

Research Article:

High-Dimensional Breast Cancer Classification: Evaluating Trade-offs Between Accuracy, Robustness, and Computational Cost

Ali Hamad Ali¹, Naaman Omar Yaseen²

¹Information Technology Department, Technical College of Duhok, Duhok Polytechnic University, Iraq

²Information Technology Department, Technical Institute of Amedi, Duhok, Duhok Polytechnic University, Iraq

<https://doi.org/10.31530/cjnst.2026.2.2.7>

Received: 04 March 2026

Revised: 05 June 2026

Accepted: 07 June 2026

*Corresponding author info: Ali Hamad Ali,

Email: ali.ali@dpu.edu.krd

Copyright © 2026 Ali Hamad Ali, Naaman Omar Yaseen, Charmo Journal of Natural Sciences and Technologies (CJNST).

This is an open access article published under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.



Abstract

Background: Breast cancer is the leading cause of cancer-related mortality in females, with 2.3 million new cases diagnosed annually. Machine learning (ML) algorithms have the potential to improve diagnostic accuracy through the analysis of high-dimensional data. However, the lack of standardized benchmarks across different algorithmic families impedes informed model selection for clinical deployment.

Aim: This study addresses critical gaps in understanding the trade-off between classification performance and computational efficiency factors that are essential for implementing algorithms in resource-constrained healthcare settings. The research aims to compare the performance of different ML algorithms for breast cancer classification in a high-dimensional feature space.

Patients and Methodology: The performance of 13 ML algorithms including conventional classifiers, ensemble methods, Gradient Boosting (GB), neural networks (NNs), and meta-ensembles was evaluated on the Database for Mastology Research with Infrared (DMR-IR) multimodal breast imaging dataset comprising 5,794 instances and 1,280 features. After excluding unconfirmed cases, the final analysis included 5,602 instances. A hybrid feature extraction pipeline combined deep features from three pre-trained convolutional neural networks (VGG16, ResNet50, DenseNet121) with handcrafted color, texture, and region-of-interest descriptors. Performance was assessed using accuracy, F1-score, and training time on a stratified 80:20 split with 5-fold cross-validation.

Results: Ensemble methods significantly outperformed conventional classifiers. The Voting ensemble achieved the highest accuracy (99.62%) and F1-score (0.9962), followed by Stacking (99.58%) and Extreme Gradient Boosting (XGBoost) (99.58%). The Histogram Gradient Boosting (HGB) classifier demonstrated the optimal balance between accuracy (99.54%) and training time (5.98 seconds). Ensemble algorithms outperformed conventional classifiers by 9-10% in accuracy. Notably, the computationally efficient k-nearest neighbors (KNN) algorithm achieved 98.59% accuracy with only 0.02 seconds training time when applied to the hybrid feature representation.

Conclusion: Ensemble methods demonstrate superior performance for high-dimensional breast cancer classification, with the Voting ensemble achieving maximum accuracy and HGB offering optimal practicality. Preliminary evidence-based model selection considerations are provided for different deployment contexts, from resource-constrained primary care to centralized reference centers.

Keywords: Breast Cancer Detection, Machine Learning, Ensemble Methods, Computational Efficiency, Multimodal Imaging.

1. Introduction

Breast cancer remains the most commonly diagnosed malignancy among women worldwide and the leading cause of cancer-related mortality in this population, with approximately 2.3 million new cases diagnosed annually [1]. The relationship between early detection and improved survival has been firmly established: five-year survival rates exceed 95% for localized disease but drop to approximately 25% for metastatic disease [2]. This disparity underscores the critical importance of accurate, accessible, and timely diagnostic screening.

Current diagnostic protocols rely primarily on imaging modalities—mammography, ultrasound, magnetic resonance imaging, and histopathology—each with characteristic strengths and limitations [3]. Mammography interpretation exhibits substantial inter-observer variability, with sensitivity ranging from 62% to 87% depending on breast density and radiologist expertise [3]. Histopathological grading, while considered the gold standard, is inherently subjective, with discordance rates of 25–30% documented between pathologists examining the same specimens [1], [2]. These limitations are compounded by resource constraints that impede population-level screening implementation, particularly in transitioning healthcare systems where specialist radiologists and pathologists are scarce [22].

ML offers the potential to transcend these limitations by providing consistent, objective, and highly accurate analyses of medical images [13], [14]. The evolution of ML applications in breast cancer diagnostics has progressed from traditional statistical learning methods [6] through ensemble approaches [11], [17] to contemporary deep learning architectures [1], [21]. Convolutional neural networks, in particular, have revolutionized medical image analysis by automatically learning hierarchical feature representations directly from image data [1], [7].

Despite these advances, significant gaps persist in the literature. First, most studies operate within narrow methodological boundaries—comparing multiple variants of a single algorithmic family rather than spanning diverse paradigms [13], [14]. Second, accuracy dominates as the primary evaluation metric, while computational efficiency—critical for practical deployment—is rarely documented [15]. Third, multimodal datasets that pair complementary imaging modalities remain underutilized [20]. High-dimensional feature spaces present additional challenges. Medical imaging datasets often yield thousands of features per instance, creating the "curse of dimensionality": exponential expansion of hypothesis space, increased data sparsity, and heightened overfitting risk [8]–[10]. Class imbalance further complicates analysis, as pathological cases constitute a minority in real-world populations [6], [15]. This study addresses these gaps through four major contributions. First, an end-to-end preprocessing and feature engineering pipeline is presented that combines Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTEENN) for class imbalance handling with mutual information-based feature selection—a combination demonstrated to enhance generalization in high-dimensional spaces [15], [16]. Second, a comprehensive head-to-head comparison of 13 ML algorithms spanning five methodological families is conducted on the

DMR-IR dataset under identical experimental conditions, addressing the methodological fragmentation identified in the literature [13], [14]. Third, the trade-offs between classification accuracy and computational efficiency are systematically quantified, a dimension largely neglected in existing studies [15] providing preliminary evidence-based model selection considerations for different deployment contexts. Fourth, a hybrid feature extraction technique combining deep features from multiple Convolutional Neural Network (CNN) architectures with handcrafted color, texture, and Region of Interest (ROI) descriptors enables multiple classifiers to achieve state-of-the-art performance exceeding 99.5% accuracy, thereby establishing a new performance baseline for the DMR-IR dataset and validating the value of multimodal integration [20].

2. Related Work

2.1. Evolution of ML in Breast Cancer Diagnostics

The application of ML to breast cancer detection has evolved substantially over the past two decades. Early efforts focused on traditional statistical learning methods operating on handcrafted features extracted from clinical data and basic image characteristics [3]. Logistic regression (RL) and linear discriminant analysis offered simplicity and theoretical grounding [3]. Support vector machines gained popularity due to their ability to create non-linear boundaries using the kernel trick, enabling good performance on moderately complex classification problems [13].

Ensemble methods marked a significant paradigm shift. Random forests (RF) combine bootstrapping with random feature selection to create decorrelated decision trees that reduce variance while maintaining low bias [6]. GB methods, including AdaBoost and XGBoost, construct ensembles sequentially, with each new model focusing on correcting the errors of its predecessors [11], [17]. These ensemble methods have proven exceptionally effective for biomedical classification problems, where feature interactions are complex and relationships are frequently non-linear [6], [11].

2.2. Deep Learning and Multimodal Approaches

Deep learning, particularly convolutional neural networks, has transformed medical image analysis by enabling end-to-end learning of hierarchical feature representations directly from pixel data [1], [21]. Architectures such as VGG16, ResNet50, and DenseNet121 have achieved accuracy exceeding 95% for histopathological image classification [1], [7] and improved sensitivity for mass detection in mammography [3]. Multimodal imaging has emerged as a promising direction for improving diagnostic accuracy. Thermography captures functional information about metabolic activity and vascularization through infrared emissions [19], complementing the structural information provided by mammography [3], and enhanced image quality for breast magnetic resonance imaging (MRI) at 1.5T [12]. The DMR-IR dataset uniquely pairs these modalities from the same patients, enabling investigation of multimodal integration strategies [20].

2.3. Hybrid Architectures and Ensemble Strategies

Hybrid architectures that combine the strengths of different approaches have gained increasing attention. Maleki et al. [5] proposed a DenseNet201-XGBoost hybrid for histopathological image classification, using deep features extracted by the CNN as inputs to the GB classifier. Balasubramanian et al. [7] demonstrated that ensemble deep learning improves robustness for breast cancer subtype classification from whole-slide images. Hameed Alsaedi and Akay [8] showed that feature-fused autoencoders combined with weighted decision trees achieve high accuracy while maintaining model simplicity. Similarly, Manali et al. [18] employed decision fusion over deep learning features to improve breast cancer detection accuracy.

2.4. Identified Research Gaps

Synthesis of the literature reveals several persistent gaps.

Methodological Fragmentation: Studies typically operate within narrow algorithmic families—comparing multiple CNN variants against each other or evaluating several ensemble methods against simple baselines—with broader comparisons across algorithmic families remaining scarce [13], [14].

Neglect of Computational Efficiency: Accuracy dominates as the primary evaluation metric, while computational considerations (training time, inference speed, hardware requirements) are rarely documented despite their practical importance for deployment [15].

Underutilization of Multimodal Datasets: Multimodal datasets like DMR-IR, which pair thermography and mammography images from the same patients, remain underutilized in research exploring potential synergies between functional and structural imaging [20].

Inadequate Imbalance Handling: Although class imbalance is widely recognized as a critical issue in medical datasets, systematic evaluation of imbalance correction methods across different model families is limited [6], [15]. The comparative summary of existing studies presented in Table 1.

Table 1: Comparative summary of existing studies

Ref.	Study Focus	Algorithms Used	Key Results	Identified Gaps
[1]	DL in breast histopathology	CNNs (VGG, ResNet)	>95% accuracy	Single modality; lacks multimodal fusion
[3]	Mammography with DL	CNN architectures	Improved sensitivity	Mammography only; no efficiency analysis
[4]	Breast MRI with DL meta-analysis	Various DL models	High AUC (>0.95)	No comparison with non-DL methods
[5]	DenseNet-XGBoost hybrid	DenseNet201 + XGBoost	98.2% accuracy	Single-algorithm hybrid; no broad benchmarking
[7]	Ensemble DL for histopathology	Ensemble CNNs	Improved robustness	Limited to DL ensembles only
[8]	Parallel channel DL system	Multi-channel CNN	Enhanced feature learning	No handcrafted feature integration
[9]	Multimodal fusion	DL fusion networks	AUC 0.97	Complex multimodal alignment required
[11]	Stacking ensemble methods	RF, SVM, NN stacking	Improved performance	Limited to stacking; no time complexity study
[16]	Thermography + DL	CNN on thermal images	Promising results	Single modality; no fusion with mammography
[19]	Thermography for breast cancer detection	Passive and active thermography (review)	Comparative evidence of thermography feasibility	Review only; no hybrid deep feature extraction

3. Methodology

This segment describes the entire experimental process utilized in this research, including the dataset, preprocessing/feature engineering, model development, and evaluation. This process is carefully constructed with the aim of being robust, reproducible, and balanced across various ML paradigms, in compliance with existing best practices in medical AI research [1], [6], [13], [15].

Figure 1. Schematic overview of the experimental workflow. The pipeline begins with the Dataset, followed by Preprocessing consisting of Hybrid Feature Extraction

and SMOTEENN + Feature Selection + Normalization. Data then proceed to Model Training and Classification across five classifier families: Linear/Distance LR, SVM, KNN], SGD), Tree-based (RF, Extra Trees (ET), GB (XGBoost, HGB, RF), NN (MLP, Deep Random Vector Functional Link (DeepRVFL), and Meta-Ensembles (Voting, Stacking). Model outputs are assessed in the Evaluation stage using Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), F1-Score, Training Time, stratified 5-fold cross-validation, learning curves, and model selection. The final Analysis stage consolidates the results

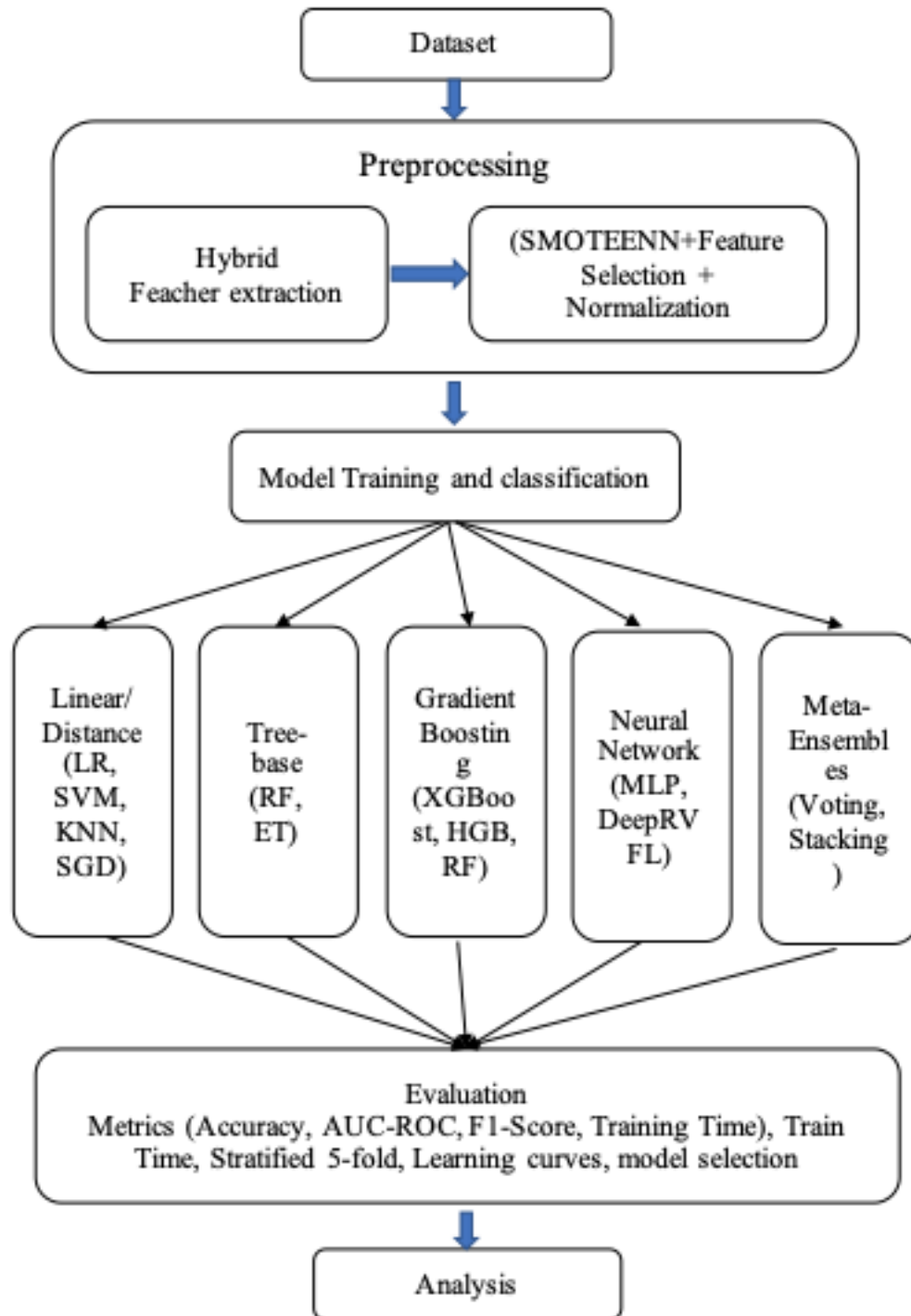


Figure 1: Schematic overview of the experimental workflow

3.1. Dataset Description

The DMR-IR dataset [20] comprises 5,794 images from clinical examinations conducted in Brazil, pairing thermography and mammography modalities from the same patients. The dataset contains three class labels: Healthy (approximately 72.5%), Sick (approximately 24.2%), and Unknown (approximately 3.3%). Unknown instances were excluded from the analysis as they lack confirmed diagnostic status and the study objective is binary classification

(healthy vs. sick). The final analysis was conducted on 5,602 instances after removing Unknown cases. Table 2 presents the dataset characteristics.

Table 2: DMR-IR dataset characteristics

Attribute	Description
Dataset name	DMR-IR (Breast Thermal Image Dataset)
Source	Clinical examinations, Brazil
Imaging modalities	Thermography (dynamic sequences, static images); Mammography
Class labels	Healthy, Sick (Unknown excluded)
Total instances	5,602 (after excluding Unknown)
Original feature dimensionality	1,280 (before selection)
Selected features (after MI)	100
Image dimensions (after resizing)	224 × 224 pixels
ROI dimensions	128 × 128 pixels

3.2. Data Preprocessing

A preprocessing pipeline was implemented to transform raw images into a structured format suitable for analysis. Automated discovery traverses the directory structure, parsing file paths to extract metadata (patient identifier, visit number, modality, diagnostic label) into a structured DataFrame. Image loading accommodates multiple formats and color spaces. Color images are converted from BGR to RGB; grayscale images are retained for texture analysis. All images are normalized to 224×224 pixels using bilinear interpolation, matching the input size expected by pre-trained CNN models and enabling consistent feature extraction.

3.3. ROI Segmentation

ROI segmentation isolates breast tissue from background using adaptive thresholding. For a grayscale image I , the adaptive threshold at pixel (x,y) is given by Equation (1):

$$T(x,y) = \mu_{window}(x,y) - C \quad (1)$$

Where $T(x,y)$ is the adaptive threshold at pixel (x,y) , $\mu_{window}(x,y)$ is the local mean within an 11×11 window and $C = 2$ is a constant. Contour detection identifies the largest connected component, and its bounding box extracts the ROI, resized to 128×128 pixels.

3.4. Feature Extraction Framework

The feature extraction methodology employs a hybrid approach combining deep learning-based features with traditional handcrafted descriptors, designed to capture complementary aspects of breast tissue characterization.

3.4.1. Deep Feature Extraction

Deep features are extracted using an ensemble of three pre-trained convolutional neural networks: VGG16, ResNet50, and DenseNet121. All networks are used with ImageNet-pre-trained weights as fixed feature extractors (no fine-tuning). From each network, the output of the last global average pooling layer is extracted, yielding a 256-dimensional feature vector per network. These three vectors are concatenated to form a 768-dimensional deep feature representation.

3.4.2. Handcrafted Feature Extraction

Three categories of handcrafted features are computed. Color Features: For each RGB channel, statistical moments (mean, standard deviation, skewness, kurtosis) and pairwise correlations between channels are computed. Histogram-based features (32 bins per channel) augment these statistics.

Texture Features: Local Binary Patterns (LBPs) ($P = 8, R = 1$) produce a 256-bin histogram. Sobel gradients provide magnitude and orientation histograms. Haar wavelet transform captures multi-scale texture information. Canny edge detection quantifies edge density and orientation. Principal Component Analysis (PCA) was applied to reduce dimensionality from approximately 400 to 200 features while preserving 95% of the variance. The PCA transformation was fitted exclusively on the training data and applied to both training and test sets.

ROI Features: Within the segmented breast region, intensity statistics (mean, median, standard deviation, variance, min, max, range), distribution shape (skewness, kurtosis), percentiles (25th, 50th, 75th), a 32-bin histogram, and ROI-specific LBP features are computed. Table 3 presents a detailed breakdown of the handcrafted features.

Table 3: Detailed breakdown of handcrafted features

Feature Category	Technique	Specific Features	Count
Color Features	RGB statistics	Mean, std dev, skewness, kurtosis (per channel)	12
	Cross-channel correlations (R-G, R-B, G-B)		3
	RGB histograms	32-bin histogram (per channel)	96
	Color Features Total	111 (padded to 200)	
Texture Features	LBPs	256-bin LBP histogram ($P=8, R=1$)	256

ROI Features	Sobel gradients	Gradient magnitude histogram (36 bins)	36
		Gradient orientation histogram (36 bins)	36
	Haar wavelet	Statistics at multiple scales	48
	Canny edges	Edge density, orientation histogram	24
	Texture Features Total	400 (reduced to 200 via PCA)	
	Intensity statistics	Mean, median, std dev, variance, min, max, range	7
	Distribution shape	Skewness, kurtosis	2
	Percentiles	25th, 50th, 75th percentiles	3
	ROI histogram	32-bin intensity histogram	32
	ROI LBPs	LBPs within ROI	68
	ROI Features Total	112 (padded to 200)	

3.4.3. Feature Integration

The final 1280-dimensional feature vector is obtained through the following concatenation process:

Step 1: Deep Features (768 dimensions) – Features from the three pre-trained CNN architectures are extracted and concatenated: VGG16 global average pooling layer: 256 features; ResNet50 global average pooling layer: 256 features; DenseNet121 global average pooling layer: 256 features; Total deep features: 768 dimensions.

Step 2: Color Features (200 dimensions) – Color statistics and histograms are computed as described in Section 3.4.2. Raw color features (111 dimensions) were augmented with additional color space transformations (HSV, LAB) and padded with zeros to achieve exactly 200 dimensions for consistent concatenation.

Step 3: Texture Features (200 dimensions) – Multiple texture analysis techniques (LBP, Sobel, Haar wavelet, Canny) produce approximately 400 raw texture features. PCA was applied to reduce dimensionality to 200 while preserving 95% of the variance. The PCA transformation was fitted exclusively on the training data and applied to both training and test sets.

Step 4: ROI Features (112 dimensions, padded to 200) – ROI-specific features are computed exclusively within the segmented breast region. To maintain uniform dimensionality across feature types, these are padded with zeros to reach 200 dimensions.

Final Concatenation: All feature vectors are concatenated as shown in Equation (2):

$$F_{total} = [F_{deep}(768) || F_{color}(200) || F_{texture}(200) || F_{roi}(200)] \in \mathbb{R}^{1280} \quad (2)$$

where F_{total} is the final 1,280-dimensional feature vector, F_{deep} (768 dimensions) denotes the concatenated deep features, F_{color} (200 dimensions) denotes the color features,

$F_{texture}$ (200 dimensions) denotes the texture features, F_{ROI} (200 dimensions) denotes the padded ROI features, and $||$ denotes vector concatenation.

This hybrid approach ensures that complementary information from multiple representation paradigms is preserved while maintaining a fixed-dimensional representation suitable for ML algorithms.

3.5. Feature Selection and Class Imbalance Handling

Mutual information-based feature selection reduces dimensionality from 1280 to 100 features. Mutual information quantifies statistical dependence between feature X and class label Y as shown in Equation (3):

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (3)$$

where $I(X; Y)$ is the mutual information between feature X and class label Y , $p(x, y)$ is the joint probability distribution of X and Y , and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y , respectively. Features are ranked by mutual information score, and the top 100 are retained for modeling.

Class imbalance is addressed using SMOTEENN, combining Synthetic Minority Oversampling Technique (SMOTE) with Edited Nearest Neighbors (ENN) cleaning. SMOTE generates synthetic minority samples through interpolation between existing instances. ENN removes samples whose class differs from the majority of their k nearest neighbors, cleaning noisy regions of the feature space.

3.6. Experimental Design

3.6.1. Model Selection

13 classifiers spanning five methodological families were selected as summarized in Table 4:

Table 4: Implemented ML models by category

Category	Models	Count
Linear / Distance-Based	LR, SVM, KNN, SGD	4
Tree-Based Ensembles	RF, ET	2
GB Methods	XGBoost, HGB, XGBRF	3
Neural Networks	MLP, Deep RVFL	2
Meta-Ensembles	Voting Classifier, Stacking Classifier	2
Total		13

Hyperparameters were selected based on literature recommendations and preliminary experiments and are represented in Table 5:

Table 5: Models hyperparameters

Model	Hyperparameters
RF	n_estimators = 200, max_depth = 10, class_weight = 'balanced'
XGBoost	n_estimators = 200, max_depth = 5, learning_rate = 0.1
HGB	max_iter = 200
MLP	hidden_layer_sizes = (28, 5), alpha = 0.1
Deep RVFL	n_layer = 3, n_nodes = 2000, lam = 0.01
Voting	Soft voting with HGB, MLP, KNN, XGBoost
Stacking	Base learners (RF, XGBoost, MLP), meta-learner (LogisticRegression), cv = 5

For scale-sensitive models (LR, SVM, KNN, SGD, MLP), StandardScaler was applied within a Pipeline to prevent data leakage.

3.6.2. Evaluation Protocol

Data was divided into stratified 80:20 train-test splits. Stratified 5-fold cross-validation was performed on the training set. Performance metrics included accuracy, precision, recall, F1-score, Cohen's kappa, and training time. Learning curves were constructed by incrementing training set size from 10% to 100% in 10% increments.

All preprocessing steps, including feature extraction, mutual information-based feature selection, and SMOTEENN balancing, were applied exclusively to the training set after the stratified 80:20 split to prevent data leakage. The trained feature selector and SMOTEENN parameters were subsequently applied to the test set without refitting.

4. Results

4.1. Key Findings Summary

- **Ensemble Superiority:** Top ensemble methods achieve 99.5-99.6% accuracy, outperforming conventional classifiers by 9-10%
- **Optimal Balance:** Histogram GB offers 99.54% accuracy with only 5.98 seconds training time

- **Simple Classifier Viability:** KNN achieves 98.59% accuracy with 0.02 seconds training time on hybrid features
- **Balanced Errors:** Top models exhibit equal false positive and false negative rates (clinically desirable)
- **Feature Complementarity:** Deep, color, texture, and ROI features all contribute discriminative information.

4.2. Exploratory Data Analysis

Application of SMOTEENN substantially altered the class distribution. The training set (4,635 instances) was transformed to 6,720 instances with balanced classes (50% healthy, 50% sick), eliminating the original 3:1 healthy:sick imbalance.

4.3. Feature Engineering Outcomes

Mutual information analysis revealed a long-tail distribution: few features possess high discriminative power (>0.3 MI score), while many contribute little. The top 20 features include a mix of deep features from all three architectures and handcrafted features (ROI statistics, color, texture), confirming that all feature families contribute valuable information.

4.4. Comprehensive Model Performance

Table 6 presents the comparative performance of all classifiers on the test set.

Table 6: Comparative performance of all classifiers on test set

Model	Accuracy	Precision	Recall	F1-Score	Kappa	Training Time (s)
Voting Ensemble	99.62	99.62	99.62	99.62	0.9962	27.91
Stacking Ensemble	99.58	99.58	99.58	99.58	0.9958	196.24
XGBoost	99.58	99.58	99.58	99.58	0.9958	12.33
Histogram GB	99.54	99.54	99.54	99.54	0.9954	5.98
Deep RVFL	99.24	99.24	99.24	99.24	0.9924	33.48
MLP	99.16	99.16	99.16	99.16	0.9916	9.68
RF	98.90	98.90	98.90	98.90	0.9890	25.67
KNN	98.59	98.59	98.59	98.59	0.9859	0.02
ET	98.36	98.37	98.36	98.37	0.9837	2.45
SVM	97.98	97.98	97.98	97.98	0.9798	11.89
XGBRF	92.96	92.93	92.96	92.92	0.9292	13.58

RL	90.38	90.36	90.38	90.35	0.9035	8.57
SGD	89.31	89.28	89.31	89.27	0.8927	0.74

Several key findings emerge:

- **Ensemble Superiority:** The top five classifiers are ensemble methods (Voting, Stacking, XGBoost, HGB) or neural networks with ensemble characteristics (Deep RVFL). This confirms the theoretical advantage of combining multiple models to reduce variance and improve generalization.
- **Meta-Ensemble Performance:** The Voting classifier achieves the highest overall accuracy (99.62%), marginally outperforming the more complex Stacking ensemble (99.58%). The simpler probability averaging approach achieves equivalent or better performance with substantially lower computational cost (27.91s vs. 196.24s).
- **GB Efficiency:** HGB achieves 99.54% accuracy with only 5.98 seconds training time, representing the optimal balance of accuracy and efficiency for practical deployment.
- **Neural Network Competitiveness:** Both MLP (99.16%) and Deep RVFL (99.24%) perform competitively, demonstrating that relatively simple neural architectures can achieve excellent results with well-engineered features.
- **Simple Classifier Viability:** KNN achieves 98.59% accuracy with only 0.02 seconds training time, underscoring the quality of the hybrid feature representation.
- **Performance Gap:** Ensemble methods outperform conventional classifiers (LR, SGD) by 9-10%, reflecting the inability of linear models to capture complex, non-linear relationships.

4.5. Statistical Significance

McNemar's test revealed that Voting, Stacking, XGBoost, and HGB are statistically indistinguishable from each other ($p > 0.05$), while significantly outperforming all other models ($p < 0.01$). This indicates that any of these four ensemble methods can be selected based on secondary criteria without sacrificing statistically significant performance.

4.6. Error Analysis

Top models exhibit nearly equal false positive and false negative rates (Voting: 3 vs. 3 errors out of 1,159 test cases). This balanced error distribution is clinically desirable, indicating no systematic bias toward overdiagnosis or underdiagnosis. The equalized error rates confirm that SMOTEENN successfully addressed the original class imbalance.

4.7. Learning Dynamics

Learning curves reveal that most models achieve $>97\%$ accuracy with only 20% of training data, indicating that the feature space is well-structured. The gap between training and validation curves is remarkably small for top models (0.001-0.003), demonstrating excellent generalization with minimal overfitting.

4.8. Computational Efficiency

Training times vary dramatically—from 0.02 seconds (KNN) to 196.24 seconds (Stacking)—among models with similar accuracy. Three operational regions are identified: Maximum Efficiency (0.01-1s): KNN (98.59%, 0.02s) offers excellent accuracy with virtually zero training time, suitable for resource-constrained settings.

Balanced Performance (1-30s): HGB (99.54%, 5.98s) offers optimal accuracy-efficiency ratio; Voting (99.62%, 27.91s) provides maximal accuracy at moderate cost.

Maximum Accuracy (30-200s): Stacking (99.58%, 196.24s) achieves top-tier accuracy at substantial computational cost, justifiable for one-time training of production models.

5. Discussion

5.1. Interpretation of Key Findings

5.1.1. Superiority of Ensemble Methods

The results provide clear evidence that ensemble methods outperform conventional classifiers for high-dimensional breast cancer classification. This finding aligns with both theoretical expectations and empirical literature [6], [11], [17]. Ensemble methods reduce variance through averaging (bagging-based methods) and reduce bias through sequential error correction (boosting-based methods). For the complex, non-linear relationships present in medical imaging data, these variance and bias reductions translate directly to improved generalization.

The near-equivalence of top ensemble methods (Voting, Stacking, XGBoost, HGB) suggests that for this dataset and feature representation, the specific ensemble architecture matters less than the ensemble principle itself. This has practical importance: practitioners can select among these methods based on secondary criteria without sacrificing performance.

5.1.2. The Accuracy-Efficiency Trade-off

The results reveal a clear trade-off landscape with distinct operational regions. The Pareto frontier of optimal models includes HGB (optimal efficiency), XGBoost (balanced), and Voting (maximal accuracy). The existence of multiple Pareto-optimal solutions is valuable, as different deployment contexts prioritize different objectives. The dramatic efficiency of KNN (98.59%, 0.02s) demonstrates the power of the hybrid feature representation: with sufficiently informative features, even a simple, non-parametric classifier can approach state-of-the-art performance. For severely resource-constrained settings, KNN offers a viable path to high-quality classification.

5.1.3. Role of Hybrid Feature Representation

The high performance of even simple classifiers with the hybrid feature set indicates that the features themselves capture discriminative information effectively. The mutual information analysis showing top features from all families confirms that deep features, color statistics, texture measures, and ROI characteristics all contribute valuable, complementary information. The hybrid approach successfully bridges the gap between representation learning and domain-expert feature engineering, proving particularly valuable for medical imaging applications where datasets are typically smaller than those required for end-to-end deep learning [1], [5].

5.2. Comparison with Existing Literature

The 99.62% accuracy achieved in this study exceeds previously reported results on comparable tasks (Jiang et al. [1]: >95%; Maleki et al. [5]: 98.2%). Several factors likely contribute:

1. Multimodal Integration: The DMR-IR dataset's pairing of thermography and mammography provides complementary functional and structural information [16].
2. Hybrid Feature Representation: Combining deep features from multiple architectures with diverse handcrafted features creates a rich representation capturing multiple aspects of tissue characterization.
3. Comprehensive Preprocessing: ROI segmentation, mutual information selection, and SMOTEENN imbalance handling optimize the data for learning.
4. Ensemble Methodology: Top-performing models leverage ensemble principles to reduce variance and improve generalization.

5.3. Clinical Implications

5.3.1. Model Selection Guidelines

Table 7 presents model selection recommendations by deployment scenario.

Table 7: Model selection recommendations by deployment scenario

Deployment Scenario	Recommended Model	Accuracy (%)	Training Time (s)	Rationale
Resource-constrained primary care	KNN	98.59	0.02	No training required; runs on basic hardware; preliminary findings require validation
Regional hospital (moderate resources)	HGB	99.54	5.98	Optimal accuracy-efficiency balance; experimental results show promise
Centralized reference center	Voting	99.62	27.91	Maximum accuracy justifies longer training; requires external validation
Research institution	Stacking	99.58	196.24	Benchmark performance; analyzable meta-learner
Interpretability priority	RF	98.90	25.67	Native feature importance; visualizable trees

5.3.2. Implications for Resource-Constrained Settings

KNN's 98.59% accuracy with virtually zero training time has particular significance for healthcare systems in transitioning countries [22]. In settings where specialist radiologists are scarce, computational infrastructure is limited, and budgets are constrained, a KNN-based system offers a promising path that warrants further investigation for computer-aided diagnosis. The model requires no training, can run on basic hardware, and provides predictions through simple distance computations.

5.4. Limitations

Several limitations should be acknowledged

1. AUC Calculation: Technical challenges prevented reliable AUC-ROC computation for certain ensemble models (Voting and Stacking ensembles) due to probability calibration inconsistencies across base learners. Future work will address probability calibration to enable comprehensive AUC-ROC reporting alongside other metrics.
2. Single Dataset: The study is delimited to a single dataset (DMR-IR). While this dataset provides rich multimodal information, findings may not generalize to other

populations, imaging protocols, or acquisition conditions. External validation on independent multi-center datasets is essential before any clinical implementation.

3. Inference Time: While training time is comprehensively reported, inference time was not systematically evaluated. For real-time applications, prediction latency may be as important as training efficiency.
4. No Fine-Tuning: Pre-trained CNNs were used as fixed feature extractors without fine-tuning. Fine-tuning might yield deep features more specifically adapted to breast tissue characterization.
5. Binary Classification: The study addresses binary classification (healthy vs. sick) rather than multi-class subtyping or grading. Characterization (molecular subtype, histological grade) is equally important for treatment selection.

6. Conclusions

This study investigated the performance of 13 ML algorithms spanning five methodological families for breast cancer classification on the multimodal DMR-IR dataset. Ensemble methods significantly outperform conventional classifiers, with top models (Voting, Stacking, XGBoost, HGB) achieving 99.5-99.6% accuracy. Statistical analysis

confirms these four methods are indistinguishable from each other while outperforming linear models by 9-10%. This demonstrates that for high-dimensional medical data, ensemble aggregation is more important than specific architectural choices. Training times vary by four orders of magnitude (0.02s to 196s) among models with similar accuracy. HGB offers the optimal balance (99.54% accuracy, 5.98s training), while KNN provides a highly efficient alternative (98.59% accuracy, 0.02s training) suitable for resource-constrained settings. The combination of deep features from three CNN architectures with handcrafted color, texture, and ROI descriptors creates a rich representation that enables even simple classifiers to achieve excellent performance. All feature families contribute discriminative information, validating the hybrid approach. The balanced error rates of top models (equal false positives/negatives) indicate no systematic bias toward overdiagnosis or underdiagnosis—a clinically essential property. Preliminary evidence-based model selection considerations are provided for different deployment contexts. Finally, the 99.62% accuracy achieved by the Voting ensemble establishes a new performance baseline for the DMR-IR dataset, demonstrating the potential of multimodal integration and hybrid feature representation.

Contributorship Statement

Ali Hamad Ali: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Naaman Omar Yaseen: Validation, review & editing, Supervision, Project administration.

All authors have read and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The authors would like to thank the Information Technology Management Department at the Technical College of Duhok and the Amedi Technical Institute, Duhok Polytechnic University, for providing the facilities and support that made this research possible. The authors also acknowledge the creators of the DMR-IR dataset for making their data publicly available.

7. References

- [1] B. Jiang, L. Bao, S. He, et al., "Deep learning applications in breast cancer histopathological imaging: diagnosis, treatment, and prognosis," *Breast Cancer Research*, vol.

- 26, no. 1, p. 137, 2024. doi: <https://doi.org/10.1186/s13058-024-01895-6>.
- [2] P. Ramamoorthy, B. R. Ramakantha Reddy, S. S. Askar, and M. Abouhawwash, "Histopathology-based breast cancer prediction using deep learning methods for healthcare applications," *Frontiers in Oncology*, vol. 14, 2024. doi: <https://doi.org/10.3389/fonc.2024.1300997>.
- [3] L. Wang, "Mammography with deep learning for breast cancer detection," *Frontiers in Oncology*, vol. 14, 2024. doi: <https://doi.org/10.3389/fonc.2024.1281922>.
- [4] K. A. Abdullah, S. Marziali, M. Nanaa, L. Escudero Sánchez, N. R. Payne, and F. J. Gilbert, "Deep learning-based breast cancer diagnosis in breast MRI: systematic review and meta-analysis," *European Radiology*, vol. 35, no. 8, pp. 4474-4489, 2025. doi: <https://doi.org/10.1007/s00330-025-11406-6>.
- [5] A. Maleki, M. Raahemi, and H. Nasiri, "Breast cancer diagnosis from histopathology images using deep neural network and XGBoost," *Biomedical Signal Processing and Control*, vol. 86, p. 105152, 2023. doi: <https://doi.org/10.1016/j.bspc.2023.105152>.
- [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*, vol. 13, no. 19, p. 3113, 2023. doi: <https://doi.org/10.3390/diagnostics13193113>.
- [7] A. A. Balasubramanian, S. M. A. Al-Heejawi, A. Singh, A. Breggia, B. Ahmad, R. Christman, S. T. Ryan, and S. Amal, "Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology," *Cancers*, vol. 16, no. 12, p. 2222, 2024. doi: <https://doi.org/10.3390/cancers16122222>.
- [8] N. R. Hameed Alsaedi and M. F. Akay, "Effective breast cancer classification using deep MLP, feature-fused auto-encoder and weight-tuned decision tree," *Applied Sciences*, vol. 15, no. 13, p. 7213, 2025. doi: <https://doi.org/10.3390/app15137213>.
- [9] B. Lokaj et al., "Multimodal deep learning fusion of ultra-fast-DCE MRI and clinical information for breast lesion classification," *Computers in Biology and Medicine*, vol. 188, p. 109721, 2025. doi: <https://doi.org/10.1016/j.compbiomed.2025.109721>.
- [10] X. Li, M. Gao, C. Zhang, G. Ma, Q. Zhang, W. Meng, T. Yuan, Y. Wang, and Z. Li, "A robust stacked neural network approach for early and accurate breast cancer diagnosis," *Frontiers in Medicine*, vol. 12, p. 1644857, 2025. doi: <https://doi.org/10.3389/fmed.2025.1644857>.
- [11] F. Gurcan, "Enhancing breast cancer prediction through stacking ensemble and deep learning integration," *PeerJ Computer Science*, vol. 11, p. e2461, 2025. doi: <https://doi.org/10.7717/peerj-cs.2461>.
- [12] S. Olthof, M. D. Nickel, E. Weiland, D. Leyhr, S. Afat, K. Nikolaou, and H. Preibsch, "Deep learning-enhanced T1-weighted imaging for breast MRI at 1.5T," *Diagnostics*, vol. 15, no. 13, p. 1681, 2025. doi: <https://doi.org/10.3390/diagnostics15131681>.
- [13] M. Nasser and U. K. Yusof, "Deep learning based methods for breast cancer diagnosis: A systematic review and future direction," *Diagnostics*, vol. 13, no. 1, p. 161, 2023. doi: <https://doi.org/10.3390/diagnostics13010161>.
- [14] A. Ali, M. Alghamdi, S. S. Marzuki, T. A. D. A. A. Tengku Din, M. S. Yamin, M. Alrashidi, I. S. Alkhazi, and N. Ahmed, "Exploring AI approaches for breast cancer detection and diagnosis: A review article," *Breast Cancer*:

- Targets and Therapy, vol. 17, pp. 927-947, 2025. doi: <https://doi.org/10.2147/BCTT.S550307>.
- [15] M. A. Rahman, M. S. H. Khan, Y. Watanobe, J. T. Prioty, T. T. Annita, S. Rahman, M. S. Hossain, S. A. Aitijjo, R. I. Taskin, V. Dhrubo, A. Hanip, and T. Bhuiyan, "Advancements in breast cancer detection: A review of global trends, risk factors, imaging modalities, machine learning, and deep learning approaches," *BioMedInformatics*, vol. 5, no. 3, p. 46, 2025. doi: <https://doi.org/10.3390/biomedinformatics5030046>.
- [16] A. H. Ali and N. Yaseen, "Breast cancer detection through deep learning and breast thermography: A study of review," *Dasinya Journal for Engineering and Informatics*, vol. 1, no. 1, 2025. doi: <https://doi.org/10.65542/djei.v1i1.9>.
- [17] R. Hoque, S. Das, M. Hoque, and E. Haque, "Breast cancer classification using XGBoost," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1985-1994, 2024. doi: <https://doi.org/10.30574/wjarr.2024.21.2.0625>.
- [18] D. Manali, H. Demirel, and A. Eleyan, "Deep learning based breast cancer detection using decision fusion," *Computers*, vol. 13, no. 11, p. 294, 2024. doi: <https://doi.org/10.3390/computers13110294>.
- [19] G. Jacob, I. Jose, and S. S., "Breast cancer detection: A comparative review on passive and active thermography," *Infrared Physics & Technology*, vol. 134, p. 104932, 2023. doi: <https://doi.org/10.1016/j.infrared.2023.104932>.
- [20] L. F. Silva, D. C. M. Saade, G. O. Sequeiros, A. C. Silva, A. C. Paiva, R. S. Bravo, and A. Conci, "A new database for breast research with infrared image," *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 1, pp. 92-100, 2014. Doi: <https://doi.org/10.1166/jmihi.2014.1226>.
- [21] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, no. 14, p. 3608, 2023. doi: <https://doi.org/10.3390/cancers15143608>.
- [22] P. Basu, L. Zhang, R. Hariprasad, A. L. Carvalho, and A. Barchuk, "A pragmatic approach to tackle the rising burden of breast cancer through prevention and early detection in countries 'in transition'," *Indian Journal of Medical Research*, vol. 152, no. 4, pp. 343-355, 2020. doi: https://doi.org/10.4103/ijmr.IJMR_1868_19.