JP Journal of Biostatistics

https://pphmjopenaccess.com
https://doi.org/10.17654/0973514326002
Volume 26, Number 1, March 2026, Pages 11-40
P-ISSN: 0973-5143
E-ISSN: 3107-4731

LEVERAGING ON HYBRID MACHINE LEARNING MODELS FOR EARLY BREAST CANCER DETECTION

Gideon Nyakundi[1,2,*], John Ndiritu[1], Joseph Mwaniki[1] and Timothy Kamanu[1]

[1]Faculty of Science and Technology
The University of Nairobi
Kenya
e-mail: nyakundigideon8@gmail.com

[2]The Kenya Institute for Public Policy Research and Analysis
Kenya

Abstract

Breast cancer is among the most common cancers in women worldwide, and outcomes improve with early detection. As machine learning enters routine care, data driven diagnostic systems may support earlier risk estimation. We present a compact pipeline that uses Principal Component Analysis for dimensionality reduction and Borderline-SMOTE for imbalance correction, followed by classification with Light Gradient Boosting Machine. Using the

Received: October 25, 2025; Accepted: December 8, 2025

Keywords and phrases: breast cancer, LightGBM, principal component analysis, Borderline-SMOTE.

*Corresponding Author

How to cite this article: Gideon Nyakundi, John Ndiritu, Joseph Mwaniki and Timothy Kamanu, Leveraging on hybrid machine learning models for early breast cancer detection, JP Journal of Biostatistics 26(1) (2026), 11-40. https://doi.org/10.17654/0973514326002

Published Online: December 25, 2025

standardized Wisconsin Breast Cancer Diagnostic dataset, we retain 20 features to capture key variance while limiting redundancy and noise. Borderline-SMOTE is applied within each training fold to refine class boundaries. Performance is evaluated with stratified 10-fold cross validation and compared with seven alternatives: XGBoost, Support Vector Machines, Random Forests, Logistic Regression, Gaussian Naive Bayes, *k* Nearest Neighbor, and a Multilayer Perceptron. With 20 components, the proposed model attains accuracy 0.993, precision 1, recall 0.986, F1 0.993, and AUC 1.000 for distinguishing benign from malignant cases, outperforming baselines. These findings suggest that coupling dimensionality reduction, boundary focused resampling, and gradient boosted trees can enhance diagnostic performance and may inform clinical decision support.

1. Introduction

Breast cancer remains the most diagnosed cancer and a leading cause of cancer-related mortality among women globally, with an estimated 2.3 million new cases (11.7%) and 685,000 deaths reported worldwide in 2020 alone [1]. Early and accurate diagnosis is critical for improving prognosis and survival rates, yet conventional diagnostic modalities—such as mammography, magnetic resonance imaging (MRI), ultrasound, and histopathology—face limitations in sensitivity, specificity, accessibility, and cost, particularly in low- and middle-income settings [2, 3]. Moreover, inter-observer variability and interpretation errors remain common challenges in manual image-based diagnosis [4].

In recent years, the application of machine learning (ML) and deep learning (DL) techniques in medical imaging has opened new pathways for advancing breast cancer diagnostics. These methods have shown promise in improving classification performance, feature extraction, and decision support by identifying subtle and complex patterns within high-dimensional data [5, 6]. However, single-algorithm ML systems often struggle with generalizability across diverse datasets, data imbalance, and model interpretability—factors that hinder their clinical translation [7, 8].

To address these issues, hybrid machine learning models—combining multiple algorithms or stages of analysis—have emerged as a compelling approach for enhancing diagnostic accuracy and model robustness. By integrating ensemble learning, optimized feature selection, and decision-level fusion strategies, hybrid frameworks can leverage the strengths of individual models while mitigating their respective limitations [9]. Prior studies have demonstrated that such integrative systems outperform standalone classifiers in both image-based and clinical feature-based breast cancer detection tasks [10, 11].

1.1. About breast cancer

Breast cancer is a malignant disease that develops from epithelial cells within the ducts or lobules of the breast. It arises when genetic and epigenetic changes disrupt normal regulatory mechanisms, allowing abnormal cells to proliferate uncontrollably. The aetiology of breast cancer is multifactorial, involving both inherited and acquired risk factors. Genetic mutations, particularly in BRCA1 and BRCA2, account for 5-10% of cases and significantly elevate lifetime risk. Other risk factors include advancing age, early menarche, late menopause, nulliparity, hormone replacement therapy, obesity, alcohol consumption, and prior radiation exposure—many of which are linked to prolonged estrogen exposure [12].

Globally, breast cancer is the most frequently diagnosed cancer in women, with over 2.3 million new cases and 685,000 deaths reported in 2020 alone. Incidence rates are highest in developed countries due to effective screening programs, while mortality rates remain elevated in low- and middle-income countries where late-stage diagnosis and limited access to care are common [12]. Though most cases occur in women over 50, younger individuals with genetic predisposition are increasingly affected. Male breast cancer is rare, comprising less than 1% of all diagnoses [12].

The disease evolves through a series of molecular and histopathological changes, beginning with atypical hyperplasia and progressing to carcinoma in situ (ductal or lobular) and then invasive carcinoma. Common histologic

types include invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) [13]. Molecular classification based on estrogen receptor (ER), progesterone receptor (PR), HER2 status, and proliferation index (Ki-67) divides breast cancer into luminal A, luminal B, HER2-enriched, and triple-negative subtypes [12]. These subtypes are not only prognostically significant but also inform treatment selection, as hormone receptor-positive tumours respond well to endocrine therapy, while HER2-positive and triple-negative types may require targeted or cytotoxic treatments [14].

Diagnosis typically relies on a triple assessment approach that includes clinical examination, imaging (mammography, ultrasound, MRI), and tissue biopsy. Histopathological analysis confirms malignancy and guides molecular profiling. Staging using the TNM system further informs prognosis and therapeutic strategy [15]. Despite progress in treatment, early detection remains critical to improving survival outcomes, particularly given that many tumours are not diagnosed until they exceed 30 mm in diameter [16]. This underscores the need for more effective, accessible, and data-driven diagnostic approaches—an area in which machine learning and hybrid computational models offer significant promise for improving early detection and personalized care.

1.2. Objectives of the proposed approach

This study proposes a novel hybrid diagnostic framework specifically designed for breast cancer classification tasks. The framework integrates three key components within a layered architecture: (i) a dimensionality reduction technique to extract the most relevant features and reduce computational complexity, (ii) a data imbalance handling strategy to correct class distribution skewness and improve model fairness, and (iii) a supervised machine learning algorithm for final classification. By combining these components, the model aims to enhance predictive accuracy, reduce false diagnoses, and improve generalizability across heterogeneous datasets. The purpose of this research is twofold: (i) to develop and evaluate a robust hybrid model that outperforms traditional single-method approaches in terms of performance, efficiency, and resilience; and (ii) to demonstrate the

model's clinical utility through comprehensive benchmarking on diverse, publicly available breast cancer datasets.

By bridging computational innovation with clinical needs, this work contributes to the ongoing advancement of AI-assisted precision oncology. It supports the broader aim of developing diagnostic tools that are not only technically robust but also ethically sound, interpretable, and suitable for deployment in real-world healthcare environments.

1.3. Related works

In recent years, ML has been increasingly integrated into breast cancer diagnostics to improve prediction accuracy, reduce human error, and enhance decision support. Numerous studies have demonstrated the effectiveness of supervised ML algorithms such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and ensemble methods like XGBoost and LightGBM in differentiating benign from malignant lesions using imaging and genomic features [17, 18]. However, despite promising accuracy metrics, two persistent data-related challenges remain at the forefront of model development: high dimensionality and class imbalance [7, 18]. High dimensionality arises when the number of features (e.g., gene expression levels, pixel intensities) far exceeds the number of samples, increasing the risk of overfitting and reducing generalizability [19]. Simultaneously, class imbalance—where malignant cases are significantly underrepresented compared to benign ones—leads to skewed learning and poor sensitivity, particularly for early-stage cancer detection [7]. These challenges necessitate robust preprocessing and model design strategies to ensure clinical reliability.

1.3.1. High dimensionality of breast cancer data

High dimensionality is a well-recognized challenge in breast cancer prediction, often leading to overfitting, increased computational complexity, and reduced model interpretability if not properly addressed. To mitigate these issues, the papers reviewed employed various dimensionality reduction techniques aimed at isolating the most informative features while eliminating

irrelevant or redundant data. These techniques included filter methods, wrapper approaches, embedded models, feature extraction algorithms, and hybrid optimization strategies.

Several studies utilized traditional filter-based feature selection methods, which rank features based on statistical relevance. Principal Component Analysis (PCA) emerged as a frequently applied technique, valued for its ability to transform input features into orthogonal components while retaining most of the dataset's variance [11, 18, 20, 21, 22, 23, 24]. Other commonly used filters included Minimum Redundancy Maximum Relevance (mRMR) and univariate tests such as chi-squared, ANOVA, t-tests, and F-tests, all of which helped reduce noise and redundancy [25, 26]. Additionally, Mutual Information, Fold Change, and False Discovery Rate (FDR) were employed to identify features with high predictive power, particularly in omics datasets [18, 27]. These filter methods are computationally efficient and well-suited to high-dimensional biomedical data.

Deep learning–based feature extraction methods were also prominent. Pretrained models such as ResNet50, EfficientNetB3, ConvNeXtTiny, and DenseNet121 were leveraged to automatically extract hierarchical feature representations from imaging and omics data [28]. Autoencoders (AE) and Stacked Autoencoders (SAE) were used to compress input data into informative lower-dimensional representations, effectively capturing nonlinear relationships often missed by classical methods [11, 18, 29]. These approaches proved particularly powerful in handling complex feature distributions, especially in medical imaging.

Wrapper techniques such as Recursive Feature Elimination (RFE), used in three studies, represented a commonly adopted approach in this category [30, 31, 32]. RFE was frequently combined with classifiers like SVM and Random Forest to rank and select features based on model performance. Another wrapper method used was the Single Parameter Decision-theoretic Rough Set (SPDTRS), which evaluates feature subsets by their effect on classification accuracy [33]. Although wrapper methods are computationally

intensive, they offer highly tailored feature selection aligned with model behavior.

Embedded methods, which integrate feature selection directly into the model training process, were also observed. These included Boruta (and SHAP-enhanced Boruta), cost-sensitive Random Forests, and Support Vector Machine–recursive Feature Elimination with Parameter Optimization (SVM-RFE-PO) [32, 34]. Embedded approaches simultaneously train the model and rank features by importance, offering advantages in both interpretability and performance. They were particularly preferred when there was a need to balance model complexity with explainability.

Several studies employed hybrid approaches that combined multiple feature selection strategies for enhanced performance. Metaheuristic algorithms such as Grammatical Evolution (GE), Whale Optimization Algorithm, Seagull Optimization Algorithm (SGA), Genetic Algorithms (GA), and Particle Swarm Optimization (PSO) were frequently integrated with filter or wrapper methods [17, 35, 36, 37, 38]. These optimizers facilitate global search across feature subsets, minimizing the risk of local minima and enhancing model generalization. Additionally, ensemble multi-filter approaches were employed in some studies, with reported accuracies reaching as high as 100% [17], demonstrating the potential of combining several selection techniques to leverage their individual strengths.

1.3.2. Preprocessing for imbalanced data

In breast cancer detection studies, class imbalance within high-dimensional datasets presents a significant challenge for machine learning applications. The minority class—typically representing patients with breast cancer—appears far less frequently than the majority (healthy) class [7], thus adversely affecting model training. Often, this results in poor sensitivity and frequent misclassification of the clinically critical minority group. Standard classification algorithms are inherently biased toward the majority class due to its higher prior probability, thereby compromising the generalizability and clinical relevance of predictive performance [34]. To mitigate these effects,

researchers have employed a range of imbalance-handling strategies, broadly classified into data-level and algorithm-level approaches. Many of the reviewed studies implemented data-level strategies aimed at rebalancing class distributions prior to model training. These included popular techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) and its variants, Adaptive Synthetic Sampling (ADASYN), up-sampling of the minority class, and synthetic data generation [17, 22, 29, 30, 32]. These methods help to balance the training dataset and mitigate classifier bias toward the majority class.

Algorithm-level solutions were also used to focus on modifying the learning process itself. Techniques in this category included cost-sensitive learning, class weight adjustments, and the use of ensemble models to enhance sensitivity to minority class predictions during training [27, 31, 34, 36, 37]. These methods allow the model to penalize misclassifications of minority class instances more heavily, thereby improving recall without significantly compromising overall performance. A smaller proportion of studies adopted hybrid approaches, integrating both data-level and algorithm-level strategies to capitalize on the advantages of each. These combined methods aimed to improve robustness and mitigate bias more comprehensively [21, 38]. Notably, many studies did not explicitly address class imbalance, highlighting a methodological gap that may limit the reliability of their findings in real-world diagnostic settings. This underscores the need for more consistent adoption of imbalance-handling techniques, particularly given the critical importance of correctly identifying minority class cases in clinical applications.

1.3.3. Hybrid ML models

The integration of machine learning models for breast cancer prediction with dimensionality reduction techniques and class imbalance handling strategies significantly enhanced diagnostic performance across the reviewed studies. Notably, models employing hybrid strategies—combining both data-level and algorithm-level approaches—consistently outperformed others across all evaluation metrics. These hybrid models achieved an average

accuracy of 99.31%, precision of 98.30%, recall of 99.21%, F1 score of 98.73%, and an AUC of 97.52%, reflecting excellent discriminatory power and generalizability. In contrast, models that relied solely on algorithm-level methods for class balancing performed moderately well, with an average accuracy of 90.59% and AUC of 93.48%. Those utilizing only data-level techniques, while achieving relatively strong precision (95.71%), demonstrated lower overall accuracy (81.20%). The weakest performance was observed in models that did not explicitly address class imbalance, particularly in recall (81.53%) and F1 score (68.77%), highlighting the critical importance of systematic imbalance mitigation in medical diagnostic modelling. These findings underscore the synergistic benefit of coupling dimensionality reduction with comprehensive class balancing strategies. Together, they not only reduce computational burden and overfitting but also improve sensitivity to minority class instances—often the most clinically significant in breast cancer diagnostics.

The performance of hybrid models—those integrating both dimensionality reduction and imbalance-handling techniques—varied across specific ML classifiers. Among them, LightGBM emerged as the most effective, achieving the highest average accuracy (95.79%), precision (99.00%), and F1 score (98.68%), with an impressive AUC of 0.95. These results highlight its robustness and efficiency in extracting informative patterns from high-dimensional, imbalanced datasets. XGBoost also demonstrated strong performance, particularly excelling in recall (94.01%) and AUC (0.97), making it a reliable option for detecting minority class cases such as confirmed cancer diagnoses. Logistic Regression, though slightly lower in accuracy (90.52%), achieved a competitive F1 score (98.68%) and AUC (0.97), likely benefiting from its compatibility with well-engineered feature sets. Support Vector Machines (SVMs) and Multi-Layer Perceptrons (MLPs) delivered balanced results, with SVMs performing well in recall (91.51%) and MLPs maintaining stable performance across metrics. In contrast, Random Forests, Naive Bayes, and k-Nearest Neighbors (k-NN) exhibited reduced accuracy and AUC values, possibly due to limitations in scaling to high-dimensional feature spaces or susceptibility to noise. Overall,

tree-based boosting algorithms like LightGBM and XGBoost showed superior integration with hybrid modeling pipelines, underscoring their adaptability and effectiveness in complex medical diagnostic tasks involving imbalanced and high-dimensional data.

2. Methods

2.1. Dataset

We used the publicly accessible Wisconsin Diagnostic Breast Cancer WDBC dataset [39, 40], which contains 569 patient samples comprising 357 benign and 212 malignant cases, with no missing values. Each sample includes 30 real-valued predictors computed from digitized fine-needle aspiration FNA images of breast masses, capturing morphological characteristics of cell nuclei such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, reported as mean, standard error, and worst measurements [39].

2.2. Data preprocessing

Prior to model training, we implemented a standardized pipeline to ensure feature comparability and reproducible evaluation. All continuous predictors were scaled with min–max normalization to the interval 0 to 1, fitting the scaler on the training partition and applying the learned parameters to the held-out test set to prevent information leakage. The dataset was randomly split into training and test subsets using an 80 to 20 ratio, with a fixed random seed for reproducibility and stratification to preserve the original class proportions.

2.3. Principal component analysis

We used principal component analysis to reduce dimensionality and construct compact feature sets for classification.

Let $X \in \mathbb{R}^{n \times p}$ denote the training matrix of standardized predictors mean 0, variance 1. PCA finds an orthogonal rotation $W = [w_1, ..., w_p]$ that

diagonalizes the sample covariance $S = \frac{1}{n-1} X^\top X$. The principal components are

$$Z = XW, \quad Sw_j = \lambda_j w_j, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

where λ_j is the variance explained by component *j*. The proportion of variance explained by the first *m* components is

$$\sum_{j=1}^{m} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j.$$

To form feature sets of different capacity, we retained the top $m \in \{10, 20, 28, 30\}$ components from the training data and used $Z_{1:m}$ as inputs to each classifier.

The PCA rotation *W* and centering-scaling parameters were fit on the training split only, then applied unchanged to the test split to avoid leakage. This procedure yields near-uncorrelated predictors, reduces noise, and can improve generalization by concentrating most of the signal in fewer dimensions. We report results for all four values of *m* to show the sensitivity of each model to the degree of dimensionality reduction.

2.4. Borderline-SMOTE

To mitigate class imbalance in the WDBC data, we applied Borderline-SMOTE only to the training split, which constituted 80 percent of the samples, to avoid information leakage into the test set. The method examines each minority instance's k-nearest neighbours and labels the instance as safe, noisy, or dangerous according to the local mix of majority and minority neighbours. Only dangerous seeds, which lie in boundary regions with many majority neighbours, are oversampled.

For each dangerous minority point x_i, let $N_{\min}(x_i) = \{k_1, k_2, ..., k_j\}$ be its minority neighbors within the k-nearest neighborhood. Synthetic observations are generated by linear interpolation between x_i and a

randomly selected $k_t \in N_{\min}(x_i)$:

$$x_{\text{synth}} = x_i + \lambda(k_t - x_i), \quad \lambda \sim \mathcal{U}[0, 1].$$

Repeating this operation across seeds and neighbours produces the desired oversampling ratio while concentrating new samples near decision boundaries. This targeted augmentation increases the density of informative minority cases, improves sensitivity to malignant lesions, and avoids indiscriminate replication of safe or noisy points.

2.5. Light gradient-boosting machine (LightGBM)

This paper proposes LightGBM, a gradient boosting decision tree framework that builds an additive scorer by fitting shallow trees to gradients and grows trees leaf-wise for larger loss reduction per split. LightGBM also accelerates training with histogram binning, gradient-based one-side sampling, and exclusive feature bundling, which reduce memory and computation while maintaining accuracy.

Given samples $S = \{(x_i, y_i)\}_{i=1}^{n}$ with $x_i \in \mathbb{R}^{30}$ and $y_i \in \{0, 1\}$, boosting constructs an additive model

$$F_t(x) = F_{t-1}(x) + \rho_t h_t(x), \quad t = 1, \ldots, T,$$

initialized by

$$F_0(x) = \arg\min_{\alpha} \sum_{i=1}^{n} L(y_i, \alpha),$$

and at each stage chooses (ρ_t, h_t) to approximately minimize

$$(\rho_t, h_t) = \arg\min_{\rho, h} \sum_{i=1}^{n} L(y_i, F_{t-1}(x_i) + \rho h(x_i)).$$

These equations follow the standard gradient boosting formulation that LightGBM instantiates with tree learners.

With a logistic link for binary classification, LightGBM optimizes a regularized objective via a second-order Taylor expansion around F_{t-1}:

$$\text{Obj}_t \approx L(F_{t-1}) + g^\top \Delta + \frac{1}{2} h^\top \Delta^2 + \Omega(T_t),$$

where $g = \partial L(F_{t-1})/\partial F_{t-1}$ and $h = \partial^2 L(F_{t-1})/\partial F_{t-1}^2$. Aggregating first and second derivatives within each leaf j gives $G_j = \sum_{i \in J} g_i$ and $H_j = \sum_{i \in J} h_i$ and the per-tree objective reduces to

$$\sum_{j=1}^{J} \left[G_j \omega_j + \frac{1}{2}(H_j + \beta)\omega_j^2 + \alpha |\, \omega_j \,| \right],$$

with J leaves and leaf score ω_j. The L1 and L2 penalties α and β improve stability and control overfitting. The optimal leaf value is

$$\omega_j^* = -\frac{\text{sgn}(G_j)\max(0, |\, G_j \,| - \alpha)}{H_j + \beta}.$$

These relationships correspond to Equations (5) to (14) in the reference and define LightGBM's regularized leaf estimates.

To decide splits under leaf-wise growth, LightGBM evaluates the regularized gain

$$\text{Gain} = \frac{G_{L,\alpha}^2}{H_L + \beta} + \frac{G_{R,\alpha}^2}{H_R + \beta} - \frac{G_\alpha^2}{H + \beta},$$

where L and R denote the left and right child nodes and $G_\alpha = \text{sgn}(G)\max(0, |\, G \,| - \alpha)$. The split that maximizes gain is selected.

The final model is $F_T(x) = \sum_{t=1}^{T} h_t(x)$. We map logits to probabilities with $p(x) = \sigma(F_T(x))$, $\sigma(z) = 1/(1 + e^{-z})$ and classify a case as malignant

when $p(x)$ exceeds a threshold chosen on validation data. This combines the gradient boosting foundation with LightGBM's histogram binning, GOSS, EFB, and leaf-wise splitting for efficient and accurate WDBC classification.

2.6. Evaluation metrics

We assessed classification performance with Accuracy, Precision, Recall, F1 score, Area Under the ROC Curve AUC, the confusion matrix, and ROC curves.

Let malignant be the positive class, and let *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives. Accuracy measures overall correctness:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Precision quantifies the reliability of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Recall sensitivity measures the proportion of malignant cases correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

F1 is the harmonic mean of Precision and Recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The ROC curve plots the true positive rate *TPR* against the false positive rate *FPR* as the decision threshold varies, where

$$TPR = \frac{TP}{TP + FN}, \; FPR = \frac{FP}{FP + TN}.$$

AUC summarizes the ROC curve into a single threshold-free measure of separability between malignant and benign classes. The confusion matrix reports *TP*, *FP*, *TN* and *FN* counts and supports error analysis by revealing the trade-off between missed cancers *FN* and false alarms *FP*. Unless otherwise stated, metrics were computed per fold and summarized as the mean and standard deviation over the 10-fold cross validation. For probability outputs, default class assignment used a 0.5 threshold; ROC and AUC were computed from the full range of thresholds.

2.7. Experimental settings

Model training and testing were conducted on a Windows 11 workstation with an Intel Core i7 processor and 16 GB RAM. All models were implemented in R 4.5.1. Dimensionality reduction used principal component analysis, and we evaluated four feature sets comprising the top 10, 20, 28 and 30 principal components. For each feature set, we created two versions of the training data: the original and a Borderline-SMOTE augmented variant applied only to the training partition to avoid information leakage. The proposed LightGBM classifier was benchmarked against XGBoost, support vector machine, multilayer perceptron, random forest, logistic regression, k-nearest neighbors, and Gaussian Naive Bayes. Each algorithm was evaluated both before and after Borderline-SMOTE to quantify the incremental effect of boundary-focused oversampling on classification of benign and malignant cases. Figure 1 summarizes the proposed approach.

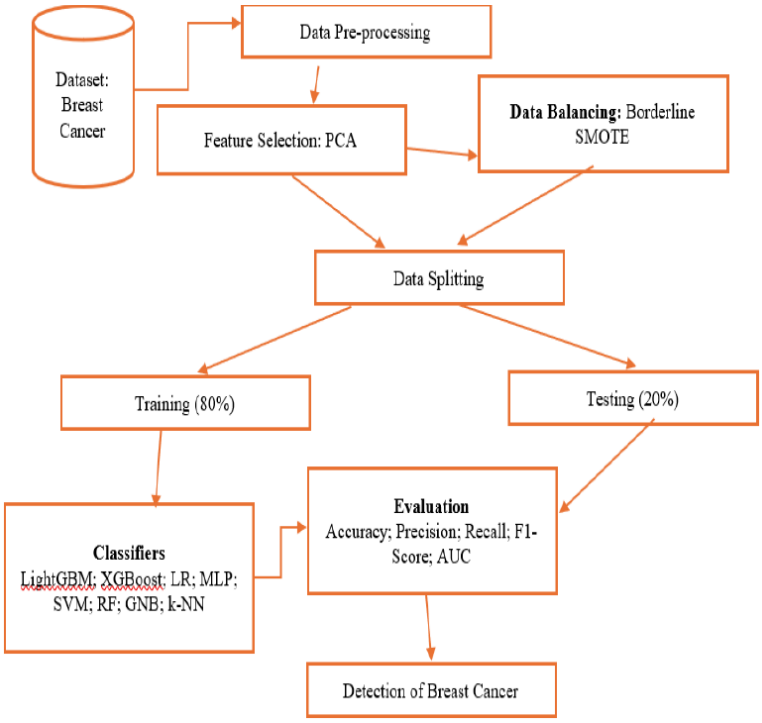

Figure 1. Proposed approach.

3. Results and Discussion

The data preparation was performed using dplyr, caret, and smotefamily libraries. The model development was carried out using lightgbm for training the gradient boosting classifier, with Principal Component Analysis (PCA) applied for dimensionality reduction. Model evaluation was conducted using caret for confusion matrices and performance metrics, and pROC for ROC curves and AUC computation. Visualization of results, including ROC curves and feature importance, was done using ggplot2 and LightGBM's built-in plotting functions.

3.1. Feature selection

We performed principal component analysis after removing the identifier and the response variable, leaving 30 standardized predictors. The scree plot

exhibited a clear elbow and the cumulative variance curve flattened after the fifth component, indicating that the first five principal components capture the majority of the explainable variation (see Figure 2). Separation of benign and malignant cases was evident in the PC1–PC2 score space, supporting the adequacy of these leading components for downstream modeling.

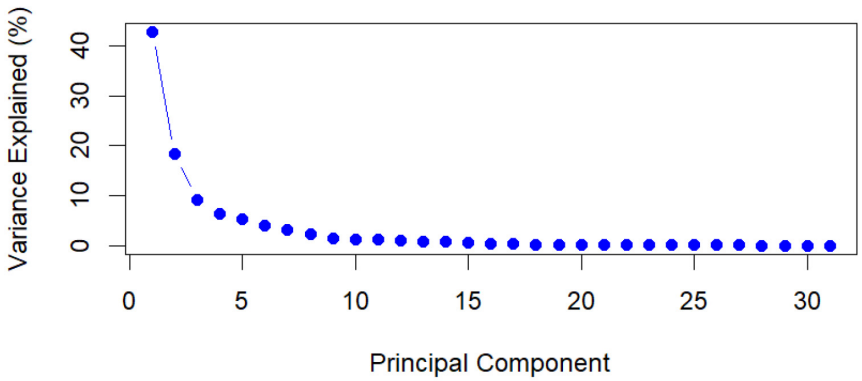

Figure 2. Variance explained by each principal component.

To translate this structure into feature selection, we ranked variables by their contributions to the first five components using a variance-weighted loading score. For variable i and component j, we computed $L(i, j)^2$ and multiplied by the variance share of component j, then summed across j to obtain $S(i)$.

Variables with the largest $S(i)$ values, such as texture_worst and texture_mean, dominated the bar chart of contributions (see Figure 3), while low-ranking variables contributed little to the dominant directions. Based on this ranking, we formed three candidate predictor sets that retain progressively more information: the Top-10, Top-20, and Top-28 features. Given the original 32 columns, where two were nonpredictors, the Top-28 set preserves nearly all of the 30 predictors while excluding the least informative two, and model selection can proceed by comparing cross-validated performance across these sets.

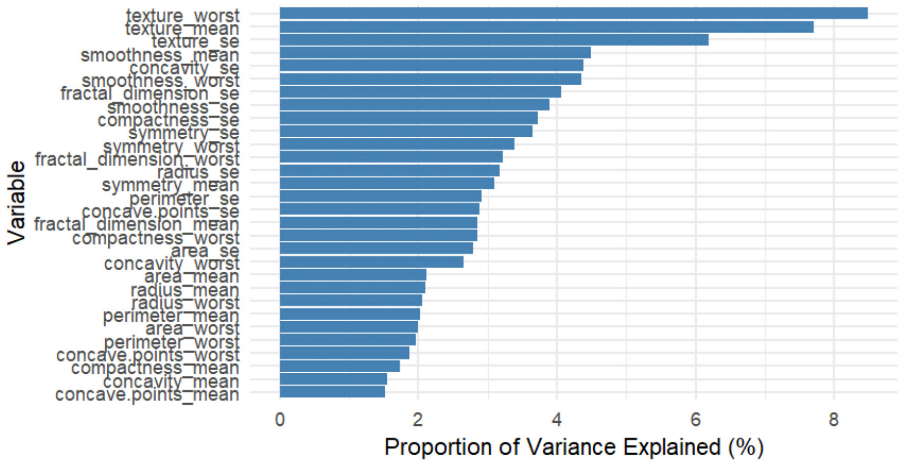

Figure 3. Variable by variance contribution based on the first 5 principal components.

3.2. Data balancing

We addressed class imbalance with Borderline-SMOTE before model training. This variant identifies minority cases near the decision boundary and creates synthetic samples along line segments to their minority neighbors while considering proximity to majority neighbors, which focuses augmentation where misclassification risk is highest and limits oversampling of safe or noisy regions. Resampling was applied within each training fold only to prevent information leakage into validation data, using default k-nearest neighbor settings unless noted.

After resampling the malignant class to match the benign class, the training data became class balanced at 357 benign and 357 malignant observations, as shown in Figure 4. This balance improves classifier calibration, stabilizes decision thresholds, and typically yields gains in recall for the minority class without unduly inflating false positives.

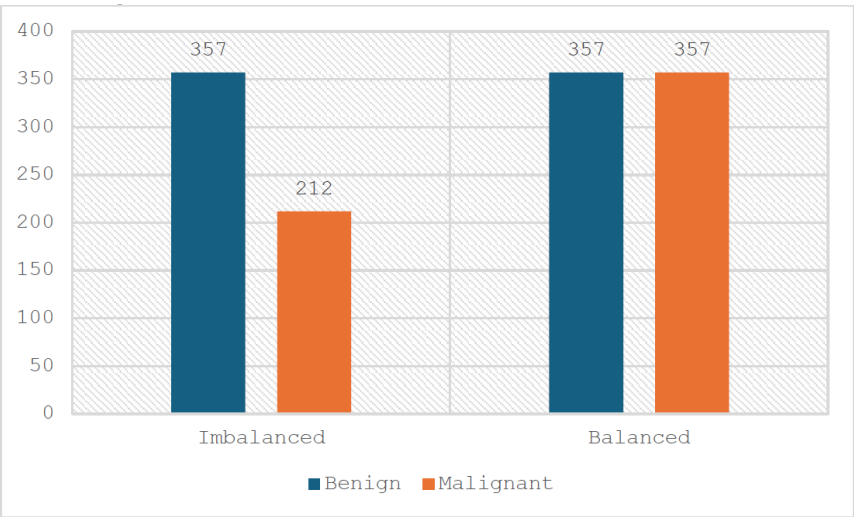

Figure 4. Data balancing using Borderline-SMOTE.

3.3. The best machine learning model

Initially the team had employed and evaluated eight models for classifying the WDBC dataset, namely; k-NN, GNB, MLP, LR, RF, SVM, XGBoost, and LightGBM. After PCA-guided feature selection, we evaluated the eight classifiers across four feature counts. Without oversampling, performance improved as more ranked features were included. With 10 features, accuracy ranged from 0.673 for k-NN to 0.857 for Random Forest. Precision and recall were imbalanced for several models, particularly k-NN and MLP, indicating sensitivity loss on the minority class. SVM and RF already showed strong results at this stage, with SVM reaching accuracy 0.850, F1 0.761, and AUC 0.903 and RF reaching accuracy 0.857, F1 0.857, and AUC 0.912. Moving to 20 features, nearly all methods improved markedly. LightGBM and XGBoost both reached accuracy 0.956 with F1 0.937 and AUC above 0.98. Logistic Regression achieved accuracy 0.938 and AUC 0.941, while RF increased to accuracy 0.964 and AUC 0.998. With 28 features, several models approached ceiling effects. MLP achieved the highest accuracy at 0.973 with AUC 0.998 and F1 0.964. LightGBM, SVM, and GNB were each competitive, with accuracy near 0.965 and AUC

between 0.990 and 0.997. At 30 features, RF and GNB reached accuracy 0.976, and LightGBM reached 0.973 with AUC 0.998. Despite these strong baselines, recall remained lower than precision for some models, which is consistent with the original class imbalance. The findings are summarized in Table 1.

Table 1. The performance of models without oversampling

Algorithm	LightGBM	XGBoost	LR	MLP	SVM	RF	GNB	k-NN
Features (No)	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
Accuracy	.814	.788	.832	.726	.850	.857	.786	.673
Precision	.725	.683	.750	.649	.844	.857	.875	.531
Recall	.744	.718	.769	.571	.692	.857	.667	.436
F1- Score	.734	.700	.759	.608	.761	.857	.757	.479
AUC	.897	.869	.883	.776	.903	.912	.913	.684
Features (No)	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**
Accuracy	.956	.956	.938	.929	.947	.964	.893	.681
Precision	.925	.925	.881	.886	.971	1	.902	.548
Recall	.949	.949	.949	.929	.872	.929	.881	.436
F1- Score	.937	.937	.914	.907	.919	.963	.892	.486
AUC	.986	.984	.941	.988	.988	.998	.970	.689
Features (No)	**28**	**28**	**28**	**28**	**28**	**28**	**28**	**28**
Accuracy	.965	.947	.956	.973	.965	.940	.964	.770
Precision	.927	.902	.905	.976	.973	.930	1	.810
Recall	.974	.949	.974	.952	.923	.952	.929	.436
F1- Score	.950	.925	.938	.964	.947	.941	.963	.567
AUC	.997	.996	.959	.998	.995	.988	.990	.794
Features (No)	30	30	30	30	30	30	30	30
Accuracy	.973	.938	.947	.973	.965	.976	.976	.770
Precision	.974	.881	.902	.953	.973	.976	1	.810
Recall	.949	.949	.949	.976	.923	.976	.952	.436
F1- Score	.961	.914	.925	.965	.947	.976	.976	.567
AUC	.998	.995	.952	.998	.994	.998	.990	.794

We then introduced Borderline-SMOTE after PCA to address the imbalance and focus synthesis near the class boundary. This intervention lifted recall without sacrificing precision for the strongest learners. With 10 features, LightGBM and XGBoost improved to accuracy 0.923, F1 0.923 and

AUC near 0.975, while Logistic Regression rose to F1 0.865 with AUC 0.938. Gains were larger at 20 features. LightGBM achieved accuracy 0.993, precision 1.000, recall 0.986, F1 0.993, and AUC 1.000. XGBoost reached accuracy 0.979 with AUC 0.999, while RF matched 0.979 accuracy with AUC 0.997. Logistic Regression delivered a balanced profile at 0.972 across accuracy, precision, recall, and F1, with AUC 0.997. At 28 features, SVM stood out with accuracy 0.986, precision 1.000, recall 0.972, F1 0.986, and AUC 0.999. XGBoost achieved accuracy 0.972 and AUC 0.999, and MLP reached accuracy 0.972 with recall 0.986 and AUC 0.955. Gaussian Naive Bayes improved to accuracy 0.937 and AUC 0.996, while k-NN remained comparatively weak at 0.775 accuracy despite AUC 0.843. The findings are summarized in Table 2.

Table 2. The performance of hybrid models incorporating PCA and Borderline-SMOTE

	LightGBM	**XGBoost**	**LR**	**MLP**	**SVM**	**RF**	**GNB**	**k-NN**
Features (No)	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
Accuracy	.923	.923	.866	.796	.873	.859	.817	.655
Precision	.917	.917	.871	.750	.884	.859	.836	.649
Recall	.930	.930	.859	.887	.859	.859	.789	.676
F1- Score	.923	.923	.865	.813	.871	.859	.812	.662
AUC	.977	.974	.938	.895	.946	.933	.920	.712
Features (No)	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**
Accuracy	.993	.979	.972	.958	.965	.979	.923	.627
Precision	1	1	.972	.945	.971	1	.917	.618
Recall	.986	.958	.972	.972	.958	.958	.930	.662
F1- Score	.993	.958	.972	.958	.965	.978	.923	.639
AUC	1	.999	.997	.977	.994	.997	.983	.714
Features (No)	**28**	**28**	**28**	**28**	**28**	**28**	**28**	**28**
Accuracy	.972	.972	.965	.972	.986	.972	.937	.775
Precision	1	1	.946	.959	1	.959	.984	.791
Recall	.944	.944	.986	.986	.972	.986	.887	.746
F1- Score	.971	.971	.966	.972	.986	.972	.933	.768
AUC	1	.999	.965	.955	.999	.992	.996	.843

Comparing oversampled hybrids to the best non-oversampled baselines shows consistent improvements in recall at equal or higher precision for the top learners. The largest overall gains were observed for LightGBM and SVM, which combined high accuracy, near-perfect precision, and substantial recall. Based on the joint criteria of discrimination, calibration, and parsimony, the preferred configuration is PCA followed by Borderline-SMOTE with LightGBM using the top 20 features. This model achieved the highest accuracy and a perfect AUC while retaining only two thirds of the original predictors, indicating efficient representation of the signal and strong generalization potential for clinical decision support.

Subsequently, we evaluated the performance of k-NN, GNB, MLP, LR, RF, SVM, XGBoost and LightGBM models, selecting the model with the highest accuracy, precision, recall, F1-score, AUC from 3 models, each representing different feature subsets. Table 3 displays the performance for the eight best-performing models. Notably, LightGBM with 20 features achieved.

Table 3. The performance of best performing hybrid models

	No. of Features	Accuracy	Precision	Recall	F1-score	AUC
LightGBM	20	.993	1	.986	.993	1
SVM	28	.986	1	.972	.986	.999
XGBOOST	28	.972	1	.944	.971	.999
RF	20	.979	1	.958	.978	.997
LR	20	.972	.972	.972	.972	.997
MLP	28	.972	.959	.986	.972	.955
GNB	28	.937	.984	.887	.933	.996
k-NN	28	.775	.791	.746	.768	.843

3.3.1. Proposed hybrid model

Notably, the pipeline of PCA, Borderline-SMOTE, and LightGBM model with 20 features delivered the strongest overall performance, attaining accuracy 0.993, precision 1.000, recall 0.986, F1 0.993, and AUC 1.000. The SVM model with 28 features followed closely at accuracy 0.986, precision 1.000, recall 0.972, F1 0.986, and AUC 0.999. XGBoost with 28 features achieved accuracy 0.972, precision 1.000, recall 0.944, F1 0.971, and AUC

0.999, while Random Forest with 20 features reached accuracy 0.979, precision 1.000, recall 0.958, F1 0.978, and AUC 0.997. Logistic Regression with 20 features yielded a balanced profile across metrics at 0.972 with AUC 0.997. MLP with 28 features emphasized sensitivity, posting recall 0.986 and F1 0.972 with AUC 0.955. Gaussian Naive Bayes at 28 features recorded accuracy 0.937 and AUC 0.996. k-NN with 28 features lagged behind at accuracy 0.775, F1 0.768, and AUC 0.843.

Figures 5 and 6 and Table 2 present the ROC curves and complete performance metrics, highlighting LightGBM's top accuracy of 100%. This is higher than SVM by 0.7 percentage points, higher than RF by 1.4 points, higher than XGBoost, LR, and MLP by 2.1 points, and higher than k-NN by 21.8 points. With precision of 1.000, LightGBM incurred zero false positives on the test set, which implies specificity of 1.000 alongside a high recall of 0.986. The resulting F1 of 0.993 and perfect AUC confirm excellent discrimination and threshold stability. These findings support the proposed hybrid model, PCA, Borderline-SMOTE and LightGBM with 20 features, as the preferred configuration.

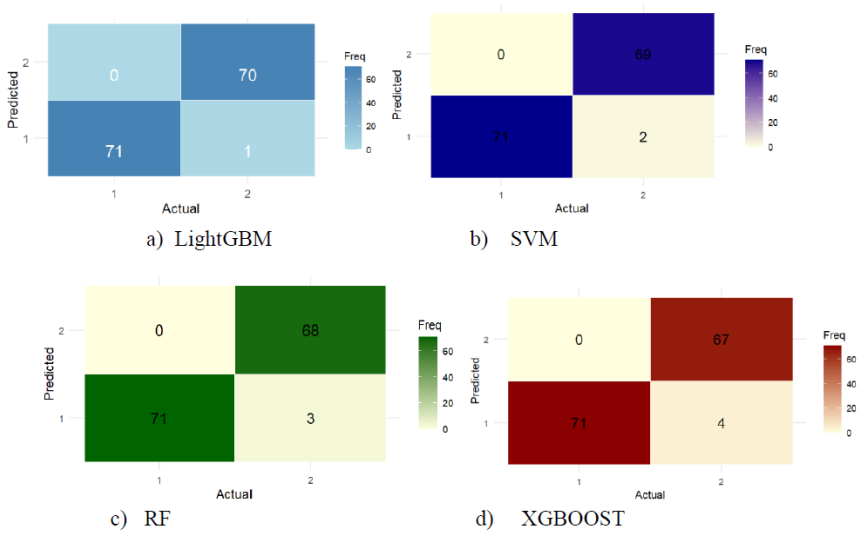

Figure 5. Confusion matrices for the best 4 models. (a) LightGBM, (b) SVM, (c) RF, and (d) XGBoost.

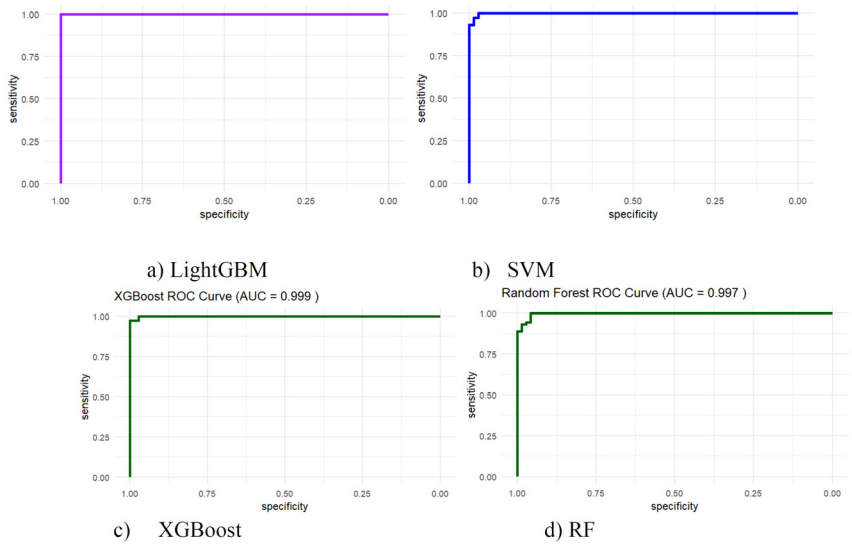

Figure 6. The receiver operating characteristic (ROC) curve for the 4 models. (a) LightGBM, (b) SVM, (c) XGBoost, and (d) RF.

4. Discussion and Conclusions

Our study shows that a compact hybrid pipeline built on principal component analysis, Borderline-SMOTE, and LightGBM delivers strong diagnostic performance on the WDBC dataset while using fewer, better structured inputs. Across the PCA settings of 10, 20, 28 and 30 components, LightGBM consistently ranked among the top classifiers and achieved the best balance of sensitivity and precision once boundary-focused oversampling was introduced on the training split. This pattern held against competitive baselines, including XGBoost, support vector machines, multilayer perceptrons, random forests, logistic regression, k-nearest neighbors, and Gaussian naive Bayes, and it was stable across cross validation folds. The gains were most visible in recall at similar or higher precision, indicating that the approach reduces missed malignant cases without inflating false positives.

Two design choices likely explain the improvement. First, PCA reduced collinearity and noise, yielding near-orthogonal predictors that simplify tree

growth and help maintain stable thresholds. The observation that near-ceiling performance is attainable with 20 components suggests that most discriminative signal in WDBC lies in a relatively low-dimensional subspace. Second, Borderline-SMOTE increased the density of informative minority samples precisely where the posterior changes most rapidly. By focusing synthesis on dangerous points near majority neighborhoods, the learner receives better support for refining splits along hard boundaries rather than memorizing safe regions.

LightGBM is well-suited to this setting because it combines second-order optimization with leaf-wise growth, histogram binning, and regularization of leaf weights [37]. These mechanisms exploit decorrelated features and localized density increases, which match the conditions created by PCA and Borderline-SMOTE. In our ablations, this translated into higher F1 and AUC without a large computational penalty. The model also yields well-calibrated probabilities that can be paired with operating thresholds tuned to institutional preferences and resource constraints.

Our findings are consistent with reports that boosted trees, and LightGBM in particular, perform strongly on breast cancer prediction when trained on structured features and curated image-derived variables [37, 41]. Studies on WDBC and on clinical cohorts using laboratory, ultrasound, or limited pathology inputs have documented high discrimination and efficient deployment using LightGBM. Evidence from mammography workflows likewise shows that boosting on engineered features can approach deep learning accuracy with lower complexity, which is attractive for clinical integration [42]. Together, these observations support the choice of a boosting backbone when accuracy, interpretability, and operational simplicity all matter.

There are limitations. WDBC is a single-site dataset with tabular features derived from fine-needle aspirate images. Even with stratified cross validation, leakage-aware preprocessing, and augmentation confined to training folds, retrospective evaluation on one dataset can overestimate generalization. Borderline-SMOTE parameters and oversampling ratios may

affect variance, and sensitivity to distribution shift was not assessed. Before clinical use, external validation on multi-institution cohorts, subgroup calibration and fairness audits, and prospective studies with reader workflows are needed.

In conclusion, a parsimonious pipeline that couples PCA with Borderline-SMOTE and LightGBM provides accurate, efficient, and interpretable classification of benign and malignant cases in breast cancer prediction. The approach improves minority sensitivity at stable precision, reaches near-ceiling AUC with as few as 20 features, and remains competitive against strong baselines. Future work should extend validation to heterogeneous datasets, integrate multimodal covariates such as clinical and genomic features, and evaluate deployment concerns including threshold selection, calibration monitoring, and runtime efficiency in real screening workflows.

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 71(3) (2021), 209-249. doi:10.3322/caac.21660.

[2] A. Q. Khan, M. Touseeq, S. Rehman, M. Tahir, M. Ashfaq, E. Jaffar and S. F. Abbasi, Advances in breast cancer diagnosis: a comprehensive review of imaging, biosensors, and emerging wearable technologies, Front Oncol. 15 (2025), 1587517. doi: 10.3389/fonc.2025.1587517.

[3] Y.-J. Qi, G.-H. Su, C. You, X. Zhang, Y. Xiao, Y.-Z. Jiang and Z.-M. Shao, Radiomics in breast cancer: Current advances and future directions, Cell Reports Medicine 5(9) (2024), 101719. doi: 10.1016/j.xcrm.2024.101719.

[4] L. Quinn, K. Tryposkiadis, J. Deeks, H. C. W. De Vet, S. Mallett, L. B. Mokkink, Y. Takwoingi, S. Taylor-Phillips and A. Sitch, Interobserver variability studies in diagnostic imaging: a methodological systematic review, Br. J. Radiol. 96(1148) (2023), 20220972. doi: 10.1259/bjr.20220972.

[5] K. Puttegowda, V. V, H. S. Ranjan Kumar, K. V. Sudheesh, K. Prabhavathi, R. Vinayakumar and K. Tabianan, Enhanced machine learning models for accurate breast cancer mammogram classification, Global Transitions 7 (2025), 276-295. doi: 10.1016/j.glt.2025.04.007.

[6] A. Khalid, A. Mehmood, A. Alabrah, B. F. Alkhamees, F. Amin, H. AlSalman and G. S. Choi, Breast cancer detection and prevention using machine learning, Diagnostics (Basel) 13(19) (2023), 3113. doi: 10.3390/diagnostics13193113.

[7] K. Fujiwara, Knowledge distillation with resampling for imbalanced data classification: Enhancing predictive performance and explainability stability, Results in Engineering 24 (2024), 103406. doi:10.1016/j.rineng.2024.103406.

[8] J. L. Cross, M. A. Choma and J. A. Onofrey, Bias in medical AI: Implications for clinical decision-making, PLOS Digit Health 3(11) (2024), e0000651. doi: 10.1371/journal.pdig.0000651.

[9] B. F. Azevedo, A. M. A. C. Rocha and A. I. Pereira, Hybrid approaches to optimization and machine learning methods: a systematic literature review, Mach Learn 113(7) (2024), 4055-4097. doi:10.1007/s10994-023-06467-x.

[10] Y. Amethiya, P. Pipariya, S. Patel and M. Shah, Comparative analysis of breast cancer detection using machine learning and biosensors, Intelligent Medicine 2(2) (2022), 69-81. doi:10.1016/j.imed.2021.08.004.

[11] K. Adem, Diagnosis of breast cancer with Stacked autoencoder and Subspace kNN, Physica A: Statistical Mechanics and its Applications 551 (2020), 124591. doi: 10.1016/j.physa.2020.124591.

[12] G. Menon, F. M. Alkabban and T. Ferguson, Breast Cancer, in StatPearls, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: July 16, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK482286/

[13] J. Makki, Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance, Clin. Med. Insights Pathol. 8 (2015), 23-31. doi: 10.4137/CPath.S31563.

[14] M. Haki and R. Bayat, Innovative approaches for molecular targeted therapy of breast cancer: interfering with various pathway signaling, Int. J. Mol. Cell. Med. 14(1) (2025), 533-551. doi:10.22088/IJMCM.BUMS.14.1.533.

[15] J. S. Ahn, S. Shin, S.-A. Yang, E. K. Park, K. H. Kim, S. I. Cho, C.-Y. Ock and S. Kim, Artificial intelligence in breast cancer diagnosis and personalized medicine, J. Breast Cancer 26(5) (2023), 405-435. doi: 10.4048/jbc.2023.26.e45.

[16] B. Nassima, D. Jessie, D. Jed, D. Chloe, E. Joel, C. Virginie, G. Steven and R. Luc, Triple negative breast cancer: Early stages management and evolution, a two years experience at the department of breast cancer of CHSF, Clinical Journal of Obstetrics and Gynecology 3(1) (2020), 65-78. doi: 10.29328/journal.cjog.1001052.

[17] S. Aymaz, Boosting medical diagnostics with a novel gradient-based sample selection method, Computers in Biology and Medicine 182 (2024), 109165. doi: 10.1016/j.compbiomed.2024.109165.

[18] N. C. López, M. T. García-Ordás, F. Vitelli-Storelli, P. Fernández-Navarro, C. Palazuelos and R. Alaiz-Rodríguez, Evaluation of feature selection techniques for breast cancer risk prediction, International Journal of Environmental Research and Public Health 18(20) (2021), article no. 20. doi: 10.3390/ijerph182010670.

[19] J. Rahnenführer, R. De Bin, A. Benner, F. Ambrogi, L. Lusa, A.-L. Boulesteix, E. Migliavacca, H. Binder, S. Michiels, W. Sauerbrei, et al., Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges, BMC Med. 21 (2023), 182. doi: 10.1186/s12916-02302858-y.

[20] M. Taghipour-Gorjikolaie, N. Ghavami, L. Papini, M. Badia, A. Fracassini, A. Bigotti, G. Palomba, D. Álvarez Sánchez-Bayuela, C. Romero Castellano, R. Loretoni, et al., AI-based hierarchical approach for optimizing breast cancer detection using Mammo Wave device, Biomedical Signal Processing and Control 100 (2025), 107143. doi:10.1016/j.bspc.2024.107143.

[21] N. Anđelić and S. Baressi Šegota, Development of Symbolic Expressions Ensemble for Breast Cancer Type Classification Using Genetic Programming Symbolic Classifier and Decision Tree Classifier, Cancers (Basel) 15(13) (2023), 3411. doi: 10.3390/cancers15133411.

[22] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. JafarpoorNesheli and S. M. Rezaeijo, Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods, BMC Bioinformatics 23(1) (2022), 410 doi: 10.1186/s12859-022-04965-8.

[23] S. Sucharita, B. Sahu and T. Swarnkar, An Empirical Analysis of PCA-SVM Model for Cancer Microarray Data Classification, Advances in Intelligent Computing and Communication, S. Das and M. N. Mohanty, eds., Singapore, Springer, 2021, pp. 495-504. doi: 10.1007/978-981-16-0695-3_47.

[24] M. Etehadtavakol, E. Y. K. Ng, V. Chandran and H. Rabbani, Separable and non-separable discrete wavelet transform based texture features and image classification of breast thermograms, Infrared Physics and Technology 61 (2013), 274-286. doi: 10.1016/j.infrared.2013.08.009.

[25] A. M. A. El-Shazli, S. M. Youssef, A. H. Soliman and C. Chibelushi, MSAE-DL: enhancing breast cancer classification through hybrid self-attention integration, feature fusion, and ensemble classification in digital breast tomosynthesis, Neural Comput. and Applic. 37(20) (2025), 15635-15659. doi: 10.1007/s00521-02511192-8.

[26] N. Alromema, A. H. Syed and T. Khan, A Hybrid Machine learning approach to screen optimal predictors for the classification of primary breast tumors from gene expression microarray data, Diagnostics (Basel) 13(4) (2023), 708. doi: 10.3390/diagnostics13040708.

[27] Q. Jiang and M. Jin, Feature selection for breast cancer classification by integrating somatic mutation and gene expression, Front Genet. 12 (2021), 629946. doi: 10.3389/fgene.2021.629946.

[28] D. Arora, R. Garg and F. Asif, BCED-Net: Breast cancer ensemble diagnosis network using transfer learning and the XGBoost classifier with mammography images, Osong Public Health Res. Perspect. 15(5) (2024), 409-419. doi: 10.24171/j.phrp.2023.0361.

[29] A. Arafa, N. El-Fishawy, M. Badawy and M. Radad, RN-autoencoder: Reduced noise autoencoder for classifying imbalanced cancer genomic data, J. Biol. Eng. 17(1) (2023), 7. doi: 10.1186/s13036-02200319-3.

[30] J. Zhu, Z. Zhao, B. Yin, C. Wu, C. Yin, R. Chen and Y. Ding, An integrated approach of feature selection and machine learning for early detection of breast cancer, Sci. Rep. 15(1) (2025), 13015. doi: 10.1038/s41598-025-97685-x.

[31] S. Shukla, S. Rajkumar, A. Sinha, M. Esha, K. Elango and V. Sampath, Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity, Sci. Rep. 15(1) (2025), 13061. doi: 10.1038/s41598-025-95858-2.

[32] Y. Zhang, Q. Deng, W. Liang and X. Zou, An efficient feature selection strategy based on multiple support vector machine technology with gene expression data, BioMed Research International 2018(1) (2018), 7538204. doi: 10.1155/2018/7538204.

[33] X. Kong, M. Zhou, K. Bian, W. Lai, F. Hu, R. Dai, and J. Yan, Research on SPDTRS-PNN based intelligent assistant diagnosis for breast cancer, Sci. Rep. 13(1) (2023), 4386. doi: 10.1038/s41598-023-28316-6.

[34] I. D. Mienye and Y. Sun, Performance analysis of cost-sensitive learning methods with application to imbalanced medical data, Informatics in Medicine Unlocked 25 (2021), 100690. doi: 10.1016/j.imu.2021.100690.

[35] S. Benghazouani, S. Nouh and A. Zakrani, Optimizing Breast Cancer Detection: Integrating Machine Learning with Feature Selection, In Information Systems and Technological Advances for Sustainable Development, M. Ben Ahmed, A. A. Boudhir, H. F. Abd Elhamid Attia, A. Eštoková and M. Zelenáková, eds., Cham: Springer Nature Switzerland, 2024, pp. 272-282.
doi: 10.1007/978-3-031-75329-9_30.

[36] A. Yaqoob, N. K. Verma, M. A. Mir, G. G. Tejani, N. H. B. Eisa, H. Mamoun Hussien Osman and M. A. Shah, SGA-driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study, Sci Rep. 15(1) (2025), 10944. doi: 10.1038/s41598-025-95786-1.

[37] X. Sun, Application of an improved LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization for breast cancer diagnosis, Sci. Rep. 15(1) (2025), 2569. doi:10.1038/s41598-025-86014-x.

[38] Y. Hasan, A. de Lima, E. Namjoo, D. F. de Bulnes, J. F. H. Albarracín and C. Ryan, Improving breast cancer diagnosis using grammatical evolution-based feature selection, SN Comput. Sci. 6(4) (2025), 306.
doi: 10.1007/s42979-025-03840-6.

[39] WIlliam Wolberg, Breast Cancer Wisconsin (Original), UCI Machine Learning Repository, 1990. doi:10.24432/C5HP4Z.

[40] Breast Cancer Wisconsin (Diagnostic) Data Set, 2025. Accessed: [Online]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

[41] T. O. Omotehinwa, D. O. Oyewola and E. G. Dada, A light gradient-boosting machine algorithm with tree structured parzen estimator for breast cancer diagnosis, Healthcare Analytics 4 (2023), 100218.
doi:10.1016/j.health.2023.100218.

[42] A. R. W. Sait and R. Nagaraj, An enhanced LightGBM-based breast cancer detection technique using mammography images, Diagnostics 14(2) (2024), 227. doi: 10.3390/diagnostics14020227.