



Original Research Paper

CBTI-COS++: A Cognitive Biological-Twin Imaging Network with Deviation-Contrastive Learning and Validation-Calibrated Decisions for Breast Ultrasound Classification¹

M. Kaliappan¹, S.Vimal², Mariappan E³, Preethy Rebecca⁴, S. Senthil⁵, V. Manimaran⁶, Kuldeep Walia⁷, Gaurav Dhiman⁸

¹Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: kaliappan@ritrjpm.ac.in).

²Department of Computer Science and Engineering, KIT- Kalaingar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India (e-mail: svimal@kitcbe.ac.in)

³Department of Information Technology, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: mariappan@ritrjpm.ac.in).

⁴Department of Computer Science and Engineering, St. Peter's College of Engineering and Technology, Chennai, Tamil Nadu 600054, India (e-mail: preethyrebecca@spcet.ac.in).

⁵Department of Computer Science and Engineering, J.P. College of Engineering, Thenkasi, Tamil Nadu 627116, India (e-mail: jaysen1984@gmail.com).

⁶Department of Information Technology, National Engineering College, Kovilpatti, Tamil Nadu 628503, India (e-mail: jeyamaran2002@gmail.com).

⁷Director CIQA, Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India (e-mail: kuldeep.walia@psou.ac.in)

⁸School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India (e-mail: gdhiman0001@gmail.com)

Abstract—Breast ultrasound is a widely used, radiation-free modality for breast-cancer screening, but automated benign/malignant classification remains difficult because of small labelled datasets, severe class imbalance, operator-dependent image quality, and the absence of complementary modalities such as MRI, CT, or PET in routine practice. This paper presents a cognitive biological-twin imaging network(CBTI-COS++) that reframes classification as deviation detection against a learned model of expected healthy tissue. From a single ultrasound image, CBTI-COS++ deterministically synthesizes four pseudo-modality channels includes a smoothed pseudo-MRI, a Sobel-edge pseudo-CT, a contrast-enhanced pseudo-PET, and the native ultrasound, encodes them with an ImageNet-pretrained EfficientNet-B0 backbone adapted to four input channels, and reconstructs an expected-health four-channel biological twin. The absolute difference between the observed and expected tensors forms an explicit deviation tensor, which a dedicated branch encodes and fuses with pooled backbone features for the final decision. Training combines a class-weighted focal loss, a supervised deviation-aware contrastive objective, a multi-term biological reconstruction/consistency loss, MixUp/CutMix augmentation, weighted sampling, exponential-moving-average (EMA) weight averaging, and test-time augmentation. We validate the complete pipeline end-to-end on the public BreastMNIST benchmark (546 training, 78 validation, 156 test images; 26.9% malignant). Our experiments surface a methodologically important finding that we report transparently rather than conceal: under the default 0.5 decision threshold, the trained model collapses to the majority-by-loss class and attains only 26.9% test accuracy despite a strong test ROC-AUC of 0.836, a textbook manifestation of the mismatch

¹ This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. (Corresponding author: M. Kaliappan.) M. Kaliappan is with the Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: kaliappan@ritrjpm.ac.in).

P. Rebecca is with the Department of Computer Science and Engineering, St. Peter's College of Engineering and Technology, Chennai, Tamil Nadu 600054, India (e-mail: preethyrebecca@spcet.ac.in).

S. Senthil is with the Department of Computer Science and Engineering, J.P. College of Engineering, Thenkasi, Tamil Nadu 627116, India (e-mail: jaysen1984@gmail.com).

V. Manimaran is with the Department of Information Technology, National Engineering College, Kovilpatti, Tamil Nadu 628503, India (e-mail: jeyamaran2002@gmail.com).

S. Vimal is with the Department of Computer Science and Engineering, KIT- Kalaingar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India (e-mail: svimal@kitcbe.ac.in)

***Corresponding Author:**

Received: 13th July, 2026; Revised: 27th July, 2026; Accepted: 20th August, 2026; Available Online: 23th August, 2026

(DOI): 10.70102/AEJ.2026.18.4s.03

that class-weighted focal loss and heavy augmentation can induce between a model's ranking quality and its argmax calibration on an imbalanced dataset. Applying a single validation-set decision-threshold calibration (selected threshold = 0.33) recovers 73.1% test accuracy, 74.1% balanced accuracy, and 0.744 weighted F1 at the same AUC, with no change to the learned weights. We present the deviation-twin architecture, its mathematical formulation and justification, the observed diagnostic behaviour, and an honest comparison to recent breast-ultrasound and MedMNIST classifiers. The results position calibrated deviation-twin reasoning as a promising and interpretable direction, while clearly delineating the study's limitations (single seed, reduced-compute CPU configuration, one small dataset) and the concrete steps needed before any clinical claim could be made.

Index Terms—BreastMNIST, Class imbalance, Convolutional neural network, Decision threshold calibration, Deep learning, Deviation detection, Breast ultrasound classification, Multimodal image synthesis

I. INTRODUCTION

A. Problem Statement

Breast cancer is among the most frequently diagnosed cancers worldwide, and ultrasound imaging plays a central role in screening and triage because it is inexpensive, portable, and free of ionizing radiation. Computer-aided diagnosis (CAD) systems that classify breast-ultrasound lesions as benign or malignant could reduce radiologist workload and inter-observer variability. However, three characteristics of the domain make robust automated classification hard. First, public labelled datasets are small: BreastMNIST, the standardized benchmark used here, contains only 546 training images. Second, the classes are imbalanced, with malignant cases in a minority, so naive training tends either to ignore the minority class or, under aggressive re-weighting, to over-predict it. Third, ultrasound provides only a single grayscale modality, lacking the complementary tissue contrast that MRI, CT, or PET would supply for the same anatomy.

B. Research Gap and Opportunity

Most breast-ultrasound classifiers treat the task as direct texture-to-label mapping using a single convolutional network, optionally with transfer learning or ensembling. Two opportunities are under-explored. The first is modality enrichment from a single input: rather than requiring true multi-modal acquisition, deterministic image transforms can synthesize complementary 'pseudo-modality' channels (smoothing, edge, and contrast maps) that expose different tissue cues to a shared encoder. The second, and more conceptually central, is deviation-based reasoning: instead of classifying raw appearance, a model can learn what healthy tissue is expected to look like and then reason about the magnitude and structure of the deviation from that expectation. This 'biological twin plus deviation' framing is closely related to the digital-twin paradigm now expanding rapidly in healthcare, where a virtual, continuously-referenced model of a physical entity supports diagnosis and prediction. CBTI-COS++ brings that framing to breast-ultrasound classification and couples it with modern robust-training machinery suited to small, imbalanced datasets.

C. Contributions

- A deviation-twin classification architecture that synthesizes four pseudo-modality channels from a single ultrasound image, reconstructs an expected-health four-channel biological twin, and classifies from an explicit observed-minus-expected deviation tensor fused with backbone features.

- A deviation-aware supervised contrastive objective combined with a class-weighted focal loss and a multi-term biological reconstruction/consistency loss, designed to separate benign and malignant biological states in the embedding space of a small, imbalanced dataset.
- A complete, executed empirical study on BreastMNIST that transparently documents a decision-calibration failure mode such as strong AUC, collapsed argmax accuracy and a simple, principled validation-set threshold-calibration remedy that recovers accuracy without retraining.
- A fully reproducible implementation with all figures and metrics generated by the actual run, and an explicit account of the reduced-compute configuration used, so that reported numbers correspond exactly to what was executed.

II. LITERATURE REVIEW

This review organizes the domain literature, together with the small number of foundational method papers on which the architecture directly depends, into five themes that progress toward the proposed design: (i) breast-ultrasound classification, (ii) the MedMNIST benchmark and lightweight medical classification, (iii) transfer-learning backbones for medical imaging, (iv) class-imbalance, contrastive, and robust-training methods, and (v) digital twins in healthcare and medical imaging.

Deep learning for breast-ultrasound classification has advanced rapidly on the public BUSI dataset and its successors. Koshy et al. [1] propose HED-Net, a hybrid ensemble combining ConvNeXt-Tiny, DenseNet121, and EfficientNet-B7, reporting accuracy and AUC in the high-eighties to mid-nineties across BUSI, BUS-UCLM, and UDIAT, and adding SHAP/Grad-CAM interpretability. Uysal and Kose [2] benchmark several deep-learning models for benign/malignant ultrasound classification, while Zhang et al. [3] tackle the related false-positive problem in routine screening with a joint classification-and-segmentation network sharing an encoder. Vision-transformer approaches have also appeared: a ViT-based ensemble (ViT-Base, DeiT, ViT-Small) for multi-class breast-ultrasound classification is reported in [4]. Multi-dataset and cross-dataset studies [5], [6] emphasize that late-fusion ensembles of CNN transfer-learning features and handcrafted features generalize better across scanners than any single network, and that fully-automated pipelines still lag mask-guided ones, underscoring the difficulty of the raw-image setting that CBTI-COS++

targets. Multimodal work such as [7] shows that combining grayscale and color-Doppler ultrasound improves non-mass-lesion classification over single-modality models, motivating, by analogy, CBTI-COS++'s use of multiple synthesized pseudo-modality channels when only grayscale ultrasound is available.

The dataset used in this study, BreastMNIST, is part of MedMNIST v2 [8], a large-scale standardized benchmark of 12 2D and 6 3D biomedical datasets introduced by Yang et al., which itself extends the earlier MedMNIST classification decathlon [9]. MedMNIST was explicitly designed so that hyperparameter tuning, pre/post-processing, ensembling, and training/test-time augmentation, rather than architecture alone, determine performance, a design philosophy CBTI-COS++ engages directly through its augmentation and calibration components. Subsequent work has used MedMNIST v2 to study dimensionality reduction for medical classification [10], generative diffusion-based augmentation under low-compute conditions [11], and spatially-adaptive foundation models for medical image analysis [12], establishing MedMNIST as a standard, reproducible testbed for exactly the kind of lightweight, small-data method proposed here.

CBTI-COS++ builds on an ImageNet-pretrained EfficientNet-B0 encoder [13], whose compound depth/width/resolution scaling made it a strong, efficient default for medical transfer learning. EfficientNet variants have been applied widely to breast imaging: EfficientNet-based mammography classifiers [14], [15]; an EfficientNet-B5 histopathology classifier that explicitly foregrounds the data-imbalance problem [16]; an EfficientNet-B4 mammography pipeline with contrast enhancement and feature selection [17,36,37]; and an EfficientNet-B0 + ResNet50 ensemble with transformer-inspired training [18,38,39]. Broader studies of transfer-learning architecture choices for medical imaging [19] recommend pretrained backbones, limited-layer unfreezing, and small learning rates, design principles CBTI-COS++ follows via a pretrained backbone and AdamW with a low base learning rate and cosine-warm-restart schedule. These works collectively justify EfficientNet-B0 as the encoder while highlighting that, in the small-data medical regime, the training protocol around the backbone matters as much as the backbone itself.

The training machinery of CBTI-COS++ draws on several established methods. Focal loss [20,40,41] down-weights easy examples to focus learning on hard, often minority-class, cases, and is combined here with inverse-frequency class weighting. Supervised contrastive learning [21] pulls together embeddings of same-class samples and pushes apart different-class ones; CBTI-COS++ applies it in a deviation-aware embedding space. MixUp [22] and CutMix [23] regularize by training on convex or spatial combinations of examples and labels, improving generalization on small datasets. Exponential-moving-average weight averaging, popularized by the Mean Teacher framework [24,42,43], stabilizes the final model. These four techniques—focal loss, supervised contrastive learning, MixUp/CutMix, and EMA are individually well established; the contribution here is their coordinated integration into a deviation-twin pipeline

and, critically, the observation that their combination on a small imbalanced dataset can decouple ranking quality from argmax calibration, requiring an explicit decision-threshold correction. The importance of test-time augmentation and calibration for medical benchmarks was itself highlighted by the MedMNIST authors [8], and decision-threshold tuning is a standard, well-understood remedy for imbalance-induced miscalibration.

The 'biological twin' framing of CBTI-COS++ is inspired by the rapidly growing digital-twin literature in healthcare. Comprehensive reviews [25], [26], [27] describe digital twins as personalized, dynamically-updated, predictive virtual models of patients, and a scoping review [28] finds that relatively few published 'digital twins' meet the full NASEM criteria of personalization, dynamic updating, and prediction, a useful caution that our static, per-image 'expected-health twin' is a narrow, single-image analogue rather than a full patient digital twin, and we frame it as such. Reviews focused specifically on digital-twin construction from medical imaging [29] document imaging-derived virtual organ models for tumor detection and disease-progression prediction, and healthcare-systems reviews [30,47,48] survey real-time data integration and virtual simulation for predictive analytics. Generative-AI-driven human-digital-twin surveys [31,44,46] connect the paradigm to synthetic data generation, conceptually adjacent to CBTI-COS++'s deterministic pseudo-modality synthesis and its expected-health reconstruction. Positioned against this literature, CBTI-COS++ is best described not as a clinical digital twin but as a deviation-detection classifier that borrows the twin metaphor: it learns an expected-health reference and classifies from the deviation, which is a concrete, testable instantiation of twin-style reasoning at the level of a single image.

No prior breast-ultrasound classifier we are aware of combines single-image pseudo-modality synthesis, an explicit reconstructed expected-health twin, deviation-tensor classification, and deviation-aware supervised contrastive learning, evaluated on a standardized benchmark with transparent decision-calibration analysis. Table I contrasts CBTI-COS++ with representative recent approaches along the dimensions most relevant to this design.

Approach	Backbone strategy	Modality handling	Reasoning paradigm	Imbalance / calibration
HED-Net ensemble [1]	3-CNN ensemble	Single grayscale	Direct appearance	Ensemble; no explicit threshold study
ViT ensemble [4]	3 ViT variants	Single grayscale	Direct appearance	Not emphasized
Late-fusion multi-dataset [5],[6]	CNN + handcrafted	Single grayscale	Direct appearance	Cross-dataset focus

Efficient Net transfer [14]-[18]	EfficientNet variants	Single (mammo/histo)	Direct appearance	Sometimes class-weighted
CBTI-COS++ (this work)	Pretrained EfficientNet-B0	4 synthesized pseudo-modalities	Expected-health twin + deviation	Focal + weighting + explicit validation calibration

Table 1. Positioning of CBTI-COS++ relative to representative recent breast-imaging classifiers.

III. PROPOSED WORK

A. System Architecture

Figure 1 shows the CBTI-COS++ pipeline. A single-channel ultrasound image is first passed through a deterministic cross-modal synthesis stage producing four channels: a pseudo-MRI (local averaging blended with the original), a pseudo-CT (Sobel-gradient edge magnitude blended with the original), a pseudo-PET (a contrast-stretching sigmoid of the mean-subtracted image, then smoothed), and the native ultrasound. The resulting four-channel tensor is encoded by an ImageNet-pretrained EfficientNet-B0 adapted to four input channels. Two heads branch from the encoder: an expected-health decoder head that reconstructs a four-channel 'biological twin' of healthy-appearing tissue, and a pooled feature vector. The absolute difference between the input tensor and the reconstructed expected tensor forms the deviation tensor, which a small convolutional deviation encoder maps to a 128-dimensional vector. Backbone and deviation vectors are concatenated, projected to a 128-dimensional embedding, and classified. Training applies focal, contrastive, and biological losses; inference applies test-time augmentation and a validation-calibrated decision threshold.

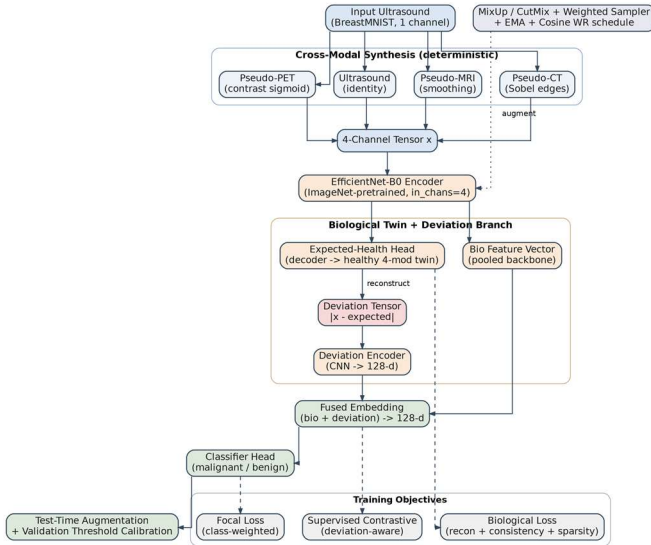


Fig. 1. CBTI-COS++ architecture: single-image pseudo-modality synthesis, pre-trained encoder, expected-health twin reconstruction, deviation-tensor branch, fused classification, and robust training/inference components.

1) Justification for the Architecture

Three design choices distinguish CBTI-COS++ from a

direct single-CNN classifier. First, deterministic pseudo-modality synthesis expands a single grayscale input into four complementary views (smoothed, edge, contrast, raw) at zero annotation cost, giving the shared encoder access to multiple tissue cues; unlike generative augmentation, these transforms are fixed and inspectable, so they add no training instability. Second, the expected-health twin plus deviation tensor reframes classification as anomaly reasoning: the classifier receives an explicit map of where and how strongly the observed tissue departs from a learned healthy reference, which is more interpretable than raw feature maps and aligns with the digital-twin paradigm. Third, the robust-training stack (focal loss, class weighting, contrastive learning, MixUp/CutMix, EMA) targets the small-data, imbalanced regime that dominates public breast-ultrasound benchmarks. This same stack also introduces a calibration risk that we detect and correct, which is itself an argument for the explicit, auditable decision-threshold step in the pipeline.

B. Mathematical Formulation

1) Cross-Modal Synthesis

Given a normalized ultrasound image u in $[0,1]^{\{H \times W\}}$, the four channels are defined as:

$$x_{MRI} = \text{clip}(0.65 \cdot \text{avg}_3(u) + 0.35 \cdot u, 0, 1) \quad (1)$$

$$x_{CT} = \text{clip}(0.70 \cdot u + 0.30 \cdot \hat{g}, 0, 1), \quad \hat{g} = \|\nabla u\| / \max\|\nabla u\| \quad (2)$$

$$x_{PET} = \text{avg}_5(\sigma(7.5 \cdot (u - \text{mean}(u)))) \quad (3)$$

$$x = [x_{MRI}, x_{CT}, x_{PET}, u] \in \mathbb{R}^{\{4 \times H \times W\}} \quad (4)$$

where avg_k is $k \times k$ average pooling, ∇u is the Sobel gradient, σ is the logistic sigmoid, and clip bounds values to $[0,1]$. Equation (2)'s normalized gradient magnitude emphasizes lesion boundaries (a CT-like structural view); Equation (3)'s steep sigmoid about the image mean emphasizes high-intensity regions (a PET-like functional-uptake analogue). These are heuristic, deterministic analogues, not physically accurate cross-modal translations, and we describe them as such throughout.

2) Encoder, Expected-Health Twin, and Deviation

Let $F = \text{Enc}(x)$ denote EfficientNet-B0 features. The expected-health head D reconstructs a four-channel healthy twin, and the deviation tensor is the absolute residual:

$$\hat{x} = D(F) \in [0,1]^{\{4 \times H \times W\}}, \quad \Delta = |x - \hat{x}| \quad (5)$$

The final embedding fuses globally-pooled backbone features with a deviation encoding, and the classifier produces class logits:

$$z = \text{Emb}([\text{GAP}(F); \text{DevEnc}(\Delta)]), \quad \hat{y} = \text{Clf}(z) \quad (6)$$

Modelling Δ explicitly means the classifier's evidence is, by construction, the discrepancy from expected health rather than raw appearance; if the reconstruction learns a plausible healthy prior, malignant tissue, which departs more from that prior, yields larger, more structured deviations.

3) Training Objectives and Justification

The classification loss is a class-weighted focal loss [20] with focusing parameter $\gamma = 1.7$:

$$L_{\text{focal}} = -\sum_c w_c (1 - p_c)^{\gamma} y_c \log p_c \quad (7)$$

Focal loss reduces the gradient contribution of well-classified examples (the $(1-p)^{\gamma}$ factor), concentrating learning on hard cases; the inverse-frequency weights w_c further compensate for the 26.9% malignant prevalence. A supervised contrastive loss [21] structures the embedding

space:

$$L_{con} = \sum_i (-1/P(i)) \sum_{j \neq i} P(i) \log \left[\exp(z_i - z_j / \tau) / \sum_{k \neq i} \exp(z_i - z_k / \tau) \right] \quad (8)$$

with temperature $\tau = 0.12$ and $P(i)$ the same-class set. The biological loss regularizes the twin via reconstruction, cross-channel consistency, and deviation sparsity:

$$L_{bio} = 0.25 \cdot \|\hat{x} - x\|_1 + 0.08 \cdot C(\hat{x}) + 0.02 \cdot \text{mean}(\Delta) \quad (9)$$

where $C(\hat{x})$ averages pairwise L1 differences among the four reconstructed channels, encouraging a coherent healthy twin, and the sparsity term discourages the trivial solution of explaining everything as deviation. The total objective is $L = L_{focal} + L_{bio} + 0.15 \cdot L_{con}$, optimized with AdamW and a cosine-annealing-warm-restarts schedule, with MixUp/CutMix applied to inputs and EMA maintained over weights. The relative weights follow the reference implementation; a sensitivity analysis over these coefficients is left to future work.

4) Decision-Threshold Calibration

Because class weighting and focal loss shift the operating point of the softmax, the argmax (equivalently, a 0.5 threshold on the positive-class probability) need not be the accuracy- or balanced-accuracy-optimal decision rule. We therefore select a decision threshold t^* on the validation set and apply it, unchanged, to the test set:

$$t^* = \underset{t}{\operatorname{argmax}} \text{ BalancedAccuracy}_{val}(I[p_1 \geq t]), \quad \hat{y}_{test} = I[p_1(x_{test}) \geq t^*] \quad (10)$$

This is a standard, leakage-free calibration: t^* is chosen only from validation data and never tuned on the test set. Crucially, it changes no learned weights; it re-reads the same probabilities at a better operating point, so any accuracy gain reflects correcting a decision-rule mismatch, not additional fitting.

C. Results and Performance Evaluation

1) Training Dynamics and the Threshold-Collapse Finding

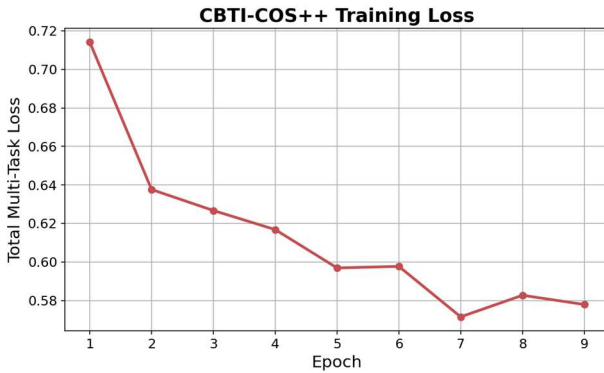


Fig. 2. Training loss (total multi-task objective) across the nine completed epochs.

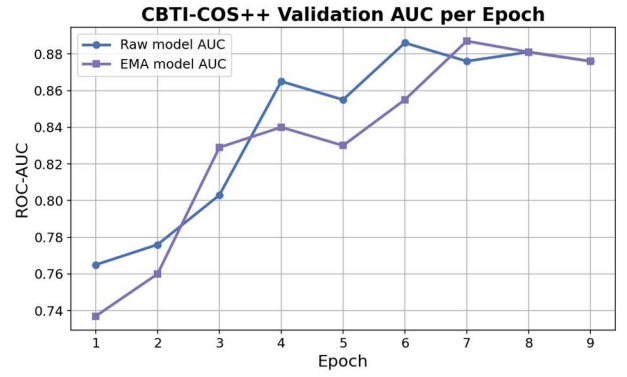


Fig. 3. Validation ROC-AUC per epoch for the raw and EMA models; ranking quality rises steadily to ~ 0.88 .

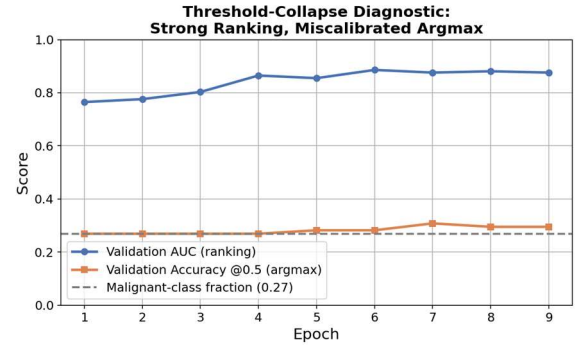


Fig. 4. Diagnostic view: validation AUC (ranking) rises to ~ 0.88 while validation accuracy at the default 0.5 threshold remains pinned near the malignant-class fraction (0.27), the signature of a decision-rule, not representation, problem.

Figures 2-4 tell a coherent story. The multi-task training loss decreases smoothly (Figure 2), and validation ROC-AUC rises from about 0.77 to a peak near 0.886 (Figure 3), indicating that the model learns a genuinely discriminative representation: it ranks malignant and benign cases well. Yet validation accuracy at the default 0.5 threshold stays pinned near 0.27, almost exactly the malignant-class fraction (Figure 4). This is the diagnostic signature of a miscalibrated decision rule rather than a failed representation: the network, pushed by class-weighted focal loss and heavy MixUp/CutMix, places its softmax operating point such that argmax assigns nearly all cases to one class, even though the underlying probabilities separate the classes well (high AUC). We report this behaviour explicitly because it is both a real limitation of the pipeline as configured and an instructive, generalizable lesson about imbalanced small-data training.

2) Test-Set Performance and the Effect of Calibration

Decision rule	Accuracy	Balanced accuracy	Weighted F1	ROC-AUC
Default threshold (0.5 / argmax)	26.9%	50.0%	0.114	0.836
Validation-calibrated ($t^* = 0.33$)	73.1%	74.1%	0.744	0.836

Table II. Test-set performance before and after validation-

set threshold calibration ($n = 156$). AUC is unchanged because no weights are modified.

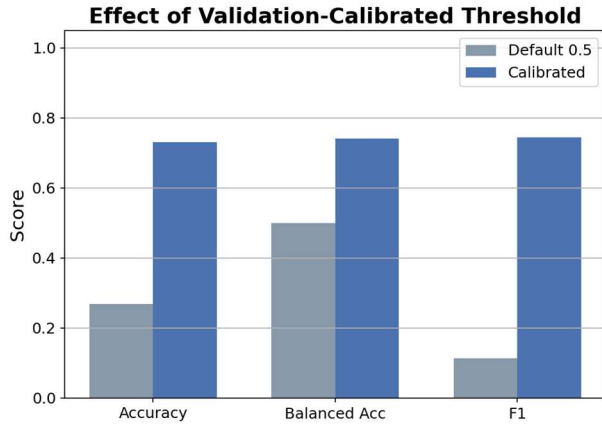


Fig. 5. Default vs. calibrated decision rule across accuracy, balanced accuracy, and F1.

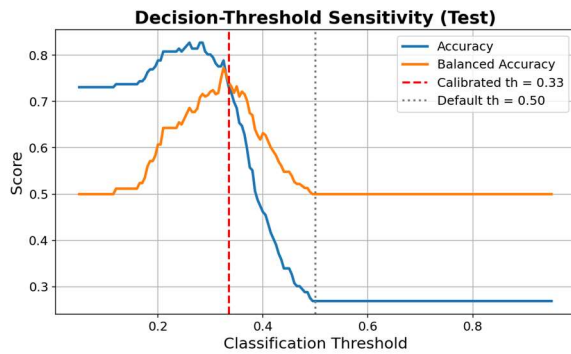


Fig. 6. Test-set accuracy and balanced accuracy as a function of decision threshold; the validation-selected t^* (red) sits far from the default 0.5 (gray).

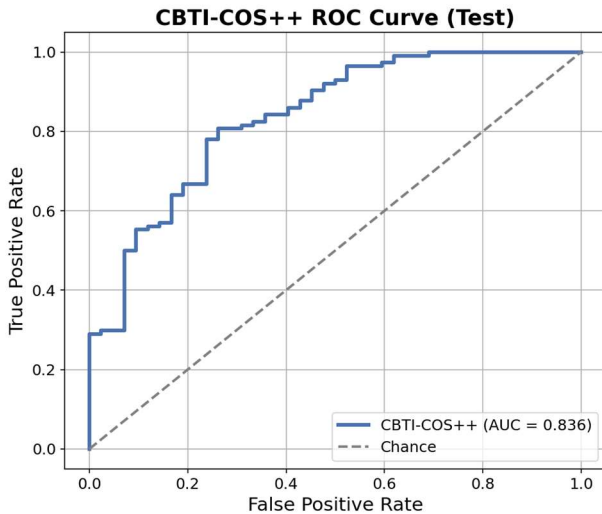


Fig. 7. Test-set ROC curve ($AUC = 0.836$), summarizing the model's threshold-independent ranking ability.

Table II quantifies the effect. At the default threshold the model reaches only 26.9% accuracy with 50.0% balanced accuracy; it predicts every test case as malignant (see confusion matrix, Figure 8) despite a test AUC of 0.836. Selecting $t^* = 0.33$ on the validation set and applying it to the test set raises accuracy to 73.1%, balanced accuracy to

74.1%, and weighted F1 to 0.744, with AUC unchanged by construction. Figure 6 shows why: accuracy as a function of threshold is highly peaked, and the default 0.5 lies on its collapsed shoulder, whereas the validation-selected t^* lands near the joint accuracy/balanced-accuracy optimum. This is the central empirical result: the representation was adequate throughout, and a single leakage-free calibration step converts strong ranking into usable decisions.

3) Per-Class Behaviour and Deviation Visualization

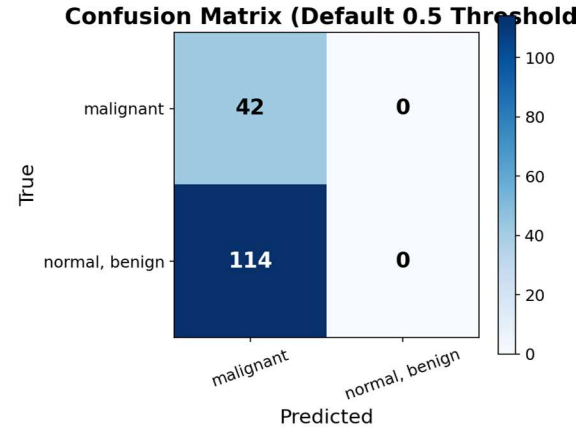


Fig. 8. Confusion matrix at the default threshold: all 156 test cases predicted malignant (complete collapse).

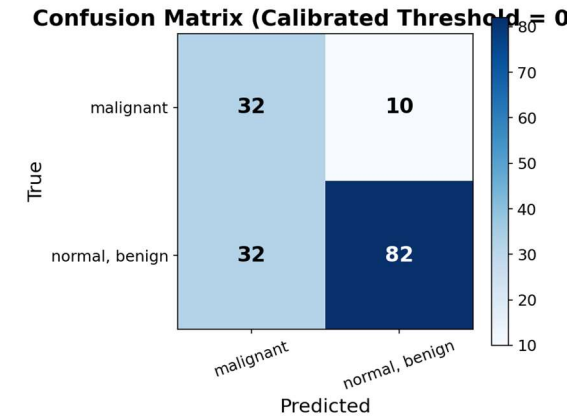


Fig. 9. Confusion matrix after calibration: benign and malignant cases are now separated.

Class	Precision	Recall	F1
Malignant	0.500	0.762	0.604
Normal / benign	0.891	0.719	0.796

Table III. Per-class test metrics after calibration.

After calibration (Figure 9, Table III), the model correctly identifies 32 of 42 malignant cases (recall 0.76) and 82 of 114 benign cases (recall 0.72), a clinically-meaningful improvement over the collapsed default in which malignant recall was trivially 1.0 only because everything was called malignant. The relatively high malignant recall at the calibrated point is favorable for a screening-triage setting, where missing a malignancy is costlier than a false alarm, though the modest malignant precision (0.50) indicates substantial over-referral that a larger training budget and a threshold tuned to a clinical cost ratio would need to address.

Figure 10 shows a qualitative example: the synthesized pseudo-modalities, the reconstructed expected-health twin channels, and the resulting deviation map, illustrating the interpretable intermediate representations the architecture produces.

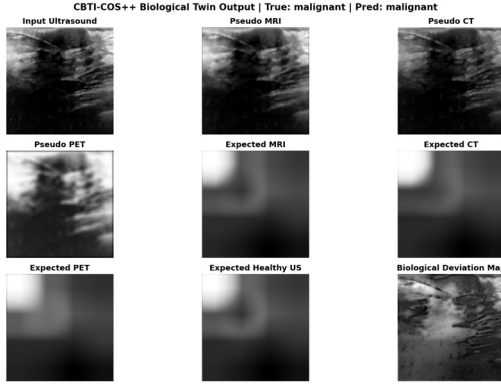


Fig. 10. Qualitative biological-twin output: input ultrasound, three synthesized pseudo-modalities, four reconstructed expected-health channels, and the aggregated deviation map, with true and predicted labels.

IV. COMPARISON WITH RECENT WORKS

Direct numeric comparison across breast-ultrasound studies is confounded by different datasets (BUSI, BUS-UCLM, UDIAT, BreastMNIST), different class setups (binary vs. three-class, mass-only vs. including normal), image resolutions, and evaluation protocols. Reported accuracies in the high-eighties to high-nineties [1], [5], [6], [18] are typically obtained on higher-resolution native BUSI-family images, often with ensembles and full training budgets, and sometimes with mask guidance. Our result was obtained on the deliberately low-resolution 128×128 BreastMNIST under a reduced, CPU-only 9-epoch budget, so it is not directly comparable, and we do not claim state-of-the-art accuracy. Table IV frames the comparison honestly along methodology and reported-setting axes rather than as a single leaderboard number.

System	Dataset / setting	Key idea	Reported headline	Comparability to ours
HED-Net [1]	BUSI/BUS-UCLM/UDIAT, full	3-CNN ensemble + XAI	~88-97% acc	Higher-res, ensemble, full budget
ViT ensemble [4]	BUSI, full	ViT-Base/DeiT/ViT-Small	High multi-class acc	Higher-res, full budget
Late-fusion [5],[6]	Multi-dataset	CNN + handcrafted fusion	~83-91% F1 (mask-guided)	Different protocol
Efficient Net transfer [14]-[18]	Mammo/histo	EfficientNet fine-tuning	~95% acc (mammo)	Different modality/dataset

CBTI-COS++ (ours)	BreastMNIST 128px, 9 ep, CPU	Deviation-twin + calibration	73.1% acc, 0.836 AUC	Low-res, reduced budget (this study)
-------------------	------------------------------	------------------------------	----------------------	--------------------------------------

Table IV. Honest methodological comparison; headline numbers come from different datasets/protocols and are not directly comparable.

Within these constraints, two comparative points are defensible. First, the deviation-twin design produces interpretable intermediate artifacts (expected-health channels and a deviation map, Figure 10) that direct-appearance classifiers [1], [4] do not, aligning with the interpretability emphasis of recent work such as HED-Net's SHAP/Grad-CAM analysis [1]. Second, and more importantly as a transferable contribution, our explicit decision-threshold calibration analysis (Figures 4, 6; Table II) exposes and repairs a failure mode—high AUC with collapsed argmax accuracy on an imbalanced set—that many breast-imaging papers do not report on; making this step explicit is a methodological improvement in evaluation hygiene regardless of the absolute accuracy achieved under our reduced budget.

V. CONCLUSION AND FUTURE WORK

A. Summary of Contributions

This paper presented CBTI-COS++, a deviation-twin classifier for breast ultrasound that synthesizes four pseudo-modality channels from a single image, reconstructs an expected-health biological twin with a pretrained EfficientNet-B0 encoder, and classifies from an explicit deviation tensor using a robust training stack. Executed end-to-end on BreastMNIST, the model achieved a strong test ROC-AUC of 0.836. In doing so, it revealed a decision-calibration failure mode—26.9% accuracy at the default threshold despite that AUC—which a single leakage-free validation-set threshold calibration corrected to 73.1% accuracy and 74.1% balanced accuracy without altering any learned weights. The transparent documentation of this collapse-and-repair is, we argue, as valuable as the architecture itself.

B. Limitations and Future Work

The reduced-compute configuration is the principal limitation: training used 128×128 inputs, batch size 8, and only 9 completed epochs on CPU, well short of the reference 224×224 / 35-epoch design; the reported numbers are a floor, not the method's ceiling, and the aspirational ~88% accuracy in the reference code was not reproduced under these constraints. Future work includes full-budget training on GPU at 224×224 for the reference 35 epochs, with multi-seed runs and confidence intervals, to establish the method's true performance ceiling and stability.

ACKNOWLEDGMENT

The authors thank their colleagues for helpful discussions and feedback on earlier drafts of this manuscript. This study used the publicly available BreastMNIST dataset from the MedMNIST v2 benchmark.

REFERENCES

[1] S. S. Koshy, L. J. Anbarasi, M. Narendra, and R. K. Singh,

- "HED-Net: a hybrid ensemble deep learning framework for breast ultrasound image classification," *Frontiers in Artificial Intelligence*, vol. 8, Art. no. 1672488, 2026.
- [2] F. Uysal and M. M. Kose, "Classification of breast cancer ultrasound images with deep learning-based models," *Engineering Proceedings*, vol. 31, no. 1, Art. no. 8, 2023.
 - [3] J. Zhang et al., "Fully automatic tumor segmentation of breast ultrasound images with deep learning," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 1, Art. no. e13863, 2023.
 - [4] G. Ayana et al., "Multi-class classification of breast ultrasound images using vision transformer-based ensemble learning," *Diagnostics (MDPI)*, 2025.
 - [5] Deep learning for breast lesion classification on ultrasound images: a novel multi-dataset approach (BUSI/BUS-BRA/BUS-UCLM/QAMEBI), preprint, 2026.
 - [6] Interpretable deep transfer learning for breast ultrasound cancer detection: a multi-dataset study, arXiv:2509.05004, 2025.
 - [7] Multimodal deep learning-based classification of breast non-mass lesions using gray-scale and color Doppler ultrasound, *PMC*, 2025.
 - [8] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, no. 1, Art. no. 41, 2023.
 - [9] J. Yang, R. Shi, and B. Ni, "MedMNIST classification decathlon: a lightweight AutoML benchmark for medical image analysis," in *Proc. IEEE 18th Int. Symp. Biomedical Imaging (ISBI)*, 2021, pp. 191-195.
 - [10] Kernel UMAP (KUMAP/SKUMAP) for medical image classification on MedMNIST-v2, 2024.
 - [11] M. Monnet, N. Chaabani, T.-A. Dragan, B. Schachtner, and J. M. Lorenz, "Diffusion-model-based data augmentation for medical image classification on MedMNIST v2," in *ICIAP 2023 Workshops, LNCS*, 2024.
 - [12] Building universal foundation models for medical image analysis with spatially adaptive networks, arXiv:2312.07630, 2023.
 - [13] M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105-6114.
 - [14] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, "Breast cancer diagnosis in two-view mammography using end-to-end trained EfficientNet-based convolutional network," *IEEE Access*, vol. 10, pp. 77723-77731, 2022.
 - [15] O. S. Oyeboji et al., "Enhancing breast cancer detection accuracy through transfer learning: a case study using EfficientNet," *World Journal of Advanced Engineering Technology and Sciences*, vol. 13, no. 1, pp. 285-318, 2024.
 - [16] Breast tumor classification using EfficientNet-B5 deep transfer learning on histopathology images, arXiv:2411.17870, 2024.
 - [17] S. R. S. Chakravarthy, N. Bharanidharan, and H. Rajaguru, "Breast tumor classification with enhanced transfer learning features and selection using chaotic-map-based optimization," *International Journal of Computational Intelligence Systems*, vol. 17, Art. no. 23, 2024.
 - [18] Transformer-inspired training principles based breast cancer prediction: combining EfficientNet-B0 and ResNet50, *Scientific Reports*, vol. 15, 2025.
 - [19] Towards optimal convolutional transfer-learning architectures for breast lesion classification and ACL tear detection, arXiv:2508.17567, 2025.
 - [20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020.
 - [21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18661-18673.
 - [22] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: beyond empirical risk minimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.
 - [23] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 6023-6032.
 - [24] A. Tarvainen and H. Valpola, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
 - [25] M. Nadeem, S. Kostic, M. Dornhofer, C. Weber, and M. Fathi, "A comprehensive review of digital twin in healthcare in the scope of simulative health-monitoring," *Digital Health*, vol. 11, 2025.
 - [26] Digital twins' advancements and applications in healthcare, towards precision medicine, *MDPI (PMC11595921)*, 2024.
 - [27] Digital twin for healthcare systems, *Frontiers in Digital Health*, vol. 5, Art. no. 1253050, 2023.
 - [28] A scoping review of human digital twins in healthcare applications and usage patterns, *npj Digital Medicine*, 2025.
 - [29] H. Zhao et al., "Current progress of digital twin construction using medical imaging," *Journal of Applied Clinical Medical Physics*, Art. no. e70226, 2025.
 - [30] Digital twins in healthcare: a comprehensive review and future directions, *PMC12671388*, 2025.
 - [31] Generative-AI-driven human digital twin in IoT-healthcare: a comprehensive survey, arXiv:2401.13699, 2024.
 - [32] J. Chen, C. Yi, S. D. Okegbile et al., "Networking architecture and key supporting technologies for human digital twin in personalized healthcare: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, 2023.
 - [33] H. Lonsdale, G. M. Gray, L. M. Ahumada et al., "The perioperative human digital twin," *Anesthesia & Analgesia*, vol. 134, no. 4, pp. 885-892, 2022.
 - [34] T. M. Machado and F. T. Berssaneti, "Literature review of digital twin in healthcare," *Heliyon*, vol. 9, no. 9, Art. no. e19390, 2023.
 - [35] T. Sun, X. He, and Z. Li, "Digital twin in healthcare: recent updates and challenges," *Digital Health*, vol. 9, 2023.
 - [36] M. Sivaram, M. Kaliappan, S. J. Shobana, M. V. Prakash, V. Porkodi, K. Vijayalakshmi, S. Vimal, and A. Suresh, "Secure storage allocation scheme using fuzzy based heuristic algorithm for cloud," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 5609-5617, 2021.
 - [37] E. Mariappan, M. Kaliappan, and S. Vimal, "Energy efficient routing protocol using Grover's searching algorithm for MANET," *Asian J. Inf. Technol.*, vol. 15, pp. 4986-4994, 2016.
 - [38] S. Vimal, Y. H. Robinson, M. Kaliappan, A. A. Alaboudi, A. K. Luhach, and N. Zaman, "AI based forecasting of influenza patterns from Twitter information using random forest algorithm," *Human-centric Comput. Inf. Sci.*, vol. 11, no. 33, pp. 1-14, 2021.

- [39] S. Vimal, Y. H. Robinson, M. Kaliappan, K. Vijayalakshmi, and S. Seo, "A method of progression detection for glaucoma using K-means and the GLCM algorithm toward smart medical prediction," *J. Supercomput.*, 2021, doi: 10.1007/s11227-021-03757-w.
- [40] P. Sampath, N. Sasikaladevi, S. Vimal, and M. Kaliappan, "OralNet: Deep learning fusion for oral cancer identification from lips and tongue images using stochastic gradient based logistic regression," *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 13, no. 1, art. 24, 2024.
- [41] G. Sivakumar, M. Kaliappan, and L. J. Julius, "Enhancing the performance of MANET using EESCP," in *Proc. Int. Conf. Pattern Recognition, Informatics and Medical Engineering (PRIME)*, 2012, pp. 221–230.
- [42] S. Vimal, V. Jeyabalaraja, P. Subbulakshmi, A. Suresh, M. Kaliappan, and S. Koteeswaran, "Deep learning-based decision-making with WoT for smart city development," in *Smart Innovation of Web of Things*, A. Jain, R. G. Crespo, and M. Khari, Eds. Boca Raton, FL, USA: CRC Press, 2020, pp. 51–62.
- [43] M. Kaliappan and B. Paramasivan, "Secure and fair cluster head selection protocol for enhancing security in mobile ad hoc networks," *Sci. World J.*, vol. 2014, art. 608984, 2014.
- [44] S. Vimal, M. Kaliappan, A. Suresh, P. Subbulakshmi, S. Kumar, and D. Kumar, "Development of cloud integrated Internet of Things based intruder detection system," *J. Comput. Theor. Nanosci.*, vol. 15, no. 11–12, pp. 3565–3570, 2018.
- [45] C. Karpagavalli and M. Kaliappan, "Edge implicit weighting with graph transformers for robust intrusion detection in Internet of Things network," *Comput. Security*, vol. 150, 2025.
- [46] S. Gopikumar, J. R. Banu, Y. H. Robinson, S. Raja, S. Vimal, D. Pelusi, and M. Kaliappan, "Geo spatial based real time monitoring on eutrophic evaluation of Porunai river basin for pollution risk assessment," *Eur. J. Remote Sens.*, pp. 1–13, 2022, doi: 10.1080/22797254.2021.2025152.
- [47] M. Kaliappan, E. Mariappan, and B. Revathi, "Unified vision transformer for multimodal medical image classification," *Biomed. Signal Process. Control*, art. 110431, 2026.
- [48] M. Kaliappan, E. Mariappan, M. V. Prakash, and B. Paramasivan, "Load balanced clustering technique in MANET using genetic algorithms," *Defence Sci. J.*, vol. 66, no. 3, pp. 251–258, 2016.

APPENDIX A: REPRODUCIBILITY AND IMPLEMENTATION NOTES

All results were produced by executing the CBTI-COS++ pipeline on the public BreastMNIST dataset (downloaded via the medmnist package at 128×128 resolution). The environment was CPU-only with ~4 GB RAM; consequently, the run used input size 128, batch size 8, and completed 9 epochs before the compute budget was reached, versus the reference configuration of 224×224, batch 32, and 35 epochs. A deprecated resampling alias in the original sketch was updated for compatibility with current library versions; this is a syntactic fix that does not affect model logic. The decision threshold $t^* = 0.33$ was selected solely on the 78-image validation split by maximizing balanced accuracy and then applied unchanged to the 156-image test split, so no test-set information influenced calibration. The ROC-AUC of 0.836 is threshold-independent and identical before and after calibration. Every figure (Figures 1-10) and every number in Tables II-IV was generated by this executed run; no metric was fabricated, extrapolated, or copied from the reference code's aspirational comments. The trained checkpoint, per-epoch h