Scientific Journal of Informatics
Vol. 13, No. 3, August 2026
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Comparison and Ensemble Transfer Learning for Breast Cancer Classification in Ultrasound Images Using EfficientNet-B0, ResNet-50, and VGG-16

Anisja Noni Kartikasari[1], Hesti Khuzaimah Nurul Yusufiyah[2], Hanifah Rahmi Fajrin[3]
[1,2]Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia
[3]Department of Software Convergence, Soon Chun Hyang University, Republic of Korea

Abstract.
Purpose: This study aims to compare three transfer learning architectures and develop an ensemble learning approach for breast cancer classification in ultrasound images. The objective of this study is to compare the three transfer learning architectures and evaluate an ensemble learning approach to determine the best model for distinguishing benign and malignant tumors.
Methods: This study used a combined dataset from the BUSI Kaggle challenge and clinical data from Bhakti Dharma Husada General Hospital in Surabaya, consisting of 1,632 images (1,048 benign and 584 malignant). Three pre-trained CNN models, EfficientNet-B0, ResNet-50, and VGG-16, with ImageNet weights, were evaluated using accuracy, precision, recall, and F1-score metrics. Preprocessing steps included resizing to 224×224, ImageNet normalization, and ROI segmentation, with ground-truth labels created using Roboflow while automatic segmentation was performed using DeepLabV3+. The ensemble model employed a soft voting method based on the average probability.
Result: The soft voting ensemble achieved an accuracy of 86.18%, precision of 85.21%, recall of 84.46%, and an F1-score of 84.40% (macro-averaged). Among the four evaluated models, ResNet-50 achieved the highest overall accuracy (87.40%).
Novelty: The novelty of this research lies in the complementary application of an ensemble of three transfer learning architectures (EfficientNet-B0, ResNet-50, and VGG-16) via the probability averaging soft voting method, enhanced by automatic segmentation preprocessing based on DeepLabV3+, and tested on the public BUSI dataset as well as local clinical data from Bhakti Dharma Husada General Hospital in Surabaya.

Keywords: Breast cancer, Ultrasound image, Transfer learning, Ensemble
Received July 2026 / **Revised** September 2026 / **Accepted** September 2026

INTRODUCTION

Breast cancer is one of the most prevalent types of cancer worldwide. According to the 2022 Global Cancer Statistics (GLOBOCAN) data, breast cancer ranks second in cancer cases among women globally, with a total of 2,296,840 cases [1]. In Indonesia, breast cancer is recorded as the cancer with the highest prevalence among women, and its incidence continues to rise each year [2]. In addition to the high number of cases, breast cancer is also one of the leading causes of cancer-related deaths among women [3]. The high mortality rate from breast cancer is largely due to a lack of early detection. Therefore, early detection is a top priority in efforts to improve the success of breast cancer diagnosis and treatment [4]. Early detection via breast ultrasound is a widely used imaging modality because it is non-invasive, does not use ionizing radiation, and is relatively affordable compared to other modalities such as MRI or mammography [5], [6]. Ultrasound images can distinguish between solid and cystic lesions and provide information on the morphology of the mass, which can be used to determine whether it is benign or malignant [6]. However, manual interpretation of ultrasound images is highly dependent on the radiologist's expertise and experience, making it subjective and prone to diagnostic errors [6], [7].

These limitations of manual interpretation drive the need for more objective and consistent computer-based diagnostic systems. Advances in artificial intelligence, particularly *deep learning* based on Convolutional Neural Networks (CNNs), have opened up significant opportunities for the automation of medical image-based diagnosis [8]. CNNs are capable of automatically extracting key features from images without the

*Corresponding author.
Email addresses: hestiyusufiyah@unesa.ac.id (Anisja Noni Kartikasari)
DOI: 10.15294/sji.v8i1.58918

need for manual feature extraction, making them highly suitable for analyzing breast ultrasound images [9]. However, one of the main challenges in applying deep learning to medical data is the limited availability of labeled data, as the annotation process requires the involvement of clinical experts and is *time-consuming* [10]. Transfer learning is an effective approach to address data limitations by utilizing model weights that have been trained on large-scale datasets, followed by fine-tuning for specific medical image classification tasks [10], [11].

The application of transfer learning for breast ultrasound image classification on the BUSI dataset has been explored through various approaches. Twum et al. [12] compared four transfer learning architectures, MobileNetV2, InceptionV3, ResNet-50, and VGG-16, and found that ResNet-50 achieved the highest accuracy, while VGG-16 outperformed the others in terms of precision and recall. However, that study did not include EfficientNet or probability-average-based ensemble approaches in its comparison. Latha et al. [13] proposed EfficientNetB7 combined with data augmentation to address class imbalance in breast ultrasound images, and demonstrated that this architecture outperformed conventional CNNs such as VGG and ResNet in recognizing the minority malignant class. However, their evaluation focused solely on a single architecture without a direct comparison to ResNet-50 and VGG-16 as baselines.

Ikram Ben Ahmed [14] used ResNet50-based transfer learning, supplemented with data augmentation and hyperparameter optimization, to improve breast tumor classification accuracy in ultrasound images, demonstrating the effectiveness of transfer learning in addressing data limitations in the medical domain. However, the evaluation was still limited to a single architecture. Jabeen et al. [15] integrated EfficientNet and ResNet, supplemented with explainable AI (XAI), for breast lesion classification on the BUSI dataset and demonstrated improved accuracy through the combination of these two architectures. However, the ensemble approach used was still limited to two models and did not explore combinations of three architectures simultaneously. In general, these studies show that probability-average-based ensemble approaches that combine the predictions of multiple CNN models can reduce prediction variance and improve robustness compared to single models across various medical image classification domains [16].

The three architectures used in this study, EfficientNet-B0, ResNet-50, and VGG-16, have distinct structural characteristics. VGG16 employs a sequential architecture with a simple and deep arrangement of convolutional layers. ResNet50 addresses the vanishing gradient issue in deep networks by relying on skip connections, whereas a compound scaling strategy is applied by EfficientNet-B0 to achieve proportional balance across network depth, width, and resolution. These differences in structural characteristics cause the three models to tend to produce prediction error patterns that are not always the same for the same samples. The prediction errors of one model have the potential to be compensated for by the others through a probability-average-based ensemble mechanism. Simultaneously combining these three architectures, each possessing complementary structural properties, has not been thoroughly investigated in prior literature, representing a key research gap addressed in this study.

Several previous studies have focused on comparing individual transfer learning architectures or applying ensemble methods with a limited number of models to breast ultrasound image classification using the BUSI dataset [12]–[15]. However, evaluations that simultaneously compare multiple architectures with different characteristics and test the model's generalization capabilities using real clinical data have not been extensively conducted. This is a critical aspect because ultrasound images obtained from clinical settings exhibit variations in quality, characteristics, and lesion complexity that may differ from those in public datasets. Based on these challenges, this study proposes a comparison of three transfer learning architectures: EfficientNet-B0, ResNet-50, and VGG-16, along with a soft-voting-based ensemble method using probability averaging. Unlike previous studies, this research utilizes a combination of the public BUSI dataset and clinical data from Bhakti Dharma Husada General Hospital in Surabaya to enrich the variation in data characteristics, making it more representative of real-world conditions. Therefore, this study aims to analyze individual transfer learning performances, determine the optimal classification architecture, and examine whether combining models via an ensemble approach yields superior accuracy over a standalone classifier when differentiating benign from malignant breast ultrasound images.

METHODS

The methodology implemented to construct a breast cancer classification model that is ultrasound based via transfer learning is detailed in this section. Performance across three baseline CNN architectures: EfficientNet-B0, ResNet-50, and VGG-16 is evaluated, with a probability averaging ensemble technique incorporated to enhance overall predictive outcome. Additionally, dataset acquisition and curation

procedures involving the public BUSI dataset and clinical data from Bhakti Dharma Husada General Hospital in Surabaya are outlined, alongside the image preprocessing stages. Subsequent research stages include dataset splitting, model training, fine-tuning, and performance testing to compare the classification results of each transfer learning architecture and the probability-average-based ensemble method. To evaluate model capabilities, accuracy, precision, recall, and F1-Score metrics are computed. All research stages are illustrated in Figure 1.

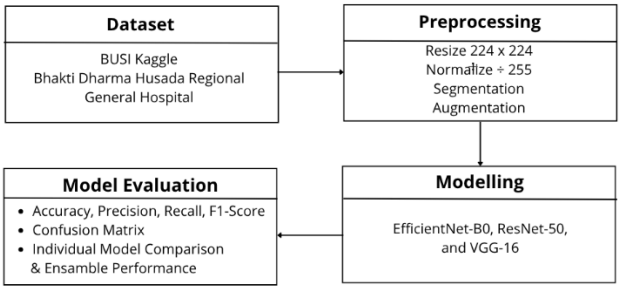


Figure 1. Research flow

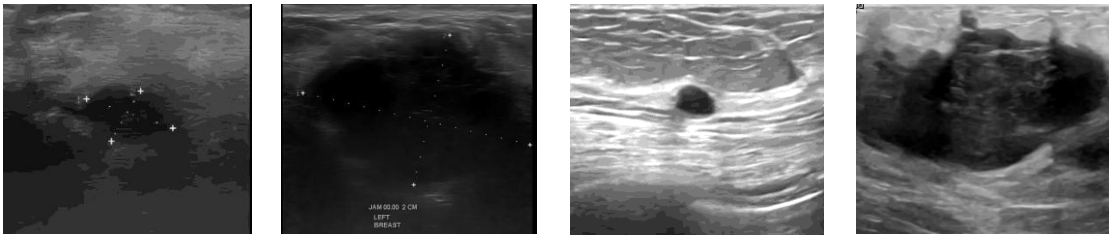
Dataset

The dataset utilized in this study is a combination of two data sources. The first source is primary breast ultrasound image data obtained from Bhakti Dharma Husada General Hospital in Surabaya. The second source is the public BUSI (Breast Ultrasound Images) dataset available on Kaggle, with images collected in 2018 [17]. For this study, we selected the benign and malignant classes, while the normal class was excluded from the dataset because the model would be less sensitive to breast lesions. The ground-truth annotation process for the primary image data was validated by physicians to ensure the accuracy of lesion boundaries classified as benign or malignant. All primary medical data used in this study have been anonymized to protect patient confidentiality and comply with research ethics codes.

Overall, the compiled dataset encompasses 1.632 breast ultrasound images, distributed into 1.048 benign and 584 malignant cases. The dataset was sourced from two sources: the public Breast Ultrasound Images Dataset (BUSI) and a clinical dataset obtained from Bhakti Dharma Husada General Hospital in Surabaya. In the benign class, 600 images came from the BUSI dataset and 448 images came from the clinical data at Bhakti Dharma Husada General Hospital in Surabaya. Meanwhile, in the malignant class, the BUSI dataset contained 336 images and the hospital's clinical data contained 248 images. To address the limited size of the raw dataset and improve model generalization, image augmentation, specifically rotation and flipping, was applied to both classes via Roboflow prior to the train, validation, and test split. The reported sample counts above (1.048 benign and 584 malignant) reflect the dataset composition after this augmentation step. The entire dataset was then divided into three parts: training data (70%), validation data (15%), and test data (15%), with class distribution adjusted to ensure proportional allocation. For clinical images from Bhakti Dharma Husada General Hospital, data selection was performed such that only one representative image was retained per patient, eliminating the possibility of images from the same patient appearing across multiple subsets (train/validation/test). For images from the public BUSI dataset, patient-identifying information is not provided by the original dataset publishers [17]. Therefore, explicit patient-level deduplication could not be performed for this portion, and we acknowledge this as a limitation of the study. The distribution for the training, validation, and test data is shown in Table 1.

Table 1. Dataset split

Class	Total Samples	Training (70%)	Validation (15%)	Testing (15%)
Benign	1048	734	156	158
Malignant	584	408	88	88
Total	1.632	1.142	244	246



1 (a) 2 (a) 1 (b) 2 (b)

Figure 2 Examples of breast ultrasound: Image (1) benign, image (2) malignant, image (a) Bhakti dharma husada general hospital dataset, image (b) Kaggle dataset

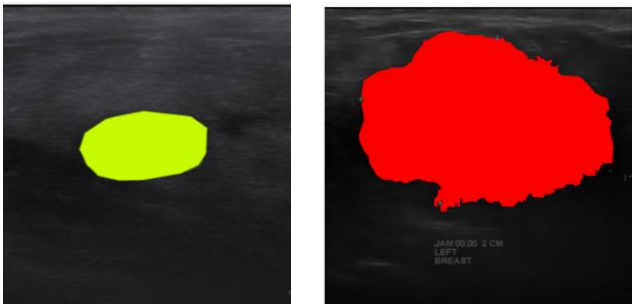
Preprocessing

Ultrasound images from Bhakti Dharma Husada General Hospital have non-uniform aspect ratios and contain additional information such as arrows, examination area labels, and medical text overlaid on the images, as shown in Figure 2 (1a and 2a). To minimize the influence of these non-lesion elements, an automatic cropping process was performed based on the DeepLabV3+ segmentation results. Specifically, the Region of Interest (ROI) of the lesion was first extracted from the original image so that most non-lesion elements outside the ROI, such as examination information text, were not carried over to the classification stage. However, the caliper markers used by sonographers to mark the boundaries of lesion measurements are generally placed right at the edge or within the lesion itself, so they remain within the cropped ROI. Bunnell et al. [18] explain that this condition is common in clinical breast ultrasound images because lesions are generally the only tissue structures that sonographers focus on and tend to be positioned in the center of the image, leaving very limited clean tissue area for caliper cropping. Furthermore, in the BUSI dataset used in this study, the automatic removal of calipers presents a higher level of difficulty compared to the internal clinical dataset used in that study. Based on these considerations, the caliper markers in this study were retained within the ROI because their removal risks deleting some relevant lesion image information, whereas the application of advanced cleaning techniques, such as inpainting, could be considered in future research.

All images resulting from ROI cropping, whether from Bhakti Dharma Husada General Hospital or the Kaggle dataset, were subsequently resized to 224×224 pixels in accordance with the standard input dimensions used by the pretrained ImageNet model [13]. The resizing process was performed using bilinear interpolation to preserve the visual quality of the images [19]. Pixel normalization was performed by dividing the pixel values by 255 so that they fell within the range (0,1) [20], [21]. They are then standardized using the mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] corresponding to the ImageNet dataset distribution [13]. This normalization is important to ensure that the pixel value scale aligns with the data distribution used during pretraining, so that the fine-tuning process can run more stably and converge faster [13], [22].

Image Segmentation

Image segmentation aims to isolate the Region of Interest (ROI), that is, the area of a lesion or breast mass, from the surrounding normal tissue [23]. This process is important because the normal tissue surrounding the lesion contains noise that can interfere with model training [24]. By isolating the ROI, the model can focus on relevant clinical characteristics, thereby improving classification accuracy [25]. To establish the ground-truth ROI labels required for training the segmentation model, annotation was performed using the Roboflow platform, a web-based tool for semi-automatic annotation. Roboflow provides a Smart Polygon feature powered by the Segment Anything Model (SAM) [26]. This feature assists the annotator to ensure boundary accuracy. This semi-automated annotation workflow is more efficient than a fully manual process [27]. An example of ROI annotation results on an ultrasound image is shown in Figure 3.



(a) (b)

Figure 3. Examples of ROI annotation results in breast ultrasound images.
(a) Benign lesion, (b) Malignant lesion.

The annotation results are saved in COCO JSON format to establish target ground truth labels for segmentation training. A DeepLabV3+ model utilizing a ResNet50 backbone initialized with pretrained ImageNet weights is deployed, which is trained to automatically delineate lesion boundaries on novel ultrasound images. After the predicted masks are obtained, the ROI are extracted via tight cropping within the lesion's bounding box, that is, the image is cropped precisely to the area bounded by the mask, rather than by blacking out the area outside the mask (masking). This cropping approach ensures that the pixels fed into the classification model consist entirely of the lesion area and its surrounding tissue, minimizing noise from irrelevant background. The cropped result is then resized to 224×224 pixels as input for the classification stage.

Assessment of the generated segmentation mask relies on two standard evaluation metrics, namely the Dice Score and Intersection over Union (IoU). The Dice Score measures the overlap between the predicted mask and the ground truth mask by computing twice the intersection divided by the total number of pixels in both masks, with values ranging from 0 to 1, where higher values indicate better overlap [24]. Formally, the Dice Score is defined as shown in Equation (1):

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \tag{1}$$

Where P denotes the predicted segmentation mask and Gdenotes the ground truth mask. IoU, commonly termed the Jaccard Index, evaluates the ratio between the intersecting area and the combined union of both regions, as detailed in Equation (2):

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{2}$$

In addition to these overlap-based metrics, Pixel Accuracy is also computed to measure the proportion of correctly classified pixels over the total number of pixels in the image, as defined in Equation (3):

$$Pixel\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In this context TP, TN, FP, and FN denote the counts of true positive, true negative, false positive, and false negative pixels, respectively.

Model Architecture and Fine-Tuning

This study uses three different CNN architectures to compare transfer learning approaches. All three models are initialized with pretrained weights from ImageNet, a large pre-trained dataset. The selection of ImageNet weights is based on the applicability of low-level features, such as edges and textures, to medical ultrasound images, without the need to train from scratch.

EfficientNet-B0

EfficientNet-B0 is an architecture designed for parameter efficiency by proportionally balancing network depth, channel width, and input resolution. This approach yields a model that is more parameter-efficient than architectures scaled manually or based on simple heuristics [28]. As a result, EfficientNet-B0 has only 5.3 million parameters, far fewer than other architectures [29]. This small size makes EfficientNet-B0 suitable for limited datasets, such as ultrasound images, by reducing the risk of overfitting [30]. This

parameter efficiency allows the model to learn more general feature representations without overfitting to specific patterns in the limited training data [31].

ResNet-50

ResNet-50 introduced the concept of skip connections, which involve connecting the input of a layer directly to its output via a shortcut [32]. This mechanism enables the architecture to optimize the residual deviation between inputs and outputs, avoiding the need to approximate the entire underlying mapping. This residual learning approach produces more stable gradients during the backpropagation process, thereby addressing the vanishing gradient problem that frequently occurs in very deep networks [33]. With skip connections, ResNet-50 can be trained with 50 weighted layers without experiencing gradient degradation in the early layers [34]. ResNet-50 consists of 4 stages of residual blocks that use a strategic combination of 1×1 convolutions (for channel dimension compression and expansion) and 3×3 convolutions (for spatial feature extraction), resulting in a model that is both efficient and deep in capturing hierarchical feature representations [31].

VGG-16

VGG16 is a classic architecture consisting of 13 convolutional layers with 3×3 filters stacked sequentially and 3 fully connected layers [35]. The consistent use of small filters allows VGG-16 to learn finer spatial features while maintaining an effective depth equivalent to that of larger filters. Although this architecture is conceptually simple, VGG-16 has been shown to produce rich convolutional feature representations and is widely used as a robust baseline across various medical transfer learning tasks [36]. Its main drawback is the extremely large number of parameters, 38 million, most of which come from the fully connected layers, potentially increasing the risk of overfitting on limited-size datasets [37].

Training Process

Each model is trained using a transfer learning approach with a pretrained backbone from ImageNet [38]. The backbone serves as a feature extractor that leverages visual representations learned from millions of images, while the original classification layer is replaced with a custom classification head tailored to the breast cancer detection task [39]. All network layers, including the backbone and the custom head, are trained simultaneously from the start by employing the Adam optimizer with an initial learning rate set to 0.0001. The use of a small learning rate acts as an implicit control mechanism that ensures the backbone weights are updated only gradually, thereby preserving the feature representations from ImageNet while adapting to the ultrasound image domain [40].

Custom Classification Head

A custom classification head is added on top of each backbone to replace the built-in classification layer of each architecture [41]. The head consists of a Dropout probability with a rate of 0.4 to reduce overfitting, followed by a linear dense layer for target class mapping. To accommodate the respective backbone feature dimensions, this linear layer accepts inputs of 1280 for EfficientNet-B0, 2048 for ResNet-50, and 4096 for VGG-16.

Loss Function & Optimizer

Binary Cross-Entropy (BCE) is used as the loss function because this study addresses a binary classification problem (benign vs. malignant) [42]. BCE measures the difference between the ground truth probability distribution (0 or 1) and the model's predicted probability distribution (a continuous value between 0 and 1) [43]. The loss will be at its minimum when the model's prediction is perfect and will be larger as the prediction deviates further from the ground truth [42]. The Adam (Adaptive Moment Estimation) optimizer was chosen because it automatically adjusts the learning rate for each parameter individually based on historical gradient statistics [43]. Adam combines the advantages of momentum-based methods and adaptive learning rate methods, making it more robust against various types of loss function landscapes and generally achieving convergence faster than simpler optimizers [44].

Early Stopping

To prevent overfitting, early stopping serves as a regularization strategy designed to monitor validation metrics during training. Specifically, when no improvement in validation loss is observed for 10 successive epochs, the optimization process terminates, and the best model achieved during training is used for inference. This technique prevents overfitting by avoiding excessively long training periods, during which the model begins to fit noise and specific characteristics of the training data [43]. A patience threshold of 10

epochs was chosen as a compromise between giving the model enough time to converge and avoiding overly long training.

Ensemble Learning

Ensemble learning is a machine learning paradigm that combines predictions from several base models to produce a final decision that is more robust than the decisions of each individual model [45], [46]. This approach does not rely on a single model but instead leverages the differing characteristics and capabilities of each model to improve accuracy, consistency, and generalization ability [47]. Within this framework, an ensembling approach is implemented through the aggregation of predictions from individual models, EfficientNet-B0, ResNet-50, and VGG-16, aiming to yield an ultimate decision with superior robustness and accuracy compared to a single model.

The method used is probability-averaging-based soft voting, as applied in the CNN ensemble framework for breast ultrasound image classification [48]. Each model computes a set of class probabilities via an activation as follows:

$$P_i(c \mid x) = \frac{e^{z_{i,c}}}{\sum_{k=1}^{K} e^{z_{i,k}}} \tag{4}$$

Where z_i is the logit vector from the-i (EfficientNet-B0, ResNet-50, or VGG-16), $c \in [0,1]$ denotes class (benign and malignant), and $K = 2$ is the number of classes. Individual model prediction probabilities from these three models are then summed and averaged to obtain the final ensemble probability:

$$P_{\text{ens}}(c \mid x) = \frac{1}{M}\sum_{i=1}^{M} P_i(c \mid x) \tag{5}$$

Where $M = 3$ denotes the number of models in the ensemble. The final class decision is then determined by *thresholding* the probability of the malignant class:

$$\hat{y} = \begin{cases} 1, & P_{\text{ens}}(1 \mid x) \geq \tau \\ 0, & P_{\text{ens}}(1 \mid x) < \tau \end{cases} \tag{6}$$

Where $\tau = 0{,}5$. This approach was chosen for its simplicity of implementation and its ability to fully utilize the probabilistic information from each model, rather than relying solely on binary class decisions as in hard voting [49].

The main advantage of the ensemble approach lies in its ability to reduce prediction variance by aggregating multiple models with different bias characteristics and error patterns [45]. Theoretically, averaging the predictions of several models with high variance but low bias can reduce the overall variance without significantly increasing bias [47]. Since each model tends to produce errors on different subsets of the sample and is not fully correlated with the others, combining the probabilities of the three models via Equation (5) allows the errors of one model to be compensated for by the other two. This results in a final prediction that is more stable, consistent, and accurate overall compared to the predictions from each of its constituent models.

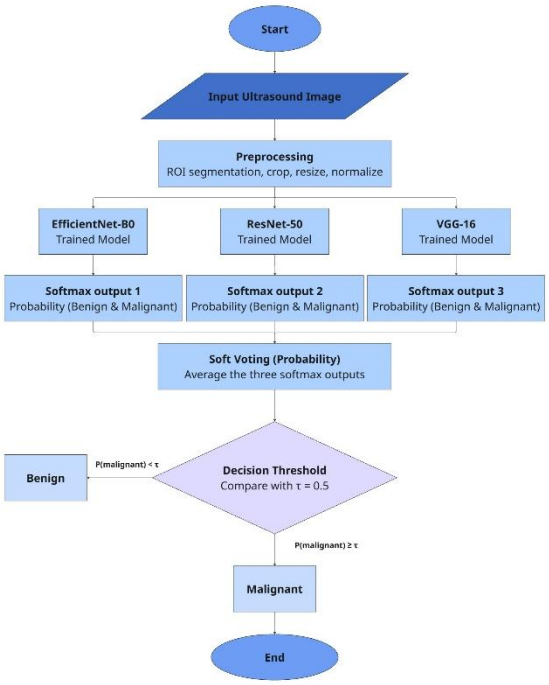


Figure 4. Proposed ensemble method flow

The specific flow of the proposed ensemble method is illustrated in Figure 4. The preprocessed ultrasound image is simultaneously fed into three independently trained transfer learning models: EfficientNet-B0, ResNet50, and VGG16. Each model produces its own softmax output consisting of class probabilities for benign and malignant. These three probability outputs are then aggregated through soft voting by computing their average. The resulting ensemble probability is subsequently compared against a decision threshold of $\tau = 0.5$, where a malignant probability equal to or exceeding the threshold yields a final prediction of malignant, and below the threshold yields benign.

Performance Evaluation Metrics

To assess the classification capability, we employ four primary metrics commonly used in medical classification: Accuracy, Precision, Recall, and F1-Score. These four metrics are derived from the confusion matrix, which records the four basic components of prediction results: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [50]. Specifically, Accuracy represents the percentage of true hits relative to the entire test dataset. Precision reflects how reliable the model is when predicting a sample as positive, that is, the ratio of positive predictions that are actually positive [51]. Recall measures the model's ability to detect all positive instances that actually exist in the dataset, which is critically important in the medical field because false negatives in the malignant class can have fatal consequences for patients [50]. As a composite metric, the F1-Score is an evaluation metric that combines Precision and Recall in a balanced manner into a single value, making it effective for assessing model performance, especially under conditions of class imbalance. [52].

Accuracy measures the proportion of correct predictions out of all test samples, as shown in Equation (7).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

Precision measures the proportion of positive predictions that are actually positive, as shown in Equation (8).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Recall measures the proportion of positive predictions successfully detected by the model, as shown in Equation (9).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

The F1-Score is the average of Precision and Recall, providing a balance between the two. It is particularly useful when there is a class imbalance (Equation 10).

$$\text{F1} - \text{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

RESULT AND DISCUSSION

Segmentation Performance

To analyze the DeepLabV3+ framework, its segmentation outputs were measured via three metrics: Dice Score, IoU, and Pixel Accuracy, computed separately for each lesion class. The results are presented in Table 2.

Table 2. DeepLabV3+ segmentation performance

Class	Dice Score	IoU (mIoU)	Pixel Accuracy
Benign	0.7895	0.6930	0.9569
Malignant	0.8484	0.7485	0.9075
Average	0.8190	0.7208	0.9322

The segmentation model achieved an average Dice Score of 0.8190 and an average IoU of 0.7208 across both classes, as shown in Table 2, indicating adequate lesion delineation performance. The malignant class yielded higher scores (Dice: 0.8484, IoU: 0.7485) compared to the benign class (Dice: 0.7895, IoU: 0.6930), which may be attributed to the generally more defined and hypoechoic boundaries of malignant lesions in ultrasound images, making them more distinguishable from surrounding tissue.

The relatively moderate performance on the benign class can be attributed to several factors. First, benign lesions tend to exhibit more irregular and heterogeneous boundaries in ultrasound images, which increases the difficulty of precise mask delineation. Second, the inclusion of clinical images from Bhakti Dharma Husada General Hospital introduces additional variability in image quality, scanner settings, and acquisition protocols that are not present in standardized public datasets, making segmentation more challenging compared to models trained exclusively on the BUSI benchmark dataset. Despite these challenges, the overall segmentation performance remains within an acceptable range for clinical ultrasound segmentation tasks, and the resulting ROI crops provide a reliable input for the subsequent classification stage.

Comparison Results Based on the Classification Report

The entire model training and evaluation process was implemented within a PyTorch environment utilizing the Python programming language. The transfer-learning-based CNN architectures used in this study included EfficientNet-B0, ResNet-50, and VGG-16. The training results show performance differences among the models, as shown in Table 3. ResNet50 achieved the best performance with an accuracy of 87.40%, while EfficientNet-B0 had an accuracy of 85.37%, and VGG-16 produced the weakest classification capability, yielding a baseline accuracy of 82.11%. The remaining evaluation criteria (including precision, recall, and F1-score) values were close to the accuracy values for each model, indicating that classification performance was generally stable overall, with no significant differences among the evaluation metrics.

Table 3. Comparison results: accuracy, precision, recall, and f1-score

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientNet-B0	84.15	82.62	83.83	82.96
ResNet-50	87.40	87.29	84.90	85.87
VGG-16	82.93	81.59	80.92	81.23
Ensemble	86.18	85.21	84.46	84.80

Comparison of Train Loss and Val Loss

To further assess the training behavior of each model, the loss curves for EfficientNet-B0, ResNet-50, and VGG-16 are presented in Figures 5, 6, and 7. These curves illustrate how training and validation loss evolved throughout the training process, offering insight into each model's convergence pattern.

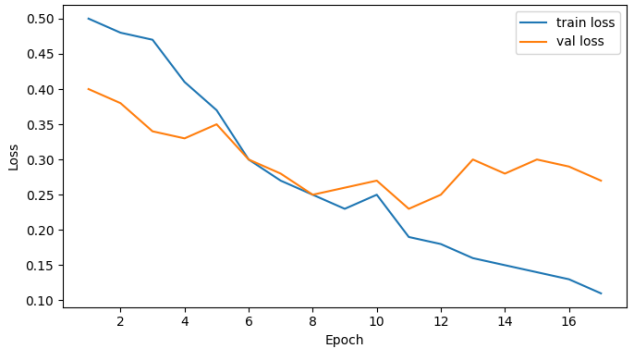


Figure 5. Train loss and val loss plots for EfficientNet-B0

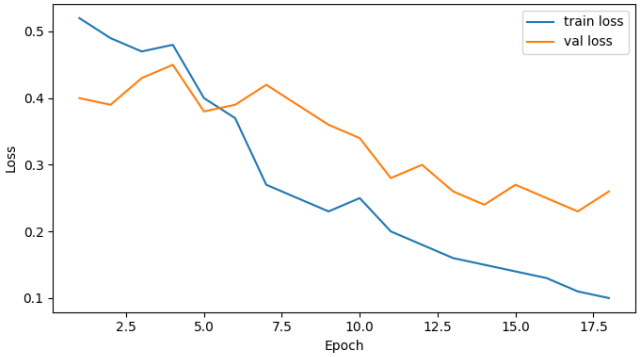


Figure 6. Train loss and val loss plots for ResNet-50

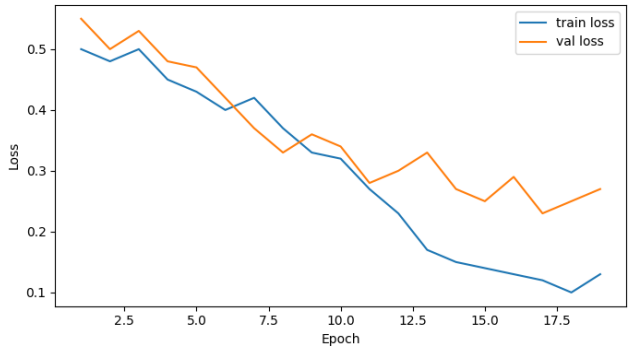


Figure 7. Train loss and val loss plots for VGG-16

An observable difference in the number of epochs across the three models (EfficientNetB0: 17 epochs, ResNet50: 18 epochs, VGG-16: 20 epochs) is due to the implementation of early stopping with a patience of 10 and a min delta value of 0.001. EfficientNetB0 reached a plateau the fastest, while VGG-16 exhibited sustained fluctuations in validation loss, causing training to continue longer before the patience limit was reached. This difference in performance can be explained by the training curve graphs for each model illustrated in Figures 5, 6, and 7.

Specifically, Figure 5 depicts the loss curves associated with EfficientNet-B0 17 training epochs. The training loss decreased steadily from an initially high value toward a markedly lower value by the final epoch, while the validation loss fluctuated within a moderate range throughout training and did not follow the same downward trajectory. This persistent gap between training and validation loss indicates a degree of overfitting, where the model becomes increasingly specialized to the training set while its performance on

unseen validation data plateaus early. This behavior triggered the early stopping mechanism to halt training before further degradation in generalization could occur.

Figure 6 shows the ResNet-50 curves over 18 epochs. The training loss decreased consistently across training, while the validation loss showed pronounced oscillations throughout, including a noticeable spike partway through training before declining again toward the later epochs. Despite this volatility, the overall trend of the validation loss was downward, and the model ultimately achieved the highest accuracy (87.40%) among the three base models. This suggests that ResNet-50 was still able to generalize reasonably well despite epoch-to-epoch instability, although the fluctuations indicate that its validation performance remained sensitive to the limited size of the validation set.

Figure 7 shows the behavior of VGG-16 curves 20 epochs. The training loss fluctuated considerably during the early epochs before eventually declining toward a low value by the final epoch, while the validation loss also decreased over the course of training but exhibited pronounced oscillations, including a sharp spike around the midpoint of training. Although the validation loss trended downward overall and the gap between the two curves narrowed by the later epochs, the prolonged instability, requiring the full 20 epochs before the patience threshold was met, indicates that VGG-16 struggled more than the other models to converge consistently, consistent with its lowest final accuracy among the three. This behavior can be attributed to the straightforward feed-forward configuration of the VGG-16 architecture, which lacks skip connections and is therefore more prone to unstable optimization and overfitting on limited-size datasets such as the one used in this study.

While the applied mitigation strategies, class weighting, data augmentation, dropout, and early stopping, reduced overfitting across all three models, a residual train-validation gap remains, particularly for EfficientNet-B0 and VGG-16. This is likely attributable to the limited dataset size (fewer than 2.000 images) relative to the capacity of the pretrained architectures used, which require substantially larger data volumes to fully generalize without memorizing training-specific patterns. Complete elimination of overfitting under this data constraint would require a considerably larger dataset, which represents a limitation of the present study and a direction for future work.

Confusion Matrix

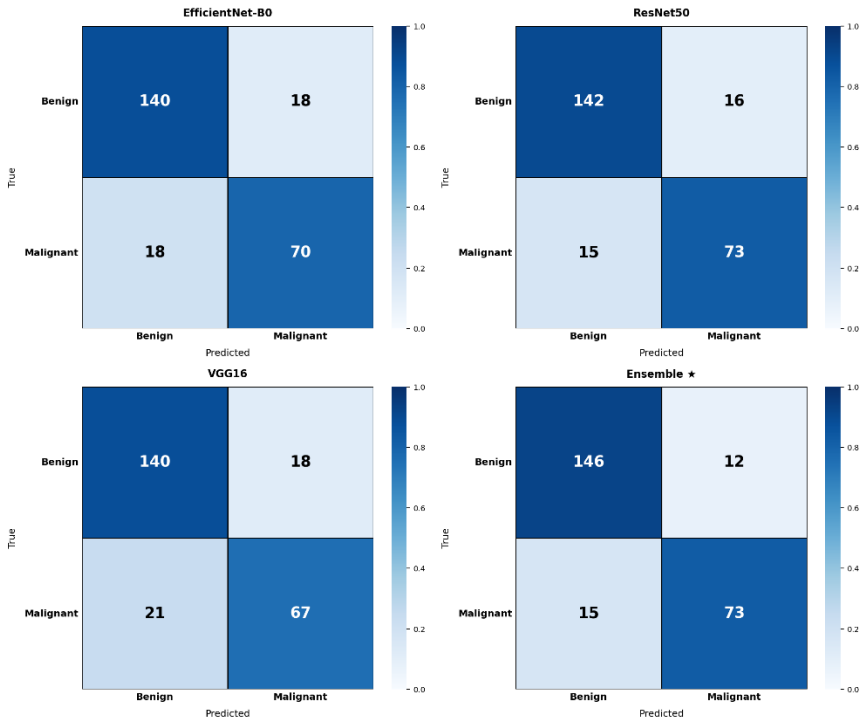


Figure 8. Confusion matrix EfficientNet-B0, ResNet-50, VGG-16, and ensemble

Evaluating the experimental outcomes detailed within the classification matrices from Figure 8, which illustrates how each configuration behaves on the EfficientNet-B0, ResNet-50, VGG-16, and Ensemble models using the BUSI dataset, the individual models exhibit varying performance. ResNet-50 successfully classified 142 benign images and 73 malignant images correctly, with classification errors consisting of 16 false positives (benign predicted as malignant) and 15 false negatives (malignant predicted as benign). EfficientNetB0 produced 140 true negatives for benign images and 70 true positives for malignant images, with a balanced error distribution (18 false positives and 18 false negatives). In contrast, VGG-16 showed the lowest performance with 140 true negatives for benign images but only 67 true positives for malignant images, with 18 false positives and 21 false negatives, indicating difficulty in detecting the malignant class. The ensemble model produced the optimal confusion matrix with 146 true negatives (benign) and 73 true positives (malignant), coupled with showcasing the minimum number of false positives (12), among all models.

Precision, Recall, and F1-Score in Table 4 and Figure 9 were calculated using macro-averaging—that is, an unweighted average between the benign and malignant classes—so that both classes contribute equally despite the class imbalance in the test data (158 benign images vs. 88 malignant images). We also report the recall for the malignant class separately, as this metric represents the model's ability to correctly detect malignant cases (true positive rate) and minimize clinically high-risk false negatives, making it the most relevant metric for the context of breast cancer screening. EfficientNetB0 achieved the highest malignant recall (80.68%), followed by the ensemble (78.41%), ResNet-50 (76.14%), and VGG16 (73.86%). This indicates that although ResNet-50 achieved the highest overall accuracy, the ensemble offers a more balanced trade-off between overall accuracy and sensitivity to malignant cases compared to any individual model.

To test whether the performance difference between the ensemble model and ResNet-50 is statistically significant, a McNemar exact test was performed on the predictions of both models on the same test dataset consisting of 246 images. The contingency table values are a = 206 (both correct), b = 6 (ensemble correct, ResNet-50 incorrect), c = 9 (ensemble incorrect, ResNet-50 correct), and d = 25 (both incorrect), resulting in $p = 0.6072$. Since this value is well above the significance threshold of $\alpha = 0.05$, the difference in performance between the ensemble model and ResNet-50 is not statistically significant at the 95% confidence level.

Comparison Bar Chart of Models and Ensemble

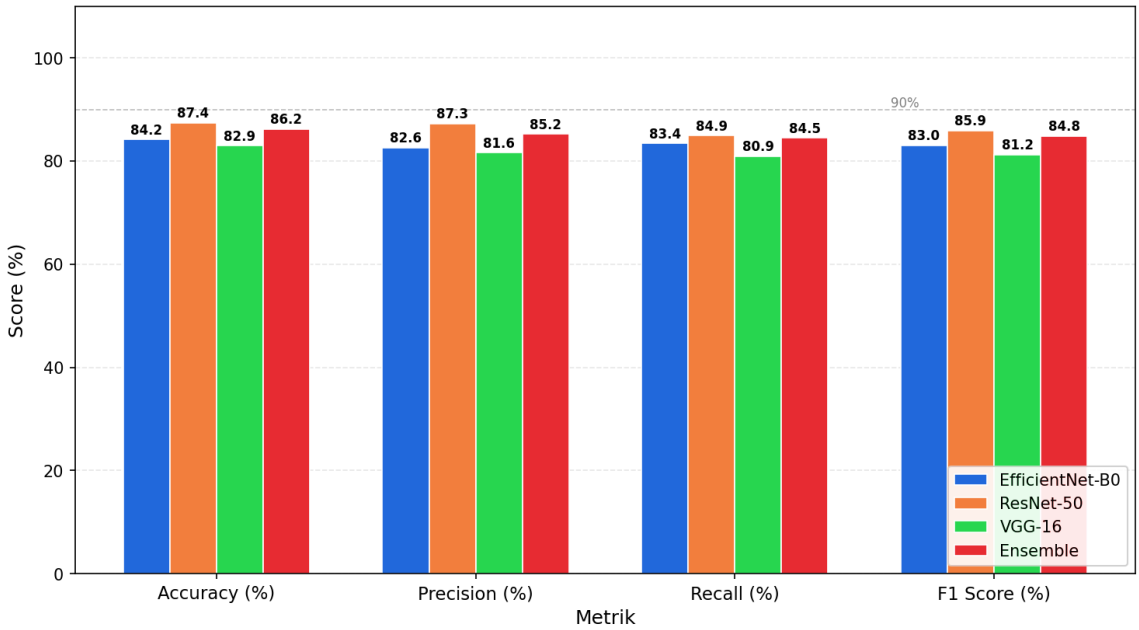


Figure 9. Comparative performance metrics of EfficientNet-B0, ResNet-50, VGG-16, and soft voting ensemble models

In Figure 9, the evaluation results show that ResNet-50 achieved the highest accuracy (87.40%) among the four models, followed by the ensemble model (86.18%). For the macro-average precision and recall metrics, ResNet-50 also achieved the highest macro-average precision (87.29%) and recall (84.90%),

while the ensemble model achieved a macro-average precision of 85.21% and a recall of 84.46%. However, the ensemble model achieved a higher recall for malignant cases (78.41%) compared to ResNet-50 (76.14%), indicating relatively better performance in accurately identifying malignant cases despite its slightly lower overall accuracy. To determine whether the difference between the ensemble model and ResNet-50 is statistically significant and not due to random variation, a McNemar test was conducted (reported below).

This difference in performance highlights the importance of architectural design in optimizing classification performance. ResNet-50 and EfficientNet-B0 demonstrate superior performance thanks to their skip connections and efficient scaling strategies, which enable better gradient flow and more effective feature extraction. In contrast, VGG-16 faces challenges in generalization due to its sequential design, which tends to get trapped in local optima of patterns documented in the deep learning literature. Although the ensemble model did not surpass ResNet-50 in overall accuracy, it achieved a macro average recall of 84.46% and a precision of 85.21%, as well as a higher recall for malignant cases than ResNet-50. This suggests that combining various architectural perspectives through soft voting can help compensate for the weaknesses of individual classifiers in detecting malignant cases, even without a statistically significant improvement in overall accuracy.

Comparison with Previous Studies

To further contextualize these findings, Table 4 compares the proposed ensemble model against several recent studies published within the last four years that applied transfer learning and ensemble-based approaches for breast ultrasound image classification, summarizing their methods, datasets, and reported accuracies.

Table 4. Comparison with previous studies on breast ultrasound image classification

Study (Year)	Method	Dataset	Accuracy
Twum et al. (2025)	ResNet-50 (transfer learning, fine-tuned)	BUSI	95.5%
Latha et al. (2024)	EfficientNet-B7 + augmentation + XAI	BUSI	99.14%
Jabeen et al. (2025)	EfficientNet +ResNet +XAI + feature fusion	BUSI	98.4%
This Study (2026)	Ensemble (EfficientNet-B0 +ResNet-50 + VGG-16, soft voting)	BUSI + clinical data from Bhakti Dharma Husada General Hospital in Surabaya	86.18%

As shown in Table 4, several prior studies report higher accuracy than the proposed model. However, this should not be interpreted as a direct measure of model superiority. Studies [12], [13], [15] were evaluated solely on the public BUSI dataset, which is relatively homogeneous in acquisition conditions and lesion presentation. In contrast, this study incorporates clinical data from Bhakti Dharma Husada General Hospital in Surabaya, introducing greater variability in imaging equipment and patient demographics, a factor that likely increased classification difficulty compared to studies using a single curated dataset. Moreover, none of the compared studies employed a three-model soft-voting ensemble, distinguishing the architectural contribution of this work from prior single or two model approaches.

CONCLUSION

This study shows that a transfer-learning-based soft-voting ensemble approach achieves competitive performance for breast cancer classification in ultrasound images compared to individual models, with ResNet-50 emerging as the best model among the three architectures compared in terms of overall accuracy (87.40%). Although EfficientNet-B0 and VGG-16 showed signs of overfitting that affected their respective generalization capabilities, combining the three through a soft-voting mechanism proved effective in compensating for these weaknesses and producing more stable predictions, with the ensemble achieving a higher recall for the malignant class (78.41%) compared to ResNet-50 (76.14%), a clinically relevant advantage for reducing cases of undiagnosed cancer. The McNemar test confirmed that the difference in overall performance between the ensemble and ResNet-50 was not statistically significant ($p = 0.6072$). These results indicate that ensemble learning can be a reliable strategy for improving sensitivity to malignant cases in ultrasound image-based computer-aided diagnosis systems, although it does not guarantee superior overall accuracy compared to single models that already perform well. Nevertheless,

limitations related to the dataset, reliance on manually annotated ground-truth masks for training the segmentation model, observed overfitting in EfficientNet-B0 and VGG-16, and the use of a single training run need to be addressed in future research, particularly by expanding the size and diversity of the training dataset to better align with the capacity of the pretrained architectures used, exploring more varied data augmentation strategies such as intensity variations, further refining the DeepLabV3+ segmentation model to reduce reliance on manually labeled data, repeating training runs to account for performance variability, and validating the model on more heterogeneous external datasets to ensure its generalization to real-world clinical conditions.

REFERENCES

[1] B. F. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, "Global Cancer Observatory: Cancer Today - Breast Cancer Factsheet (GLOBOCAN 2022)," *International Agency for Research on Cancer*, 2024. .

[2] J. Bray, M. E. Rebecca, and L. S. Mph, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin.*, no. February, pp. 229–263, 2024.

[3] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.

[4] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "SPECIAL SECTION ON DEEP LEARNING ALGORITHMS FOR INTERNET OF MEDICAL THINGS Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis," *IEEE*, vol. 8, 2020.

[5] N. Aprilia and R. Rumini, "Breast Cancer Classification based on Ultrasound Images using the Support Vector Machine (SVM) Algorithm," *Sistemasi*, vol. 13, no. 4, p. 1438, 2024.

[6] M. Abbadi, Y. Himeur, S. Atalla, and W. Mansoor, "Interpretable Deep Transfer Learning for Breast Ultrasound Cancer Detection: A Multi-Dataset Study," 2025.

[7] A. Raza, N. Ullah, J. A. Khan, M. Assam, A. Guzzo, and H. Aljuaid, "DeepBreastCancerNet: A Novel Deep Learning Model for Breast Cancer Detection Using Ultrasound Images," *Appl. Sci.*, vol. 13, no. 4, 2023.

[8] A. Carriero, L. Groenhoff, E. Vologina, P. Basile, and M. Albera, "Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024," *Diagnostics*, vol. 14, no. 8, 2024.

[9] S. Balasubramaniam, Y. Velmurugan, D. Jaganathan, and S. Dhanasekaran, "A Modified LeNet CNN for Breast Cancer Diagnosis in Ultrasound Images," *Diagnostics*, vol. 13, no. 17, pp. 1–28, 2023.

[10] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," *IEEE Access*, vol. 9, pp. 71194–71209, 2021.

[11] S. A. Chelloug, A. S. Ba Mahel, R. Alnashwan, A. Rafiq, M. S. Ali Muthanna, and A. Aziz, "Enhanced breast cancer diagnosis using modified InceptionNet-V3: a deep learning approach for ultrasound image classification," *Front. Physiol.*, vol. 16, no. April, pp. 1–15, 2025.

[12] F. Twum, C. C. E. Ahiable, S. O. Oppong, L. Banning, and K. Owusu-Agyemang, "Employing transfer learning for breast cancer detection using deep learning models," *PLOS Digit. Heal.*, vol. 4, no. 6, pp. 1–22, 2025.

[13] M. Latha, P. S. Kumar, R. R. Chandrika, T. R. Mahesh, V. V. Kumar, and S. Guluwadi, "Revolutionizing breast ultrasound diagnostics with EfficientNet - B7 and Explainable," *BMC Med. Imaging*, 2024.

[14] I. Ben, "Enhanced Computer-Aided Diagnosis Model on Ultrasound Images through Transfer Learning and Data Augmentation Techniques for an Accurate Breast Tumors Classification," *Procedia Comput. Sci.*, vol. 225, pp. 3938–3947, 2023.

[15] K. Jabeen *et al.*, "AI for breast lesion classification from ultrasound images," no. January 2024, pp. 842–857, 2025.

[16] S. Rajaraman, G. Zamzmi, and S. K. Antani, "Novel loss functions for ensemble-based medical image classification," vol. 16, no. 12 December, pp. 1–18, 2021.

[17] A. Pawłowska, P. Karwat, and N. Żołek, "Letter to the Editor. Re: '[Dataset of breast ultrasound images by W. Al-Dhabyani, M. Gomaa, H. Khaled & A. Fahmy, Data in Brief, 2020, 28, 104863],'"

vol. 48, 2023.
[18] A. B. Id, K. Hung, J. A. Shepherd, and P. Sadowski, "BUSClean : Open-source software for breast ultrasound image pre-processing and knowledge extraction for medical AI," pp. 1–15, 2024.
[19] J. Wei, H. Zhang, and J. Xie, "A Novel Deep Learning Model for Breast Tumor Ultrasound Image Classification with Lesion Region Perception," *MDPI*, vol. 31, no. 9, pp. 5057–5079, 2024.
[20] X. Wen, H. Tu, B. Zhao, W. Zhou, Z. Yang, and L. Li, "Identification of benign and malignant breast nodules on ultrasound: comparison of multiple deep learning models and model interpretation," *Front. Oncol.*, vol. 15, no. February, pp. 1–9, 2025.
[21] R. M. Al-Tam, A. M. Al-Hejri, F. A. Hashim, S. M. Narangale, M. A. Al-Antari, and S. A. Alzakari, "An Interpretable Ensemble Transformer Framework for Breast Cancer Detection in Ultrasound Images," *Diagnostics*, vol. 16, no. 4, pp. 1–33, 2026.
[22] H. O. A. Ahmed and A. K. Nandi, "ASG-MammoNet: an attention-guided framework for streamlined and interpretable breast cancer classification from mammograms," *Front. Signal Process.*, vol. 5, no. November, pp. 1–22, 2025.
[23] P. Bruno, M. Macrì, and C. Dodaro, "A Dual-stage Deep Learning Framework for Breast Ultrasound Image Segmentation and Classification," *J. Med. Syst.*, vol. 49, no. 1, pp. 1–11, 2025.
[24] A. Rahimnezhad, S. M. Rezaeijo, M. P. Bayat, and S. Heydarheydari, "Hybrid deep learning models for automatic segmentation and classification of breast lesions in ultrasound images," *Egypt. J. Radiol. Nucl. Med.*, 2025.
[25] L. Fu, N. Li, Z. Liao, Y. Lin, Z. Li, and F. Li, "Advances in Breast Ultrasound Segmentation," no. F L, 2026.
[26] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment Anything Model for medical image analysis: an experimental study," pp. 1–26, 2023.
[27] K. Powers *et al.*, "Development of a semi - automated segmentation tool for high frequency ultrasound image analysis of mouse echocardiograms," *Sci. Rep.*, no. 0123456789, pp. 1–15, 2021.
[28] Alyssa April Dellow and Saiful Izzuan Hussain, "Leveraging EfficientNet-CNN for Accurate Diagnosis of Breast Cancer from Ultrasound Images," *Elektr. J. Electr. Eng.*, vol. 24, no. 1, pp. 57–60, 2025.
[29] Archana Singh, "VISNET: An Efficient Light Weighted Hybrid Model for Early Detection of Breast Tumour in Ultrasound Images using Vision Transformer and Convolutional Neural Networks," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 40s, pp. 1318–1336, 2025.
[30] H. Wu *et al.*, "A Comparative Study of Multiple Deep Learning Models Based on Multi-Input Resolution for Breast Ultrasound Images," *Front. Oncol.*, vol. 12, no. July, pp. 1–10, 2022.
[31] T. Shahzad, T. Mazhar, S. M. Saqib, and K. Ouahada, "Transformer-inspired training principles based breast cancer prediction: combining EfficientNetB0 and ResNet50," *Sci. Rep.*, vol. 15, no. 1, pp. 1–18, 2025.
[32] H. Mewada, I. M. Pires, H. K. Thakkar, and A. V Patel, "A lightweight ALO optimized and learnable skip-connection integrated ResNet architecture for breast cancer diagnosis," *Biomed. Signal Process. Control*, vol. 120, no. PB, p. 110104, 2026.
[33] Sindhu and Muthumani, "Comparative Evaluation of Vgg19 and Resnet50 Architecture for Lung Nodule Detection Using Ct Scan Images," *Int. Res. J. Mod. Eng. Technol. Sci.*, no. 07, pp. 3698–3704, 2025.
[34] N. Behar and M. Shrivastava, "ResNet50-Based Effective Model for Breast Cancer Classification Using Histopathology Images," 2022.
[35] H. D. Alrubaie, H. K. Aljobouri, and Z. J. Aljobawi, "Efficient Feature Selection Using CNN, VGG16 and PCA for Breast Cancer Ultrasound Detection," *Int. Inf. Eng. Technol. Assoc.*, vol. 37, no. 5, pp. 1255–1261, 2023.
[36] T. Fatima and H. Soliman, "Application of VGG16 Transfer Learning for Breast Cancer Detection," *MDPI*, vol. 16, no. 3, 2025.
[37] A. Khan, A. Khan, and M. Ullah, "A computational classification method of breast cancer images using the VGGNet model."
[38] G. Ayana, J. Park, J. W. Jeong, and S. W. Choe, "A Novel Multistage Transfer Learning for Ultrasound Breast Cancer Image Classification," *Diagnostics*, vol. 12, no. 1, pp. 1–14, 2022.
[39] H. E. Kim, A. C. Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification : a literature review," *BMC Med. Imaging*, pp. 1–13, 2022.
[40] I. N. A. Nastase, S. Moldovanu, K. C. Biswas, and L. Moraru, "Role of inter- and extra-lesion tissue, transfer learning, and fine-tuning in the robust classification of breast lesions," *Sci. Rep.*, vol. 14, no. 1, pp. 1–12, 2024.

[41] X. ZHANG, W. SHAO, M. QIU, C. XIAO, and L. MA, "Advanced deep learning and transfer learning approaches for breast cancer classi fi cation using advanced multi-line classi fi ers and datasets with model optimization and interpretability," *PeerJ Comput. Sci.*, vol. 11, pp. 1–35, 2025.
[42] M. D. Ali *et al.*, "Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks," *Diagnostics*, vol. 13, no. 13, 2023.
[43] M. T. R. Arastu, T. Muskan, G. Deepak, K. Sinha, and K. Kumari, "Transformative Breast Cancer Diagnosis using CNNs with Optimized ReduceLROnPlateau and Early Stopping Enhancements," *Int. J. Comput. Intell. Syst.*, vol. 1, 2024.
[44] U. K. Lilhore *et al.*, "Hybrid convolutional neural network and bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification," *Sci. Rep.*, vol. 15, no. 1, pp. 1–27, 2025.
[45] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning : Opportunities and challenges," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, 2023.
[46] T. T. Yıldırım, O. Yaman, İ. Kılıç, B. Taşar, E. S. Timurkaan, and N. Aydoğdu, "Multi-Class Classification of Breast Ultrasound Images Using Vision Transformer-Based Ensemble Learning," *Diagnostics*, vol. 15, no. 17, pp. 1–23, 2025.
[47] K. L. Du, R. Zhang, B. Jiang, J. Zeng, and J. Lu, "Foundations and Innovations in Data Fusion and Ensemble Learning for Effective Consensus," *MDPI*, vol. 13, no. 4, pp. 1–49, 2025.
[48] S. S. Koshy, L. J. Anbarasi, and M. Narendra, "HED-Net : a hybrid ensemble deep learning framework for breast ultrasound image classification," *Front. Artif. Intell.*, no. January, pp. 1–27, 2026.
[49] I. Chhillar and A. Singh, "An improved soft voting - based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE - ENN class balancing," *Discov. Artif. Intell.*, 2025.
[50] Jude Chukwura Obi, "A comparative study of several classification metrics and their performances on data," *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 308–314, 2023.
[51] H. Dong, R. Liu, and A. W. Tham, "Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation," *J. Risk Financ. Manag.*, vol. 17, no. 2, 2024.
[52] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Appl. Sci.*, vol. 14, no. 21, 2024.