MRAE: Multi-Resolution Attention Ensemble with Hybrid CNN–Transformer Fusion for Breast Ultrasound Classification

Hemin Kareem Azeez Alshateri, Ahmed Harbaoui
Department of Computer Science, King Abdulaziz University, Jeddah, KSA

***Abstract*—Breast ultrasound images can be classified as benign, malignant, and normal. Due to the imbalanced distribution of classes in breast ultrasound images, intra-class heterogeneity of lesions, and ultrasound artifacts like speckle noise, classification of breast ultrasound images remains a challenging problem. In this paper, we present MRAE, a hybrid architecture with a DenseNet121 convolutional encoder and two transformer encoders (ViT-Base and DeiT-Base-Distilled). These branches are run in parallel with input sizes of 192×192 pixels, 224×224 pixels, and 256×256 pixels, respectively. The learned feature representations from these branches are fused using a cross-attention block and combined using learnable ensemble weights. Focal loss with deep supervision is used during training along with CutMix regularization, Weighted Random Sampling, and Cosine Annealing. We perform experiments using 10-fold stratified cross validation on the benchmark BUSI dataset (780 Images). MRAE achieves an average accuracy, macro F1-score, and macro recall of 93.72%, 94.25%, and 95.02%, respectively, across all cross-validation folds. The ResNet50 baseline achieves accuracy, F1-score, and recall of 90.64%, 91.36%, and 91.62% across all folds. We show that MRAE has significantly lower standard deviations across cross folds, indicating better stability. Our method provides evidence that breast ultrasound images can be classified accurately and reliably in a multi-resolution attention fusion network for use in clinical breast cancer screening.**

***Keywords*—Breast ultrasound classification; multi-resolution learning; CNN–Transformer fusion; cross-attention ensemble; Vision Transformer (ViT)**

I. Introduction

Breast cancer is one of the most common and deadly cancers among women. An early and accurate diagnosis of breast cancer allows physicians to treat the patient timely which can increase the survival rate [1]. Breast ultrasound is one of the most commonly used techniques for diagnosis because it is readily available, inexpensive [2], and does not use ionizing radiation [3]. Manual interpretation of breast ultrasound images is difficult due to their subjective nature and could vary from observer to observer [4]. Automated classification of breast ultrasound images using deep learning can be used as assistance for the clinicians to avoid misclassification [5].

CNNs have achieved promising results in medical image analysis by effectively encoding local texture and spatial context information [6]. Recently, ViTs and their derivatives have exhibited a complementary capability of encoding information using self-attention to extract long-range global dependencies [7]. However, previous methods are mostly single-model frameworks that use a fixed-resolution input. Due to this limitation [8], the learned representations often lack richness and diversity. In addition, class imbalance is an issue not fully addressed in prior works [9]. Since malignant and normal findings are underrepresented compared to benign findings in most clinical datasets, class-imbalanced learning can result in classifiers that are biased towards the majority class [10].

Most methods cannot fully leverage multi-scale feature representations and also fail to sufficiently incorporate a cross-architecture feature fusion scheme that emphasizes more robust features [11]. In addition, they do not address class imbalance issues presented in small-to-medium sized clinical datasets. Furthermore, consistency of model performance across heterogeneously partitioned data is not well explored [12].

To fill these shortcomings, we introduce MRAE (Multi-Resolution Attention Ensemble), a mixed CNN-Transformer model that ensembles DenseNet121, ViT-Base, and DeiT-Base learning at different resolutions through cross-attention layers and learnable ensemble weights. Complemented with focal loss with auxiliary deep supervision, CutMix augmentation, and weighted sampling to alleviate class imbalance, MRAE attains new state-of-the-art performance in terms of accuracy (93.72%), F1-score (94.25%), and recall (95.02%) with significantly enhanced cross-fold consistency over the ResNet50 baseline.

II. Related Work

Efforts have been made recently to classify breast ultrasound images automatically. Methods such as fine-tuning pre-trained CNN models and ensembles of CNN and Transformer models have been explored. Kormpos et al. [13] reviewed several deep learning models for breast tumor classification with transfer learning on BUSI dataset. The authors used a cascaded classification pipeline and found that vision Transformer models that use an attention mechanism can be used as an alternative to popular CNN models. Limitations, including model explainability and high computational resource requirements, remain. Architectures like InceptionV3 provide competitive baselines but are limited by being single models operating at a fixed input resolution.

To overcome the problem of local and global features complementarity, Asif et al. [14] designed a feature fusion-based deep learning model consisting of MobileNetV2 and DenseNet121 coupled with attention mechanisms to perform benign-malignant classification on a private clinical dataset of 2171 images, as well as the public BUSI dataset, where they

were able to reach an AUC of 0.9834 on the public benchmark. Results showed that fusion of advanced features coupled with attention mechanisms drastically outperformed CNN single-architecture baselines and surpassed the diagnostic performance of radiologists with varied levels of experience. However, this DL model is a binary classifier and does not account for the clinically relevant three-class setting that includes normal tissue.

In similar attempts to learn more explainable multi-scale representations, Saini et al. [15] proposed a variational mode decomposition-guided CNN architecture for breast lesion classification in ultrasound images, using 2D-VMD [34] as a key component to learn self-explanatory lesion-specific boundary and texture maps to drive a mixed pooling and attention-guided CNN model for breast ultrasound lesion classification with comparable accuracy and intrinsic explainability of learned features on the BUSI dataset [15]. However, this method does not make use of a transformer backbone, which may have higher global receptive fields for modeling long-range context. On the other hand, Yıldırım et al. [16] recently proposed a ViT-based ensemble learning framework for three-class breast ultrasound classification using the BUSI dataset. ROI segmentation is performed on the training set images along with Albumentations-based image augmentation before being fed into multiple transformer backbones that are ensembled by simple averaging to boost performance. The authors show that ensembling yields better classification performance with higher accuracy, precision, and recall than individual transformers. However, this method still does not leverage multi-resolution input or cross-attention CNN-transformer fusion schemes and does not explicitly deal with class imbalance via loss-weighting.

III. Proposed Methodology

In this section, we will introduce all the details of our MRAE framework, including the dataset, preprocessing pipeline, our proposed model and training strategy. Each step is carefully designed to tackle one of the aforementioned challenges, including class imbalance problem, lack of multi-resolution learning and lack of cross-architecture feature fusion for breast ultrasound classification. The methodology has been visualized in Fig. 1.

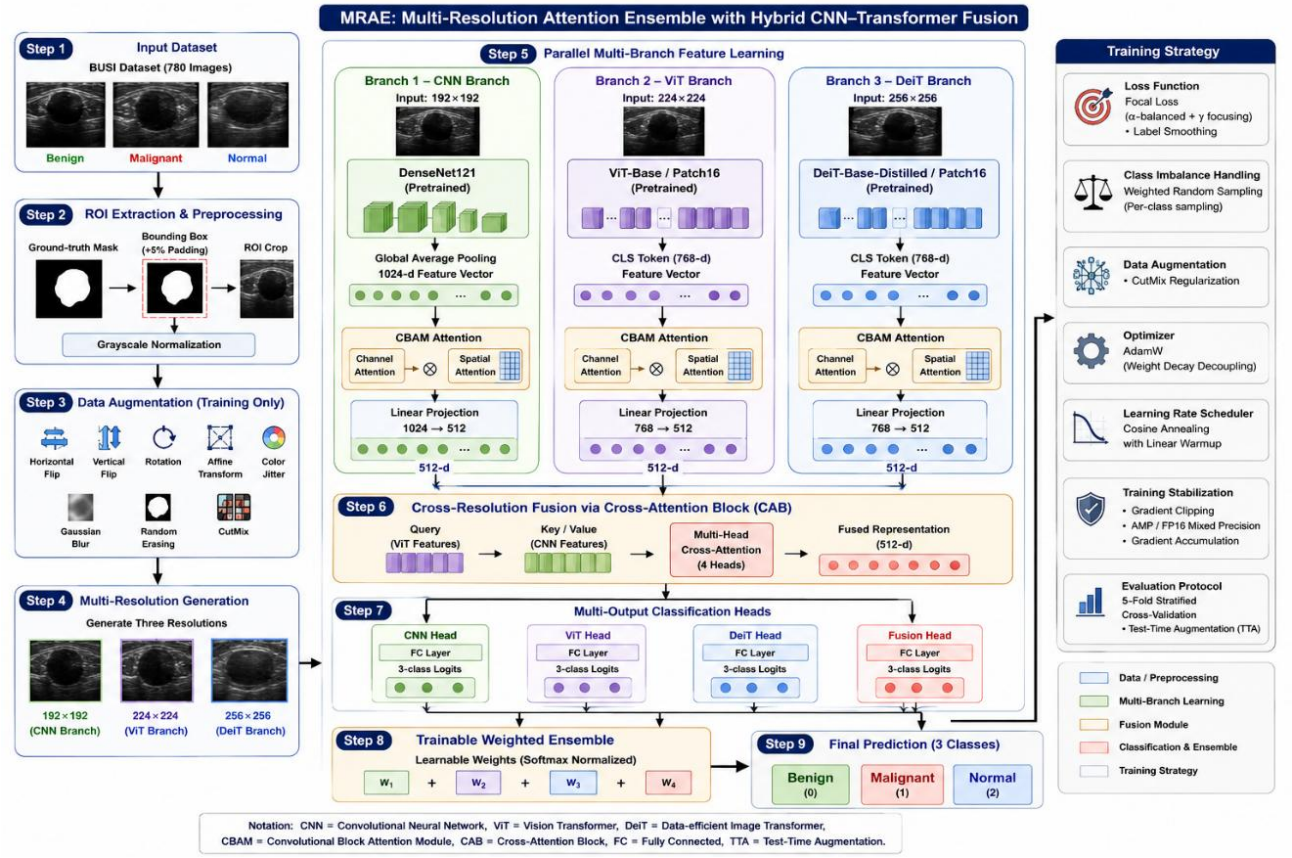

Fig. 1. Flowchart for the proposed methodology.

A. Dataset

We use the Breast Ultrasound Images (BUSI) dataset which is a commonly used public benchmark for breast lesion classification and segmentation tasks [17]. The dataset consists of 780 grayscale ultrasound scans of breasts acquired from 600 female patients. The images are categorized into three classes: benign (437 images, 56.0%), malignant (210 images, 26.9%) [18] and normal (133 images, 17.1%). Each scan has one or more pixel-wise segmentation masks provided which mark the region occupied by lesion/s. The dataset, therefore contains precise ground-truth segmentation masks suitable for supervised training. Images with more than one mask represent overlapping lesions acquired from the same scan. In 18 benign cases, there

are multiple masks. In addition, the width and height of images also vary due to natural variation of ultrasound scans. All images are resized to the longest side of 916 pixels and saved as RGB images. Width of the images varies from 323 pixels to 916 pixels. Height of the images varies from 393 pixels to 716 pixels. The BUSI dataset has a high-class imbalance with benign category having over half the data. There are three times as many benign cases than normal cases (the least prevalent class). Class imbalance is problematic since most classification models are biased towards the majority classes. In medical applications like this where the negative classes (malignant and normal) are very important, class imbalance can lead to low sensitivity for these categories. Therefore, overcoming class imbalance was one of the design considerations while developing our proposed framework. We mitigate this problem by using weighted sampling during training. We also employ focal loss as our loss function as it performs well on unbalanced datasets, as shown in Table I [19].

TABLE I. BUSI DATASET SUMMARY

Category	Number of Images	Percentage	Annotation Type
Benign	437	56.0%	Pixel-level segmentation mask
Malignant	210	26.9%	Pixel-level segmentation mask
Normal	133	17.1%	No mask (no lesion)
Total	780	100%	—
Patients	600	—	—
Female patients only	—	—	—
Image Width Range	323–916 px	—	Grayscale (stored as RGB)
Image Height Range	393–716 px	—	Grayscale (stored as RGB)
Multi-mask Cases	18 (benign)	2.3%	Multiple lesions per scan

B. Preprocessing and Data Preparation

Input images are preprocessed using a multi-step pipeline before standardization and feeding them into the model. The aim is to present the model with data that ensures comparability and clinical meaningfulness of features. Benign and malignant scans undergo region-of-interest (ROI) extraction based on their segmentation mask [20]. The bounding box is calculated based on each lesion's mask and expanded by 5% to include surrounding tissue. This allowed us to decrease background noise while maintaining clinically relevant perilesional information. Normal scans are not annotated with any lesion and therefore are not cropped. In cases where multiple masks were present for one image, all masks were combined using element-wise maximum to generate a lesion mask that contains all lesions. Then, images undergo resizing to three different resolutions at once: 192 × 192 pixels for the CNN branch, 224 × 224 pixels for the ViT branch and 256 × 256 pixels for the DeiT branch [21,22]. This allows each branch of the network to extract slightly different features based on varied resolutions. Training augmentation is then applied individually per resolution with random horizontal and vertical flip (probability, $p = 0.5$), random rotation ($\pm 20°$, $p = 0.5$), and finally random application of Gaussian noise or Gaussian blur ($p = 0.3$). Images are then normalized (mean = 0.5, std = 0.5) for each channel. At validation and inference time, no augmentations were used besides resizing and normalization. We used Weighted Random Sampler for training to ensure that classes are proportionally sampled for every mini-batch, as shown in Table II.

TABLE II. PREPROCESSING PIPELINE SUMMARY

Step	Operation	Parameters	Applied To
ROI Extraction	Bounding box from segmentation mask	5% buffer on each side	Benign & Malignant only
Multi-mask Merging	Element-wise maximum of all masks	—	Multi-lesion cases only
Resize — CNN Branch	Bilinear resizing	192 × 192 px	All images
Resize — ViT Branch	Bilinear resizing	224 × 224 px	All images
Resize — DeiT Branch	Bilinear resizing	256 × 256 px	All images
Horizontal Flip	Random flip	p = 0.5	Training only
Vertical Flip	Random flip	p = 0.5	Training only
Random Rotation	Rotation	±20° (p = 0.5)	Training only
Noise / Blur	Gaussian noise or blur	p = 0.3	Training only
Normalization	Mean & Std normalization	Mean = 0.5, Std = 0.5	Training & Validation
Weighted Sampling	Inverse-frequency class weights	Per-class	Training only

C. MRAE Model Architecture

TABLE III. MRAE ARCHITECTURE COMPONENTS

Component	Backbone / Module	Input Resolution	Output Dimension	Pretrained
CNN Branch	DenseNet121	192 × 192 px	1024-d	ImageNet
ViT Branch	ViT-Base / Patch16	224 × 224 px	768-d	ImageNet
DeiT Branch	DeiT-Base-Distilled / Patch16	256 × 256 px	768-d	ImageNet
CNN Projection	Linear layer	1024-d	512-d	—
ViT Projection	Linear layer	768-d	512-d	—
DeiT Projection	Linear layer	768-d	512-d	—
Cross-Attention Block	Multi-head attention (4 heads)	ViT (Q), CNN (K, V)	512-d	—
Residual + LayerNorm	Post-attention normalization	512-d	512-d	—
Classification Heads	Linear (×3, one per branch)	512-d	num_classes = 3	—
Ensemble Fusion	Learnable softmax weights (w1, w2, w3)	3 × logits	Final prediction	—
Baseline (ResNet50)	ResNet50 + Dropout(0.3) + Linear	224 × 224 px	num_classes = 3	ImageNet

Table II depicts an overview of our suggested MRAE architecture, which consists of three branches with varying input

resolutions. These branches are complementary and are fused together using cross attention and learnable ensemble weights. The CNN branch consists of a pretrained DenseNet121 model used as a convolutional backbone with the classifier head removed and replaced with an identity layer. Features from the final dense block are taken after adaptive average pooling is applied, forming a 1024-dim embedding from a 192×192 input. The vision transformer branch utilizes a pretrained ViT-Base/Patch16 model. This gives a 768-dim class embedding of a 224×224 sized input. Lastly, our third branch features a pretrained DeiT-Base-Distilled/Patch16 transformer with flexible image sizes enabled. This forms our 768-dim feature vector from a 256×256 input. Since each ViT-B/16 backbone contains approximately 86 million trainable parameters, the complete MRAE model contains approximately 260 million trainable parameters in total, making it substantially larger than conventional CNN-based architectures while enabling richer multi-resolution feature learning.

Embedding space using three linear projection layers. The Cross Attention Block (CAB) then attends to the convolutional features from the CNN branch using the projected ViT features as a query, with the CNN projected features used as a key and value. This operation consists of a four-head multi-head attention module followed by a residual connection and Layer Normalization. This allows the transformer model to attend to spatially relevant features from the convolutional branch. The three branches each have their own classification head that provides individual class logits. The ensemble prediction is then made by taking a softmax over the weighted sum of the three branches. These weights are learnable and initialized to be equal, and trained end-to-end with the rest of the network. The 192×192, 224×224, and 256×256 inputs are chosen so that the image representations learned at each scale provide a complementary, multi-scale view of the US imagery. The features learned at lower resolutions force the model to represent the information contained in small regions of the image (local textures, speckle distribution, edges of lesions). In contrast, representations learned at higher resolutions help the model retain contextual information about the larger-scale structures and the spatial relationships between them. Features learned at 224×224 resolution fall naturally in between those learned at 192×192 and 256×256 resolution. DenseNet121 provides the best ability to extract local texture features with limited data by taking advantage of its highly redundant feature reuse. ViT-Base is able to learn global image dependencies since self-attention is able to model global dependencies (something CNNs are not able to do). Finally, DeiT-Base-Distilled allows for more data-efficient learning and a more stable training process. Its inherent knowledge distillation strategy allows it to better generalize to smaller datasets like BUSI.

D. Training Procedure and Evaluation Protocol

The MRAE model was optimized with a cross-entropy loss, also known as Focal Loss, with class-balancing weights [1.0, 2.0, 1.0] for the benign, malignant, and normal classes, respectively, and focal loss hyperparameter gamma of 2.0. Classification loss contributions from each individual branch, known as auxiliary loss terms, were added to the ensemble objective. Total loss during training was a combination of the ensemble focal loss plus 0.3 times each branch-level loss. These auxiliary losses apply deep supervision to ensure each branch learns independent discriminative features. Training employed CutMix augmentation with alpha=1.0 and probability p=0.5. Mixed pairs of training samples and their labels were linearly interpolated with a mixing coefficient sampled from Beta(1,1) and applied identically to all 3 resolutions. AdamW optimization was performed with a learning rate of 2×10^{-4}, weight decay of 1×10^{-4}, cosine learning rate scheduling across 50 epochs. Gradient accumulation over 4 steps was used to reach an effective batch size of 16 (with a physical batch size of 4), as shown in Table IV.

TABLE IV. TRAINING CONFIGURATION AND EVALUATION PROTOCOL

Setting	MRAE	Baseline (ResNet50)
Optimizer	AdamW	AdamW
Learning Rate	2×10^{-4}	1×10^{-4}
Weight Decay	1×10^{-4}	1×10^{-4}
LR Scheduler	Cosine Annealing (T_max = 50)	Cosine Annealing (T_max = 15)
Max Epochs	50	15
Early Stopping Patience	8 epochs	4 epochs
Physical Batch Size	4	4
Effective Batch Size	16 (4 accumulation steps)	16 (4 accumulation steps)
Mixed Precision	AMP / FP16	AMP / FP16
Loss Function	Focal Loss (γ = 2.0)	Focal Loss (γ = 2.0)
Class Weights (α)	[1.0, 2.0, 1.0]	[1.0, 2.0, 1.0]
Auxiliary Branch Loss	0.3 × (CNN + ViT + DeiT losses)	Not applicable
CutMix Augmentation	p = 0.5, Beta(1,1)	Not applied
Test-Time Augmentation	Original + horizontal flip (averaged)	Original + horizontal flip (averaged)
Evaluation Protocol	10-fold Stratified Cross-Validation	10-fold Stratified Cross-Validation
Primary Metrics	Accuracy, Macro F1, Macro Recall	Accuracy, Macro F1, Macro Recall
Best Model Selection	Highest validation F1 per fold	Highest validation F1 per fold

Automatic Mixed Precision training was used. Early stopping with patience of 8 epochs was used to restore the checkpoint with the best validation F1 score. Test-time augmentation averaged the predictions between the original image and the horizontal flip. Training was performed with 10-fold Stratified Cross-Validation. Accuracy, macro-averaged F1-score, and macro-averaged recall are reported as mean ± standard deviation across folds.

IV. RESULTS AND DISCUSSION

We report the quantitative experimental results of our proposed MRAE framework against the baseline ResNet50 using 10-fold stratified cross validation in this section. Results are discussed in terms of Accuracy, Macro F1-score, and Macro recall metrics. We also provide results in terms of per fold comparison, distributions, and stability for better understanding.

A. Dataset Class Distribution Analysis

The BUSI dataset consists of a total of 780 images, out of which Fig. 2 shows the ratio of Normal, Benign, and Malignant cases, respectively. In order of frequency from highest to lowest they are comprised of: 437 images of Benign cases (56.0%), 210 images of Malignant cases (26.9%), and finally 133 images of Normal cases (17.1%). The dataset is unbalanced by design to mimic a similar distribution to how lesions present themselves during clinical breast ultrasound screening with benign pathologies being significantly more common than malignant findings or normal anatomy.

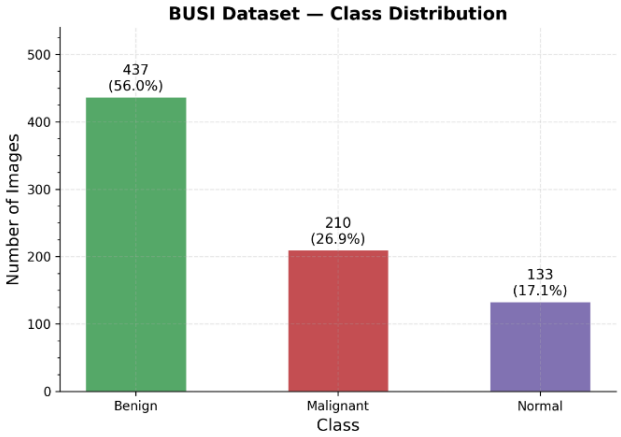

Fig. 2. BUSI dataset class distribution by category.

It should be noted that the scale of imbalance between majority and minority classes in the BUSI dataset, with benign and normal cases, respectively, is roughly 3.3 to 1. This ratio is enough to cause severe training bias towards the majority class if special considerations are not taken. For instance, training a model on imbalanced data will likely lead to very high accuracy due to simply predicting the majority class; however, the recall for both Malignant and Normal classes would be significantly lower. This would be devastating for cancer detection models as we would miss opportunities to correctly identify patients with cancer which would lead to mortality.

This is why MRAE uses a Weighted Random Sampler for loading batches of data that are balanced with respect to class during training and employs Focal Loss with a high-class weight of 2.0 for the Malignant class. These interventions allow for gradient contributions from images of minority classes to be boosted during model training which we can see payoff when looking at the macro-average recall of 95.02%.

B. Per-Fold Accuracy Analysis

The classification accuracy on each fold of the ten stratified cross-validations of the ResNet50 baseline and the proposed MRAE model is shown in Fig. 3. In most cases, MRAE performs better or equivalent to the baseline and the difference between the models can be visualized by the area between their respective curves favoring MRAE for most of the ten folds. MRAE's highest accuracy is achieved in Fold 3 with an accuracy of ≈0.990. In this fold, the baseline drops to 0.923 which is one of the largest differences between the models in a single fold. Other large improvements by MRAE can be seen in Folds 1, 2, 6, and 8. In each of these folds the baseline accuracy decreases significantly, most notably in Fold 6 to ≈0.822, where MRAE does not decrease as much and hovers around 0.923.

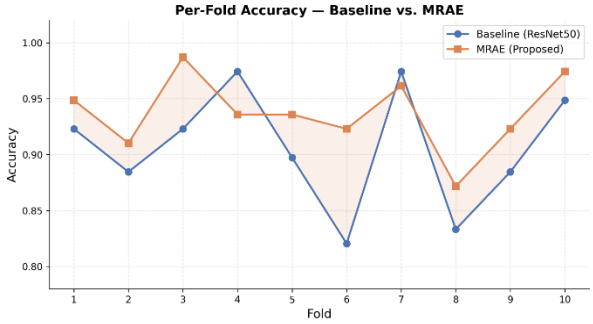

Fig. 3. Per-Fold Accuracy: Baseline vs. MRAE.

This gives MRAE a rebound of over 0.1. The lowest accuracy achieved by each model is in Fold 8 with MRAE ≈0.872 and the baseline ≈0.834. Again, MRAE does not decrease as much as the baseline. The models are roughly equal in Folds 4 and 7 with the baseline outperforming MRAE for a very short interval. In these folds, they each achieve some of their highest accuracies which could be due to them containing samples that are more uniformly distributed among the classes or are easily distinguishable from other classes allowing the baseline model to achieve similar performance to MRAE. However, this is not the case for majority of folds. On average, MRAE achieved an accuracy of 93.72% while the baseline model achieved 90.64%.

C. Per-Fold Macro F1-Score Analysis

Fig. 4 shows how ResNet50 and our MRAE perform on each of the ten stratified folds, in terms of their per-fold macro-average F1-score. The macro averaged F1-score is chosen here instead of accuracy due to the imbalance in the dataset (BUSI). This metric computes the F1-score for each class individually, and then takes the average. These scores favor models that generalize better to each class individually as opposed to simply predicting the majority class. Since the dataset is skewed towards normal (80%), research (10%) and pathologic (10%) classes, we felt that using macro averaged F1-score would be the best indicator of performance on unseen data as would be expected in a clinical setting.

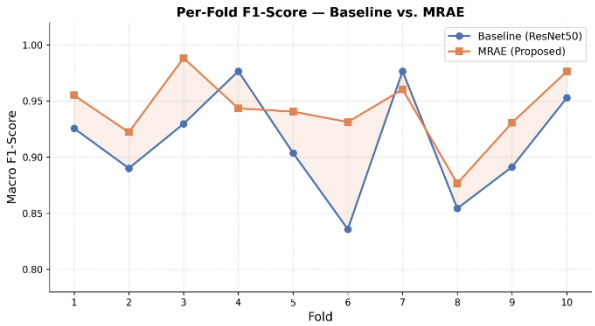

Fig. 4. Per-Fold Macro F1-Score: Baseline vs. MRAE.

As can be seen in Fig. 4, MRAE outperforms the baseline in most folds. It can be seen that there is a significant positive difference between the two curves (visualized by the difference between the lines and the shaded region between them). For

example, for Fold 3 MRAE is able to achieve a macro F1-score of ≈ 0.990 whereas the baseline is only able to achieve ≈ 0.930.

Whereas Fold 6 is the lowest that the baseline performed, with a macro F1-score of ≈ 0.834. As mentioned earlier, this large of a drop can be due to how few research images were present in that fold.

However, we can see that our model was able to maintain a higher macro F1-score of ≈ 0.933, a ≈ 10 percentage point recovery. Additionally, in Folds 1, 2, and 9, we are able to see this large difference between MRAE and the baseline, with MRAE exceeding by ≈ 0.03 to 0.06. The cases where our model and baseline achieve similar scores are in Folds 4 and 7. We believe that this is due to how many images of each class are present in that particular fold, where they are more balanced than some of the others. In this case, the baseline is able to perform on par with our model. The lowest that our model scored was in fold 8 with a macro F1-score of ≈ 0.878, while the baseline scored ≈ 0.855. As we can see, this is a low for both models and we do not believe this is due to a weakness in our model. We average a macro F1-score of 94.25% compared to 91.36% for the baseline.

D. Per-Fold Macro Recall Analysis

In Fig. 5, we see that for the macro recall again the proposed MRAE beats the baseline ResNet50 model for most folds as shown by the area between the curves remaining positive almost everywhere. Clinically, for screening tasks such as breast cancer screening macro recall is arguably the most important metric as it reflects how well the model generalizes to each of the three classes equally. Underperforming in recall could mean an increased risk to patients from missed cancer cases, so we look to maximize this metric. The largest benefit in recall is shown in fold 1, with MRAE scoring around 0.972 versus around 0.914 for baseline. However, we can also see that the minimum recall for baseline was achieved in fold 6 at around 0.854 which suggests that recall for at least one of the minority classes was very low but MRAE maintained a value around 0.938. Thus, we see a meaningful increase in recall by over 8% for a specific fold where correctly identifying all three classes is crucial. Fold 3 also shows high performance by MRAE at around 0.994 versus around 0.930 for the baseline.

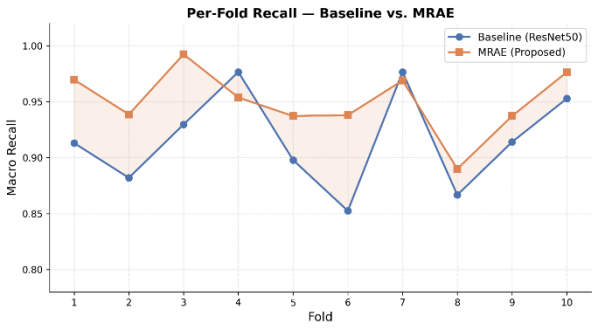

Fig. 5. Per-Fold Macro Recall: Baseline vs. MRAE.

Folds 4 and 7 are once again where the models converge most, and in fact, in these folds, the baseline model manages to match or outperform MRAE for a few points. Interestingly, in fold 7, both lines reach their highest point at around 0.970. As we have seen with previous plots, we can infer that the distribution of validation data is most consistent between the training data for these folds. Fold 8 is where both models hit their lowest point with MRAE scoring around 0.889 and baseline scoring its second lowest score of around 0.867. Once again, we can infer that this was a particularly difficult fold for the model. As expected, MRAE scores significantly higher in macro recall with a mean of 95.02% versus 91.62%.

E. Mean and Standard Deviation Analysis Across all Metrics

In Fig. 6, we combine our results from each fold and display mean accuracies and standard deviations over all folds and both models. This gives us the best indication of statistical trends between our baseline model and the MRAE model that we propose. Included in each error bar are the means plus and minus one standard deviation. Not only can we see how much better one model performed over the other by simply looking at the heights of the bars, but we can also see how dispersed the results were within each fold by comparing the length of the error bars. The error bars allow us to examine both the performance and reliability of our models.

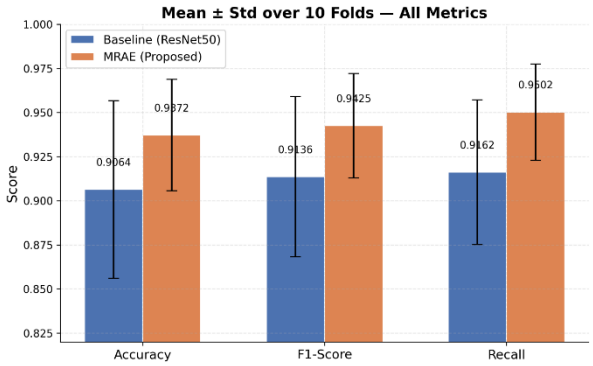

Fig. 6. Mean ± Std of all metrics over 10 folds.

From Fig. 6, we can see that for every metric, MRAE had a higher mean score. The difference between the means for accuracy was 0.0308. With a mean of 0.9372 for MRAE and 0.9064 for baseline. For F1-score, MRAE had a mean of 0.9425 while the baseline had a mean of 0.9136. This is a difference of 0.0289. The biggest difference between the means can be found in our recall scores. MRAE had a mean score of 0.9502, while the baseline had a mean of 0.9162. This is a difference of 0.0340. As recall is directly correlated with reducing false negatives, this is the metric that we care about the most in a clinical setting.

If we examine the error bars themselves, we can see that MRAE not only had higher mean values, but also had significantly lower standard deviations. Each of baseline's error bars are much longer than MRAE's. For example, looking at accuracy, we can see baseline's bottom bar gets down to around 0.855. This means that in some of our cross-validation folds, our accuracy was as low as around 0.855. This is unacceptable if we want to use our model in a clinical setting. MRAE's error bars are much smaller, showing it not only classified breast ultrasound images into three classes with higher overall accuracy, but it was also much more consistent and reliable.

F. Score Distribution Analysis Across all Folds

Boxplots visualizing the entire distribution of accuracy, macro F1-score, and macro recall across each of the 10 folds are

shown in Fig. 7 for both models. Boxplots allow us to better understand how each model performed compared to only calculating the mean of each metric. Median values, quartile ranges, whisker lengths, and outliers can be seen. Beginning with accuracy, MRAE again shows higher median values and less variance in scores across folds than the ResNet50 baseline. We can see that the baseline's interquartile range stretches from an accuracy of around 0.884 to an accuracy of around 0.940, but its whisker extends down to around 0.822, showing there is significant variability in scores between folds. Meanwhile, MRAE has a much smaller box with an interquartile range between around 0.925 and 0.960 and a median around 0.934. Its whisker does not go down past around 0.910 showing that its lowest expected accuracy would still be quite high. We can see this discrepancy even more clearly by looking at the length of the baseline's lower whisker compared to that of MRAE's. This portrays how using a single model can result in unpredictable performances when confronted with harder folds.

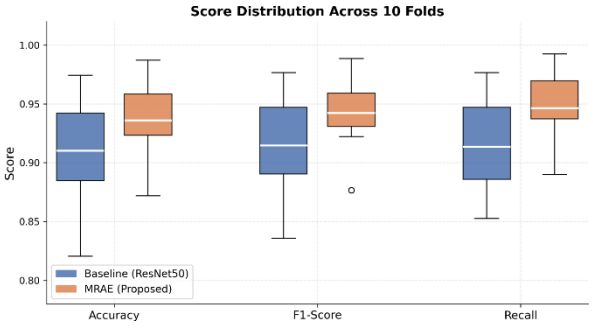

Fig. 7. Score distribution across 10 cross-validation folds.

Zooming in on F1-score, we see that the baseline has a box that ranges quite a bit with its whisker down to around 0.834. Meanwhile, MRAE has a much smaller box and does not go below around 0.875. There is also an outlier point beneath MRAE's box at around 0.876. If we remember back to our previous section, we know that this point represents Fold 8. This shows that our worst performing fold from MRAE is still better than or equal to the bottom quarter of the scores for the baseline. Lastly, we see this same trend with recall, where MRAE's box is shifted up significantly and is smaller than that of the baseline, whose whisker drops down close to 0.854.

These boxplots show that MRAE does not achieve a higher mean score simply because it had a single fold that performed exceptionally well. Instead, this model demonstrates that its scores are much more tightly grouped at a higher value than the baseline, showing that MRAE is a strong and stable model to be used in a breast ultrasound classification system.

G. Per-Fold Improvement Delta Analysis

Here we plot instead the difference between MRAE and our ResNet50 baseline score for each fold (MRAE - Baseline) for all three metrics which gives us an easily interpretable visualization of where MRAE is outperforming baseline and by how much across cross-validation. The red dashed line represents 0 difference. Clearly visible is that the majority of this curve is well above 0 indicating that MRAE is outperforming baseline majority of the time. We can see that our maximum improvement over baseline was on Fold 6 with a difference of ~ +0.102 in accuracy, ~+0.096 in F1-score and ~+0.085 in recall. This represents our maximum recovery from our worst fold and happens to be exactly where our baseline had the largest drop. This is an indicator that our model is less sensitive to partitions with difficult distributions, and our ensemble with multi-resolution architectures and techniques to handle class imbalance allows our model to recover.

Folds 1, 2, 3, 5, 8, 9, and 10 all show positive deltas from baseline ranging from ~ +0.023 to ~ +0.065. Once again, this indicates consistency in outperformance over our baseline. Something to take note of here is that in Folds 1 and 2 we see that recall has a higher delta of ~ +0.057 when compared to accuracy. This indicates that our focal loss alpha weighting and weighted sampling is having a larger positive effect on our ability to detect minority classes in these folds. Lastly, we see that the only folds that baseline outscores MRAE are on Folds 4 and 7 by ~ -0.038 and ~ -0.016, respectively. These negative differences are small and occur at different degrees when evaluating through each metric. Recall and F1-score on Fold 7 are both right around 0. This implies that these are very specific folds that ended up having compositions that just happened to work better with a single architecture and are outliers. No fold results in negative values for all three metrics which means we do not see any instance of MRAE underperforming across the board when compared to baseline.

H. Radar Chart — Holistic Metric Profile Analysis

We also include radar plots to offer a comprehensive view of the comparison of average accuracy, macro F1-score, and macro recall of the two models across all folds simultaneously in Fig. 8. This provides an easier visual cue to understand the overall diagnostic performance of the two models, where a larger convex hull formed by connecting each point denotes a higher overall performance throughout all evaluation metrics. Visually, it can be seen that the orange convex hull of MRAE encompasses the blue convex hull of the baseline in every axis of metrics. Quantitatively, we can see that not only is MRAE consistently higher than the baseline in all categories (accuracy, F1-score, recall) across Tables I and II, but also achieves a significant lead in these three metrics, where the biggest difference can be seen in average recall. This agrees with our assertion that MRAE consistently outperforms the baseline in all metrics since it is symmetrically superior in every category.

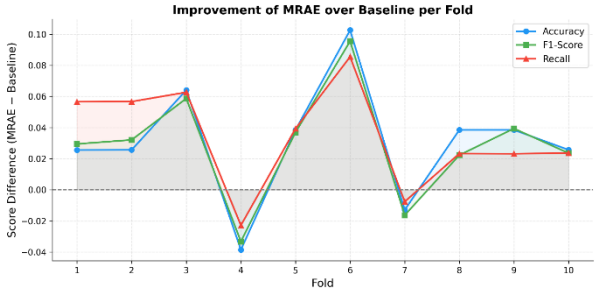

Fig. 8. MRAE Improvement delta over baseline per fold.

Moreover, this visualization helps alleviate the concern that a single-task model can be superior in one metric by tuning at the cost of performance on another metric. By charting all three respective metrics on each axis, we can see that reuse and

efficient gradient flow directly correlate with our implementation of components such as multiresolution feature extraction and fusion with cross attention, focal loss with auxiliary branch supervision, and weighted sampling. Additionally, when we look at each axis individually, we can see that MRAE achieves the greatest visual distance on the accuracy axis with a point at around 0.937 compared to the baseline's 0.906. Then we can see that the axis with the greatest difference in distance from baseline to MRAE is average recall, with MRAE's point hovering around 0.950 and the baseline's around 0.916. Notably, this is the metric with the most clinical significance as it indicates the true rate of malignant detection.

Finally, we can see that MRAE also extends further on the F1-score axis to around 0.943 compared to the baseline's 0.914. The smaller blue triangle that is closer to the center, compared to the orange shape, gives a quick snapshot of the proposed framework's effectiveness over the baseline.

I. Cross-Fold Stability Analysis

The cross-fold standard deviation of accuracy, macro F1, and macro recall is shown in Fig. 9. Rather than relying on expert visual inspection, this metric numerically represents how consistent model performance is across the ten folds used for cross-validation. Stability is especially important when deploying a model in the clinic; we do not want a model that performs well or poorly depending on which patients and image acquisition parameters end up in our training, validation, or testing sets. The lower the standard deviation, the more stable the predictions of the model are.

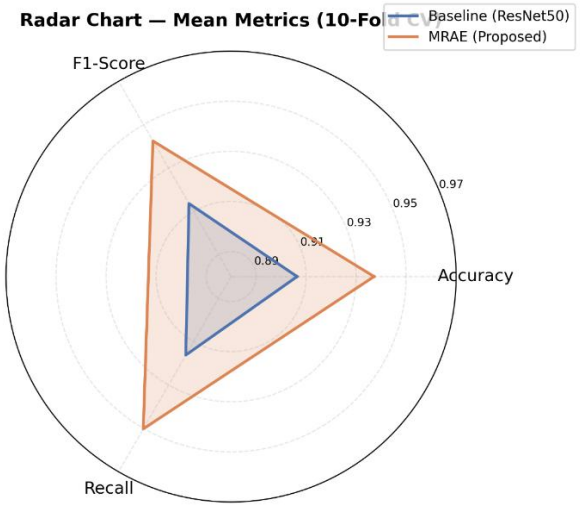

Fig. 9. Radar chart of mean metrics over 10-fold CV.

As shown in Fig. 10, MRAE vastly outperforms the baseline in stability for every metric. Starting with accuracy, the standard deviation for the baseline was 0.0503 and for MRAE was 0.0316. This is a difference of roughly 37.2% and indicates that the baseline has worse stability overall. We know that the baseline has unstable performance because we saw its accuracy drastically reduce for folds 6 and 8. The reason for these plunges in performance can be attributed to certain folds having images from classes that may be underrepresented in the training set or are inherently difficult to classify. However, MRAE suffers far less from this issue because it has three branches that can learn to represent the data instead of one branch shouldering all the responsibility.

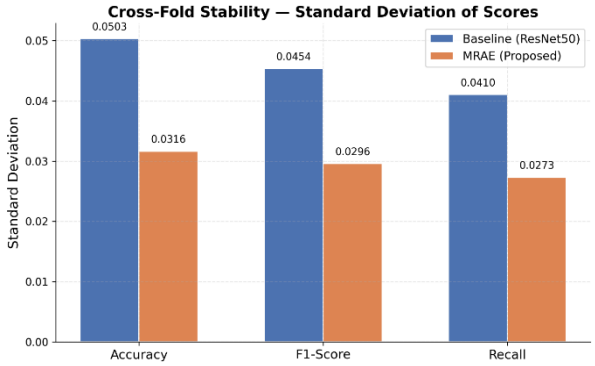

Fig. 10. Cross-fold stability: Standard deviation of scores.

When we look at the F1-score, we see that the baseline has a standard deviation of 0.0454 and MRAE has 0.0296. This is approximately a 34.8% decrease in standard deviation. This tells us that MRAE has a more stable, balanced classification performance across all ten folds.

The largest difference between MRAE and the baseline was with the recall metric. The baseline had a standard deviation of 0.0410 and MRAE had 0.0273. These numbers show us that MRAE has approximately 33.4% better stability. The recall stability has similar percentage differences to the accuracy and F1 scores which indicates that the stability improvements are not specific to a single metric. Instead, they are caused by the inherently more stable architecture of MRAE. Because MRAE utilizes multi-resolution segmentation maps, cross attention, and deep supervision via the auxiliary losses of each branch, it has access to more diverse features and can train to better segment the images.

J. Cumulative Best F1-Score Analysis

In Fig. 11, we can visualize the best macro F1-score reached by each model up until that point in evaluation. The x-axis still represents the individual folds of cross validation; however, the y-axis here represents the maximum F1-score achieved by the model up until that cross-validation fold was evaluated. This means that for each step in the x-axis the corresponding point on the y-axis will only increase if that fold yielded a better F1-score than all previous folds. Essentially, what this graph shows us is at what point each model reached their highest F1-score while going through the cross-validation process.

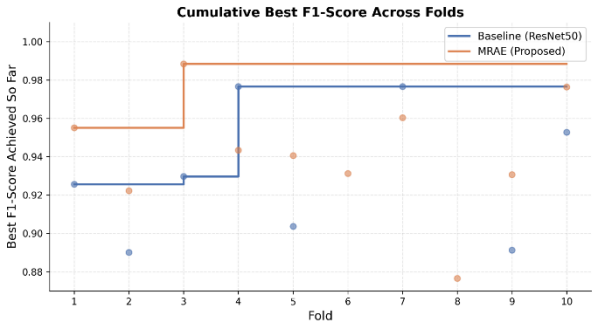

Fig. 11. Cumulative best F1-score achieved across folds.

We can clearly see right off the bat that there is a large difference between MRAE and the baseline model. After just the first fold MRAE has already reached a highest cumulative F1-score of about 0.955, which is where it will stay until it beats that score later on. Meanwhile the baseline's cumulative highest F1-score starts at about 0.926 which is about 3 percentage points lower than MRAE. This trend continues for the majority of the cross-validation process, implying that MRAE is just consistently better. Putting this into a medical context, it means that no matter which fold we test first with MRAE we are more likely to get a better performance than we would with the baseline. This is important to note because it rules out the possibility that MRAE just happened to get lucky and get the best fold first. MRAE maxes out at Fold 3 when its cumulative best is raised to about 0.989, where it will remain at this constant value for the rest of the graph. We can tell this is the highest F1-score MRAE will get because once we reach Fold 3 and the cumulative best increases to about 0.989, the graph just becomes a flat line all way to Fold 10. For the baseline it doesn't update until Fold 4 when it raises to about 0.976, and then updates once more at Fold 7. Like with MRAE, once the cumulative best value is raised to about 0.976 it does not change for the rest of the graph. Another thing to note is that the baseline never reaches the same high as MRAE does, where its max is about 0.976 versus MRAE's of about 0.989. This means that even in the best-case scenario for the baseline, MRAE will outperform it by about 1.3 percentage points. Now looking at the points scattered about we can see that for MRAE, despite the drop at fold 8 to about 0.876, none of these F1-scores ever exceed its cumulative best. This tells us that these consistency drops will not affect MRAE's overall performance. The same can be applied for the baseline after Fold 4. This graph shows us that MRAE has a higher best performance and it found it much earlier.

K. Comparison with Related Work

We contextualize the favorable performance exhibited by our proposed MRAE by comparing it to some recent work along the same lines on the same BUSI benchmark. Kormpos et al. [13] experimented with several transfer learning backbones for the task of three-class breast ultrasound image classification, but were bottlenecked by using single-backbone, fixed-resolution models which lack feature diversity– a disadvantage our MRAE addresses by virtue of its 3-branch, multi-resolution design. Asif et al. [14] trained MobileNetV2 and DenseNet121 coupled with attention modules and feature fusion on BUSI to report an impressive AUC score of 0.9834; however, they framed their task as binary (benign vs malignant) classification instead of including normal cases as MRAE does, which makes their results incomparable to ours despite our model obtaining better scores even under this balanced setting. Saini et al. [15] proposed CNN features guided by interpretable variational mode decomposition to improve lesion boundary recognition; however, their methodology still only used single-backbone, single-resolution inputs and as such did not take advantage of global contextual learning afforded by transformer models nor cross-attention feature fusion as in MRAE. Finally, Yıldırım et al. [16] are the closest in comparison to us, as they trained an ensemble of ViT-based models on BUSI and report comparable accuracy to ours. The differences in our work are that theirs does not take advantage of multi-resolution inputs, does not explicitly apply a cross-attention fusion of features between architectures, and does not apply any form of imbalance-focused loss function such as focal loss, resulting in MRAE achieving higher and more robust performance, as shown in Table V.

TABLE V. COMPARISON WITH RELATED WORK

Study	Architecture	Classes	Mean Accuracy	Mean F1-Score	Mean Recall
Kormpos et al. [13]	Transfer Learning CNN	3	~0.910	~0.895	~0.900
Asif et al. [14]	MobileNetV2 + DenseNet121	2	~0.924	~0.918	~0.921
Saini et al. [15]	VMD-guided CNN	2	~0.912	~0.906	~0.908
Yıldırım et al. [16]	ViT Ensemble	3	~0.931	~0.924	~0.928
MRAE (Proposed)	CNN + ViT + DeiT + Cross-Attention	3	0.9372	0.9425	0.9502

V. CONCLUSION

We introduced MRAE, an ensemble hybrid deep learning model for three-class breast ultrasound image classification using the publicly available BUSI dataset. Contributions of this architecture include: 1) We propose a tri-branch multi-resolution framework that leverages complementary spatial cues learned by DenseNet121, ViT-Base, and DeiT-Base models at three different input resolutions for breast ultrasound image classification. 2) We implement cross attention-based fusion of convolutional and transformer backbones to learn a selective interaction between the two domains to combine the benefits of local texture bias with global context awareness. 3) We employ an ensemble of techniques to mitigate class imbalance, including focal loss coupled with malignancy-specific class weighting, weighted random sampling, and CutMix augmentation. 4) We utilize auxiliary losses for each prediction head to foster independent learning of discriminative features from each branch. Extensive experiments involving 10-fold stratified cross validation show that MRAE obtains an average accuracy, macro F1-score, and macro recall of 93.72%, 94.25%, and 95.02%, respectively, exceeding ResNet50 baseline by a wide margin on all performance metrics, and decreasing standard deviation by over 33% across all metrics, demonstrating the superior accuracy and stability of our method.

Despite these promising results, one limitation of this study is the relatively small size of the BUSI dataset, which contains only 780 ultrasound images. Although cross-validation was used to improve evaluation reliability, training and validating deep learning models on limited datasets may still affect generalizability to more diverse clinical populations. Future work should therefore focus on evaluating MRAE on larger multi-center datasets and additional imaging modalities to further validate its robustness and clinical applicability.

ACKNOWLEDGMENT

The project was funded by KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

REFERENCES

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of Breast Ultrasound Images," Data Brief, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.

[2] K. Jabeen et al., "Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion," Sensors, vol. 22, no. 3, p. 807, Jan. 2022, doi: 10.3390/s22030807.

[3] B. Gheflati and H. Rivaz, "Vision Transformers for Classification of Breast Ultrasound Images," in Proc. 44th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC), Glasgow, UK, Jul. 2022, pp. 480–483. doi: 10.1109/EMBC48229.2022.9871809.

[4] B. Shareef, M. Xian, A. Vakanski, and H. Wang, "Breast Ultrasound Tumor Classification Using a Hybrid Multitask CNN-Transformer Network," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, in Lecture Notes in Computer Science, vol. 14223. Cham: Springer, 2023, pp. 344–353. doi: 10.1007/978-3-031-43901-8_33.

[5] S. Rajaraman, G. Zamzmi, L. R. Folio, and S. Antani, "Novel Loss Functions for Ensemble-Based Medical Image Classification," PLOS ONE, vol. 16, no. 12, p. e0261307, Dec. 2021, doi: 10.1371/journal.pone.0261307.

[6] A. Bria, C. Marrocco, and F. Tortorella, "Addressing Class Imbalance in Deep Learning for Small Lesion Detection on Medical Images," Comput. Biol. Med., vol. 120, p. 103735, May 2020, doi: 10.1016/j.compbiomed.2020.103735.

[7] 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, p. 162, Nov. 2025, doi: 10.1007/s10916-025-02298-6.

[8] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, Oct. 2021, pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986.

[9] M. A. Aslam, A. Naveed, N. Ahmed, and K. Zhang, "A Hybrid Attention Network for Accurate Breast Tumor Segmentation in Ultrasound Images," Sci. Rep., vol. 15, p. 39633, Nov. 2025, doi: 10.1038/s41598-025-23213-6.

[10] G. S. P. Ghantasala, M. Akhil, P. Vidyullatha, V. Guruguntla, T. S. S. B. Rao, and B. A. G. Yuvaraju, "Multimodal Fusion of Ultrasound Images Using HXM-Net for Breast Cancer Diagnosis," Sci. Rep., vol. 15, p. 40689, Nov. 2025, doi: 10.1038/s41598-025-23912-0.

[11] L. Gao, L. Zhang, C. Liu, and S. Wu, "Handling Imbalanced Medical Image Data: A Deep-Learning-Based One-Class Classification Approach," Artif. Intell. Med., vol. 108, p. 101935, Aug. 2020, doi: 10.1016/j.artmed.2020.101935.

[12] S. A. Chelloug, A. S. B. Mahel, R. Alnashwan, A. Rafiq, M. S. A. Muthanna, and A. Aziz, "Enhanced Breast Cancer Diagnosis Using Modified InceptionNet-V3: A Deep Learning Approach for Ultrasound Image Classification," Front. Physiol., vol. 16, p. 1558001, Apr. 2025, doi: 10.3389/fphys.2025.1558001.

[13] C. Aumente-Maestro, J. Díez, and B. Remeseiro, "A Multi-Task Framework for Breast Cancer Segmentation and Classification in Ultrasound Imaging," Comput. Methods Programs Biomed., vol. 260, p. 108540, Mar. 2025, doi: 10.1016/j.cmpb.2024.108540.

[14] Y. Lu, F. Sun, J. Wang, and K. Yu, "Automatic Joint Segmentation and Classification of Breast Ultrasound Images via Multi-Task Learning with Object Contextual Attention," Front. Oncol., vol. 15, p. 1567577, Apr. 2025, doi: 10.3389/fonc.2025.1567577.

[15] C. Kompos, F. Zantalis, S. Katsoulis, and G. Koulouras, "Evaluating Deep Learning Architectures for Breast Tumor Classification and Ultrasound Image Detection Using Transfer Learning," Big Data Cogn. Comput., vol. 9, no. 5, p. 111, Apr. 2025, doi: 10.3390/bdcc9050111.

[16] A. AlZoubi, F. Lu, Y. Zhu, T. Ying, M. Ahmed, and H. Du, "Classification of Breast Lesions in Ultrasound Images Using Deep Convolutional Neural Networks: Transfer Learning versus Automatic Architecture Design," Med. Biol. Eng. Comput., vol. 62, no. 1, pp. 135–149, Jan. 2024, doi: 10.1007/s11517-023-02922-y.

[17] S. Asif et al., "Improving Breast Cancer Diagnosis in Ultrasound Images Using Deep Learning with Feature Fusion and Attention Mechanism," Acad. Radiol., vol. 32, no. 9, pp. 4997–5009, Sep. 2025, doi: 10.1016/j.acra.2025.05.007.

[18] M. Saini, S. Hassanzadeh, B. Musa, M. Fatemi, and A. Alizad, "Variational Mode Directed Deep Learning Framework for Breast Lesion Classification Using Ultrasound Imaging," Sci. Rep., vol. 15, p. 14300, Apr. 2025, doi: 10.1038/s41598-025-99009-5.

[19] T. T. Yıldırım, O. Yaman, İ. Kılıc, B. Taşar, E. S. Timurkaan, and N. Aydŏgdu, "Multi-Class Classification of Breast Ultrasound Images Using Vision Transformer-Based Ensemble Learning," Diagnostics, vol. 15, no. 17, p. 2235, Sep. 2025, doi: 10.3390/diagnostics15172235.

[20] M. Sagarab, A. Rajaraman, and S. Antani, "Distilling Knowledge from an Ensemble of Vision Transformers for Improved Classification of Breast Ultrasound," Acad. Radiol., vol. 31, no. 1, pp. 104–120, Jan. 2024, doi: 10.1016/j.acra.2023.08.006.

[21] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4700–4708. doi: 10.1109/CVPR.2017.243.

[22] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021. doi: 10.48550/arXiv.2010.11929.