DOI: 10.4274/forbes.galenos.2025.02360
Forbes J Med 2025;6(2):147-55

Original Article / Orijinal Araştırma

Multi-CNN Deep Feature Fusion and Stacking Ensemble Classifier for Breast Ultrasound Lesion Classification

Meme Ultrasonu Lezyon Sınıflandırması için Multi-CNN Derin Özellik Füzyonu ve Yığınlama Topluluk Sınıflandırıcısı

Kemal PANÇ[1], Sümeyye SEKMEN[2]

[1]Karakoçan State Hospital, Clinic of Radiology, Elazığ, Türkiye
[2]Patnos State Hospital, Clinic of Radiology, Ağrı, Türkiye

Cite as: Panç K, Sekmen S. Multi-CNN deep feature fusion and stacking ensemble classifier for breast ultrasound lesion classification. Forbes Med J. 2025;6(2):147-55

ABSTRACT

Objective: To develop and validate a robust machine learning model for classifying breast ultrasound images into benign, malignant, and normal categories, aiming to enhance diagnostic accuracy using advanced feature extraction and ensemble learning techniques.

Methods: A dataset comprising 2233 images from five public datasets was utilized. After masking regions of interest, deep features were extracted using pre-trained VGG16, ResNet50V2, and EfficientNetB3 models, and concatenated. A multi-step feature selection process involving principal component analysis, recursive feature elimination with LightGBM, and partial least squares discriminant analysis was applied. A stacking ensemble classifier, integrating LightGBM, XGBoost, CatBoost, and random forest with a logistic regression meta-learner, was trained using 5-fold cross-validation on a 75% training set (balanced with synthetic minority oversampling technique), and evaluated on a 25% test set.

Results: The model achieved a macro average area under the curve-receiver operating characteristic (AUC-ROC) of 0.956 and an F1-score of 0.88 on the test set. Benign class results were AUC: 0.984, F1: 0.93, and normal class results were AUC: 0.969, F1: 0.92. The results for the malignant class were AUC: 0.916, F1 score: 0.79. Feature importance analysis showed that ResNet50V2 had the highest contribution to the model's performance.

Conclusion: The proposed approach, combining multi-convolutional neural network deep feature fusion, optimized feature selection, and ensemble stacking, shows significant potential for automated breast ultrasound classification, especially for benign and normal cases. While promising for clinical decision support, the model's lower sensitivity for malignant lesions necessitates further refinement.

Keywords: Breast ultrasound, deep learning, computer-aided diagnosis, ensemble learning, image classification

ÖZ

Amaç: Gelişmiş özellik çıkarma ve topluluk öğrenmesi teknikleri kullanarak tanısal doğruluğu artırmayı hedefleyen, meme ultrason görüntülerinin benign, malign ve normal kategorilerine sınıflandırılması için güçlü bir makine öğrenmesi modeli geliştirmek ve doğrulamaktır.

Yöntem: Beş halka açık veri setinden oluşan 2233 görüntülük bir veri seti kullanılmıştır. İlgili bölgeler maskelendikten sonra, önceden eğitilmiş VGG16, ResNet50V2 ve EfficientNetB3 modelleri kullanılarak derin özellikler çıkarılmış ve birleştirilmiştir. Temel bileşen analizi, LightGBM ile özyinelemeli özellik eleme ve kısmi en küçük kareler ayırt edici analizi içeren çok adımlı bir özellik seçme süreci uygulanmıştır. Lojistik regresyon meta-öğrenicisi ile LightGBM, XGBoost, CatBoost ve random forest'ı entegre eden bir yığınlama topluluk sınıflandırıcısı, %75'lik (sentetik azınlık aşırı örnekleme tekniği ile dengelenmiş) eğitim seti üzerinde *5-fold cross-validation* kullanılarak eğitilmiş ve %25'lik test seti üzerinde değerlendirilmiştir.

Received/Geliş: 19.04.2025
Accepted/Kabul: 19.06.2025
Epub: 17.07.2027
Yayınlanma Tarihi/ Publication Date: 06.08.2025

Corresponding Author/ Sorumlu Yazar:
Kemal PANÇ, MD,
Karakoçan State Hospital, Clinic of Radiology, Elazığ, Türkiye
✉ kemal.panc@gmail.com
ORCID: 0000-0002-3951-7344

Bulgular: Model, test seti üzerinde makro ortalama işlem karakteristik eğrisi altındaki alan (EAA) değeri 0,956 ve F1 skoru 0,88 elde etmiştir. Benign sınıf sonuçları EAA: 0,984, F1: 0,93 ve normal sınıf sonuçları EAA: 0,969, F1: 0,92. Malign sınıf sonuçları EAA: 0,916, F1: 0,79. Özellik önem analizi ResNet50V2'nin en yüksek katkıyı sağladığını göstermiştir.

Sonuç: Çoklu evrişimli sinir ağları derin özellik birleştirme, optimize edilmiş özellik seçimi ve topluluk yığınlamayı birleştiren önerilen yaklaşım, özellikle benign ve normal olgular için otomatik meme ultrason sınıflandırması açısından önemli bir potansiyel göstermektedir. Klinik karar desteği için umut verici olmakla birlikte, modelin malign lezyonlar için daha düşük duyarlılığı, daha fazla iyileştirme gerektirmektedir.

Anahtar Kelimeler: Meme ultrasonu, derin öğrenme, bilgisayar destekli tanı, topluluk öğrenmesi, görüntü sınıflandırma

INTRODUCTION

Breast cancer remains a leading cause of cancer-related mortality in women worldwide.[1] Early detection is critical, as it not only improves survival rates but also leads to more effective treatment options. Ultrasonography has emerged as a key imaging modality in this context owing to its accessibility, low cost, lack of ionizing radiation, and ability to provide real-time visualization of breast tissue architecture.[1,2] However, despite its advantages, ultrasound imaging is inherently operator-dependent, and its image interpretation can be highly challenging, which may result in diagnostic variability.[1] To overcome these limitations, deep learning offers a novel solution for automated lesion classification.

Several studies have demonstrated that deep learning models, including convolutional neural networks (CNNs), can effectively identify and classify regions of interest within ultrasound images by learning hierarchical representations of features.[3-5] For instance, Cao et al.[3] compared multiple deep learning architectures for lesion detection and classification, underscoring the potential of CNNs to delineate lesion boundaries more consistently than manual methods. Similarly, Vigil et al.[4] introduced a dual-purpose deep learning model that concurrently detects and diagnoses breast lesions in ultrasound images, highlighting the benefits of integrated approaches for improving diagnostic consistency.

Advances in feature extraction through discriminative deep learning frameworks further enhance the performance of automated systems. Yu et al.[5] demonstrated that employing deep feature extraction from targeted regions in ultrasound images can lead to improved accuracy in differentiating between benign, malignant, and normal tissues. Moreover, incorporating attention mechanisms, as proposed by Kalafi et al.,[6] helps the model focus on the most diagnostically relevant parts of the image, thereby addressing the ambiguity inherent in ultrasound interpretation. Such strategies may ultimately reduce the incidence of unnecessary biopsies while ensuring high sensitivity in malignancy detection.

The integration of these models not only promises greater diagnostic consistency but also reduces operator variability and enhances clinical decision support systems. Furthermore, as shown by Yap et al.,[2] automated approaches based on CNNs offer scalable solutions that facilitate rapid and reliable lesion detection, potentially contributing to earlier intervention and improved patient outcomes.

This study aimed to develop and validate a machine learning model to classify breast ultrasound images as benign, malignant, or normal, thereby enhancing diagnostic accuracy and supporting radiologists. Our approach is distinguished from previous work through several key innovations: First, we systematically integrate deep features from three complementary CNN architectures (VGG16, ResNet50V2, EfficientNetB3) rather than relying on single architectures. Second, we implement a comprehensive multi-step feature optimization pipeline combining principal component analysis (PCA), recursive feature elimination (RFE) with LightGBM, and partial least squares discriminant analysis (PLS-DA), a more sophisticated approach than typically employed in breast ultrasound classification. Third, we utilize a stacking ensemble methodology that integrates four diverse base learners (LightGBM, XGBoost, CatBoost, random forest) with logistic regression meta-learning, going beyond simple voting or averaging approaches. Most importantly, our model is trained and validated on a robust, heterogeneous dataset created by systematically merging five publicly available collections, representing diverse imaging conditions, patient populations, and clinical settings, addressing the generalizability limitations inherent in single-dataset studies. This comprehensive approach was designed to achieve high accuracy and interpretability while improving classification robustness across diverse clinical scenarios, ultimately supporting radiologists in making accurate diagnoses and improving patient outcomes.

METHODS

The study was conducted using five publicly available and anonymized breast ultrasound datasets. Ethical approval for this specific analysis was waived as it involved secondary use of non-identifiable data.

Data Collection and Preprocessing

Our study utilized a comprehensive dataset created by merging five publicly available breast ultrasound

image collections: Breast Ultrasound Dataset from Universidad de Castilla-La Mancha, breast ultrasound lesion segmentation dataset[7], breast ultrasound images dataset (BUSI)[8], breast ultrasound images database[9], breast ultrasound classification dataset[10], and breast-lesions- ultrasonography dataset.[11] Details of the datasets are shown in Table 1. This approach allowed us to address the limitations of individual datasets while creating a more robust and diverse collection for training our classification model. These lesions were defined as separate cases In the presence of multiple masks belonging to an image containing more than one lesion. All data were classified as benign, malignant, and normal and organised them in separate directories according to their labels with corresponding masks.

Preprocessing was performed using Python 3.8 with OpenCV (version 4.5.5). Images were processed by applying corresponding masks to isolate regions of interest and increase focus on clinically relevant areas. The masked images were resized to a uniform size of 224x224 pixels and converted to red-green-blue colour space, standardized for compatibility with pre-trained deep learning models used in feature extraction. The experiments were conducted on an Apple M4 chip with 16 GB random-access memory, without a dedicated graphics processing unit.

Deep Feature Extraction

Three distinct widely-used CNN architectures, pre-trained on the ImageNet dataset, were selected as feature extractors: VGG16, ResNet50V2, and EfficientNetB3 implemented using TensorFlow 2.10.0. These models were chosen for their proven performance in medical imaging tasks and varying architectural complexities. Models were loaded without their final classification layers, allowing access to the rich, hierarchical feature representations learned during their original training.

Each image was preprocessed (e.g., normalized and scaled) for model compatibility. The CNNs then processed the masked ultrasound images, generating feature representations that captured patterns ranging from low-level textures to high-level semantic features. The outputs from all three models were concatenated into a composite feature vector per image, providing a comprehensive representation of lesion characteristics.

Feature Selection

To address the high dimensionality of the concatenated feature vectors, a multi-step feature selection process was implemented using Scikit-learn 1.0.2. Initially, the feature matrix was standardized using StandardScaler to ensure uniform scaling across features.

Then, PCA was applied to reduce dimensionality while preserving 95% of the variance. This step eliminated redundant and noisy features, transforming the high-dimensional feature vectors (200,704 dimensions) into a more manageable set of principal components, facilitating subsequent analysis.

Next, we used RFE with a LightGBM model (version 3.3.2) to select the top 50 features based on their importance scores. RFE iteratively removes the least significant features, ensuring that only the most informative features

Table 1. Composition of the breast ultrasound datasets

	BUS-UCLM[7]	Breast ultrasound dataset (BUSI)[8]	Breast ultrasound images database[9]	BUSC dataset[10]	Breast-lesions-USG[11]
Total images	683	780	232	250	256
Normal	419	133	0	0	4
Benign	174	437	109	100	154
Malignant	90	210	123	150	98
Number of patients	38	600	Not specified	Not specified	256
Number of radiologists	2	Not stated	1	Not stated	4
Ultrasound scanner	Siemens ACUSON S2000TM	GE LOGIQ E9 and LOGIQ E9 Agile ultrasound system	AixPlorer Ultimate, Supersonic Imagine ultrasound machine	Not stated	• Hitachi ARIETTA 70 • Esaote 6150 • Samsung RS85 • Philips Affiniti 70G and EPIQ 5G
Image file format	.png	.png	.bmp (image), .tif (mask)	.png	.png
Histopathological confirmation	Yes	Not stated	Yes	Not stated	Yes

BUS-UCLM: Breast ultrasound dataset from Universidad de Castilla-La Mancha, BUSC: Breast ultrasound classification dataset, USG: Ultrasonography

are retained. This step refined the feature set by focusing on those most relevant to the classification task.

Finally, we applied PLS-DA to the selected 50 features to project them into a lower-dimensional space optimized for class separation. Unlike PCA, which maximizes variance, PLS-DA prioritizes features that maximize the distinction between benign, malignant, and normal classes, enhancing the discriminative power of the feature set for the stacking classifier.

Following feature selection, we prepared the dataset for model training to ensure robust performance.

Data Splitting and Model Training

The dataset, comprising 2233 breast ultrasound images (1005 benign, 672 malignant, 556 normal), exhibited class imbalance, with benign images constituting 45.0%, malignant 30.1%, and normal 24.9% of the total. We split the dataset into training (75%) and test (25%) sets using stratified sampling to maintain these class proportions. This resulted in a training set of 1675 images (754 benign, 504 malignant, 417 normal) and a test set of 558 images (251 benign, 168 malignant, 139 normal). To address the class imbalance in the training set, where the normal and malignant classes were underrepresented compared to the benign class, we applied the synthetic minority oversampling technique (SMOTE). SMOTE generated synthetic samples for the minority classes (malignant and normal), balancing the training set while preserving the original data distribution in the test set for unbiased evaluation.

We developed a stacked ensemble classifier using four base models: LightGBM (version 3.3.2), XGBoost (version 1.6.2), CatBoost (version 1.0.6), and random forest (scikit-learn 1.0.2). These models were trained on selected features to leverage their unique decision boundaries. Then, their predictions were integrated using logistic regression as a meta-learner, with 5-fold cross-validation ensuring robustness during training. Model hyperparameters are shown in Table 2.

Model performance is assessed on the test set through multiple metrics. A confusion matrix evaluates classification accuracy per class, identifying potential misclassifications. Receiver operating characteristic (ROC) curves are plotted for each class, with area under the curve (AUC) scores calculated to quantify discriminatory ability. Model pipeline is summarized in Figure 1.

RESULTS

The dataset consisted of 2233 breast ultrasonography images, categorized as benign (1005), malignant (672), and normal (556). All images were successfully processed, with features extracted from the regions of interest defined by their corresponding masks. Feature extraction was performed using pre-trained VGG16, ResNet50V2, and EfficientNetB3 models, and the resulting features were concatenated to form a high-dimensional feature vector of 200,704 dimensions for each image. Dimensionality reduction was then applied using PCA to retain features explaining 95% of the variance, followed by exploration of class-discriminative dimensionality reduction via PLS-DA (Figure 2).

We evaluated the model's performance using metrics such as AUC-ROC, AUC-precision-recall (PR), F1-score, precision, and recall for each class, plus macro averages (Table 3). Notably, the system yielded high precision and recall for the benign and normal classes, with particularly strong performance in identifying normal cases, as evidenced by a recall of 1.000.

The model achieved high AUC scores: 0.984 for benign, 0.916 for malignant, and 0.969 for normal classes (Figure 3). The benign class achieved the highest AUC of 0.984, with a sharp curve near the top-left corner, indicating excellent sensitivity and specificity with minimal false positives. The normal class followed with an AUC of 0.969, reflecting strong performance consistent with its perfect recall. The malignant class had an AUC of 0.916, with a more gradual curve suggesting a higher balance between sensitivity and specificity due to the complexity of identifying malignant lesions. These high AUC scores highlight the model's effectiveness, especially for benign and normal cases. Predictive reliability was further assessed using positive predictive value (PPV) and negative predictive value (NPV), yielding strong results, for benign (PPV: 0.924, NPV: 0.949), malignant (PPV: 0.875, NPV: 0.899), and normal (PPV: 0.863, NPV: 1.000). The F1 scores, which balance precision and recall, were 0.93 for benign, 0.79 for malignant, and 0.92 for normal, with a macro average of 0.88.

Table 2. Hyperparameters of the base models

Model	Key parameters
LightGBM	num_leaves=63, max_depth=15, n_estimators=50
Random forest	max_depth=20, n_estimators=200
XGBoost	eval_metric='logloss'
CatBoost	verbose=0
Stacking	cv=5, final_estimator=LogisticRegression(max_iter=1000)

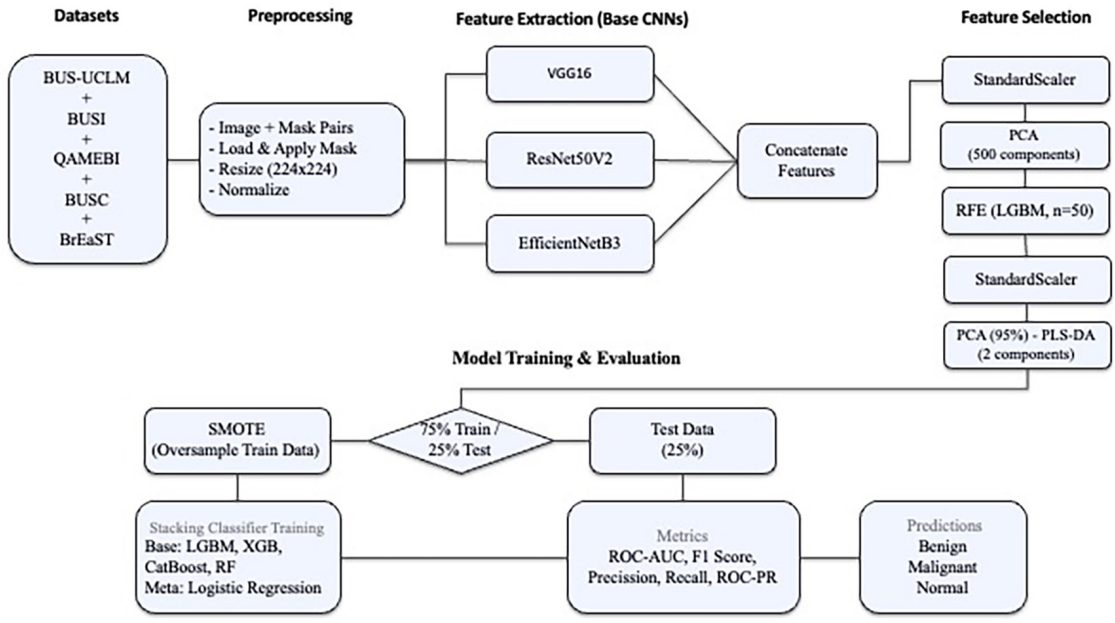

Figure 1. Summary of the model pipeline

BUS-UCLM: Breast ultrasound dataset from Universidad de Castilla-La Mancha, BUSI: Breast ultrasound images dataset, QAMEBI: Breast ultrasound images database, BUSC: Breast ultrasound classification, BrEaST: Breast-lesions-ultrasonography dataset, PCA: Principal component analysis, RFE: Recursive feature elimination, LGBM: Light gradient boosting machine, PLS-DA: Partial least squares discriminant analysis, SMOTE: Synthetic minority oversampling technique, XGB: eXtreme gradient boostin, RF: Random forest, ROC-AUC: Receiver operating characteristic-area under the curve, PR: Precision-recall

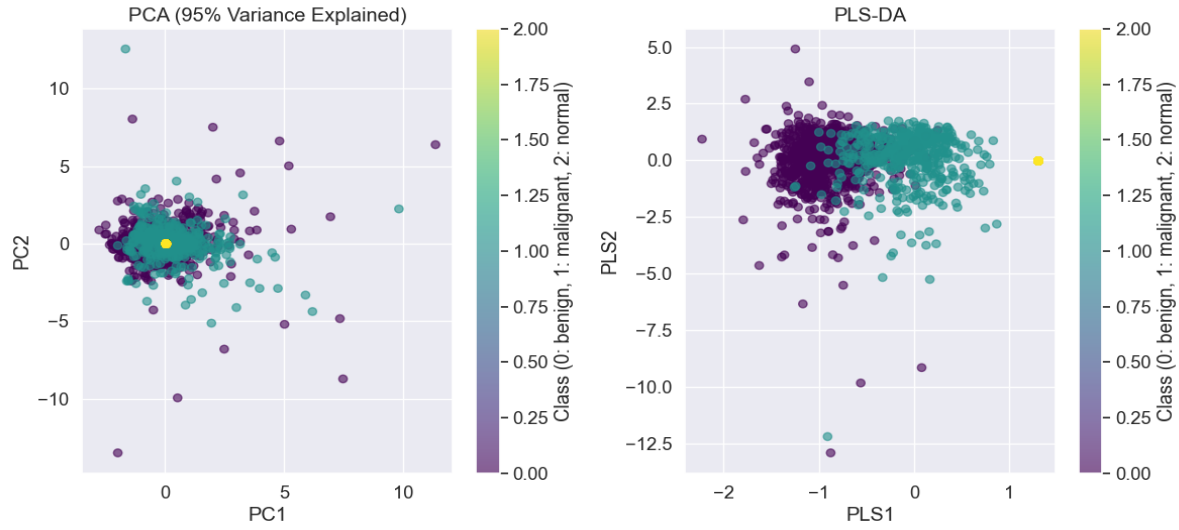

Figure 2. Comparison of dimensionality reduction techniques applied to the selected and scaled image features. (Left) PCA projection onto the first two principal components (PC1, PC2), capturing 95% of the total variance in the features used for classification. (Right) PLS projection onto the first two latent variables (PLS1, PLS2), derived specifically to maximize the separation between classes based on the same feature set

PCA: Principal component analysis, PLS-DA: Partial least squares discriminant analysis

Table 3. Performance metrics of the stacking classifier on the test set					
Class	**AUC-ROC**	**AUC-PR**	**F1-score**	**Precision**	**Recall**
Benign	0.984	0.98	0.93	0.92	0.93
Malignant	0.916	0.83	0.79	0.87	0.71
Normal	0.969	0.86	0.92	0.86	1.00
Macro Avg.	0.956	0.89	0.88	0.88	0.88
AUC-ROC: Area under the curve-receiver operating characteristic, PR: Precision-recall, Avg.: Average					

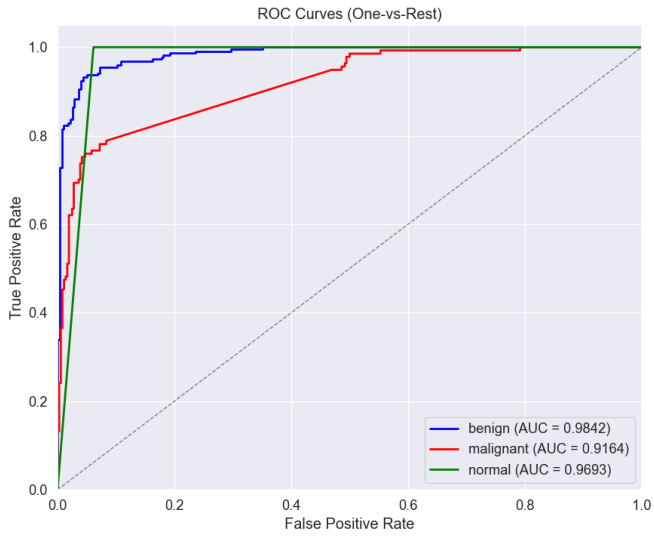

Figure 3. ROC curves for benign (AUC=0.9842), malignant (AUC=0.9164), and normal (AUC=0.9693) classes on the test set

ROC: Receiver operating characteristic, AUC: Area under the curve

The high F1 scores for benign and normal classes reflect the model's ability to achieve both high precision and recall for these classes, while the lower F1 score for the malignant class (0.79) indicates a challenge in balancing precision and recall.

The model's performance was further assessed using PR curves (Figure 4), which show the balance between precision and recall for each class. The PR curves show the model's ability to maintain high precision across varying recall levels. The benign class achieved the highest average precision (AP) score of 0.9803, reflecting the model's strong performance in correctly identifying benign cases with minimal false positives, as indicated by the curve maintaining high precision even at high recall values. The malignant class had an AP score of 0.8305, with the curve showing a more noticeable decline in precision as recall increases, suggesting a trade-off due to the complexity of distinguishing malignant lesions. The normal class, with an AP score of 0.8634, demonstrated a stable PR trade-off,

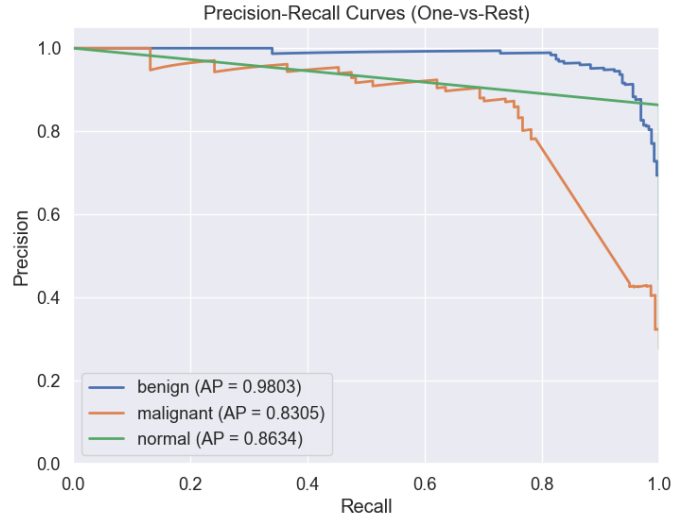

Figure 4. Precision-recall curves for benign (AP=0.9803), malignant (AP=0.8305), and normal (AP=0.8634) classes on the test set

AP: Average precision

reflecting the smaller sample size of normal cases in the dataset, though with a slightly steeper drop in precision at higher recall compared to the benign class.

A summary of the confusion matrix on the test set (251 benign, 168 malignant, 139 normal) reveals the following: 237 benign correctly predicted, with 15 misclassified as malignant and 0 as normal; 111 malignant correctly predicted, with 34 misclassified as benign and 23 as normal; and 139 normal correctly predicted, with 0 misclassified as benign or malignant. This corresponds to 0 false negatives for the benign and normal classes, and 57 false negatives in total for the malignant class (34 as benign, 23 as normal), representing 33.9% of malignant cases. The most frequent confusions occurred in malignant cases, where 34 were misclassified as benign and 23 as normal, highlighting a challenge in distinguishing malignant lesions from other classes.

The feature importance analysis of the 50 selected features reveals a clear contribution ranking among the models: ResNet50V2, VGG16, and EfficientNetB3 (Figure 5). ResNet50V2 leads with a total importance of 3,467 across

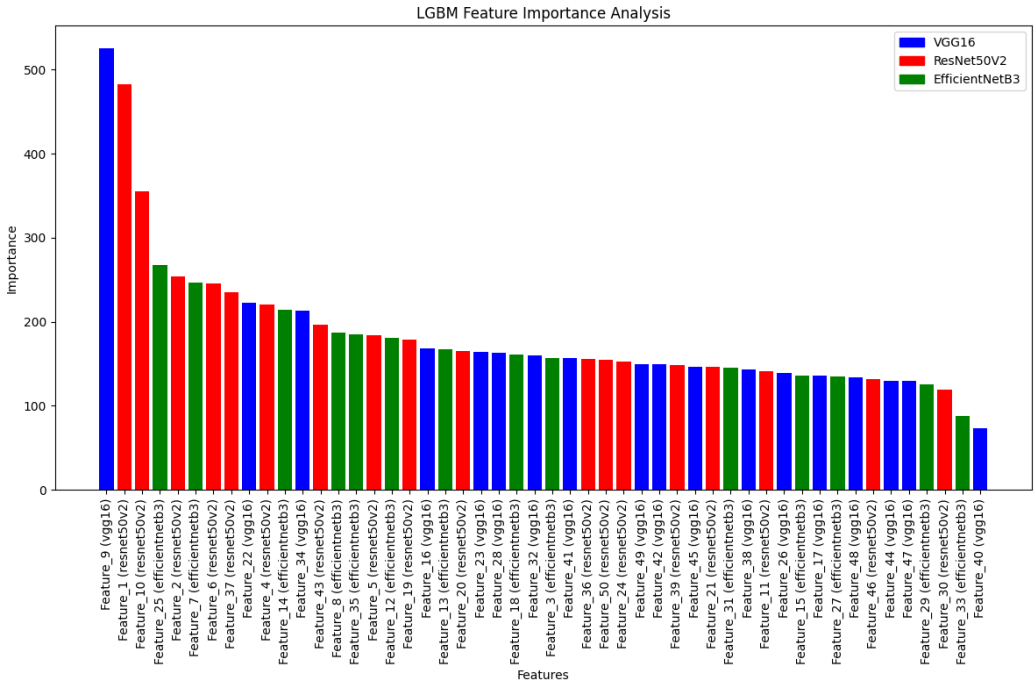

Figure 5. Feature importance analysis of the top 50 selected features from ResNet50V2, VGG16, and EfficientNetB3 models

LGBM: Light gradient boosting machine

18 features, yielding the highest average importance per feature, at 192.61. VGG16 follows closely in importance with a total score of 3,103, contributing 18 features and an average importance per feature of 172.39. EfficientNetB3 ranks third, with a total importance of 2,395 across 14 features, resulting in an average importance per feature of 171.07. These results highlight ResNet50V2's dominant influence in the LightGBM model.

DISCUSSION

Our study successfully developed and evaluated a machine learning model aimed at classifying breast ultrasound images into benign, malignant, and normal categories by leveraging deep features extracted from multiple pre-trained CNNs (VGG16; ResNet50V2; EfficientNetB3), employing sophisticated feature selection methods, and utilizing a stacking classifier. The model was trained and evaluated on a comprehensive dataset aggregated from five distinct public collections; this final dataset represents a heterogeneous population across multiple medical centers, enhancing the potential generalizability of our model for breast lesion classification in diverse clinical settings, a crucial aspect highlighted by studies like Gu et al.,[12] which demonstrated the value of large multi-center datasets. The findings presented in the results section indicate a robust overall performance, highlighted by a macro average AUC-ROC of 0.956 and an F1-score of 0.88 on the unseen test data. The model demonstrated particularly high efficacy in classifying benign lesions (AUC 0.984, F1 0.93) and normal tissue (AUC 0.969, F1 0.92), achieving perfect recall for the normal class. Despite the model's overall strong performance, classifying malignant lesions proved more challenging, as is evidenced by lower metrics (AUC: 0.916; F1-score: 0.79) compared to benign and normal classes.

The model's overall success can be attributed to several key factors inherent in the methodology. Firstly, the extraction of deep features using transfer learning from three powerful, pre-trained CNNs (VGG16; ResNet50V2; EfficientNetB3) provided rich, hierarchical representations of the ultrasound images. Concatenating features from these diverse architectures likely created a more comprehensive feature pool than relying on a single network, capturing a wider range of patterns relevant to classification. For example, Cao et al.[3] achieved 87.5% accuracy using DenseNet on 1043 images with binary classification, while Ellis et al.[13] reported 77.77% accuracy for ResNet50 and 73.80% for VGG-19 on 3-class classification with 571 images. However, direct performance comparison is inappropriate due to different dataset sizes (our dataset of 2,233 images vs. their smaller datasets); preprocessing methods, and validation approaches.

Feature importance analysis confirmed the contribution of all three networks, with ResNet50V2 features showing the highest aggregate importance in the final LightGBM selection step. Secondly, the use of a stacking ensemble classifier, integrating predictions from LightGBM, XGBoost, CatBoost, and random forest via a logistic regression meta-learner, leveraged the strengths of multiple algorithms, while compensating for individual model weaknesses and enhancing predictive robustness and generalization. Ensemble methods, like the one used by Ragab et al.[14] (achieving 97.52% accuracy with VGG-16/19/SqueezeNet + multilayer perceptron with 780 images), often enhance predictive robustness and generalization. Thirdly, employing SMOTE during training helped mitigate the inherent class imbalance in the dataset, likely contributing to the strong performance observed, particularly for the benign and normal classes, which achieved excellent precision, recall, and AUC scores.

However, while the model demonstrated high precision (0.87) for the malignant class, indicating that positive predictions for malignancy are likely correct, the recall (sensitivity) of 0.71 presents a significant concern from a clinical perspective. This recall value implies that approximately 28-29% of actual malignant cases in the test set were misclassified as benign or normal (false negatives). This contrasts sharply with some studies reporting very high recall/sensitivity, such as Yadav et al.[15] who achieved 98.55% overall recall and 90.32% malignant recall using a modified ResNet-101 with 780 images, and Ragab et al.[14] reporting 96.01% overall sensitivity with their ensemble. Even Kalafi et al.,[6] who used an Attention-VGG16 for binary classification with 439 images, reported 96% sensitivity. In clinical practice, failing to detect malignancy has far more severe consequences than misclassifying a benign lesion as malignant, (false positive). The 29% false-negative rate for malignant cases could lead to delayed diagnoses, allowing disease progression that may result in advanced-stage cancer, increased mortality risk, and reduced treatment efficacy. Our lower sensitivity might stem from the inherent subtlety of some malignant lesions, potential limitations of SMOTE for this complex class, or the aggressive feature selection potentially removing crucial subtle features. Future improvements could focus on addressing class imbalance or incorporating additional features to better distinguish malignant characteristics.

Our approach aligns with trends in the literature that utilize deep learning for breast ultrasound analysis.[3-6] We extended the deep feature extraction concept used by Yu et al.,[5] by fusing features from three distinct architectures rather than focusing on specific regions within one architecture (Inception-V3). Our model achieved strong performance, particularly for the benign (AUC: 0.984) and normal (AUC: 0.969) classes. While direct comparison is challenging due to variations in datasets, preprocessing, metrics, and architectures across studies, these results are competitive with or exceed those reported previously (e.g., Zhang et al.[1] 90% AUC using Breast Imaging Reporting and Data System features; Vigil et al.[4] 78.5% accuracy using autoencoder/radiomics; Cao et al.[3] 87.5% accuracy using DenseNet for ternary classification; Gu et al.[12] 0.91 AUC using VGG19 for binary prediction). However, achieving the near-perfect scores reported by Jabeen et al.[16] (99.1% accuracy with augmented BUSI dataset using pre-trained DarkNet-53) or Kiran et al.[17] (100% accuracy with EfficientKNN on a small 780-image dataset) remains challenging, and may depend heavily on their reliance on extensive data augmentation and smaller, potentially less diverse datasets, which may not generalize as effectively to our ternary classification task across a larger, multi-center dataset of 2233 images.

Study Limitations

Several limitations should be acknowledged in this study. The multi-step feature selection, while necessary, might have discarded valuable information; the clinical relevance of the final 50 features needs further investigation. Although the dataset was compiled from multiple sources to enhance diversity, the model's performance was evaluated only on an internal test split. External validation on completely independent datasets from different institutions and ultrasound machines is crucial to assess its generalizability.

Future research should prioritize enhancing the sensitivity (recall) for malignant lesion detection. This may involve exploring cost-sensitive learning algorithms that better address the misclassification of malignant cases, experimenting with alternative data augmentation techniques techniques, advanced oversampling methods (e.g., ADASYN, class weighting within models), or optimizing feature selection. Investigating attention mechanisms (as in Kalafi et al.[6] or Lyu et al.[18] for segmentation) within the feature extractors, exploring architectures known for high performance, such as advanced ResNet and variants, could yield improvements. Investigating alternative approaches, such as end-to-end deep learning models that learn features and classify directly without explicit feature extraction/selection steps, or autoencoders for dimensionality reduction, could also be beneficial.

CONCLUSION

In conclusion, this study demonstrates the considerable potential of combining deep feature extraction from multiple CNNs with advanced feature selection and ensemble learning techniques for classifying breast ultrasound images. This approach represents a promising

frontier in breast cancer diagnostics. The developed model achieved high overall accuracy and discriminatory power, particularly excelling in the classification of benign and normal cases on a diverse, multi-center dataset. These results underscore the model's efficacy and potential as a robust tool for clinical decision support. By mitigating challenges associated with operator dependency and subjective interpretation, such automated methods can offer a reproducible approach to enhance early detection and potentially reduce unnecessary invasive procedures. However, the critical challenge of lower sensitivity (recall) for malignant lesions must be addressed through further research and refinement. While promising as a component of a computer-aided diagnosis system to support radiologists and enhance consistency, this model requires significant improvements in malignant detection and rigorous external validation are essential before it can be reliably integrated into clinical workflows to ultimately support improved patient outcomes.

Ethics

Ethics Committee Approval and Informed Consent: This study used anonymized, publicly available datasets with no personally identifiable information. As it involved no human interaction or new data collection, neither ethics committee approval nor patient consent was required.

Footnotes

Authorship Contributions

Concept: K.P., Design: K.P., Data Collection or Processing: K.P., S.S., Analysis or Interpretation: K.P., S.S., Literature Search: K.P., S.S., Writing: K.P., S.S.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Zhang E, Seiler S, Chen M, Lu W, Gu X. BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. Phys Med Biol. 2020;65:125005.
2. Yap MH, Pons G, Martí J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform. 2018;22:1218-26.
3. Cao Z, Duan L, Yang G, Yue T, Chen Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. BMC Med Imaging. 2019;19:51.
4. Vigil N, Barry M, Amini A, et al. Dual-intended deep learning model for breast cancer diagnosis in ultrasound imaging. Cancers (Basel). 2022;14:2663.
5. Yu H, Sun H, Li J, et al. Effective diagnostic model construction based on discriminative breast ultrasound image regions using deep feature extraction. Med Phys. 2021;48:2920-8.
6. Kalafi EY, Jodeiri A, Setarehdan SK, et al. Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics. 2021;11:1859.
7. Vallez N, Bueno G, Deniz O, Rienda MA, Pastor C. BUS-UCLM: breast ultrasound lesion segmentation dataset. Mendeley Data, V1. 2024. Available from: https://data.mendeley.com/datasets/7fvgj4jsp7/1
8. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief. 2020;28:104863.
9. Abbasian Ardakani A, Mohammadi A, Mirza-Aghazadeh-Attari M, Acharya UR. An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. Comput Biol Med. 2023;152:106438.
10. Iqbal A. BUSC Dataset. Mendeley Data, V1. 2023. Available from: https://data.mendeley.com/datasets/vckdnhtw26/1
11. Pawłowska A, Ćwierz-Pieńkowska A, Domalik A, et al. Curated benchmark dataset for ultrasound based breast lesion analysis. Sci Data. 2024;11:148.
12. Gu Y, Xu W, Lin B, et al. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. Insights Imaging. 2022;13:124.
13. Ellis J, Appiah K, Amankwaa-Frempong E, Kwok SC. Classification of 2D ultrasound breast cancer images with deep learning. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2024;5167-73.
14. Ragab M, Albukhari A, Alyami J, Mansour RF. Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images. Biology (Basel). 2022;11:439.
15. Yadav A, Kolekar M, Zope M. ResNet-101 Empowered Deep Learning for Breast Cancer Ultrasound Image Classification. 2024;763-9.
16. Jabeen K, Khan MA, Alhaisoni M, et al. Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. Sensors (Basel). 2022;22:807.
17. Kiran A, Ramesh JVN, Rahat IS, Khan MAU, Hossain A, Uddin R. Advancing breast ultrasound diagnostics through hybrid deep learning models. Comput Biol Med. 2024;180:108962.
18. Lyu Y, Xu Y, Jiang X, Liu J, Zhao X, Zhu X. AMS-PAN: breast ultrasound image segmentation model combining attention mechanism and multi-scale features. Biomed Signal Process Control 2023;81:104425.