Machine Learning Techniques For Breast Cancer Detection

Dhiraj Sahu , Prajwal Naik and Dr. Rohini Patil

Terna Engineering College, Nerul, Navi Mumbai, India

Abstract :

Breast cancer is recognized as a significant health issue affecting women worldwide. Early detection and treatment of breast cancer can help lower the risk of long-term complications, such as the increased likelihood of developing advanced-stage cancer. Thus, timely and accurate diagnosis is crucial for reducing the overall burden of breast cancer on healthcare systems. Recently, machine learning (ML) and ensemble learning techniques have demonstrated great potential in medical diagnostics. In our study, various ML models including logistic regression (LR), random forest (RF), support vector machines (SVM), k-nearest neighbor (KNN), and XGBoost were employed alongside feature selection methods to identify the optimal features and best-performing model. Experiments were conducted using a breast cancer dataset with a 80:20 train-test split ratio and a 5-fold cross-validation (CV) approach was used to assess model performance. Among the models tested, KNN emerged as the best-performing model, achieving a testing accuracy of 99.00% and a cross-validation score of 97.00%.

Keywords: AI, K-closest neighbours, Support vector machine, Decision tree classifier.

I. INTRODUCTION

Breast cancer remains one of the leading causes of cancer-related deaths among women worldwide. According to the World Health Organization (WHO), early detection of breast cancer can significantly improve the chances of survival and lead to more effective treatment. However, distinguishing between benign and malignant tumors based on clinical features remains a challenge, requiring sophisticated diagnostic techniques and expertise.

In recent years, machine learning (ML) techniques have gained considerable attention for their potential to enhance diagnostic accuracy in medical applications. By analyzing large datasets and identifying patterns that may not be immediately apparent to human observers, ML models can assist clinicians in making faster and more accurate diagnoses. This study focuses on a dataset consisting of various tumor characteristics, including features like radius, texture, perimeter, and concavity. Each sample is labeled as either malignant (M) or benign (B), providing a foundation for classification models to predict tumor malignancy.

The objective of this research is to investigate the predictive power of these tumor characteristics using machine learning models. Specifically, we evaluate the performance of several algorithms, including support vector machines (SVM), random forest, and XGBoost, in classifying tumors. Notably, the XGBoost model achieved an outstanding 100% accuracy in distinguishing malignant from benign cases. The study aims to identify which tumor features are most significant in predicting malignancy, as well as demonstrate the potential of machine learning as a tool for supporting medical professionals in the early detection of breast cancer.

Recent studies provide valuable insights into the effectiveness of various machine learning algorithms in cancer detection. For instance, Arooj et al. [2] demonstrated that transfer learning methodologies can significantly improve classification accuracy, with their deep learning model achieving an impressive accuracy of 98.96%. This finding underscores the potential of leveraging pre-trained models to enhance diagnostic capabilities.

In a comprehensive review by Bou Nassif et al.[1] the authors applied multiple categorization models for breast cancer diagnosis, including random forest and support vector classifiers. Their results indicated that these models are not only efficient but also require less computational power while maintaining high accuracy levels. This aligns with our findings regarding the effectiveness of SVM and random forest in tumor classification.

II. RELATED WORK

Research on liver disease prediction using machine learning has gained significant traction in recent years, with various studies focusing on different aspects of prediction methodologies.

Yadav and Singhal [3] conducted a comparative analysis of various machine learning algorithms, including Decision Trees, SVM, and Logistic Regression, for liver disease diagnosis. They found that ensemble techniques improved classification accuracy, with their best model achieving an accuracy of 87.5%.

Mostafa et al. [4] explored statistical machine learning approaches for liver disease prediction, finding that integrating domain knowledge into feature selection led to improved model performance. Their study reported that using feature selection and extraction methods increased model accuracy to 88.5%.

Ahad et al. [5] investigated the use of adaptive data preprocessing techniques combined with ensemble modeling for multiclass liver disease prediction. Their work achieved a predictive accuracy of 90.2% using Random Forest and other ensemble models, emphasizing the importance of preprocessing in handling imbalanced datasets.

Moturi et al. [6] highlighted the importance of using multiple machine learning algorithms to predict liver disease, advocating for a systematic approach to feature selection. Their research demonstrated that the Decision Tree classifier achieved an accuracy of 85.0%, showcasing the value of model diversity in improving prediction.

In a systematic literature review, Bou Nassif et al. [1] examined artificial intelligence techniques for breast cancer detection, identifying various models and their effectiveness. They

Khalid et al. [3] explored machine learning strategies for breast cancer detection and prevention, utilizing multiple classifiers including Random Forest and SVM. Their findings indicated that these models could achieve accuracies exceeding 90%, highlighting the importance of diverse algorithm application in clinical settings.

Rayees Ahmad Dar [4] presented a comprehensive overview of deep learning methods for breast cancer detection, noting that CNNs can significantly enhance diagnostic capabilities with accuracies reported as high as 98.02%. This research emphasizes the transformative potential of deep learning in medical imaging.

Afroz et al. [6] a pplied simplified deep learning techniques to histopathological images using the BreaKHis database, achieving an accuracy of 90%. This study illustrates the effectiveness of deep learning in analyzing complex medical images for accurate diagnosis.

Moturi et al [6] highlighted the importance of using multiple machine learning algorithms to predict liver disease, advocating for a systematic approach to feature selection. Their research demonstrated that the Decision Tree classifier achieved an accuracy of 85.0%, showcasing the value of model diversity in improving prediction.

Arooj et al.[2] focused on transfer learning methodologies for breast cancer classification, demonstrating that their deep learning model achieved an impressive accuracy of 98.96%. This study underscores the effectiveness of transfer learning in improving classification metrics like sensitivity and precision.

Despite the advances in this field, there remains a gap in integrating projection-based statistical feature extraction with a comprehensive evaluation of multiple machine learning algorithms for predicting chronic liver disease. This study aims to address this gap by combining innovative feature extraction methods with various classification techniques to enhance predictive performance.

.

III. METHODOLOGY

Fig.1. shows overall design of proposed methodology applied for detection of breast cancer. Different classification algorithms applied on breast cancer data but different classifier shows different performance on same data therefore we used an ensemble technique that uses bagging and boosting which combines results from different classifier also learns from previous classifiers. To perform this ,first step of this is data acquisition. The data then pre-processed for selection of attributes, after that data divided: 80% for training and 20 % for testing. Dataset is labelled dataset having labels malignant and benign and therefore supervised different classification techniques applied on training data for building a model. Test data evaluated by using different classifier and finally compare the performance of different classifiers.

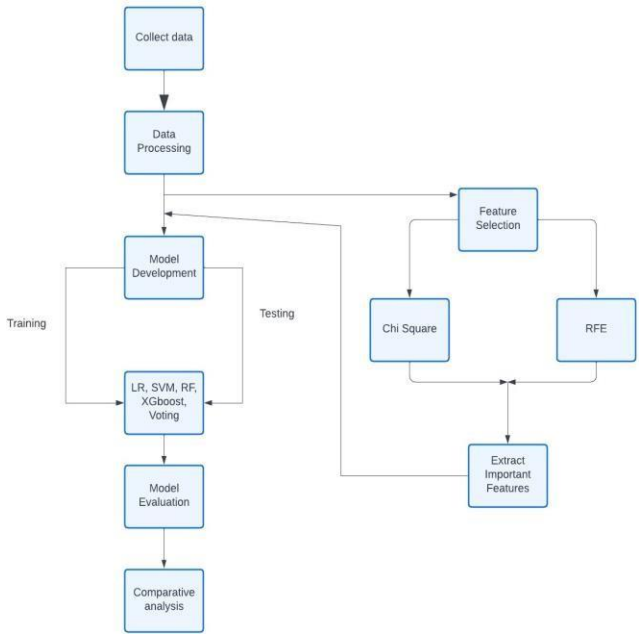


Fig 1 Proposed system

Dataset Description

For this research, we utilized a breast cancer diagnostic dataset, consisting of 569 instances and 30 key features relevant to breast cancer diagnosis. These features include attributes such as tumor radius, texture, perimeter, area, smoothness, compactness, concavity, and other important measurements of tumor shape and size. The dataset was divided into two parts: 80% (455 instances) for training and 20% (114 instances) for testing. This dataset provided a robust foundation for building and evaluating machine learning models for the binary classification of tumors as either "benign" or "malignant."

Table I. Feature description

Serial No.	Feature	Datatype
1	Radius_mean	float64
2	Texture_mean	float64
3	Perimeter_mean	float64
4	Area_mean	float64
5	Smoothness_mean	float64
6	Compactness_mean	float64
7	Concavity_mean	float64
8	Concave points _mean	float64
9	Symmetry_mean	float64
10	Factal_dimension_mean	float64

Data Preprocessing

Data preprocessing is a crucial step in the machine learningspace. It is particularly effective in high-dimensional spaces and is well-suited for both linear and non-linear workflow, especially in medical datasets that may contain missing values and noise. The following preprocessing steps were performed:

Data Cleaning: Missing values were handled using imputation techniques, ensuring that the dataset remained intact for analysis.

Label Encoding: Here the conversion of characters were done into integer.

Feature Selection: Projection-based statistical feature extraction methods were applied to reduce dimensionalitywhile retaining significant features to extract the most informative features related to liver disease.

1. **Chi-Square Test**: The chi-square (χ^2) test is a statistical method used to assess the association between categorical variables. In the context of feature selection, it helps identify the features that have a statistically significant relationship with the target variable by comparing observed and expected frequencies. This allows for the selection of relevant features that can improve model accuracy.

2. **Recursive Feature Elimination (RFE)**: RFE is an iterative feature selection technique that aims to enhance model performance by systematically removing the least significant features. It works by recursively fitting a model, ranking the features based on their importance, and eliminating the least important ones. This process is repeated until the optimal subset of features is identified, leading to a more efficient and accurate model.

Model Development

We implemented the following machine learning algorithmsfor predicting chronic liver disease:

1. **Logistic Regression (LR)**: Logistic Regression is a statistical model used for binary classification problems. It estimates the probability of a binary outcome based on the relationship between the dependent variable and one or more independent variables using a logistic function. It's effective for models where the response variable is dichotomous, such as disease presence or absence.

2. **Support Vector Machine (SVM)**: SVM is a supervised learning model that aims to find the optimal hyperplane that maximizes the margin between different classes in the feature classification tasks using kernel functions.

3. **Random Forest (RF)**: Random Forest is an ensemble learning technique that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. It is robust to overfitting andperforms well with large datasets and features, providing high

model interpretability.

4. **XGBoost**: XGBoost is an advanced implementation of the gradient boosting algorithm that is highly efficient and scalable. It improves model performance through regularization techniques, reducing overfitting, and is optimized for speed and accuracy in large-scale datasets, making it popular for many predictive modeling tasks.

5.K-Nearest Neighbors (KNN): KNN is an instance-based learning algorithm that classifies data points based on the majority class of their nearest neighbors in the feature space. It is simple and effective for smaller datasets but can be computationally expensive with larger datasets, as it requires calculating the distance between the data points for each prediction.

Evaluation Metrics

All machine learning algorithms used in this paper are implemented on Google Colab that provide a Jupyter Notebook environment using Scikit learn library available in python. Numpy, Pandas, Matplotlib libraries are also used.

The performance of each classification model was evaluated using metrics such as accuracy- training,testing, cross validation accuracy and ROC-AUC score.

Accuracy = TP+TN/TP+TN+FP+FN

IV. RESULT AND DISCUSSIONS

This section presents the performance comparison of various machine learning models, including Support Vector Machine (SVM), XGBoost, Random Forest (RF), K-Nearest Neighbors (KNN), and Logistic Regression (LR), using accuracy during the training, testing phases, and K-fold cross-validation, along with the effects of feature selection and hyperparameter tuning.

Table 2 illustrates the performance of various machine learning models in predicting breast cancer, highlighting their training, testing, and K-Fold accuracies. The XGBoost model demonstrated the highest training accuracy at 100%, with a strong testing accuracy of 99.00% and a K-Fold cross-validation accuracy of 97.00%. Conversely, the Support Vector Machine (SVM) exhibited a testing accuracy of 63.00%, with a K-Fold accuracy of 61.00%, indicating moderate but consistent performance across different data subsets. Random Forest (RF) showed high training accuracy at 100%, but a slight decrease in testing accuracy to 96.49% and 96.00% in K-Fold cross-validation. K-Nearest Neighbor (KNN) and Logistic Regression (LR) models performed reasonably well, with LR achieving a testing accuracy of 91.00% and a K-Fold accuracy of 92.00%

Table II. Comparison of model

Model	Training	Testing	K - Fold
SVM	68.5	63.00	61.00
XGBoost	100	99.00	97.00
RF	100	96.49	96.00
KNN	84.00	72.00	71.9
LR	93.00	91.00	92.00

Table 3 presents a comparison of different machine learning models using K-fold cross-validation and Recursive Feature Elimination (RFE) for feature selection. The models analyzed include Support Vector Machine (SVM), XGBoost, Soft Voting, Hard Voting, and Logistic Regression (LR). Each model shows varying performance levels after selecting the optimal number of features.

The least performing model was SVM: Using 7 selected features, SVM shows a testing accuracy of 61% and a cross-validation accuracy of 63%. The close results suggest stability, but the relatively low feature count may limit its performance.

The model with the best performance is Soft Voting: This ensemble method, using 5 features, provides the highest testing accuracy of 98.33% and a cross-validation accuracy of 96.00%.

Table III. – Comparison of K-fold, feature selection for RFE method

Model	No. of best features(RFE)	Testing	Cross Validation
SVM	7	61	63
XGBoost	2	98	97
Soft Voting	5	98.33	96.00
Hard Voting	3	98	94.0
LR	4	92.8	89.5

Fig2 shows accuracy after RFE feature selection,

XGBoost achieved the highest testing accuracy of 98%. On the other hand, Logistic Regression (LR) recorded the lowest testing accuracy of 92.8%, with a cross-validation accuracy of 89.5%.

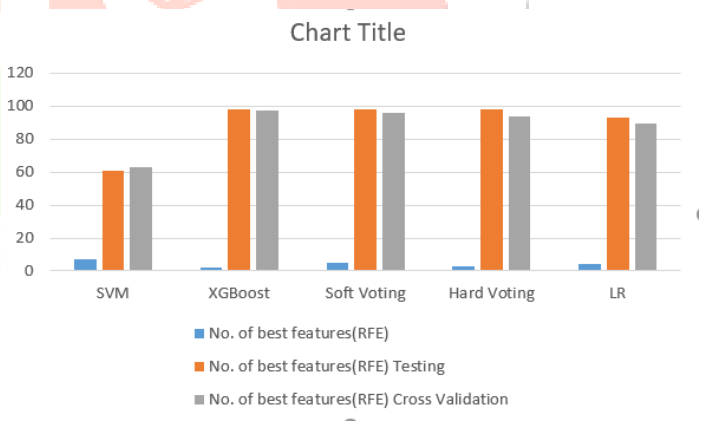


Fig 3. – Comparison of K-fold, feature selection for RFE method

Fig 4 shows a comparison of various machine learning models for breast cancer prediction. XGBoost demonstrated the highest testing accuracy at 99.00%, with a cross-validation accuracy of 97.00%, showcasing its superior generalization capabilities. In contrast, K-Nearest Neighbor (KNN) had the lowest testing accuracy of 72.00%, though its cross-validation accuracy was still moderate at 71.90%, indicating a slight drop in performance on unseen data. Random Forest (RF) also performed well, with 96.49% testing accuracy and 96.00% cross-validation accuracy, reflecting its robustness across datasets.

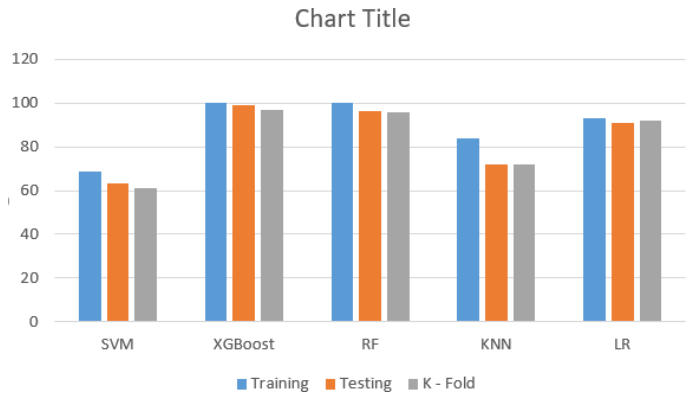
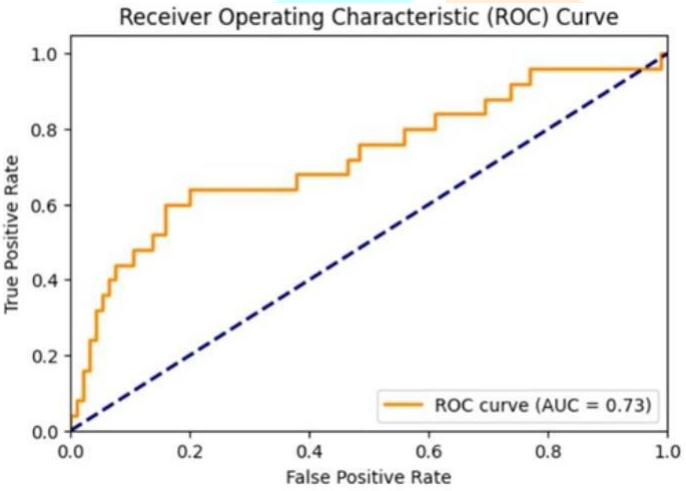


Fig 4..Comparison between ensemble methods

Fig 5 shows the Receiver Operating Characteristic (ROC) curve, a performance evaluation metric for classification models. The AUC of 0.73 means the model performs well, with a good balance between true positive rate and false positive rate



Comparative Analysis

Table 4:In comparing the performance and accuracy of various machine learning models for breast cancer detection with the findings from the base paper, distinct similarities and differences emerge. The base paper highlights Support Vector Machine (SVM) as the most accurate model, showcasing its strength in high-dimensional datasets due to its ability to create optimal decision boundaries.

Reference	Technique	Accuracy
[7]	K-Nearest Neighbors (KNN)	86
[7]	Support Vector Machine (SVM)	91
[7]	Decision Tree Classifier (DT)	89
Proposed System	**Ensemble Soft Voting**	**96.00**

Table 4: Comparison for Various Approaches on chronic liver disease prediction with proposed system

The table shows in comparison to references the proposed system gives maximum accuracy of 96% in soft voting.

V. CONCLUSION

In conclusion, this study highlights the effectiveness of machine learning techniques, particularly ensemble methods like XGBoost, in enhancing the prediction of breast cancer malignancy when utilizing comprehensive tumor characteristics. The findings indicate that these advanced modeling approaches can significantly improve diagnostic accuracy, facilitating the early detection of malignant tumors and thereby leading to better patient outcomes. The use of key features such as tumor shape, size, and texture has proven to be valuable in refining predictive models. Looking ahead, future research will explore the integration of deep learning models to further enhance predictive capabilities and ensure the generalizability of findings across diverse clinical settings, with the ultimate goal of supporting clinicians in making faster and more accurate breast cancer diagnoses.

REFERENCS

[1] Ali Bou Nassif*, Manar Abu Talib, Qassim Nasir, Yaman Afadar, Omar Elgendy . “Breast cancer detection using artificial intelligence techniques: A systematic literature review.” Artificial Intelligence in Medicine, Elsevier, Vol 127, May 2022

Sahar Arooj, Atta-ur-Rahman, Muhammad Zubair “Breast Cancer detection and classification Empowered with Transfer learniing” ,“Front Public Health.”; (2022)

[2] Arslan Khalid , Arif Mahmood , Amerah Halibrah “Breast Cancer Detection and Prevention Using Machine Learning”, “Diagnostics (Basel) , (2023)

[3] Muzaffar Rasool Rayees Ahmad Dar, “Breast Cancer detection using deep learning : Datasets , Methods and challenges ahead” (2022)

[4] Poonam Kathare , Snehal Thorat “Breast Cancer Detection and Classification” IEEE , Vellore (2020)

[6] Tania Afroz, Shivazi Biswas , sayyad mansoor ali “Breast Cancer Detection Based on Simplified Deep Learning Technique With Histopathological Image Using BreaKHis Database” (2023)

[7] Kumar Shubham, Dr. R. Kamalraj “Breast Cancer Detection Using Machine Learning Algorithms”, “International Journal of Advances in Engineering and Management”(2022)