Original Research Article

Machine Learning Models for Breast Cancer Diagnosis Using Ultrasound Images

Arshia Ghatak[1], Jeremy Hitt[2]

[1]*Dougherty Valley High School, 2980 Bollinger Canyon Road, San Ramon, CA 94583, United States;*
[2]*Form Energy Inc., 30 Dane St, Somerville, MA 02143, United States*

ABSTRACT

Breast cancer is one of the most commonly diagnosed and deadliest diseases among women worldwide, and early detection is critical for improving survival outcomes. Ultrasound imaging is widely used in breast cancer screening due to its safety, low cost, and effectiveness in visualizing soft tissue. Image interpretation is highly dependent on radiologist expertise and, therefore, can be subject to variability. This study explores the use of classical machine learning approaches for automated breast ultrasound classification, emphasizing interpretability and reliability in small, clinically annotated datasets. Using a publicly available breast ultrasound dataset, images were preprocessed and transformed into engineered feature representations capturing texture, shape, and intensity characteristics commonly used in clinical assessment. Three supervised machine learning models - Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) - were trained and evaluated to classify images as benign, malignant, or normal tissue. To address class imbalance, random oversampling was applied, and model performance was assessed using class-specific accuracy metrics and confusion matrices. The ensemble achieved improved overall classification performance and enhanced malignant tissue detection compared with each individual classifier. These results demonstrate that ensemble-based classical machine learning methods offer a practical, interpretable, and low-resource approach for automated breast cancer detection using ultrasound imaging. The individual models, Random Forest demonstrated the strongest and most balanced performance across all tissue classes, with texture- and shape-based features contributing most significantly to its predictions. A weighted ensemble voting classifier was then implemented to combine the strengths of all three models, assigning greater influence to the Random Forest based on validation performance. A deployable graphical user interface was also developed to make the system accessible to both clinicians and non-experts.

Keywords: Breast Cancer; Ultrasound Imaging; Machine Learning; Support Vector Machine; Random Forest; Multilayer Perceptron

INTRODUCTION

Among women worldwide, breast cancer is one of the most common and lethal diseases; it is diagnosed in millions of patients every year (1). Early diagnosis significantly improves survival because treatment is more effective when made before metastasis. Ultrasound has become a modality of great importance in current diagnostic imaging methods due to its safety, low

Corresponding author: Arshia Ghatak, E-mail: arshia.ghatak@gmail.com.
Copyright: © 2026 Arshia Ghatak *et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Accepted February 18, 2026
https://doi.org/10.70251/HYJR2348.41777784

cost, and good visualization of soft-tissue structures without radiation exposure (2). However, ultrasound interpretation is highly dependent on the expertise of the radiologist, which means variability and potential human error may arise.

The integration of machine learning (ML) and artificial intelligence (AI) into breast ultrasound analysis is a promising approach to achieve more uniform and automated detection of cancer. These models are trained to find subtle textural or morphological variations in imaging data that may be hard for human observers to detect (3). Machine learning has already demonstrated remarkable success in other aspects of medical diagnostics, such as the prediction of cardiovascular and pancreatic diseases, using algorithms including random forests, neural networks, and support vector machines that have attained outstanding performance in identifying patterns of disease from clinical data (4). Previous studies on breast ultrasound classification further show that both traditional machine-learning algorithms and deep neural networks can reliably distinguish benign from malignant lesions with high accuracy, achieving Area under the curve (AUC) values above 90% (5). Such findings strengthen the validity of ML systems in complementing radiologists and reducing diagnostic subjectivity in breast cancer screening (6).

Given the variability of human interpretation in ultrasound imaging and the limitations of deep learning approaches on small, clinically annotated datasets, this study addresses a gap in the literature by conducting a comparative analysis of interpretable, classical machine learning models for breast ultrasound classification. While these methods have shown strong performance in medical imaging tasks, they typically require large datasets and substantial computational resources. Three models were implemented and evaluated on a labeled dataset of ultrasound images: support-vector machines (SVM), random forests (RF), and multilayer perceptrons (MLP). Each of these models brings different strengths: SVMs perform well on high-dimensional datasets, while random forests are good for noisy and nonlinear data; finally, MLPs are able to learn complex spatial representations within the images. By comparing these models, the goal was to find the best algorithm for tumor detection and type identification. By systematically comparing these models and deploying a weighted ensemble classifier, this study aims to identify the most robust and interpretable classical machine learning approach for automated breast ultrasound classification, providing a practical, low-resource framework suitable for exploratory research, without making claims about performance relative to deep learning models. Additionally, a graphical user interface (GUI) was developed to enable users to interact with the trained models and obtain classification results without requiring code execution. Results indicate that Random Forest models and the weighted ensemble achieve the most balanced performance, particularly in malignant tissue detection, while maintaining interpretability and low computational requirements.

METHODS AND MATERIALS

Dataset and Preprocessing

This study used the Breast Ultrasound Images Dataset published by Al-Dhabyani et al. (7), accessed through its publicly available Kaggle distribution (8); all images originate from this single dataset. The dataset contains 597 images across these three classes: 309 benign, 159 malignant, and 128 normal (2)(9). Each image was manually labeled by medical professionals and cropped to focus on the lesion region. Representative examples of benign, malignant, and normal ultrasound images from the dataset are shown in Figure 1.

All images were preprocessed in two main stages. During image loading, each file was opened as a red-green-blue (RGB) image using the Python Pillow library, resized to 224 × 224 pixels for dimensional consistency, and converted to a NumPy array for downstream processing. In the feature extraction stage, grayscale conversion and texture analysis were performed using OpenCV. The extracted features included the mean and standard deviation of each RGB channel, texture statistics such as mean, standard deviation, and energy calculated from Sobel gradient magnitudes, shape features including area, perimeter, aspect ratio, circularity, and compactness derived from thresholder contours, and the first five bins from each hue-saturation-value (HSV) color histogram channel. These features

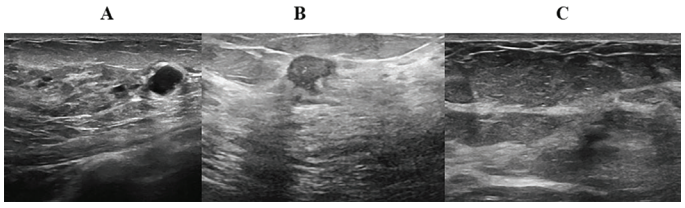

Figure 1. *Example ultrasound images from the Breast Ultrasound Dataset showing (A) benign on the left, (B) malignant in the middle, and (C) normal tissue on the right.*

were selected to capture both the macroscopic shape and the microscopic texture characteristics of breast lesions, which radiologists commonly use to distinguish between benign, malignant, and normal tissue in ultrasound imaging. All features were standardized using StandardScaler to ensure uniform scaling across attributes before model training.

Data Balancing and Oversampling

The original dataset was highly imbalanced, with 309 samples in the benign class, 159 in the malignant class, and 128 in the normal class. To prevent bias during training, a simple random oversampling strategy was applied to the minority classes using resampling with replacement. The goal was to equalize all classes to the size of the majority class (309 samples). The final balanced dataset contained 927 samples (309 per class). The class distributions before and after oversampling are summarized in Table 1. This process did not generate new synthetic images but ensured equal class representation during training.

Model Development

Three supervised machine learning models were trained and compared in terms of classification performance. The MLP was included to evaluate the ability of a neural network to learn complex, nonlinear relationships in the extracted features (10).

The performance implications of these model configurations are evaluated and compared in the Results section. A weighted ensemble voting system was created to combine predictions from the SVM, Random Forest, and MLP models. Each model generated probability scores for the benign, malignant, and normal classes. These probabilities were combined using predefined weights (Table 2), which gave greater weight to the Random Forest model because it performed best during validation. The weighted probabilities were added together, and the class with the highest total score was selected as the final prediction. This method allowed the system to balance all three classifiers and produce more stable and reliable overall predictions.

Table 1. Distribution of breast ultrasound images before and after random oversampling.

Class	Before	After	Duplicates Added
Normal	128	309	181
Malignant	159	309	150
Benign	309	309	0

Graphical User Interface (GUI) Design

To make the trained models accessible for practical use, a standalone Graphical User Interface (GUI) was developed using Tkinter with ttk widgets for styling and Pillow for image rendering. The application integrates a weighted ensemble classifier that combines the three trained models, favoring Random Forest predictions based on validation performance.

Upon launch, the GUI will present a clean, labeled interface titled *"Breast Cancer Tumor Classifier."* Users can select a model (Random Forest, SVM, or MLP) from a dropdown, choose an ultrasound image to classify, and preview it in a dedicated display pane. Once the *"Predict"* button is clicked, the application asynchronously generates classification results, confidence scores, and per-model probabilities. The interface also includes an animated progress bar, status updates, and a read-only output panel summarizing ensemble decisions and confidence levels. This interface allows researchers and students to interact with the trained models and obtain classification outputs for exploratory and educational purposes without navigating command-line environments. A screenshot of the implemented graphical user interface is shown in Figure 2. The GUI implementation and its performance results are presented in the Results section to demonstrate practical application and usability.

Table 2. Machine learning models and their key parameters used for breast ultrasound classification.

Model	Key Parameters	Weight from Ensemble Voting Model
Support Vector Machine (SVM)	radial basis function (RBF) kernel, C = 2.0, gamma = 'scale'	15%
Random Forest (RF)	100 estimators, default max depth, random_state = 42	60%
Multilayer Perceptron (MLP)	two hidden layers (100, 50), ReLU activation, Adam optimizer, max_iter = 500	25%

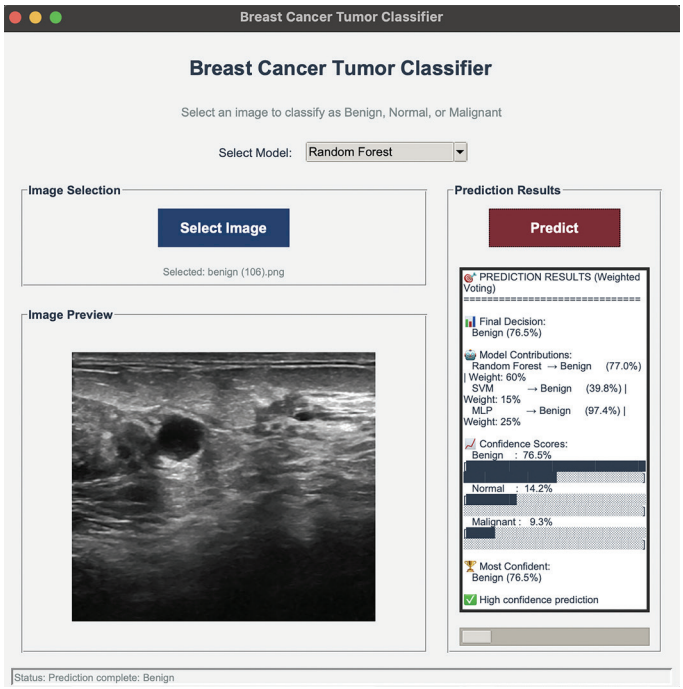

Figure 2. *Screenshot of the graphical user interface (GUI) displaying a classification result for an uploaded ultrasound image.*

RESULTS

Model Validation

All models were implemented using Scikit-learn and trained on the extracted feature vectors. The dataset was split using stratified sampling to preserve class proportions, with 70% allocated for training and 30% for testing. This 70/30 split is a common practice in machine learning, balancing sufficient data for model training with a representative portion for testing and evaluation. Model training and evaluation were performed using the Python package scikit-learn. Each model was evaluated using accuracy and a confusion matrix to visualize the distribution of true versus predicted classes.

Model Performance

The performance of the three supervised machine learning models - Support Vector Machine (SVM), RF, and MLP - was evaluated on the test dataset using accuracy and precision. Confusion matrices were generated to visualize true versus predicted classifications across benign, malignant, and normal tissue classes. A quantitative summary of class-specific classification accuracy for each model is provided in Table 3.

Figure 3 presents the confusion matrices for the SVM, Random Forest, MLP, and ensemble models arranged side by side. The SVM correctly classified 51.6% of benign samples, 57.0% of normal samples, and 67.7% of malignant samples, while misclassification was more frequent between benign and malignant classes as well as between benign and normal classes. This indicates that SVM struggled to fully capture the nonlinear and high-dimensional patterns in the dataset, particularly when distinguishing between benign and normal tissue. These results are shown in the confusion matrix in Figure 3A.

The Random Forest model demonstrated the highest individual performance, with 67.7% of benign, 89.2% of normal, and 78.5% of malignant samples correctly classified. Misclassifications occurred primarily between benign and malignant classes, but overall, the model produced balanced predictions across all tissue types. The corresponding confusion matrix is shown in Figure 3B. Based on the feature importance analysis, texture and shape were the strongest discriminators of tumors in the RF model, so those may be more pronounced in the normal images than the other two.

Feature Importance Analysis

To understand how the Random Forest model makes its predictions, a feature importance analysis was performed. The top ten most influential features identified by the Random Forest model are shown in Figure 4, ranked from greatest to least importance,

Table 3. Summary of class-specific classification accuracy (%) for all models on the test dataset.

Model	Benign (%)	Normal (%)	Malignant (%)
Support Vector Machine (SVM)	51.6	57.0	67.7
Random Forest (RF)	67.7	89.2	78.5
Multilayer Perceptron (MLP)	53.8	87.1	82.8
Weighted Ensemble	59.1	89.2	82.8

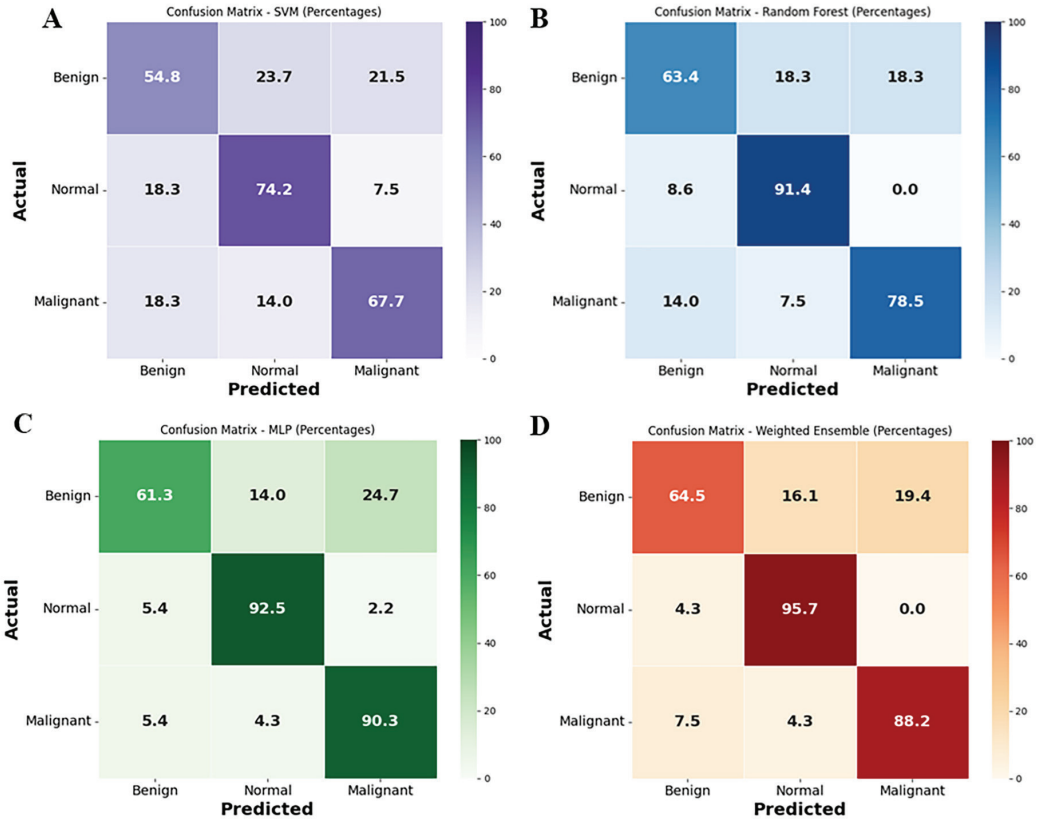

Figure 3. *Confusion matrices for (A) SVM, (B) Random Forest, (C) MLP models and (D) for the weighted ensemble voting system, which showed improved overall classification relative to the individual models.*

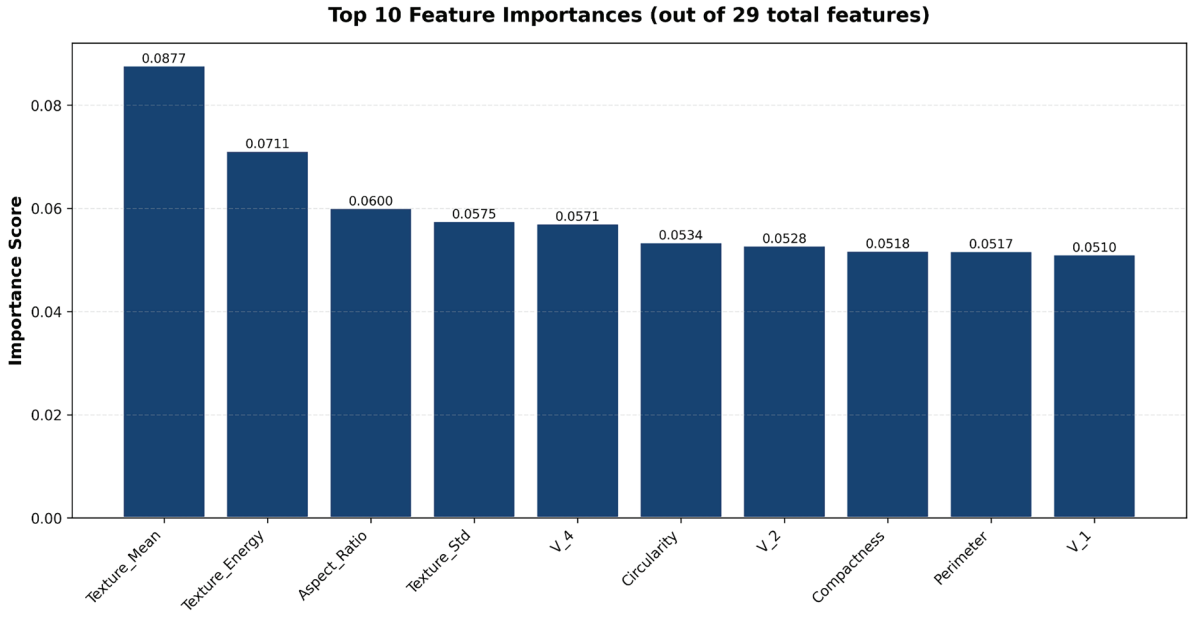

Figure 4. *Bar graph of the top ten features used by the Random Forest model, ranked by importance. Texture- and shape-related features contributed most to classification.*

including Texture_Mean, Texture_Energy, Aspect_Ratio, and Texture_Std. Texture- and shape-related features such as Texture_Mean, Texture_Energy, Aspect_Ratio, and Texture_Std contributed most to the model's decisions, while color histogram features (V_1, V_2, and V_4) and other geometric features were relatively less important. The bottom ten features with the lowest importance values are shown in Figure 5, including S_5, H_4, S_3, and H_3. Hue- and saturation-related features such as S_5, H_4, and H_3 contributed very little to the model's decisions, with importance values near zero. This indicates that color-based information had minimal impact on classification, while texture- and shape-related features dominated the predictive power, a finding that is consistent with radiological assessment frameworks such as Breast Imaging Reporting and Data System, which emphasize lesion shape, margin characteristics, and internal texture over color-based information. This analysis provides insight into which characteristics of the ultrasound images are most predictive of tissue type and helps explain why the Random Forest model performs strongly. These insights informed the construction of the weighted ensemble classifier, which gives more influence to the Random Forest model due to its relatively stronger performance within this dataset.

The MLP performed moderately, correctly classifying 53.8% of benign, 87.1% of normal, and 82.8% of malignant samples. Although it was slightly less consistent than Random Forest in distinguishing benign and malignant tissues, its predictions could offer complementary perspectives that the ensemble classifier is designed to incorporate. Its confusion matrix is presented in Figure 3C.

In addition to overall classification performance, the effect of MLP architectural depth was evaluated. The MLP's overall performance was slightly lower than that of the Random Forest model, particularly in distinguishing malignant cases. The MLP was tested with five hidden layers; however, its performance differed by less than 3% from that of the two-hidden-layer configuration. Therefore, only two hidden layers were used for the remainder of the analysis. Nonetheless, it contributed to the ensemble classifier by providing complementary perspectives on the data.

Ensemble Classification Performance

A weighted ensemble voting system was implemented to combine the predictions of all three models. Probabilities generated by each model were weighted by validation performance, with Random Forest receiving greater influence due to its superior accuracy. The ensemble demonstrated improved overall classification relative to individual classical models on this dataset, particularly for malignant tissue, with 59.1% of benign,

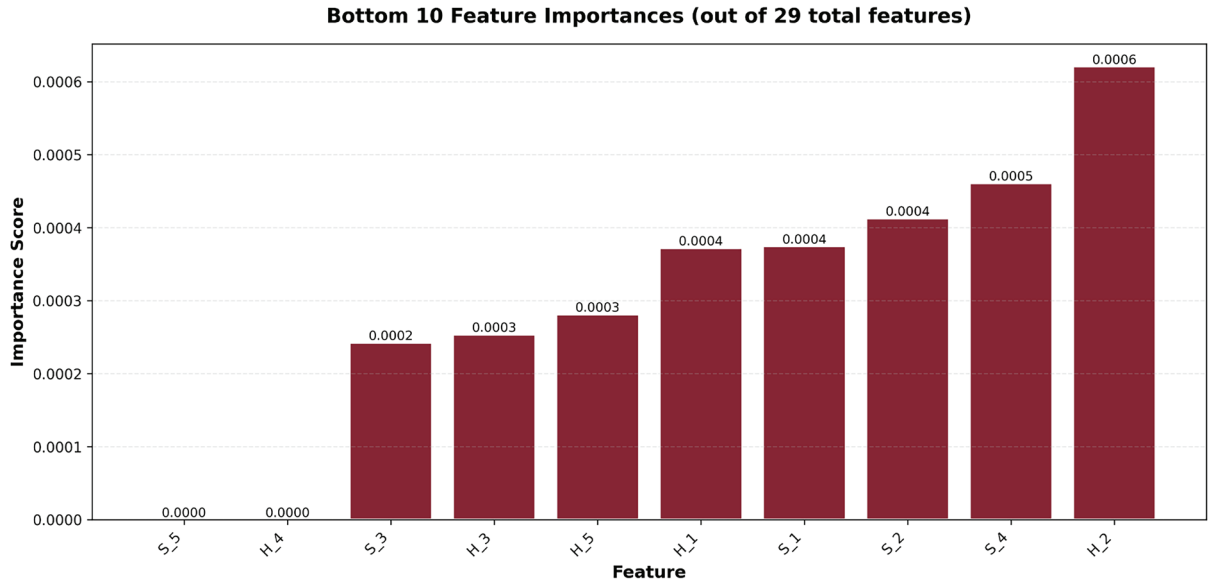

Figure 5. *Bar graph of the bottom ten features used by the Random Forest model, ranked by importance. Hue- and saturation-related features contributed least to classification.*

89.2% of normal, and 82.8% of malignant samples correctly classified; these results are exploratory and not clinically deployable. Notably, the ensemble slightly reduced the accuracy for benign tissue compared with the Random Forest alone (59.1% vs. 67.7%), reflecting a trade-off between balancing overall classification performance and improving malignant detection. Although the SVM's lower performance, particularly in distinguishing benign from malignant samples, slightly reduced the accuracy of some ensemble predictions, the combined system successfully leveraged the strengths of Random Forest and the complementary insights from MLP. Figure 3D shows the confusion matrix for the ensemble predictions. Each row represents the true class (benign, normal, malignant), while each column represents the predicted class. The numbers in each box indicate the number of samples classified correctly or misclassified, enabling a visual assessment of which classes are most frequently confused.

The weighted ensemble voting system demonstrated improved overall classification performance compared with individual models, particularly in detecting malignant tissue, while misclassifications were still most frequent between benign and malignant classes. These results establish a foundation for further analysis and interpretation in the Discussion.

DISCUSSION

The evaluation of the three machine learning models - SVM, Random Forest, and MLP - along with the weighted ensemble system provides several key insights into automated breast ultrasound classification. Random Forest consistently exhibited the strongest individual performance, particularly in detecting malignant tissue, and also determined the most predictive feature importances, which included texture- and shape-based features. The MLP also performed comparably for malignant tissue detection (82.8%), highlighting that it provides complementary predictive value despite slightly lower overall consistency across tissue types. The MLP captured complementary nonlinear patterns in the data, contributing unique perspectives that strengthened the ensemble classifier. All three models tended to perform worse on benign images, indicating that distinguishing benign tissue from malignant or normal tissue remains the greatest challenge. Its weaker performance slightly limited some ensemble predictions, highlighting the importance of carefully weighing individual model contributions.

Beyond individual model performance, many existing approaches to automated breast ultrasound classification focus on deep learning-based models, such as convolutional neural networks, which operate directly on raw pixel data and aim to learn hierarchical feature representations. In contrast, this work emphasizes classical machine learning models trained on carefully engineered texture and shape features, which offer improved interpretability and robustness in small-data settings. By systematically comparing multiple classifiers and integrating them into a weighted ensemble, this study highlights how complementary models can improve classification stability, particularly for malignant tissue detection. Additionally, the inclusion of a deployable graphical user interface extends this work beyond offline evaluation toward practical usability.

Overall, the Random Forest model produced the most accurate and balanced predictions, while MLP contributed additional nonlinear information. The ensemble classifier benefited from combining these complementary strengths, achieving better overall accuracy and stability than any single model. This was most evident in malignant tissue detection, where the ensemble reduced misclassifications compared with the SVM or Random Forest alone.

An important consideration in medical classification tasks is the balance between false positives and false negatives, as each type of error carries different clinical implications. False negatives, particularly in malignant cases, are of greatest concern because they may delay diagnosis and treatment, while false positives can lead to unnecessary follow-up procedures and patient anxiety. In this study, misclassification was most frequently observed between benign and malignant tissue across all models, reflecting the visual similarity of these classes in ultrasound imaging. To mitigate these errors, a weighted ensemble strategy was employed to reduce reliance on any single model's predictions and to improve stability in malignant tissue detection. By assigning greater weight to the Random Forest classifier, which demonstrated stronger performance and more reliable feature utilization, the ensemble reduced false negatives for malignant cases relative to individual models (malignant accuracy: SVM 67.7%, RF 78.5%, MLP 82.8%, ensemble 82.8%). Additionally, stratified sampling and class balancing were used to minimize bias toward majority classes, further supporting more equitable distribution of errors across tissue types.

Additionally, the study reinforces the value of feature analysis for interpretability. By identifying the most influential features, such as Texture_Mean, Texture_

Energy, Aspect_Ratio, and Texture_Std, the Random Forest model provides insight into which characteristics of ultrasound images are most predictive. Understanding these contributions can inform both future model optimization and clinical decision-making, as knowing which features are most important may help doctors prioritize imaging tests or measurements that emphasize these characteristics.

Collectively, the findings suggest that ensemble learning, when carefully constructed, is a promising approach for automated breast cancer detection, offering improved reliability and interpretability while leveraging the complementary strengths of different machine learning models.

CONCLUSION

This study evaluated the performance of three supervised machine learning models - Support Vector Machine, Random Forest, and Multilayer Perceptron - on the classification of breast ultrasound images into benign, malignant, and normal tissue. Random Forest demonstrated the highest individual predictive accuracy, with texture- and shape-based features contributing most significantly to its decisions. The MLP also provided complementary nonlinear insights, while the SVM exhibited limitations in distinguishing benign from malignant tissue.

By integrating these models into a weighted ensemble classifier, the study achieved improved classification stability and enhanced detection of malignant tissue compared with any individual model. The ensemble approach highlights the benefits of combining diverse algorithms to leverage their respective strengths and mitigate weaknesses.

These findings demonstrate the potential of classical machine learning and ensemble methods as interpretable research tools for studying automated breast ultrasound classification in constrained data settings. Future work may focus on further optimizing model architectures, exploring additional feature representations, and validating performance on larger and more diverse datasets to facilitate clinical deployment.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to this work.

REFERENCES

1. World Health Organization: WHO, World Health Organization: WHO. Breast cancer [Internet]. 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 2025-9-24)
2. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data in Brief.* 2020 Feb; 28: 104863. https://doi.org/10.1016/j.dib.2019.104863
3. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*. 2020 Sep 29; 10 (1). https://doi.org/10.1038/s41598-020-72685-1
4. Byra M, Karwat P, Ryzhankow I, Komorowski P, Klimonda Z, Fura L, et al. Deep meta-learning for the selection of accurate ultrasound based breast mass classifier [Internet]. arXiv.org. 2022. Available from: http://arxiv.org/abs/2211.01892 (accessed on 2025-8-11)
5. Qasrawi R, Daraghmeh O, Thwib S, Qdaih I, Issa G, Polo SV, et al. Advancing breast cancer detection in ultrasound images using a novel hybrid ensemble deep learning model. *Intelligence-Based Medicine.* 2025; 11: 100222. https://doi.org/10.1016/j.ibmed.2025.100222
6. Rezazadeh A, Jafarian Y, Kord A. Explainable Ensemble Machine Learning for Breast Cancer Diagnosis based on Ultrasound Image Texture Features [Internet]. arXiv.org. 2022. Available from: http://arxiv.org/abs/2201.07227 (accessed on 2025-8-11)
7. Shah A. Breast Ultrasound Images Dataset [Internet]. www.kaggle.com. 2018. Available from: https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset (accessed on 2025-7-10)
8. The Cancer Imaging Archive. BrEaST Lesions USG. [Internet]. Cancerimagingarchive.net. 2023. Available from: http://www.cancerimagingarchive.net/wp-content/uploads/BrEaST-Lesions-USG-clinical-data-Dec-15-2023.xlsx (accessed on 2025-7-10)
9. Uysal F, Köse MM. Classification of Breast Cancer Ultrasound Images with Deep Learning-Based Models. Engineering Proceedings [Internet]. 2022; 31 (1): 8. Available from: https://www.mdpi.com/2673-4591/31/1/8#B7-engproc-31-00008 (accessed on 2025-8-11)
10. weDevise. 5.1 Benign lesions | Ultrasound Cases [Internet]. Ultrasoundcases.info. Ultrasound Cases; 2025. Available from: http://www.ultrasoundcases.info/cases/breast-and-axilla/benign-lesions/ (accessed on 2025-7-10)