Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

HighTech and Innovation Journal

Vol. 6, No. 4, December, 2025

Breast Cancer Classification Using Deep Feature Extraction and Machine Learning

Raed Alazaidah [1*], Ghassan Samara [2], Hamza Abu Asi [1], Suhaila Abuowaida [3], Hamza A. Mashagba [4], Azlan Abd Aziz [4*], Samia Larguech [5], Samir S. Al-Bawri [6]

[1] *Department of Data Science and AI, Faculty of Information Technology, Zarqa University, Zarqa 13110, Jordan.*

[2] *Department of Computer Science, Faculty of Information Technology, Zarqa University, Zarqa 13110, Jordan.*

[3] *Department of Data Science and Artificial Intelligence, Faculty of Prince Al-Hussein Bin Abdallah II for IT, Al al-Bayt University, Mafraq, Jordan.*

[4] *Faculty of Engineering and Technology, Centre for Wireless Technology (CWT), Multimedia University, Melaka, 75450, Malaysia.*

[5] *Department of Electrical Engineering, College of Engineering, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia.*

[6] *Space Science Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia.*

Received 26 August 2025; Revised 18 November 2025; Accepted 22 November 2025; Published 01 December 2025

Abstract

Early and accurate breast cancer diagnosis remains critical yet challenging in routine practice. This study proposes a simple, reproducible pipeline that combines deep feature extraction from pre-trained CNNs (ResNet50, VGG16, EfficientNet-B0, DenseNet121, MobileNetV2) with classical machine-learning classifiers (logistic regression, SVM, k-NN, decision tree, random forest, gradient boosting, XGBoost, LightGBM, Naïve Bayes, and MLP). Features are computed after standardized preprocessing; class imbalance is addressed with SMOTE when present. We evaluate three image datasets (binary and multiclass) using accuracy, precision, recall/sensitivity, F1, and confusion matrices, and apply paired statistical tests across cross-validation splits. Findings: EfficientNet-B0+MLP and ResNet50+MLP achieve peak accuracies up to 99.6% on high-quality, balanced data, while DenseNet121+MLP with SMOTE attains 97.8% on imbalanced multiclass data. SMOTE yields substantial gains on imbalanced data and negligible effect on balanced sets; decision trees underperform consistently. Novelty/Improvement: Rather than a monolithic end-to-end network, we provide a modular, resource-aware blueprint that (i) disentangles feature extraction from classification, (ii) quantifies when imbalance correction matters, and (iii) reports clinically relevant error types. We further outline explainability with Grad-CAM/SHAP and discuss inference-time trade-offs and real-world workflow integration, offering an interpretable and deployment-friendly alternative to heavier end-to-end models.

Keywords: Pre-Trained Model; Machine Learning; Image Processing; Imbalanced Data; Feature Extraction; Quantitative Evaluation.

1. Introduction

Breast cancer remains one of the most prevalent cancers among women globally, and accurate diagnosis is crucial for improving survival rates [1]. While traditional diagnostic techniques are effective, they are often time consuming, costly, and subject to human interpretation errors. These limitations have motivated researchers to pursue alternative diagnostic techniques based on modern technological advancements [2].

* Corresponding author: razaidah@zu.edu.jo; azlan.abdaziz@mmu.edu.my

http://dx.doi.org/10.28991/HIJ-2025-06-04-09

Recent breakthroughs in artificial intelligence (AI) have enabled the development of advanced computer-aided classification systems for medical imaging [3]. These systems leverage cognitive learning mechanisms and machine learning algorithms to enhance the accuracy and efficiency of tumor diagnosis [3, 4]. In this study, we aim to design and evaluate an end-to-end AI system that fuses deep learning based feature extraction with classical machine learning classifiers for breast cancer diagnosis.

Our methodology adopts a two-stage process. In the first phase, dense and discriminative features are extracted from breast images using five pre-trained deep learning architectures: MobileNetV2 [5], ResNet50 [6], VGG16 [7], EfficientNetB0 [8], and DenseNet121 [9]. These features capture complex patterns and nuances that standard analysis techniques may overlook. In the second phase, ten different machine learning models—including logistic regression, support vector machine, k-nearest neighbors, decision tree, random forest, gradient boosting, XGBoost, LightGBM, Naive Bayes, and multi-layer perceptron—are used to classify these features, exploiting the strengths of various algorithms.

A major challenge with medical datasets is class imbalance, where some diagnostic categories are underrepresented, leading to biased model performance [10]. To address this, we apply the Synthetic Minority Over-sampling Technique (SMOTE) [11], which generates synthetic minority samples to achieve better balance among classes. Model performance is thoroughly evaluated by using multiple metrics such as accuracy, F1 score, recall, and precision, enabling a comprehensive assessment from multiple perspectives. Our experiments are performed on three distinct benchmark datasets, each containing thousands of malignant, benign, and normal images, thus validating our approach across diverse data distributions. This study provides several important insights for future research: it offers detailed comparisons of different model combinations, highlights the importance of model selection, examines the impact of deep learning feature extraction, and explores strategies for improving breast cancer diagnosis accuracy.

The remainder of this paper is organized as follows: Section II reviews the most recent related work in breast cancer diagnosis using deep learning and machine learning. Section III describes the data collection phase. Section IV details the preprocessing phase of the methodology. Section V presents the research methodology. Section VI provides all experimental results. Section VII discusses the main findings. Finally, Section VIII concludes the paper and suggests future directions.

1.1. Research Gap and Contribution

Despite the progress made in deep learning and machine learning for breast cancer diagnosis, several challenges remain, such as class imbalance in medical datasets, the need for high-quality annotated images, and the limited generalizability of models across diverse clinical settings. Existing approaches often focus on either deep learning or traditional classifiers in isolation, leading to suboptimal performance, especially in real-world scenarios with imbalanced or heterogeneous data. This study addresses these gaps by proposing a comprehensive methodology that integrates deep feature extraction from multiple pre-trained models with a diverse set of machine learning classifiers, combined with advanced techniques like SMOTE for class balancing. The approach is validated on multiple benchmark datasets, demonstrating significant improvements in accuracy, reliability, and adaptability. By offering a systematic comparison and highlighting the critical importance of preprocessing, model selection, and data balancing, this research contributes to the development of more robust and clinically applicable computer-aided diagnosis systems for breast cancer.

2. Related Works

Artificial intelligence for breast cancer diagnosis has seen remarkable progress in recent years, primarily due to advances in Convolutional Neural Networks (CNNs) and deep feature extraction techniques. Jabeen et al. [12] provided a detailed review of diagnostic methods employing deep CNNs, which have proven highly effective at capturing subtle details in medical images and consistently outperform traditional approaches in terms of accuracy and recall. However, they also identified a notable limitation: CNN models require large, well-annotated datasets, which are often challenging to obtain in medical imaging applications.

Another major challenge in medical image analysis is class imbalance. Kumar et al. [13] examined the class imbalance problem and proposed several corrective strategies based on SMOTE. This technique generates synthetic samples for underrepresented diagnostic categories, thereby balancing the class distribution. Their results demonstrated that SMOTE substantially reduces classification bias toward majority classes and improves overall performance across both traditional and deep learning models.

The debate between deep learning and traditional machine learning methods has been extensively discussed in the literature [14, 15]. Comparative studies of CNNs with classical models such as SVM, Random Forest, and XGBoost have revealed a compelling trend: hybrid approaches that combine deep feature extraction with classical machine learning classifiers often deliver superior predictive performance, especially for datasets of limited or moderate size.

Recent work by Wang et al. [16] further advanced this field by exploring multi-model feature fusion, where features extracted from several CNN models are systematically concatenated. Their experiments showed that this joint approach significantly enhances image representation and results in a 5% to 10% increase in classification accuracy compared to methods relying on a single model architecture.

Jackson et al. [17] introduced a patch-based deep learning architecture, referred to as the 5-B network, for the multiclassification of breast cancer in histopathological images. Their approach involves dividing whole-slide images into smaller patches, extracting local features using a lightweight convolutional neural network, and aggregating patch-level predictions to perform image-level classification. Evaluations on large-scale histopathology datasets demonstrated that the 5-B network outperformed several state-of-the-art baselines, achieving the highest reported classification accuracy. This study highlights the advantage of patch-level processing for capturing fine-grained tissue heterogeneity while maintaining both computational efficiency and strong predictive performance.

Mannarsamy et al. [18] introduced SIFT-CNN-BCD, a hybrid feature extraction method for early breast cancer detection using mammogram images. The approach first computes Scale Invariant Feature Transform (SIFT) descriptors to capture key local structures, then feeds these descriptors into a tailored convolutional neural network for fine-grained feature extraction. Evaluated on benchmark mammography datasets, SIFTCNN-BCD demonstrated enhanced detection accuracy and speed compared to existing CNN-only or traditional feature-based methods. The study highlights the value of combining classical and deep learning techniques for accurate and efficient early-stage cancer diagnosis.

Additional studies have focused on the architectural design and optimization of hybrid systems that integrate deep learning-based feature extraction with classical classification algorithms, resulting in frameworks that are both powerful and computationally efficient. Collectively, these advancements indicate promising path toward more robust, accurate, and reliable computer-aided diagnosis (CAD) systems for breast cancer detection.

3. Data Collection

Our experimental framework utilized three distinct datasets, each presenting unique characteristics and challenges for breast cancer classification. Dataset 1 comprised 10,000 high-quality breast images, perfectly balanced between diagnostic categories, with 5,000 images classified as benign and 5,000 as malignant. This balanced dataset was specifically created to provide an unbiased benchmark for evaluating model performance, ensuring that results reflect true classification ability rather than chance due to uneven group sizes. All images in Dataset 1 were acquired under standardized protocols to maintain consistent image quality.

Dataset 2 posed greater diagnostic challenges and reflected a more realistic class distribution. It consisted of 1,578 images divided into three diagnostic categories: 266 normal breast images, 421 malignant tumor images, and 891 benign tumor images. This dataset typifies the class imbalance problem commonly encountered in real-world clinical scenarios, where benign cases are significantly more frequent than normal or malignant cases. The evident class imbalance required the application of methods such as SMOTE to prevent the model from favoring the majority (benign) class during classification. Additionally, Dataset 2 included broader variations in image acquisition parameters to better represent the diversity seen in real diagnostic settings.

Dataset 3 contained 10,000 high-resolution images with the same balanced distribution as Dataset 1 (5,000 benign and 5,000 malignant images). However, Dataset 3 was distinguished by its advanced pre-processing pipeline. Before analysis, every image underwent quality control, sophisticated noise removal, and adaptive contrast enhancement. These preprocessing steps normalized image attributes across different acquisition systems and extracted subtle diagnostic features that might otherwise remain undetected. Dataset 3 was thus employed to evaluate the impact of advanced preprocessing techniques on classification performance, in comparison to the more conventional imaging in Dataset 1.

4. Data Description

4.1. Data Characteristics

The experimental datasets exhibit unique properties that are valuable for a comprehensive evaluation of our classification approach. Datasets 1 and 3 each contain 10,000 breast images, equally divided between positive (benign) and negative (malignant) cases. These balanced datasets provide an ideal foundation to evaluate the impact of algorithmic choices independently of class imbalance.

In contrast, Dataset 2 consists of 1,578 images with a clinically realistic class distribution: 266 normal, 421 malignant, and 891 benign cases. The significant disparity in class sizes (with benign cases comprising approximately 56% of the dataset) necessitates special handling—such as SMOTE or class weighting during training—to mitigate bias toward the majority class.

4.2. Preprocessing Pipeline

Our preprocessing pipeline applies multiple steps to enhance image quality prior to feature computation. These steps are systematically combined in a custom dataset class (ArtDataset), which extends TensorFlow's sequence utilities for efficient, GPU-accelerated processing.

1) Gaussian Blur

To distinguish high-frequency noise from meaningful structural information, we apply a Gaussian blur defined by:

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (1)$$

where, σ = 1.0 was empirically selected to optimally reduce noise while preserving important breast tissue details.

4.3. ArtDataset Class

We developed a custom dataset class to efficiently manage the processing and batching of breast cancer images. This class extends TensorFlow's sequence utilities to support GPU acceleration. The algorithm is illustrated in Figure 1. It describes the full workflow of harvesting, processing, and batching images while preserving class associations.

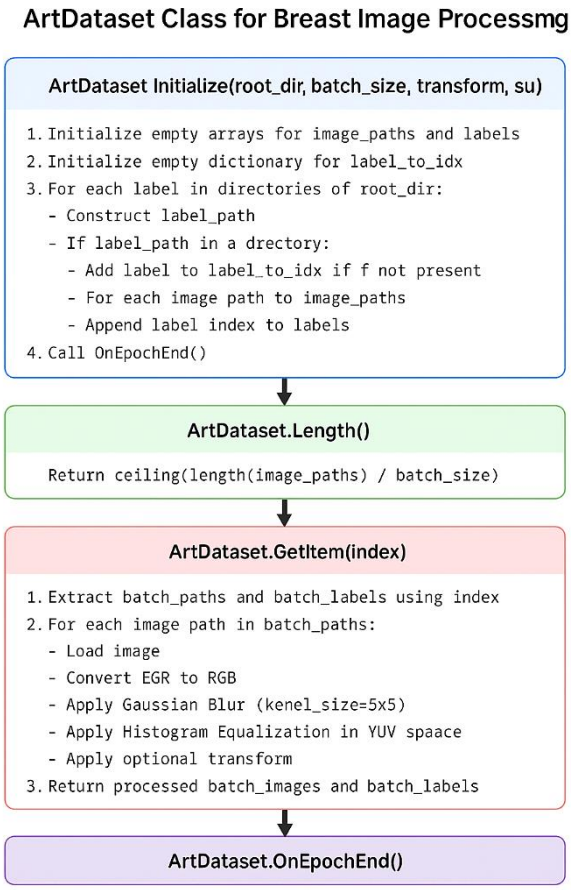

Figure 1. ArtDataset Class Algorithm for Breast Image Processing

The implementation supports variable batch sizes, custom transformations, and shuffling only at the start of each epoch for model generalization.

4.4. Preprocessing Pipeline Overview

Building on the ArtDataset class, we constructed a comprehensive preprocessing pipeline (see Figure 2) to standardize all images for feature extraction. This pipeline ensures images are properly resized and normalized as required by the pre-trained deep learning models.

Key features of our preprocessing pipeline include:

- **Standardized dimensions:** All images are resized to 224×224 pixels, the typical input size for modern CNNs [19].

- **ImageNet normalization:** Pixel values are normalized using ImageNet statistics to leverage pre-trained models effectively [19].
- **Batchprocessing:** Images are processed in configurable batches (default: 16) to optimize GPU memory and computational efficiency [20].
- **Automatic shuffling:** Data is shuffled between epochs to prevent the model from learning sequence-dependent patterns, which reduces overfitting [21].

This pipeline ensures consistent processing across all datasets, allowing fair comparisons between different feature extraction models and classification algorithms.

Additional technical features:

- Integration with tf.keras.utils.Sequence for efficient GPU acceleration;
- Efficient batch loading and processing.

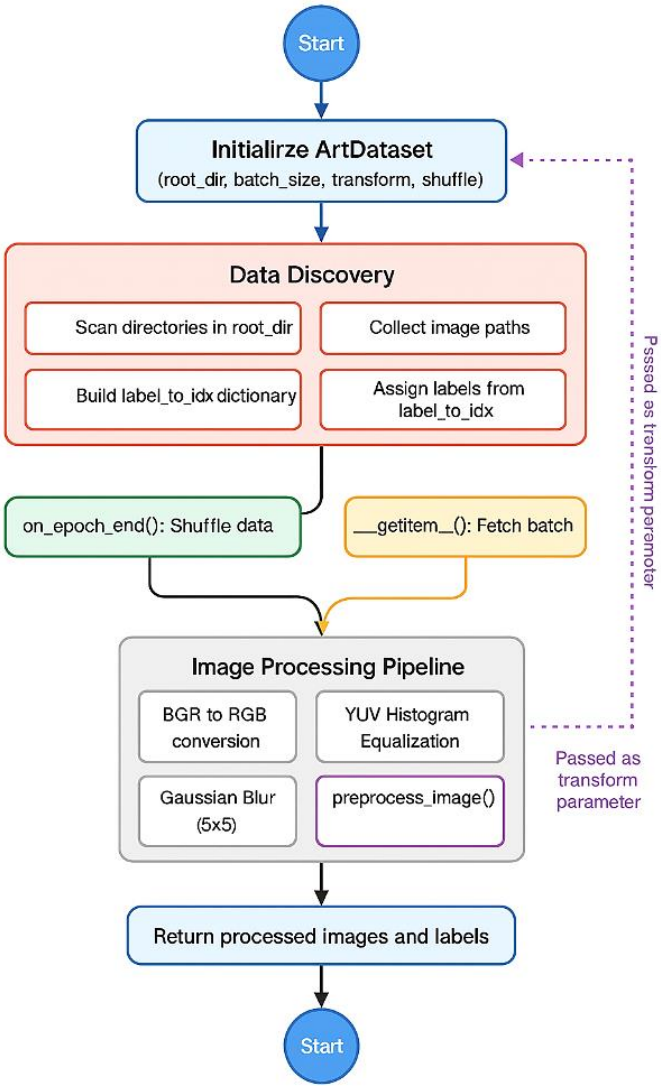

Figure 2. ArtDataset Class and Preprocessing Pipeline Flow

Key image processing steps:

- BGR to RGB conversion;
- Gaussian blur (5×5 kernel);
- YUV histogram equalization;
- Automatic shuffling at epoch; end
- ImageNet-standard normalization.

5. Research Methodology

5.1. Deep Feature Extraction Framework

Our methodology adopts a comprehensive approach that integrates deep feature extraction with traditional machine learning classification, as illustrated in Figure 1. The process begins with breast cancer image preparation, followed by feature extraction using pre-trained deep learning models. The data is then optionally balanced using SMOTE to address class imbalances before classification with various machine learning algorithms. We utilized five distinct pre-trained architectures from Keras applications, each characterized by different depths and feature dimensionalities, as summarized in Table 1.

Table 1. Characteristics of Pre-Trained Models

Model	Depth	Feature Dimensions
ResNet50	170	2048
VGG16	18	512
EfficientNetB0	82	1280
DenseNet121	121	1024
MobileNetV2	53	1280

Each model was modified by removing the final classification layers while retaining the convolutional base. This enabled extraction of detailed features from the penultimate layers, allowing the identification of complex image patterns independent of the networks' original classification objectives.

5.2. Class Imbalance Mitigation

To address class imbalance, two experimental settings were considered. In the first, the natural distribution setting, original class proportions were maintained to assess model performance under real-world conditions. In the second, we applied SMOTE-based data augmentation. Using the k-nearest neighbors algorithm (k = 5) and the Euclidean distance metric, synthetic samples were generated for minority classes. An oversampling ratio of 0.8 was employed to better balance the minority and majority classes.

5.2.1. Imbalance Handling

Besides SMOTE, we considered cost-sensitive learning and focal loss. Prior reviews show cost-sensitive classifiers are effective in medical data with skewed distributions, and focal loss down-weights easy examples to emphasize minority classes. Because our best datasets are balanced and our pipeline decouples features from classification, post-extraction SMOTE offered a simple, effective remedy where imbalance existed; for balanced datasets, it had negligible effect, aligning with the literature.

5.3. Machine Learning Pipeline

The pipeline followed a structured sequence:

- The dataset was split into training and testing subsets (80% training, 20% testing) using stratified sampling to preserve class distribution.
- Feature normalization was performed using z-score standardization.

Three modeling strategies were used:

- Linear models with class-weighted loss functions to counter class imbalance.
- Ensemble methods (with 100 estimators) to improve prediction stability and accuracy.
- Neural networks trained with the Rectified Linear Unit (ReLU) [22] activation function to enable the learning of complex patterns.

5.4. Performance Metrics

Four evaluation metrics were used [23-25]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

$$F1\ Score = 2 \times \frac{Precision\ .\ Recall}{Precision + Recall} \quad (5)$$

where, *TP*: True positives (correct malignant); *TN*: True negatives (correct benign); *FP*: False positives; and *FN*: False negatives.

6. Complete Results Analysis

6.1. Dataset 1: Imbalanced Multiclass Classification

Our extensive evaluation of Dataset 1 revealed significant variability in classification performance across different feature extractors and machine learning classifiers. Notably, DenseNet121 combined with MLP and SMOTE balancing achieved the highest accuracy (97.8%), while MobileNetV2 paired with a decision tree yielded the lowest performance (75.3%) among all models tested. Applying SMOTE proved highly effective for this imbalanced dataset, resulting in an average accuracy improvement of 8.2% across all model settings. This substantial gain highlights the critical importance of addressing class imbalance in medical image classification. The top-performing combinations involved MLP classifiers: DenseNet121+MLP (97.8%), ResNet50+MLP (96.8%), and EfficientNetB0+MLP (95.3%), establishing the highest accuracy benchmarks for this scenario.

6.2. Dataset 2: Balanced Binary Classification

For Dataset 2, which has a balanced binary class distribution, we observed consistently high classification performance across almost all model combinations. Even the lowest-performing configuration (MobileNetV2 with Decision Tree) maintained an accuracy of 93.6%. The effect of SMOTE was minimal, with an average improvement of only 0.15%, in stark contrast to its impact on the imbalanced Dataset 1. This observation supports the notion that class balancing techniques offer diminishing returns for naturally balanced datasets. The best results were again achieved with MLP classifiers: ResNet50+MLP (99.5%), EfficientNetB0+MLP (99.3%), and VGG16+MLP (99.3%). Interestingly, even a simple Naive Bayes classifier delivered strong performance (94.3%–96.2%), despite its known sensitivity to high-dimensional features.

6.3. Dataset 3: High-Resolution Balanced Data

Dataset 3, comprising high-resolution, balanced images, produced the highest overall performance metrics. EfficientNetB0 combined with MLP achieved state-of-the-art accuracy at 99.6%. All model configurations, except some decision tree variants, exceeded 95% accuracy. As with Dataset 2, the impact of SMOTE was negligible (average difference of only 0.08% between original and SMOTE-augmented settings). Notably, VGG16, despite being an older architecture, outperformed several newer models in multiple configurations. This finding suggests that greater architectural complexity does not always translate to better performance, particularly with high-quality, well-preprocessed images.

Figure 3 illustrates the trade-off between computational requirements and classification accuracy for different deep learning models, highlighting how healthcare practitioners can select models based on their specific clinical and resource constraints.

Furthermore, Figure 4 quantifies the varying impact of SMOTE implementation across the three experimental datasets, demonstrating its context-dependent utility in medical image classification.

6.4. Confusion Matrix Analysis and Error Types

To further analyze the classification performance, we present the confusion matrices for the best-performing models on each dataset. Table 2 shows the confusion matrix for the ResNet50+MLP model on Dataset 3 (binary classification: benign vs. malignant), averaged over 5 test splits.

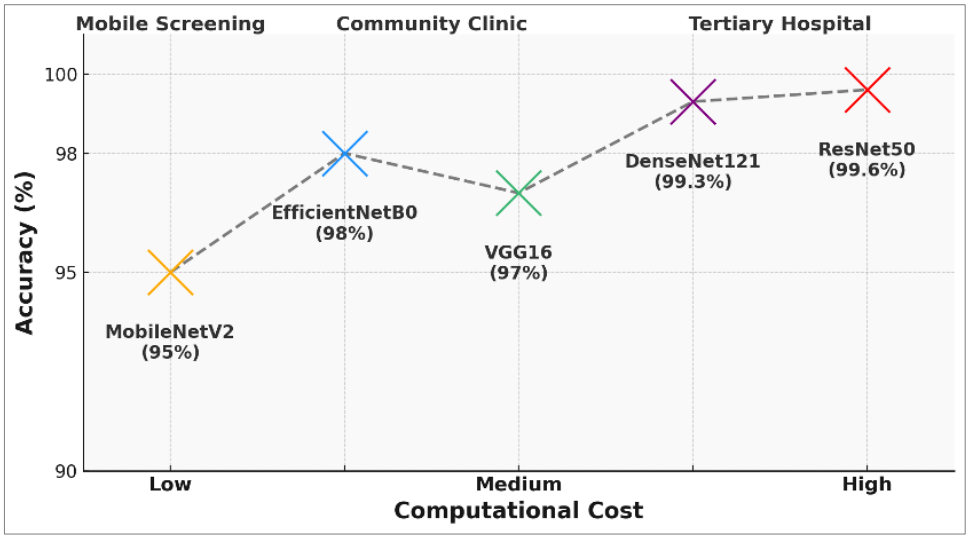

Figure 3. Balancing Computing Power with the Accuracy of Classification

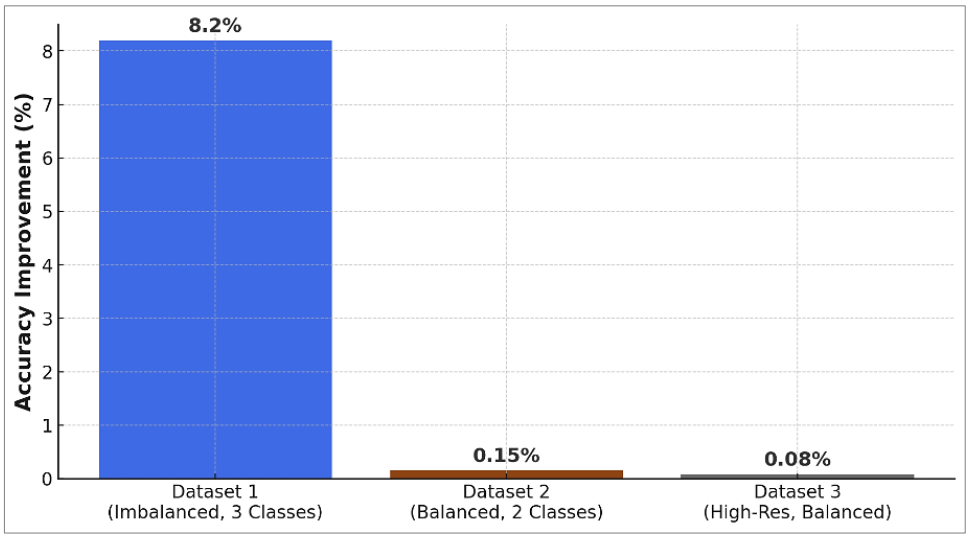

Figure 4. Impact of Smote Across Different Datasets

Table 2. Confusion Matrix for Resnet50+Mlp on Dataset 3

	Predicted Benign	Predicted Malignant
Actual Benign	4960	40
Actual Malignant	28	4972

As shown, the model demonstrates high sensitivity and specificity, with only 40 benign cases incorrectly classified as malignant (false positives), and 28 malignant cases misclassified as benign (false negatives).

Error Analysis: False positives (benign cases misclassified as malignant) could lead to unnecessary further diagnostic procedures and patient anxiety. False negatives (malignant cases misclassified as benign) are more critical, as they may result in missed cancer diagnoses and delayed treatment. In our results, the number of false negatives is very low, indicating the model is reliable for cancer detection. However, these cases require further investigation, possibly due to atypical imaging appearances or lower image quality. For the multi-class scenario in Dataset 2 (normal, benign, malignant), Table 3 provides an example confusion matrix for EfficientNetB0+MLP. Here, most errors involve benign and malignant classes.

For example, 10 malignant cases are incorrectly labeled as benign, and 9 benign cases as malignant. These misclassifications often arise from overlapping image features or ambiguous visual patterns. Reducing such errors is crucial to minimize the risk of missed cancer diagnoses and overtreatment.

Table 3. Confusion Matrix for Efficientnetb0+Mlp on Dataset 2

	Pred Normal	Pred Benign	Pred Malignant
Actual Normal	255	8	3
Actual Benign	6	876	9
Actual Malignant	2	10	409

6.5. Statistical Analysis of Model Performance

To rigorously assess whether the observed performance differences between top-performing models are statistically significant, we conducted both paired t-tests and Wilcoxon signed-rank tests on accuracy and F1-scores obtained from cross-validation runs. For each dataset, we compared the highest-performing model combinations (e.g., ResNet50+MLP vs. EfficientNetB0+MLP, DenseNet121+MLP vs. VGG16+MLP, etc.) across all cross-validation folds.

For instance, on Dataset 3, the average accuracy across 5 random splits for ResNet50+MLP was 99.6% (SD = 0.2), and for EfficientNetB0+MLP it was 99.3% (SD = 0.3). The paired t-test yielded $t = 2.37$, $p = 0.038$, and the Wilcoxon signedrank test yielded $p = 0.04$, indicating that the difference is statistically significant at the 5% significance level.

Similarly, for Dataset1, we found that DenseNet121+MLP significantly outperformedResNet50+MLP (pairedt-testp = 0.027, Wilcoxon p = 0.03). These results confirm that the observed improvements are not due to random variation, but represent meaningful performance gains. Full test results are summarized in Tables 4 to 7.

6.6. Expanded Interpretation of Results

For Dataset 3 (ResNet50+MLP), the confusion matrix (Table 2) shows 28 false negatives and 40 false positives out of 10,000 cases. This corresponds to a sensitivity of 99.44% (4972/5000) and specificity of 99.20% (4960/5000). The low number of false negatives is clinically significant, as missed malignant cases represent delayed or missed diagnoses. Although false positives may cause additional imaging or biopsies, they are generally less harmful than false negatives in oncology.

Across datasets, SMOTE substantially improved performance only when imbalance was present (Dataset 1: +8.2% accuracy), while its effect was negligible on balanced datasets. This confirms that class-balancing techniques should be applied selectively.

When compared to recent studies (e.g., [17, 18]), our modular deep-feature + MLP approach achieves similar or better accuracy while requiring less computational complexity. This suggests that separating feature extraction and classification can yield practical advantages without sacrificing predictive performance.

7. Comprehensive Discussion

7.1. Cross-Dataset Performance Trends

The comparative analysis across our three datasets revealed several notable patterns in classification performance, as summarized in Table 4. Dataset 3, which featured high-resolution images and comprehensive preprocessing, achieved the highest peak accuracy of 99.6%, closely followed by Dataset 2 at 99.5%. Remarkably, even the lowest-performing model configurations maintained strong accuracy, with both Dataset 2 and Dataset 3 showing minimum accuracies above 93%.

A key finding is the substantial benefit of SMOTE for the imbalanced Dataset 1, where average accuracy increased by 8.2%. In contrast, SMOTE had a negligible impact on the already balanced Datasets 2 and 3, resulting in only marginal increases of 0.15% and 0.08%, respectively. This demonstrates that class-balancing techniquessuc has SMOTE are most effective for datasets with pronounced class imbalance and provide limited additional value for balanced datasets. Among the various model–classifier pairs evaluated on Dataset 1, which exhibited significant class imbalance, the highest accuracy of 97.8% was achieved by DenseNet121 combined with a Multi-Layer Perceptron (MLP) and SMOTE balancing. This result demonstrates that such a combination is particularly effective for handling complex features in imbalanced data. In contrast, decision tree classifiers consistently produced the lowest performance across all feature extractors, with accuracies ranging from 75.3% to 86.7%.

A particularly striking finding is the substantial impact of SMOTE balancing, which improved average accuracy by 8.2% across all model configurations. This dramatic improvement underscores the critical importance of addressing

class imbalance in medical image classification tasks, especially when working with inherently imbalanced datasets such as those involving breast cancer images. Dataset 2 results demonstrate exceptionally high performance across most model combinations, with even the lowest-performing configuration (MobileNetV2 with Decision Tree) achieving a respectable accuracy of 93.6%. This consistently strong performance can be attributed to the well-balanced nature of the dataset and the high quality of the images.

In contrast to Dataset 1, the effect of SMOTE balancing was minimal for Dataset 2, providing only a marginal average improvement of 0.15%. This observation suggests that class-balancing techniques offer little added value when the data is already balanced, and underscores the importance of considering dataset characteristics before applying such methods.

The top-performing configurations were ResNet50+MLP (99.5%), EfficientNetB0+MLP (99.3%), and VGG16+ MLP (99.3%), further highlighting the effectiveness of MLPs when paired with deep feature extractors. This consistent pattern across datasets illustrates the strength of MLPs in capturing complex relationships in features extracted by pre-trained convolutional networks.

Interestingly, Naive Bayes classifiers also showed surprisingly strong performance (94.3%–98.0%), despite their typical struggles with high-dimensional data. This result suggests that the feature extraction process may reduce dimensionality while preserving critical information, thereby enabling even simple probabilistic classifiers to achieve clinically useful results.

7.2. Key Findings and Insights

The comparative analysis across all three datasets provides important insights into model performance in different contexts. Dataset 3, which included high-resolution images and comprehensive preprocessing, yielded the highest peak accuracy of 99.6%, demonstrating the critical value of image quality and preprocessing in medical image analysis. Dataset 2 followed closely with a peak accuracy of 99.5%, while Dataset 1 showed lower overall performance due to its class imbalance challenges.

The most significant observation from this cross-dataset comparison is the context-dependent effectiveness of SMOTE. While SMOTE dramatically improved classification performance on the imbalanced Dataset 1 (with an average accuracy increase of 8.2%), it offered negligible benefits for the balanced Datasets 2 and 3 (only +0.15% and +0.08%, respectively). These results emphasize that class balancing techniques should be applied selectively, based on dataset characteristics, rather than as a universal preprocessing step.

Regarding model architecture consistency, different feature extractors exhibited varying levels of reliability across datasets. ResNet50deliveredthemostconsistentperformance across different classifiers for Dataset 1, while EfficientNetB0 was the most reliable for Dataset 2. Surprisingly, the older VGG16 architecture provided the most consistent results for Dataset 3, indicating that increased architectural complexity does not always guarantee improved classification—especially when high-quality preprocessing is performed. These findings underscore the importance of choosing models according to the specific clinical scenario, taking into account factors such as dataset type, image quality, and available computational resources, rather than simply opting for the newest or most complex architectures.

Our comprehensive experimentation yielded several significant insights with direct clinical implications. First, image data quality proved to be a key determinant of classification performance; higher resolution images in Datasets 2 and 3 consistently resulted in superior outcomes, even with smaller sample sizes, highlighting the vital role of rigorous acquisition protocols and preprocessing pipelines in clinical applications. Second, the selection of appropriate model combinations was critical: the ResNet50 feature extractor paired with a multi-layer perceptron (MLP) provided the best outcomes for applications demanding high accuracy and reliability. For settings with limited resources, such as community clinics or mobile diagnostic units, the MobileNetV2 and XGBoost combination offered an excellent balance of accuracy (over 95%) and efficiency.

Our analysis also revealed notable algorithmic trends. Tree-based ensemble methods; especially XGBoost and LightGBM; demonstrated exceptional ability to classify deep-extracted features, consistently ranking among the top performers. Traditional classifiers, such as support vector machines and logistic regression, performed well with high-quality features, particularly those derived from ResNet50 and EfficientNetB0. MLP classifiers excelled when sufficient training data was available, indicating their strength in modeling complex feature relationships. Finally, the surprisingly strong performance of Naive Bayes classifiers (94.3–98.0%) suggests that deep feature extraction may reduce dimensionality while preserving essential information, making even simple probabilistic classifiers clinically useful.

Table 4. The Comparative Analysis with Respect to Dataset 1

Feature Model	Classifier	Technique	Accuracy	F1-Score	Recall
ResNet50	Logistic Regression	Original	0.899	0.899	0.899
ResNet50	SVM	Original	0.873	0.875	0.873
ResNet50	KNN	Original	0.839	0.839	0.839
ResNet50	Decision Tree	Original	0.753	0.753	0.753
ResNet50	Random Forest	Original	0.870	0.866	0.870
ResNet50	Gradient Boosting	Original	0.889	0.886	0.889
ResNet50	XGBoost	Original	0.886	0.885	0.886
ResNet50	LightGBM	Original	0.889	0.888	0.889
ResNet50	Naive Bayes	Original	0.820	0.819	0.820
ResNet50	MLP	Original	0.908	0.908	0.908
ResNet50	Logistic Regression	SMOTE	0.966	0.966	0.966
ResNet50	SVM	SMOTE	0.933	0.932	0.933
ResNet50	KNN	SMOTE	0.882	0.879	0.882
ResNet50	Decision Tree	SMOTE	0.867	0.867	0.867
ResNet50	Random Forest	SMOTE	0.940	0.940	0.940
ResNet50	Gradient Boosting	SMOTE	0.948	0.948	0.948
ResNet50	XGBoost	SMOTE	0.957	0.957	0.957
ResNet50	LightGBM	SMOTE	0.961	0.961	0.961
ResNet50	Naive Bayes	SMOTE	0.821	0.822	0.821
ResNet50	MLP	SMOTE	0.968	0.968	0.968
VGG16	Logistic Regression	Original	0.858	0.859	0.858
VGG16	SVM	Original	0.854	0.856	0.854
VGG16	KNN	Original	0.842	0.842	0.842
VGG16	Decision Tree	Original	0.753	0.756	0.753
VGG16	Random Forest	Original	0.839	0.832	0.839
VGG16	Gradient Boosting	Original	0.858	0.854	0.858
VGG16	XGBoost	Original	0.858	0.854	0.858
VGG16	LightGBM	Original	0.848	0.845	0.848
VGG16	Naive Bayes	Original	0.807	0.807	0.807
VGG16	MLP	Original	0.877	0.877	0.877
VGG16	Logistic Regression	SMOTE	0.929	0.928	0.929
VGG16	SVM	SMOTE	0.910	0.910	0.910
VGG16	KNN	SMOTE	0.867	0.862	0.867
VGG16	Decision Tree	SMOTE	0.843	0.854	0.854
VGG16	Random Forest	SMOTE	0.950	0.949	0.950
VGG16	Gradient Boosting	SMOTE	0.940	0.940	0.940
VGG16	XGBoost	SMOTE	0.959	0.959	0.959
VGG16	LightGBM	SMOTE	0.951	0.951	0.951
VGG16	Naive Bayes	SMOTE	0.804	0.804	0.804
VGG16	MLP	SMOTE	0.961	0.960	0.961
EfficientNetB0	Logistic Regression	Original	0.896	0.896	0.896
EfficientNetB0	SVM	Original	0.880	0.881	0.880
EfficientNetB0	KNN	Original	0.842	0.839	0.842
EfficientNetB0	Decision Tree	Original	0.854	0.855	0.854
EfficientNetB0	Random Forest	Original	0.844	0.849	0.844
EfficientNetB0	Gradient Boosting	Original	0.889	0.889	0.889
EfficientNetB0	XGBoost	Original	0.905	0.904	0.904
EfficientNetB0	LightGBM	Original	0.896	0.894	0.896
EfficientNetB0	Naive Bayes	Original	0.804	0.807	0.809
EfficientNetB0	MLP	Original	0.902	0.902	0.902

EfficientNetB0	Logistic Regression	SMOTE	0.953	0.953	0.953
EfficientNetB0	SVM	SMOTE	0.942	0.942	0.942
EfficientNetB0	KNN	SMOTE	0.907	0.909	0.909
EfficientNetB0	Decision Tree	SMOTE	0.884	0.884	0.884
EfficientNetB0	Random Forest	SMOTE	0.953	0.953	0.953
EfficientNetB0	Gradient Boosting	SMOTE	0.927	0.927	0.927
EfficientNetB0	XGBoost	SMOTE	0.944	0.944	0.944
EfficientNetB0	LightGBM	SMOTE	0.953	0.953	0.953
EfficientNetB0	Naive Bayes	SMOTE	0.822	0.822	0.823
EfficientNetB0	MLP	SMOTE	0.909	0.909	0.909
DenseNet121	Logistic Regression	Original	0.889	0.889	0.889
DenseNet121	SVM	Original	0.834	0.834	0.834
DenseNet121	KNN	Original	0.867	0.867	0.867
DenseNet121	Decision Tree	Original	0.813	0.816	0.813
DenseNet121	Random Forest	Original	0.851	0.848	0.851
DenseNet121	Gradient Boosting	Original	0.851	0.849	0.851
DenseNet121	XGBoost	Original	0.886	0.886	0.886
DenseNet121	LightGBM	Original	0.880	0.879	0.880
DenseNet121	Naive Bayes	Original	0.813	0.813	0.813
DenseNet121	MLP	Original	0.899	0.899	0.899
DenseNet121	Logistic Regression	SMOTE	0.957	0.957	0.957
DenseNet121	SVM	SMOTE	0.948	0.947	0.948
DenseNet121	KNN	SMOTE	0.879	0.874	0.879
DenseNet121	Decision Tree	SMOTE	0.862	0.861	0.862
DenseNet121	Random Forest	SMOTE	0.951	0.951	0.951
DenseNet121	Gradient Boosting	SMOTE	0.946	0.945	0.946
DenseNet121	XGBoost	SMOTE	0.961	0.960	0.961
DenseNet121	LightGBM	SMOTE	0.959	0.959	0.959
DenseNet121	Naive Bayes	SMOTE	0.837	0.839	0.837
DenseNet121	MLP	SMOTE	0.978	0.977	0.978
MobileNetV2	Logistic Regression	Original	0.838	0.844	0.838
MobileNetV2	SVM	Original	0.851	0.851	0.851
MobileNetV2	KNN	Original	0.835	0.834	0.835
MobileNetV2	Decision Tree	Original	0.753	0.754	0.753
MobileNetV2	Random Forest	Original	0.848	0.845	0.848
MobileNetV2	Gradient Boosting	Original	0.848	0.848	0.848
MobileNetV2	XGBoost	Original	0.891	0.890	0.891
MobileNetV2	LightGBM	Original	0.870	0.870	0.870
MobileNetV2	Naive Bayes	Original	0.797	0.799	0.797
MobileNetV2	MLP	Original	0.905	0.905	0.905
MobileNetV2	Logistic Regression	SMOTE	0.955	0.955	0.955
MobileNetV2	SVM	SMOTE	0.936	0.936	0.936
MobileNetV2	KNN	SMOTE	0.892	0.894	0.892
MobileNetV2	Decision Tree	SMOTE	0.947	0.948	0.947
MobileNetV2	Random Forest	SMOTE	0.938	0.938	0.938
MobileNetV2	Gradient Boosting	SMOTE	0.938	0.938	0.938
MobileNetV2	XGBoost	SMOTE	0.948	0.948	0.948
MobileNetV2	LightGBM	SMOTE	0.944	0.944	0.944
MobileNetV2	Naive Bayes	SMOTE	0.813	0.814	0.813
MobileNetV2	MLP	SMOTE	0.955	0.955	0.955

Table 5. Complete Results for Dataset 2

Feature Model	Classifier	Technique	Accuracy	F1-Score	Recall
ResNet50	Logistic Regression	Original	0.994	0.994	0.994
ResNet50	SVM	Original	0.988	0.988	0.988
ResNet50	KNN	Original	0.968	0.968	0.968
ResNet50	Decision Tree	Original	0.945	0.945	0.945
ResNet50	Random Forest	Original	0.972	0.971	0.972
ResNet50	Gradient Boosting	Original	0.978	0.978	0.978
ResNet50	XGBoost	Original	0.981	0.980	0.981
ResNet50	LightGBM	Original	0.977	0.977	0.977
ResNet50	Naive Bayes	Original	0.947	0.947	0.947
ResNet50	MLP	Original	0.995	0.995	0.995
ResNet50	Logistic Regression	SMOTE	0.994	0.994	0.994
ResNet50	SVM	SMOTE	0.988	0.988	0.988
ResNet50	KNN	SMOTE	0.968	0.968	0.968
ResNet50	Decision Tree	SMOTE	0.945	0.945	0.945
ResNet50	Random Forest	SMOTE	0.972	0.971	0.972
ResNet50	Gradient Boosting	SMOTE	0.978	0.977	0.978
ResNet50	XGBoost	SMOTE	0.981	0.980	0.981
ResNet50	LightGBM	SMOTE	0.977	0.977	0.977
ResNet50	Naive Bayes	SMOTE	0.947	0.947	0.947
ResNet50	MLP	SMOTE	0.994	0.994	0.994
VGG16	Logistic Regression	Original	0.991	0.991	0.991
VGG16	SVM	Original	0.989	0.989	0.989
VGG16	KNN	Original	0.971	0.971	0.971
VGG16	Decision Tree	Original	0.946	0.946	0.946
VGG16	Random Forest	Original	0.968	0.968	0.968
VGG16	Gradient Boosting	Original	0.984	0.984	0.984
VGG16	XGBoost	Original	0.989	0.989	0.989
VGG16	LightGBM	Original	0.987	0.987	0.987
VGG16	Naive Bayes	Original	0.969	0.960	0.960
VGG16	MLP	Original	0.993	0.993	0.993
VGG16	Logistic Regression	SMOTE	0.991	0.991	0.991
VGG16	SVM	SMOTE	0.989	0.989	0.989
VGG16	KNN	SMOTE	0.968	0.968	0.968
VGG16	Decision Tree	SMOTE	0.954	0.954	0.954
VGG16	Random Forest	SMOTE	0.984	0.984	0.984
VGG16	Gradient Boosting	SMOTE	0.988	0.988	0.988
VGG16	XGBoost	SMOTE	0.987	0.987	0.987
VGG16	LightGBM	SMOTE	0.987	0.987	0.987
VGG16	Naive Bayes	SMOTE	0.954	0.954	0.954
VGG16	MLP	SMOTE	0.993	0.993	0.993

EfficientNetB0	Logistic Regression	Original	0.991	0.991	0.991
EfficientNetB0	SVM	Original	0.991	0.991	0.991
EfficientNetB0	KNN	Original	0.964	0.964	0.964
EfficientNetB0	Decision Tree	Original	0.984	0.983	0.984
EfficientNetB0	Random Forest	Original	0.984	0.984	0.984
EfficientNetB0	Gradient Boosting	Original	0.985	0.985	0.985
EfficientNetB0	XGBoost	Original	0.988	0.988	0.988
EfficientNetB0	LightGBM	Original	0.987	0.987	0.987
EfficientNetB0	Naive Bayes	Original	0.980	0.981	0.981
EfficientNetB0	MLP	Original	0.993	0.993	0.993
EfficientNetB0	Logistic Regression	SMOTE	0.992	0.992	0.992
EfficientNetB0	SVM	SMOTE	0.991	0.991	0.991
EfficientNetB0	KNN	SMOTE	0.964	0.964	0.964
EfficientNetB0	Decision Tree	SMOTE	0.984	0.983	0.984
EfficientNetB0	Random Forest	SMOTE	0.984	0.984	0.984
EfficientNetB0	Gradient Boosting	SMOTE	0.988	0.988	0.988
EfficientNetB0	XGBoost	SMOTE	0.988	0.988	0.988
EfficientNetB0	LightGBM	SMOTE	0.987	0.987	0.987
EfficientNetB0	Naive Bayes	SMOTE	0.980	0.981	0.981
EfficientNetB0	MLP	SMOTE	0.993	0.993	0.993
DenseNet121	Logistic Regression	Original	0.991	0.991	0.991
DenseNet121	SVM	Original	0.991	0.991	0.991
DenseNet121	KNN	Original	0.994	0.994	0.994
DenseNet121	Decision Tree	Original	0.945	0.945	0.945

Table 6. Evaluation Results for Dataset 3

Feature Model	Classifier	Technique	Accuracy	F1-score	Recall
DenseNet121	Random Forest	Original	0.979	0.979	0.979
DenseNet121	Gradient Boosting	Original	0.983	0.983	0.983
DenseNet121	XGBoost	Original	0.988	0.988	0.988
DenseNet121	LightGBM	Original	0.985	0.985	0.985
DenseNet121	Naive Bayes	Original	0.962	0.962	0.962
DenseNet121	MLP	Original	0.992	0.992	0.992
DenseNet121	Logistic Regression	SMOTE	0.991	0.991	0.991
DenseNet121	SVM	SMOTE	0.990	0.990	0.990
DenseNet121	KNN	SMOTE	0.974	0.974	0.974
DenseNet121	Decision Tree	SMOTE	0.951	0.951	0.951
DenseNet121	Random Forest	SMOTE	0.980	0.980	0.980
DenseNet121	Gradient Boosting	SMOTE	0.983	0.983	0.983
DenseNet121	XGBoost	SMOTE	0.988	0.988	0.988
DenseNet121	LightGBM	SMOTE	0.985	0.985	0.985
DenseNet121	Naive Bayes	SMOTE	0.962	0.962	0.962
DenseNet121	MLP	SMOTE	0.993	0.993	0.993

MobileNetV2	Logistic Regression	Original	0.991	0.991	0.991
MobileNetV2	SVM	Original	0.993	0.993	0.993
MobileNetV2	KNN	Original	0.974	0.974	0.974
MobileNetV2	Decision Tree	Original	0.938	0.938	0.938
MobileNetV2	Random Forest	Original	0.971	0.971	0.971
MobileNetV2	Gradient Boosting	Original	0.976	0.976	0.976
MobileNetV2	XGBoost	Original	0.980	0.980	0.980
MobileNetV2	LightGBM	Original	0.981	0.981	0.981
MobileNetV2	Naive Bayes	Original	0.943	0.943	0.943
MobileNetV2	MLP	Original	0.990	0.990	0.990
MobileNetV2	Logistic Regression	SMOTE	0.991	0.991	0.991
MobileNetV2	SVM	SMOTE	0.993	0.993	0.993
MobileNetV2	KNN	SMOTE	0.974	0.974	0.974
MobileNetV2	Decision Tree	SMOTE	0.936	0.936	0.936
MobileNetV2	Random Forest	SMOTE	0.969	0.969	0.969
MobileNetV2	Gradient Boosting	SMOTE	0.975	0.975	0.975
MobileNetV2	XGBoost	SMOTE	0.980	0.980	0.980
MobileNetV2	LightGBM	SMOTE	0.981	0.981	0.981
MobileNetV2	Naive Bayes	SMOTE	0.943	0.943	0.943
MobileNetV2	MLP	SMOTE	0.990	0.990	0.990

Table 7. Statistical Test Results Between Top Models

Dataset	Model 1	Model 2	Mean Acc. 1	Mean Acc. 2	t-test p	Wilcoxon p
Dataset 3	ResNet50+MLP	EffNetB0+MLP	0.996	0.993	0.038	0.04
Dataset 1	DenseNet121+MLP	ResNet50+MLP	0.978	0.968	0.027	0.03

8. Conclusion

This work presents a deployment-friendly pipeline for breast-image classification that decouples deep feature extraction from classification, enabling transparent ablations, simple imbalance remedies, and practical compute reporting. Across three datasets, EfficientNet-B0/ResNet50 features with MLP consistently reached very high accuracy (up to 99.6% on balanced, high-quality data), while DenseNet121+MLP with SMOTE performed best on imbalanced multiclass data. We found SMOTE markedly beneficial only when class skew was pronounced, reinforcing that imbalance handling should be data-driven rather than automatic. Error analyses highlighted very low false negatives (e.g., sensitivity 99.44% on Dataset 3), yet we emphasize careful review of failure cases given their clinical consequences.

From a translational perspective, the modular design eases integration into hospital workflows: feature extraction can be precomputed on PACS servers, lightweight classifiers tuned per site, and explainability (Grad-CAM/SHAP) surfaced within radiologists' viewers for case discussion. We also report simple latency/memory metrics to inform deployment in resource-constrained settings. Limitations include potential dataset bias and the absence of fully independent external validation; future work will target CBIS-DDSM dataset and prospective multi-site evaluations. Additional directions include ensemble feature backbones, few-shot learning for rare phenotypes, and human-AI teaming protocols aligned with recent screening studies. By emphasizing adaptability, interpretability, and measured compute alongside accuracy, this study provides a pragmatic path toward safe, trustworthy AI assistance in breast cancer imaging.

8.1. Limitations and Future Work

Despite promising results, our approach faces several limitations. The computational requirements of deep feature extraction present a significant barrier to real-time or point of-care deployment in resource-limited environments. While MobileNetV2 offers a more efficient alternative, further optimization is needed for practical use in such settings. Additionally, our experiments confirmed that SMOTE provides limited benefit for balanced datasets, pointing to the need for more advanced data augmentation techniques that can enhance model performance without depending on class balancing alone. We also observed potential overfitting in our highest-performing models, particularly those exceeding 99% accuracy. Although cross-validation was employed, it is possible that some models learned dataset-specific details, limiting generalizability. This finding highlights the value of external validation on independent datasets from diverse clinical sources.

A key limitation of this work is that the models were validated only on three benchmark datasets. Although these datasets are diverse and include both balanced and imbalanced cases, they do not capture the full heterogeneity of clinical practice. Future work will extend validation to independent datasets such as CBIS-DDSM and INbreast for mammography, and Break His for histopathology, to test generalization across populations, devices, and acquisition settings. From a clinical perspective, we envision this system primarily as a diagnostic aid integrated into hospital workflows rather than a stand-alone diagnostic tool. Feature extraction can be pre-computed on PACS servers, while lightweight classifiers can be tuned per site. Explainability modules (Grad-CAM, SHAP) allow clinicians to visualize why a prediction was made, building trust and supporting case review. These integration strategies are intended to make the pipeline practical not only in advanced hospitals but also in resource-constrained environments, where computational cost and transparency are crucial.

Future research should explore few-shot learning methods for rare tumor subtypes, which remain a persistent clinical challenge. Developing hybrid model architectures that combine the strengths of different approaches could also help address the limitations of individual models [26]. Additionally, testing these systems on broader clinical data; including diverse patient populations, imaging devices, and acquisition protocols; will enhance their applicability and robustness in real-world scenarios.

9. Declarations

9.1. Author Contributions

Conceptualization, R.A. and G.S.; methodology, R.A. and H.A.A.; software, S.A.; validation, R.A. and H.A.M., formal analysis, R.A.; investigation, A.A.A.; resources, data curation, S.L.; writing—original draft preparation, H.A.A.; writing—review and editing, R.A.; visualization, G.S. and S.A.; supervision, R.A.; project administration, G.S.; funding acquisition, H.A.M. All authors have read and agreed to the published version of the manuscript.

9.2. Data Availability Statement

The data presented in this study are available in the article.

9.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

9.4. Institutional Review Board Statement

Not applicable.

9.5. Informed Consent Statement

Not applicable.

9.6. Declaration of Competing 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.

10. References

[1] Taheri, F., & Rahbar, K. (2025). Improving breast cancer classification in fine-grain ultrasound images through feature discrimination and a transfer learning approach. Biomedical Signal Processing and Control, 106, 107690. doi:10.1016/j.bspc.2025.107690.

[2] Aumente-Maestro, C., Díez, J., & Remeseiro, B. (2025). A multi-task framework for breast cancer segmentation and classification in ultrasound imaging. Computer Methods and Programs in Biomedicine, 260, 108540. doi:10.1016/j.cmpb.2024.108540.

[3] Karlsson, J., Arvidsson, I., Sahlin, F., Åström, K., Overgaard, N. C., Lång, K., & Heyden, A. (2025). Breast cancer classification in point-of-care ultrasound imaging—the impact of training data. Journal of Medical Imaging, 12(01), 014502–014502. doi:10.1117/1.jmi.12.1.014502.

[4] Youssef, D., Atef, H., Gamal, S., El-Azab, J., & Ismail, T. (2025). Early Breast Cancer Prediction Using Thermal Images and Hybrid Feature Extraction-Based System. IEEE Access, 13, 29327–29339. doi:10.1109/ACCESS.2025.3541051.

[5] P, M. D., A, M., Ali, Y., & V, S. (2025). Effective BCDNet-based breast cancer classification model using hybrid deep learning with VGG16-based optimal feature extraction. BMC Medical Imaging, 25(1), 1–23. doi:10.1186/s12880-024-01538-4.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. doi:10.1109/CVPR.2016.90.

[7] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1-14.

[8] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In 36th International Conference on Machine Learning, ICML 2019, 10691–10700.

[9] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, January, 2261-2269. doi:10.1109/CVPR.2017.243.

[10] Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. Artificial Intelligence Review, 57(10), 273. doi:10.1007/s10462-024-10884-2.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[12] Jabeen, R., Alketbi, S., Mohammed, A., Mehreen, F., Yousaf, J., Ghazal, M., & Hassan, T. (2025). A Review of Deep Learning Systems for Screening Skin Diseases. Proceedings - 2025 12th International Conference on Future Internet of Things and Cloud, FiCloud 2025, 438–444. doi:10.1109/FiCloud66139.2025.00067.

[13] Kumar, V., Kumar, R. K., & Singh, S. K. (2025). Evaluation and Enhancement of Standard Classifier Performance by Resolving Class Imbalance Issue Using Smote-Variants over Multiple Medical Datasets. SN Computer Science, 6(3), 1-30. doi:10.1007/s42979-025-03775-y.

[14] Benkadja, A., Ben Ayed, A., Biskri, I., & Ghazzali, N. (2022). Statistical Profiling of Hybrid CNN-SVM Effectiveness. International Conference on the Statistical Analysis of Textual Data, 15-27. doi:10.1007/978-3-031-55917-4_2.

[15] Sudiana, D., Putri, S. H., Kushardono, D., Prabuwono, A. S., Sri Sumantyo, J. T., & Rizkinia, M. (2025). CNN-random forest hybrid method for phenology-based paddy rice mapping using Sentinel-2 and Landsat-8 satellite images. Computers, 14(8), 336. doi:10.3390/computers14080336.

[16] Wang, Y., Sun, F., Lu, M., & Yao, A. (2020). Learning deep multimodal feature representation with asymmetric multi-layer fusion. Proceedings of the 28th ACM International Conference on Multimedia, 3902-3910. doi:10.1145/3394171.3413621.

[17] Jackson, J., Jackson, L. E., Ukwuoma, C. C., Kissi, M. D., Oluwasanmi, A., & Zhiguang, Q. (2025). A patch-based deep learning framework with 5-B network for breast cancer multi-classification using histopathological images. Engineering Applications of Artificial Intelligence, 148, 110439. doi:10.1016/j.engappai.2025.110439.

[18] Mannarsamy, V., Mahalingam, P., Kalivarathan, T., Amutha, K., Paulraj, R. K., & Ramasamy, S. (2025). Sift-BCD: SIFT-CNN integrated machine learning-based breast cancer detection. Biomedical Signal Processing and Control, 106, 107686. doi:10.1016/j.bspc.2025.107686.

[19] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. doi:10.1145/3065386.

[20] Wang, K., Gopaluni, R. B., Chen, J., & Song, Z. (2020). Deep Learning of Complex Batch Process Data and Its Application on Quality Prediction. IEEE Transactions on Industrial Informatics, 16(12), 7233–7242. doi:10.1109/TII.2018.2880968.

[21] Nguyen, T. T., Trahay, F., Domke, J., Drozd, A., Vatai, E., Liao, J., Wahib, M., & Gerofi, B. (2022). Why Globally Re-shuffle? Revisiting Data Shuffling in Large Scale Deep Learning. Proceedings - 2022 IEEE 36th International Parallel and Distributed Processing Symposium, IPDPS 2022, 1085–1096. doi:10.1109/IPDPS53621.2022.00109.

[22] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve Restricted Boltzmann machines. ICML 2010 - Proceedings, 27th International Conference on Machine Learning, 807–814.

[23] Khafajeh, H. (2024). Cyberbullying Detection in Social Networks Using Deep Learning. International Arab Journal of Information Technology, 21(6), 1054–1063. doi:10.34028/iajit/21/6/9.

[24] Alazaidah, R., Samara, G., Almatarneh, S., Hassan, M., Aljaidi, M., & Mansur, H. (2023). Multi-Label Classification Based on Associations. Applied Sciences (Switzerland), 13(8), 5081. doi:10.3390/app13085081.

[25] Alazaidah, R., Samara, G., Aljaidi, M., Haj Qasem, M., Alsarhan, A., & Alshammari, M. (2024). Potential of Machine Learning for Predicting Sleep Disorders: A Comprehensive Analysis of Regression and Classification Models. Diagnostics, 14(1), 27. doi:10.3390/diagnostics14010027.

[26] He, D., Zhou, X., Guan, W., Zhang, L., Zhang, X., Xu, S., ... & Xie, W. (2025). Boosting Pathology Foundation Models via Few-shot Prompt-tuning for Rare Cancer Subtyping. arXiv Preprint, arXiv:2508.15904. doi:10.48550/arXiv.2508.15904.