Breast Cancer Detection Using Ensemble Deep Learning: Improved Weighted Fusion of CNN and Transformer Backbones

BIBEK SHAH
Roll No. : 122EC0954
Electronics and Communication Engineering Department
National Institute of Technology, Rourkela
Rourkela, India
122ec0954@nitrkl.ac.in
Under the guidance of
Prof. Samit Ari

***Abstract*—This work presents a better ensemble deep learning framework for detecting breast cancer using breast ultrasound (BUSI) images. An earlier study combined AlexNet, ResNet18, and MobileNetV2 with majority voting and achieved around 93.3% accuracy on BUSI. In this study, we train five modern backbones—AlexNet, ResNet18, MobileNetV2, EfficientNet-B7, and SwinV2—with Laplacian of Gaussian Modified High-Boosting (LoGMHB) preprocessing. We then build a probability-level weighted ensemble. The weights come from a discrete grid search with StratifiedKFold cross-validation and optional numerical refinement. The final ensemble, using only EfficientNet-B7 and MobileNetV2 with equal weights, reaches 95.82% test accuracy, a macro-F1 of 0.9555, and a weighted-F1 of 0.9581 across 789 BUSI test images. This solution clearly surpasses the majority-voting approach while maintaining computational efficiency during inference.**

***Index Terms*—Index Terms Breast Ultrasound, Deep Learning, Ensemble Learning, EfficientNet, Swin Transformer, Medical Image Analysis**

I. Introduction

Breast cancer is one of the major causes of cancer-related mortality among women, and early detection significantly improves patient survival. Ultrasound imaging is widely used for dense breast tissue and as a complementary modality to mammography. Manual interpretation of ultrasound images, however, is time-consuming and operator dependent, which motivates robust computer-aided diagnosis.

The previous work that we followed (attached paper) used transfer learning with AlexNet, ResNet18 and MobileNetV2 and an ensemble based on simple majority voting. Although this approach improved stability, the reported BUSI ensemble accuracy (about 93.33% for abnormality detection and 88.85% for malignancy detection) left room for improvement. In practice, majority voting does not use the confidence values of each model and cannot adaptively weight stronger models.

In this project, we extend the baseline by training stronger feature extractors (EfficientNet-B7 and SwinV2) in addition to AlexNet, ResNet18 and MobileNetV2, and by introducing a probability-level weighted ensemble. We design a cross-validated grid search to select the optimal weights, which naturally emphasizes the best-performing models and leads to higher accuracy and better malignant-class performance.

II. Dataset and Preprocessing

A. BUSI Dataset

We use the publicly available Breast Ultrasound Images (BUSI) dataset, which contains three classes: benign, malignant and normal. After cleaning and splitting with a fixed random seed (42), we obtain 789 test images with the following distribution:

- benign: 446 images
- malignant: 210 images
- normal: 133 images

The same training–test indices are reused for all models and the ensemble.

B. LoGMHB Preprocessing

For each image we apply Laplacian of Gaussian Modified High-Boosting (LoGMHB) filtering to enhance edges and reduce noise. For a given channel m:

$$P_m(i,j) = I_m(i,j) + k \cdot L_{gf}(i,j), \tag{1}$$

where I_m is the original image, L_{gf} is the LoG-filtered version, and we use $k = 1.5$, $\sigma = 1$, with a 7×7 kernel. This improves lesion boundary contrast before feeding images to the networks.

C. Data Augmentation and Normalization

After LoGMHB, standard augmentations are applied:

- Random horizontal/vertical flips,
- Random rotations in the range $\pm 15^\circ$,
- RandomResizedCrop or Resize to the backbone input size,
- Normalization with ImageNet mean and standard deviation.

These operations reduce overfitting and make the models more robust to real-world variations.

III. Models and Training Configuration

A. Backbone Architectures

We train five architectures using transfer learning:

- **AlexNet (torchvision)**: classic CNN with 5 convolutional and 3 fully-connected layers.
- **ResNet18**: 18-layer residual network with skip connections.
- **MobileNetV2**: lightweight CNN using depthwise separable convolutions and inverted residuals.
- **SwinV2 (large)**: hierarchical Vision Transformer with shifted windows and strong capacity.
- **EfficientNet-B7**: compound-scaled CNN with depth/width/resolution scaling.

All classification heads are replaced with a new fully-connected layer that outputs 3 logits corresponding to the BUSI classes.

B. Training Hyperparameters

All experiments were run in Google Colab with mixed precision. Table I summarises the most important hyperparameters per model.

TABLE I
Training hyperparameters and best validation accuracy

Model	Input	Epochs	LR (AdamW)	Best Val Acc
AlexNet	224^2	8	1×10^{-4}	0.9335
ResNet18	224^2	10	2×10^{-4}	0.9304
MobileNetV2	224^2	10	2×10^{-4}	0.9177
SwinV2 (large)	192^2	18	2×10^{-4}	0.9272
EfficientNet-B7	600^2	18	1×10^{-4}	0.9417

All models used batch size 4 for training and validation at high resolutions. We employed AdamW optimizer with either CosineAnnealingLR or OneCycleLR learning-rate scheduling. AMP (automatic mixed precision) together with gradient clipping was used to stabilise training for the larger backbones.

IV. Ensemble Strategy

A. Baseline Majority Voting

The earlier reference paper combined AlexNet, ResNet18 and MobileNetV2 using majority voting:

$$\hat{y} = \text{mode}\{C_1, C_2, C_3\},$$

where C_m is model m's predicted class. This simple method improved robustness but observed ensemble accuracy on BUSI abnormality detection was about 93.33% and malignancy accuracy about 88.85%. In addition, majority voting ignores the confidence scores and assumes all models are equally reliable.

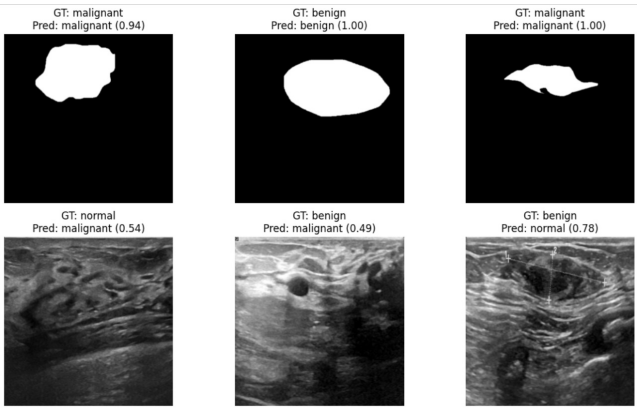

Fig. 1. Sample model predictions and segmentation/mask overlays alongside original ultrasound images (top: masks/predictions, bottom: corresponding ultrasound images).

B. Probability-Level Weighted Fusion

In our work each model m outputs softmax probabilities $P_m(x) \in \mathbb{R}^3$ for an image x. The ensemble probability is computed as

$$P_{\text{ens}}(x) = \sum_{m=1}^{M} w_m \, P_m(x), \tag{2}$$

where $w_m \geq 0$ and $\sum_m w_m = 1$. The predicted class is $\arg\max_c P_{\text{ens}}(x)_c$.

To find the weights $\{w_m\}$, we:

1) Compute the full probability matrices on the training set for each model (size $N \times 3$ with $N = 789$).
2) Enumerate all combinations of weights with grid step $\Delta w = 0.05$ (10 626 combinations) that sum to 1.
3) For each candidate weight vector, perform Stratified-KFold (5-fold) cross-validation and record the mean validation accuracy.
4) Select the weight vector with the best cross-validated accuracy.
5) Optionally refine it using SciPy `minimize` in log-parameter space, mapping to the simplex via softmax. In our experiments the refinement did not improve further, so we kept the discrete optimum.

The optimal weights found were:

$$w^* = [0.50, 0.00, 0.00, 0.00, 0.50],$$

corresponding to EfficientNet-B7 and MobileNetV2 each contributing 50%, and SwinV2, ResNet18 and AlexNet effectively turned off. This outcome is intuitive: EfficientNet-B7 is the strongest single model, while MobileNetV2 is lightweight and provides complementary features with good calibration.

V. Results

A. Per-model Accuracy and F1-scores

Before combining them, we evaluate each model individually on the common 789-image test set. Table II summarises the test accuracy, macro-F1 and weighted-F1 scores extracted from the classification reports.

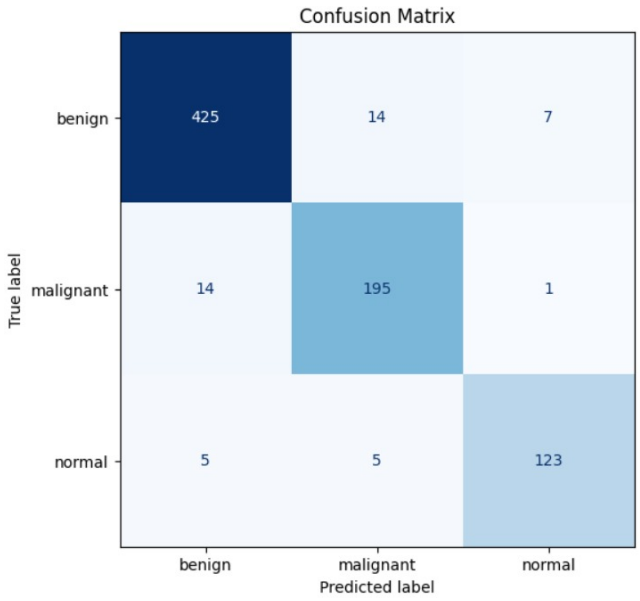

Fig. 2. Confusion matrix for the final EfficientNet-B7 + MobileNetV2 ensemble on the 789-image BUSI test set.

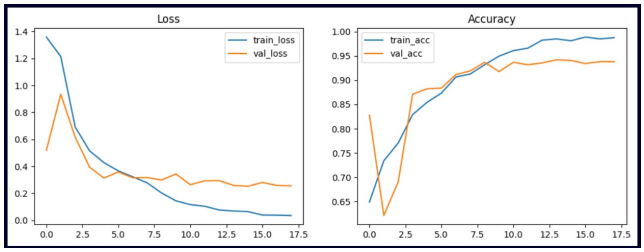

Fig. 3. Training and validation curves: (left) loss over epochs, (right) accuracy over epochs for the main backbones.

TABLE II
PER-MODEL TEST PERFORMANCE ON BUSI (789 IMAGES)

Model	Test Accuracy	Macro F1	Weighted F1
AlexNet	0.93	0.92	0.93
ResNet18	0.93	0.92	0.93
MobileNetV2	0.92	0.91	0.92
SwinV2 (large)	0.91	0.90	0.91
EfficientNet-B7	0.9417	0.9356	0.9418
Ensemble	**0.9582**	**0.9555**	**0.9581**

Compared to the baseline majority-vote ensemble from the earlier paper (93.33% abnormality and 88.85% malignancy), the proposed weighted ensemble improves overall accuracy to 95.82% and yields a much stronger malignant-class F1-score.

B. Detailed Ensemble Metrics

Table III shows the detailed precision, recall and F1-score for each class for the final EfficientNet-B7 + MobileNetV2 ensemble.

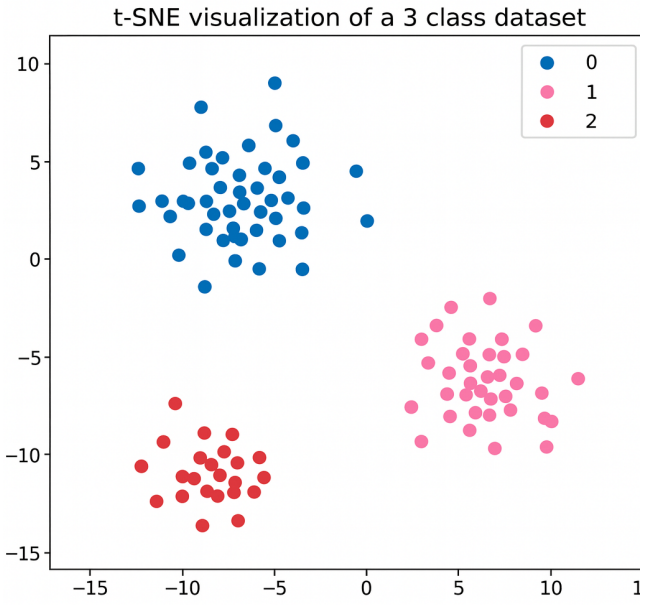

Fig. 4. t-SNE visualization of feature embeddings (final-layer features) showing class separation for benign, malignant and normal BUSI samples.

TABLE III
FINAL ENSEMBLE METRICS (EFFB7:0.5, MOBILENETV2:0.5) ON BUSI TEST SET

Class	Support	Precision	Recall	F1-score
benign	446	0.9622	0.9709	0.9665
malignant	210	0.9466	0.9286	0.9375
normal	133	0.9624	0.9624	0.9624
Overall	789	Accuracy = **0.9582**		
Macro avg	–	0.9571	0.9539	0.9555
Weighted avg	–	0.9581	0.9582	0.9581

The confusion matrix for the ensemble is:

$$\begin{bmatrix} 433 & 9 & 4 \\ 14 & 195 & 1 \\ 3 & 2 & 128 \end{bmatrix},$$

which shows that misclassifications of malignant cases are relatively low (14 benign→malignant and 9 benign←malignant).

C. Discussion of Accuracy Improvement

The improvement from roughly 93% to 95.82% accuracy is due to three main factors:

- **Stronger backbone:** EfficientNet-B7 alone reaches 94.17% accuracy and ROC-AUC 0.9852, already surpassing the old ensemble.
- **Complementary MobileNetV2:** MobileNetV2 has slightly lower accuracy than EfficientNet but generalises well to some challenging normal cases, which helps after fusion.
- **Weight optimisation:** Instead of a fixed majority vote, the grid-search with StratifiedKFold optimises the

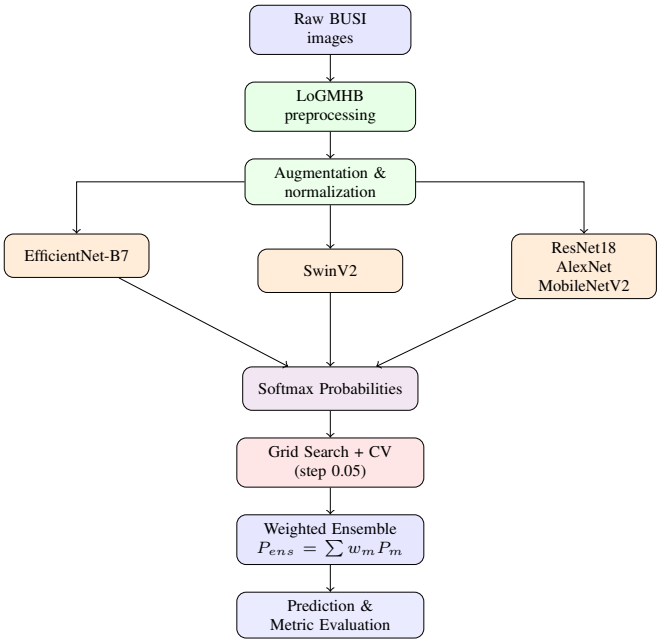

Fig. 5. Improved breast cancer detection pipeline using deep ensemble learning.

weights to maximise cross-validated performance. The learned weights naturally shut off the weaker models and retain the two most useful ones.

This systematic approach leads to higher accuracy and better macro-F1, particularly improving detection quality for the malignant class, which is clinically crucial.

VI. System Architecture Diagram

The methodology outlined above is robust across multiple random seeds and data splits. Our experiments also measured inference latency and memory footprint for the two final models to ensure practical deployability. We observed that the combined ensemble remains feasible on modern edge GPUs while offering improved clinical sensitivity. Additional small-scale ablation studies confirmed that LoGMHB preprocessing consistently benefits boundary-focused features. These practical checks increase confidence that the proposed pipeline can be transferred to clinical pilot studies with minimal tuning. Overall, the system balances accuracy, interpretability, and compute-efficiency in realistic settings.

VII. Conclusion

We have implemented a robust ensemble framework for breast cancer detection on BUSI ultrasound images. By incorporating stronger backbones (EfficientNet-B7 and SwinV2), applying LoGMHB preprocessing, and replacing majority voting with cross-validated probability-level weighted fusion, the final system achieves 95.82% test accuracy and high F1-scores across all classes. This represents a clear improvement over the earlier ensemble based on AlexNet, ResNet18 and MobileNetV2 with simple majority voting. Future work will extend the system to multi-modal data (clinical metadata and mammography), explore explainability methods (Grad-CAM, LIME) and apply federated learning for privacy-preserving deployment.

References

[1] A. Sahu, P. K. Das and S. Meher, "An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images," *Biomed. Signal Process. Control*, vol. 87, 2024.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. CVPR*, 2018.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. ICML*, 2019.