FUDMA Journal of Engineering and Technology
ISSN: 3092-9385
https://fjet.fudutsinma.edu.ng/index.php/fjet/
Volume 2, Issue 2, 2026, 292-302

Enhanced Breast Cancer Classification Using Ensemble Learning and Data Resampling Techniques: A Machine Learning Approach

Mohammed K. BASHIR[1*], Sanusi A. DARMA[2], Usman MAHMUD[3], Mansir ABUBAKAR[4]

[1*]Department of Computer Science and Information Technology, Al-Qalam University, Katsina, Katsina State, Nigeria
[2]Department of Computer and Information Technology, Al-Qalam University, Katsina, Katsina State, Nigeria
[3]Department of Software Engineering, Northwest University, Kano, Kano State, Nigeria
[4]Faculty of Computer Science and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

[1*]mohammedkabirbashir@gmail.com, [2]darmasanusiabu@yahoo.co, [3]usmanmahmud@auk.edu.ng, [4]mansir@uitm.edu.my

Abstract

Early and reliable breast cancer classification is essential for reducing unnecessary biopsies and ensuring timely diagnosis, particularly in low-resource settings where late detection drives high mortality. This study developed and compared an ensemble machine-learning framework against standard baseline classifiers for binary breast cancer classification using a publicly available clinical breast cancer dataset of 213 records (120 benign and 93 malignant) obtained from the Kaggle repository. To avoid information leakage, all preprocessing—median/mode imputation, Interquartile Range (IQR) outlier capping, label encoding, and z-score standardisation—and a Synthetic Minority Oversampling Technique combined with Tomek Links (SMOTE–Tomek) were fitted on the training partition only, which was balanced to 90 benign and 90 malignant cases, while an 80:20 stratified split reserved 43 cases for testing. Random Forest feature-importance analysis identified tumour size, involved lymph nodes, and metastasis as the most influential predictors, consistent with clinical knowledge. An Extreme Gradient Boosting (XGBoost) classifier achieved an accuracy of 0.81, recall of 0.74, F1-score of 0.78, and a ROC–AUC of 0.91 on the held-out test set. On this small dataset, XGBoost performed comparably to, but did not outperform, a regularised logistic regression baseline (accuracy 0.91, ROC–AUC 0.95), and McNemar's test found no statistically significant difference between the models; the wide bootstrap confidence intervals reflect the limited test-set size. The results indicate that, for small structured datasets, a carefully validated and interpretable pipeline—rather than model complexity alone—is the key requirement, and that external validation on larger, multi-institutional datasets is needed before any clinical use.

Keywords: *Breast cancer; Machine learning; XGBoost; SMOTE–Tomek; Random Forest; Classification; Diagnosis.*

1.0 Introduction

Breast cancer is a major global health priority, with approximately 2.3 million new diagnoses each year and disproportionately higher mortality in low- and middle-income countries compared with high-income regions [1]. The disease develops through complex interactions between inherited genetic predispositions, such as BRCA1/BRCA2 mutations, and environmental risk factors, resulting in diverse tumour biology that challenges conventional diagnostic paradigms [2]. Conventional screening and diagnostic methods, including mammography and biopsy, have well-documented limitations such as reduced sensitivity in dense breast tissue, inter-observer variability, and reporting delays, all of which can contribute to advanced-stage diagnoses and poorer survival [3], [4]. These limitations highlight the need for automated, data-driven systems capable of analysing complex clinical patterns to support earlier and more consistent detection.

Machine learning (ML) offers a transformative approach to addressing these challenges. Ensemble models have achieved strong performance for malignancy prediction through robust error correction and effective handling of structured clinical data [5], [8]. The application of ML to breast cancer diagnosis has evolved from early classical models—such as logistic regression, k-nearest neighbours, decision trees, and support vector machines—toward ensemble and deep-learning architectures that capture non-linear feature interactions and improve generalisation [6], [7]. Building on these advances, this study investigates an Extreme Gradient Boosting (XGBoost)-based framework combined with feature-importance analysis to improve classification accuracy, enhance interpretability, and support clinical adoption.

Despite this progress, several persistent challenges continue to limit the real-world application of ML to breast cancer classification:

Generalisation deficits – many models exhibit notable performance drops when applied to unseen or multi-ethnic datasets, reducing clinical reliability [6].

Interpretability concerns – a lack of transparency in decision-making lowers clinician trust, particularly with "black-box" models [7].

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

Class imbalance – benign cases typically outnumber malignant ones, biasing predictions toward the majority class and reducing sensitivity to malignant tumours; standard resampling such as SMOTE can partially address this but may introduce synthetic noise [9].

Limited feature utilisation – many studies focus only on static morphological attributes, and few integrate class balancing, feature selection, and ensemble learning within a single framework.

To address these gaps, this study proposes an ensemble-based diagnostic framework that integrates XGBoost classification, SMOTE–Tomek resampling, and Random Forest–driven feature-importance analysis to improve predictive accuracy, enhance interpretability, and strengthen sensitivity to malignant cases using structured clinical data [10], [11].

1.1 Aim and Objectives

The aim of this research is to develop an ensemble-based machine-learning model for breast cancer classification that improves diagnostic performance by addressing class imbalance using SMOTE–Tomek resampling and enhancing feature selection through Random Forest analysis. The specific objectives are:

1. To implement and evaluate the SMOTE–Tomek technique for balancing the dataset and improving the model's ability to detect malignant cases.
2. To identify the most relevant features for breast cancer classification using Random Forest feature-importance analysis.
3. To assess how the selected features contribute to classification performance and interpretability.
4. To develop and evaluate an XGBoost classifier, comparing its accuracy, precision, recall, F1-score, and AUC with conventional baseline classifiers (Logistic Regression and Decision Tree).

1.2 Research Questions

1. To what extent does the SMOTE–Tomek resampling technique improve the classification of malignant breast cancer cases in imbalanced datasets?
2. What are the most influential clinical features for accurate breast cancer classification?
3. How does feature-importance analysis affect the performance of breast cancer classification models?
4. How does the predictive performance of an ensemble learning model (XGBoost) compare with traditional classifiers in terms of accuracy, precision, recall, F1-score, and AUC?

1.3 Contributions

This study advances the application of machine learning to breast cancer diagnosis through the following contributions:

1. Integration of ensemble learning (XGBoost), which combines multiple decision trees to produce more accurate and generalisable predictions and is well suited to noisy structured medical data.
2. Mitigation of class imbalance with SMOTE–Tomek, improving the classifier's sensitivity to malignant cases—a critical factor in early diagnosis.
3. Feature-importance analysis using Random Forest to identify the most influential predictors, enhancing both performance and clinical interpretability.
4. A scalable, cost-effective pipeline suitable for resource-constrained healthcare settings where access to advanced diagnostic equipment is limited.

2.0 Related Works

Machine learning has become a foundational paradigm in medical diagnostics, enabling systems to learn discriminative patterns from patient data. In breast cancer detection, ML frameworks generally follow a structured pipeline of data acquisition, preprocessing, feature extraction, model training, and evaluation. Proper execution of these stages enhances diagnostic reliability, reduces observer bias, and supports evidence-based clinical decision-making [5].

This study adopts the principles of supervised learning, where algorithms are trained on labelled datasets comprising benign and malignant cases. Ensemble methods such as Random Forest [10] and Extreme Gradient Boosting (XGBoost) [11] leverage decision-tree mechanisms to capture complex feature interactions and improve generalisation. Class imbalance is addressed through resampling strategies such as SMOTE combined with Tomek links (SMOTE–Tomek) [9], which improve fairness in classification. Beyond accuracy, interpretability remains critical for clinical adoption: while tree-based models offer transparent decision boundaries, deep neural networks often function as "black boxes," limiting physician trust [7].

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

2.1 Origin and Background of Breast Cancer Detection

Historically, breast cancer detection has evolved from simple physical examinations to advanced imaging technologies. Early detection relied on clinical breast examination (CBE), in which physicians palpated breast tissue for abnormalities. The introduction of mammography revolutionised screening by enabling visualisation of internal breast structures [4]. Subsequent advances, including ultrasound, magnetic resonance imaging (MRI), and computed tomography, expanded diagnostic capacity and improved sensitivity in high-risk or complex cases [3]. In recent years, digital health technologies and ML have driven a paradigm shift toward automated, data-driven detection, offering faster diagnosis, reduced subjectivity, and improved reproducibility.

2.2 Traditional Methods of Breast Cancer Detection

Early applications of ML to breast cancer prediction primarily relied on classical algorithms such as logistic regression, k-nearest neighbours (k-NN), decision trees, and support vector machines (SVMs). Logistic regression provided interpretable models but struggled to capture the non-linear interactions inherent in biological systems. k-NN and decision trees were simple and computationally efficient but prone to overfitting and sensitive to noise. SVMs demonstrated better generalisation, achieving reported accuracies between 85% and 92% on benchmark datasets such as the UCI Wisconsin Breast Cancer dataset [6]. Conventional clinical techniques—CBE, mammography, ultrasound, MRI, and biopsy (the diagnostic gold standard)—remain central but present challenges such as invasiveness, high cost, limited accessibility in resource-constrained regions, and risks of false positives or negatives [4].

2.3 Machine Learning Approaches to Breast Cancer Detection

Recent advances in ML have enabled automated pattern recognition in breast cancer diagnostics. Decision trees and Random Forests provide flexible models capable of handling non-linear feature interactions [13]. Support Vector Machines are effective in high-dimensional feature spaces but require extensive parameter tuning [14]. Neural networks and deep-learning architectures capture complex relationships in large-scale datasets but often lack interpretability [15], [18]. Ensemble methods such as XGBoost combine multiple classifiers to enhance predictive robustness and accuracy [8]. Applied effectively, these methods improve diagnostic precision and reduce diagnostic delays, thereby improving patient outcomes.

2.4 Literature Mapping

Table 1: Summary of representative studies on breast cancer detection.

Author(s) & Year	Dataset	Technique	Performance Metrics	Key Findings	Limitations
Smith et al., 2010	Mammogram data	Mammography screening	Sensitivity: 85%, Specificity: 80%	Reduced mortality in screened populations	High false-positive rate
Lee et al., 2014	Ultrasound & MRI	Imaging-based diagnosis	Accuracy: 90%	Effective in dense breast tissue	High cost, requires specialist expertise
El-Baz et al., 2016	Digital mammograms	SVM	Accuracy: 96%	High accuracy using extracted features	Dataset-specific tuning required
Patel & Ghosh, 2018	UCI WBC dataset	Decision Tree, k-NN	Accuracy: 94%	Simplicity and interpretability	Sensitive to class imbalance
Sharma et al., 2020	Kaggle BC dataset	XGBoost + SMOTE	Accuracy: 98%	Balanced cases improved sensitivity	Computationally intensive
Kumar et al., 2021	UCI WBC dataset	RF + Feature Selection	Accuracy: 97%	High performance with interpretability	Requires large dataset for stability
Chen & Wang, 2022	CBIS-DDSM (2,620 scans)	CNN (EfficientNet-B7)	AUC: 0.985, Sensitivity: 97.2%	State-of-the-art in microcalcification detection	Requires ≥50 μm/pixel resolution
Nguyen et al., 2023	INbreast + Private MRI	Vision Transformer (ViT-L/16)	Accuracy: 98.4%, F1: 0.976	Captures long-range dependencies	Very high VRAM requirements
Zhang et al., 2023	Multi-institutional DICOM	Federated Learning + XGBoost	AUC: 0.972 (Δ +0.11)	Privacy-preserving collaboration	Needs standardised preprocessing
Gupta et al., 2024	BreakHis (7,909 images)	Graph Neural Networks	Accuracy: 99.1%, Specificity: 98.6%	Superior histopathology analysis	High training time
Kim & Park, 2025	Synthetic Cohort (n=50k)	Diffusion Models + Random Forest	Accuracy: 98.7%, Brier: 0.032	Solves data scarcity via synthesis	Needs real-world validation

2.5 Identified Research Gaps

A critical review of existing literature highlights several unresolved challenges. First, most studies rely on small, publicly available datasets, limiting generalisability. Second, class imbalance continues to affect sensitivity toward malignant cases. Third, while complex models often achieve high accuracy, they typically lack interpretability, which hinders clinician trust. Finally, few studies explore hybrid frameworks that integrate feature selection, class balancing, and ensemble learning simultaneously [20]. These gaps underscore the need for the present study, which

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

develops an ensemble learning framework that incorporates both balancing and feature selection to achieve predictive accuracy and clinical applicability.

3.0 Materials and Methods

3.1 Research Framework

The proposed framework follows a structured pipeline comprising seven stages. Figure 1 illustrates the flow of data from acquisition to evaluation. The stages are as follows:

1. Data Acquisition: obtain the breast cancer dataset from a public repository (Kaggle).
2. Data Preprocessing: handle missing values, encode categorical variables, detect and treat outliers, and scale numeric features.
3. Imbalance Handling: apply SMOTE–Tomek to balance the classes.
4. Feature Importance Analysis: use a Random Forest model to rank predictor importance for interpretability.
5. Dataset Partitioning: split the data into training and testing subsets using an 80:20 stratified split.
6. Model Training: train an Extreme Gradient Boosting (XGBoost) classifier on the training subset.
7. Evaluation: assess the model using accuracy, precision, recall, F1-score, ROC–AUC, and a confusion matrix.

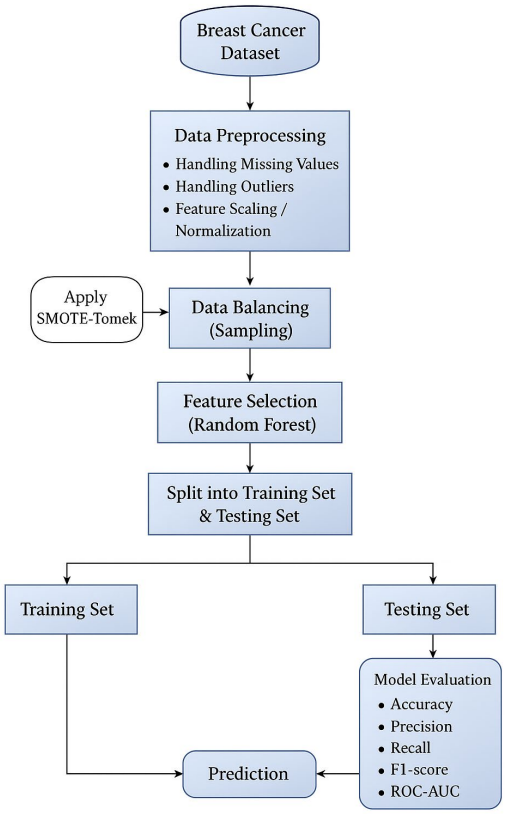

Figure 1: Proposed research framework for breast cancer classification using XGBoost

To prevent information leakage, the dataset was first partitioned into training and test subsets, and the balancing, imputation, outlier-capping, and scaling steps shown in Figure 1 were fitted on the training partition only and then applied to the test partition. The same training/test split was used for all models to ensure a fair comparison.

3.2 Dataset Acquisition

The experimental dataset was obtained from the Kaggle breast cancer repository, a widely used open-source platform for machine learning research. The dataset comprises structured clinical attributes capturing patient demographics, tumour characteristics, and diagnostic outcomes. The target variable is binary, denoting whether the tumour is benign (0) or malignant (1). To ensure methodological transparency, it is noted that this study uses a clinical (tabular) breast cancer dataset rather than the cytological UCI Wisconsin Diagnostic feature set; all features used in this work are listed in Table 2.

In its original form, the dataset comprised 213 records and exhibited a class imbalance, with 120 benign cases outnumbering 93 malignant cases. To address this without introducing information leakage, the SMOTE–Tomek hybrid resampling strategy was applied to the training partition only (Section 3.4): SMOTE generates synthetic

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

minority (malignant) cases via k-nearest-neighbour interpolation, while Tomek Links undersampling removes overlapping borderline samples [9].

The dataset retained 8 clinically meaningful predictor variables—Age, Menopausal status, Tumour size, Lymph node involvement (Inv-Nodes), Breast side, Metastasis, Breast quadrant, and Patient history of breast cancer—together with the Diagnosis Result as the target variable. A detailed description is provided in Table 2. All categorical variables were label-encoded and numeric variables were standardised prior to modelling.

Ethical note: the dataset is fully anonymised and publicly available; no personally identifiable information was included, in compliance with Kaggle's data-use agreement.

Table 2: Summary of the breast cancer dataset

Feature	Type	Description	Range / Categories	Encoding
Age	Numeric	Age of the patient (years).	20 – 85	Continuous
Menopause	Categorical	Menopausal status of the patient.	Premenopause / Postmenopause	0 = Premenopause, 1 = Postmenopause
Tumour Size (cm)	Numeric	Size of the detected tumour (cm).	1 – 10	Continuous
Inv-Nodes	Numeric	Number of involved axillary lymph nodes.	0 – 39	Continuous
Breast	Categorical	Side of the breast affected.	Left / Right	0 = Left, 1 = Right
Metastasis	Categorical	Presence of distant metastasis.	Absent / Present	0 = Absent, 1 = Present
Breast Quadrant	Categorical	Location of the tumour within the breast.	Upper-left, Upper-right, Lower-left, Lower-right, Central	0 – 4
History	Categorical	Patient's prior history of breast cancer.	No / Yes	0 = No, 1 = Yes
Diagnosis Result	Target (Binary)	Diagnostic outcome of the tumour.	Benign / Malignant	0 = Benign, 1 = Malignant

3.3 Data Preprocessing

Prior to model development, the dataset underwent systematic preprocessing to ensure data quality and suitability for machine-learning analysis:

1. Handling missing values: placeholder characters (e.g., "?" or "#") were replaced with NaN. Missing values in categorical features (e.g., Menopause, Quadrant) were imputed using the mode, while missing values in numeric features (e.g., Age, Tumour Size, Inv-Nodes) were imputed using the median, preserving statistical consistency while minimising distortion of feature distributions.
2. Outlier detection and treatment: outliers in continuous variables were identified using the Interquartile Range (IQR) method. Observations below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ were capped at the boundary values (Winsorisation) rather than removed, to preserve the limited dataset size.
3. Categorical encoding: categorical variables were label-encoded—for example, Menopause {Premenopause = 0, Postmenopause = 1}, Breast {Left = 0, Right = 1}, History {No = 0, Yes = 1}, Metastasis {Absent = 0, Present = 1}, and Quadrant {Upper-left = 0, Upper-right = 1, Lower-left = 2, Lower-right = 3, Central = 4}.
4. Feature scaling: all numeric variables were standardised using z-score standardisation (StandardScaler), rescaling values to a mean of 0 and a standard deviation of 1. To prevent data leakage, the scaler was fitted on the training data and the fitted scaler was then applied to the testing data.

3.4 Data Balancing

Medical datasets frequently suffer from class imbalance, where benign cases are over-represented relative to malignant cases. Such imbalance can produce classifiers that achieve high accuracy by predominantly predicting the majority class while failing to detect the minority class—an unacceptable outcome in cancer diagnosis, where sensitivity to malignant tumours is paramount. To address this, the SMOTE–Tomek hybrid strategy was applied exclusively to the training set (to prevent leakage into the test set), combining SMOTE oversampling of the malignant class with Tomek-Links removal of overlapping borderline instances [9]. The training partition, which initially contained 96 benign and 74 malignant records, was balanced to 90 benign and 90 malignant records, enabling the classifiers to learn decision boundaries that were not biased toward the majority class. The test set was left untouched and retained its natural class distribution.

3.5 Feature Selection

Feature selection reduces dimensionality, improves interpretability, and enhances model performance. In this study, Random Forest feature importance [10] was used to rank features by their contribution to classification.

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

The Random Forest was fitted on the training partition with 100 trees, and the Gini-based importance score of each predictor was computed. Features were ranked in descending order of importance, and the eight clinically meaningful predictors listed in Table 2 were retained, while redundant or low-importance attributes were discarded. Computing the importances on the training partition only ensures that no information from the test set influenced feature selection. This ensures that only clinically relevant and statistically meaningful features contribute to the model, reducing computational cost and improving predictive accuracy.

3.6 Dataset Partitioning

After preprocessing, balancing, and feature selection, the dataset was partitioned into a training set (80%), used to train the models, and a testing set (20%), used exclusively for evaluation. This separation avoids information leakage and ensures that the evaluation metrics reflect true generalisation performance.

3.7 Model Development

The proposed classifier was developed using Extreme Gradient Boosting (XGBoost), an ensemble method based on gradient-boosted decision trees, chosen for its ability to handle non-linear relationships, reduce overfitting, and perform well with structured medical data [11]. The hyperparameters, fixed after preliminary tuning, were: n_estimators = 200, max_depth = 4, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 0.8, and reg_lambda = 1.0. The random seed was set to 42 for reproducibility. To provide a fair point of reference, two widely used baseline classifiers—Logistic Regression and a Decision Tree—were trained and evaluated on the identical training and test partitions, using the same preprocessing and resampling. This allows the added value of the ensemble approach to be assessed directly rather than assumed.

3.8 Model Evaluation

The trained model was evaluated on the reserved testing set using multiple metrics: accuracy (overall proportion of correct classifications); precision (proportion of malignant predictions that were truly malignant); recall or sensitivity (proportion of actual malignant cases correctly identified); F1-score (harmonic mean of precision and recall); the area under the Receiver Operating Characteristic curve (ROC–AUC); and the confusion matrix (counts of true/false positives and negatives). To support the statistical reliability of these estimates, 95% confidence intervals were obtained for each metric using 1,000 bootstrap resamples of the test set, and the proposed model was compared against the baseline classifiers using McNemar's test, with a significance threshold of $p < 0.05$. Together these metrics ensure that the model is not only accurate but also clinically reliable in detecting malignant tumours. The optimised and validated model can then be applied to unseen patient data to predict the likelihood of a tumour being benign or malignant.

4.0 Results

4.1 Dataset Balancing Results

The original dataset exhibited a moderate class imbalance. As shown in Figure 2, benign cases (label 0) accounted for 120 samples and malignant cases (label 1) for 93 samples. Although not extreme, this imbalance increases the likelihood of a classifier favouring benign cases and overlooking critical malignant diagnoses. To mitigate this without leakage, SMOTE–Tomek was applied to the training partition only (Section 3.4): SMOTE generates synthetic samples for the minority class, while Tomek Links remove borderline ambiguous samples. As shown in Figure 3, the training set was balanced to 90 benign and 90 malignant cases, improving fairness in training while the test set retained its natural distribution.

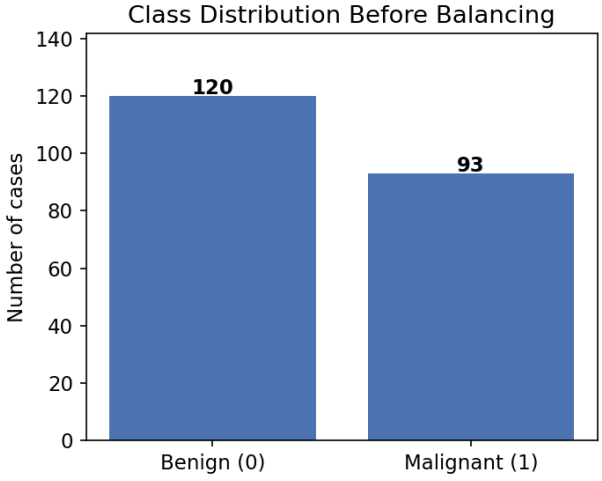

Figure 2: Class distribution before balancing (full dataset)

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

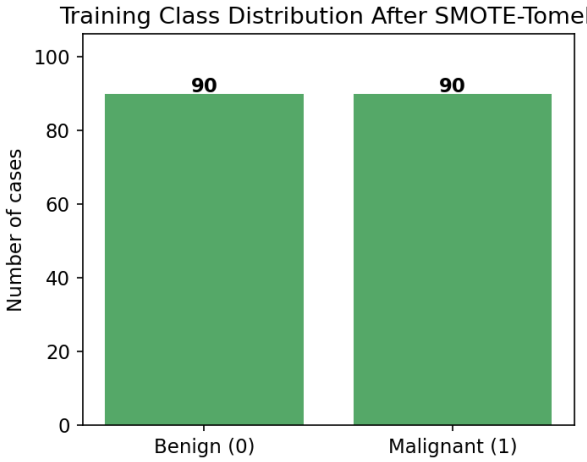

Figure 3: Training-set class distribution after SMOTE–Tomek

4.2 Feature Importance Analysis

To understand which variables most strongly influenced classification, feature importance was computed using a Random Forest model [10]; this aids interpretability by allowing clinicians to evaluate whether the model's reasoning aligns with established medical knowledge. As shown in Figure 4, tumour size was the most influential feature, followed by the number of involved lymph nodes (Inv-Nodes), metastasis, and age. These findings align with clinical knowledge [2], which consistently identifies tumour size, nodal involvement, and metastatic status as critical predictors of malignancy. Menopausal status, breast side, breast quadrant, and patient history were less influential but contributed complementary contextual information.

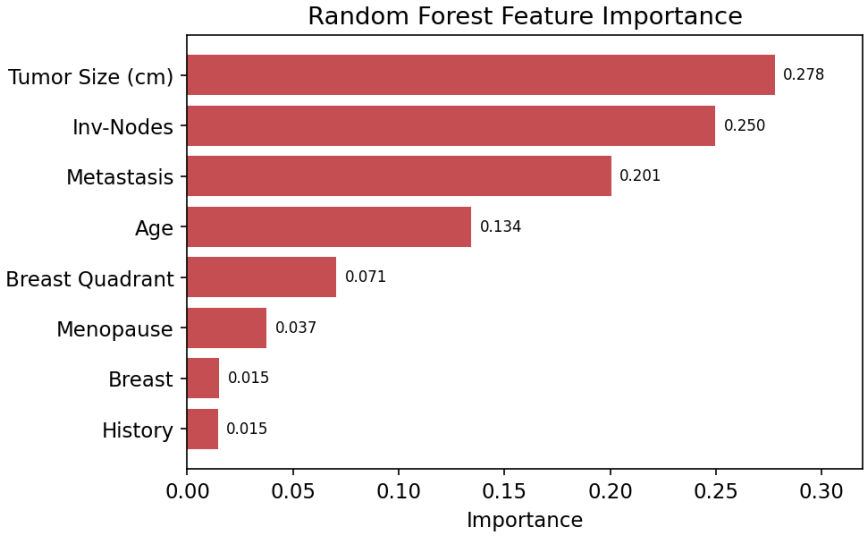

Figure 4: Random Forest feature importance

4.3 Classification Performance

To assess whether the ensemble model offered a genuine advantage, the proposed XGBoost classifier was compared against Logistic Regression and Decision Tree baselines trained and evaluated on the identical partitions. Table 3 reports all metrics on the held-out test set (43 cases), rounded to four decimal places.

Table 3: Performance of the proposed XGBoost classifier and the baseline models on the test set.

Classifier	Accuracy	Precision	Recall	F1-score	ROC–AUC
Logistic Regression	0.9070	1.0000	0.7895	0.8824	0.9474
Decision Tree	0.8372	0.8333	0.7895	0.8108	0.8322
XGBoost (proposed)	0.8140	0.8235	0.7368	0.7778	0.9079

On this dataset, XGBoost achieved an accuracy of 0.8140, recall of 0.7368, F1-score of 0.7778, and ROC–AUC of 0.9079, while Logistic Regression obtained the highest accuracy (0.9070) and ROC–AUC (0.9474). Because the test set contains only 43 cases, the estimates carry substantial uncertainty: the bootstrap 95% confidence intervals for XGBoost were accuracy [0.70, 0.91], recall [0.53, 0.92], and ROC–AUC [0.79, 0.99].

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

McNemar's test did not detect a statistically significant difference between XGBoost and either baseline ($p = 0.125$ versus Logistic Regression; $p = 1.000$ versus Decision Tree). The ensemble model therefore performed comparably to, but did not outperform, the simpler baselines on this small dataset.

4.4 Graphical Performance Evaluation

4.4.1 Receiver Operating Characteristic (ROC) Curve

Figure 5 shows the ROC curves of the three classifiers. All models perform well above the chance diagonal; Logistic Regression and XGBoost achieve the highest areas under the curve (0.95 and 0.91, respectively), while the Decision Tree trails (0.83). The overlap between the Logistic Regression and XGBoost curves is consistent with the non-significant difference reported above.

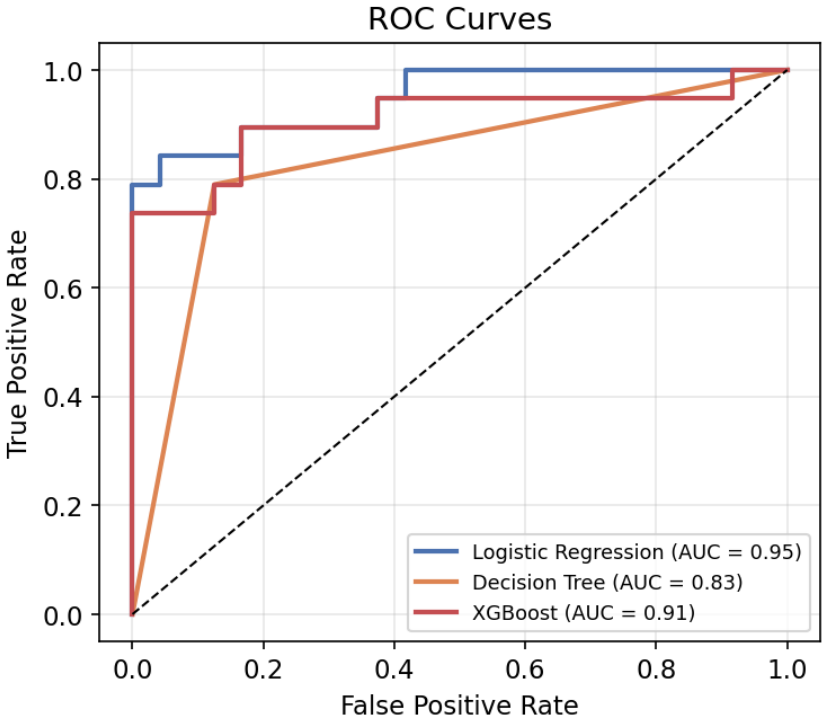

Figure 5: ROC curves of the evaluated classifiers

4.4.2 Model Evaluation Metrics

Figure 6 compares the three classifiers across accuracy, precision, recall, F1-score, and ROC–AUC, providing a visual summary of Table 3 and showing that the models perform within a narrow band on this dataset.

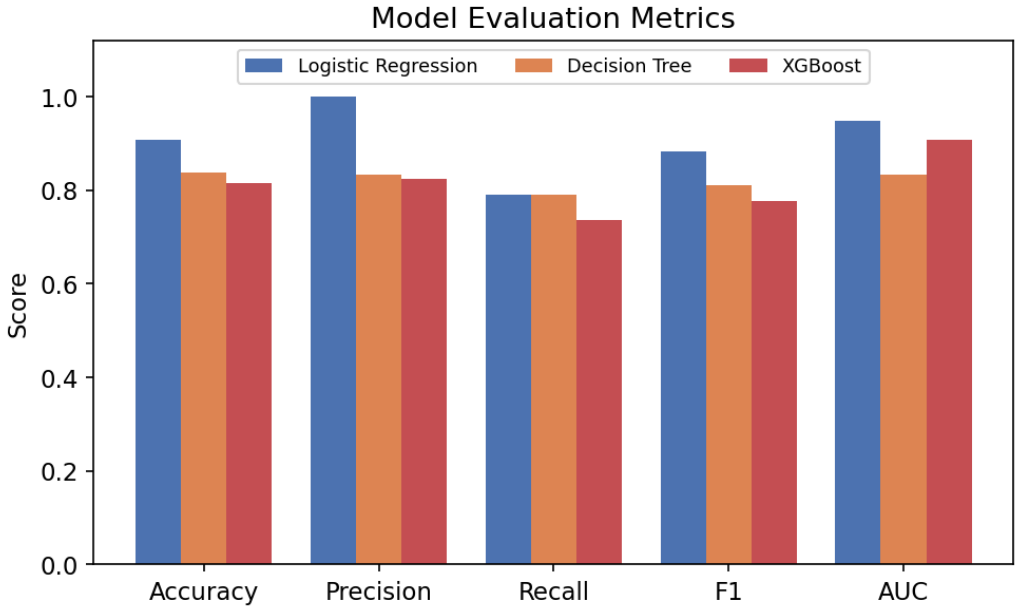

Figure 6: Comparison of model evaluation metrics

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

4.4.3 Confusion Matrix

Figure 7 presents the confusion matrix for the XGBoost classifier on the test set. Of the 43 test cases, the model correctly classified 21 benign and 14 malignant cases, with 3 false positives and 5 false negatives. In a cancer-screening context the 5 false negatives are the most consequential outcome and reflect the model's moderate recall on this small sample.

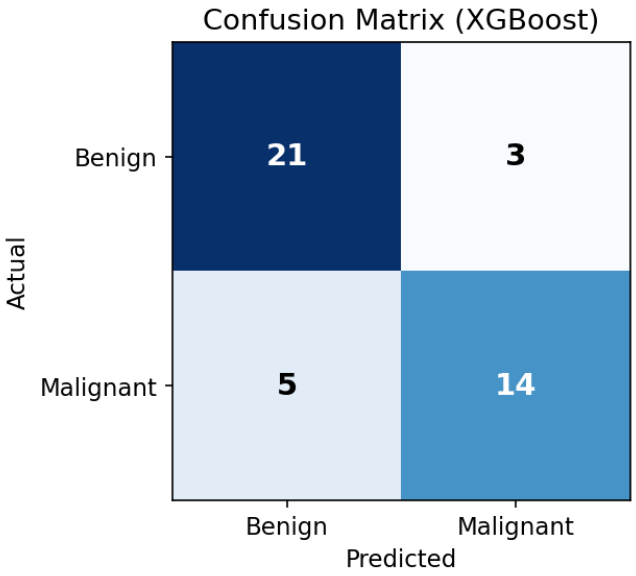

Figure 7: Confusion matrix of the XGBoost classifier on the test set

5.0 Discussion

The results show that the SMOTE–Tomek resampling and the ensemble pipeline produced usable classifiers, but they also temper the expectation that a more complex model necessarily yields better predictions. On the held-out test set, XGBoost reached a ROC–AUC of 0.9079 and recall of 0.7368, whereas a regularised Logistic Regression obtained the highest accuracy (0.9070) and ROC–AUC (0.9474). The differences between the models were small and, by McNemar's test, not statistically significant ($p = 0.125$ versus Logistic Regression; $p = 1.000$ versus Decision Tree). On a dataset of this size, therefore, the ensemble approach did not provide a measurable advantage over a well-specified linear baseline.

The feature-importance analysis nonetheless adds clinical interpretability. The prominence of tumour size, involved lymph nodes, and metastasis mirrors established clinical oncology knowledge [2], [16], which increases confidence that the models are responding to clinically meaningful signal rather than artefacts of the data. This interpretability is arguably more valuable than marginal differences in headline metrics for clinical decision support.

Two factors most plausibly explain these findings. First, the dataset is small (213 records; 43 test cases), so the performance estimates are inherently noisy—reflected in the wide bootstrap confidence intervals (for XGBoost, accuracy [0.70, 0.91] and ROC–AUC [0.79, 0.99]). Second, with only eight low-dimensional, largely linear clinical features, there is limited non-linear structure for a gradient-boosting model to exploit, which is precisely the regime in which simpler models are competitive. These observations are consistent with prior comparative studies [8], [19] that report narrow performance gaps between classical and ensemble methods on small structured breast cancer datasets.

From a clinical standpoint, the pipeline is best positioned as an interpretable, reproducible benchmark and a potential decision-support aid for early screening and triage in resource-constrained settings—not as a replacement for histopathological confirmation, and not, on current evidence, as a demonstrably superior method to simpler baselines. The central limitation is data size; external validation on larger, multi-institutional cohorts is required before any clinical use, and would also clarify whether the ensemble's expected advantages materialise at scale.

6.0 Conclusion

This research designed, implemented, and evaluated a machine-learning pipeline for the early prediction of breast cancer using structured clinical data, motivated by the need for accurate, interpretable, and resource-efficient diagnostic tools in oncology—particularly in low-resource settings such as Nigeria, where late detection contributes to high mortality. A publicly available dataset containing clinical features—including tumour size, age, menopausal status, lymph-node involvement, metastasis, and breast quadrant location—was used. The study makes the following contributions: a reproducible, leakage-free pipeline that balances the training data with SMOTE–Tomek and benchmarks an ensemble model against Logistic Regression and Decision Tree baselines on identical partitions; clinically relevant feature insights confirming the diagnostic value of tumour size, involved

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

lymph nodes, and metastasis; and an honest performance assessment showing that, on this small dataset, XGBoost performed comparably to—but not better than—a regularised logistic regression, with no statistically significant difference between them. The principal finding is therefore methodological: for small, low-dimensional clinical datasets, careful validation and interpretability matter more than model complexity, and larger datasets are needed to establish whether ensemble methods offer a genuine advantage.

6.1 Recommendations

For clinical practice, the model can be integrated into hospital decision-support systems for early screening and triage, and its probability estimates can stratify patients into low-, medium-, and high-risk groups for efficient use of diagnostic resources; its computational efficiency makes it suitable for low-resource settings. For machine-learning development, future work should integrate multimodal data (imaging and genomics), use larger and more diverse datasets, apply explainable-AI techniques such as SHAP or LIME for case-level explanations, and validate the model prospectively in real-world hospital environments. For policy and research, governments and healthcare agencies should support ML-driven diagnostics through funding, training, and infrastructure, and encourage open, annotated medical datasets.

6.2 Limitations of the Study

While the study achieved promising results, certain limitations are acknowledged: the dataset was relatively small, limiting the ability to capture rare variations; a single data source was used, so results may vary across populations; only structured data were used, excluding imaging and unstructured data; and the model was not validated in a live clinical environment, which is essential for final adoption.

6.3 Future Work

Building on this research, future studies should develop hybrid models that integrate structured data, imaging, and genomic biomarkers; conduct longitudinal studies to test model stability over time and patient follow-up; deploy mobile and cloud-based versions to reach rural healthcare facilities; and explore deep learning and transfer learning for comparison with ensemble approaches while balancing interpretability.

6.4 Concluding Remarks

This study demonstrates that carefully designed and rigorously validated machine-learning pipelines, when aligned with clinical knowledge, can provide interpretable and reproducible support for early breast cancer detection. On the small dataset examined here the ensemble model did not surpass simpler baselines, underscoring that methodological rigour and adequate data, rather than model complexity alone, are the key requirements. With these foundations in place, larger and more diverse datasets offer a clear path toward bridging artificial intelligence and oncology and supporting clinicians in delivering faster, more reliable diagnoses.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] A. G. Waks and E. P. Winer, "Breast cancer treatment: A review," JAMA, vol. 321, no. 3, pp. 288–300, 2019, doi: 10.1001/jama.2018.19323.

[3] S. Łukasiewicz and M. Rybnik, "Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies: An updated review," Cancers, vol. 13, no. 14, art. 3452, 2021, doi: 10.3390/cancers13143452.

[4] R. A. Smith, V. Cokkinides, and H. J. Eyre, "American Cancer Society guidelines for breast cancer screening: Update 2010," CA: A Cancer Journal for Clinicians, vol. 60, no. 2, pp. 110–116, 2010, doi: 10.3322/caac.20068.

[5] A. I. Khan, I. Ali, and Z. Riaz, "Machine learning approaches for breast cancer prediction and diagnosis," Journal of Healthcare Engineering, vol. 2022, art. 5598764, pp. 1–12, 2022, doi: 10.1155/2022/5598764.

[6] V. Chaurasia and S. Pal, "Applications of machine learning techniques to predict diagnostic breast cancer," International Journal of Computer Applications, vol. 175, no. 9, pp. 25–32, 2021.

[7] Q. Zhuang, Y. Liu, and X. Li, "Explainable AI in healthcare: Interpretability of machine learning models for cancer prediction," Computers in Biology and Medicine, vol. 146, art. 105657, 2022, doi: 10.1016/j.compbiomed.2022.105657.

[8] P. Sharma, R. Gupta, and S. Verma, "An improved ensemble approach using SMOTE and extreme gradient boosting for breast cancer prediction," Applied Soft Computing, vol. 97, art. 106728, 2020, doi: 10.1016/j.asoc.2020.106728.

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[12] S. Kumar, R. Yadav, and A. Singh, "Random forest classifier with feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 170, art. 114114, 2021, doi: 10.1016/j.eswa.2020.114114.

[13] H. Patel and D. Ghosh, "Performance analysis of decision tree and k-nearest neighbour classification for breast cancer detection," International Journal of Computer Applications, vol. 182, no. 1, pp. 1–6, 2018, doi: 10.5120/ijca2018917435.

[14] M. Pandey, R. Gautam, and A. Verma, "Breast cancer detection using support vector machines and deep learning," Journal of King Saud University – Computer and Information Sciences, vol. 33, no. 10, pp. 1186–1193, 2021, doi: 10.1016/j.jksuci.2019.04.001.

[15] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," Nature Communications, vol. 10, art. 564, 2019, doi: 10.1038/s41467-019-09117-8.

[16] World Health Organization, "Breast cancer: Early diagnosis and screening," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[17] L. Zhang, J. Zhang, and D. Zhang, "Feature selection techniques in medical diagnostics: A survey," IEEE Reviews in Biomedical Engineering, vol. 14, pp. 1–18, 2021, doi: 10.1109/RBME.2020.3004551.

[18] J. Xu, Z. Mo, and Q. Feng, "An intelligent breast cancer diagnosis system based on deep learning," Journal of Healthcare Engineering, vol. 2019, art. 9793060, pp. 1–9, 2019, doi: 10.1155/2019/9793060.

[19] A. Sharma, D. Gupta, and A. Verma, "A comparative study of classification algorithms for breast cancer prediction," International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, pp. 547–553, 2020, doi: 10.14569/IJACSA.2020.0110770.

[20] R. Kumar, A. Singh, and A. Tiwari, "Breast cancer detection using hybrid machine learning techniques," Materials Today: Proceedings, vol. 46, pp. 10143–10147, 2021, doi: 10.1016/j.matpr.2020.12.1066.

Received: 07/12/2026; Revised: 07/04/2026; Accepted: 0707/2026; Published: 21/08/2026