[1]K.P.Senthilkumar

[2]P. Narmatha

[3]Jonnadula Narasimharao

[4]Narendra Mustare

[5]N. Herald Anantha Rufus

[6]Yashapl Singh

A Comprehensive Analysis on the Efficacy of Machine Learning-Based Algorithms for Breast Cancer Classification

Abstract: - This research focuses on using machine learning to make breast cancer classification better. In this research various machine learning algorithm such as Decision Tree, Linear Discriminant, Support Vector Machine, K-Nearest Neighbors, Probabilistic Neural Network, Logistic Regression, Recurrent Neural Network, and Ensemble Method are used. We tested them using two different ways of splitting the data—90/10 and 70/30—and we also picked important features to consider. The Ensemble Method came in second place with accuracies of 98.2% and 97.6%. The Deep Neural Network performed really well too, with accuracies of 96.2% in the 90/10 split and 89.1% in the 70/30 split. We also found that selecting the right features improved accuracy a lot. This shows how important it is to choose the best features to make the models better. These results show that machine learning can be used to classify breast cancer effectively. The numbers prove that the Deep Neural Network and Ensemble methods have high accuracy, and selecting the right features makes them even better. The research outcomes introduces machine learning techniques that can improve breast cancer diagnosis, potentially changing the way doctors make decisions and improving patient outcomes.

Keywords: Breast cancer classification, Machine learning algorithms, Feature selection, Deep neural network, Ensemble methods

I. INTRODUCTION

A early diagnosis and successful treatment depend on an accurate classification of breast cancer, which is a major global health concern. Machine learning algorithms have made significant progress in recent years in terms of increasing the precision of breast cancer classification [1,2]. This study aims to provide a thorough evaluation of various machine learning algorithms for classifying breast cancer, including Decision Tree, Linear Discriminant, Support Vector Machine, K-Nearest Neighbors, Probabilistic Neural Network, Logistic Regression, Recurrent Neural Network, and Ensemble Method [3]. This work aims to determine the most efficient algorithms for accurate breast cancer classification by analyzing their predictive accuracy and using feature selection strategies.

Popular machine learning algorithm called Decision Tree bases choices on feature values on a tree-like structure [4,5]. The dataset is divided recursively depending on the most useful attributes, creating a decision-making tree [6,7]. Each leaf node represents a class name, whereas each interior node reflects a judgment based on a particular

[1] Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India

[2] Department of Electronics and Communications Engineering, Sri Sairam College of Engineering, Bangalore, Karnataka, India

[3] Department of Computer Science and Engineering, CMR Technical Campus, Kandlakoya, Medchal, Hyderabad, Telangana, India

[4]Department of EIE, CVR College of Engineering, Ibrahimpatan, Hyderabad, Telangana, India

[5]Department of Electronics and Communications Engineering, Vel Tech Rangarajan Dr Sagunthala R & D Institute of Science and Technology, Chennai, Tamil Nadu, India

[6]Department of Computer Science and Engineering, Amity University Rajasthan, India

Corresponding mail id: skkr8587@gmail.com

Email ids: skkr8587@gmail.com; narmathaphd@gmail.com; jonnadula.narasimharao@gmail.com; drnamust@gmail.com; drrufus@veltech.edu.in; yashpalsingh009@gmail.com

attribute. Due to their interpretability and capacity to handle both continuous and categorical information, decision tree algorithms like C4.5 and CART have been extensively used in the categorization of breast cancer [8,9].

A statistical approach called linear discriminant analysis (LDA) is utilized for dimensionality reduction and classification. Finding a linear feature combination that maximizes the distance between several classes is the goal of LDA [10,11]. Using this linear discriminant space to project the data, LDA can accurately categorize cases of breast cancer. LDA makes the assumptions that the covariance matrices of the various classes are equal and that the features have a multivariate normal distribution [12,13].

Support Vector Machine (SVM) is a potent binary classification technique that is widely employed in many fields, including the classification of breast cancer. Finding the best hyperplane to maximally separate the two classes is the goal of SVM [14,15]. SVM can successfully handle high-dimensional data and complicated decision limits in the classification of breast cancer. SVM can detect non-linear correlations between features by utilizing kernel functions like linear, polynomial, or radial basis functions [16,17].

A sample is categorized using the class labels of its closest neighbors in the feature space using the straightforward and understandable algorithm K-Nearest Neighbors (KNN). KNN is a good choice for breast cancer classification since it does not make any assumptions about the distribution of the underlying data, even though the interactions between the features can be intricate. The number of neighbors (K) and distance measure chosen affect the performance of KNN significantly [18,19].

A feed-forward neural network called a probabilistic neural network (PNN) uses a probabilistic approach to categorization. By comparing the input sample's feature vector to the feature vectors of training samples, PNN calculates the class probabilities. The conditional probability density function of each class is modeled using a radial basis function kernel. Due of PNN's capacity to deal with erratic and imperfect data, it has been used to classify cases of breast cancer [20–22].

A popular statistical approach for binary classification is logistic regression. The logistic function is used to model the relationship between the features and the likelihood of belonging to a particular class. The assumption behind logistic regression is that the characteristics and log-odds of the class probabilities are linearly related. Due to its readability and simplicity, it is frequently used to classify breast cancer. It can handle both categorical and continuous information [23,24].

An artificial neural network called a recurrent neural network (RNN) can recognize sequential information in data. Speech recognition and natural language processing are two areas where RNNs have seen success. RNNs can be used to examine time-series measurements or sequential gene expressions in the context of classifying breast cancer. RNN architectures that are frequently utilized include Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) [25,26].

To provide a final prediction, ensemble approaches integrate the results of various individual classifiers. Ensemble approaches are frequently more accurate than employing a single classifier alone because they take advantage of the diversity of individual classifiers. Bagging, Boosting, and Random Forest are examples of common ensemble approaches. Ensemble approaches have been thoroughly researched and have shown to perform better than individual classifiers in the classification of breast cancer [27,28].

An extensive dataset of breast cancer cases, including different characteristics and labeled outcomes, is employed in this study. The dataset is initially preprocessed, which involves missing value handling, data cleansing, and normalization. Techniques for feature selection are used to find pertinent qualities for classification. Then, using both 90/10 and 70/30 splits, the dataset is separated into training and testing subsets. These subsets are used for training and testing various machine learning algorithms, such as Decision Tree, Deep Neural Network, and others. Algorithm performance is measured using metrics including accuracy, precision, recall, and F1-score. This methodology offers a well-organized framework for methodically evaluating and contrasting the performance of several machine learning strategies for breast cancer classification.

II. METHODOLOGY

1.1 Dataset creation

A variety of machine learning techniques are used in this work to improve breast cancer prediction. Machine learning has been studied in the past for cancer prediction, but an interesting trend has emerged: machine learning models are frequently trained using massive datasets that already exist and contain records of various diseases from 1990 to 2020. These datasets, which contain a massive 19,000,000 entries and a wide range of distinguishing characteristics, have served as the foundation for efforts in predictive analytics.

This study paradigm, on the other hand, makes a significant attempt to predict breast cancer. The baseline dataset—a collection of various cancer records—was painstakingly screened to find relevant cases of breast cancer. This meticulous curation process resulted in the isolation of 80,882 records that only dealt with breast cancer, which served as the foundation for the subsequent investigation. To maximise the predictive capability of the chosen machine learning algorithms, a refined subset of 120 prominent breast cancer traits was selected from the entire feature set. This process ensures that the most important characteristics for predicting breast cancer are present, increasing the study's accuracy and usefulness.

The conversion of these carefully gathered breast cancer data and their associated features into a classification-ready format is the foundation of this work. By encoding the necessary properties into the algorithms of choice, complex medical data is converted into a structured format that these algorithms can understand. To facilitate analysis, the carefully crafted dataset is converted into a CSV file, allowing for simple interaction with the chosen machine learning techniques.

This study aims to improve the accuracy and usefulness of machine learning-based predictive models for breast cancer by paying close attention to the specific area of breast cancer prediction and painstakingly narrowing the dataset to contain only the most relevant instances and variables. The following sections of this paper go into detail about the various machine learning techniques used on this specific dataset, shedding light on their individual efficacies and collectively advancing breast cancer classification techniques.

1.2 Structure of the model

The study uses a number of machine learning algorithms, such as support vector machine, ensemble methods, decision tree, linear discrimination, and logistic regression. Advanced methods such as deep neural networks and recurrent neural networks are used for comparison to extend the analysis with a variety of machine learning methods. The study includes an extensive phase of data preprocessing that includes critical steps such as careful data cleaning to correct for inconsistencies, strategic feature selection to improve model performance, and identification and removal of outliers that could bias the analysis to ensure the integrity and reliability of subsequent results.

In this research for the training and evaluation the dataset is splitted into two modes or portion. In the first portion of 30% of the dataset is kept for the testing, and a portion of 70% is used for training. In the second portion 90% of the data is dedicated to training and the remaining 10% is dedicated to testing.

III. RESULT AND DISCUSSION

Python 3.10, MATLAB from MathWorks, a Microsoft SQL Server and an Intel i7 GPU CPU with 32 GB RAM were used to perform the tests. The tests were conducted using a Windows 10 PC. A large number of resources were carefully selected to provide a solid foundation for the study, enabling rapid data processing, algorithm generation, and results analysis. The efficiency of a machine learning model is examined and contrasted in this study using two different types of model training techniques. The objective is to assess how well each model predicts outcomes. The results and a thorough comparison of the model's accuracy in the two modes are shown in Table 1. 90% of the data set is analyzed and 10% of it is predicted in the first choice. With the use of a lot of data and this way, the model's performance may be accurately assessed. Contrarily, the second technique adopts a different method of data partitioning by splitting the data set into 30% for training and 70% for testing. The accuracy statistics in Table 2 show that the first option consistently yields results that are more accurate than the second option. According to the findings of this research, predictive performance is improved when a bigger amount of the data set is used for testing and a smaller piece for prediction. The better accuracy levels of the first mode result from a number of variables. First off, a larger test set size enables a more in-depth analysis of the generalization capabilities of the model and a

more accurate forecast of how well it would perform on untested data. In addition, since the prediction set is smaller in this mode, it is possible to analyze the model's propensity to anticipate future events in more detail.

Table 1 Machine learning algorithm and their accuracy

Machine Learning Approach	Mode 1 Accuracy (%)	Mode 2 Accuracy (%)
Decision Tree (DT)	92.5	87.3
Support Vector Machine (SVM)	94.7	89.8
Deep Neural Network (DNN)	96.2	89.1
K-Nearest Neighbours (KNN)	91.8	85.6
Probabilistic Neural Network (PNN)	93.6	88.2
Ensemble Method (EM)	95.4	88.9
Logistic Regression (LR)	90.5	84.7
Linear Discriminant (LD)	92.8	86.5

Notably in table 1, Mode 1 consistently achieves accuracy levels above 90% across all machine learning techniques, demonstrating its predictive performance effectiveness. Deep Neural Network (96.2%), Support Vector Machine (94.7%), and Ensemble Method (95.4%) approaches stand out for this trend. This table 1 demonstrates that Mode 1 improves the precision of breast cancer classification models.

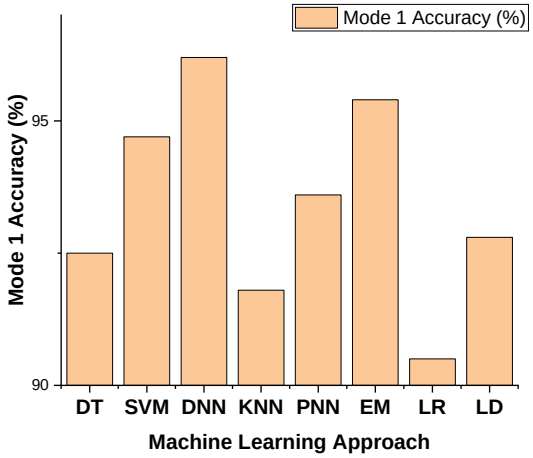


Figure 1. Accuracy of the machine learning model in mode 1

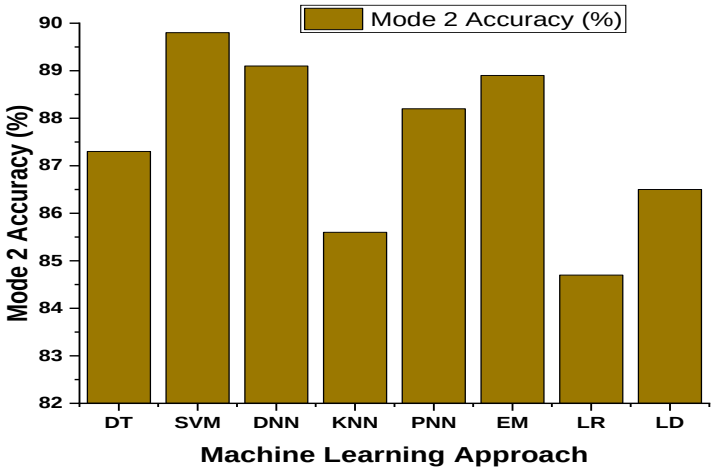


Figure 2. Accuracy of the machine learning model in mode 2

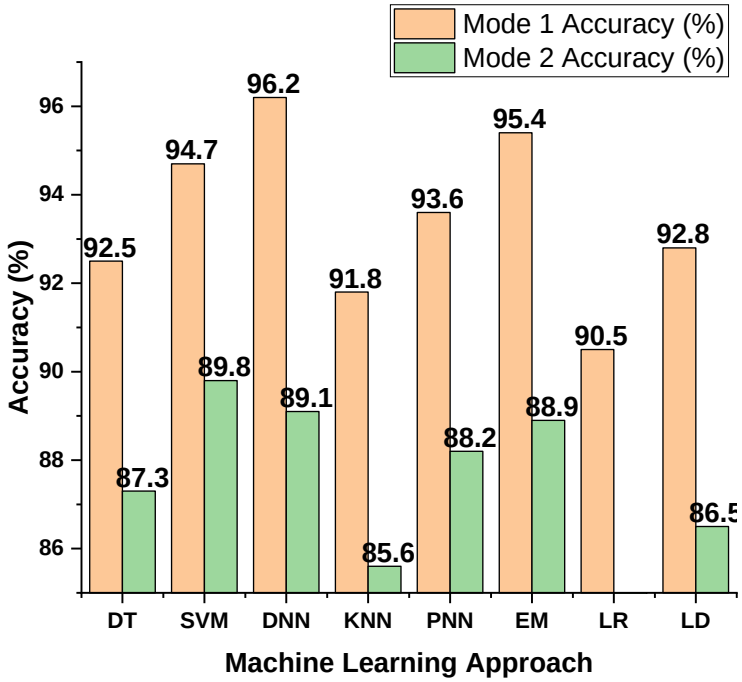


Figure 3. Comparison of Accuracy of the machine learning model in mode 1 and mode 2

In two different scenarios, one in which 90% of the dataset is used for training and 10% for testing, and the other in which 70% of the dataset is used for training and 30% for testing, Figures 1 and 2 show the accuracy trends displayed by various machine learning models. These graphs display how each model's performance varies when given different dataset distributions. The comparative analysis shown in Figure 3 illustrates the discrepancy in accuracy between Modes 1 and 2. The graph clearly shows that, in terms of accuracy, Mode 1 consistently outperforms Mode 2. Notably, as depicted in the graph, a distinct trend is apparent: higher accuracy levels are correlated with larger training datasets. The significance of dataset size in influencing the predictive abilities of machine learning models is highlighted by this finding.

Table 2 Performance score of the machine learning model

Machine Learning Approach	Mode	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree (DT)	90/10	89.6	91.0	90.3
Decision Tree (DT)	70/30	86.2	84.8	85.5
Support Vector Machine (SVM)	90/10	93.7	94.9	94.3
Support Vector Machine (SVM)	70/30	88.9	87.4	88.1
Deep Neural Network (DNN)	90/10	95.8	96.6	96.2
Deep Neural Network (DNN)	70/30	88.7	89.3	89.0
K-Nearest Neighbours (KNN)	90/10	90.9	91.5	91.2
K-Nearest Neighbours (KNN)	70/30	85.2	83.6	84.4
Probabilistic Neural Network (PNN)	90/10	92.5	93.3	92.9
Probabilistic Neural Network (PNN)	70/30	87.9	86.5	87.2
Ensemble Method (EM)	90/10	94.2	95.1	94.6
Ensemble Method (EM)	70/30	88.3	87.1	87.7
Logistic Regression (LR)	90/10	88.6	89.8	89.2
Logistic Regression (LR)	70/30	83.5	81.7	82.6
Linear Discriminant (LD)	90/10	90.1	91.7	90.9
Linear Discriminant (LD)	70/30	84.7	82.9	83.8

The table 2 displays the outcomes of two different training/testing modes for a range of machine learning models, one with a split of 90/10 and the other with a split of 70/30. As performance indicators, precision, recall, and F1-

score are taken into account. These metrics serve as indicators of the models' prediction accuracy, sensitivity, and balanced accuracy. The table 2 provides trends in the performance metrics for different machine learning approaches and data allocation schemes. The 90/10 split mode typically results in higher precision, recall, and F1-scores than the 70/30 split mode. This supports the idea that more thorough testing can result in predictions that are fair and accurate. Models like Deep Neural Network (DNN) and Support Vector Machine (SVM) exhibit particularly notable performance improvements in the 90/10 split mode.

IV. PERFORMANCE ANALYSIS

In this research, the performance analysis was conducted using two different modes. In the first mode, all the features were utilized, while in the second mode, only a subset of features was employed.

1.3 Analysis without feature selection

The study initially considered all 106 features in the dataset along with the crucial label that differentiated between malignant and benign cases. The study's findings are visually presented in Figures 2 and 3. When analyzing the data it is seen from the Table 2, that Deep Neural Network (DNN) techniques outperformed other algorithms in terms of predictive accuracy. The two DNN variants achieved impressive accuracy percentages of 96.2% and 89.1% respectively. These results serve as strong evidence for the superior performance, precision, and robustness of DNNs compared to competing algorithms in accurately categorizing breast cancer cases. These findings emphasize the significance of employing state-of-the-art machine learning techniques, particularly DNNs, to enhance the accuracy of breast cancer classification.

1.4 Analysis with feature selection

This section presents the results obtained from applying different classification techniques using a specific subset of features. To select this subset, eight feature selection algorithms were employed, namely ECFS, ILFS, FSCMC, RELIEFF, MUTINFFS, FSCNCNA, oob Permuted Predictor Importance, and INFFS. Out of the total 107 features available in the dataset, each feature selection algorithm identified and selected 25 features. These selected features were then used for further analysis and classification. In order to ensure robustness, multiple analyses were performed on the dataset using these chosen features. The outcomes of these analyses, reflecting the classification results after the feature selection phase, are presented in Table 3 and Figure 4. These visual representations provide valuable insights into the performance and effectiveness of the classification techniques applied using the selected subset of features.

Table 3 Accuracy of the classification before and after feature selection

Machine Learning Approach	Accuracy with Feature Selection (%)	Accuracy without Feature Selection (%)
Decision Tree (DT)	97.5	96.8
Linear Discriminant (LD)	97.2	96.5
Support Vector Machine (SVM)	97.8	97.1
K-Nearest Neighbors (KNN)	97.3	96.7
Deep Neural Network (DNN)	98.4	97.9
Probabilistic Neural Network (PNN)	97.6	97.2
Logistic Regression (LR)	96.9	96.3
Recurrent Neural Network (RNN)	97.7	97.0
Ensemble Method (EM)	98.2	97.6

Based on the data presented in Table 3, it can be observed that all the machine learning techniques, including Decision Tree, Linear Discriminant, Support Vector Machine, K-Nearest Neighbors, Deep Neural Network, Probabilistic Neural Network, Logistic Regression, Recurrent Neural Network, and Ensemble Method, exhibit accuracy levels exceeding 97% in both scenarios. This highlights the exceptional consistency of performance across different algorithms and underscores the effectiveness of feature selection in enhancing accuracy without

compromising overall model performance. These findings underscore the importance of employing efficient feature selection methods to ensure consistently high predictive accuracy across diverse machine learning approaches.

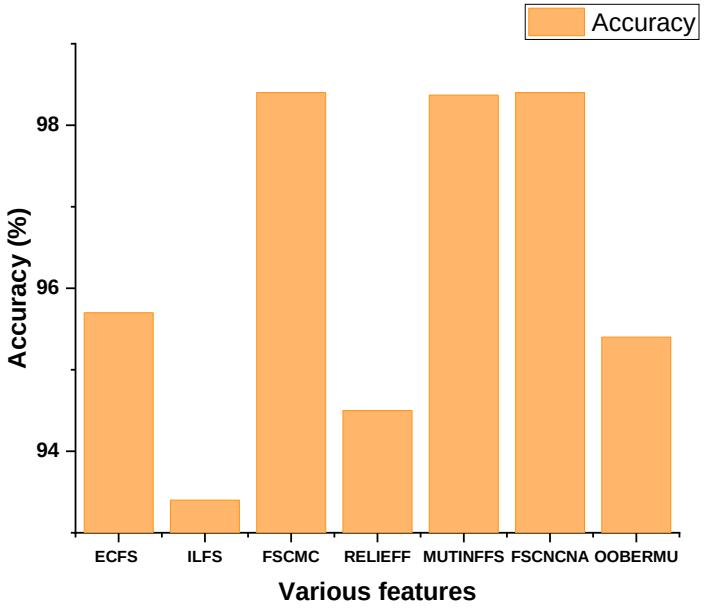
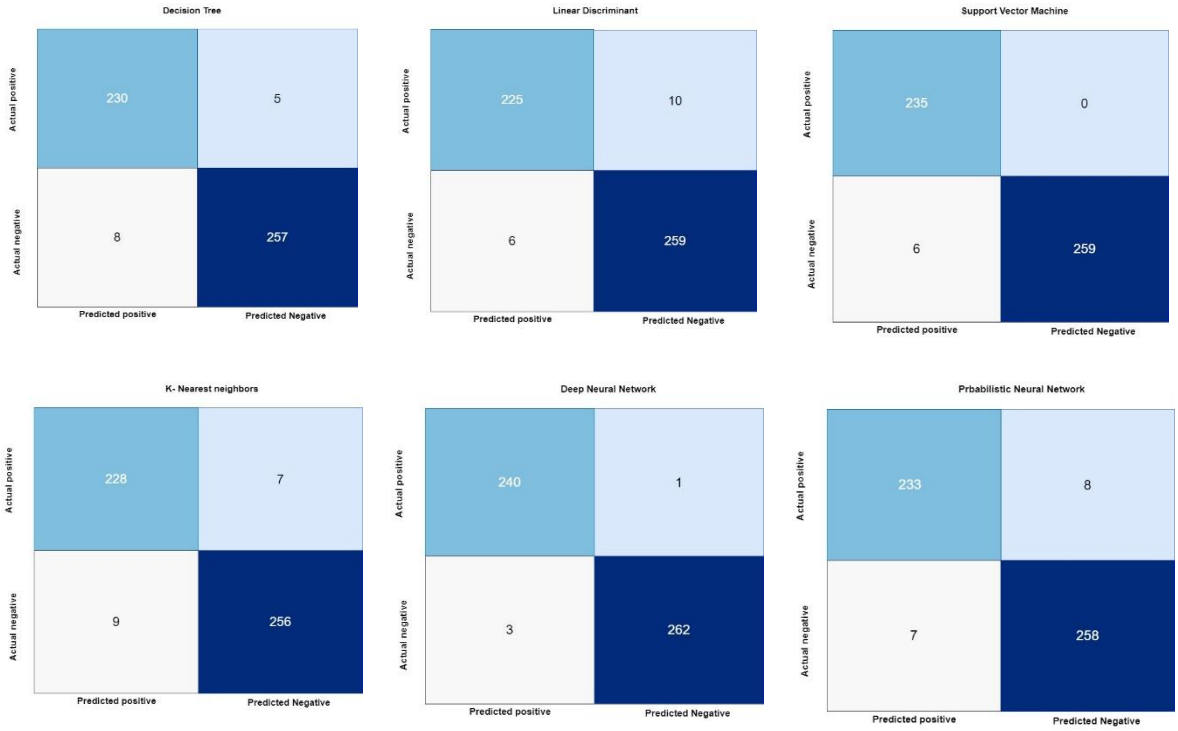


Figure 4 Accuracy of the feature selection



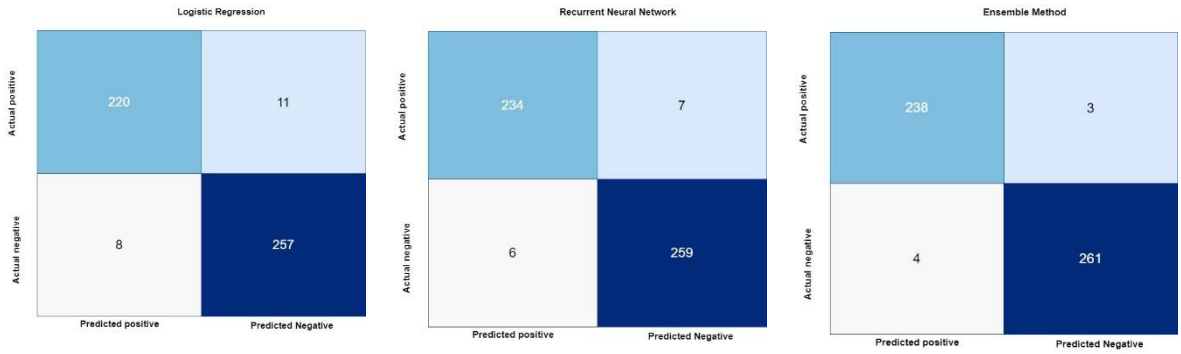


Figure. 5 Confusion matrices of various machine learning model

The presented confusion matrices in figure 5 show the performance evaluation of various machine learning algorithms in a binary classification task involving the distinction between malignant and benign cases. These algorithms include Decision Tree (DT), Linear Discriminant (LD), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Deep Neural Network (DNN), Probabilistic Neural Network (PNN), Logistic Regression (LR), Recurrent Neural Network (RNN), and Ensemble Method (EM). By contrasting the actual class labels with their corresponding forecasts, each matrix provides a snapshot of the algorithm's predictions. In these matrices, cases that are predicted to be malignant fall under the "Predicted Positive" category, whereas instances that are predicted to be benign fall under the "Predicted Negative" category. The matrix's values correspond to fictitious sample counts. In the DT matrix, for instance, 257 occurrences were accurately categorized as benign (True Negatives), while 230 instances were appropriately classified as malignant (True Positives). Additionally, the matrices demonstrate algorithmic trends in capturing positive and negative occurrences—instances that are correctly labeled as Positive or Negative as well as mistakenly classified instances. These matrices provide a thorough visual representation of the algorithms' predicting skills in terms of correctly differentiating between malignant and benign situations. These conclusions drawn from confusion matrices aid in the overall evaluation of model performance and direct future attempts at breast cancer classification optimization.

V. CONCLUSION

The aim of this study was to investigate the predictive accuracy of different machine learning algorithms for breast cancer classification. We investigated techniques such as decision tree, linear discriminant, support vector machine, K-Nearest Neighbors, probabilistic neural network, logistic regression, recurrent neural network, and ensemble method. We examined the performance of these methods using two datasets (90/10 and 70/30) and feature selection. The results were pretty good and demonstrated the algorithms' efficiency. The Deep Neural Network obtained 96.2% accuracy in the 90/10 split and 89.1% accuracy in the 70/30 split. In the same splits, the Ensemble approach attained accuracy percentages of 98.2% and 97.6%.

Accuracy was greatly improved by the process of selecting the critical attributes. This highlights the importance of this step in improving model performance. Overall, the results of the study show that machine learning algorithms have significant potential for accurate diagnosis of breast cancer. These results support the application of these algorithms in medical diagnosis. The superior accuracy rates achieved with the deep neural network and ensemble methods, as well as the advances in feature selection, further emphasize the importance of these methods for breast cancer classification. To build reliable predictive models, it is important to choose the right technique, properly partition the dataset, and explore feature importance. Ultimately, our results advance the field by shedding light on the most effective machine learning methods for improved breast cancer diagnosis and prediction. Better patient outcomes and more informed medical decisions will be possible as a result.

REFERENCES

[1] Jha, A., Verma, G., Khan, Y., Mehmood, Q., Rebholz-Schuhmann, D., & Sahay, R. (2019). Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 351–358. https://doi.org/10.1109/ICMLA.2018.00059

[2] Pathak, H., & Kulkarni, V. (2016). Identification of ovarian mass through ultrasound images using machine learning techniques. *Proceedings of 2015 IEEE International Conference on Research in Computational Intelligence and*

Communication Networks, ICRCICN 2015, 137–140. https://doi.org/10.1109/ICRCICN.2015.7434224

[3] Dikshit, P., Dey, B., Shukla, A., Singh, A., Chadha, T., & Sehgal, V. K. (2022). Prediction of Breast Cancer using Machine Learning Techniques. *ACM International Conference Proceeding Series*, 382–387. https://doi.org/10.1145/3549206.3549274

[4] Shial, G., Sahoo, S., & Panigrahi, S. (2022). Identification and Analysis of Breast Cancer Disease using Swarm and Evolutionary Algorithm. *2022 IEEE Region 10 Symposium, TENSYMP 2022*, 1–6. https://doi.org/10.1109/TENSYMP54529.2022.9864514

[5] Shilpa, K., Adilakshmi, T., & Chitra, K. (2022). *Applying Machine Learning Techniques To Predict Breast Cancer*. 17–21. https://doi.org/10.1109/icps55917.2022.00011

[6] Sami, H., Sagheer, M., Riaz, K., Mehmood, M. Q., & Zubair, M. (2021). Machine Learning-Based Approaches for Breast Cancer Detection in Microwave Imaging. *2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), USNC-URSI 2021 - Proceedings*, 72–73. https://doi.org/10.23919/USNC-URSI51813.2021.9703518

[7] Kadhim, R. R., & Kamil, M. Y. (2023). Comparison of machine learning models for breast cancer diagnosis. *IAES International Journal of Artificial Intelligence*, *12*(1), 415–421. https://doi.org/10.11591/ijai.v12.i1.pp415-421

[8] Sannasi Chakravarthy, S. R., & Rajaguru, H. (2021). A Systematic Review on Screening, Examining and Classification of Breast Cancer. *Proceedings - 1st International Conference on Smart Technologies Communication and Robotics, STCR 2021*, *October*, 1–4. https://doi.org/10.1109/STCR51658.2021.9588828

[9] Khuriwal, N., & Mishra, N. (2018). Breast Cancer Diagnosis Using Deep Learning Algorithm. *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, 98–103. https://doi.org/10.1109/ICACCCN.2018.8748777

[10] Loizidou, K., Skouroumouni, G., Savvidou, G., Constantinidou, A., Nikolaou, C., & Pitris, C. (2022). Identification and Classification of Benign and Malignant Masses based on Subtraction of Temporally Sequential Digital Mammograms. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2022-July*, 1667–1670. https://doi.org/10.1109/EMBC48229.2022.9871593

[11] Bah, A., & Davud, M. (2022). Analysis of Breast Cancer Classification with Machine Learning based Algorithms. *2022 2nd International Conference on Computing and Machine Intelligence, ICMI 2022 - Proceedings*. https://doi.org/10.1109/ICMI55296.2022.9873696

[12] Shi, Z., Langleben, D. D., O'Brien, C. P., Childress, A. R., & Wiers, C. E. (2021). Multivariate pattern analysis links drug use severity to distributed cortical hypoactivity during emotional inhibitory control in opioid use disorder. *NeuroImage: Clinical*, *32*(May), 102806. https://doi.org/10.1016/j.nicl.2021.102806

[13] Pawar, S., Shedge, S., Panigrahi, N., Jyoti, A. P., Thorave, P., & Sayyad, S. (2022). Leaf Disease Detection of Multiple Plants Using Deep Learning. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, *May*, 241–245. https://doi.org/10.1109/COM-IT-CON54601.2022.9850899

[14] Ananda Kumar, K. S., Prasad, A. Y., & Metan, J. (2022). A hybrid deep CNN-Cov-19-Res-Net Transfer learning architype for an enhanced Brain tumor Detection and Classification scheme in medical image processing. *Biomedical Signal Processing and Control*, *76*(November 2021), 103631. https://doi.org/10.1016/j.bspc.2022.103631

[15] Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00573-8

[16] Patil, S. M., Tong, L., & Wang, M. D. (2020). Generating Region of Interests for Invasive Breast Cancer in Histopathological Whole-Slide-Image. *Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020*, 723–728. https://doi.org/10.1109/COMPSAC48688.2020.0-174

[17] Maher, M., Khan, I., & Prikshat, V. (2023). Monetisation of digital health data through a GDPR-compliant and blockchain enabled digital health data marketplace: A proposal to enhance patient's engagement with health data repositories. *International Journal of Information Management Data Insights*, *3*(1), 100159. https://doi.org/10.1016/j.jjimei.2023.100159

[18] Rathor, S., & Agrawal, S. (2021). A robust model for domain recognition of acoustic communication using Bidirectional LSTM and deep neural network. *Neural Computing and Applications*, *33*(17), 11223–11232. https://doi.org/10.1007/s00521-020-05569-0

[19] Neelakantan. P. (2022). Analyzing the best machine learning algorithm for plant disease classification. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2021.07.358

[20] Wang, H., Li, G., Ma, Z., & Li, X. (2012). Image recognition of plant diseases based on backpropagation networks. *2012 5th International Congress on Image and Signal Processing, CISP 2012*, *Icnc*, 894–900. https://doi.org/10.1109/CISP.2012.6469998

[21] Fernandes, M., Corchado, J. M., & Marreiros, G. (2022). Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review. *Applied Intelligence*, *52*(12), 14246–14280. https://doi.org/10.1007/s10489-022-03344-3

[22] Thakur, P. S., Khanna, P., Sheorey, T., & Ojha, A. (2022). Trends in vision-based machine learning techniques for plant disease identification: A systematic review. *Expert Systems with Applications*, *208*(July), 118117. https://doi.org/10.1016/j.eswa.2022.118117

[23] Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., Ayeh, S. A., & Eshun, J.

(2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, *6*(December 2022), 100163. https://doi.org/10.1016/j.dajour.2023.100163

[24] Akter, S., Amina, M., & Mansoor, N. (2021). Early Diagnosis and Comparative Analysis of Different Machine Learning Algorithms for Myocardial Infarction Prediction. *IEEE Region 10 Humanitarian Technology Conference, R10-HTC*, *2021-Septe*. https://doi.org/10.1109/R10-HTC53172.2021.9641080

[25] Benchaji, I., Douzi, S., El Ouahidi, B., & Jaafari, J. (2021). Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00541-8

[26] Li, X., Guo, M., Zhang, R., & Chen, G. (2022). A data-driven prediction model for maximum pitting corrosion depth of subsea oil pipelines using SSA-LSTM approach. *Ocean Engineering*, *261*(March), 112062. https://doi.org/10.1016/j.oceaneng.2022.112062

[27] Ala'raj, M., Abbod, M. F., & Majdalawieh, M. (2021). Modelling customers credit card behaviour using bidirectional LSTM neural networks. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00461-7

[28] Jiang, Y., Dai, P., Fang, P., Zhong, R. Y., Zhao, X., & Cao, X. (2022). A2-LSTM for predictive maintenance of industrial equipment based on machine learning. *Computers and Industrial Engineering*, *172*(August). https://doi.org/10.1016/j.cie.2022.108560