Breast cancer classification and prediction methods by employing machine and deep learning approaches-A survey

Sirisha Yerraboina[1*], *Harikrishna* Bommala[1], *Vineela* Madireddy[2]

[1*]Department of CSE, Bharatiya Engineering Science and Technology Innovation University, Anantapur, Andhra Pradesh.
[2]Department of Computer Science and Technology (AI&ML), Vignan Institute of Technology and Science, Deshmukhi, Nalgonda, India

Abstract. Breast carcinoma stands as one of the most perilous afflictions affecting females, lacking an effective treatment to date. Recent advancements in deep learning techniques, coupled with artificial intelligence (AI), have demonstrated promising results in breast cancer identification. This innovation facilitates early detection, consequently enhancing patient survival rates. Deep learning necessitates minimal human intervention for feature extraction, contrasting with traditional machine learning methods. The ML and DL techniques are practised and comparison of all these techniques were shown. Specifically, emphasis is placed on genomic and histopathologic imaging data. Various algorithms, including R, SVM, logistic regression, KNN, Naïve Bayes, CNN, and ANN, are thoroughly researched and valuated to gauge their efficacy. Furthermore, many screening protocols were deployed to identify and examine the datasets. Lastly, the paper explores the challenges encountered and the most possible directions to detect the breast cancer .Hence researchers and clinicians with a thorough understanding and insights into this deep learning domain.

1 Introduction

Cancer emerges when abnormal cells in the body begin to divide and interact with healthy cells. Breast cancer stands as one of the most prevalent and perilous diseases worldwide, categorized into two types: invasive and non-invasive.

Invasive breast cancer, being cancerous, malignant, and invasive, spreads to various organs, while non-invasive cancer remains confined to the original organ until it progresses into aggressive breast cancer [1]. The glands responsible for milk transportation serve as the primary sites for breast cancer. It frequently metastasizes to other organs through the bloodstream, exhibiting diverse growth rates for each cancer type. According to WHO, breast carcinoma claimed the lives of 8, 64,000 females in 2020, constituting a significant

[*] Corresponding author: haribommala@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

global health concern, particularly prevalent in the United States. There are four main types of breast carcinoma: DCIS-Ductal Carcinoma in Situ, Secondary breast carcinoma, inflammatory breast carcinoma, and Metastatic breast carcinoma. Diagnostic imaging modalities such as mammograms, ultrasonography, MRI, and biopsy generate the necessary images for classification. Mammograms employ X-rays to detect breast cancer, with further examination conducted upon detecting suspicious areas. Ultrasonography is utilized following mammography to investigate suspicious findings, while breast MRI provides comprehensive insights into the illness. Biopsy remains the primary tool for confirming suspected cancerous sites. Fortunately, the majority of women who undergo mammograms are found not to have breast cancer. As we gather the images from the multiple sources to classify those images Machine learning techniques helps us to classify the breast images. Machine techniques are used for training the models and to predict helps in taking the decisions. For all the classification and prediction problems machine learning methods helps to retrieve the best end results.ML helps in breast cancer research to find whether the cancer in the body is benign or malignant.

2 Literature Review

DL model that makes use of the TL technique was proposed by Abeer Saber et al for diagnosing and identifying breast cancer using two methodologies: 80-20 cross-validation [02]. To preprocess the features the pre-trained CNN was developed on various image datasets. Various systems including InceptionV3, ResNet-50, VGGNet.19, VGGNet16, and Inception-V2 ResNet were employed. Results indicate that VGG16 proved to be a robust model for accurately classifying breast cancer images.

An adaptive median filter was proposed by P. Esther Jebarani et al. [03] with the goals of improving image quality and reducing noise. The study evaluated performance in distinguishing between benign and malignant tumours using the k-means and Gaussian mixture model approaches. Additionally, ANOVA test was employed for conducting multivariate analysis to predict the rate of prediction.

For the purpose of processing breast cancer photos, Mohammed Abdulla Salim Al Husaini et al. introduced two different models: artificial neural networks(ANN) and deep learning models(DL)[04].To evaluate the best accuracy we used the methods like radial basis functions,k-nearest neighbour algorithm and support vector machine They attained an accuracy rate of up to 80%.

The ensemble of classifiers is another technique used by the author Usman Naseem et al. for automated prognosis of breast cancer , achieving an accuracy of 98.83% [05]. Additionally, they utilized ensemble methods, further enhancing the accuracy to 98.83%.

Deep learning(DL) and machine learning(ML) methods for breast cancer diagnosis were introduced by Marco Repetto et al. [06]Their model utilizes federated learning, enabling training of images on a decentralized dataset. The authors suggested a multiple criteria optimization approach for federated learning.

Yongjun Wang et al. introduced a method for breast cancer detection utilizing histopathological images [07]. They employed pretrained convolutional neural networks for reducing feature dimensionality and trained the images using E-SVM. In addition, they applied dual-network orthogonal low-rank learning methods to improve performance.

Ranjini K et al. introduced deep learning and machine learning methods for investigating breast cancer. With Convolutional Neural Networks (CNN) demonstrating notable accuracy in recent studies, there has been a surge in popularity for medical image analysis utilizing this technique [08]. This research aims to compile and compare various breast cancer classifiers currently in use. The adoption of advanced deep learning techniques holds promise for enhancing performance. Future advancements in performance and accuracy are anticipated as a result of these endeavors.

A machine-learning strategy using the multilayer perceptron network (MLP) technique was presented by Huan-Jung Chiu et al.The network was configured to analyse data variations with increasing or decreasing dimensions [09]. The model had been designed to analyse high-dimensional input and subsequently generate low-dimensional data. Transfer learning techniques were employed by Support Vector Machine (SVM) to use characteristic data as classifiers after training. This allowed the models to distinguish between representative qualities and quantities. Ultimately, an accuracy of 86.97% was attained.

Jing Zheng et al. discussed the under utilization of neural networks in cancer data classification. Their paper introduces an effective AdaBoost algorithm (DLA-EABA) supported by deep learning for breast cancer detection, utilizing advanced statistical techniques[10]. Besides conventional computer vision methods, they explore tumor transfer classification methods employing deep transformational neural networks (CNNs). The results shows the highest accuracy using the above models.

Nina Youneszade et al deployed the models like Machine learning and deep learning techniques for the automatic segmentation and classification of images of cervical cytology and colposcopy images within computer-aided diagnosis systems [11]. Their work comprehensively addresses various deep learning techniques, outlining their architectures, classification methodologies, and segmentation approaches for both colposcopy and cervical cytology images.

An effective Adaboost approach with deep learning support for breast cancer detection and categorization was introduced by Jing Zheng et al. [12]. They highlighted the active utilization of deep neural networks (CNNs) in developing tumour classification methods that extend beyond conventional computer vision techniques. Transfer learning is also employed in this context. In this paper the efficient techniques like transfer learning is used to detect the breast density on the variety of histopathelogical images like breast tomosynthesis,Ultra sound scanning and magnetic resonance images for the prediction and the detection of breast classification. Based on experimental results, the BC detection system shows the highest accuracy of 97.2% compared to the other existing systems.

Nipun B Nair et al. advocated Support Vector Machine to be the most effective ML techniques [13]. When integrated with the computational prowess of Convolutional Neural Networks (CNNs), it emerges as a highly potent classification algorithm. This hybrid model has demonstrated superior accuracy compared to other image classifiers such as VGGNet.16, Res Net-50 and Inception-v3 models. Through research experiments, the (SVM) Kernels –(CNN) Kernels, VGGNet.16, Res Net-50 and Inception-v3 models were assessed, yielding reported accuracies of 93.35%, 89.54%, 92.45%, and 88.6%, respectively.

Amin Ul Haq et al. introduced a method that combines both Relief algorithm & Autoencoder, PCA algorithm systems to extract relevant features from a dataset [14]. For the BC detection the best features are used to train the model and we can classify using the svm classifier. To validate the best model we can retrieve the optimal parameters using the cross validation technique.The best evaluation metrics were used to test and classify the BC datasets.The models testing was done using breast cancer datasets. The study of experimental results revealed that features selected by the Relief algorithm exhibited greater accuracy compared to those selected by PCA algorithms and Autoencoder in breast cancer detection. Finally the RA algorithm retrieved the best accuracy of 99.91by crossing all the previous methodologies and techniques.This superiority was confirmed by the McNemar test.

Sweta Bhis et al. proposed the utilization of ConvNet (CNN) as the primary Machine Learning algorithm, which excels in feature extraction [15]. CovNet CNN is compared with many other machine learning techniques like logistic regression,K-Nearest neighbour,support vector machine?(SVM) and Naive bayes algorithms . The dataset utilized was obtained from Hist Kaggle. The primary aim is to accurately classify the data and assess the models' performance by comparing their accuracy and recall rates. Additionally, the analysis includes F1 scores and accuracy metrics. Upon conducting the experiments, the findings indicate that ConvNet demonstrates the highest accuracy with the lowest loss/error rate.

3 Breast cancer detection Techniques

ML, a subset of AI, offers the ability to retrain models and improve performance, particularly effective in processing linear data. Deep learning, a subcategory of machine learning, derives knowledge from unstructured or unlabelled data through deep neural networks comprising multiple hidden layers. ConvNet (CNN) is employed in classifying breast cancer datasets, demonstrating significant potential in improving diagnostic accuracy. To investigate the breast cancer the traditional methods which are used are very tedious and needs more observational analysis. In the recent decades Computer aided diagnosis(CAD) methods are applied on patients in intensive care units and it requires continuous monitoring of patients. By using ML many more techniques are used for classification and prediction. The techniques like multi layer perceptron, naïveBayesian and decision trees are used. All these techniques are tested on the data sets like Wisconsin breast cancer(original) data sets.

A ConvNet (CNN) is employed to categorize the images within the breast carcinoma database. Each image, along with its respective weights, is fed into the CNN. These weights undergo adjustments to minimize errors and enhance performance over time.

1.Step by step procedure for breast cancer detection and classification:

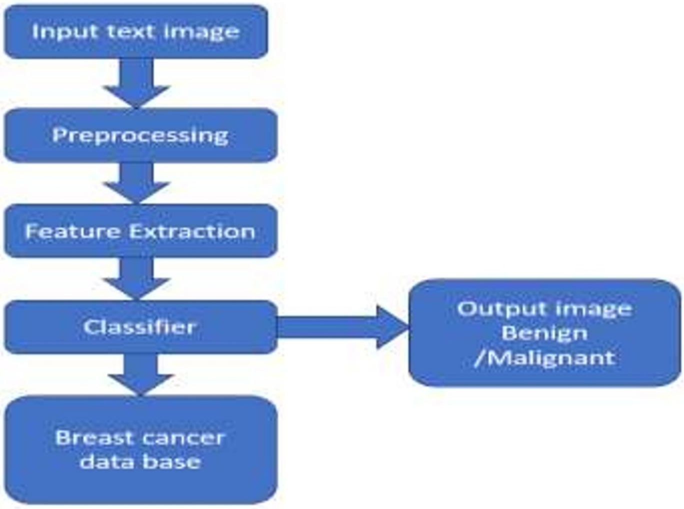

Fig.1. Flow diagram of BC detection system

Due to the evolution of technology in the medical field there are many approaches in the detection of breast cancer. The techniques are

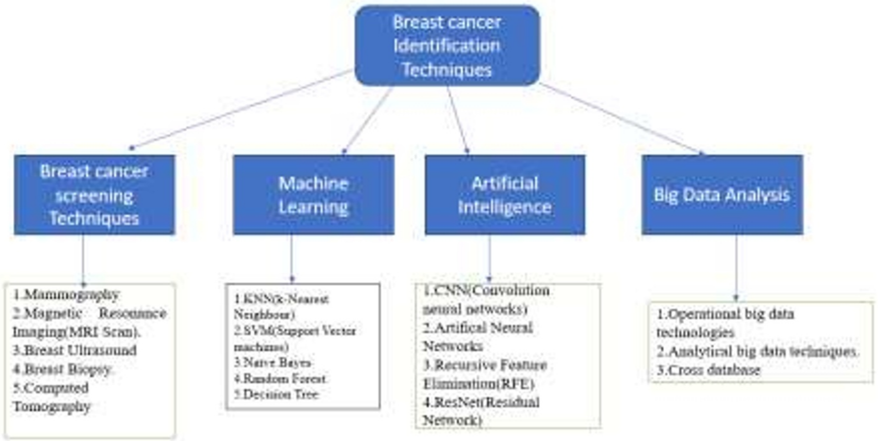
Breast cancer Identification Techniques
Breast cancer screening Techniques
Machine Learning
Artificial Intelligence
Big Data Analysis
1.Mammography
2.Magnetic Resonance Imaging(MRI Scan).
3.Breast Ultrasound
4.Breast Biopsy
5.Computed Tomography
1.KNN(k-Nearest Neighbour)
2.SVM(Support Vector machines)
3.Naive Bayes
4.Random Forest
5.Decision Tree
1.CNN(Convolution neural networks)
2.Artifical Neural Networks
3.Recursive Feature Elimination(RFE)
4.ResNet(Residual Network)
1.Operational big data technologies
2.Analytical big data techniques.
3.Cross database

Fig.2. Breast Cancer detection techniques using multiple technologies

3.1 Breast screening techniques

To diagnose the breast cancer the following tests were conducted on the patients like mammography, (MRI scan), Breast biopsy and Computed tomography All the physical examinations of the patient will be conducted on the patients using these tests.

3.1.1 Using Machine learning techniques

In Breast cancer detection ML techniques are extensively used in the classification and prediction. In machine learning the most efficient techniques like decision tree, random forest, Naive bayes,SVM,K-nearest Neighbour(KNN) algorithms were deployed.More over these methods helps us in identification and prediction of breast cancer and for best decision making.

3.1.2 Artificial Intelligence

The techniques of AI and deep learning were widely used are convolution neural networks, artificial neural networks, recursive feature elimination(RFE),RESNET(Residual Network) and many more are used. The deep learning also plays one of the predominant role in the classification and prediction of breast cancer. The accuracy is also high using these techniques.

3.1.3 Using Big data Analysis

Big data analytics plays huge role in the health sector industry. By using different analytics tools, we can analyse the images with huge data sets. The tools like Apache Hadoop, Hbase,Hive,Spark and many more tools are used for storing the huge data sets and can go for analysis. Operational big data technologies, analytical big data techniques, cross database are used in the detection of breast cancer.

3.2 Methodologies used

3.2.1 SVM

SVM is supervised machine learning technique which implements the linear discriminant function by selecting the moderate number of samples or by using support vectors.The linear boundary constraint was solved by SVM[5]. The SVM can be considered as a linearly partitionable data set exhibiting the maximum over plane margin of two classes.The new models, after selecting an appropriate map, are either linearly fitted or apparently linearly discretized in a higherlevel plane.By comparing the results on WDBC datasets among all the classifiers svm classifier achieved the highest accuracy.The (EHR) data set,achieving 93.26 % accuracy, produced two data sets using the two hybrid techniques like gray wolf optimization and support vector machine.

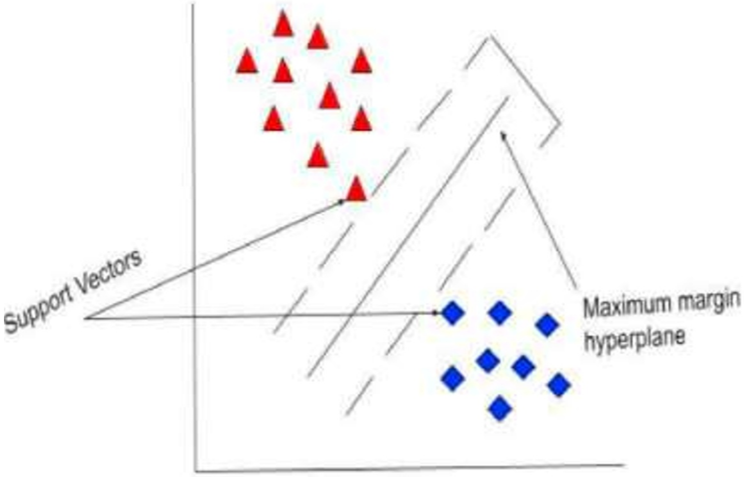

Fig. 3. Support vector machine

3.2.2 Logistic Regression

Reconstructing the posterior probabilities of the k groups in a linear function of x, when they are equal to one and ensuring that they stay in the range [0, 1],and it helps us in building the LR method,Linear regression can be explained with probabilities and log shits.

The choice of denominator is arbitrary because the estimates are evenly distributed, although the last set is the denominator of the odd ratio. When K = 2, there is only one linear function, so the procedure is straightforward. This method is commonly used in biostatic functions reproducing binary reactions.

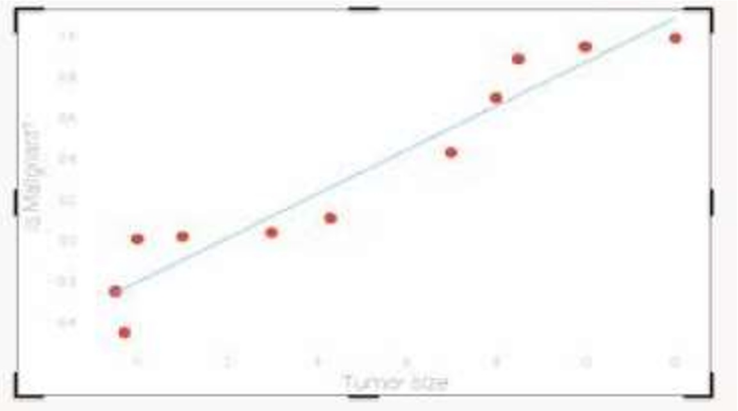

Fig. 4. Logistic regression

3.2.3 Decision Tree

Decision tree helps in grouping the breast cancer samples into the specified category of groups. An example of accurate cluster attributes being identified by combining nominal and numerical attributes is represented by an example.The features are represented as models which helps in forming a decision tree.The DT can be constructed using if then else process. The DT tree classifiers like J48 and CART algortihms are used to achieve the highest accuracy for classification of breast cancer.As J48 depends on ID3 iterative dichotamiser to classify the continuous and categorical data.

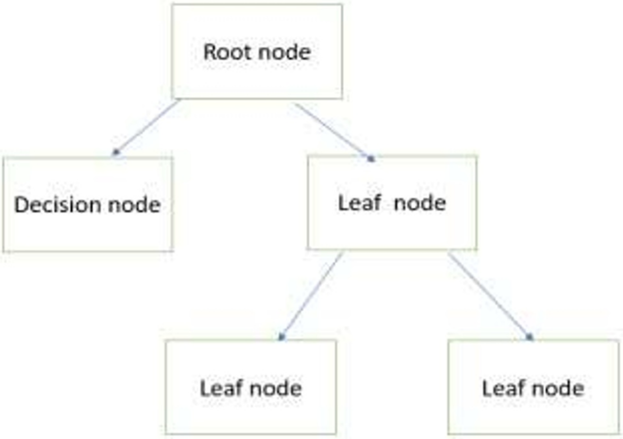

Fig. 5. Decision Tree

3.2.4 Naïve Bayes

Naïve bayes algorithm can be implemented using bayes theorem.We conclude that T is an informative training set of instances. These specimens have clusters of markings. The names of each group are K1, K2, · · · Kc.Each sample can be represented as an n-dimensional object by the formulas Y = y1, y2, · · · y_n.Given that Y has n dimensions, it is said to have n features. Formally speaking, a specimen Y is projected to belong to group K_i

if the likelihood that group K_i depends on Y it is greater than the likelihood that each of the other groups depends on Y .

$$P(K_i\,|Y) > P(K_j\,|Y)\ for\ 1 \leq j \geq k_j = i) \qquad (1)$$

$P(K_i\,|Y)$ is calculated as

$$P(K_i|Y) = P(Y|Y_i)P(K_i)/P(Y) \qquad (2)$$

3.2.5 Artificial Neural Networks

As scholars have used ANNs in recent decades, this area of study is relevant. Partnerships have contributed significantly to success, particularly in the areas of early prediction and BC classification. In the recent decades most of the work id done using the ANN techniques. Most specifically in the area of BC classification and prediction. The three layers of ANN models are :input,Hidden,output layer.To improve the nonlinear dynamics, the layers are constructed with networks of networks with nonlinear switching activation functions. First, the input layer receives the data, and then sends the findings to the hidden layer for analysis before sending them back to the output layer.The final results will be displayed using output layers. However, given the limitations, it is expected that a long unsupervised computing chain will be required to train the ANN. The ANN framework used in this work has two dropout layers and three coarse layers. In contrast, the DNN has three dropout layers and five dense layers.

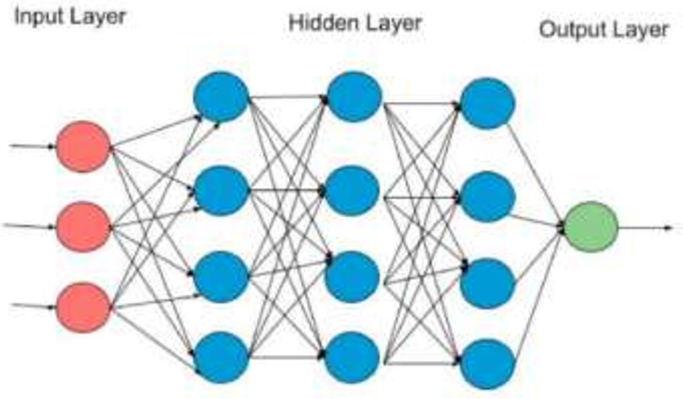

Fig. 6. Artificial Neural Network

3.2.6 Convolution Neural Networks

CNNs are generated by neurons and will always resemble traditional artificial neural networks. Multiple ANNs receiving inputs and performing operations will still serve each neuron. When the raw image vectors are translated as output into the class score as input, the network further represents the perceptual scoring function[15].The final layer is represented as loss functions and these are connected to classes.when compared to the other ML classifers preprocessing is very less in CNN.The working style of CNN is shown in the figure 3.4.we have the output of 99.67%in [16],F1-score as 98%[17] ,98% accuracy[18] applied on WDBC datasets.

In [19]they used CNN models for feature extraction and low ranking based orthogonal linear technique were used.They used ICIAR dataset and found the accuracy of

97.70%.SVM classifiers were used to classify the breast cancers.The pooling operation and gradient boosting algorithm is used after the feature extraction process for classification. We achieved the best results showing an accuracy of 93.8% and 92.50% To achieve the highest accuracy, transfer learning was used technique was implemented using CNN model (InceptionV3) and weights.

4 Performance Evaluation Metrics

4.1 Accuracy

Classification accuracy is defined as the ratio of correct prediction to total prediction.It is only useful in situations where all classes have the same amount of data and all predictions and prediction errors have the same weight, which is not always the case.

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

4.2 Sensitivity

The way a model can predict the positive cases is defined as sensitivity. The more recalls, the more positive samples can be determined.The ratio of actual positive cases that are correctly identified.The other name for sensitivity is recall.

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

4.3 Specificity

Specificity can be determined by the number of negative cases that are produced by ML models. The confusion matrix is used to determine the formula

$$\text{Specificity}=\frac{TN}{TP+FN} \qquad (1)$$

4.4 Precision

The precision is determined as the ratios of the true positives by the total number of positive cases. The model's accuracy in classifying a sample as positive is evaluated.

$$\text{Precision}=\frac{Tp}{TP+FP} \qquad (1)$$

4.5 F1 Score

The best metric to find the optimal errors is F1-Score.The average of precision and recall is F1 score. It helps us to evaluate the effectiveness of the machine learning model in terms of binary classification.

$$\text{F1 Score}=2*\frac{precision*recall}{precision+recall} \qquad (1)$$

5 Conclusion

In the recent decades multiple ML and DL algorithms are used on breast cancer Wisconsin data sets. The performance of SVM,logistic regression,KNN,Naïve Bayes,CNN,ANN are compared in this paper.The study of all these papers make us understand that convolution neural networks got the highest accuracy for the classification and prediction. In future we can employ the techniques like gaussian adversarial networks (GAN)and variational auto encoders(VAE) to get more accuracy to classify the images.

References

1. Aggarwal, R., Sounderajah, V., Martin, G. et al.npj Digit. Med. 4, **65** (2021).
2. A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk and H. Chen, in *IEEE Access*, vol. **9**, pp. 71194-71209, (2021). doi: 10.1109/ACCESS.(2021)
3. P. E. Jebarani, N. Umadevi, H. Dang and M. Pomplun, in *IEEE Access*, vol. **9**, pp. 146153-146162, (2021).
4. Mohammed Abdulla Salim Al Husaini , Mohamed Hadi Habaebi ,Shihab A. Hameed , Md. Rafiqul Islam and Teddy Surya Gunawan.(2020).
5. U. Naseem *et al.*in *IEEE Access*, vol. **10,** pp. 78242-78252, (2022)
6. M. Repetto and D. La Torre, *2022 5th International Conference on Signal Processing and Information Security (ICSPIS)*, Dubai, United Arab Emirates, , pp. **1-4**, (2022)
7. Yongjun Wang, Baiying Lei ,Ahmed Elazab , Ee-Leng Tan , Wei Wang , Fanglin Huang , Xuehao,Gong,andTianfuWang.IEEEAccessPP(**99**):(2020)
8. R. K and M. S. K, *IEEE 7th International conference for Convergence in Technology (I2CT)*, Mumbai, India, pp. **1-6**,(2022*)*
9. H. -J. Chiu, T. -H. S. Li and P. -H. Kuo, in *IEEE Access*, vol. **8**, pp. 204309-204324, (2020).
10. J. Zheng, D. Lin, Z. Gao, S. Wang, M. He and J. Fan, in *IEEE Access*, vol. **8**, pp. 96946-96954, (2020).
11. N. Youneszade, M. Marjani and C. P. Pei, in *IEEE Access*, vol. **11**, pp. 6133-6149, (2023)
12. N. B. Nair, T. Singh, A. Thakur and P. Duraisamy, *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. **1-6**, (2022).
13. A. U. Haq *et al.*in *IEEE Access*, vol. **9**, pp. 22090-22105, (2021).
14. Bhis, Shrutika Gadekar, Aishwarya Singh Gaur, Simran Bepari , Deepmala Kale, Dr. Shailendra Aswale International Journal of Engineering Research & Technology (IJERT). Vol.10(**7**):98,(2021)
15. Yadav, Rahul Kumar, Pardeep Singh, and Poonam Kashtriya. Procedia Computer Science 218 (2023).
16. N. Khuriwal and N. Mishra, International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp**98-103**,(2018).
17. S. S. Prakash and K. Visakha,2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. **88-92**(2020).

18. A. Algarni, B. A. Aldahri and H. S. Alghamdi, 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), pp. **1-5**,(2021).
19. Y. Wang et al.in IEEE Access, vol. **8**, pp. 27779-27792, (2020)