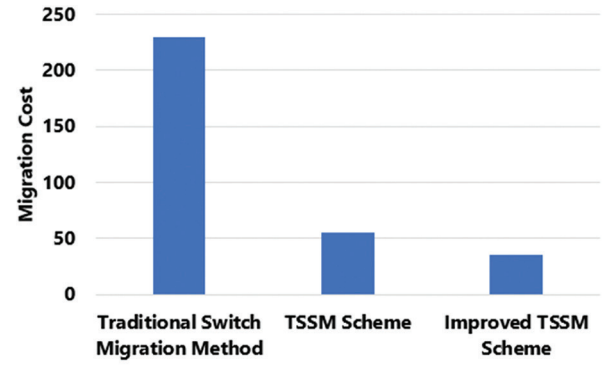


Fig. 9. Comparison of migration cost between conventional method and proposed method

Test 1: Workload of Controllers

As we have considered, each controller can process up to 800,000 PIMs per period, and if the controller has processed more than 160,000 PIMs/s then it is considered an overloaded controller. In this test, two conventional methods such as OpenFlow and TSSM are considered, and their test results are compared with the proposed method for evaluating the performance.

It is noted that switch migration is not performed in the OpenFlow method, and therefore, controllers C4, and C5 are heavily overloaded, as shown in Fig. 6a, based on PIMs generated in the switches. During this period, controllers C4 and C5 must handle about 1,165,000 PIMs/period, which is more than their maximum capacity (106 PIMs per period) and leads to unexpected issues in the networking domain. In the case of the TSSM scheme, it shares the workload between controllers through time sharing migration and ensures that all the controllers are within their threshold limits, as shown in Fig. 6b. In addition, the Ping-Pong issue (there are no high jumps, and frequently transferred switches are considered nil) is not noticed in the test results. The test results of the proposed switch migration scheme are presented in Fig. 6c. It is observed that load sharing between the controllers is much flatter (i.e., mostly all the controllers are sharing about the same load, which improves the efficiency and reduces the downtime or maintenance activity of the controllers) as compared with the TSSM scheme.

Test 2: Number of Overload Occurrences

This test is used to evaluate the number of overload occurrences for the controllers in the entire period (300 s), and it is useful for finding the performance of the switch migration method. The comparison of overload occurrence for all three methods is given in Fig. 7. It shows that the OpenFlow method provides a high number of overload occurrences since there is no action of switch migration, and therefore, controllers C1, C2, C3, and C6 are in the lightly loaded range, whereas C4, and C5 are highly loaded, and these controllers are completely overloaded during the given period.

In cases of TSSM, it has significantly reduced the number of overload occurrences for the controller since it avoids ping-pong difficulty and therefore switches that are repeatedly migrated are neglected. In the case of the proposed method, it further reduces the number of overload occurrences compared with the TSSM scheme. The proposed method uses more than one optimum controller as a lightly loaded controller during time sharing migration, which could reduce the number of overload occurrences. Because, if one controller is not enough to share the load of the switch (this controller may be initially considered excess in this situation), then it is again processed to find another controller for switch sharing in the conventional TSSM method, this might happen when the requirement for load sharing is high in the overloaded controller and lightly loaded single converters are not enough to handle this load. However, the proposed method chooses more optimum controllers based on the load sharing and avoids excessive processing and overload occurrences.

Test 3: Controller Resource Consumption

Controller resource consumption describes the utilization of a given number of controllers and switches. It is noted that if we reduce the number of controllers associated with the switches, that certainly reduces the switch migration cost of the network. OpenFlow is not performed with the switch migration event; thus, it is omitted for this evaluation study. The migration cost of the conventional TSSM is lower when compared with other switch migration methods. However, it is higher as re-

lated to the proposed switch migration method because the proposed method chooses the optimal controllers for sharing the workload based on flow characteristics, which reduces the controller's resource consumption and reduces the switch migration cost. The control resource consumption of the switch migration method is given in Fig. 8. It is observed that the proposed switch migration method provides about 34% less switch migration cost as compared with the conventional TSSM scheme.

Test 4: Migration Cost

The total number of switches that must be moved between various subnets to pass the migration cost test. Usually, controllers reassign switches to make sure loads are as evenly distributed as feasible. As switch migration is not carried out through the OpenFlow approach, it is not included. In the case of TSSM, it chooses switches for migration at random whose loads are near to tiny changes and permits two controllers to share their loads at the same time (i.e., time-sharing migration). The improved TSSM technique optimizes controller selection and enables controller workload sharing via time-sharing switch migration. The findings (Fig. 9) demonstrate that the enhanced TSSM program offers a 36% reduction in migration cost when compared with the traditional TSSM scheme.

6. CONCLUSION

This paper has presented an improved TSSM approach and mitigates the issue of increased switch migration cost in the conventional TSSM method by finding more than one optimal target controller during the time-sharing period. It has utilized the flow characteristics for finding the optimal controllers through a greedy set coverage algorithm. In addition, the proposed switch migration approach has TSSM benefits, which have conquered the ping-pong controller difficulty. The ONOS platform is considered for the performance evaluation of this study, which found that the improved TSSM scheme has better performance than the conventional TSSM method in terms of workload sharing among controllers, number of overload occurrences, and reduced controller resource consumption. Specifically, it reduces the controller's resource consumption by 34% when compared with the conventional TSSM.

ACKNOWLEDGEMENT

The authors wish to thank Annamalai University, India, for providing laboratory and experimentation resources.

7. REFERENCES

[1] N. Anerousis, P. Chemouil, A. A. Lazar, N. Mihai, S. B. Weinstein, "The origin and evolution of open programmable networks and SDN", IEEE Communications Surveys and Tutorials., Vol. 23, No. 3, 2021, pp. 1956-1971.

[2] Y.-C. Wang, H. Hu, "An adaptive broadcast and multicast traffic cutting framework to improve Ethernet efficiency by SDN", Journal of Information Science and Engineering., Vol. 35, No. 2, 2019, pp. 375-392.

[3] M. Alsaeedi, M. M. Mohamad, A. A. Al-Roubaiey, "Toward adaptive and scalable OpenFlow-SDN flow control: A survey", IEEE Access, Vol. 7, 2019, pp. 1073-1079.

[4] J. H. Cox, R. Clark, H. Owen, "Leveraging SDN and WebRTC for rogue access point security", IEEE Transactions on Network and Service Management., Vol. 14, No. 3, 2017, pp. 756-770.

[5] Y.-C. Wang, S.-Y. You, "An efficient route management framework for load balance and overhead reduction in SDN-based data center networks", IEEE Transactions on Network and Service Management., Vol. 15, No. 4, 2018, pp. 1422-1434.

[6] W. Iqbal et al. "ALAM: Anonymous lightweight authentication mechnism for SDN-enabled smart homes", IEEE Internet of Things Journal., Vol. 8, No. 12, 2021, pp. 9622-9633.

[7] Y.-C. Wang, R.-X. Ye, "Credibility-based countermeasure against slow HTTP DoS attacks by using SDN", Proceedings of the IEEE Annual Computing and Communication Workshop and Conference, NV, USA, 27-30 January 2021, pp. 890-895.

[8] F. Bannour, S. Souihi, A. Mellouk, "Distributed SDN control: Survey, taxonomy, and challenges", IEEE Communications Surveys and Tutorials. Vol. 20, No. 1, 2018, pp. 333-354.

[9] J. Lu, Z. Zhang, T. Hu, P. Yi, J. Lan, "A survey of controller placement problem in software-defined networking", IEEE Access, Vol. 7, 2019, pp. 24290-24307.

[10] H. Sufiev, Y. Haddad, L. Barenboim, J. Soler, "Dynamic SDN controller load balancing", Future Internet, Vol. 11, No. 3, 2019, pp. 1-21.

[11] Y. Wu, S. Zhou, Y. Wei, S. Leng, "Deep reinforcement learning for controller placement in software defined network", Proceedings of the. IEEE Conference on Computer Communications Workshops, Toronto, ON, Canada, 2020, pp. 1254-1259.

[12] Y.-C. Wang, Y.-C. Wang, "Efficient and low-cost defense against distributed denial-of-service attacks

in SDN-based networks", International Journal of Communication Systems., Vol. 33, No. 14, 2020, pp. 1-24.

[13] F. Tang, H. Zhang, L. T. Yang, L. Chen, "Elephant flow detectionand load-balanced routing with efficient sampling and classification", IEEE Transactions on Cloud Computing., Vol. 9, No. 3, 2021, pp. 1022-1036.

[14] Y.-C. Chan, K. Wang, Y.-H. Hsu, "Fast controller failover for multidomain software-defined networks", Proceedings of the. European Conference on Networks and Communications., Paris, France, 29 June - 02 July 2015, pp. 370-374.

[15] W. H. F. Aly, "Controller adaptive load balancing for SDN networks", Proceedings of the International Conference on Ubiquitous and Future Networks., Zagreb, Croatia, 2-5 July 2019, pp. 514-519.

[16] T. Hu, J. Zhang, L. Cao, J. Gao, "A reliable controller deployment strategy based on network condition evaluation in SDN", Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, Beijing, China, 24-26 November 2017, pp. 367-370.

[17] T. Kim, J. Myung, S.-E. Yoo, "Load balancing of distributed datastore in OpenDaylight controller cluster", IEEE Transactions on Network and Service Management, Vol. 16, No. 1, 2019, pp. 72-83.

[18] Y.-C. Wang, E.-J. Chang, "Cooperative flow management in multidomain SDN-based networks with multiple controllers", Proceedings of the International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI, Charlotte, NC, USA, 14-16 December 2020, pp. 82-86.

[19] S. Nithya, M. Sangeetha, K. N. A. Prethi, K. S. Sahoo, S. K. Panda, A. H. Gandomi, "SDCF: A software-defined cyber foraging framework for cloudlet environment", IEEE Transactions on Network and Service Management, Vol. 17, No. 4, 2020, pp. 2423-2435.

[20] K. S. Sahoo, P. Mishra, M. Tiwary, S. Ramasubbareddy, B. Balusamy, A. H. Gandomi, "Improving end-users' utility in software-defined wide area network systems", IEEE Transactions on Network and Service Management, Vol. 17, No. 2, 2020, pp. 696-707.

[21] A. Dixit, F. Hao, S. Mukherjee, T. V. Lakshman, R. Kompella, "Towards an elastic distributed SDN controller", ACM Special Interest Group on Data Communication., Vol. 43, No. 4, 2013, pp. 7-12.

[22] Z. Min, Q. Hua, Z. Jihong, "Dynamic switch migration algorithm with Q-learning towards scalable SDN control plane", Proceedings of the International Conference on Wireless Communications and Signal Processing, Nanjing, China, 11-13 October 2017, pp. 1-4.

[23] J. Cui, Q. Lu, H. Zhong, M. Tian, L. Liu, "SMCLBRT: A novel load-balancing strategy of multiple SDN controllers based on response time", Proceedings of the IEEE International Conference on High Performance Computing and Communications, Exeter, UK, 28-30 June 2018, pp. 541-546.

[24] K. S. Sahoo et al. "ESMLB: Efficient switch migration-based load balancing for multicontroller SDN in IoT", IEEE Internet of Things Journal, Vol. 7, No. 7, 2020, pp. 5852-5860.

[25] T. Hu, J. Lan, J. Zhang, W. Zhao, "EASM: Efficiency-aware switch migration for balancing controller loads in software-defined networking", Peer-to-Peer Networking and Applications., Vol. 12, 2019, pp. 452-464.

[26] Y. Hu, W. Wang, X. Gong, X. Que, S. Cheng, "BalanceFlow: Controller load balancing for OpenFlow networks", Proceedings of the IEEE International Conference on Cloud Computing and Intelligence Systems., Hangzhou, China, 30 October - 1 November2012, pp. 780-785.

[27] W. Lan, F. Li, X. Liu, Y. Qiu, "A dynamic load balancing mechanism for distributed controllers in software-defined networking", Proceedings of the International Conference on Measuring Technology and Mechatronics Automation, Changsha, China, 10-11 February 2018, pp. 259-262.

[28] S. M. Maity, C. Mandal, "Traffic-Aware Consistent Flow Migration in SDN", Proceedings of the IEEE International Conference on Communications, Dublin, Ireland, 7-11 June 2020, pp. 1-6.

[29] S. M. Maity, C. Mandal, "DART: Data Plane Load Reduction for Traffic Flow Migration in SDN," IEEE Transactions on Communications, Vol. 69, No. 3, 2021, pp. 1765-1774.

[30] B. Gorkemli, S. Tatlcıoglu, A. M. Tekalp, S. Civanlar, E. Lokman, ˇ "Dynamic control plane for SDN at scale", IEEE Journal on Selected Areas in Communications, Vol. 36, No. 12, 2018, pp. 2688-2701.

[31] F. Al-Tam, N. Correia, "Fractional switch migration in multicontroller software-defined networking", Computer Networks., Vol. 157, 2019, pp. 1-10.

[32] L. Chen, F. Tang, X. Li, "Mobility-and load-adaptive controller placement and assignment in LEO satellite networks", Proceedings of the IEEE Conference on Computer Communications, Vancouver, BC, Canada, 10-13 May 2021, pp. 1-10.

[33] X. Zhang et al. "Dynamical controller placement among SDN space-terrestrial integrated networks", Proceedings of the IEEE 22nd International Conference on High Performance Computing and Communications, 14-16 December 2020, pp. 352-359.

[34] L. Chen et al. "Dynamical control domain division for software-defined satellite-ground integrated vehicular networks", IEEE Transaction on Network Science and Engineering, Vol. 8, No.4, 2021, pp. 2732-2741.

[35] W.-K. Lai, Y.-C. Wang, Y.-C. Chen, Z.-T. Tsai, "TSSM: Time-Sharing Switch Migration to Balance Loads of Distributed SDN Controllers", IEEE Transaction on Network and Service Management, Vol. 19, No. 2, 2022, pp. 1585-1597.

[36] Y. Zhang, Y. Ran, Z. Zhang, "A simple approximation algorithm for minimum weight partial connected set cover", Journal of Combinatorial Optimization., Vol. 34, No. 3, 2017, pp. 956-963.

The Impact of Applying ISO Standards Systems on Improving the Quality of the Performance in Higher Educational Institutions in Egypt

Original Scientific Paper

Gehan Mounir
Faculty of Computers and Information Technology, Future University in Egypt, Egypt
Gehan.mounir@fue.edu.eg

Amira M. Idrees
Faculty of Computers and Information Technology, Future University in Egypt, Egypt
amira.mohamed@fue.edu.eg

El Sayed M. Khater
Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt
dr.elsied.khater@gmail.com

Eman Mosallam
High Obour Institute for Management, Computers and Information Systems, Egypt
emanmosallam@gmail.com

Ayman E. Khedr
Faculty of Computers and Information Technology, Future University in Egypt, Egypt
ayman.khedr@fue.edu.eg

Abstract *– Applying ISO 21001:2018 standard ensures that universities have a competitive advantage as well as the achievement of their objectives. This study aims to identify the impact of implementing ISO 21001: 2018 management systems standards on the performance quality of higher education institutions. The study investigates the reasons why private higher education institutions seek ISO standards certificates in general and the specifications of management systems for educational institutions in particular. The study applied a set of statistical testing methods on paired samples as well as independent samples to ensure quality assurance. The study also proposes the required prerequisites that should be considered. The study investigated a hypothesis stating that "there are no statistically significant differences before and after applying the ISO 21001:2018 management systems specification for educational institutions in improving the quality of performance in higher education institutions" which was rejected by conducting an experiment in Future University in Egypt and accepting the alternative hypothesis. The study confirmed the impact of quality which was previously investigated by prior research that has been discussed in this study. The study further presented the need to apply quality based on determined criteria which were not considered in prior studies. Moreover, the study proposed the impact of ISO standards in educational institutions in general and in Egypt in specific. This recommendation is proved by this study to enhance the quality level in educational institutions.*

Keywords*: Personalization; Performance Quality, ISO standards systems, Higher Education Institutions management*

1. INTRODUCTION

The term "quality" began as a concept in private organizations and industrial establishments [1]. This concept went beyond to include the field of services in general, on top of which are educational services and higher education [2]. It is certain that the number has multiplied a lot, as education is one of the main needs of all societies. These societies have worked to develop educational institutions to meet the needs of the present and suit the requirements of the future [3]. The rapid change in the scientific, cultural, knowledge and technological structures in the world has led to the importance of this continuous development.

Educational institutions have a vision for the outputs of education, which is to obtain qualified learners who contribute to development [4]. This is performed

by applying specifications and standards that satisfy the concerned parties and are compatible with the requirements of the labor market and in line with technical progress in education, gaining knowledge and obtaining skills in the easiest ways, discovering and refining talents, and searching for the latest theories and methods [5]. There are many studies that dealt with the impact of management systems for educational institutions on the quality of performance of higher education institutions, but from different angles [6].

Although they mainly seek to identify the application of the quality system and ways to improve and develop it. In this chapter, we present some Arab and foreign studies.

All regimes and governments seek to improve and develop the educational system as the true measure of the present civilization of the nation. It could be considered the only choice in a world of many and successive changes through which we can draw the image of the future [7]. Predicting a qualified generation that can interact with the requirements and changes of the times could be achieved. Solving problems and keeping pace with advanced countries in the fields of medicine, industry, engineering, and innovation-based science is also one of the main objectives [8]. The successive events of this century have brought about many changes in economic, social, and cultural systems. They forced many educational systems in different societies to conduct changes to promote necessary education for survival, and those who reject or hinder change are condemned to annihilation. It has imposed attention to the specifications and conditions under which ISO, EFQM, is obtained [9].

Through the years, e-learning systems proved their applicability and advancement in the higher education sector. Many Educational Organizations focus on the students' benefits by applying the e-learning process since it ensures applying a scalable, reliable, and high performance of e-learning system to obtain student satisfaction and raise the student's level of education to achieve the ultimate desire of the educational process. Moreover, New Information Technology directs to increase the success of the e-learning process. An adaptive educational process for an e-learning system was proposed by applying e-learning on a cloud computing environment to support the load balance technique seeking the highest performance in the educational process which leads to a positive impact on the students' user satisfaction covered through an approach to evaluate where the results demonstrated [10].

Quality is the force needed to effectively drive the university education system to achieve its goals and mission by the community and interested parties in university education [11]. Recent trends in quality measurement and management enhance the directional, cognitive, professional, and behavioral characteristics of graduates, as well as the quality of the elements of the educational service delivery system [12]. Quality is one of the most important ways and means to improve the structure of the educational system with its physical and human components and to improve its performance in the present era, which some intellectuals call "the age of quality" [13].

The quality of educational work is no longer a theoretical or a kind of luxury sought by educational institutions or an alternative to educational systems. It is now a reality that no institution can abandon or stand away from entering. It is an urgent necessity dictated by the contemporary movement of life and refers to the specifications and characteristics expected in the educational product and in the processes and activities through which these specifications are achieved. The availability of integrated tools and methods that help educational institutions also achieve satisfactory, desired results, and quality in education which are related to the learning and education processes [10]. These achievements connect education to the needs of the community, make meaningful educational changes, and build and develop the queen of innovation among learners. Learning occurs when the learner interacts with the environment. Learning occurs when a change in the behavior of the learner occurs, and the role of education is to allow the interaction to take place so that learning can take place. The role of education is to provide an opportunity for interaction to require the development of process standards, including a specific education quality assurance system, so improving the quality of education has become a fundamental goal for all societies to improve current educational policies. But ensuring the education provided is of high quality.

The main objective of the current study is to identify the similarities and differences between the applications of ISO standards at private universities in Egypt. The study also aims at exploring the objectives, requirements, and challenges of applying the standards of management systems for educational institutions. To achieve the objectives of this study, a questionnaire which is composed of seven standard items is conducted. The study community participants included forty-seven participants from the teaching staff members and the heads of departments from the Faculty of Computers and Information Technology, Future University in Egypt.

The study presents the impact and challenges of implementing the ISO 21001:2018 standards for educational institutions. The study discusses a proposed vision for activating the ISO standards on educational institutions' management systems and the proposed procedures to meet their requirements and implementation steps. A hypothesis was examined which is stated: "There are no statistically significant differences before and after applying the ISO 21001:2018 management systems standard for educational institutions in improving the quality of performance in higher education institutions". The study confirmed the rejection of the hypothesis and accepted the alternative hypothesis with a confidence level of 99%.

The rest of the paper is organized as follows. Section 2 presents the background of the ISO 21001:2018 education organization management system. Section 3 discusses the literature review. The fourth section includes the research contribution. The fifth section discusses the developed research methodology while the results' analysis and discussions are discussed in section 6. The final section discusses the conclusion and implications for the future.

2. BACKGROUND OF ISO 21001:2018 EDUCATION ORGANIZATION MANAGEMENT SYSTEM

ISO 21001:2018 Education Organization Management System (EOMS) is an independent management system standard, consistent with ISO 9001 and focuses on the management systems of educational organizations as well as their impact on learners and other relevant stakeholders [14]. The system provides a common management tool for organizations that provide educational products and services capable of meeting the needs of learners and other beneficiaries.

Educational organizations are urgently and continuously needed to assess their degree of fulfillment of the needs of learners and other beneficiaries, as well as other relevant stakeholders, and to improve their ability to continue doing so.

Although educational organizations and educators around the world are the main beneficiaries of this document, all interested parties will benefit from unified management systems in educational institutions, e.g. employers who support and encourage employees to participate in educational services can also benefit from this document. The potential benefits of an organization implementing an EOMS based on this document are:

Ensure the alignment between the operated activities with the organization's mission and vision with encouraging the participants' innovation.

Fulfilling the organization's social obligation by offering a suitable quality level.

Providing a higher level of efficient learning system by considering the personalization features as well as online education with a more focus on the students with special needs and ensuring the continuous offering of educational opportunities.

Focus on raising the system performance through applying systematic and effective processes and efficiently applying the evaluation methods for continuous monitoring.

Raise the organization's credibility level and improve the stakeholder's culture which extends their willingness in the processes' participation.

Aligning the organization standards with the international framework.

The management system of the educational organization follows a set of principles including learners having the optimal focus, stakeholders' engagement, enhancement willingness, innovative leaders, a defined systematic secured approach and relationships management, proofed decisions, ethical conduct and equity, and society consideration and admitting responsibility.

This paper focuses on the standards for a qualified educational process. The paper proposes the alignment with ISO 9001 standards highlighting a focus on the required management standards that provide a positive impact on the educational process as well as on the participating stakeholders. The proposed ISO 9001 standards base educational process fulfills the implementation requirements from different perspectives with the ability to adapt to other aligned standards. Finally, the most important recommendations that have been reached and the most important research proposals will be presented

3. LITERATURE REVIEW

Different researchers have discussed the impact of quality on educational institutions. In [15] a study was applied to analyze the policies that control quality assurance at Ghanian Public University. The study followed the traditional methods for requirements determination including Interviews which have been conducted as well as analyzing the institution's documents with the stakeholders' observation. The conducted experiment revealed that the institution follows a robust road for raising the level of quality in the institution while the study also highlighted the requirement of other stakeholders' participation besides the management team by highlighting the factors that affect quality assurance such as examining the strategic plan, the available resources, the stakeholders' culture, …etc.

In [16], a hierarchical model was proposed for assessing the level of quality of the educational institutions' services in Mauritius based on both managerial and technical services' quality. The model included a set of fifty-three members representing the quality factors. These factors are revealed based on a conducted literature review, then assessing these factors' impact was conducted to confirm the validity of the proposed model. They reached a conclusion of five main factors including the level of quality of the administration activities, the institution's internal environment, the level of the educational processes quality, the level of quality of the available supportive services as well as the interaction.

In [17], research was conducted in Ghana for discussing the role of quality assurance in the stakeholders' satisfaction. The experiment was conducted at the University of Energy and Natural Resources. They highlighted the role of quality in the satisfaction level including the required services and roles as well as the regulations' executions. According to the conducted experiment,

the quality role has been confirmed with the objective of continuously improving the learning services' quality as well as raising the staff skills. Also, they highlighted the main requirement of quality awareness among the stakeholders with a further objective of adopting the institution's setting to be based on ensuring the required quality and continuous monitoring.

In [18], research was conducted in Tanzania aiming at illustrating the main parameters that have an impact on the stakeholders' satisfaction level. The experiment was conducted at the University of Dar es Salaam. The study applied the model proposed in [19] with a total of 153 participants. The results revealed the positive impact of the quality of the whole organization system, the staff performance, and the services' level of quality on the student's satisfaction. While no relation has been revealed between the course's level of quality and the satisfaction level. The research finally suggested considering the revealed relations in enhancing the quality level of educational institutions.

In [20], research was conducted in Germany with the objective of identifying the level of quality in educational institutions. The study focused on measuring the effective management process in raising the quality level and implemented a regression model based on questionnaire results which confirmed the focused relation between the management process activities and the significant role of the managers and quality assurance.

In [21], research focused on the impact of the quality culture in the organization and its internal quality assurance level. The study investigated this relationship in educational institutions located in Ho Chi Minh City, Vietnam. They highlighted that the activities of quality assurance are a pillar for setting a quality culture in educational institutions. The experiment included eight institutions with a total of 222 staff members. The study considered the variety of the institutions' types to be both public and private to avoid experiment bias. On the other hand, the conducted questionnaire included three main topics, the activities of quality assurance, the participants' awareness of these activities, and the institution's culture. While the research conclusion was detecting a correlation between the culture and the activities, however, it varied between the public and private sectors with a higher correlation in the private sector.

In [22], research focused on two main aspects of the private educational institutions in Bangladesh. The relation between the human resources activities and the educational institutions' quality level was the first aspect while the second aspect was the impact of the participants' commitments to this relation. Also, they focused on a set of human resources activities including securing the employees' jobs, employees' compensation, and the employees' independent behavior. A survey was conducted in twenty institutions and the results revealed the positive relationships between the two factors with the positive impact of the participants' commitments to the relationship.

4. RESEARCH CONTRIBUTION

The research contribution could be summarized in the following points:

1. Identify the impact of applying the ISO 21001:2018 Management Systems specification on the quality of performance of higher education institutions.
2. Identify why private higher education institutions seek ISO certification in general and the standards of management systems for educational institutions in particular.
3. Identifying the similarities and differences between applying ISO standards to Egyptian universities' stakeholders.
4. Develop a proposed vision for activating the ISO 21001:2018 Management Systems specification.
5. Identify the objectives, requirements, and constraints of applying the standards of management systems for educational institutions.

5. DEVELOPED RESEARCH METHODOLOGY

Since we are interested in identifying the impact of implementing ISO 21001:2018 management systems on the higher education institutions' performance quality. So, we can summarize our research contributions as follows: identify the impact of applying the ISO 21001:2018 management systems specification on the quality of performance of higher education institutions, identify the reason that private higher education institutions seek ISO certification in general, and the standards of management systems for educational institutions in particular, identify the similarities and differences between applying ISO standards at Egyptian Universities, develop a proposed vision for activating the ISO 21001:2018 management systems specification, and identify the objectives, requirements, and constraints of applying the standards of management systems for educational institutions. According to our research contributions, the proposed research methodology consists of six steps as follows:

Step 1: Preparing a data collection questionnaire (Pre and Post) applying the ISO 21001:2018 Management Systems specification.

Step 2: Application of the survey to the study sample (Pre and Post) application of the ISO 21001:2018 Management Systems specification.

Step 3: Analysis of survey results (Pre and Post) implementation of ISO 21001:2018 Management Systems specification.

Step 4: Make the necessary comparisons for the study.

Step 5: Measuring the impact of the implementation of the ISO 21001:2018 Management Systems Standard on the quality of performance of higher education institutions in Egypt.

Step 6: Submission of recommendations and proposals for future research in light of the results.

This study is based on the descriptive approach to fit the nature of the research which is based on the data collection resolution tool and to design the resolution in light of the research objectives and questions. The research is directed in particular to teaching staff members and department managers in the Faculty of Computers and Information Technology, Future University in Egypt (Private University). The sample of the study was applied to teaching staff members and department managers in the Faculty of Computers and Information Technology, Future University in Egypt. The total sample was equal to fifty-five members, thirty-eight are full-time professors, and seventeen are full-time technical staff. SPSS was used to analyze data, repetitions, arithmetic averages, and standard deviations. Testing has been extracted by statistical testing (T) for paired samples test and statistical testing (T) for independent samples test.

6. RESULTS ANALYSIS AND DISCUSSIONS

The First stage was examining differences between the opinions of Teaching Staff Members and Managers – Future University (Pre & Post) which results are illustrated in Table1.

Table 1. Paired Samples T-Test Results – Future University in Egypt

Standard Items	T	df	Sig. (2-tailed)
Context of the organization	-12.991	46	.000
Leadership	-15.898	46	.000
Planning	-16.165	46	.000
Support	-16.123	46	.000
Operation	-15.839	46	.000
Performance Evaluation	-20.175	46	.000
Improvement	-13.978	46	.000
Total	**-20.170**	**46**	**.000**

It is clear from the table that there are statistically significant differences Pre and Post the implementation of the ISO 21001:2018 Management Systems specification, where morale (0.000) was below 0.05, which is evidence of the rejection of the zero hypothesis that "There are no statistically significant differences before and after the application of the Management Systems specification For educational institutions ISO 21001:2018 and improving performance" at Future University. Consequently, the research accepts the alternative hypothesis that there are statistically significant differences between tribal and dimensional applications in favor of the dimensional application. The same examination is conducted in the second sample and the results are illustrated in Table 2.

It is clear from the table that there are statistically significant differences before and after the implementation of the ISO 21001:2018 Management Systems specification, where morale (0.000) is below 0.05, which is proof that we reject zero-based imposition, "there are no statistically significant differences before and after the implementation of the ISO 21001:2018 Management Systems specification and between education institutions Improving the quality of performance".

Table 2. Paired Samples T-Test Results

Standard Items	T	df	Sig. (2-tailed)
Context of the organization	-13.345	51	.000
Leadership	-17.509	51	.000
Planning	-15.183	51	.000
Support	-15.562	51	.000
Operation	-14.521	51	.000
Performance Evaluation	-15.415	51	.000
Improvement	-11.491	51	.000
Total	**-18.781**	**51**	**.000**

From the above, it is clear that the first hypothesis that "there are no statistically significant differences before and after the implementation of the ISO 21001:2018 Management Systems specification for educational institutions in improving the quality of performance in higher education institutions" is rejected and the alternative hypothesis is accepted.

The second stage is measuring the differences between the opinions of Teaching Staff Members and Managers - Future University in the application (Pre) and the opinions of Teaching Staff Members and Managers in the departments of another sample. The results are illustrated in Table 3.

Table 3. Independent Samples T Test—for Pre-Application

Standard Items	T	df	Sig. (2-tailed)
Context of the organization	1.079	97	.283
Leadership	.849	97	.398
Planning	.046	97	.964
Support	2.351	97	.021
Operation	.770	97	.443
Performance Evaluation	-.963	97	.338
Improvement	.960	97	.340
Total	**.897**	**97**	**.372**

It is clear from the table that there is no difference between the opinions of faculty members and directors of departments in future University (private university) samples before applying the management systems specification for educational institutions ISO 21001:2018 Prior to the implementation of ISO 21001:2018 Management Systems Specification, this is evidence of the acceptance of zero-enforcement prior to the implementation of the ISO 21001:2018 Management Systems specification and the improvement of performance quality in higher education institutions.

Then, another measure is conducted which included the differences between the opinions of Teaching Staff Members and Managers - Future University samples in the application (Post) and the opinions of Teaching Staff Members and Managers of the departments in the application (Post). The results of the conducted stage are illustrated in Table 4.

Table 4 confirms that there are statistically significant differences after the implementation of the ISO 21001:2018 Management Systems specification (0.000), which is less than (0.05) which is evidence of statistically significant differences between the opinions of different Teaching Staff Members and Managers of departments - Future University (Private University). Following the implementation of ISO 21001:2018 Management Systems Specification and the opinions of Teaching Staff Members and department managers after applying the Management Systems specification for educational institutions ISO 21001:2018 to improve the quality of performance in higher education institutions. Moreover, it is also confirmed that there are no statistically significant differences in private universities in applying the ISO 21001:2018 Management Systems specification to educational institutions in partially improving performance quality".

Table4. Independent Samples T Test—for Post Application

Standard Items	T	df	Sig. (2-tailed)
Context of the organization	3.604	97	.000
Leadership	4.084	97	.000
Planning	4.658	97	.000
Support	4.008	97	.000
Operation	4.530	97	.000
Performance Evaluation	6.183	97	.000
Improvement	6.573	97	.000
Total	5.474	97	.000

7. CONCLUSION AND FUTURE WORK

This paper focused on the standards for a qualified educational process, the paper proposed ISO 9001 standards base educational process fulfills the implementation requirements in different perspectives with the ability to adapt to other aligned standards. In this research, a hypothesis was examined which stated: "There are no statistically significant differences before and after applying the ISO 21001:2018 management systems standard for educational institutions in improving the quality of performance in higher education institutions".

The result of the T-test for double samples at Future University in Egypt. The results confirmed the existence of statistically significant differences towards the specification items (improving the quality of performance) at a confidence level of 99%, as the significance of the test was all less than the level of significance of 1%, and these differences were in favor of the test.

The post-test and the percentage of improvement in the ETA coefficient ranged between 72.14% to 85.74% and at the level of overall performance quality 87.37%, and the improvement rates ranged from 25.19% to 33.61%, which reflects the impact of applying the management systems specification for educational institutions ISO 21001: 2018 in improving the quality of performance in educational institutions higher.

From the above, it is clear that the first hypothesis is rejected, stating that "there are no statistically significant differences before and after applying the ISO 21001:2018 management systems specification for educational institutions in improving the quality of performance in higher education institutions" and accepting the alternative hypothesis.

Fulfilling the study highlighted a set of recommendations which can be summarized as follows:

- Update the functional standards and mechanisms for assessing performance at universities periodically and continuously, to keep up with future work developments and needs to make the evaluation process and its results in college management the basis of individuals' desire to work, taking into account the privacy and responsibility of each job.
- Make the evaluation process more than once a year, make initial models within departments, and then make key models between colleges and universities.
- University departments should promote a culture of performance appraisal among all employees, but performance evaluation is not an oversight function of employees, but rather a response to and treatment of negative points. The aim of a system of assessment of developmental and structuring performance should be more than an objective through which accountability and accountability are carried out in order to punish or reward.
- Performance evaluators should be qualified to track the performance of staff; both heads and supervisors; all the time to discuss their needs, problems, and proposals at work. This leads to developing them, eliminating negative points in the college, and increasing positive points.
- The need for feedback follows the performance appraisal process so that employees can identify the results of their assessment, identify weaknesses, attempt to eliminate them, avoid them, identify strengths, encourage them, and motivate them so that the outcome of the assessment is more meaningful to the employee's actual performance.
- The activation and application of the system of moral and material rewards and incentives at university colleges and their link to the performance appraisal process, which has a significant impact on the progress and development of the level of performance.

- Linking training programs and plans to the results of performance analysis and the need to prepare well for courses of action, to cover performance deficiencies, improve performance, increase skills, and develop their behavior.
- Keep in mind that the performance appraisal of workers is based on job efficiency and not on position so that everyone is equal in the assessment process.
- Universities must promote a culture of performance appraisal among all workers that performance appraisal is not a waste for employee monitoring, but rather for addressing and addressing negative points, with the goal of performance assessment being more developmental and constructive than an accountable and accountable goal for punishment Or reward.
- Quality attention should be self-interested, generated from within the entire staff of the Organization, quality-aware and cost-effective, and quality-oriented programs needed to continuously improve quality.
- For administrators and members of the administrative machinery within universities, their well-trained computer training can help them to organize the work at the university in an excellent way that enables them to meet the demands of both students and university students.

Future directions following this research could focus on the application of the management systems specification to educational institutions 21001-2018 which has affected the career satisfaction of staff in higher education institutions. Another direction is developing a strategy to overcome the obstacles to the implementation of management systems for educational institutions 21001-2018 in educational institutions. Moreover, comparing theoretical and practical colleges could be conducted.

8. REFERENCES

[1] A. Mostafa, Y. Helmy, A. M. Idrees, "A Configurable Mining Approach for Learning Services", Future Computing and Informatics Journal, Vol. 6, No. 1, 2021, pp. 16-24.

[2] A. E. Khedr, A. M. Idrees, A. E.-F. Hegazy, S. El-Shewy, "A proposed configurable approach for recommendation systems via data mining techniques", Enterprise Information Systems, Vol. 12, No. 2, 2018.

[3] A. E. Khedr, A. M. Idrees, R. Salem, "Enhancing the e-learning system based on a novel tasks' classification load-balancing algorithm", PeerJ Computer Science, Vol. 7:e669, 2021.

[4] A. M. Mostafa, Y. M. Helmy, A. E. Khedr, A. M. Idrees, "A Proposed Architectural Framework For Generating Personalized Users' Query Response", Journal Of Southwest Jiaotong University, Vol. 55, No. 5, 2020.

[5] N. Sultan, A. E. Khedr, A. M. Idrees, S. Kholeif, "Data Mining Approach for Detecting Key Performance Indicators", Journal of Artificial Intelligence, Vol. 10, No. 2, 2017, pp. 59-65.

[6] D. H. A. Hassouna, A. E. Khedr, A. M. Idrees, A. I. ElSeddawy, "Intelligent Personalized System for Enhancing the Quality of Learning", Journal of Theoretical and Applied Information Technology, Vol. 98, No. 13, 2020, pp. 2199-2213.

[7] A. E. Khedr, A. M. Idrees, "Adapting Load Balancing Techniques for Improving the Performance of e-Learning Educational Process", Journal of Computers, Vol. 12, No. 3, 2017, pp. 250-257.

[8] A. E. Khedr, A. M. Idrees, E. Shabaan, "Automated Ham-Spam Lexicon Generation Based on Semantic Relations Extraction", International Journal of e-Collaboration, Vol. 16, No. 2, 2020, pp. 45-64.

[9] A. E. Khedr, A. I. El Seddawy, A. M. Idrees, "Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS", International Journal of Innovative Research in Computer Science & Technology, Vol. 2, No. 6, 2014, pp. 111-118.

[10] A. E. Khedr, A. M. Idrees, "Enhanced e-Learning System for e-Courses Based on Cloud Computing", Journal of Computers, Vol. 12, No. 1, 2017.

[11] S. Naiem, A. E. Khedr, M. Marie, A. M. Idrees, "Distributed Denial of Services Attacks and Their Prevention in Cloud Services", Journal of Theoretical and Applied Information Technology, Vol. 100, No. 4, 2022, pp. 1170-1181.

[12] A. E. Khedr, A. M. Idrees, F. K. Alsheref, "A Proposed Framework to Explore Semantic Relations for Learning Process Management", International Journal of e-Collaboration, Vol. 15, No. 4, 2019.

[13] A. M. Idrees, A. E. Khedr, "A Collaborative Mining based Decision Support Model for Granting Personal Loans in Banking Sector", International Journal of E-Services and Mobile Applications, Vol. 14, No. 3, 2021.

[14] A. M. Idrees, A. I. El Seddawy, M. EL Moaaz, "A Proposed Mining Based Business Continuity Information

System for Educational Institutes", Journal of Computer Science, Vol. 15, No. 8, 2019, pp. 1133-1149.

[15] J. A. Seniwoliba, R. N. Yakubu, "An Analysis of the Quality Assurance Policies in a Ghanian University", Educational Research and Reviews, Vol. 10, No. 16, 2015, pp. 2331-2339.

[16] V. Teeroovengadum, T. J. Kamalanabhan, A. K. Seebaluck, "Measuring service quality in higher education: Development of a hierarchical model (HESQUAL)", Quality Assurance in Education, Vol. 24, No. 2, 2016, pp.244-258.

]17] P. Swanzy, A. Potts, "Quality Assurance Strategies in Higher Education The Case of Ghanaian Polytechnics", Educational Research and Prespectives, Vol. 44, 2017, pp. 100-127.

[18] J. S. Mtebe, C. Raphael, "Key factors in learners' satisfaction with the e-learning system at the University of Dar es Salaam", Australasian Journal of Educational Technology, Vol. 34, No. 4, 2018, pp.107-122.

[19] W. H. Delone, E. R. McLean, "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update", Journal of Management Information Systems, Vol. 19, No. 4, 2003, pp. 9-30

[20] M. Seyfried, P. Pohlenz, "Assessing quality assurance in higher education: quality managers' perceptions of effectiveness", European Journal of Higher Education, Vol. 8, No. 3, 2018, pp. 258-271.

[21] D. T. Do, C. L. Le, T. V. Giang, "The correlation between internal quality assurance and the formation of quality culture in Vietnam higher education: A case study in Ho Chi Minh city",. European Journal of Educational Research, Vol. 9, No. 2, 2020, pp. 499-509.

[22] M. H. R. Joarder, M. A. Ashraf, S. R. A. Ratan, "Quality education from faculty management perspectives in private higher education: does faculty commitment behavior mediate?". International Journal of Education and Practice, Vol. 8, No. 1, 2020, pp. 190-206.

Elimination of CM Noise from SMPS Circuit using EMI Filter

Case Study

Venkata Sai Charishma Pathala
Department of EECE, Gandhi Institute of Technology and Management
Visakhapatnam, India
cpathala@gitam.in

V.Y. Jayasree Pappu
Department of EECE, Gandhi Institute of Technology and Management
Visakhapatnam, India
jpappu@gitam.edu

Abstract *– The electronic devices are exposed to external electromagnetic signals that produce an unwanted signal called noise in the circuit, which causes electromagnetic interference [EMI] problems. It occurs in two modes: radiated mode and conducted mode. In the radiation mode, the shielding technique is used for radiation mode, in conduction mode filtering technique is used. The design of an EMI filter depends upon the type of noise generated by the Switched Mode Power supply circuit [SMPS]. The SMPS circuit used in this paper is a DC-DC power converter, the Boost converter is a step-up converter and Buck converter is step down converter are considered as equipment for generation of noise, the Line Impedance Stabilization Network [LISN]is used for generating the common output impedance to the power converters, the EMI filters are designed to eliminate noise generated by the circuits. There noise generated by this power converters is Common Mode [CM] noise and Differential Mode [DM] noise. The separation of noise from the equipment is done by using a noise separator. In this paper, CM noise generated by these power converters is eliminated by designing an EMI filter called an inductor filter and a PI filter. The comparison between the LC inductor filter and the PI filter for the boost and buck converters is observed. The PI filter has better performance characteristics when compared to the inductor filter for both SMPS circuits as per the Comité International Special des Perturbations Radioélectriques [CISPR] standards. This standard gives the conducted emission range for different electronic devices.*

Keywords*: electromagnetic interference [EMI], Switched Mode Power Supply circuit [SMPS], Boost converter, Buck converter, Common Mode and Differential Mode Noise, LC Inductor filter, PI filter, CISPR standards*

1. INTRODUCTION

The electrical and electronic systems are affected by the external electromagnetic signals, which produce electromagnetic interference [EMI] in the circuit. This interference occurs from one electronic system to another due to the electromagnetic fields generated during the operation of electronic systems [1]. EMI obtained in two ways Man –Made Source and Natural Source, it occurs in two ways Radiated mode and Conducted mode [2]. The mitigation of EMI through radiation mode is done by shielding technique [3]. The EMI filters are used to eliminate interference in the conduction mode [4]. The interference that occurs in the conducted mode is due to the noises present in the circuit. The noises are DM noise and CM noise [5]. Power converters, power electronic devices, and power supply devices are widely used in many applications, like military, industrial, and aerospace. The compatibility of the power converter is based on the noise generated and its propagation path. The different types of power converters are considered for the analysis of their electromagnetic compatibility. The DC-DC power converter is more compatible than AC-DC power converters [6]. In this paper, the DC-DC power converters called Boost converters and Buck converters are considered as equipment that generates noise inside the circuit [7]. A boost converter increases the voltage at the output stage, whereas a buck converter decreases the output voltage. The LISN is considered an input to the equipment, which has a constant output impedance of 50 ohms. The power converters generate noises called CM noise and DM noise [8]. The separation of these noises from this equipment is done using a noise separator. In this paper, DM noise is eliminated by using a DM noise separator. The EMI filter is designed to eliminate the CM noise generated in the DC-DC power converter [9]. The electronic equipment has some EMI specification lim-

its, and to satisfy those limits, EMI filters are designed. There are two types of EMI filters. Active and passive EMI filters, the performance of the filter depends upon the insertion loss, cut-off frequency of the circuit, and input and output impedance [9]. The attenuation of the signal at the filter output of a given frequency is measured by its insertion loss. Insertion loss is defined as the ratio of the signal at the input stage of the filter [V1] to the signal at the output stage of the filter [10].

$$\mathrm{IL[dB]} = 20\log_{10}\frac{V_1}{V_2} \quad (1)$$

The design of power converters with LISN is done in the MULTISIM software. The noise generated by this power converter is measured by using the noise separator, and the elimination of noise is done by designing the EMI filters like the inductor and PI filter. The total design process and calculation of noise and insertion loss are obtained from the MULTISIM software. The EMI filter results with EMI specification limits are plotted in the MATLAB tool.

2. MEASUREMENT SETUP

The measurement setup consists of power converters called Boost and Buck converters, which are considered the equipment for testing the two-stage LISN circuit. DC power supply to run the equipment. The noise generated by this power converter is separated using a noise separator, which produces CM and DM noise. The elimination of these noises is done by using an EMI filter.

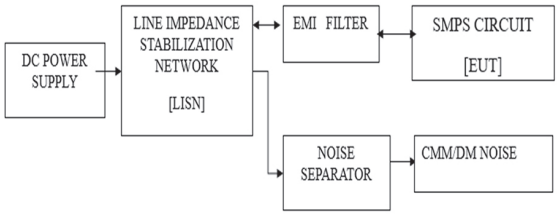


Fig.1. Block Diagram for total Measurement Setup

2.1. LINE IMPEDANCE STABILIZATION NETWORK

The LISN is placed between the power supply and the power converter [equipment] to maintain the known impedance. The LISN has an output impedance of 50 ohms; it acts as a low-pass filter.

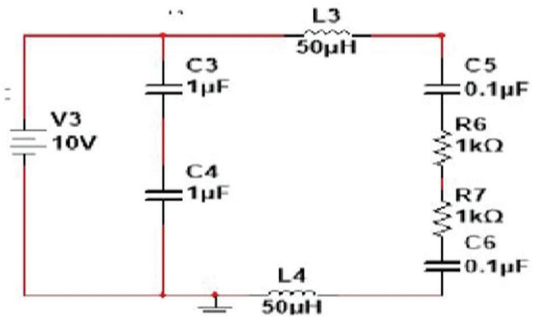


Fig. 2. LISN circuit

2. 2. POWER CONVERTERS

The converters used as equipment for the generation of noise inside the circuit are the Boost converter and the Buck converter; these are DC-DC converters. The noise generation is due to the switching action of diodes and transistors; the Boost circuit acts as a step-up converter, and the Buck circuit acts as a step-down converter. The selection of components for the power converter is calculated by considering the input voltage and frequency of operation. The conducted signal has a frequency range of 150 kHz to 3 MHz.

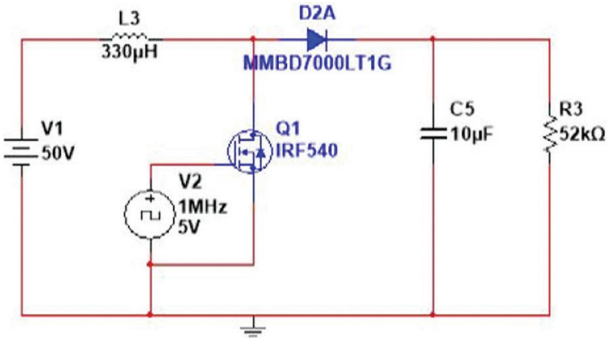


Fig. 3. Boost circuit

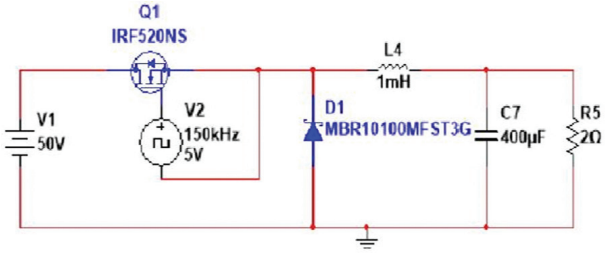


Fig. 4. Buck circuit

2. 3. NOISE SEPARATOR

The calculation of total noise inside the equipment is measured by using the noise separator; there are different types of noise separators. In this paper, the DM noise separator is considered. The total noise is obtained at the output of the LISN when it is connected to the equipment, and the CM noise and DM noise are calculated using the below formulae

$$V_{CM} = \frac{V_2+V_1}{2} \quad (2)$$

$$V_{DM} = \frac{V_2-V_1}{2} \quad (3)$$

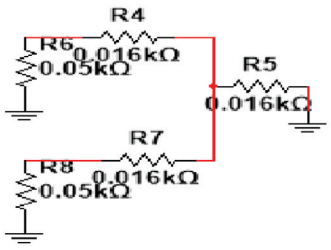


Fig. 5. DM noise separator

2. 4. EMI FILTER

The EMI filters are used to protect the electronic devices from high-power signals generated by other electronic devices. The design of filter characteristics depends upon the noise generated by the devices. Active and passive EMI filters are used for the reduction of noise in electronic devices. The LC inductor filter and PI filter are designed in this paper for the elimination of noise generated by the devices. The PI filter is considered because it gives a high output voltage and a very good ripple factor. The LC inductor has better voltage regulation and a very low ripple factor.

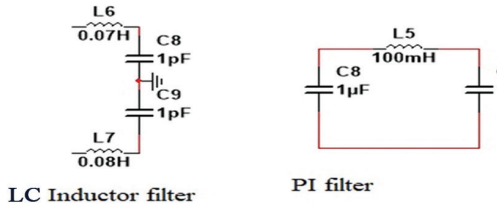


Fig. 6. Filters used in measurement setup

3. NOISE MEASUREMENT FOR BOOST CIRCUIT

The measurement of CM noise from EUT using LISN is shown in Fig. 2. The EUT is supplied by DC voltage of 50V through LISN circuit. The noise generated due to diodes and switches in the EUT, the noise is measured at the output of the noise separator by connecting through LISN, and the CM noise is obtained at the end of the noise separator by eliminating the DM noise [10]. The design process and noise calculation are done in Multisim software. The output graphs for CM noise with and without filter are plotted in the MATLAB simulation tool.

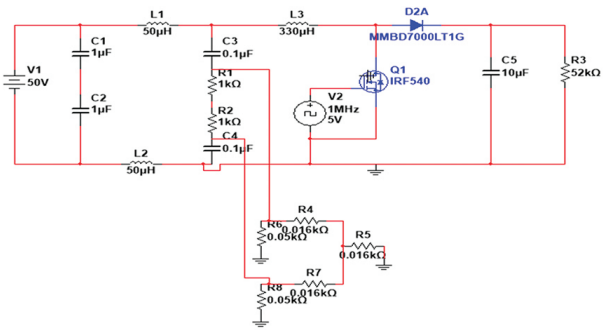


Fig. 7. Noise Measurement Setup for Boost circuit

The conducted emission ranges from 150 kHz to 3 MHz, so the filters are designed to satisfy the EMI regulation limits. For conducted emissions, the CISPR 22 standard is used as a limit line for EMI regulation. This standard deals with the conducted emissions range for different devices, like class A and class B devices.

The measurement of EMI in conducted mode requires a line impedance stabilisation network of 50 mH to be placed between the EUT and DC power supply to find impedance for noise measurement. The EUT used is an SMPS circuit of a boost converter that acts as a noise source due to diodes, transistors, and switching elements [11-13]. The noise separator is placed across the output of the LISN when EUT is connected to the LISN circuit. DM and CM noises are separated at the output of the noise separator. The noise separator used is a DM noise separator, which eliminates DM noise and produces CM noise.

Table 1. The limits for class A device in conducted mode as per CISPR standard

FREQUENCY [MHz]	µV	dB[µV]
0.15-0.5	8912.5	68
0.5-30	4467	62

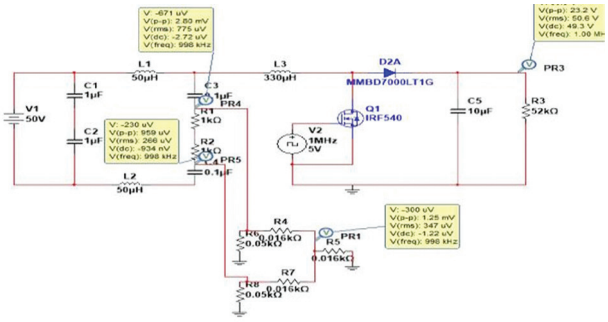


Fig. 8. CM noise obtained at the output of the Noise Separator

The noise obtained at the output of the noise separator is CM noise with a value of 266 mV, so it is converted to check whether the circuit meets the EMI specification limit [14]. By using equation 4, it is converted into a value of 108. So, it does not satisfy the CISPR standard limits shown in Table 1.

$$1\mathrm{dB}_{\mu\mathrm{V}} = 20\log_{10}[\mu\mathrm{V}] \tag{4}$$

So, the design of an EMI filter is necessary for the reduction of noise in the circuit. The insertion loss required for the design of an EMI filter depends upon how much the amount of noise increases as per the EMI limit. In order to get better performance out of the circuit, the filter should satisfy these conditions: the CM noise obtained at the output of the noise separator, in dBµV, is the noise limit as per the CISPR standard.

$$IL_{CM} >= V_{CM}(\mathrm{dB\mu V}) - \mathrm{V_{LIMIT}}(\mathrm{dB\mu V}) \tag{4}$$

3. 1. EMI FILTER DESIGN FOR BOOST CIRCUIT

The passive EMI filter was designed after measuring insertion loss; the filter used for elimination of CM noise is the LC inductor filter. This filter is placed between LISN and EUT. The noise measurement is done at the output of the noise separator, and in the MULTISIM software, it is observed that 296 is obtained when a filter is added to the equipment. The obtained noise is converted into dB with a value of 49 dB by using equation 4, and graphs are plotted in the MATLAB simulation tool. From the simulations, it is observed that the output noise is decreased and is within the CISPR limit after placing the LC inductor filter.

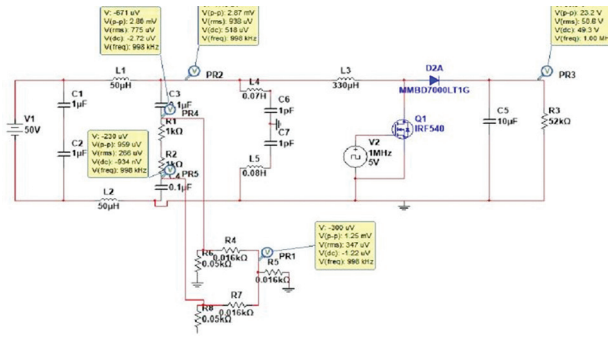


Fig. 9. CM noise obtained at the output of the Noise Separator when LC Inductor filter is connected

The PI filter is considered for the measurement of CM noise, and it is placed in between EUT and LISN. The design is done in Multisim software. The noise measurement is done at the output of the noise separator in the MULTISIM software, and it is observed that 62 is obtained when filtering is added to the equipment. So, it is converted into dBµV by using the equation 4, and the value obtained is 35 dBµV. Graphs are plotted in the MATLAB simulation tool.

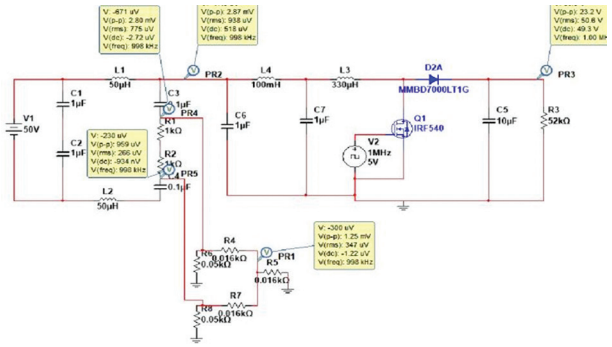


Fig. 10. CM noise obtained at the output of the Noise Separator when PI filter is connected

In Fig. 11. Shows the graph for CM noise measurement with filter and without filters. As per CISPR standards, it is observed that the PI filter has better performance characteristics when compared to the LC inductor filter.

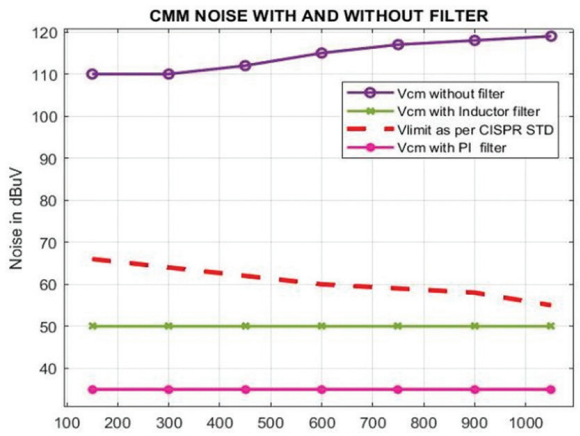


Fig. 11. Noise measurement as per CISPR standard with and without EMI filter

The filters designed for the elimination of noise are the PI filter and the inductor filter, which generate an insertion loss greater than the insertion loss without a filter.

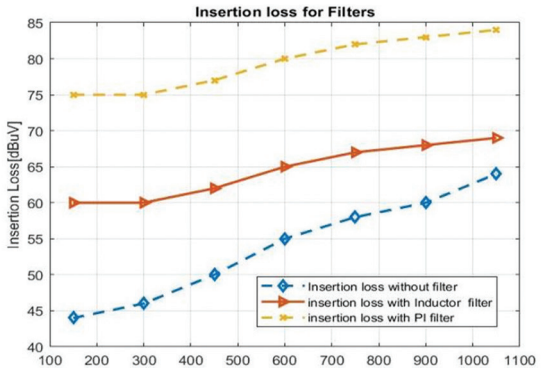


Fig.12. Insertion loss for LC Inductor filter, PI filter and without filter

4. NOISE MEASUREMENT SETUP FOR BUCK CIRCUIT

The noise is measured from the buck circuit by placing the noise separator between EUT and LISN. The total measurement is done with software simulation tools. The buck converter acts as a noise source in this setup. The noise separator used in the circuit is a DM noise separator. The EUT is supplied by DC voltage of 50V through LISN circuit. The CM noise is obtained at the end of the noise separator by eliminating the DM noise. The design process and noise calculation are done in Multisim software. The output graphs for CM noise with and without filter are plotted in the MATLAB simulation tool.

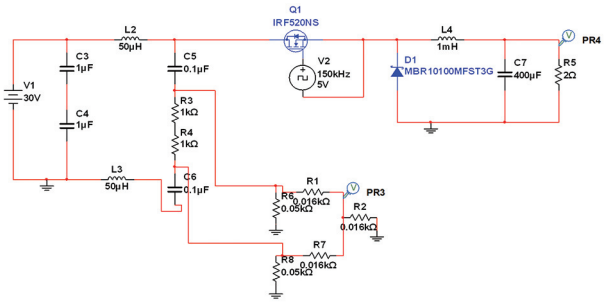


Fig.13. Noise Measurement Setup for Buck circuit

The EUT used is an SMPS circuit of a Buck converter, which acts as a noise source due to diodes, transistors, and switching elements. The noise separator is placed across the output of the LISN. The CM noise is generated at the output of the noise separator.

The noise obtained at the output of the noise separator is CM noise at 421 mV, so it is converted into dBµV by using equation 4, to check whether the circuit meets the EMI specification limit. The noise obtained is 112 dBµV at the output, so it does not satisfy the CISPR standard limits shown in table 1. So, the design of an EMI filter is necessary for the reduction of noise

in the circuit. The insertion loss required for the design of an EMI filter depends upon how much the amount of noise increases as per the EMI limit. In order to get better performance out of the circuit, the filter should satisfy these conditions, is the CM noise obtained at the output of the noise separator, in dBμV, is the noise limit as per the CISPR standard.

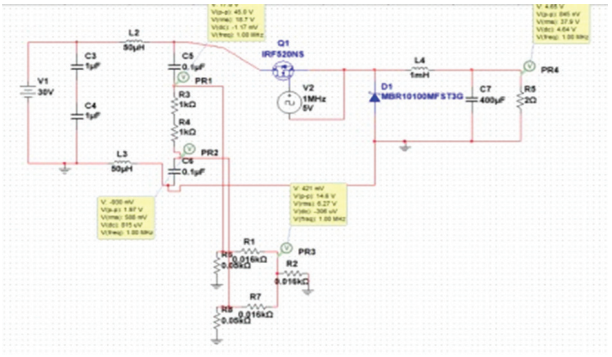


Fig.14. CM noise obtained at the output of the Noise Separator for Buck circuit

4.1. EMI FILTER DESIGN FOR BUCK CIRCUIT

The passive EMI filter was designed after measuring insertion loss, the filter used for elimination of CM noise is the LC inductor filter. This filter is placed between LISN and EUT. This filter is placed between LISN and EUT. The noise measurement is done at the output of the noise separator in the MULTISIM software. It is observed that this is obtained when a filter is added to the equipment. The obtained noise is converted into dB with a value of 56 dB by using equation 4, and graphs are plotted in the MATLAB simulation tool. From the simulations, it is observed that the output noise is decreased and is within the CISPR limit after placing the LC inductor filter.

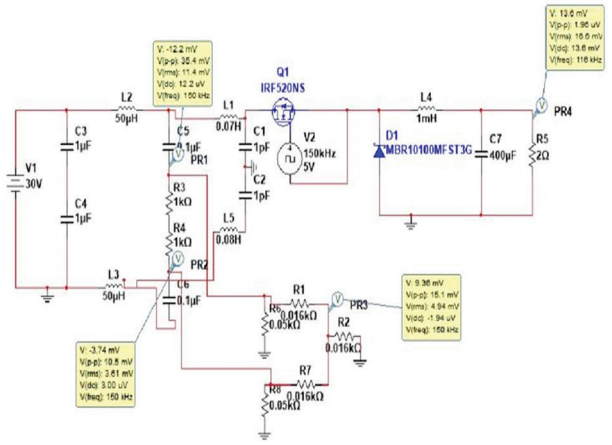


Fig.15. CM noise obtained at the output of the Noise Separator when LC Inductor filter is connected

The PI filter is considered for the measurement of CM noise, and it is placed in between EUT and LISN. This filter is placed between LISN and EUT. The noise measurement is done at the output of the noise separator, and in the MULTISIM software, it is observed that 59 is obtained when filtering is added to the equipment. The obtained noise is converted into dB with a value of 38 dB by using equation 4, and graphs are plotted in the MATLAB simulation tool. From the simulations, it is observed that the output noise is decreased and is within the CISPR limit after placing the PI filter.

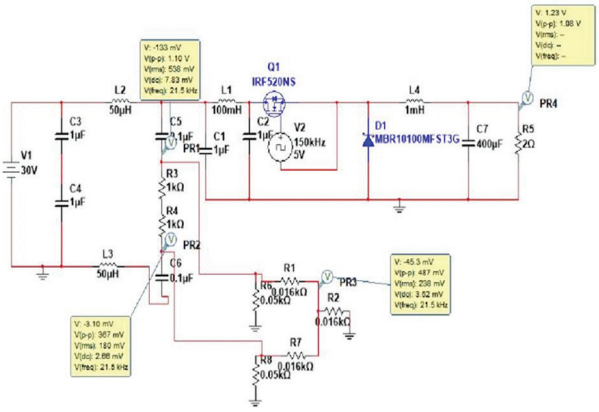


Fig.16. CM noise obtained at the output of the Noise Separator when PI Inductor filter is connected

In Fig. 17, which shows the graph for CM noise measurement with and without filters as per CISPR standards, it is observed that the PI filter has better performance characteristics when compared to the LC inductor filter.

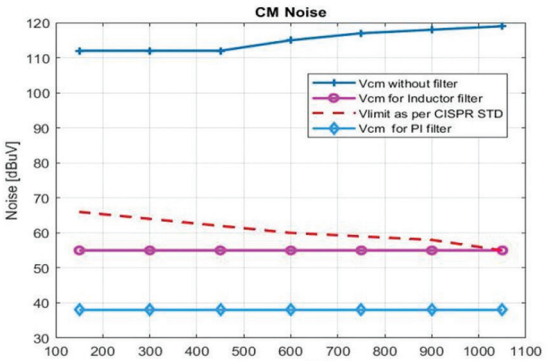


Fig.17. Noise measurement as per CISPR standard with and without EMI filter

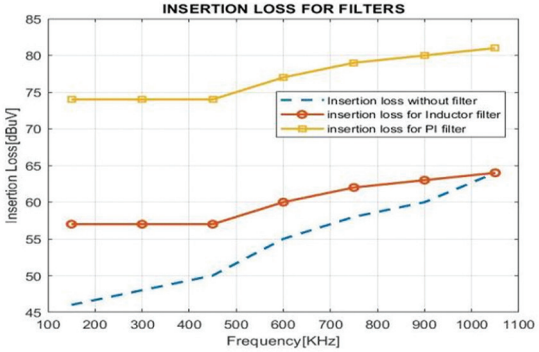


Fig.18. Insertion loss for LC Inductor filter, PI filter and without filter

5. RESULTS

The conducted signal radiates in the frequency range of 150 kHz to 30 MHz. The measurement setup consists of Boost and Buck power converters treated as equipment with LISN, and the noise generated by these power converters without filter is shown in the below table. The insertion loss without a filter is around 40-45dBμV for a frequency range of 150–1100 kHz for both power converters. The insertion loss for the designed filters, the PI filter and the inductor filter, is higher than the insertion loss for the setup with no filter. The inductor filter for the boost and buck converters produces an insertion loss of around 55–60 dBμV in the 150KHz–1MHz range for both the SMPS circuits. The PI filter produced an insertion loss of 75 dBμV for 150 KHz–1 MHz. The use of Multisim software made easier for the design and calculation of noise. The graphs are plotted in the MATLAB tool. The results obtained for SMPS circuits are given below. The measured noise and filters used for elimination of noise with their values are mentioned in Table 2.

Table 2. Noise levels in the measurement setup

Specifications	Boost circuit	Buck circuit
NOISE measured without filter	108 dBμV	112 dBμV
NOISE at LC filter	49 dBμV	56 dBμV
NOISE at PI filter	35 dBμV	38 dBμV

From the table, it was observed that the buck circuit generates more noise than the boost circuit, and the filter used for noise elimination is designed as per the CISPR standard. The PI filter has better performance than the LC inductor filter

6. CONCLUSION

The conductive emission ranges from 150 KHz to 30 MHz, the measurement setup generates noise inside the equipment, and the elimination of noise is done by designing the appropriate EMI filter, which depends upon the frequency of operation and insertion loss. By measuring the insertion loss, the filter was designed. The LC inductor filter and PI filter generate more insertion loss than a measurement setup without filters. The noise generated by the boost circuit is 108 dBμV and by the buck circuit is 112 dBμV.

The filter design should meet the EMI standards. The CISPR 22 standard is considered for the filter design, and the designed filter satisfied the CISPR STD with these values. The noise at the output of the LC inductor filter for the boost converter is 49 dBμV. The noise at the PI filter output for the Boost converter is 35 dBμV. The noise at the output of the LC inductor filter for the Buck converter is 56 dBμV. The noise at the PI filter output for the Buck converter is 38 dBμV. Hence, PI filters meet good EMI standards in the conducted emissions range of 150 kHz–1 MHz

7. REFERENCES

[1] C. R. Paul, "Introduction to Electromagnetic Compatibility", 2nd Edition, Hoboken, Wiley, New Jersey, 2006.

[2] V. S. C. Pathala, P. V. Y. Jayasree, "Design of a Shielded Room against EMP Signal as per MIL-STD 461", Engineering, Technology & Applied Science Research, Vol. 13, No. 1, 2023, pp. 9837-9842.

[3] P. V. Y. Jayasree, V. Baba, P. Rao, P. Lakshman, "Analysis of Shielding Effectiveness of Single, Double and Laminated Shields for Oblique Incidence of EM Waves", Progress in Electromagnetics Research B, Vol. 22, 2010, pp. 187-202.

[4] A. Rabat, P. Bonnet, K. E. K. Drissi, S. Girard, "Analytical Formulation for Shielding Effectiveness of a Lossy Enclosure Containing Apertures", IEEE Transactions on Electromagnetic Compatibility, Vol. 60, No. 5, 2017, pp. 1384-1392.

[5] X. Chang, W. Chen, Y. Yang, K. Wang, X. Yang, "Research and realization of a novel active common-mode EMI filter", Proceedings of the IEEE Applied Power Electronics Conference and Exposition, Charlotte, NC, USA, 15-19 March 2015, pp. 1941-1945.

[6] K. Mainali, R. Oruganti, "Conducted EMI Mitigation Techniques for Switch-Mode Power Converters: A Survey", IEEE Transactions on Power Electronics, Vol. 25, No. 9, 2010, pp. 2344-2356.

[7] D. Hamza, M. Sawan, P. K. Jain, "Suppression of common-mode input electromagnetic interference noise in DC-DC converters using the active filtering method", IET Trans. Power Electron., Vol. 4, No. 7, 2011, pp. 776-784.

[8] S. Wang, Y. Y. Maillet, F. Wang, D. Boroyevich, and R. Burgos, "Investigation of hybrid EMI filters for common-mode EMI suppression in a motor drive system", IEEE Transactions on Power Electronics, Vol. 25, No. 4, 2010, pp. 1034-1045.

[9] D. Shin et al. "Analysis and design guide of active EMI filter in a compact package for reduction of common-mode conducted emissions", IEEE Transactions on Electromagnetic Compatibility, Vol. 57, No. 4, 2015, pp. 660-671.

[10] V K. Prasad, "Engineering Electromagnetic Compatibility", Wiley-IEEE Press, 2001.

[11] S. Banerjee, A. Bandyopadhyay, A. Mukherjee, A. Das, R. Bag, “Random Valued Impulse Noise Removal Using Region Based Detection Approach”, Engineering, Technology & Applied Science Research, Vol. 7, No. 6, 2017, pp. 2288-2292.

[12] M. C. Caponet, F. Profumo, L. Ferraris, A. Bertoz, D. Marzella, "Common and DMnoise separation: comparison of two different approaches", Proceedings of the IEEE 32nd Annual Power Electronics Specialists Conference, 2001, pp. 1383-1388.

[13] Wenjie Chen, Xu Yang, Zhaoan Wang, “An active EMI Filtering Technique for Improving Passive Filter LowFrequency Performance”, IEEE Transactions on Electromagnetic Compatibility, Vol. 48, No. 1, 2006.

[14] K. Raggl, T. Nussbaumer, J. W. Kolar, “Guideline for a Simplified Differential-Mode EMI Filter Design”, IEEE Transactions on Industrial Electronics, Vol. 48, No. 7, 2009, pp. 1865-1872.

Design and analysis of three phase inverter based Solar PV powered single switch Buck-Boost converter with reduced THD for industrial applications

Case Study

Maheshwari L.
Vels Institute of Science Technology and Advanced Studies,
Chennai, Tamilnadu, India
saimahi.eee@gmail.com

Premila T. R.
Vels Institute of Science Technology and Advanced Studies,
Chennai, Tamilnadu, India
trpremilashaji@gmail.com

Abstract *– In recent times development of economical & feasible eco-friendly renewable source powered power electronic converters have become more attractive in multiple areas such as automotive, house appliances and industrial applications etc., Bucking and boosting of voltage according to the requirement is also much needed. So, this work proposes a solar photovoltaic (SPV) powered single switch buck-boost converter for industrial applications which reduces implementation cost, minimal voltage and current stress across the capacitors and diodes and less switching power losses. The work structure comprises of solar PV source with modified perturbation and observation (P&O) algorithm based maximum power point tracker (MPPT), single switch buck-boost dc-dc converter, battery backup to store excess energy, three phase inverter with sinusoidal pulse width modulation (PWM) to find optimal switching angles for harmonic control and 3 phase (Φ) induction motor load. Along with the proposed single switch buck-boost converter, an EMI input filter is included to investigate the harmonic reduction. It also uses the Fast Fourier Transform (FFT) analysis with the reduction of supply voltage harmonics generated by the 3Φ induction motor drive.The DC link is connected with the VSI (voltage source inverter) and then it is converted to AC to feed the AC load. The VSI is controlled with a new PWM based total harmonic distortion (THD) reduction method. MATLAB/SIMULINK platform is used for performance analysis. This work also presents a steady state analysis of a rectifier-inverter topology. Using the method caled linearization, the dynamic equations of the proposed system is obtained.; The proposed single switch buck-boost topology provides an output voltage and current of 363V, 45.5A DC from 520V, 35A PV array. The proposed converter is employed to run a 3Φ induction motor with the rating of 440V, 15A AC. From the simulation results, it is found that the solar powered single switch buck-boost with MPPT is stable, efficient with minimal losses and less THD with better quality output.*

Keywords*: Solar PV source, Modified P&O algorithm based MPPT, Single switch Buck-boost dc-dc converter, Total Harmonic Distortion, Steady state analysis, Matlab/Simulink*

1. INTRODUCTION

Recent years, Renewable energy source powered dc-dc converters have become the most appropriate sources for any kind of power generation because of its abundant availability, less maintenance requirements, pollution free, and lower operational costs. Moreover, solar PV sources have got great attention over other energy sources. To increase the efficiency of the SPV system, MPPT algorithm is much needed. Though there are many MPPT algorithms in the literature, P&O algorithm is the simplest one which works towards maximum power point using operating point irrespective of the atmospheric conditions, aging, etc., Generally this method compares the SPV current (Ipv) and SPV voltage (Vpv), introduce delay then controls the values of Ipv and Vpv, resulting in less system response, fluctuation in finding MPP during steady state [1-3]. To overcome the issue mentioned, this work proposes a modified P&O algorithm that adjusts hysteresis bandwidth & reference step size. It uses hysteresis band and

auto-tuning perturbation step method. Here, the MPPT controllers drive the proposed converter by changing the dc-dc converter duty cycle with variable steps.

According to the authors, Islam et al [4] and Yasmin et al [5], DC/DC/AC converter topologies are mostly preferred in industries and manufacturing companies. In industrial systems, converters with DC-link and battery system, inverters are commonly preferred to provide better power quality, high voltage capability, lesser component stress and lower switching losses [6,7]. It is evident that the stepdown or/ step up converters are used in high power electronic systems for bucking or boosting purposes [8]. Khan et al., [9], Abdikarimuly et al., [10], and Pawar et al., [11] have modelled and simulated various types of buck-boost converters like fly back converter, SEPIC converter and cuk converter. However these converters easily stress the switches, high leakage inductance and less efficiency. Literature [12] also presents KY converters for buck and boost purpose. As the topology uses four power switches and capacitors, in some operational modes, it is suddenly charged. It results a current stress on capacitors, switches and diodes which leads to implementation problems. In this work, the authors propose a novel buck-boost converter topology with a power switch and two capacitors in series. The proposed topology can be operated in continuous conduction mode (CCM) with three different modes. One is ON mode and OFF. OFF mode has two working categories. It will not allow the parallel connected capacitors. The proposed topology uses one power switch, it has the simple control scheme and reduced switching power losses.

As said by the authors Dogga et al., [13] and Aly et al., [14], AC signal conversion from DC signal is done using an inverter with pulse width modulation technique (PWM). PWM signal generation is done with the comparison of a reference signal and a carrier signal. Sinusoidal or square wave signals are considered as reference signal. Saw tooth or a triangular wave signals are taken as carrier signal. PWM signals are generated by triggering the carrier signals by turning ON/OFF the inverter switches. This work uses a sinusoidal PWM technique which further reduces the switching losses [15].

Literature on steady state/ small signal analysis of SPV powered 3Φ induction motor have been published over the years. It was analysed by the authors [16] and [17]. Root locus technique and frequency dependent method was considered by them. Fallside et al., [18] have analyzed the stability of the proposed system. This method eliminates the effect of harmonics. The stability study of a induction motor based inverter and rectifier system have been analysed by the authors Lipo et al., [19]. This method eliminates stator voltage harmonics. It uses Nyquist stability criterion method. The relationship between voltage and current is considered for the analysis of stability and small signal modelling. Dc-dc converter connected with the inverter and 3Φ induction motor is considered and their stability is analysed by deriving the d-q voltage-flux-linkage. It has many advantages over previously discussed methods. Reduced variables are used in the steady state equations. The flux linkages are continuous in nature. It derives a transfer function is derived from the ratio between input and output signals.

The main disadvantages of power electronic loads are harmonics [20,21]. Voltages and currents present in the fundamental frequencies produce harmonics. A lot of literature dealt about the enhancement of power quality using various methods. According to the literature [22-24] there are several techniques which addresses the mitigation of harmonics. THD techniques are classified as passive, active, hybrid passive and active filter etc., Pires et al., [25] found that the passive filter technique is suitable for asymmetrical system having 10kV rating. The authors have used the optimization technique named genetic algorithm (GA) and the Ls and Rs values are taken as objective function. Banaei et al., [26] and Van de Sype et al., [27] have presented a novel passive filter technique for THD reduction. They have followed series and shunt passive filter method where variable impedance controlled by the thyristor is used as shunt filter. The combination of shunt with series type passive filter is called hybrid filter.

Low-voltage networks mostly prefer active harmonic filters due to the limitation posed on power converter rating. This method of harmonic elimination can be used in aircraft power system [28]. Active power filter with the asymmetrical inverter is introduced by the authors An, Le, and Dylan Dah-Chuan Lu [29]. In their findings, at high switching and line frequency is operate at line frequency one leg operated by the inverter legs. High frequency switching can be used to track the compensated current command. Number of researchers have worked to solve time delay problem with the help of neural network (NN) based active filters. p-q and id-iq theory can be controlled instantly with these control technique methods. Chakraborty et al., [30] and Akhil et al., [31] have analyzed p-q and id-iq techniques and also performed a comparative analysis between the techniques. Then they have concluded that better results are achieved by id-iq method. Filters installing methods are so versatile or unique in power electronics topologies. These disadvantages are rejected in switching methods. So this work tries to implement PWM based THD reduction technique.

This work is organized as follows. The first section deals about the view of the previous work done. Second section has block diagram, system description and results discussion of the proposed single switch buck-topology, stability analysis and THD reduction techniques. In third section simulation results and efficiency is discussed. Finally, the findings, disadvantages and the future works of the works are highlighted.

2. BLOCK DIAGRAM OF THE PROPOSED SYSTEM

Fig. 1. presents the pictorial representation of the proposed work. It consists of solar PV power generat-

ing unit, MPPT controller algorithm, single switch buck-boost converter, 3Φ VSI, sinusoidal PWM controller and AC load. SPV system uses a MPPT controller algorithm to track maximum power output. Here modified P&O algorithm based MPPT controller is used.

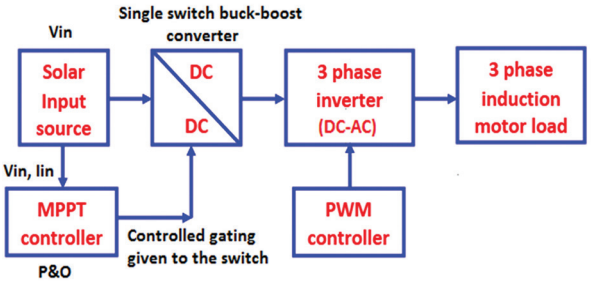


Fig. 1. Block diagram of three phase inverter based Solar PV powered single switch Buck-Boost converter

It utilizes a main inverter circuit to connect dc link to the induction motor. By controlling the switching period of the PWM controller, the inverter using IGBT, reduce harmonics, increase reliability, reduce the complexity, low switching stress power and less cost.

2.1. TOPOLOGY OF THE PROPOSED SYSTEM

The topology of three phase inverter based SPV powered single switch buck-boost converter with reduced THD for industrial applications is presented in Fig.2. It consists of SPV source voltage, single switch buck-boost, battery backup, three phase VSI and 3Φ induction motor.

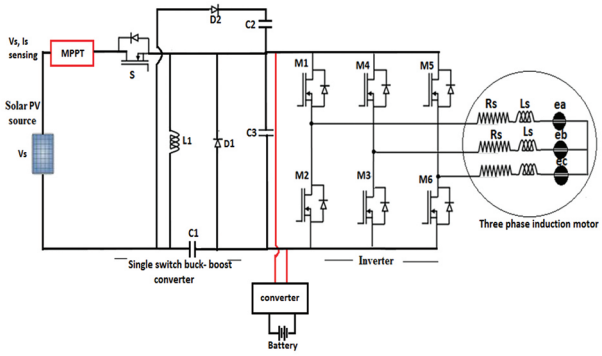


Fig. 2. Topology of the proposed system

2.2. MODELLING OF SOLAR PV SOURCE

SPV system and the single switch buck-boost converter is connected together. Fig.3. presents the equivalent circuit of a SPV cell. The Eq. (1) gives the I-V characteristics of a solar cell,

$$I = I_s - I_d \left[\exp\left(\frac{q * V}{a * k * T} \right) - 1 \right] \quad (1)$$

Where, k is the Boltzman constant (1.3806 x 10^{-23} J/K), Is is the Incident sunlight current, q is the electron charge (1.602 x 10-19 C), T is the Temperature of the p-n junction, a is diode ideality constant and I_d is Shockley diode equation. If the availability of SPV power is lesser/greater than the required dc power, then shortage/ excess energy charges/ discharges the energy storage system respectively. An dc voltage is inverted as ac voltage using three phase VSI. It is used to drive an 3Φinduction motor load torque of 57.3N-m.

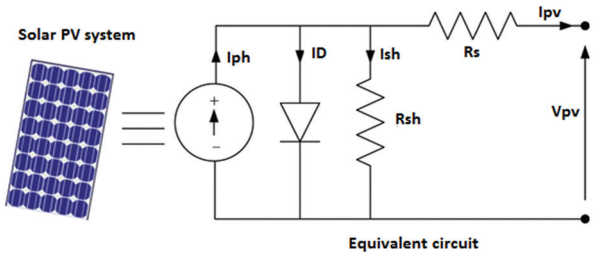


Fig. 3. Solar cell equivalent circuit

SPV system use modified P&O algorithm to track the maximum power obtained from the panel. Gate pulses obtained from the modified algorithms are used as a switching pulses for single switch buck-boost converter. 900 W/m^2 of insolation, 520V, 35A, 18.2kW is taken for simulation purposes.

The outcome from the simulation SPV voltage, current and power is presented in Fig. 4a, 4b and 4c. From the simulation outcome, it is observed that the obtained total solar power is 18.2 kW.

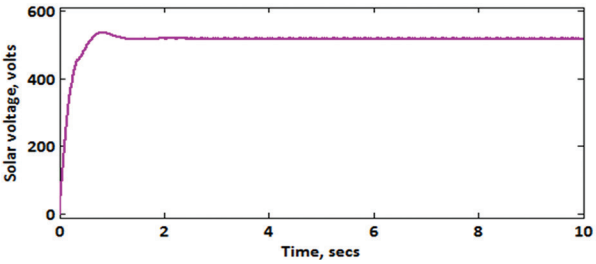


Fig. 4a. Solar voltage

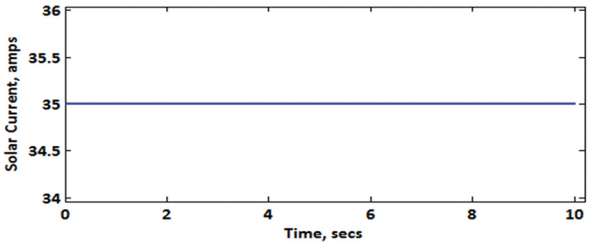


Fig. 4b. Solar current

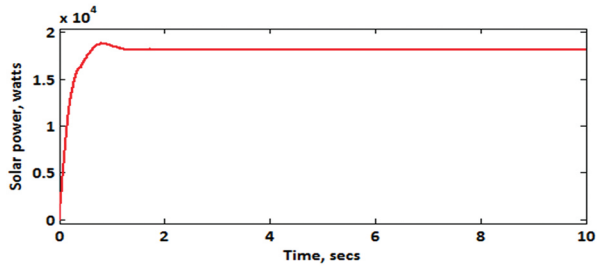


Fig. 4c. Solar power

Modeling of modified P&O algorithm is discussed here. Pictorial representation of the modified P&O is in Fig. 4d. In modified P&O tracking small perturbation is given for the duty cycle modification. This process is repeated till the maximum power is tracked. Till the system finds the steady state and maximum peak point, the algorithm oscillates around.

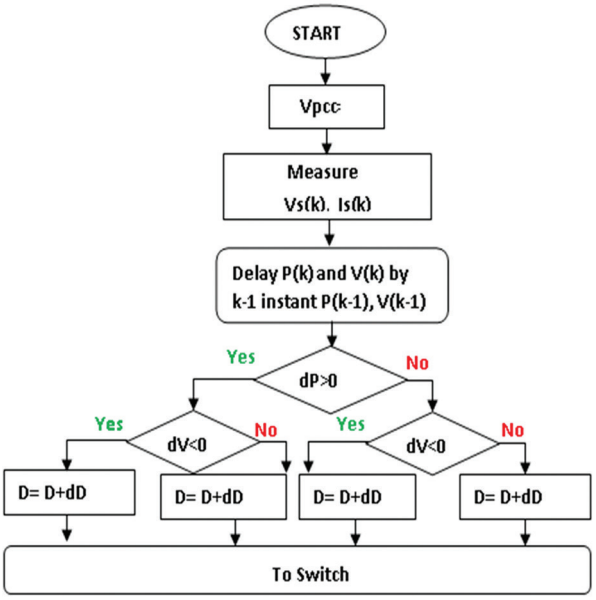


Fig. 4d. Flow chart for modified P &O algorithm

2.3. MODELLING OF 3Φ INDUCTION MACHINE

The electrical and mechanical equations of 3Φ induction machine drive system is expressed as follows,

$$\begin{bmatrix} V_{qs}^e \\ V_{ds}^e \\ V_{qr}^e \\ V_{dr}^e \end{bmatrix} = \begin{bmatrix} r_S + \frac{\rho}{\omega_r} X_{SS} & \frac{\omega_s}{\omega_r} X_{SS} & \frac{\rho}{\omega_r} X_M & \frac{\omega_s}{\omega_r} X_M \\ -\frac{\omega_s}{\omega_r} X_{ss} & r_s + \frac{\rho}{\omega_r} X_{ss} & -\frac{\omega_s}{\omega_r} X_M & \frac{\rho}{\omega_r} X_M \\ \frac{\rho}{\omega_r} X_M & s\frac{\omega_s}{\omega_r} X_M & r_r + \frac{\rho}{\omega_r} X_{rr} & s\frac{\omega_s}{\omega_r} X_{rr} \\ -s\frac{\omega_s}{\omega_r} X_M & \frac{\rho}{\omega_r} X_M & -s\frac{\omega_s}{\omega_r} X_{rr} & r_r + \frac{\rho}{\omega_r} X_{rr} \end{bmatrix} \begin{bmatrix} i_{qs}^e \\ i_{ds}^e \\ i_{qr}^e \\ i_{dr}^e \end{bmatrix} \quad (2)$$

Where, r_r – rotor resistance, r_s – stator resistance, ρ is the operator d/dt, ω_s and ω_r are the base electrical and rotor angular velocity,

$$s = \frac{\omega_s - \omega_r}{\omega_s}, X_{ss} = X_{ls} + X_M \text{ and } X_{rr} = X_{lr} + X_M$$

$$T_e - T_L = J\frac{2}{P}\rho\omega_r \quad (3)$$

$$T_e = \left(\frac{3}{2}\right)\left(\frac{P}{2}\right)\frac{X_M}{\omega_s}\left(i_{qs}^e * i_{dr}^e - i_{ds}^e * i_{qr}^e\right) \quad (4)$$

Where, P - number of poles, J - Inertia of the rotor (kg-m^2) and T_L, T_e is the load torque and electrical torque of the induction motor.

The configuration model of the proposed single switch buck-boost converter is shown in Fig. (5a). It has one power switch (S), three capacitors (C_1, C_2 & C_3), two inductors (L_1 & L_2) and two diodes (D_1 & D_2).

Various modes of operation of the proposed converter is given in Fig. (5b). It has three modes of operation.

When the switch S is in ON position, it is Mode-1. OFF position of the switch has two modes named Mode-2 and Mode-3. With the capacitors named C_2 and C_3 in the topology, dc load is connected. This makes the dc output voltage as two times D/ (1-D) that of the input voltage.

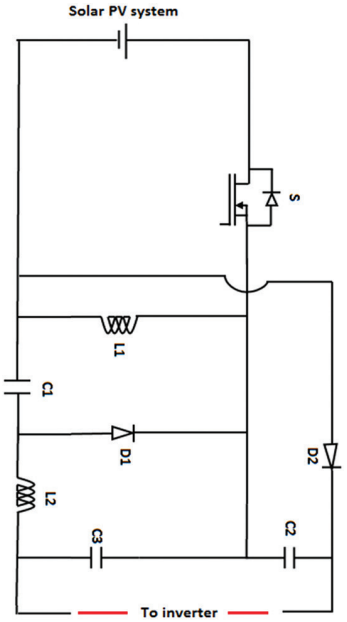


Fig. 5a. Topology of the single switch buck-boost converter

a. Mode 1: Switch S - ON condition

Fig. (6a). shows the Mode-1 operation (S is ON). Switch S is conducting which turns off the diodes D_1 and D_2. So D_1 and D_2 are in reverse biased. Then inductor L_1 is magnetized through the solar PV source. Capacitors C_1 and C_3 are used to magnetize/charge the Inductor L_2. Capacitors C_2 and C_3 feeds the load or energy deposited in the capacitors C_2 and C_3 are demagnetized through the load. In Fig. (6a), Applying KVL to the loops following equations are obtained.

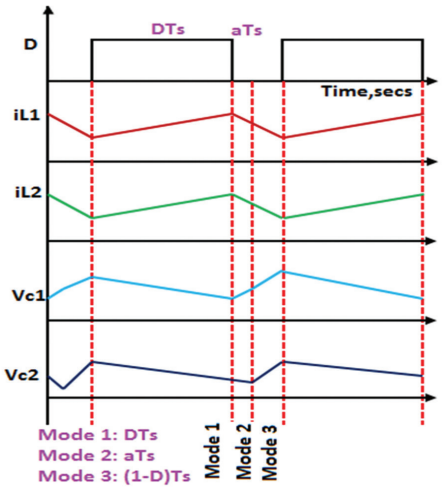


Fig. 5b. Modes of operation

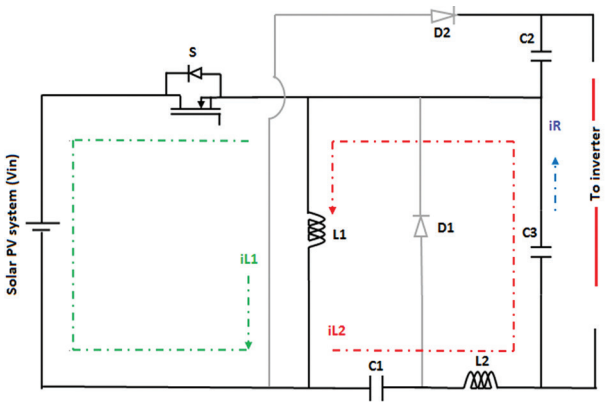


Fig. 6a. Mode-1 S- ON

$$\frac{\mathrm{di_{L_1}}}{\mathrm{dt}} = \frac{1}{\mathrm{L_1}} V_{\mathrm{in}} \tag{5}$$

$$\frac{\mathrm{diL_2}}{\mathrm{dt}} = \frac{1}{\mathrm{L_2}} (V_{in} + V_{C1} - V_{C3}) \tag{6}$$

$$V_{\mathrm{Load}} = V_{\mathrm{C2}} + V_{\mathrm{C3}} \tag{7}$$

By applying KCL to the circuit, the current flows through the capacitors are obtained

$$\mathrm{I_{C1}} = -\mathrm{I_{L2}} \tag{8}$$

$$\mathrm{I_{C2}} = -\mathrm{I_R} \tag{9}$$

$$\mathrm{I_R} = \mathrm{I_{L1}} - \frac{\mathrm{C_1 + C_2}}{\mathrm{C_1}} \mathrm{I_{C1}} \tag{10}$$

$$\begin{bmatrix} \frac{dV_{C1}}{dt} \\ \frac{dV_{C2}}{dt} \\ \frac{dV_{C3}}{dt} \\ \frac{di_{L1}}{dt} \\ \frac{di_{L2}}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & -\frac{1}{C_1} \\ 0 & 0 & 0 & -\frac{1}{C_2} & 0 \\ \frac{C_1 + C_2}{C_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 & 0 \end{bmatrix} * \begin{bmatrix} \mathrm{V_{C1}} \\ \mathrm{V_{C2}} \\ \mathrm{V_{C3}} \\ \mathrm{i_{L1}} \\ \mathrm{i_{L2}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{\mathrm{L_1}} \\ \frac{1}{\mathrm{L_3}} \end{bmatrix} * \tag{11}$$

$$[0 \quad 0 \quad 0 \quad V_1 \quad V_2]$$

$$\text{Where } A_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & -\frac{1}{C_1} \\ 0 & 0 & 0 & -\frac{1}{C_2} & 0 \\ \frac{C_1 + C_2}{C_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 & 0 \end{bmatrix} \text{ and } B_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{L_1} \\ \frac{1}{L_3} \end{bmatrix}$$

b. Mode 2: Switch S - OFF condition.

Fig. 6b shows the Mode-2 operation. Switch S is in non-conducting condition which turns on the diode $D1$ and diode D_2 is still in off position. It makes the diode D_1 in biased forward mode and D_2 in reverse biased condition. In this mode, capacitors C_1 is demagnetizing the inductors L_1 and L_2. *I*t also charges the capacitor C_3.

Energy that are stored already in the capacitor C_2 is discharged through the load. This mode continues till the charge in capacitor C_1 is equal to capacitor C_2. This avoids the current spikes through capacitors and diodes. In Fig. 6b, Applying KVL to the loops following equations are obtained.

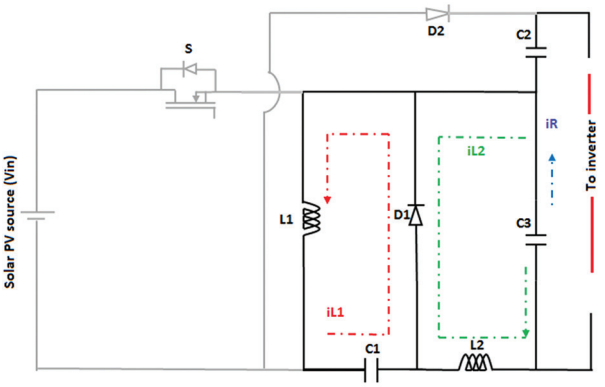


Fig. 6b. Mode- 2 S-OFF

$$\frac{\mathrm{diL_1}}{\mathrm{dt}} = -\frac{1}{\mathrm{L_1}} \mathrm{V_{C1}} \tag{12}$$

$$\frac{\mathrm{diL_2}}{\mathrm{dt}} = -\frac{1}{\mathrm{L_2}} \mathrm{V_{C3}} \tag{13}$$

$$\mathrm{V_{Load}} = \mathrm{V_{C2}} + \mathrm{V_{C3}} \tag{14}$$

By applying KCL to the circuit, the current flows through the capacitors are obtained

$$\mathrm{I_{C1}} = \mathrm{I_{L1}} \tag{15}$$

$$\mathrm{I_{C2}} = -\mathrm{I_R} \tag{16}$$

$$\mathrm{I_R} = \mathrm{I_{L1}} - \frac{\mathrm{C_1 + C_2}}{\mathrm{C_1}} \mathrm{I_{C2}} \tag{17}$$

$$\begin{bmatrix} \frac{dV_{C1}}{dt} \\ \frac{dV_{C2}}{dt} \\ \frac{dV_{C3}}{dt} \\ \frac{di_{L1}}{dt} \\ \frac{di_{L2}}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{C_1} & 0 \\ 0 & 0 & 0 & -\frac{1}{C_2} & 0 \\ \frac{C_1 + C_2}{C_1} & 0 & \frac{-1}{L_2} & 0 & 0 \\ \frac{1}{L_1} & 0 & 0 & 0 & 0 \\ \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 & 0 \end{bmatrix} * \begin{bmatrix} \mathrm{V_{C1}} \\ \mathrm{V_{C2}} \\ \mathrm{V_{C3}} \\ \mathrm{i_{L1}} \\ \mathrm{i_{L2}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{\mathrm{L_1}} \\ \frac{1}{\mathrm{L_3}} \end{bmatrix} * \tag{18}$$

$$[0 \quad 0 \quad 0 \quad V_1 \quad V_2]$$

$$\text{Where } A_2 = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{C_1} 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{C_2} & 0 \\ \frac{C_1 + C_2}{C_1} & 0 & -\frac{1}{L_2} & 0 & 0 \\ \frac{1}{L_1} & 0 & 0 & 0 & 0 \\ \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 & 0 \end{bmatrix} \text{ and } B_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{L_1} \\ \frac{1}{L_3} \end{bmatrix}$$

c. Mode 3: Switch S is still in OFF condition.

Fig. (6c). shows the Mode-position. Switch S is in non-conducting mode and the charge remaining in capacitor C_1 is equal to C_2. It turns on the diodes D_1 and D_2. So D_1 & D_2 are in forward biased. In this mode, inductor L_1 is demagnetized through C_1 and C_2. Inductor L_2 charges the capacitor C_3. Energy that are stored already in the capacitor C_2 is discharged via dc load.

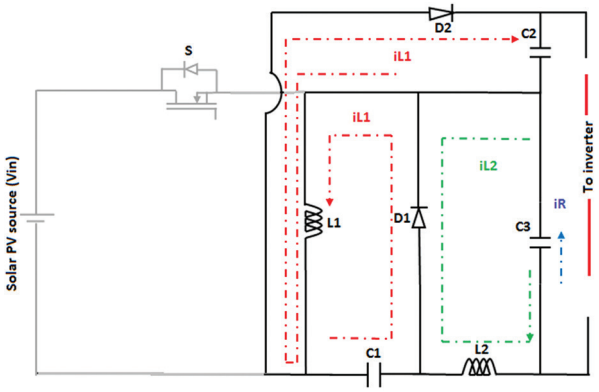


Fig. 6c. Mode- 2 S-OFF

$$\frac{\mathrm{di}_{L1}}{dt} = -\frac{1}{L_1} V_{\mathrm{C1}} \tag{19}$$

$$\frac{\mathrm{di}_{L2}}{dt} = -\frac{1}{L_2} V_{\mathrm{C3}} \tag{20}$$

$$\begin{bmatrix} \frac{dV_{C1}}{dt} \\ \frac{dV_{C2}}{dt} \\ \frac{dV_{C3}}{dt} \\ \frac{di_{L1}}{dt} \\ \frac{di_{L2}}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{C_1} & 0 \\ 0 & 0 & 0 & -\frac{1}{C_2} & 0 \\ \frac{C_1+C_2}{C_1} & 0 & \frac{-1}{L_2} & 0 & 0 \\ \frac{1}{L_1} & 0 & 0 & 0 & 0 \\ \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 & 0 \end{bmatrix} * \begin{bmatrix} \mathrm{V_{C1}} \\ \mathrm{V_{C2}} \\ \mathrm{V_{C3}} \\ \mathrm{i_{L1}} \\ \mathrm{i_{L2}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{L_1} \\ \frac{1}{L_3} \end{bmatrix} * \begin{bmatrix} 0 & 0 & 0 & V_1 & V_2 \end{bmatrix} \tag{21}$$

$$\text{Where } A_3 = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{C_1} & 0 \\ 0 & 0 & 0 & -\frac{1}{C_2} & 0 \\ \frac{C_1+C_2}{C_1} & 0 & \frac{-1}{L_2} & 0 & 0 \\ \frac{1}{L_1} & 0 & 0 & 0 & 0 \\ \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 & 0 \end{bmatrix} \text{ and } B_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{L_1} \\ \frac{1}{L_3} \end{bmatrix}$$

The proposed single switch buck-boost converter is assumed to be CCM mode. Let us assume the switch S1 is in the ON position individually for the time period d1TS & d2TS. Switch S1 is in OFF position for the time period (1- d1) TS & (1- d2) TS. In this case d1 and d2 are taken as the duty cycle and TS is considered as the switching duration. Here the capacitor voltages (VC1 and VC2) are equal and the average inductor current (IL2) is same as the output current. The inductor time constant is stated as,

$$t_L = \frac{1}{2}(1-D)^2 \tag{22}$$

By applying volt-sec balance equation on inductor L_1

$$\mathrm{DV_{in}} - \alpha V_{\mathrm{C1}} - \left((1-\mathrm{D}-\alpha)(\mathrm{V_{C1}})\right) = 0 \tag{23}$$

By solving this equation,

$$\mathrm{V_{C1}} = \mathrm{V_{C2}} = \mathrm{V_{in}} \frac{\mathrm{D}}{1-D} \tag{24}$$

Volt-sec balance equation is applied on inductor L_2,

$$D(\mathrm{V_{in}} + \mathrm{V_{C1}} - \mathrm{V_{C3}}) - \alpha \mathrm{V_{C3}} - \left((1-\mathrm{D}-\alpha)(\mathrm{v_{C3}})\right) = 0 \tag{25}$$

By solving this equation,

$$\mathrm{V_{C3}} = \mathrm{V_{in}} \frac{\mathrm{D}}{1-D} \tag{26}$$

By applying ampere-sec balance equation on capacitor C1,

$$-(\mathrm{D}+\alpha)(\mathrm{I_R}) + \left((1-\mathrm{D}-\alpha)\left(\frac{C_1}{C_1+C_2}(\mathrm{I_{L1}}-\mathrm{I_R})\right)\right) = 0 \tag{27}$$

Ampere-sec balance equation is applied on capacitor C2,

$$-(\mathrm{D}+\alpha)(\mathrm{I_R}) + \left((1-\mathrm{D}-\alpha)\left(\frac{\mathrm{C_1}}{\mathrm{C_1}+\mathrm{C_2}}(\mathrm{I_{L1}}-\mathrm{I_R})\right)\right) = 0 \tag{28}$$

By solving these equations,

$$\mathrm{I_{L2}} = \mathrm{I_R} \tag{29}$$

The proposed system is operated at CCM operation. So the average voltage of the proposed system is calculated as follows,

$$\frac{\mathrm{V_{C2}}}{\mathrm{V_{in}}} = \frac{2\mathrm{D}}{1-\mathrm{D}} \tag{30}$$

During transient period, the inductor voltgae and current, capacitor voltage and current and dc voltage and current may vary due to time. For any dc-dc converter having commutating capacitance, inductance and with no phase delay is,

$$\mathrm{V_{dc}} = \mathrm{V_d} + X_{Ldc}\rho \mathrm{I_d} \tag{31}$$

$$\rho \mathrm{V_{dc}} = \omega_b X_{Cdc}(\mathrm{I_d} - \mathrm{I_{dc}}) \tag{32}$$

$$\mathrm{V_d} = \left(\frac{3\sqrt{3}}{\pi}\mathrm{V_q}^G\right) - \left(\frac{3}{\pi}\frac{\mathrm{X_{Lc}}}{\omega_b}\omega_{\varepsilon 0}\mathrm{I_d}\right) - \frac{2}{\omega_b}X_{Lc}\rho\mathrm{I_d} \tag{33}$$

Substituting the equation (33) in (31),

$$\mathrm{V_{dc}} = \left(\frac{3\sqrt{3}}{\pi}\mathrm{V_q}^G\right) - \left(\frac{3}{\pi}\frac{\mathrm{X_{Lc}}}{\omega_b}\omega_g\mathrm{I_d}\right) - \frac{1}{\omega_b}(X_{Lc}+2X_{Lc})\rho\mathrm{I_d} \tag{34}$$

Where, V_q^G - q component of the dc voltage and ω_g - system frequency, d- duty cycle of modulation.

Neglecting the harmonics in the VSI source with sin-triangle pulse width modulation, voltages in the rotating reference frame become,

$$V_{qs}^e = \frac{d}{2} V_{dc} \tag{35}$$

$$V_{ds}^e = 0 \tag{36}$$

The power balance equation is,

$$V_{dc} * I_{dc} = \left(\frac{3}{2}\right) V_{qs}^e I_{qs}^e \tag{37}$$

$$I_{dc} = \left(\frac{3d}{4}\right) I_{qs}^e \tag{38}$$

$$\rho V_{dc} = \omega_b X_{Cdc} \left(I_d - \left(\frac{3d}{4}\right) I_{qs}^e\right) \tag{39}$$

The above equations are the dynamic equations of DC system.

d. Design equations of the proposed converter

The parameter value of the components inductor and capacitor is calculated as follows,

$$\Delta I_{L1} = \frac{V_{in} D}{L_1 f} \tag{40}$$

$$\Delta I_{L2} = \frac{V_{C3} D}{L_2 f} \tag{41}$$

$$\Delta V_{C1} = \frac{I_{in}(1-D)}{C_1 f} \tag{42}$$

$$\Delta V_{C2} = \frac{I_R(1-D)}{C_2 f} \tag{43}$$

$$\Delta V_{C3} = \frac{I_R(1-D)}{C_3 f} \tag{44}$$

From the above equations (40) to (44), the simulation parameters are calculated and the circuit is simulated. Let the duty cycle D=0.74, input voltage V_{in}= 520V, input current Iin=35A and frequency f= 20kHz

Table.1. Simulation parameters

Components	Values
L1	400 µH
L2	84.16 µH
C1	0.01 µF
C2	0.01 µF
C3	0.01 µF
Input	520V, 35A
Switching frequency	20kHz

Backup power can be compensated load demand when needed.

2.5. REDUCTION OF THD IN 3PHASE INVERTER TOPOLOGY

Fig. 7. Presents the proper layout of three phase VSI. It has six controlled semiconductor switches. By controlling the switching pattern, the output ac voltage and line currents can be obtained as pure sinusoidal. Instead of using installing active/ passive filters. Some of the harmonics are eliminated using switching pattern techniques. PWM is a different switching technique which gives versatile results by varying modulation ratio and index and switching frequency. Here increasing the switching frequency can reduce the current harmonics which further increases the switching losses. THD is very much useful in finding the harmonics order presented in the voltage or current waveforms. It helps in observing the quality of non-sinusoidal wave. THD is the division between all harmonic components RMS value and the fundamental component RMS.

$$THD = \frac{\sqrt{\sum_{n=2}^{\infty} V_{n\_rms}^2}}{V_{fund\_rms}} \tag{45}$$

Where, V_{n_rms} is the nth harmonic RMS voltage and V_{fund_rms} is the RMS voltage of the fundamental frequency.

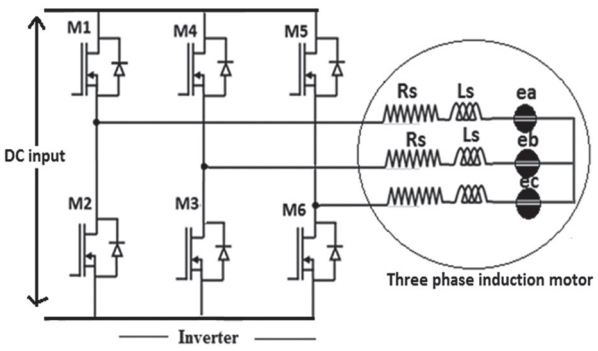


Fig. 7. Three phase inverter with 3Φ induction motor

In this work, the steady state analysis of the proposed system have been performed. Linearization of systems about an operating point with nonlinear differential equations are taken. The small signal model of linearization of mechanical equation of induction motor is given by,

$$J\Delta\omega_b\rho = \left(\frac{3}{2\omega_b}\right)\left(\frac{P}{2}\right)^2 \begin{pmatrix} X_M I_{qs0}^e \Delta I_{dr}^e + X_M \Delta I_{qs}^e I_{dr0}^e \\ -X_M I_{ds0}^e \Delta I_{qr}^e - X_M \Delta I_{ds}^e I_{qr0}^e \end{pmatrix} - \frac{P}{2}\Delta T_L \tag{46}$$

Here 0 denotes the steady state quantities at an operating point. The fundamental form of the linear differential equation is

$$\rho\Delta X = A\Delta X + B\Delta U \tag{47}$$

Where, $A=E^{-1}F$, $B=E^{-1}B'F$, ΔX and ΔU are state and control variables. A is the state matrix of the proposed system.

$$E = \frac{1}{\omega_b}\begin{bmatrix} X_{ss} & 0 & X_M & 0 & 0 & 0 & 0 \\ 0 & X_{ss} & 0 & X_M & 0 & 0 & 0 \\ X_M & 0 & X_{rr} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & X_{rr} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & J\omega_b & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \omega_b \end{bmatrix} \quad (48)$$

$$B' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{P}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{3\sqrt{3}}{\pi(X_{Ldc}+2X_{Lc})}\omega_b & 0 & -\frac{3X_{Lc}I_{d0}}{\pi(X_{Ldc}+2X_{Lc})} \end{bmatrix} \quad (49)$$

Linear combination of state and control variables form output vector of the system,

$$\Delta Y = C\Delta X + D\Delta U \quad (50)$$

In induction motor electromagnetic torque T_e and rotor velocity ω_r are considered as output.

$$\Delta T_\theta = C_1\Delta X + D_1\Delta U \quad (51)$$

Where,

$$C_1 = \frac{3P}{4\omega_b}X_M[I^e_{dr0} \quad -I^e_{qr0} \quad -I^e_{ds0} \quad I^e_{qs0} \quad 0 \quad 0 \quad 0]$$
$$D_1 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \quad (52)$$
$$\Delta\omega_r = C_2\Delta X + D_2\Delta U$$

Where,

$$C_2 = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0]$$
$$D_2 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

With the state and output equations, the transfer function of the system is written as,

$$\frac{\Delta Y}{\Delta U} = C(SI - A)^{-1}B + D \quad (53)$$

The steady state curents are given as,

$$\begin{bmatrix} i^e_{qs} \\ i^e_{ds} \\ i^e_{qr} \\ i^e_{dr} \end{bmatrix} = \begin{bmatrix} r_s & \frac{\omega_s}{\omega_r}X_{ss} & 0 & \frac{\omega_s}{\omega_r}X_M \\ -\frac{\omega_s}{\omega_r}X_{ss} & r_s & -\frac{\omega_s}{\omega_r}X_M & 0 \\ 0 & s\frac{\omega_s}{\omega_r}X_M & r_r & s\frac{\omega_s}{\omega_r}X_{rr} \\ -s\frac{\omega_s}{\omega_r}X_M & 0 & -s\frac{\omega_s}{\omega_r}X_{rr} & r_r \end{bmatrix} \begin{bmatrix} V^e_{qs} \\ V^e_{ds} \\ V^e_{qr} \\ V^e_{dr} \end{bmatrix} \quad (54)$$

Table 2 show the induction motor specifications.

Table. 2. Simulation parameters of induction motor

Parameters	Values
Power	18kW
Voltage	440V
Speed	1820rpm
Stator and rotor resistance	0.214Ω and 0.22 Ω
Xls and Xlr	9e-3H and 9e-3H
Mutual inductance	0.0114H
Inertia(J)	0.102Kg/m²

3. DISCUSSION ON SIMULATION RESULTS

To analyze the proposed single switch buck-boost topology based 3Φ induction motor, a detailed MATLAB/SIMULINK is carried out in R2013a version. The dc link voltage is set at 360V and the dc link current at 45A. Fig. 8a and 8b shows the current and voltage across the dc load. It is noted that the voltage across topology is measured as 363V and the current through the topology is measured as 45.5A.

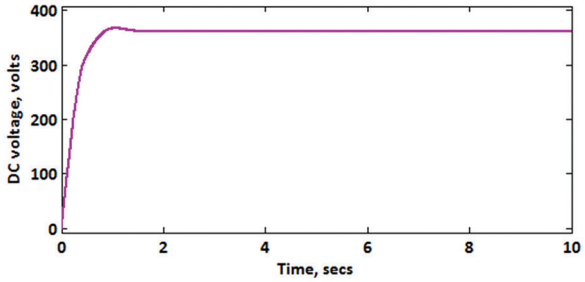


Fig. 8a. Voltage across DC load

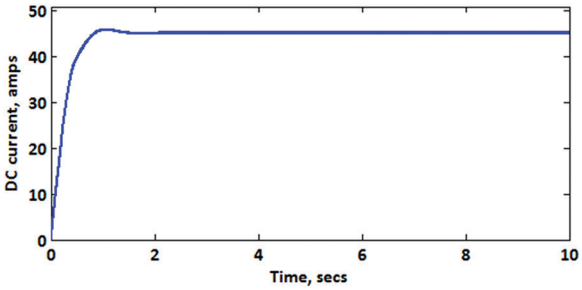


Fig. 8b. Current through DC load

Fig. 9a and 9b presents the ac output voltage and current of three phase inverter respectively. From the graph, it is noted that 440V of output voltage and 15A of output current is observed.

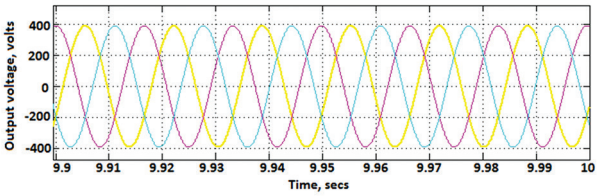


Fig. 9a. AC output voltage

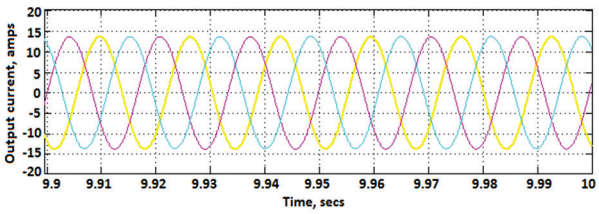


Fig. 9b. AC output current

Fig. 10 shows the results of gate pulse generation process in three phase inverter circuit. Here, two sine waves

are joined to generate the reference signal. Switching loss of the inverter can be eliminated by using three times frequency of the sine wave compared to other. It cancels the harmonics by utilizing the dc supply more. It also shows the switching pulses of inverter switches.

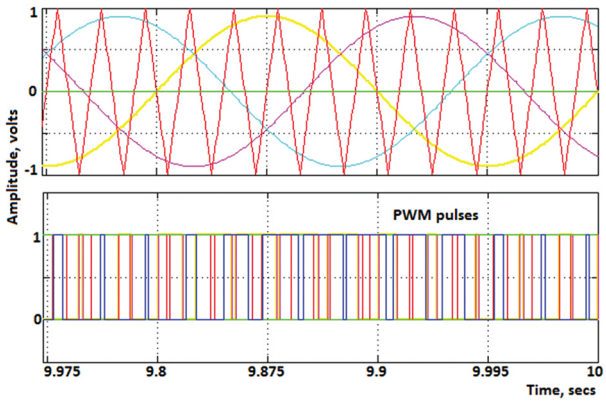


Fig.10. Gate pulse generation of inverter switches

Fig.11 and 12 shows the motor speed and Te (electromagnetic torque).

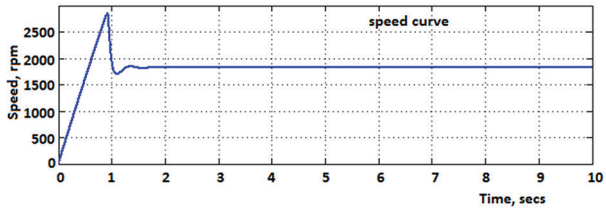


Fig. 11. Motor speed

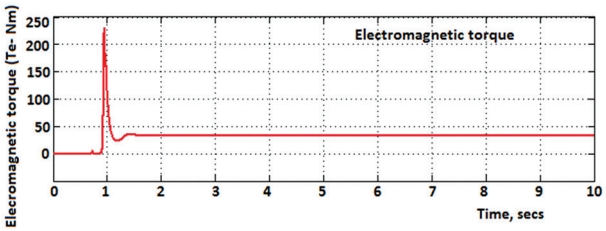


Fig. 12. Electromagnetic torque (Te)

The obtained transfer function mentioned in equation (53) is done using MATLAB/SIMULINK, R2013a version with the same simulation values for investigating its stability. By using LQG controller, it performs many attempts for the designed closed loop model. The time domain and frequency domain responses of the system are observed. The step and impulse response is shown in Fig. 13 a and 13 b. From the step response figure, zero overshoot, no peak and no transients is seen in the closed loop. Impulse response figure shows 1.01 peak response with zero overshoot. The settling time is 0.3 sec. Fig. 14 a and 14 b shows the frequency response of the system.

The main condition for the stable system in root locus is that the placement of the poles and zeros should be on the left hand side. It can be able to provide response characteristics of the system. From the obtained result of the proposed system with LQG synthesis controller, it is shown that the system is stable with the gain and phase margins of -0.0202 dB, 0.0374 rad/sec and 0.737 deg, 47.7 rad/sec respectively.

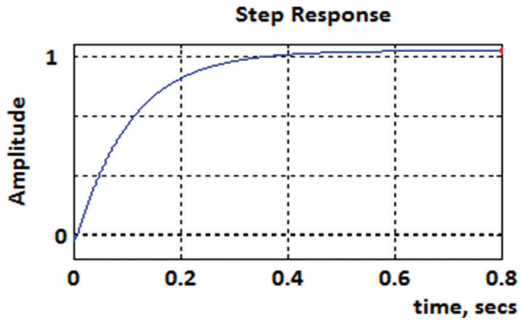


Fig. 13a. Step response

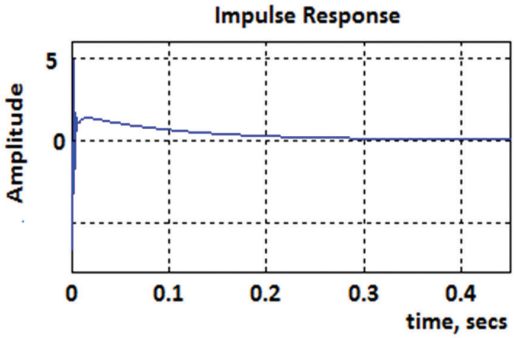


Fig. 13 b. Impulse response

Then performance analysis is carried out depending on the value of THD %. Deviation in the THD is read by changing the frequency and modulation index of the carrier signal. The obtained THD values of output current and voltage is shown in Fig.15 a and 15 b in which we got the THD values of 5.43% and 7.14% respectively. Efficiency of the proposed converter is the total output voltage divided by the input voltage. It is obtained as [(363*45.5)/ (520*35)]*100= 90.5%.

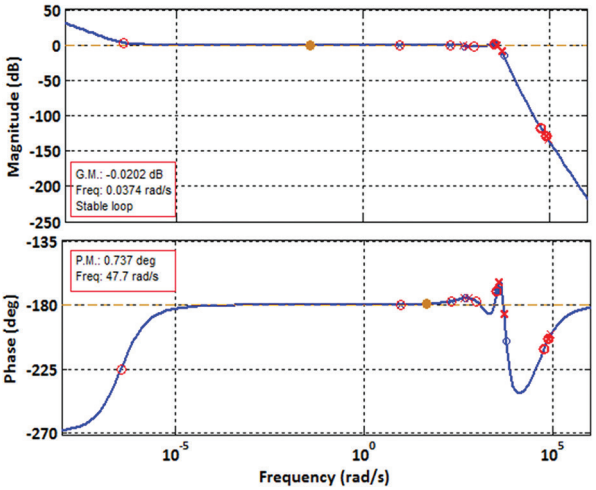


Fig. 14 a. Frequency response- Bode plot

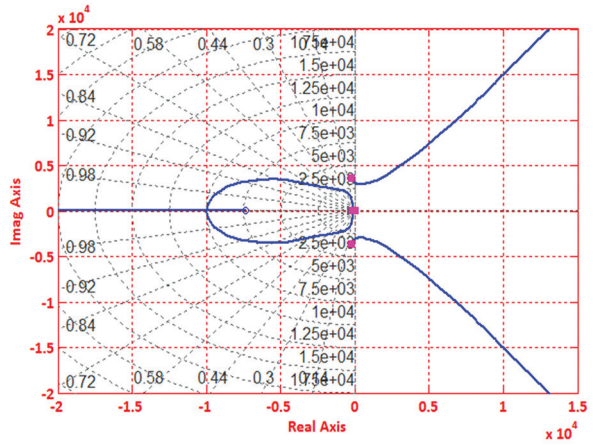


Fig. 14 b. Frequency response- Pole-Zero plot

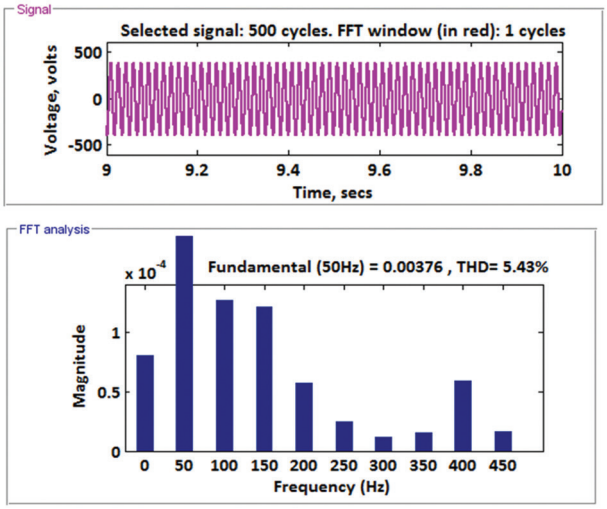


Fig. 15 a. THD of AC Output voltage

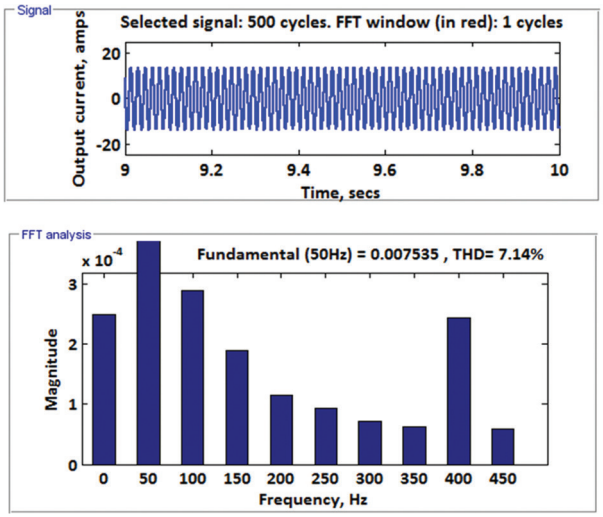


Fig. 15 b. THD of AC output current

4. CONCLUSION

Solar PV powered single switch buck-boost converter is simulated with reduced cost, minimal voltage and current stress across the capacitors and diodes, less switching power losses and increased efficiency. In this work solar PV source with modified P & O algorithm based MPPT, single switch buck-boost dc-dc converter, battery backup to store excess energy, three phase inverter with sinusoidal PWM to find optimal switching angles for harmonic control and 3Φ induction motor load is used. Reduction of THD is applied to the line to line voltage and current of the inverter and observed the voltage and current THD of 5.43% and 7.14% respectively. Performance analysis of the topology proposed is done using MATLAB/SIMULINK 2013a platform. The proposed single switch buck-boost converter has produced an output voltage and current of 363V, 45.5A DC from 520V, 35A PV array. The designed converter is then connected with a three phase VSI with 440V, 15A AC. The obtained efficiency is 90.5%. Here small signal steady state analysis is performed for the proposed system. In the extension of this work several new strategies in the MPPT algorithms like Ant colony, firefly etc., can be preferred. Recent control algorithms for VSI can be used. Fault analysis of 3Φ induction motor can be performed.

5. REFERENCES:

[1] M. R. Banaei, H. Ardi, A. Farakhor, "Analysis and implementation of a new single-switch buck–boost DC/DC converter", IET Power Electronics, Vol. 7, No. 7, 2014, pp. 1906-1914.

[2] S.-Y. Tseng, J.-H. Fan, "Buck-Boost/Fly back Hybrid Converter for Solar Power System Applications", Electronics, Vol. 10, No.4, 2021, p. 414.

[3] J. M. Alonso, J. Vina, D. G. Vaquero, G. Martinez, R. Osorio, "Analysis and design of the integrated double buck–boost converter as a high-power-factor driver for power-LED lamps", IEEE Transactions on Industrial Electronics, Vol. 59, No. 4, 2011, pp. 1689-1697.

[4] Md T. Islam, S. I. Ayon, "Performance Analysis of Three-Phase Inverter for Minimizing Total Harmonic Distortion Using Space Vector Pulse Width Modulation Technique", Proceedings of the 23rd International Conference on Computer and Information Technology, Dhaka, Bangladesh, 19-21 December 2020, pp. 1-4.

[5] F. Yalcin, U. Arifoglu, I. Yazici, K. Erin, "Robust single-phase inverter based on the buck–boost converter through an efficient hybrid control", IET Power Electronics, Vol. 13, No. 1, 2020, pp. 50-59.

[6] M. Schubert, R. W. De Doncker. "Instantaneous phase voltage sensing in PWM voltage-source

inverters", IEEE Transactions on Power Electronics, Vol. 33, No. 8, 2017, pp. 6926-6935.

[7] T. T. Atly, "Notice of Removal: Active buck-boost inverter for inverter air conditioner applications", Proceedings of the International Conference on Electrical, Electronics, Signals, Communication and Optimization, Visakhapatnam, India, 24-25 January 2015, pp. 1-5.

[8] M. E. Ibrahim, A. S. Mansour, A. M. Abd-Elhady, "A novel single-stage single-phase buck–boost inverter", Electrical Engineering, Vol. 99, No. 1, 2017, pp. 345-356.

[9] M. M. S. Khan, Md S. Arifin, Md R. T. Hossain, Md A. Kabir, A. H. Abedin, M. A. Choudhury, "Input switched single phase buck and buck-boost AC-DC converter with improved power quality", Proceedings of the 7th International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 20-22 December 2012, pp. 189-192.

[10] R. Abdikarimuly, A. Ruderman, B. Reznikov, "Calculation of current total harmonic distortion for a three-phase two-level inverter with LCL-filter", Proceedings of the 19th International Conference on Electrical Drives and Power Electronics, Dubrovnik, Croatia, 4-6 October 2017, pp. 100-105.

[11] P. S. Pawar, J. S. Bagi, G. R. Shinde, "A Review on Harmonic Reduction Techniques in Three-Phase Power Generation in PV Solar Plants", International Journal of Electrical and Electronics Engineering, Vol. 9, No. 1, 2019, pp. 1-6.

[12] P. J. B. Rao, B. Srikanth, "Harmonic reduction in grid connected solar PV system by using SVPWM technique", International Journal of Applied Engineering Research, Vol. 11, No. 6, 2016, pp. 3853-3858.

[13] R. Dogga, M. K. Pathak, "Recent trends in solar PV inverter topologies", Solar Energy, Vol. 183, 2019, pp. 57-73.

[14] M. Aly, H. A. Ramadan, "Design and implementation of adaptive SVPWM algorithm for multilevel inverters in renewable energy applications", Solar Energy, Vol. 183, 2019, pp. 745-754.

[15] R. Barzegarkhoo, E. Zamiri, N. Vosoughi, H. M. Kojabadi, L. Chang, "Cascaded multilevel inverter using series connection of novel capacitor-based units with minimum switch count", IET Power Electronics, Vol. 9, No. 10, 2016, pp. 2060-2075.

[16] P. C. Krause, T. A. Lipo, "Analysis and simplified representations of a rectifier-inverter induction motor drive", IEEE Tarnsactions on Power Apparatus and Systems, Vol. PAS-88, No. 5, 1969, pp. 588-596.

[17] R. H. Nelson, T. A. Lipo, P. Krause, "Stability analysis of a symmetrical induction machine", IEEE Trans. Power Apparatus and Systems, Vol. PAS-88, No. 11, 1969, pp. 1710-1717.

[18] F. Fallside, A. T. Wortley, "Steady-state oscillations and stabilization of variable frequency inverter fed induction motor drives", IEEE Proceedings, Vol. 116, No. 6, 1969, pp. 991-999.

[19] T. A. Lipo, P. Krause, "Stability analysis of a rectifier–inverter induction motor drives", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-88, No. 1, 1969, pp. 55-66.

[20] V. F. Pires, A. Cordeiro, D. Foito, J. F. Silva, "Three-phase multilevel inverter for grid-connected distributed photovoltaic systems based in three three-phase two-level inverters", Solar Energy, Vol. 174, 2018, pp. 1026-1034.

[21] M. Jayabalan, B. Jeevarathinam, T. Sandirasegarane, "Reduced switch count pulse width modulated multilevel inverter", IET Power Electronics, Vol. 10, No. 1, 2017, pp. 10-17.

[22] V. F. Pires, D. Foito, Ar. Cordeiro, J. F. Silva, "A single-switch DC/DC buck-boost converter with extended output voltage", Proceedings of the 7th International Conference on Renewable Energy Research and Applications, Paris, France, 14-17 October 2018, pp. 791-796.

[23] M. R. Banaei, H. Ardi, A. Farakhor, "Analysis and implementation of a new single-switch buck–boost DC/DC converter", IET Power Electronics, Vol. 7, No. 7, 2014, pp. 1906-1914.

[24] M. Delshad, H. Farzanehfard, "A new single switch Buck-Boost type dc-dc converter", Proceedings of the IEEE Symposium on Industrial Electronics & Applications, Vol. 2, Kuala Lumpur, Malaysia, 4-6 October 2009, pp. 823-826.

[25] V. F. Pires, D. Foito, A. Cordeiro, J. F. Silva, "A single-switch DC/DC buck-boost converter with extend-

ed output voltage", Proceedings of the 7th International Conference on Renewable Energy Research and Applications, Paris, France, 14-17 October 2018, pp. 791-796.

[26] M. R. Banaei, H. Ardi, F. Bonab, "A novel structure for single-switch non- isolated transformer less buck–boost DC–DC converter", IEEE Transactions on Industrial Electronics, Vol. 64, No. 1, 2016, pp. 198-205.

[27] D. M. Van de Sype, K. De Gusseme, W. R. Ryckaert, A. P. Van de Bossche, J. A. Melkebeek, "A single switch buck-boost converter with a high conversion ratio", Proceedings of the European Conference on Power Electronics and Applications, Dresden, Germany, 11-14 September 2005, p. 10.

[28] E. Jamshidpour, S. Jovanovic, P. Poure, "Equivalent Two Switches and Single Switch Buck/Buck-Boost Circuits for Solar Energy Harvesting Systems", Energies, Vol. 13, No. 3, 2020, p. 583.

[29] L. An, D. D.-C. Lu, "Design of a single-switch DC/DC converter for a PV-battery-powered pump system with PFM+ PWM control", IEEE Transactions on Industrial Electronics, Vol. 62, No. 2, 2014, pp. 910-921.

[30] S. Chakraborty, S. I. Annie, M. A. Razzak, "Design of single-stage buck and boost converters for photovoltaic inverter applications", Proceedings of the International Conference on Informatics, Electronics & Vision, Dhaka, Bangladesh, 23-24 May 2014, pp. 1-6.

[31] M. M. Akhil, V. Aishwarya, K. G. Sheela, "An improved sepic-based single switch buck-boost pfc converter fed brushless dc motor drive", Materials Today: Proceedings, Vol. 24, 2020, pp. 1855-1864.

Trends and Challenges in Electric Vehicle Motor Drivelines - A Review

Review Paper

Ashwin Kavasseri Venkitaraman
University of Cincinnati
MS – Electrical Engineering, Electrical Engineering Department
2600 Clifton Ave, Cincinnati, Ohio, USA
ashwin.venkit@gmail.com

Venkata Satya Rahul Kosuru
Lawrence Technological University
MS – Electrical & Computer Engineering, Electrical Engineering Department
21000 W 10 Mile Rd, Southfield, MI, USA
venkosuru@gmail.com

Abstract *– Considering the need to optimize electric vehicle performance and the impact of efficient driveline configurations in achieving this, a brief study has been conducted. The drivelines of electric vehicles (EV) are critically examined in this survey. Also, promising motor topologies for usage in electric vehicles are presented. Additionally, the benefits and drawbacks of each kind of electric motor are examined from a system viewpoint. The majority of commercially available EV are powered by a permanent magnet motor or single induction type motors and a standard mechanical differential driveline. Considering these, a holistic review has been performed by including driveline configurations and different battery types. The authors suggest that motors be evaluated and contrasted using a standardized driving cycle.*

Keywords*: Electric vehicles, motor drivelines, electric vehicle propulsion, traction motor, power systems*

1. INTRODUCTION

Electric vehicles are gaining large-scale acceptance, because of which there has been growing interest in optimizing driveline topologies for hybrid and fully electric vehicles. Studies on common motor drive for EV and HEV have been presented in [1, 2]. The article examines the development of the EV with a focus on commercially available and upcoming electric motors. The study covers the benefits and drawbacks of current technological advances and describes the automatic components of the EV driveline.

The automotive sector has been a key force in research since the Ford T began to be produced in large quantities. Currently seen as the car of the future, electric vehicles (EVs) are rapidly acquiring industrial traction [3]. The history of the EV may be divided into three major phases. IC motors gained initial momentum. Steam cars were risky, unclean, and costly to repair. There were several technological benefits to electric vehicles. Since only major cities had fully paved roads and those lengthy trips were unusual, the limited range of EVs was less of a restriction [4].

The improvement of internal combustion (IC) and the decline in pricing due to mass manufacturing made IC automobiles the favored and only technology for years. Early in the development of the HEV, it was recognized that better efficiencies might be attained if IC motors were used in permutation with electric traction motors. The development of power electronics led to the second surge of EVs. It is also vital to note that the automotive industry was the first to study motor control for electric vehicles. The oil crisis helped to preserve interest and funding for EV development. Modern electric cars are built on the foundation of prototypes created during this time [5, 6].

The poor energy density and increase in cost of battery made EV less viable in comparison with IC automobiles [7]. However, in the current transportation landscape, HEVs and EVs are gaining traction owing to rising pollution. Additionally, there is legislative support for eco-friendly transportation, as seen by the subsidies and tax breaks for HEVs and EVs [8]. The appeal of EVs is increased by social and economic aspects. This achievement reveals resurgence in interest for effective electric drives. Specific criteria for vehicle propulsion set stationary and on-board motors apart [9]. Every kilogram carried aboard causes the system to lose energy due to friction and increases the structural stresses. High efficiency equates to lower energy requirements,

which reduces battery weight. The most effective alternative would thus seem to be Permanent Magnet (PM) motors. If PM motors are made of rare earth elements, it is hypothesized that a switch to EVs and HEVs for all developed nations might result in a rise in pricing and a lack of raw materials [10].

This paper introduces different models of the motors in section 2. Survey on models of motors, followed by a study on drivelines and its configuration in section 3. Driveline Configuration, 4. Battery and a brief overview in section 5. Overview of Market and 6. Electric Vehicle Machine Trends. Finally summarizing and concluding in section 7. Conclusion.

2. SURVEY ON MODELS OF MOTORS

Modern vehicles have over 100 distinct types of electric motors [11]. While only traction motors are addressed here, it is necessary to note that the subject matter is quite broad. With direct current (DC) motors, Synchronous Permanent Magnet (PM), Induction machine (IM), and Synchronous Brushed (SB) motors, currently available; the market for electric vehicles is fragmented due to the wide range of motor architecture and requirements. The Reluctance motor (RM) is a suggested fifth architecture with promising attributes but has not seen widespread commercial use in EVs. Both the nominal speed and power of variable-speed motors are artificial constructs [11]. A vehicle's catalog power is equal to its maximum power output from the drive system, or the upper limit set by the control system as a compromise among production and battery life. The motor strikes a balance between lightweight and powerful. The peak power capacity of the motor exceeds the rating of the system. Electric vehicles' power ranges from a kilowatt (kW) for mini quads to well over three hundred kW for higher production. As the electric vehicle market expands, the number of available niches increases, but the industry is still a long way from standardization. Early prototypes' power was governed by technical specifications, but it is today driven by market demands [12]. Fig. 1 depicts the time-dependent development of the power placed in the traction motors.

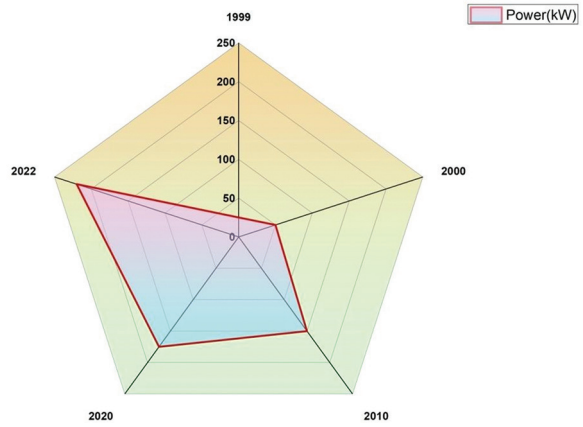


Fig. 1. Representation of power rating in EV

The operational point that every drive cycle imparts to an electric motor determines its efficiency, as is the case with IC motors. Variable speed motors do not have an industry-standard definition for efficiency rating. Efficiency maps for power and torque are used to describe them. Based on the motor type, the efficiency goes down at working points outside of the optimal region [13]. The design determines the motor's execution throughout a broad variety of speeds and powers, even though every kind of motor has a unique torque-speed relationship. For motors of similar peak efficiency, it is found that PM motors are more efficient in overload transients at constant speed, whereas RM motors perform better under high-speed overloads. The control of RMs enables for high-speed operation, but at low speeds, efficiency quickly degrades. Despite being less efficient overall than PM motors, SB motors are nonetheless highly efficient across a large operating range and can operate at high speeds because of its control. The different motor types and options for EV applications are shown in Fig. 2.

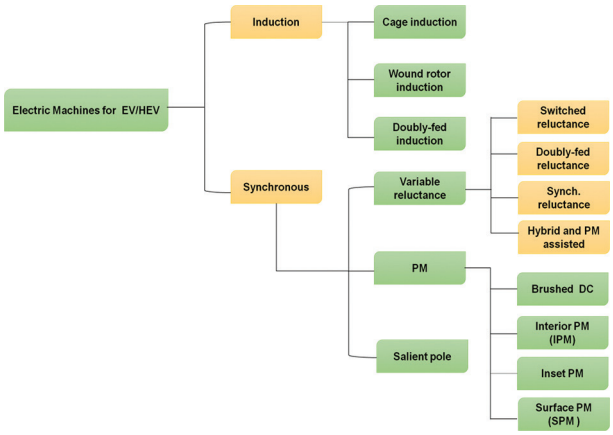


Fig. 2. Available motor types for EVs and HEVs

There are many different machine options in Fig. 2. Double-fed, wound rotor induction motors are not presently used in any automotive application within the IM category [14]. Doubly fed induction machines are often employed in wind energy production. The barrier to entry into automotive application is due to comparatively lower efficiency and power density. High peak torque, strong dynamic responsiveness, and minimal maintenance requirements across the board are some other significant benefits of these machines. It has been determined that wound-field synchronous machines might provide the increase in specific power requirement for electric propulsion. Despite being explored for several prospective uses, machine topologies such as doubly fed reluctance are not being used in any vehicles [15].

2.1. DC MOTOR

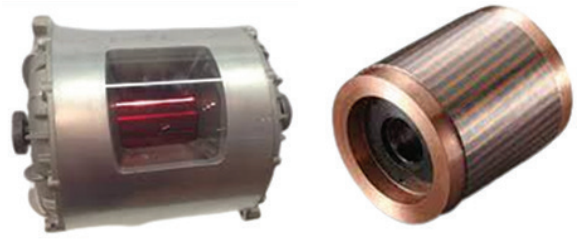
DC motors have a wound rotor and a stator. The rotor is equipped with a brush commutation mechanism and the stator possesses a stable field. Although

tiny machines may have a permanent magnet excitation, the field in the stator is often produced by coils. Depending on the desired characteristics, the field winding may be linked to the rotor coils in series or shunt. The copper segments that make up the commutator create greater friction than the slip rings, which results in the production of dust. The technology is well established, the motors are reliable, affordable, and offer an easy-to-use control system as its key benefits [16].

Prior to the development of sophisticated power electronics, DC motors were the chosen choice for variable speed operating applications. Low power density (PD) in comparison to competing technologies and expensive maintenance of the coal brushes are the primary drawbacks. The coal brushes practically need no maintenance because of the low use rate of private automobiles. Lower and intermediate power range commutation vehicles are still a significant market for DC motors [17].

2.2. INDUCTION MACHINES

The most popular devices in industry are squirrel cage induction machines. They are a solid choice including traction, owing to the simple yet robust design. The features of high peak torque, strong dynamic responsiveness, and minimal maintenance requirements across the board are some of these machines' other major benefits. The technology for induction machines is old [18]. Fig. 3 shows an example of the parts of an induction machine.



Tesla Motor Cut-Out Die-Cast Cooper Rotor

Fig. 3. Components in induction machine [16]

This technology is seen on initial Tesla car models and General Motors (GM) EV-1. Die-casting techniques that made copper rotors more affordable and gave these machines a big boost. Copper die-cast rotors have an increase in electrical conductivity than aluminum die-cast rotors, by about 60%. This means that overall motor losses are reduced by about 15–20%. Die-cast copper rotor technology may boost machine efficiency due to decreased rotor losses, reducing cooling requirements. [19]. Copper gives weight and strength to machines. Copper rotor innovation does not overcome the fundamental difficulties of the IM as a vehicle technology, especially given increased economy and PD [20]. Graphical representation of torque and power in IM as depicted in Fig. 4.

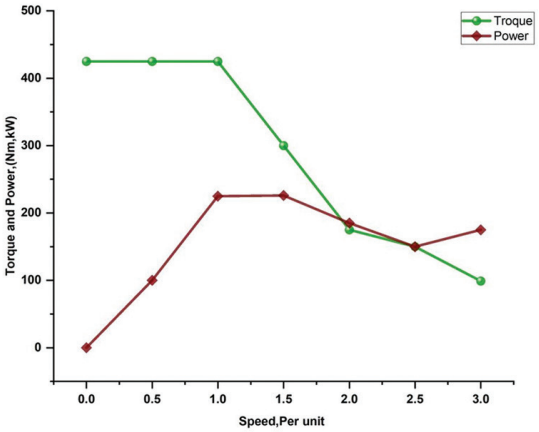


Fig. 4. Graphical representation of torque and power in IM

2.3. PERMANENT SYNCHRONOUS MAGNET MACHINE

PM machines are used in most of the machines in cars and trucks today. Most car machines employ permanent magnets. Increasing demands for high efficiency, specific power, and power density led to a move toward PMM, such as the Tesla Model 3's switch from IM to PM as shown in Fig. 5.

2010 Prius V-Shaped rotor 2017 Prius double U rotor

2016 Chevy Volt IPM rotor

Fig. 5. Rotor's IPM design for production traction motors [21]

There are many different architectures and types of PMM, but the rotor design serves as a fundamental characteristic of categorization into two major categories: permanent and surface magnet machines. Numerous crucial aspects of the machinery that have the fixed power speed range, are influenced by the rotor design [22]. The rotor design affects the fixed power speed range. Synchronous permanent magnet machines (SPMM) have a basic design/structure, however the rotor magnet causes a wider air gap, affecting performance, notably constant power speed range (CPSR).

Even though SPMM may be constructed with focused windings to increase CPSR, their use in automobiles is restricted due to the trend toward high torque, high PD machines with decreased magnet content. Interior per-

manent magnet (IPM) stators for traction machines can either have focused winding or scattered winding [23]. Concentrated windings feature fewer end turns, which reduces Joule loss and copper loss.

The dispersed windings might be randomly woven using strands or hairpin-wound bars. Comparing this winding design to the random wound, it is said to have greater slot fill, shorter end turns, and better production in thermal. Axial flux machines (AxFM) are a sort of PM machine that is receiving more and more negative press. For traction applications, the AxFMs offer qualities like high power density, high efficiency, small and modular construction, low weight, and high fault tolerance. They may take benefits of torque generation on numerous surfaces and have least current routes inside the machines [24].

2.4. RELUCTANCE

Synchronous reluctance machines (SyRMs) and the switching reluctance machine are two significant machine architectures that use the reluctance principle to generate torque. Both machines have a simple architecture with a rotor made simply of thin steel laminations and devoid of magnets. The SRM has relevant pole formation, but the SyRMs is normally non-salient, though it may be constructed with saliency [25]. Fig. 6 depicts the development of the SyRMs.

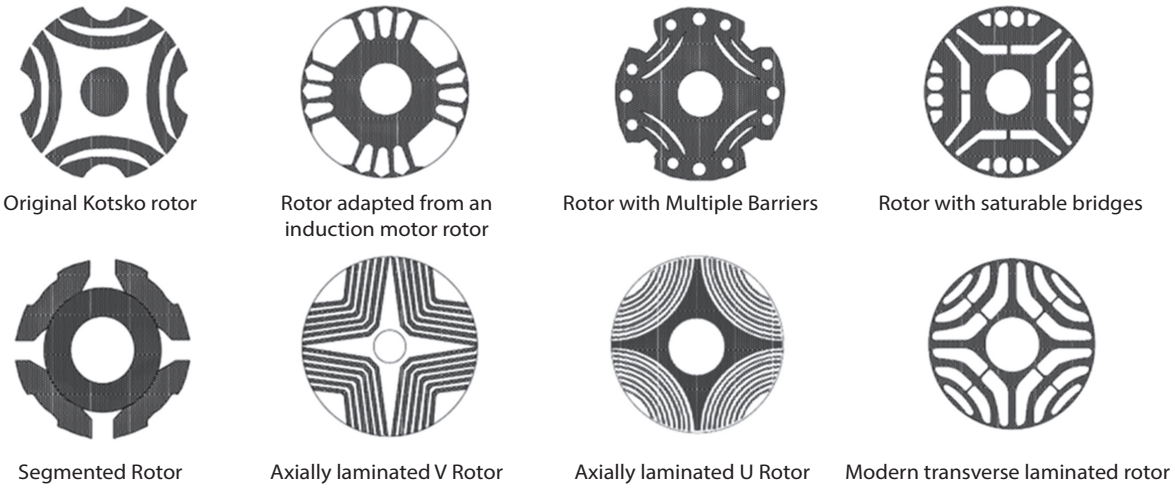


Fig. 6. Development of the SyRMs [43]

SyRMs are desirable because of their dependability, high efficiency, little torque ripple, and ease (cheap cost) of control. Due to their popularity as a perfect substitute for variable speed-controlled induction machines, and recently, these devices became available for industrial use [26].

While trying to improve their performance, SRMs are being evaluated for vehicle industries. These machines have high CPSR and great efficiency, but noise and transient response are still hurdles [27]. Low torque ripple (TR) is critical for EVs, especially in EV designs where the electric motor serves as the primary propulsion source. Significant study has been done in recent years on decrease in TR designs, including their eradication via the design of power electronic controllers. There have been several efforts made to improve the efficiency and PDof the SRM, including the usage of double stator and double-sided architectures.

By inserting magnets into stator poles, a new generation of machines is researched. It has been suggested that SRMs be used for traction drives, and several manufacturers have reportedly used or planned to employ SRMs in their drive trains. Prototypes for traction have been created. [28]. Fig. 7 shows an SRM built to match a Prius IPM. SRM may be a potential rival for applicant motors as magnet-free devices are in demand. These machines are now used in trucks and heavy machinery, and it is anticipated that they will eventually be used in light cars [29].

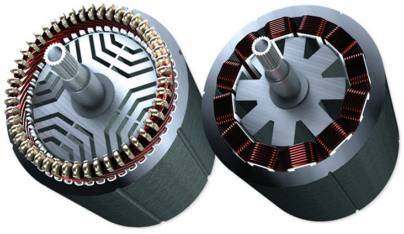


Fig. 7. SRM in traction [30]

2.5 IN WHEEL

The space available within the wheel places a restriction on the outside diameter of in-wheel motors (IWMs). Some designs use a planetary gear and a brake disc, IWMs may also be operated directly [31]. The space available within the wheel places a restriction on the outside diameter of in-wheel motors (IWMs). Some IWMs employ a planetary gear and brake disc, whereas others are actuated directly. In concept, any architecture may be used, however PM motors with outer ro-

tors or axial flux have higher PD and volume usage. There are several variants for reluctance and in wheel induction motors [32]. Table 1 indicates the benefits and drawbacks of various motors.

Table 1. Benefits and Drawbacks of Various Motors

No.	Motors	Benefits	Drawbacks
1	DC motor	Affordable speed controls	Excessive starting torque may harm reducers
2	Induction Machines	Induction motors function efficiently	The motor can't be used for traction or lifting when strong beginning torque is required
3	Permanent synchronous magnet	Power supply not required	No ability to regulate field strength
4	Reluctance Machines	These motors work at very high speeds.	It has High TR and low power factor
5	In-wheel	The response of the accelerator is great.	The life expectancy is reduced.

3. DRIVELINE CONFIGURATION

Compared to IC engines, electric motors provide greater configuration flexibility. One electric motor per wheel provides a simple, lightweight, and more effective transmission without a differential. New automobile body shapes are possible because of the great variety of geometries in which electric motors may be built. Single traction motor, single speed gear reduction, and differential are employed for propulsion [33]. A different setup would have a motor in each wheel. In-wheel motors save transmission area, weight, and friction losses.

When turning, the interior wheel moves more slowly than the exterior wheel [34]. Fig. 8 displays inlet and outlet wheel trajectories during a turning regime. To prevent slide and provide stability, drivelines with 2 motors need separate control and an electronic differential. Induction motors are also affected, even though it is clear for synchronous motors.

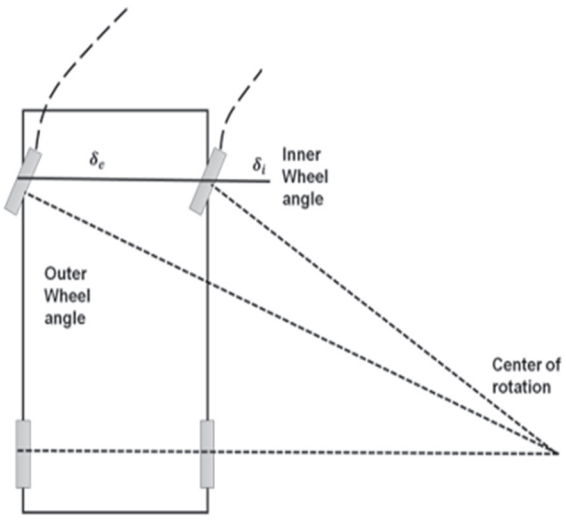


Fig. 8. Various trajectory of the inlet and outlet wheel

Induction motors' torque/slip characteristics point to a turning regime behavior that is unstable. The slower wheel has higher torque but risks losing grip, whereas the rapid wheel has less slip and consequently less torque [35]. Concept automobiles have been suggested using in-wheel engines. In particular at high speeds, the in-wheel motor has inferior dynamic performance than conventional power trains. The vehicle's unsprung mass is significantly increased by the in-wheel motor. In-wheel motors are more efficient and weigh less than mechanical drive systems. In-wheel motors are used in situations where performance is valued above comfort, like as sport automobiles, and unbeatable in solar car racing [36].

4. BATTERY

The battery selection for HEVs and Plug-in Hybrid Electric Vehicles (PHEVs) has been evaluated to achieve the appropriate balance between the electric drive and the range-extending IC motor. The mechanical design and overall cost of the vehicle are both compromised by the battery's size. According to economic research, the ideal battery size correlates to a limited range of electric battery [37]. For EVs, a variety of battery technologies are available. Before lithium-ion technology was developed, nickel metal hydride (NMH) was the favored choice for higher performance when compared to lead acid batteries.

According to the market overview, the current market is practically divided into lead-acid and lithium batteries. This battery is inexpensive and safer, but they have a lower PD. They are becoming less common, and their current uses are limited to small vehicles. Lithium-ion batteries require extensive charging cycles and are highly combustible, but their greater PD makes them the preferred solution for most general and superior efficiency vehicle producers [38].

The battery's chemistry results in nonlinear equivalent circuit behavior. Power transients significantly shorten life duration. The PD of batteries built to tolerate power transients is lower. EV batteries have longer operational cycles than identical HEV batteries [39]. Table 2 summarizes the energy and PD achieved for various methods.

Table 2. Features of Several Battery Types

Battery	Application	Wh/kg	$/kW	W/kg
Nickel Metal			1500	usage dependent
Shin-Kobe	EV	140	usage dependent	3920
Shin-Kobe	HEV	77	usage dependent	1344
Saft	EV	105	usage dependent	1550
Saft	HEV	56	usage dependent	476

Battery	Application	Wh/kg	$/kW	W/kg
	Lithium ion		2000	usage dependent
Ovonic	EV	45	usage dependent	1000
Ovonic	HEV	68	usage dependent	200
Panasonic	EV	68	usage dependent	1093
Panasonic	HEV	46	usage dependent	240
	Lead acid		150	usage dependent
Panasonic	EV	34,2	usage dependent	250
Panasonic	HEV	26,3	usage dependent	389

The battery's capacity is the only determinant of an EV's range. As a result, the vehicle's application and a safety buffer must be considered while choosing a battery. A key influencer of consumer attitudes regarding EVs is the driver's concern that the battery would run out before arriving at the destination. From quadricycles to very capable sports automobiles, there are many different values [40].

5. OVERVIEW OF MARKET

Environmental concerns and strict emission regulations focus research on low emission and fuel-efficient automobiles. The hybrid arrangements have re-emerged in this situation. The electronic industry's leadership in the development of battery technology has revived interest in fully electric automobiles. With a limited market for zero-emission vehicles, the sector has become more established [41]. The biggest technological drawbacks of electric automobiles compared to ICs are their short range and lengthy recharging times. Combining electric and IC propulsion, manufacturers strive to make EVs more appealing from a business standpoint.

With a high efficiency hybrid powertrain and a limited electric range, PHEVs are possible. The Extended Range Electric Vehicle (EREV) includes an all-electric driveline and a small IC engine that only kicks in when the battery is running low. As a result of drivers' concern of the electric range, the EREV IC engine has benefits. Two market trends apply to pure EVs. Commuter devices have light batteries and restricted ranges [42]. Two market trends apply to pure EVs. On the one hand, models made for commuting have a light battery and limited range. These city automobiles are lightweight and have a limited maximum speed.

Electric cars with a long range and large batteries are available. This product is targeted at the high-performance market because of the batteries' weight and cost [43].

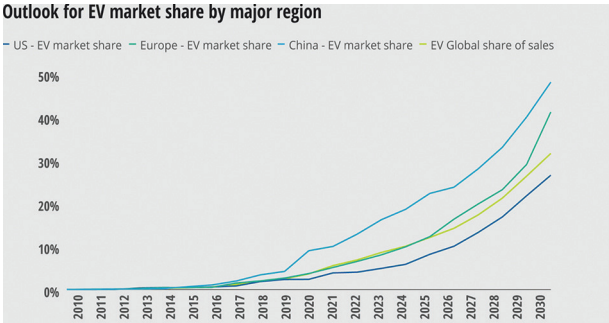


Fig. 9. EV market share projection by region [44]

In recent times, the EV market is booming with multiple alternatives in terms of range, efficiency and utility, with many products from established automakers and startups. This is further affirmed by market study and projections [44].

6. ELECTRICAL VEHICLE MACHINE TRENDS

This part includes some of the electric machine trends and challenges in electric vehicles.

6.1. ABSENCE OF RARE EARTH MAGNETS IN TRACTION MACHINES

The concurrent development of non-rare earth machine replacements is a significant trend in the development of machines for EVs. In addition to improving motor cost reduction, removing the rare-earth magnets also eliminates the reliance on this essential component. Induction machines have a decent chance of filling this requirement; however, the rising needs for high specific power and PD requirements make induction machines no longer a choice. The SRM and SyRMs machine topologies are two more crucial ones. Both machines feature a simple design with rotors made entirely of thin steel laminations. SRM controls are difficult and costly [45].

SyRMs are resilient, efficient, have minimum torque ripple, and are simple (cheap) to regulate. However, they have a decreased power factor, which impacts converter cost and size and a limited CPSR. Higher saliency ratios may mitigate these disadvantages. SynRM without magnets may be a low-cost motor and inverter if developed appropriately. SynRM and SRMs may allow for very efficient traction machines without rare earths. For non-rare-earth or reduced rare-earth drivetrains, improve SRM and SynRM [46].

6.2. IPM DEVICES USING RARE-EARTH MAGNETS

Motors with substantial rare-earth content may fulfill critical traction motor performance parameters, and most of these machines include rare earth magnets. IPM machines' ability to generate reluctance and permanent magnet torque and perform a broad operation has increased for traction drive systems. IPM can pro-

duce reluctance torque between 40 and 50% or higher [46]. This feature makes IPM ideal for certain applications. With the ability to extract high resistance torque, one would think magnet utilization in these machines would have stabilized over time, particularly as magnets make about 20–30% of a motor's cost and manufacturers would want to reduce that cost. However, it was stated that the usage of magnets has not reduced through time and may have even risen [47].

6.3. THERMAL MANAGEMENT SYSTEMS AND INTEGRATION OF MACHINERY

Another growing trend that will persist is the integration of power electronics and machines. Innovative integration approaches become increasingly crucial as space requirements for vehicle comfort rise, needing powertrain systems to become more compact. Advantages of density or reduced size, reduction in number of parts, shortened cable runs and busbars, and lower electromagnetic interference, and significant cost savings can be realized by enclosing power electronic drive and electric machine into a single compartment [48].

Integrated motor and power electronics may increase PD by at least 10% and reduce manufacturing and installation costs 30–40%. However, enclosing the electric machine and power electronics in the same compartment gives rise to system issues. Compound thermal management concerns are a challenge. Electronic boards are fragile and less tolerant to movement and roughness than motors. Combining them is difficult. Considering this, several tradeoff studies have been conducted to produce solutions that maximize integration advantages while decreasing system faults. Study [49] described four integration techniques and captured in Fig. 10.

Fig. 10. Option for motor and inverter integration [50,51]

The methods basically include attaching the power electronics to the axis of the motor utilizing radial and axial mounting approach in Fig. 9. Fig. 9(a) displays the power electronic inverter on the motor's casing, whereas Fig. 9(c) displays it on the motor's end shield. Fig. 9b shows the motor stator's perimeter fitted with power electronics, while Fig. 9d shows the stator's end attached with power electronics. For instance, the more typical technique depicted in Fig. 9a is straightforward to construct but only partially capable of providing high PD. The IM is mounted on the motor housing in this form, or on the side of the housing in other versions of the same idea [52]. Additionally, the motor and drive might employ a combined or independent cooling system [53].

The cooling system may be separate in certain circumstances, which results in less than ideal system volume utilization. The cooling system may be separate in certain circumstances, which results in less than ideal system volume utilization [54]. Although the stator curvature prevents these other designs from simply providing a flat surface to install electrical components, they do provide superior integration. The design of the EV, such as the number of motors and axles, will determine the integration method and degree of integration. The Chevy Volt serves as a typical illustration of radial housing implementation in study [55,56].

To create a longitudinal instead of a vertical layout, the Tesla cars with 2 axle motors and rear motor combinations toward the axial end plate mount [57,58]. Due to the significant benefits already described, it is generally anticipated that ever-tighter integration will be sought. Advanced thermal management systems are needed to maintain the increase in PD and specific power of vehicle traction drive systems. The paper provides a thorough examination of the cooling techniques used with traction motors, their analyses, and the calculation techniques. The study gave a description of the transmission approach used in vehicle traction motors, along with their advantages and disadvantages, as well as the conditions needed to optimize cooling performance for each method discussed.

The study examined the transmission approach used in vehicle traction motors, their trends and limits, and the conditions needed to maximize cooling efficiency. As machine power and vehicle range expand, efficient cooling systems become increasingly vital. Recent trends of compact motor drive integration make the adoption of contemporary thermal management systems that can cool motor drive components and other powertrain and automotive parts sensible.

7. CONCLUSION

This study provides a summary of significant advances toward high-PD machines used for traction in vehicles, with an emphasis on existing technology and the trends that are expected in future. Since most traction

machines are currently PMM and the development of devices with high specific power and PD is a major focus. The development of permanent magnet machines that produce greater reluctance torque appears to be a suitable field of work since this trend is predicted to continue in the future.

Alternative traction devices made without rare earth materials are becoming more popular. Based on the data, permanent magnet reluctance machines (switching and synchronous type) are preferred.

Based on motor performance, it can be concluded that PM motors are better than RMs, SBMs and IM, with IM being the least efficient in terms of performance. The cost of PM motor raw materials will decide whether PM motors become the industry standard or RM and SBM achieve a breakthrough in the market. The right choice of electric motor helps make design and packaging of the powertrain components easier owing to the size and lack of extensive thermal constraints. In-wheel motors make cars lighter and free up space inside, which allows for new body styles. This research can be further extended to include power converter components, as modern-day skateboard architecture encompasses drives, motors, and power management as a single unit. An integrated skateboard architecture in an electric vehicle, has closer integration with the power and thermal management system.

8. FUTURE WORK

This manuscript provides a review of motors and driveline topology used in electric vehicles. This study can be used for off-road vehicles, utility and long-haul vehicles as well. Based upon the traction, power and torque needed the use case of the motors can vary. The study can be extended to combine alternate energy sources such as hydrogen fuel cells. This review paper can be converted to a research study where different drivelines and configurations are modeled based on the physical parameters and the results can be documented.

9. REFERENCES

[1] JP. F. Mande, M. C. Trovao, M. HC. Ta, T. Van Do, "Dual-Source Bidirectional Quasi-Z-Source Inverter Development for Off-Road Electric Vehicles", World Electric Vehicle Journal, Vol. 13, 2022, pp. 174-174.

[2] C. S. Goli, S. Essakiappan, P. Sahu, M. Manjrekar, N. Shah, "Review of Recent Trends in Design of Traction Inverters for Electric Vehicle Applications", Proceedings of the IEEE 12th International Symposium on Power Electronics for Distributed Generation Systems, 2021.

[3] S. Sarkar, A. Das, "An Isolated Single Input-Multiple Output DC-DC Modular Multilevel Converter for Fast Electric Vehicle Charging", IEEE Journal of Emerging and Selected Topics in Industrial Electronics, 2022, pp. 1-10.

[4] R. S. Bharatkumar, S. Patel, S. Chakrabarti, S. Joshi, "Predictive Shift of Vehicles by Recent Market Trends", Lecture Notes in Electrical Engineering, 2022, pp. 909-924.

[5] L. Lander, E. Kallitsis, A. Hales, J. S. Edge, A. Korre, G. Offer, "Cost and carbon footprint reduction of electric vehicle lithium-ion batteries through efficient thermal management", Applied Energy, Vol. 289, 2021, p. 116737.

[6] M. Inci, M. Büyük, M. H. Demir, G. İlbey, "A review and research on fuel cell electric vehicles: Topologies, power electronic converters, energy management methods, technical challenges, marketing and future aspects", Renewable and Sustainable Energy Reviews, Vol. 137, 2021, p. 110648.

[7] I.-W.-M. E. "Vehicles and their Drivetrains", International Journal of Traffic and Transportation Engineering, Vol. 10, No. 4, 2020.

[8] M. Fichtner, "Recent Research and Progress in Batteries for Electric Vehicles", Batteries & Supercaps, Vol. 5, No. 2, 2021.

[9] S. Ebrie, Y. J. Kim, "Investigating Market Diffusion of Electric Vehicles with Experimental Design of Agent-Based Modeling Simulation", Systems, Vol. 10, No. 2, 2022, pp. 28-28.

[10] D.-K. Ngo, M.-F. Hsieh, "Performance analysis of synchronous reluctance motor with limited amount of permanent magnet", Energies, Vol. 12, No. 18, 2019, p. 3504.

[11] X. Wang, "A critical review on thermal management technologies for motors in electric cars", Applied Thermal Engineering, Vol. 201, 2022, p. 117758.

[12] K. Diao, X. Sun, G. Bramerdorfer, Y. Cai, G. Lei, L. Chen, "Design optimization of switched reluctance machines for performance and reliability enhancements: A review", Renewable and Sustainable Energy Reviews, Vol. 168, 2022, p. 112785.

[13] M. Burston, K. Ranasinghe, A. Gardi, V. Parezanovic, R. Ajaj, R. Sabatini, "Design principles and

digital control of advanced distributed propulsion systems", Energy, Vol. 241, 2022, p. 122788.

[14] N. Meddour, A. R. Babin, "The influence of driving cycle characteristics on motor optimization for electric vehicles", Proceedings of the 30th Mediterranean Conference on Control and Automation, 2022.

[15] M. N. Ibrahim, P. Sergeant, E. M. Rashad, "Influence of rotor flux-barrier geometry on torque and torque ripple of permanent-magnet-assisted synchronous reluctance motors", Proceedings of the XXII International Conference on Electrical Machines, 2016.

[16] Z. Zhang, J. Chen, X. Liu, "Hybrid-Excited Multi-Tooth Flux Switching Brushless Machines for EV Propulsion", Proceedings of the IEEE Transportation Electrification Conference and Expo, 2019.

[17] Z. Cao, A. Mahmoudi, S. K. Wen, L. Soong. "An overview of electric motors for electric vehicles." Proceedings of the 31st Australasian Universities Power Engineering Conference, 2021, pp. 1-6.

[18] J. Falck, C. Felgemacher, A. Rojko, M. Liserre, P. Zacharias, "Reliability of Power Electronic Systems: An Industry Perspective", IEEE Industrial Electronics Magazine, Vol. 12, No. 2, 2018, pp. 24-35.

[19] N. Novas et al. "Global Perspectives on and Research Challenges for Electric Vehicles", Vehicles, Vol. 4, No. 4, 2022, pp. 1246-1276.

[20] S. Chowdhury, E. Gurpinar, G. J. Su, T. Raminosoa, T. A. Burress, B. Ozpineci, "Enabling Technologies for Compact Integrated Electric Drives for Automotive Traction Applications", Proceedings of the IEEE Transportation Electrification Conference and Expo, 2019.

[21] J. Pyrhonen, J. Nerg, P. Kurronen, U. Lauber, "High-Speed High-Output Solid-Rotor Induction- Motor Technology for Gas Compression", IEEE Transactions on Industrial Electronics, Vol. 57, No. 1, 2010, pp. 272-280.

[22] M. Doppelbauer, P. Winzer, "A lighter motor for tomorrow's electric car", IEEE Spectrum, Vol. 54, No. 7, 2017, pp. 26-31.

[23] A. Habib, "A systematic review on current research and developments on coreless axial-flux permanent magnet machines", IET Electric Power Applications, Vol. 16, No. 10, 2022, pp. 1095-1116.

[24] X. Sun, K. Diao, G. Lei, Y. Guo, J. Zhu, "Study on Segmented-Rotor Switched Reluctance Motors With Different Rotor Pole Numbers for BSG System of Hybrid Electric Vehicles", IEEE Transactions on Vehicular Technology, Vol. 68, No. 6, 2019, pp. 5537-5547.

[25] S. Amin, S. Khan, S. S. H. Bukhari. "A comprehensive review on axial flux machines and its applications", Proceedings of the 2nd International Conference on Computing, Mathematics and Engineering Technologies, 2019, pp. 1-7.

[26] C. Liu, K. Wang, S. Wang, Y. Wang, J. Zhu, "Torque ripple reduction of synchronous reluctance machine by using asymmetrical barriers and hybrid magnetic core", CES Transactions on Electrical Machines and Systems, Vol. 5, No. 1, 2021, pp. 13-20.

[27] J. Mohanraj, B. Gopalakrishnan, L. Chokkalingam, and Mihet-Popa, "Critical Aspects of Electric Motor Drive Controllers and Mitigation of Torque Ripple-Review", IEEE Access, Vol. 10, 2022, pp. 73 635-73 674.

[28] N. Taran, D. Klink, G. Heins, V. Rallabandi, D. Patterson, D. M. Ionel, "A Comparative Study of Yokeless and Segmented Armature Versus Single Sided Axial Flux PM Machine Topologies for Electric Traction", IEEE Transactions on Industry Applications, Vol. 58, No. 1, 2022, pp. 325-335.

[29] T. Jahns, "Getting Rare-Earth Magnets Out of EV Traction Machines: A review of the many approaches being pursued to minimize or eliminate rare-earth magnets from future EV drivetrains", IEEE Electrification Magazine, Vol. 5, No. 1, 2017, pp. 6-18.

[30] W. Lee, R. A. Torres, H. Dai, T. M. Jahns, B. Sarlioglu, "Integration and Cooling Strategies for WBG-based Current-Source Inverters-Based Motor Drives", Proceedings of the IEEE Energy Conversion Congress and Exposition, 2021, pp. 5225-5232.

[31] M. Bostanci, A. Moallem, B. Parsapour, B. Fahimi, "Opportunities and Challenges of Switched Reluctance Motor Drives for Electric Propulsion: A Com-

parative Study", IEEE Transactions on Transportation Electrification, Vol. 3, No. 1, 2017, pp. 58-75.

[32] Y. Cha, K. Hyun, K. Yi, J. Park, "An integrated control of front in-wheel motors and rear electronic limited slip differential for high-speed cornering performance", Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, Vol. 236, No. 7, 2021, pp. 1355-1374.

[33] A. Habib, "A systematic review on current research and developments on coreless axial-flux permanent-magnet machines", IET Electric Power Applications, Vol. 16, No. 10, 2022, pp. 1095-1116.

[34] M. Henriksen, W. Throndsen, M. Ryghaug, T. M. Skjølsvold, "Electric vehicle charging and end-user motivation for flexibility: a case study from Norway", Energy, Sustainability and Society, Vol. 11, No. 1, 2021.

[35] C. Madariaga, J. Tapia, N. Reyes, W. Jara, M. Degano, "Influence of constructive parameters on the performance of an axial-flux induction machine with solid and magnetically anisotropic rotor." Proceedings of the IEEE Energy Conversion Congress and Exposition, 2021, pp. 4037-4044.

[36] A. Ahmed, H. A. Eissa, A. M. Faraj, A. Albagul, M. Belrzaeg, A. Alsharif, "Suspension System Modeling And Control For An Electric Vehicle Driven by In-Wheel Motors", Proceedings of the IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, 2021.

[37] X. Liu, F. Zhao, Z. Liu, "Energy-saving cost-effectiveness analysis of improving engine thermal efficiency and extending all-electric range methods for plug-in hybrid electric vehicles", Energy Conversion and Management, Vol. 267, 2022, p. 115898.

[38] A. Kebede, "Techno-economic analysis of lithium-ion and lead-acid batteries in stationary energy storage application", Journal of Energy Storage, Vol. 40, 2021, p. 102748.

[39] C. Li, N. Cui, Z. Cui, C. Wang, C. Zhang, "Novel equivalent circuit model for high-energy lithium-ion batteries considering the effect of nonlinear solid-phase diffusion", Journal of Power Sources, Vol. 523, 2022, p. 230993.

[40] G. Adu-Gyamfi, H. Song, E. Nketiah, B. Obuobi, M. Adjei, D. Cudjoe, "Determinants of adoption intention of battery swap technology for electric vehicles", Energy, Vol. 251, 2022, p. 123862.

[41] P. Ahmadi, A. Khoshnevisan, "Dynamic simulation and lifecycle assessment of hydrogen fuel cell electric vehicles considering various hydrogen production methods", International Journal of Hydrogen Energy, Vol. 47, No. 62, 2022, pp. 26758-26769.

[42] B. Xiao, J. Ruan, W. Yang, P. D. Walker, N. Zhang, "A review of pivotal energy management strate- gies for extended range electric vehicles", Renewable and Sustainable Energy Reviews, Vol. 149, 2021, p. 111194.

[43] Z. Wang, T. W. Ching, S. Huang, H. Wang, T. Xu, "Challenges Faced by Electric Vehicle Motors and Their Solutions", IEEE Access, Vol. 9, 2021, pp. 5228-5249.

[44] E. Schaltz, D. I. Stroe, K. Norregaard, L. S. Ingvardsen, A. Christensen, "Incremental Capacity Analysis Applied on Electric Vehicles for Battery State-of-Health Estimation", IEEE Transactions on Industry Applications, Vol. 57, No. 2, 2021, pp. 1810-1817.

[45] M. E. Hadraoui, A. Zegrari, O. Chebak, N. Laayati, and Guennouni, "A Multi-Criteria Analysis and Trends of Electric Motors for Electric Vehicles", World Electric Vehicle Journal, Vol. 13, No. 4, 2022, pp. 65- 65.

[46] F. Al-Amyal, L. Szamel, M. Hamouda, "An enhanced direct instantaneous torque control of switched reluctance motor drives using ant colony optimization", Ain Shams Engineering Journal, 2022, p. 101967.

[47] M. Murataliyev, M. Degano, M. D. Nardo, N. Bianchi, C. Gerada, "Synchronous Reluctance Machines: A Comprehensive Review and Technology Comparison", Proceedings of the IEEE, Vol. 110, No. 3, 2022, pp. 382-399.

[48] S. Zheng, X. Zhu, Z. Xiang, L. Xu, L. Zhang, C. H. T. Lee, "Technology trends, challenges, and opportunities of reduced-rare-earth PM motor for modern electric vehicles", Green Energy and Intelligent Transportation, Vol. 1, No. 1, 2022, p. 100012.

[49] S. Chowdhury, E. Gurpinar, G. J. Su, T. Raminosoa, T. A. Burress, B. Ozpineci, "Enabling Technologies for Compact Integrated Electric Drives for Automotive Traction Applications", Proceedings of the IEEE Transportation Electrification Conference and Expo, 2019.

[50] T. He, "Permanent Magnet Machines for High-Speed Applications", World Electric Vehicle Journal, Vol. 13, No. 1, 2022, p. 18.

[51] X. Wang, "A critical review on thermal management technologies for motors in electric cars", Applied Thermal Engineering, Vol. 201, 2022, p. 117758.

[52] R. Abebe, "Integrated motor drives: state of the art and future trends", IET Electric Power Applications, Vol. 10, No. 8, 2016, pp. 757-771.

[53] Y. Gai, "Cooling of Automotive Traction Motors: Schemes, Examples, and Computation Methods", IEEE Transactions on Industrial Electronics, Vol. 66, No. 3, 2019, pp. 1681-1692.

[54] I. Aghabali, J. Bauman, P. J. Kollmeyer, Y. Wang, B. Bilgin, A. Emadi, "800-V Electric Vehicle Powertrains: Review and Analysis of Benefits, Challenges, and Future Trends", IEEE Transactions on Transportation Electrification, Vol. 7, No. 3, 2021, pp. 927-948.

[55] E. Agamloh, A. von Jouanne, A. Yokochi, "An Overview of Electric Machine Trends in Modern Electric Vehicles", Machines, Vol. 8, No. 2, 2020, p. 20.

[56] H. El Hadraoui, O. Laayati, N. Guennouni, A. Chebak, M. Zegrari, "A data-driven Model for Fault Diagnosis of Induction Motor for Electric Powertrain", Proceedings of the IEEE 21st Mediterranean Electrotechnical Conference, Palermo, Italy, 2022, pp. 336-341.

[57] Deloitte, Electric vehicles, https://www2.deloitte.com/us/en/insights/focus/future-of-mobility/electric-vehicle-trends-2030.html (accessed: 2023)

[58] J. Wang, Q. Miao, X. Zhou, L. Sun, D. Gao, H. Lu. "Current Control Method of Vehicle Permanent Magnet Synchronous Motor Based on Active Disturbance Rejection Control", World Electric Vehicle Journal, Vol. 14, No. 1, 2023, p. 2.

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

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

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to https://ijeces.ferit.hr. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper:
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as "the Work"):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author's personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

_____ _____

Author/Authorized Agent **Date**

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