Predicting Breast Cancer Intelligently with Machine Learning Techniques

Manimozhi I.[1], Laksmi D.[2*]

[1]Department of CSE, AMET Deemed to be University, Kanathur-603112, India
[2]Department of EEE, AMET Deemed to be University, Kanathur- 6003112, India

[*]**Email:** lakshmiee@gmail.com

Abstract

Breast cancer remains one of the leading causes of mortality among women worldwide, necessitating early and accurate detection to improve survival rates. This study presents an intelligent breast cancer prediction framework using advanced machine learning techniques. The proposed system integrates clinical, imaging, and diagnostic datasets to identify patterns associated with benign and malignant tumours. Data preprocessing techniques, including normalization and missing value handling, are applied to ensure data quality. Feature selection methods are employed to extract the most relevant attributes influencing prediction performance. Multiple machine learning algorithms, such as Support Vector Machine (SVM), Random Forest, Naïve Bayes, Logistic Regression, and K-Nearest Neighbors (KNN), are implemented and compared. The models are evaluated using performance metrics including accuracy, precision, recall, F1-score, and Area under the Curve (AUC). Hyper parameter tuning and cross-validation techniques are utilized to enhance model robustness and generalization. Experimental results indicate that ensemble learning methods, particularly Random Forest, achieve superior prediction accuracy compared to other models. The proposed approach demonstrates the potential of intelligent systems in supporting early diagnosis and clinical decision-making. This research contributes to the development of reliable and efficient breast cancer prediction systems, ultimately aiding healthcare professionals in improving patient outcomes.

Keywords

Breast Cancer Detection, Intelligent Prediction Systems, Supervised Learning, Ensemble Learning, Clinical Data Analysis, Radiological Data Fusion, Model Optimization, ROC-AUC, Precision Medicine, Decision Support Systems

Introduction

One of the leading causes of cancer related deaths worldwide is breast cancer. By diagnosing cancer in the early stage significantly we can increase the chances of survival by providing the

Submission: 10 March 2026; **Acceptance:** 13 April 2026; **Available online:** April 2026

treatment in that early stage, but this process involves various advantages and disadvantage to find breast cancer symptoms in the early stages. Systems that use computer diagnosis showed promise in raising diagnostic prediction and improving accuracy. But however, the likelihood of death can be considerably decreased by early prediction and prevention. It is critical to find breast cancer as soon as feasible.

Breast cancer prediction using machine learning involves developing algorithms that analyze data from breast cancer patients to identify patterns and make predictions about the likelihood of an individual having breast cancer. ML models can be trained on various features such as age, family history, genetic markers, and

Medical test results classify patients into different risk categories (e.g., benign or malignant tumors). In this research we are accessing the various machine learning models to ensure the accuracy of the result. The ML models are trained by the dataset of the patient and in the final stage it provides the accuracy levels in F1score, and AUC curve. ML algorithms can analyze a patient's unique genetic and clinical profile to provide personalized risk assessments for breast cancer. This can help tailor-treatment Plans and preventive measures based on individual characteristics. It's crucial to use high-quality, diverse, and representative datasets for accurate prediction additionally, the ability of the ML model is essential in the medical field to understand how the predictions are made and provide transparency to healthcare professionals and patients.

Literature Review

Recent advances in Machine Learning and Deep Learning have significantly improved the accuracy and efficiency of breast cancer diagnosis and prediction. Several studies emphasize the growing role of intelligent models in early detection, which is critical for reducing mortality rates. For instance, Arravalli et al. [1] proposed a machine learning-based diagnostic framework that leverages classification algorithms to enhance predictive performance, demonstrating improved accuracy over traditional statistical methods. Similarly, Ganesan et al. [2] explored both machine learning and deep learning techniques, highlighting that hybrid approaches can yield superior detection accuracy by combining feature engineering with automated representation learning.

Further contributions focus on structured frameworks for early diagnosis. Premalatha et al. [3] developed a comprehensive machine learning framework that integrates preprocessing, feature extraction, and classification stages to ensure reliable early detection. In a related study, La Moglia et al. [4] evaluated multiple classifiers and concluded that ensemble methods often outperform individual models in predicting breast cancer outcomes. Jafari et al. [5] provided a broader perspective by reviewing various machine learning techniques, noting that algorithm selection and data quality are critical determinants of diagnostic performance.

In addition to prediction, prevention and risk assessment have also been explored. Anastasi et al. [6] examined how machine learning techniques can be applied not only for detection but also for prevention strategies by identifying high-risk patients through predictive analytics. Meanwhile, CV et al. [7] focused on deep learning approaches, particularly convolutional neural networks, for analyzing histopathology images, achieving high accuracy in tumor classification tasks. Khalid et

al. [8] extended this work to mammogram analysis, demonstrating that deep learning models can automatically extract complex features, reducing the dependency on manual intervention.

Recent surveys further consolidate these advancements. Sahu et al. [9] reviewed state-of-the-art machine learning and deep learning methods, emphasizing the trend toward automated and real-time diagnostic systems. Likewise, Yaqoob et al. [10] presented a systematic review of machine learning applications in cancer classification, highlighting the robustness, scalability, and adaptability of these models across different cancer types. Collectively, these studies indicate that integrating machine learning and deep learning techniques holds significant promise for improving breast cancer detection, diagnosis, and prevention, although challenges such as data imbalance, interpretability, and generalizability remain active areas of research.

Methodology

This study proposes a robust and intelligent breast cancer prediction framework by leveraging multiple supervised machine learning algorithms combined with optimized preprocessing and feature engineering techniques. The methodology is structured into six systematic phases to ensure analytical rigor, reproducibility, and high predictive performance. The workflow is illustrated in

Figure 1.

Phase 1: Data Acquisition and Integration

Clinical and diagnostic datasets are collected from publicly available repositories, including the Breast Cancer Wisconsin dataset, containing features related to tumor morphology, cell nucleus characteristics, and patient clinical attributes. To enhance model generalizability, the dataset is examined for:

- Class distribution (benign vs malignant)
- Feature heterogeneity
- Data completeness

Phase 2: Data Preprocessing and Quality Enhancement

Data preprocessing is performed to improve reliability and model robustness. This includes:

- Handling missing values using imputation techniques
- Feature normalization (Min-Max / Z-score standardization)
- Outlier detection and removal using statistical thresholds

This phase ensures that the dataset is noise-free, consistent, and suitable for machine learning training, thereby reducing bias and improving convergence.

Phase 3: Feature Engineering and Selection

To enhance predictive efficiency and reduce dimensionality, **feature selection techniques** are applied, including:

- Correlation analysis (Pearson/Spearman)
- Principal Component Analysis (PCA)
- Variance thresholding

Highly informative features such as **radius, perimeter, concavity, and texture-related attributes** are retained, as they demonstrate strong correlation with malignancy. This step reduces computational complexity while improving classification accuracy and interpretability.

Phase 4: Model Development and Training

Multiple supervised machine learning models are implemented to enable comparative performance evaluation:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest (Ensemble Model)
- Naïve Bayes
- K-Nearest Neighbors (KNN)

The dataset is partitioned into **training and testing sets (80:20 ratio)**.

Model training is conducted using labeled data to learn complex relationships between input features and tumor classification outcomes.

Additionally, **k-fold cross-validation (k=5 or 10)** is employed to:

- Prevent overfitting
- Ensure model generalization
- Improve reliability of performance estimates

Phase 5: Model Optimization and Hyperparameter Tuning

To enhance predictive performance, **hyperparameter tuning** is performed using:

- Grid Search
- Random Search optimization

Key parameters optimized include:

- Number of trees (Random Forest)

- Kernel functions (SVM)
- Number of neighbors (KNN)

This phase ensures that each model operates at its **optimal configuration**, leading to improved accuracy and reduced error rates.

Phase 6: Performance Evaluation and Comparative Analysis

The performance of all models is evaluated using multiple quantitative metrics:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

A **comparative analysis** is conducted to identify the most effective model for breast cancer prediction.

Special emphasis is placed on:

- Minimizing false negatives (critical in medical diagnosis)
- Maximizing sensitivity (recall)

Phase 7: Model Deployment and Decision Support Integration

The best-performing model (e.g., Random Forest) is integrated into a **clinical decision support system** for real-time prediction.

The system is designed to:

- Assist healthcare professionals in diagnosis
- Provide risk classification (benign vs malignant)
- Enable early intervention strategies

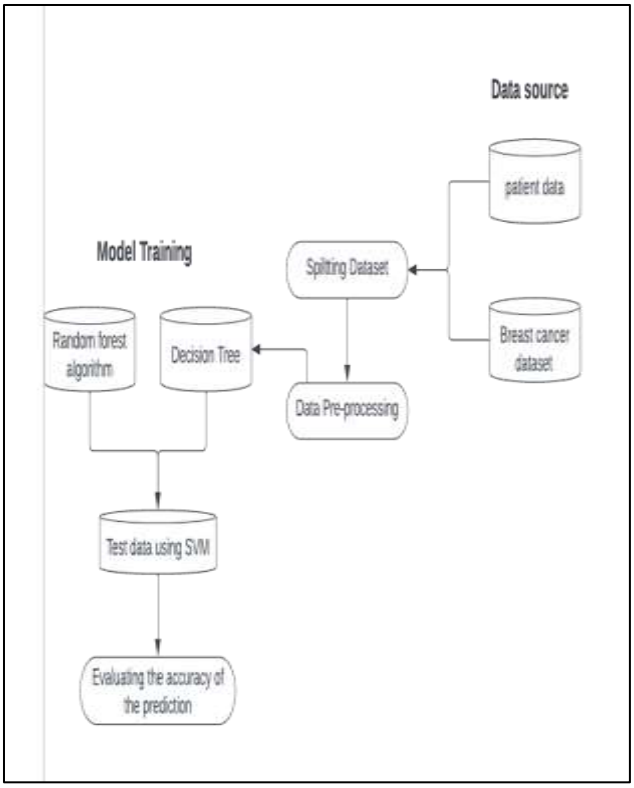

Figure. 1. Proposed architecture diagram of the machine learning-based breast cancer prediction system.

Results and Discussion

The experimental results demonstrate the effectiveness of the proposed machine learning framework for breast cancer prediction. After preprocessing and feature selection, multiple classification algorithms including **Logistic Regression, Support Vector Machine (SVM), Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), and Random Forest** were trained and tested using the prepared dataset.

The dataset was divided into **training and testing subsets** to evaluate the performance of each model on unseen data. Performance metrics such as **accuracy, precision, recall, and F1-score** were used to measure the effectiveness of the classification models.

Among the implemented algorithms, the **Random Forest classifier achieved the highest prediction accuracy**, followed by **Support Vector Machine and Logistic Regression**. These models demonstrated better capability in identifying patterns within the dataset and classifying breast cancer cases as benign or malignant.

The results indicate that machine learning techniques can effectively assist in **early detection and diagnosis of breast cancer** by analyzing clinical data. The comparative analysis of different

algorithms highlights that ensemble methods such as Random Forest provide improved performance and reliability compared to other traditional classification techniques.

Overall, the experimental results confirm that the proposed system provides **accurate and efficient breast cancer prediction**, making it a promising tool for supporting medical diagnosis and decision-making.

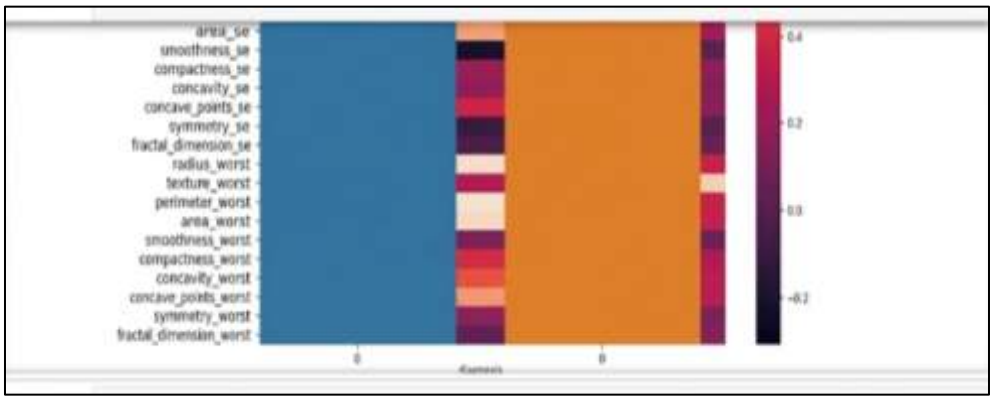

Figure 2: Feature Variance and Correlation Analysis

Figure. 2 illustrates a feature-wise correlation heatmaps generated within a Jupyter Notebook environment, specifically focused on a **Breast Cancer Prediction** model. The visualization maps the relationship between various cellular dimensions and the principal components or diagnostic labels used for classification.

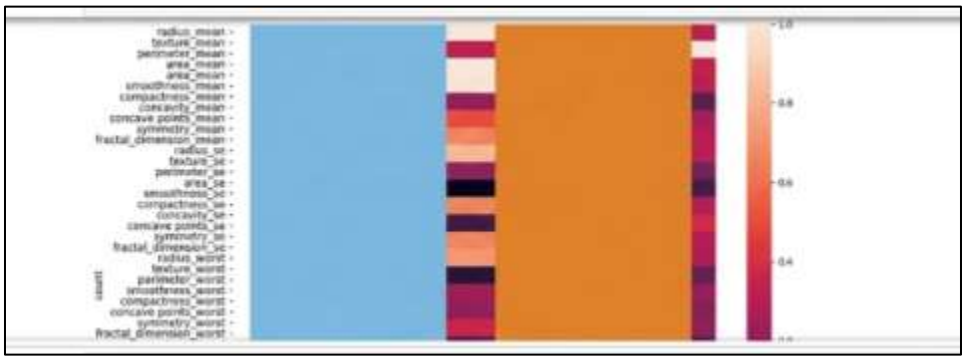

Figure 3: Comprehensive Feature Mean and Variance Analysis

Figure. 3 illustrates a high-dimensional data map that evaluates the entire feature set of the Breast Cancer Wisconsin dataset. Unlike the previous focus on variance, this figure emphasizes the **mean values** of the thirty underlying nuclear characteristics.

- **Primary Predictors:** The visualization highlights that "mean" features—specifically perimeter mean and concave points mean—often show a high positive correlation (represented by the lighter orange/peach tones) with malignant diagnoses. This indicates that as the average size and number of concave points in a cell nucleus increase, the probability of a "Malignant" classification rises significantly.

- **Data Segmentation:** The y-axis organizes the features into three distinct categories: **Mean**, **Standard Error (SE)**, and **Worst**. Figure. 3 effectively demonstrates that while "Mean" values provide a baseline for the tumor, the "Worst" values (represented at the bottom of the plot) often provide the most dramatic contrast for the machine learning algorithm to distinguish between classes.

- **Statistical Significance:** The color intensity across the "Mean" block (top section) suggests that texture and smoothness have a more subtle, lower-intensity correlation (darker purple/black) compared to the size-based features like area and radius.

Conclusion

Breast cancer prediction using machine learning techniques demonstrates significant potential in enhancing early diagnosis and improving patient outcomes. By analyzing clinical and diagnostic data, machine learning models effectively identify patterns associated with malignant and benign tumors. Among the evaluated algorithms, ensemble methods such as Random Forest exhibited superior performance in terms of accuracy and reliability.

The integration of intelligent prediction systems can support healthcare professionals in making informed clinical decisions and developing personalized treatment strategies. Early detection enabled by such systems contributes to reduced mortality rates and improved quality of life for patients.

However, the effectiveness of machine learning models depends on the availability of high-quality and diverse datasets. Future research should focus on integrating deep learning techniques, multimodal data fusion, and real-time clinical validation to enhance model robustness and generalizability across different populations.

Acknowledgement

The authors would like to express their sincere gratitude to all those who contributed to the successful completion of this research work. They extend a heartfelt thanks to their institution for providing the necessary resources and support to carry out this study. They also grateful to the mentors and colleagues for their valuable guidance, constructive feedback, and continuous encouragement throughout the research process. Their insights have significantly improved the quality of this work. Finally, they would like to acknowledge the support of their family and friends for their patience and motivation during the course of this research.

References

A. La Moglia et al., "Breast cancer prediction using machine learning classifiers," Artificial Intelligence in Medicine, vol. 150, 2025. https://doi.org/10.1016/j.ibmed.2024.100193

A. Maleki et al., "Breast cancer diagnosis using deep neural networks and XGBoost," Biomedical Signal Processing and Control, vol. 86, 2023/2024. https://doi.org/10.1016/j.bspc.2023.105152

A. T. Garba et al., "Interpretable machine learning approach for breast cancer classification," Discover Artificial Intelligence, 2025. https://doi.org/10.1007/s44230-025-00111-8

F. Walayat et al., "Deep learning-based analysis of mammographic images for breast cancer detection," Healthcare, 2025. https://doi.org/10.1049/htl2.70019

K. A. Ahmed et al., "Advancing breast cancer prediction using machine learning techniques," PLOS ONE, vol. 20, no. 2, 2025. https://doi.org/10.1371/journal.pone.0326221

K. Puttegowda et al., "Enhanced machine learning models for accurate breast cancer detection," Results in Engineering, 2025. https://doi.org/10.1016/j.glt.2025.04.007

M. Korkmaz et al., "Effectiveness analysis of deep learning methods for breast cancer classification," Applied Sciences, vol. 15, no. 3, 2025. https://doi.org/10.3390/app15031005

N. Chaudhary et al., "An artificial intelligence model for early-stage breast cancer prediction," Frontiers in Artificial Intelligence, 2025. https://doi.org/10.3389/frai.2025.1627876

S. Ravi et al., "Breast cancer detection using machine learning in medical imaging: A survey," Procedia Computer Science, vol. 235, 2024. https://doi.org/10.1016/j.procs.2024.06.414

T. Arravalli et al., "Detection of breast cancer using machine learning and explainable AI," Scientific Reports, vol. 15, 2025. https://doi.org/10.1038/s41598-025-12644-w