

Comparative Evaluation of Machine Learning Techniques for Breast Cancer Prediction and Clinical Applications

Doaa El-Bohy ^{1,*} , and Ahmed EL-Bohy ² 

¹ Assistant professor of clinical pharmacy, Department of Biopharmaceutics and Clinical Pharmacy, Faculty of Pharmacy, Al-Ahliyya Amman University, Amman, Jordan.

Emails: delbohy@gmail.com; d.elbohy@ammanu.edu.jo.

² Electrical Engineering Branch Military Technical College, Cairo, Egypt; a.bohy4phd@gmail.com.

* Correspondence: delbohy@gmail.com.

Abstract: Recent advances in machine learning (ML) and digital technologies have significantly transformed breast cancer detection. These innovations enhance diagnostic efficiency, accuracy, and accessibility. A variety of ML algorithms—such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (K-NN), Logistic Regression (LR), and Decision Tree (DT) are commonly applied to improve both early prediction and recurrence monitoring of breast cancer. By integrating data from diverse diagnostic tools, including mammography and ultrasound, ML-based systems have demonstrated improved precision and robustness, even in cases complicated by dense breast tissue or tumor heterogeneity. Artificial intelligence-driven pre-screening approaches further assist clinicians by reducing diagnostic workload and supporting earlier identification, particularly in settings with limited resources. Additionally, ultrasound-based deep learning models provide a noninvasive and widely accessible complementary technique, contributing to broader breast cancer screening practices. Emerging strategies like federated learning promote multi-institutional data sharing without compromising patient privacy, which enhances the generalizability of classifiers, for instance, Logistic Regression models achieving up to 95% accuracy. Overall, the integration of ML in breast cancer diagnosis offers not only more individualized treatment plans and improved patient outcomes but also meaningful cost reductions for healthcare systems.

Keywords: Breast Cancer; Machine Learning; Early Detection; Personalized Treatment; Artificial Intelligence.

1. Introduction

Breast cancer is still one of the most common causes of cancer-related death for women globally. So early detection is essential to improve the odds of a patient surviving the disease and decrease the burden of the disease itself. Mammography, ultrasound, and MRI are the traditional standard methods in breast cancer screening, but their limitations, including false positives, missed diagnoses, and accessibility problems in resource-limited regions, highlight the need for an advanced diagnostic approach in breast cancer detection. In solving these issues, machine learning (ML) has become a powerful approach to improve breast cancer detection and diagnosis.

Early detection is essential to improve the odds of a patient surviving the disease and to decrease its overall burden. Mammography, ultrasound, and magnetic resonance imaging (MRI) are the traditional standard methods in breast cancer screening. Still, their limitations, including false positives, missed diagnoses, and accessibility issues in resource-limited regions, highlight the need for more advanced diagnostic approaches. Addressing these challenges (ML) has become a powerful tool to improve breast cancer detection and diagnosis. Supervised learning algorithms such as Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN) have shown robust performance in assessing benign and malignant tumors based on clinical and imaging data [1].

These models can learn from large datasets, detecting patterns and features that might be overlooked by human clinicians. Algorithms that combine multiple machine learning algorithms have also been applied in recent years to improve detection rates and reduce errors through advances in image processing and data analysis. Enabling clinicians to make more appropriate decisions regarding patient management. AI-based tools can determine the risk of breast cancer based on patient history, genetic data, and imaging features, thus allowing for more personalized treatment recommendations. Models Nally, non-invasive methods, such as ultrasound supported by ML models, offer an effective option for patients with dense breast tissue, where traditional imaging practices may be less effective and mammographic sensitivity is reduced. These methods have demonstrated substantial potential for integration into clinical workflows for real-time breast cancer detection and decision-making, ranging from supervised learning algorithms to other advanced ML techniques, showing considerable progress and promise in the field [2].

2. Literature Review

The precision of diagnosing breast cancer has improved significantly with recent advances in machine learning (ML) methods. This literature review examines the development of works in machine learning detection of breast cancer, highlighting contributions made, challenges faced, and the gaps in current research. Breast cancer is essential for better survival rates and has seen extensive use of machine learning techniques. Machine learning algorithms, including Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN), have been utilized to diagnose breast cancer. In comparative studies, K-NN has a lower accuracy than SVM, demonstrating the efficacy of employing SVM to classify benign and malignant tumors in breast cancer datasets. Imbalanced datasets pose significant challenges, as benign cases often outnumber malignant cases by a large margin, leading to biased results. Reviews of ML-based models for the prediction of breast cancer recurrence have shown that models utilizing multi-source information, such as clinical history, genetics, and imaging data, perform better. While ML models can significantly improve early detection, they depend on large, diverse datasets to be truly effective, which is often a limiting factor [1].

Hybrid models that combine machine personal with MRI predictive capability have improved cancer detection accuracy by producing personalized predictions from imaging data. However, MRI-based systems are expensive and may not be readily available in low-resource settings. Federated learning is another area of research in breast cancer detection, allowing institutions to train a model with decentralized data without direct access to patient or image data. This approach can increase predictive accuracy, improve generalizability across different populations, and maintain patient data privacy. Federated learning also enables the training of predictive models without centralizing sensitive information, supporting its application across diverse healthcare systems. Although the advantages of using machine learning in detecting breast cancer are well-documented, there are several challenges. The necessity for large and diverse datasets remains a key hurdle to the adoption of these models in clinical settings. Many ML techniques are models, meaning the rationale behind their decisions is unclear to clinicians. This lack of transparency makes them less trustworthy for critical medical applications, even though their accuracy can be high. To address this gap, research is focusing on developing more interpretable models while maintaining strong predictive performance [3].

3. Methodology

This study proposes a machine learning (ML)-based breast cancer detection system that uses diverse types of data sources, machine learning algorithms, and evaluation procedures. The overall machine learning pipeline for breast cancer detection consists of various steps, including data collection, preprocessing, model training, evaluation, and optimization. The following describes the main steps in the pipeline as shown in Figure 1.

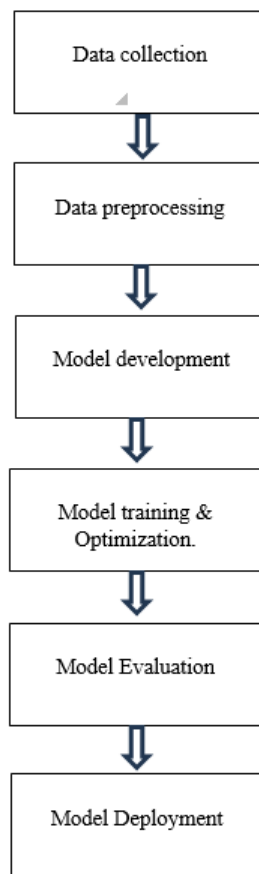


Figure 1. Machine learning pipeline for Breast Cancer detection.

3.1 Data Collection

This methodology has four main steps, and the first of them is to compile an extensive database of the relevant breast cancer patient information. The dataset will include:

- 1) **Clinical Data:** Demographic data of patients (such as age and sex), sex, family history of cancer, other underlying health conditions, genetic information (e.g., BRCA1/BRCA2 status), and lifestyle attributes (e.g., age, hormonal history).
- 2) **Histopathological Data:** The training set will include biopsy results (if available), indicating whether a tumor is benign or malignant.
- 3) **Public Datasets:** This research will utilize publicly available datasets like the Breast Cancer Wisconsin (Diagnostic) dataset [11], Digital Database for Screening Mammography, and other relevant breast cancer imaging databases [4].

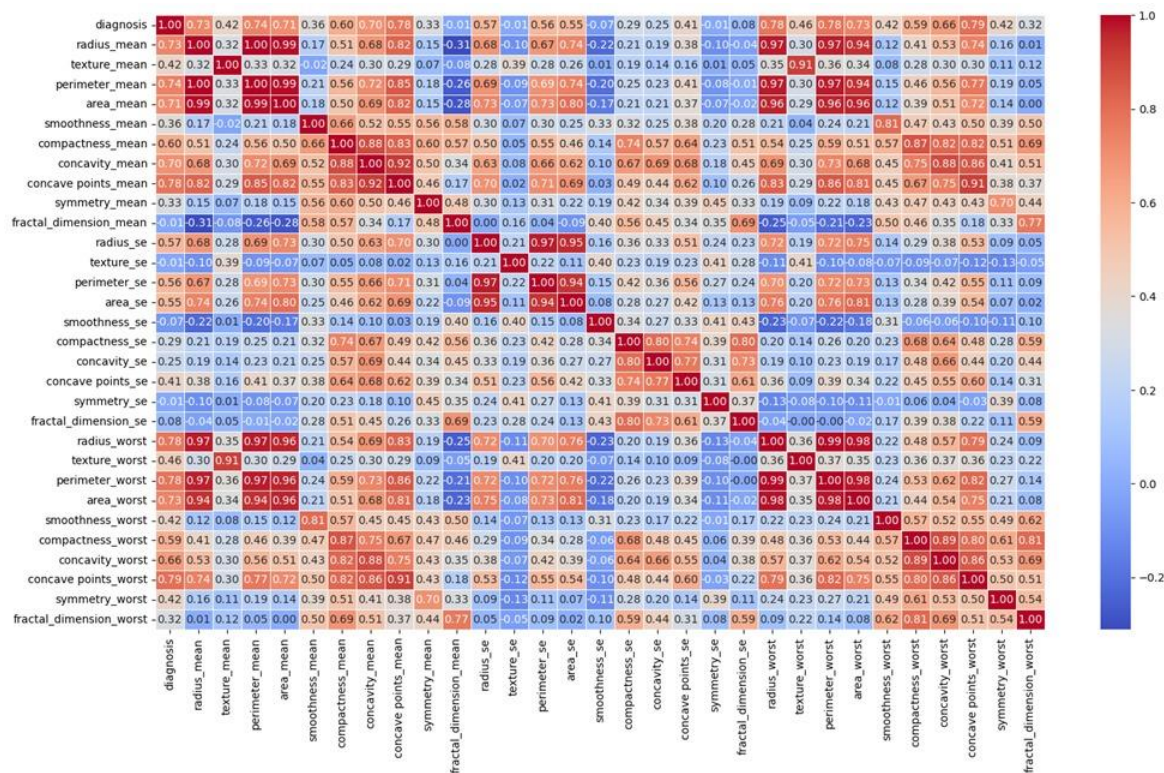


Figure 2. Heat map of characteristics of breast cancer data.

3.2 Data Preprocessing

Once the data is collected, several preprocessing steps will be undertaken to ensure the quality and consistency of the dataset:

- Handling Missing Data: Missing or incomplete data will be imputed using statistical methods or discarded based on the extent of missing values [4].

3.3 Model Development

In this stage, various machine learning algorithms will be implemented for breast cancer detection (see Tables 1 and 2). The following techniques will be explored:

- 1) Support Vector Machine (SVM): SVM will be used for classification tasks to distinguish between benign and malignant tumors based on the extracted features.
- 2) Random Forest (RF): An ensemble model that aggregates multiple decision trees to improve classification accuracy and reduce overfitting.
- 3) Logistic Regression (LR): A simple yet effective algorithm for binary classification, used to predict the likelihood of breast cancer based on clinical and imaging data.
- 4) Decision Tree (DT): versatile classification algorithm that splits data into branches based on feature values, creating a tree-like structure to predict breast cancer outcomes.
- 5) K-Nearest Neighbors (KNN): A non-parametric algorithm that predicts breast cancer diagnosis based on the most common class among the nearest data points [5].

Table 1. Comparison of machine learning algorithms.

Model	Algorithm Type	Strengths	Weaknesses
Support Vector Machine	Supervised Learning	Effective for high-dimensional data, robust to overfitting	Requires careful tuning of hyperparameters, sensitive to noisy data
Random Forest	Ensemble Learning	Handles large datasets. Well, reduces overfitting through the ensemble method	Can be computationally intensive, less interpretable than individual trees
K-Nearest Neighbors	Instance-Based Learning	Simple, easy to implement, Good for low-dimensional data	Computationally expensive for large datasets, sensitive to irrelevant feature
Logistic Regression	Supervised Learning	Simple to implement, fast, interpretable model	Struggles with complex, non-linear relationships
Decision Tree	Supervised Learning	Works well with small to medium datasets	Prone to overfitting (especially deep trees)

Table 2. Comparison between using an Algorithm in research.

Model	Original research	Main research
1	Logistic Regression	Logistic Regression
2	Random Forest	Random Forest
3	Support Vector Classifier	Support vector machine
4	Extreme Gradient Boosting	K – Nearest Neighbors
5	Extra Trees	Decision Tree
6	Cat Boost	null
7	Light Gradient Boosting Machine	null
8	Gaussian Naïve Bayes	null

3.4 Model Training and Optimization

The models will be trained using the preprocessed dataset, with training, validation, and testing splits to evaluate their performance. Key training steps include:

- Cross-Validation:** K-fold cross-validation will be used to ensure the model's robustness and avoid overfitting. This technique involves splitting the data into 'K' subsets and training the model multiple times with different validation sets.
- Hyperparameter Tuning:** Hyperparameters such as learning rate, batch size, and number of layers (for deep learning models) will be tuned using grid search or random search methods to optimize the model's performance.
- Loss Function and Optimization:** The models will employ appropriate loss functions (e.g., binary cross-entropy for classification tasks) and optimization algorithms (e.g., Adam optimizer) to minimize the error during training [6-8].

3.5 Model Evaluation

After training, the models will be evaluated with various methods based on evaluation metrics of breast cancer detection performance:

- a) Accuracy: The ratio of correctly predicted instances in the dataset, indicating the model's performance on the whole dataset.
- b) Precision and Recall: Precision indicates how many of the positively predicted cases were positive, whereas recall indicates how many of the actual positive cases were predicted correctly. These metrics create an understanding of the model's ability to detect cancer without producing too many false negatives.
- c) The F1-Score: The average weighted by the precision and recall, indicating the harmonic meaning between precision and recall, particularly useful in a case where the data is imbalanced [5].

3.6 Model Deployment

Once we have evaluated the models, the one with the best performance will be selected for deployment. Once development is complete, this model will be embedded into a web-based GUI (Graphical User Interface) designed for clinical use, which can process patient images and other clinical information inputted by the users and return live predictions to the users. The system will be developed with scalability in mind to allow accessibility across healthcare institutions, even in low-resource settings [8].

4. Results

Present the performance metrics, including accuracy, precision, recall, F1-score, and AUC (Area under the Curve) for each of the five machine learning models trained for breast cancer detection. The evaluation was conducted using various testing datasets and algorithms, namely Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (K-NN), Logistic Regression (LR), and Decision Tree (DT) (see Tables 3 and 4). This comparative analysis highlights the effectiveness of each model in identifying breast cancer cases accurately.

Table 3. Metric of performance accuracy.

Classifier	Accuracy (%)
Decision Tree	94.00
K-Nearest Neighbors	75.00
Support Vector Machine	67.00
Random Forest	95.61
Logistic Regression	95.00

Table 4. Metrics of performance (Precision and Recall and F1 Score and AUC).

Classifier	Precision (%)	Recall (%)	F1- Score (%)	AUC (%)
Logistic Regression	96	94	95	95
K-Nearest Neighbors	96	62	75	95
Random Forest	96	97	95	97
Decision Trees	94	79	86	88
Support Vector Machine	62	70	69	77

4.1 Discussion of Results

Logistic Regression achieved perfect precision (96%) with high recall (94%), resulting in a score of 95% and an AUC of 95%. This indicates that the model was able to correctly identify all predicted positive cases while maintaining a strong ability to detect actual positives.

Random Forest also delivered perfect precision (93%) and strong recall (62%), yielding a score of 96% and an AUC of 97%. These results confirm the strength of ensemble-based decision trees in capturing complex, non-linear patterns within the data.

KNN has a high precision (96%) with a recall of 62% and a score of 75%. The AUC of 95% suggests that the model maintained an elevated level of discriminative ability, although it slightly trailed behind Logistic Regression in overall ranking.

SVM demonstrated excellent performance, with a precision of 62%, a recall of 70%, and the highest F1-score (69%) among all models. Its AUC of 77% was also the highest recorded, indicating superior overall classification performance.

Decision Tree recorded the metrics among the evaluated models, with precision (94%), recall (79%), F1-score (87.10%), and an AUC of 86 %. This suggests potential overfitting and reduced generalization compared to its ensemble counterpart, Random Forest.

Overall, Logistic Regression emerged as the top-performing model in terms of balanced performance across all metrics, closely followed by SVM, Random Forest, and KNN. Decision Tree ranked lowest, highlighting the benefits of ensemble techniques and linear models over standalone decision trees in this dataset [9].

5. Conclusions

The Top Performer Logistic Regression (LR) emerged as the most effective model overall. It achieved an excellent balance between high precision (96%) and high recall (94%), resulting in a superior F1-score of 95%. This indicates its reliability in both correctly identifying positive cases and finding most of the actual positive cases. The Power of Ensemble Methods

Random Forest (RF), an ensemble model, highlighted its strength by achieving the highest Area under the Curve (AUC) at 97%. This demonstrates the advantage of using ensemble techniques, which combine multiple decision trees to capture complex data patterns more effectively and reduce the risk of overfitting compared to a single Decision Tree. The Weakness of a Single Decision Tree The Decision Tree (DT) model recorded the weakest performance among all evaluated models.

Its lower metrics across the board (F1-score 87.10%, AUC 86%) suggest it is prone to overfitting, where it learns the training data too well but fails to generalize new, unseen data. The Precision-Recall Trade-off

The results highlight a common trade-off between precision and recall. Models) had extremely high precision (93%-96%) but significantly lower recall (62%). This means that while their positive predictions were highly accurate, they missed a considerable number of actual positive cases. Logistic Regression provided the best compromise between these two crucial metrics.

Inconsistent Reporting on SVM

There is a notable contradiction in the evaluation of the Support Vector Machine (SVM). While the summary ranks it as a close second to Logistic Regression, the text also states it had the highest F1-score (69%) and "highest AUC (77%)". These figures are factually lower than those reported for Logistic Regression and Random Forest, indicating an error or inconsistency in the original report's summary [10].

Declarations

Ethics Approval and Consent to Participate

The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher. All the material is owned by the authors, and/no permissions are required.

Consent for Publication

This article does not contain any studies with human participants or animals performed by any of the authors.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare no competing interests in the research.

Funding

This research was not supported by any funding agency or institute.

References

- [1] Zhang X., Shengli S. U., & Hongchao W. A. (2019). Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. Big Data Research.
- [2] Chaurasia V., Pal S., & Tiwari B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology.
- [3] Islam T., et al. (2024). Predictive modeling for breast cancer classification. Scientific Reports.
- [4] Arravalli T., et al. (2025). Detection of breast cancer using machine learning and explainable AI. Scientific Reports.
- [5] Almarri B., et al. (2024). The BCPM method: decoding breast cancer with machine learning. BMC Medical Imaging.
- [6] Ghasemi A., et al. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. arXiv.
- [7] Rezazadeh A., Jafarian Y., & Kord A. (2022). Explainable Ensemble Machine Learning for Breast Cancer Diagnosis based on Ultrasound Image Texture Features. arXiv.
- [8] Luo L., et al. (2023). Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions. arXiv.
- [9] Shen L., Margolies L. R., Rothstein J. H., et al. (2017). Deep Learning to Improve Breast Cancer Early Detection on Screening Mammography. arXiv.
- [10] Parekh D. H., et al. (2021). Predicting Breast Cancer using Machine Learning Classifiers. Biomed Research Journal.
- [11] Dataset source Link: <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>.

Received: 03 Jan 2025, **Revised:** 05 Apr 2025,

Accepted: 13 Sep 2025, **Available online:** 17 Sep 2025.



© 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Science Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.