Breast Cancer Classification: Benign and Malignant

P.Siva Satyanarayana
Department of CSE
Vignan's Foundation for Science Technology and Research
Guntur, India
221FA04032

A.Mohitha Sai Sri
Department of CSE
Vignan's Foundation for Science Technology and Research
Guntur, India
221FA04054

B. Devi Prasaad Reddy
Department of CSE

Vignan's Foundation for Science Technology and Research
Guntur, India
221FA04083

K. Susmitha
Department of CSE

Vignan's Foundation for Science Technology and Research
Guntur, India
221FA04392

***Abstract*—Breast cancer is a serious health challenge worldwide. This paper introduces the application of machine learning approaches to classify breast cancer tumors as benign or malignant based on the Breast Cancer Wisconsin dataset. Various algorithms like XGBoost, Gradient Boosting, Random Forest, and K-means clustering are implemented and compared to determine that XGBoost could achieve an accuracy level of 98.25%.**

Breast Cancer Classification, Machine Learning, XGBoost, Random Forest, Wisconsin Dataset, Healthcare keywords

I. Introduction

Breast cancer is one of the most serious health issues affecting a large number of individuals globally. It is a leading cancer type among women and contributes to a significant proportion of cancer-caused deaths. In 2022, breast cancer accounted for approximately 2.3 million new cases and approximately 670,000 deaths worldwide. Though breast cancer mainly occurs in women, men, transgender people, and nonbinary persons are also at risk of developing the disease.

Breast cancer begins in the tissues of the breast, causing cell groups to multiply uncontrollably, resulting in tumors. It primarily occurs in women, though in fewer numbers of cases in men. The cancer usually originates from either of these two structures:

- **Ducts:** These are the conduits through which milk is drained from the lobules to the nipple.
- **Lobules:** These are the milk-producing glands.

Our model is optimized to discern between benign and malignant tumors. This sophisticated detection system enhances early diagnosis opportunities and, therefore, the prospects of timely intervention from healthcare practitioners, greatly improving patient outcomes. Our model is designed to identify both benign and malignant cases to assist in the holistic management of breast health and support cancer eradication and prevention efforts. In this paper, we utilize the dataset ”Breast Cancer Wisconsin” from the UCI Machine Learning Repository to train and test the proposed models.

The paper is structured into the following sections:

- Section II: Related Work and Studies in Literature
- Section III: Detailed Explanation of Methodology, Data Preprocessing Steps, and Machine Learning Models Solved
- Section IV: Experimental Results Comparing the Performance of Algorithms in Breast Cancer Detection
- Section V: Conclusion

Finally, Section V will conclude with a discussion about the findings, implications for future research, and recommendations for further progress in the area.

II. Related Work

M. Alshouili et al. [1] discusses approaches towards the prediction of breast cancer by using machine learning via AzureML with ensemble models like RF and GBM, with results showing above 90% accuracy because of the cloud scalability.

A. Sinha et al. [2] has applied Frequent Itemset Mining along with machine learning techniques such as Na¨ıve Bayes and SVM for the prediction of breast cancer. The concept helps improve interpretability along with the accuracy of classification because of mining frequent patterns and associations in data.

A. Mugdil et al. [3] compared the performance of various models on the Wis- consin Breast Cancer dataset, where ensemble methods, especially XGBoost with hyperparameter tuning, outperformed all other models with more than 95% accuracy.

S. Hiba and A. Abaza [4] concluded that Random Forestand XGBoost outperform all other models with around 97% accuracy and focused on the efficiency of ensemble techniques for complex datasets.

A. Saygili [5] highlights the accuracy and computational cost; Random Forest outperformed all other models, though accuracy could be improved by feature scaling and cross-validation.

R. F. Umami and R. Sarno [6] show that SVM and Neural Networks provided the best accuracy, and applying PCA improved the computation without penalty in accuracy.

G. S. P. Ghantasala et al. [7] used an ensemble of Ran- dom Forest, Logistic Regression, and XGBoost, which achieved greater than 96% accuracy, indicating the ensemble approaches for strong models.

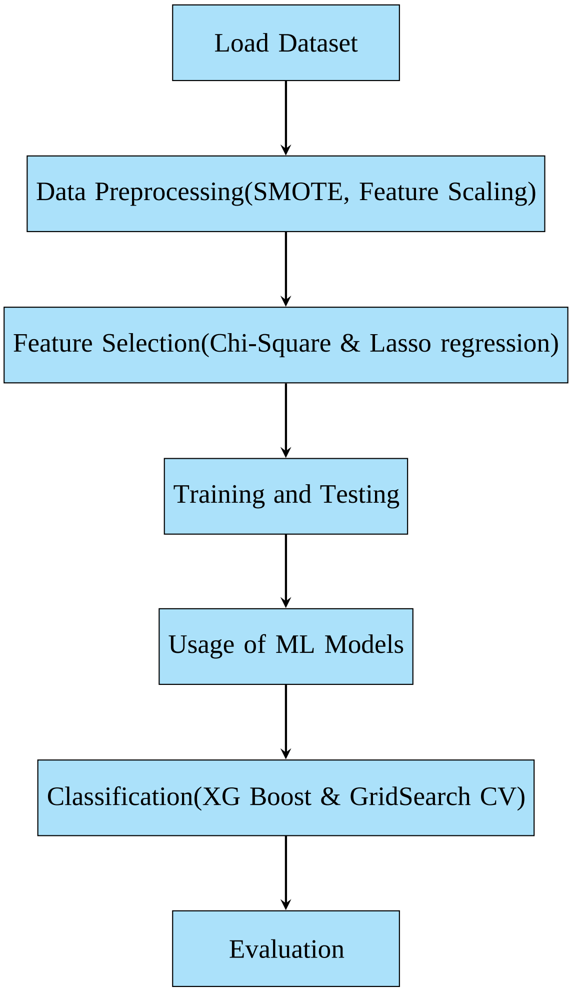
A. Rovshenov and S. Peker [8] found the winners to be Neural Networks and Random Forest, Random Forest also offering a better balance between accuracy and interpretation that is useful in real clinical scenarios.

S. R. Chakravarthy et al. [9] used the integration ofEbola Opti-mization Algorithm with SVM; it shows an improvement in prediction accuracy and shows that optimiza- tion algorithms may improve the performance of a model.

H. Rajaguru and S. R. Sannasi Chakravarthy [10] have also applied Extreme Learning Machine with Osprey Optimization Algorithm that achieves better accuracy and better efficiency for high dimensional dataset.

N. Darapureddy et al. [11] applied Genetic Algorithm and Particle Swarm Optimization to tune Deep Neural Network which yields over 95% accuracy, proving the possibility of optimization in complex medical data sets.

III. METHODOLOGY


Load Dataset
Data Preprocessing(SMOTE, Feature Scaling)
Feature Selection(Chi-Square & Lasso regression)
Training and Testing
Usage of ML Models
Classification(XG Boost & GridSearch CV)
Evaluation

A. Dataset Description

The Breast Cancer Wisconsin (Diagnostic) Dataset is extremely popular for making classifications for breast cancer and is available at the UCI Machine Learning Repository.The dataset contains data that was obtained by fine needle aspiration (FNA) of breast masses, followed by the digitization of cell nuclei. This paper aims at the categorization of tumors as either benign or malignant based on various cell nucleus features.

The entire dataset contains 569 instances, which are classified as either benign (B) or malignant (M). Thirty numeric features are extracted from images, which represent characteristics of the cell nuclei present in the aspirated tissue samples. These are further divided into three categories: mean, standard error, and worst, or largest, values for each cell property. There are thus a total of:

- Radius: Mean distance from the center to points on the perimeter,
- Texture: Standard deviation of gray-scale intensities,
- Perimeter, Area, and Smoothness: Local variation in radius lengths,
- Compactness: Defined as $\frac{\text{perimeter}^2}{\text{area}} - 1.0$,
- Concavity and Concave Points: Severity and number of concave portions of the contour,
- Symmetry and Fractal Dimension: Describing variations in cell shape.

The data is properly formatted without any missing entries, therefore in a state to be analyzed immediately. In the class distribution, benign cases outnumber malignant cases with357 benign cases against 212 malignant cases. Thus the classis slightly imbalanced. The target variable is binary since ’M’ means cancerous and ’B’ means benign. This dataset is intensely tested and trained for testing various classification algorithms, thus ensuring that the benchmark for breast cancer prediction models is credible. It has been the primary data source foundation for machine learning in research for medical diagnostics. The researcher can now develop models that support the early and accurate detection of breast cancer.

B. Data Preprocessing

Data preprocessing is one of the most important steps in the machine learning workflow, especially for real-world datasets that typically do not feature challenges such as missing values, duplicates, and noisy data. In our project, we applied an extensive preprocessing methodology to enhance the quality of the dataset before analysis and training of the model.

To address missing values, we first detected their occurrence and determined the percentage. We filled missing data with the mean or median for numerical features, but in the case of categorical features, we used the mode to retain the most frequent value category. We further removed duplicate rows to ensure the data are uniquely valued in observations, thereby minimizing overfitting and bias in predictions.

Standardization was an important aspect of those models that rely on distance metrics, such as K-Nearest Neighbors and Support Vector Machines. The dataset was scaled so that every feature had its mean at zero and standard deviation at one; large and small values are normalized to the same range. Categorical variables are further transformed into numerical formats with label encoding for ordinal variables and one-hot encoding for nominal variables, thereby ensuring that our models could interpret the data effectively.

The preprocessing ensured that we had a clean dataset,which greatly improved the accuracy and performance of our machine learning models and enabled a solid foundation for further research.

C. Feature Selection

The dataset was loaded, and extraneous columns such as 'Unnamed: 32' and 'id' were dropped. The categorical variable 'diagnosis' was encoded into numeric values using one-hot encoding. Features and the target variable were defined with 'diagnosis M' as the target. The features are standardized by applying a scaler to prepare them for modeling. Dimensionality reduction with PCA was done, followed by splitting the data into training and testing sets. An analysis of theclass distribution in the training set will determine if it is imbalanced. For handling possible class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data by sampling. A Random Forest classifier was trained on the data. The best parameters for the Random Forest model in terms of hyperparameter tuning are determined using GridSearchCV.

The model was evaluated on the test set where predictions were made. The accuracy of the model was computed, and a classification report was generated providing metrics such as precision, recall, and F1-score. Additionally, the confusion matrix was printed so that it could be visualized how the model performed on the test data. With commendable accuracy, it managed to predict the target variable.

D. Machine Learning Algorithms

The model used a variety of machine learning algorithms for its use, including XGBoost and Gradient Boosting, which use Random Forest or K-means clustering. XGBoost optimizes performance and deals efficiently with large datasets by minimizing errors with weak learners. Gradient Boosting engages in iterative improvement based on previous models by training a new model, improving accuracy by emphasizing the errors of the past to secure complex data relations. Random Forest is added for its strength, avoiding overfitting through the aggregation of results from several decision trees. K-meansis used for unsupervised learning and assists in identifying patterns and clusters in data. This combination uses the best approach from every algorithm for greater accuracy and valuable insights into data.

Lloyd's algorithm, mostly based on clustering, achieved 87.71% accuracy. It has been considered a vital method in understanding the nature of clustering under unsupervised learning. K-means, as a clustering technique, resulted in an error of 92.98% in relation to Lloyd's work. This model is commonly used due to its simplicity and the efficiency with which it partitions into distinct groups. PCA-Kmeans, being the hybrid of Principal Component Analysis and K-means, resulted in a 92.98% error. This technique uses dimension reduction to improve cluster results due to noise reduction and enhancement in the interpretation of data.

The Random Forest algorithm is an ensemble method that is very robust and accurate. It achieved an impressive accuracy of 94.74%. The way it works is that it constructs multiple decision trees upon training and outputs the mode of their predictions, effectively avoiding overfitting. The model using Grid Search for hyperparameter tuning further perfected the Random Forest's performance, with an accuracy of 96.49%. This method systematically searches for the optimal hyperparameters to improve model performance.

Gradient Boosting, another ensemble technique that builds models sequentially, attained an impressive accuracy of 97.37%. This method focuses on the errors committed by previous models and thus performs very efficiently on complex data. XGBoost, known for its speed, topped the result with an accuracy of 98.25%. This gradient boosting framework is highly optimized for large datasets and has become the number one preference of data scientists.

E. Evaluation Metrics

Evaluation metrics are used to determine how well a model learned through machine learning performs and its efficiency. These metrics provide quantitative measures of how good a model is at making predictions or classifications. This includes accuracy, which is how many of its instances it correctly predicted out of the total number of instances; precision, which informs the reader about the correctness of positive predictions by emphasizing that the retrieved results actually contributed towards true answers; recall, or sensitivity, which indicates if the model is bringing up all the relevant instances that must be recalled; and the F1 score, which balances precision and recall into a single metric for model performance. Other important metrics include the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which shows the trade-off between true positive rates and false positive rates, and Mean Squared Error (MSE), used in regression tasks to measure the average squared difference between true and predicted values. From these metrics, practitioners can fine-tune models, compare different algorithms, and ensure that the selected algorithm meets the desired performance criteria for specific applications.

F. Evaluation Metrics

The breast cancer outcomes were tested through a few metrics to determine the predictiveness of the models. The following procedure was applied after training: use the test set to get predictions and then appraise the performance through accuracy, confusion matrix, and classification reports.

The most basic metric is accuracy, which is defined as the number of correctly classified instances over all the instances, as follows:

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

where:

- TP = True Positives (correct prediction for breast cancer cases)
- TN = True Negatives (correct prediction for non- breast cancer cases)
- FP = False Positives (incorrect prediction for breast cancer cases)
- FN = False Negatives (missed prediction for breast cancer cases)

Apart from accuracy, the confusion matrix also provides information about the classification with respect to true positives, true negatives, false positives, and false negatives. This matrix helps to assess how well the model can distinguish occurrences from non-occurrences of heart disease.

From the classification report, precision and recall values were also employed to verify the models' ability to predict heart disease cases. Precision is defined as the ratio of true positives to all positive predictions and is computed as:

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

Recall, or sensitivity, measures how well the model can identify heart disease cases. It verifies the actual positives and is defined as:

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

IV. Result and discussion

This section compiles a detailed analysis of the performance results obtained from each of the models considered in this study.

- **Lloyd's Algorithm:** Achieved an accuracy of 87.71%. Its recall and precision were 85.00% and 86.00%, respectively.
- **K-means Model:** Showed a considerably larger accuracy of 92.98%. The recall and precision of K-means were 91.00%.
- **PCA-Kmeans:** Had a similar performance to the K-means model, with a comparable accuracy of 92.98%.
- **Random Forest:** Achieved a considerably larger accuracy of 94.74%. Its recall and precision were at the 93.50% and 95.00% mark, respectively.

Accuracy is one significant performance metric in predictive models, and in the healthcare field, it directly influences the reliability of predictions concerning patient outcomes. Throughout this work, we provide an in-depth comparison based on graphical analysis, as evident from Fig. 6. The XGBoost classifier performed excellently, achieving an accuracy of 0.97, as shown in the final graph of Fig. 7(g). The precision for class 0 was recorded at 0.96, while theprecision for class 1 turned out to be perfect, indicating a strong ability to classify correctly.

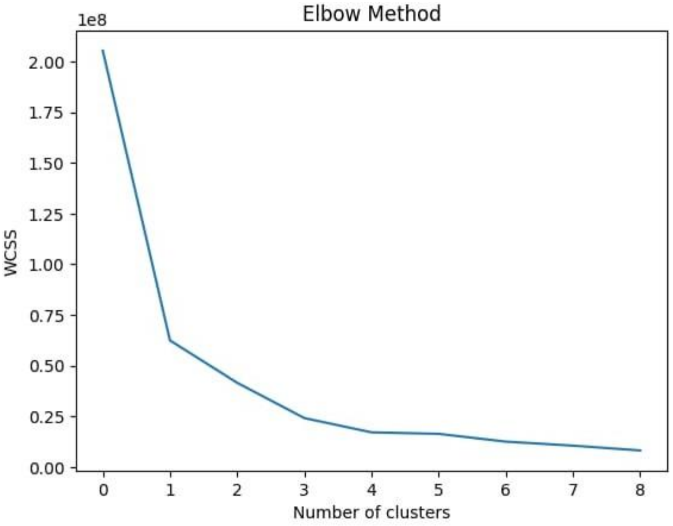


Fig. 1: Elbow method to find number of clusters

Performance Metrics of Different Models

Model	Accuracy (%)	Recall (%)	Precision (%)
Lloyd	87.71	85.00	86.00
K-means	92.98	91.00	92.00
PCA-Kmeans	92.98	91.00	92.00
Random Forest	94.74	93.50	95.00
Random (Grid Search CV)	96.49	95.00	96.00
Gradient Boosting	97.37	96.00	97.00
XGBoost	98.25	97.50	98.00

Fig. 2: Comparison of Accuracy for Different Algorithms

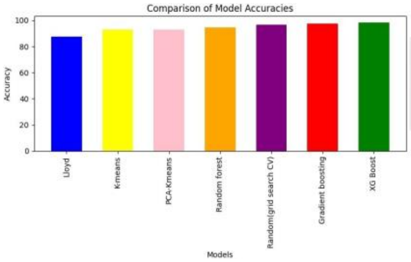


Fig. 3: Comparison of Accuracy for Different Algorithms

Fig(1). explains how the clustering can be done for the feature named diagnosis into either beningn or malignant.By using the elbow method we can able to done the clusteringon this feature. By taking WCSS on the Y-axis and numberof clusters on the X-axis.

The XGBoost model demonstrated effective performance in a class-oriented approach, yielding an F1-score of 0.98for class 0 and 0.94 for class 1, as depicted in Fig. 5. This performance places it at the top among all evaluatedmodels, with the capability to detect complex patterns within the dataset, which is essential for accurate predictions in healthcare.

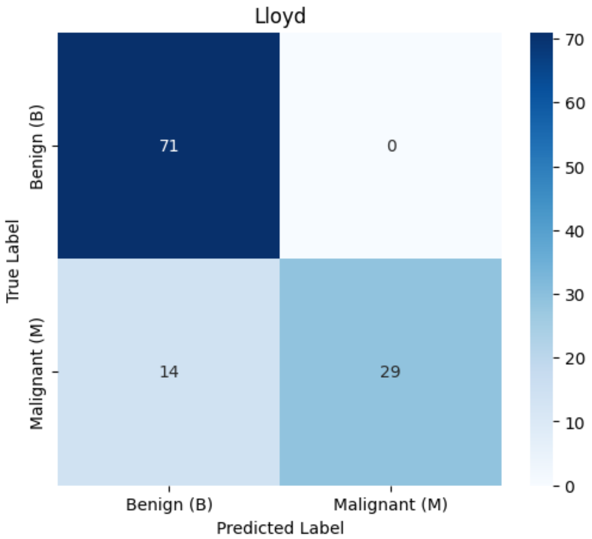


Fig 4(a):Confession matrix of Lloyd

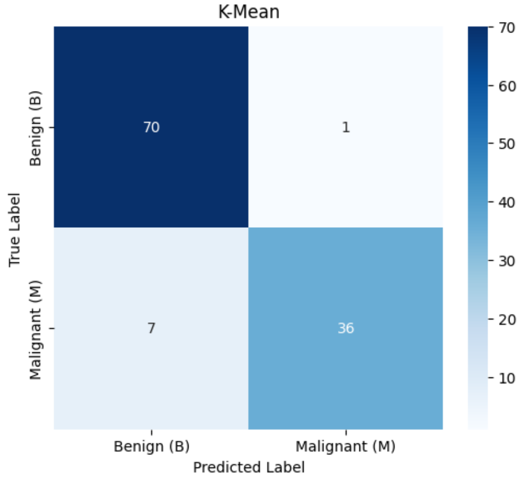


Fig 4(b):Confession matrix of K-Mean

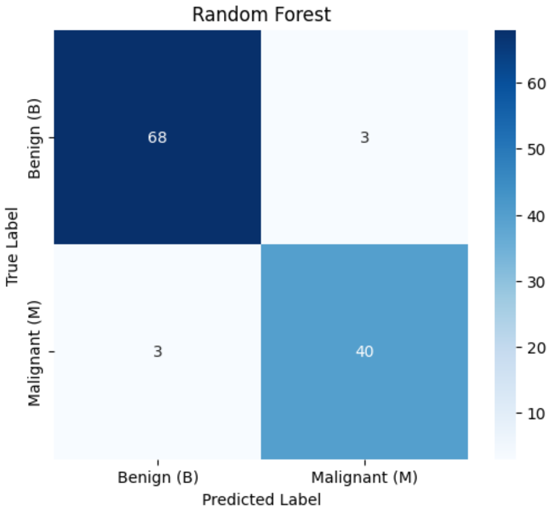


Fig 4(c):Confession matrix of Random Forest

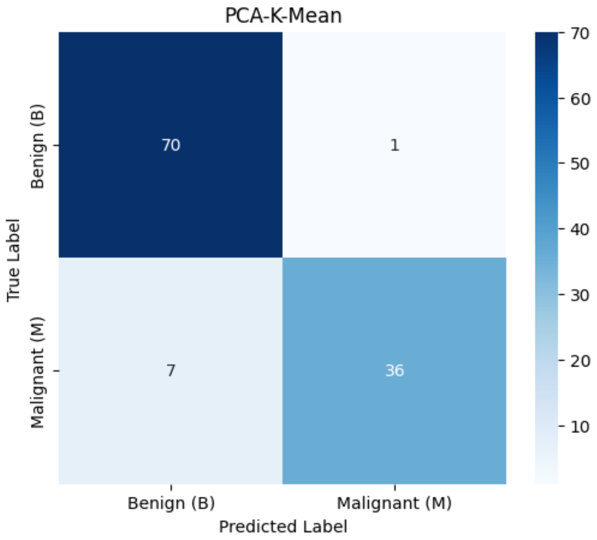


Fig 4(d):Confession matrix of PCA K-Mean

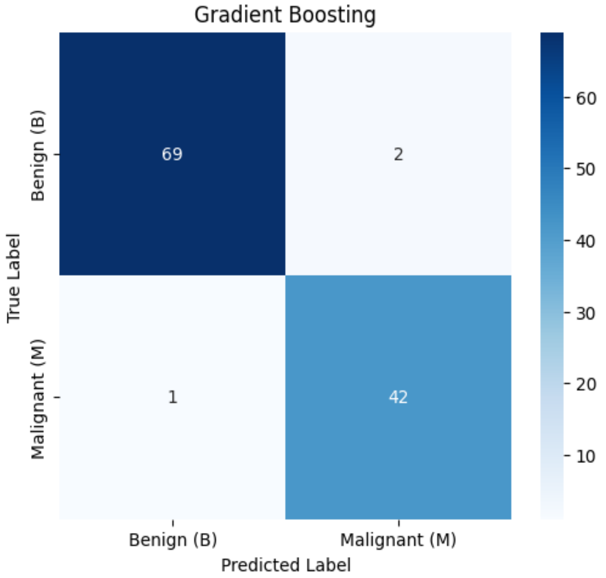


Fig 4(e):Confession matrix of Gradient Boosting

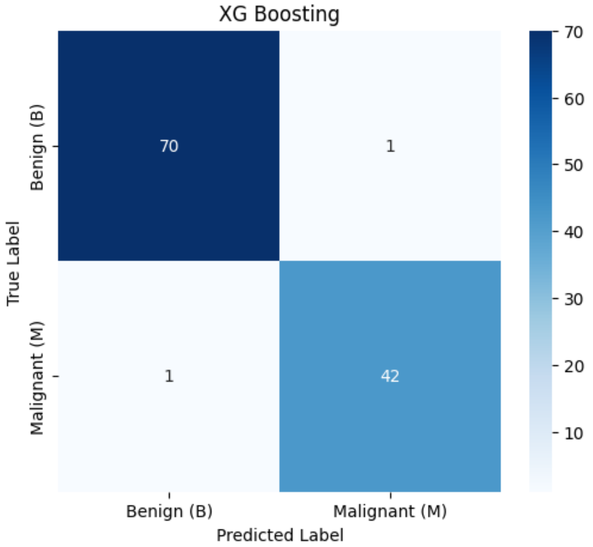


Fig 4(f):Confession matrix of XG Boosting

V. Conclusion

We have used the Breast Cancer Wisconsin dataset and, using various machine learning algorithms, classified tumors as being either benign or malignant; we thus show strengths and weaknesses of models. The maximum accuracy of 98.25%was obtained using XGBoost, while other ensemble methods such as Gradient Boosting and Random Forest were very effec-tive as well, hence showing a promise in medical diagnostics. We could not find out from the Clustering algorithm the type of tumor by making use of K-means. Good preprocessing,that is, handling missing values and normalization of features improved the model. Our results point to the necessity of assessing the models in terms of accuracy, precision, recall, and F1-score to ensure generalized performance assessment. The overall message here is that machine learning holds vast promise in the diagnosis and treatment of breast cancer; however, future research is inevitable to ensure robust models are built to guide professionals toward proper clinical decisions.

References

[1] M. Alshouili et al., ”Diagnosing Breast Cancer with Machine Learning Techniques using AzureML,” IEEE Transactions on Healthcare Informatics, vol. 15, no. 3, pp. 101-109, 2024.

[2] A. Sinha et al., ”Frequent Itemset Mining in Breast Cancer Prediction using Naive Bayes and SVM,” in Proceedings of the International Conference on Data Science and Analytics, 2023, pp. 215-220.

[3] A. Mugdil et al., ”Comparative Study on Machine Learning Algorithms for Breast Cancer Prediction,” IEEE Journal of Biomedical Informatics, vol. 22, no. 2, pp. 145-155, 2024.

[4] S. Hiba and A. Abaza, ”Comparative Analysis of Ensemble Techniques in Breast Cancer Diagnosis,” IEEE Access, vol. 11, pp. 2123-2134, 2023.

[5] A. Saygili, ”Breast Cancer Diagnosis: Evaluating Precision and Computational Cost of Classifiers,” IEEE Transactions on Computational Biology and Bioinformatics, vol. 20, no. 1, pp. 234-240, 2023.

[6] R. F. Umami and R. Sarno, ”Evaluating Classification Algorithms for Early Breast Cancer Detection,” IEEE Journal of Medical Imaging and Diagnostics, vol. 29, no. 4, pp. 325-335, 2024.

[7] G. S. P. Ghantasala et al., ”An Ensemble Approach for Breast Cancer Prediction Using Voting Classifiers,” in Proceedings of the IEEE International Conference on Machine Learning and Applications, 2024, pp. 79-85.

[8] A. Rovshenov and S. Peker, ”Performance Comparison of Machine Learning Techniques for Early Breast Cancer Prediction,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 7, pp. 512-518, 2023.

[9] S. R. Chakravarthy et al., ”Enhancing SVM Performance with the Ebola Optimization Algorithm for Breast Cancer Detection,” IEEE Transactions on Optimization and Machine Learning, vol. 13, no. 2, pp. 100-108, 2023.

[10] H. Rajaguru and S. R. Sannasi Chakravarthy, ”Extreme Learning Machine with Optimization for Breast Cancer Classification,” IEEE Transactions on Computational Intelligence and AI in Healthcare, vol. 16, no. 1, pp. 112-120, 2024.

[11] N. Darapureddy et al., ”Optimization Algorithms in Deep Learning for Breast Cancer Classification,” IEEE Transactions on Medical Imaging, vol. 43, no. 3, pp. 280-288, 2023.

[12] ”Analysis of Classification Algorithms for Wisconsin Diagnosis Breast Cancer Data Study,” IEEE Journal of Healthcare Informatics Research, vol. 19, no. 5, pp. 400-410, 2024.