Acta Informatica Pragensia
2027, Volume 16, Issue 1, Forthcoming article
https://doi.org/10.18267/j.aip.325

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Article Open Access

Mask-Prior Knowledge Distillation for Efficient Binary Breast Ultrasound Classification

Windra Swastika, Paulus Lucky Tirma Irawan

Faculty of Technology and Design, Universitas Ma Chung, Malang, Indonesia

Corresponding author: Windra Swastika (windra.swastika@machung.ac.id)

Editorial Record

First submission received:
April 14, 2026

Revisions received:
June 6, 2026
July 14, 2026

Accepted for publication:
July 29, 2026

Academic Editor:
Zdenek Smutny
Prague University of Economics and Business, Czech Republic

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

How to cite this article:
Swastika, W., & Irawan, P. L. T. (2027). Mask-Prior Knowledge Distillation for Efficient Binary Breast Ultrasound Classification. *Acta Informatica Pragensia*, 16(1), Forthcoming article. https://doi.org/10.18267/j.aip.325

Copyright:
© 2027 by the author(s). Licensee Prague University of Economics and Business, Czech Republic. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

Abstract

Background: Binary classification of breast ultrasound (BUS) images into benign and malignant categories is a clinically critical task. Standard knowledge distillation (KD) compresses large teacher models into compact student models but transfers knowledge in a spatially blind manner, leaving ground-truth segmentation masks, routinely available in BUS datasets, entirely unexploited.

Objective: This study proposes mask-prior knowledge distillation (MP-KD), a novel distillation framework that integrates ground-truth lesion segmentation masks as spatial priors to improve both classification performance and attention interpretability in a lightweight student model.

Methods: A multi-teacher ensemble of EfficientNet-B4 and ResNet50, pre-trained and fine-tuned on the Breast Ultrasound Image (BUSI) dataset (697 images; 487 benign, 210 malignant), generates soft targets and attention guidance distilled into an EfficientNet-B0 student (5.3M parameters). The proposed mask-weighted attention transfer (MWAT) loss combines two components: MWAT-AT, which applies mask-derived spatial weights to the attention transfer loss to amplify distillation signals within lesion regions; and MWAT-MA, an auxiliary mask alignment objective that directly supervises student attention towards the ground-truth mask. Loss weights were selected via grid search (18 combinations) on the validation set. All the experiments were repeated with three random seeds (42, 123, 456) on a stratified 70/10/20 split; performance is reported as mean ± standard deviation on the held-out test set.

Results: MP-KD achieves the highest AUC of 0.954 ± 0.004 among all the student configurations, outperforming standard KD (0.944 ± 0.002) and the no-KD baseline (0.949 ± 0.004). MP-KD also yields the highest mask-attention alignment score among all the distillation-based configurations (MAAS = 0.608 ± 0.006), substantially recovering the attention alignment lost during standard KD (0.577) while maintaining superior classification performance. Ablation studies show that MWAT-AT primarily drives accuracy (0.908) and specificity (0.921) improvements, while MWAT-MA yields the highest single-run sensitivity (0.952). The combined MP-KD model provides the most robust performance across all the seeds (MCC = 0.732 ± 0.007, AUC std = 0.004, smallest among all the configurations). Compared to the distilled ViT approach of Zhou and Mosadegh (AUC = 0.972, ~86M parameters), MP-KD achieves competitive performance with 16× fewer parameters.

Conclusion: MP-KD demonstrates that ground-truth segmentation masks can be effectively integrated as spatial priors within the knowledge distillation pipeline, improving both the AUC and attention alignment of a compact student model compared to standard KD. The proposed MWAT mechanism offers a practical and interpretable solution for deploying breast cancer classification in resource-constrained clinical environments and is generalizable to other medical imaging tasks where segmentation annotations are available.

Index Terms

Breast ultrasound; Knowledge distillation; Attention transfer; Binary classification; EfficientNet; Segmentation mask; Deep learning; Medical imaging.

1 INTRODUCTION

Breast cancer is the most commonly diagnosed malignancy among women worldwide, accounting for approximately 2.3 million new cases annually (Sung et al., 2021). Early detection through imaging-based screening significantly improves patient outcomes and breast ultrasound (BUS) has emerged as an important adjunct to mammography, particularly for patients with dense breast tissue (Burnside et al., 2009). BUS is widely accessible, radiation-free and capable of discriminating solid from cystic lesions; however, its interpretation is highly operator-dependent and subject to inter-observer variability (Stavros et al., 1995). Automated deep learning systems for BUS classification offer the potential to standardize and augment radiologist decision-making.

Convolutional neural networks (CNNs) have demonstrated strong performance on BUS binary classification (benign vs malignant), with recent studies reporting AUC values exceeding 0.94 on the BUSI dataset (Ali et al., 2023; Gheflati & Rivaz, 2022; Jabeen et al., 2022; Latha et al., 2024). More recently, vision transformers (ViT) have achieved state-of-the-art results on this task, with ensemble ViT models reaching AUC of 0.972 (Zhou & Mosadegh, 2024). However, these high-performing architectures typically involve tens to hundreds of millions of parameters, limiting their practicality in resource-constrained clinical environments such as point-of-care ultrasound settings, mobile diagnostic units and hospitals in low-to-middle income countries.

Knowledge distillation (KD) (Gou et al., 2021; Hinton et al., 2015) addresses this challenge by training a compact student model to mimic the soft-label outputs of a larger teacher model. The resulting student retains much of the teacher's predictive power at a fraction of the computational costs. In medical image segmentation, KD from diverse teacher ensembles produces student networks that match ensemble-level performance (Noothout et al., 2022). However, standard KD transfers knowledge through output-level distributions alone, without accounting for spatial structure. Ground-truth segmentation annotations, which delineate lesion location and extent, constitute a rich spatial prior that standard KD leaves entirely unexploited.

The BUSI dataset (Al-Dhabyani et al., 2020), the benchmark most widely used for BUS classification, provides ground-truth segmentation masks for each image. These masks identify the lesion boundary with clinical precision. Yet to our knowledge, no prior KD framework for BUS classification has made use of these masks to spatially guide the knowledge transfer process. Feature-based distillation methods such as attention transfer (AT) (Zagoruyko & Komodakis, 2017) and FitNets (Romero et al., 2015) transfer intermediate representations but do not incorporate domain-specific spatial priors. Spatial attention mechanisms in KD have been explored in natural image domains (Park et al., 2019), but their adaptation to medical imaging with ground-truth mask guidance remains unexplored.

Knowledge distillation has been increasingly adopted in medical image analysis beyond classification. Channel-wise distillation can effectively compress U-Net architectures while preserving segmentation accuracy (Qin et al., 2021). Structured KD for medical image segmentation has been proposed to reduce redundant information transfer between teacher and student networks (Zhao et al., 2023). Teacher-student distillation has been applied specifically to on-device breast cancer image classification, achieving efficient inference on mobile hardware, a deployment scenario that closely aligns with the clinical motivation of this study (Sepahvand & Abdali-Mohammadi, 2023). However, none of these approaches have incorporated domain-specific spatial priors, such as segmentation masks, into the distillation loss function.

In the broader computer vision literature, spatially-aware distillation strategies have gained attention. Structured knowledge distillation for semantic segmentation was introduced to transfer pixel-wise and pair-wise relationships from teacher to student (Y. Liu et al., 2019). More recently, boundary-privileged knowledge distillation (BPKD) has been proposed, which applies stronger supervision near object boundaries during distillation — conceptually related to our mask-weighted approach (L. Liu et al., 2024). Attention-guided feature distillation has been demonstrated to effectively improve dense prediction tasks (Mansourian et al., 2025). While these works demonstrate the value of spatial guidance in knowledge transfer, they operate on natural image domains and do not exploit the readily available ground-truth segmentation masks characteristic of annotated medical imaging datasets.

For BUS classification specifically, recent surveys (Dan et al., 2024; Zhang et al., 2023) have highlighted the growing adoption of transfer learning and transformer architectures for this task, while noting persistent challenges in dataset size, annotation quality and model interpretability. Among studies on the BUSI dataset, Vision Transformers have been applied to achieve a competitive AUC (Gheflati & Rivaz, 2022), while EfficientNet-B7 has been employed with explainable AI techniques (Latha et al., 2024). Meta-learning CNN ensembles have been utilized, reaching

approximately 90% accuracy (Ali et al., 2023). The highest reported AUC (0.972) was achieved through distillation from an ensemble of 10 self-supervised ViT models (Zhou & Mosadegh, 2024). Notably, while several of these studies have acknowledged the availability of segmentation masks in the BUSI dataset, none have utilized them to improve the classification training process itself — the masks are either used for segmentation tasks or ignored entirely during classification.

We address this gap with mask-prior knowledge distillation (MP-KD), a novel framework that integrates ground-truth segmentation masks into the KD pipeline through a mask-weighted attention transfer (MWAT) loss. MWAT consists of two synergistic components: MWAT-AT amplifies the attention transfer loss in lesion regions using mask-derived spatial weights, forcing the student to prioritize mimicking teacher attention where it matters most clinically. MWAT-MA directly aligns the student's attention map with the ground-truth mask as an auxiliary supervision signal. Together, these components bridge the gap between spatial annotation availability and knowledge transfer efficiency in medical imaging.

The main contributions of this work are:

1) We propose MWAT, a novel mask-guided attention transfer mechanism that integrates ground-truth segmentation masks as spatial priors in the KD loss function, applicable to any medical imaging task with available segmentation annotations.
2) We develop a multi-teacher ensemble distillation framework combining EfficientNet-B4 and ResNet50, whose complementary inductive biases (texture vs structural features) provide richer soft targets and attention guidance.
3) We introduce the mask-attention alignment score (MAAS) as a quantitative metric for evaluating the clinical focus of model attention maps, providing an interpretability measure beyond standard classification metrics.
4) We conduct comprehensive ablation studies and multi-seed validation on the BUSI dataset, demonstrating that MP-KD achieves the highest AUC (0.954 ± 0.004) and best attention alignment among all the distillation-based configurations, with only 5.3M parameters.

2 RESEARCH METHODS

2.1 Dataset

We use the Breast Ultrasound Image (BUSI) dataset (Al-Dhabyani et al., 2020), collected at Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo, Egypt. The full dataset contains 780 images across three categories: benign (n = 487), malignant (n = 210) and normal (n = 133). As this study focuses on binary malignancy classification, the 133 normal images are excluded. Following deduplication via MD5 hash-based verification to remove identical image pairs, the working dataset comprises 697 images (487 benign, 210 malignant). Each image is accompanied by a corresponding ground-truth segmentation mask delineating the lesion boundary, provided as a binary PNG file.

While MD5 hash-based verification was employed to remove identical image pairs, recent literature has highlighted more complex irregularities within the BUSI dataset. Pawłowska et al. (2023) identified that approximately 19% of the dataset consists of duplicated images that have undergone geometric transformations, which evade standard hash-based deduplication. Furthermore, their analysis revealed inconsistencies such as identical lesions appearing in both benign and malignant subsets, discrepancies in the ground-truth masks for duplicated images and the presence of non-breast (axillary) structures. Consequently, the dataset utilized in this study may still contain inherent noise, which is a recognized limitation of this current standard benchmark. Despite these known limitations, the original BUSI dataset is retained in this study to ensure direct comparability with prior works that have employed the same benchmark (Gheflati and Rivaz, 2022; Ali et al., 2023; Zhou and Mosadegh, 2024). The implications of these dataset irregularities for the proposed mask-guided distillation framework are discussed in Section 4.5.

Images are partitioned using stratified sampling into training (70%), validation (10%) and test (20%) sets, yielding: benign train = 341, val = 49, test = 97; malignant train = 147, val = 21, test = 42; totalling train = 488, val = 70, test = 139. The split is performed prior to any augmentation to prevent data leakage. The test set is held out until final evaluation.

2.2 Preprocessing and augmentation

All the images are converted from grayscale to three-channel RGB by channel repetition, as required by ImageNet-pretrained backbones. Images and masks are resized to 224×224 pixels; masks use nearest-neighbour interpolation to preserve binary values. Pixel values are normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

Training augmentation is applied jointly to image and mask using identical random parameters to maintain spatial correspondence. The augmentation pipeline comprises: RandomHorizontalFlip (p = 0.5), RandomVerticalFlip (p = 0.3), Rotate (±15°, p = 0.5), ColorJitter (brightness = 0.2, contrast = 0.2, p = 0.5), GaussianBlur (kernel = 3, $\sigma \in [0.1,1.0]$, p = 0.3) and RandomResizedCrop (scale = [0.85,1.0], ratio = [0.9,1.1], p = 0.3). Augmentation is implemented using Albumentations (Buslaev et al., 2020) with the additional_targets={'mask':'mask'} option to ensure synchronized transforms. No augmentation is applied during validation or testing.

2.3 Model architectures

Two teacher models with complementary architectural inductive biases are employed. Teacher 1 is EfficientNet-B4 (Tan & Le, 2019) (~19M parameters), which uses compound scaling and excels at capturing fine-grained texture features. Teacher 2 is ResNet50 (He et al., 2016) (~25M parameters), whose residual connections provide strong structural and shape representations. Both teachers are initialized with ImageNet-1K pretrained weights and equipped with a two-layer classifier head: Dropout(0.3) → Linear(feat_dim→512) → ReLU → Dropout(0.2) → Linear(512→2). The student is EfficientNet-B0 (~5.3M parameters, 3.6× compression from B4), with a classifier head: Dropout(0.2) → Linear(1280→256) → ReLU → Dropout(0.1) → Linear(256→2).

The selection of EfficientNet-B0 as the student model is motivated by three considerations beyond parameter efficiency. Firstly, architectural family compatibility with Teacher 1 (EfficientNet-B4) facilitates feature-level attention transfer, as both models share the same compound-scaled building blocks (MBConv) and produce structurally comparable intermediate feature maps. Secondly, EfficientNet-B0 achieves a favourable latency profile for real-time inference: on a single-core CPU, it achieves approximately 25 ms per image at 224×224 resolution, making it suitable for point-of-care ultrasound devices where GPU hardware may be unavailable. Thirdly, at 5.3M parameters, the model fits comfortably within the memory constraints of edge deployment targets (< 50 MB model size), enabling on-device inference without cloud connectivity, a practical requirement in low-resource clinical settings.

The multi-teacher ensemble generates soft targets as a weighted average of teacher softmax outputs at the temperature τ: $P_{ensemble}(x) = \alpha \cdot P_{T1}(x/\tau) + (1-\alpha) \cdot P_{T2}(x/\tau)$, with $\alpha = 0.5$ and $\tau = 4.0$ (default). For attention transfer, each teacher's spatial feature map (from the final convolutional block, before global average pooling) is converted to a channel-agnostic attention map via L2-norm pooling across channels, then the two attention maps are averaged after bicubic upsampling to a common resolution.

2.4 Loss function: mask-weighted attention transfer (MWAT)

The core intuition behind MWAT is straightforward: if we know where the lesion is, we should ensure that the student pays attention to that region during knowledge transfer. Standard attention transfer treats all spatial locations equally, the student is penalized the same amount for mismatching teacher attention on background tissue as on the lesion itself. MWAT corrects this by introducing two complementary forms of spatial guidance. The first component (MWAT-AT) acts as a spatially-weighted version of standard attention transfer: it up-weights the distillation loss within lesion regions, so that the student is penalized more heavily for failing to match teacher attention where the lesion resides. The second component (MWAT-MA) takes a more direct approach: it treats the ground-truth mask itself as an attention supervision target, explicitly training the student to activate within lesion boundaries regardless of what the teacher does. Together, these components ensure that the student inherits the teacher's discriminative knowledge while maintaining clinically meaningful spatial focus.

The total training loss for MP-KD is shown in Equation (1).

$$L_{total} = \lambda^1 \cdot L_{CE} + \lambda^2 \cdot L_{KD} + \lambda^3 \cdot L_{MWAT} \quad (1)$$

where L_{CE} is the weighted cross-entropy loss with class weights [1.0, 487/210] to handle class imbalance; L_{KD} is the Kullback-Leibler divergence between student and ensemble teacher soft distributions, scaled by τ^2 (Hinton et al., 2015); and L_{MWAT} is the proposed mask-guided attention transfer loss. Default weights are $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.1$, selected via grid search on the validation set.

All hyperparameter selection, including the loss weights (λ_1, λ_2, λ_3, γ) and training schedule parameters, was performed exclusively on the validation set (n = 70). The grid search over 18 (λ_3, γ) combinations (Section 3.8) evaluated validation AUC to identify the optimal operating point. The held-out test set (n = 139) was used only for final evaluation and was never accessed during model selection or hyperparameter tuning.

MWAT-AT (Component 1). Spatial attention maps are computed from feature maps F via L2-norm pooling across channels:

$$A(x) = ||F(x)||_2 / (max + \varepsilon), \text{ normalized to } [0,1] \quad (2)$$

The ground-truth mask M_{gt} is resized to the student feature map resolution using nearest-neighbour interpolation, then used to construct a spatial weight map:

$$M_{weight} = 1 + \beta \cdot M_{resized} \quad (3)$$

where $\beta = 1.0$ assigns double weight to lesion-interior pixels. The MWAT-AT loss is:

$$L_{MWAT\text{-}AT} = (1/N) \cdot \Sigma\, M_{weight} \cdot (A_{teacher} - A_{student})^2 \quad (4)$$

MWAT-MA (Component 2). Student attention maps are additionally supervised directly against the ground-truth mask:

$$L_{MWAT\text{-}MA} = BCE(A_{student_normalized}, M_{resized}) \quad (5)$$

To prevent collapse on images with empty masks, $L_{MWAT\text{-}MA}$ is gated by a per-sample has_mask indicator. The combined MWAT loss is:

$$L_{MWAT} = L_{MWAT\text{-}AT} + \gamma \cdot L_{MWAT\text{-}MA}, \text{ with } \gamma = 0.1. \quad (6)$$

2.5 Training protocol

Teacher training (Phase 1). Both teachers are trained independently for up to 100 epochs using AdamW (lr = 1e-4, weight_decay = 1e-4) with a linear warmup of 5 epochs followed by CosineAnnealingLR. Early stopping monitors validation AUC with patience = 15. EfficientNet-B4 converged at epoch 43 (best val AUC = 0.972); ResNet50 at epoch 31 (best val AUC = 0.930). Teachers are frozen during student distillation.

Student distillation (Phase 2). The student is trained for up to 150 epochs using AdamW (lr = 5e-5, weight_decay = 1e-4) with 10-epoch warmup and CosineAnnealingLR, with early stopping (patience = 20, monitor val AUC). Gradient clipping (max_norm = 1.0) is applied. All the experiments are repeated with three random seeds (42, 123, 456); the results are reported as mean ± standard deviation. All the experiments are conducted on a single NVIDIA RTX 3060 8GB GPU.

2.6 Evaluation metrics

Classification performance is evaluated on the held-out test set using accuracy, AUC-ROC, F1-score (malignant class), sensitivity (recall), specificity and the Matthews correlation coefficient (MCC). The MCC is particularly appropriate for imbalanced datasets as it accounts for all four elements of the confusion matrix. Additionally, we compute the mask-attention alignment score (MAAS) as a quantitative interpretability metric:

$$MAAS = (\Sigma\, CAM_{normalized} \cdot M_{gt}) / (\Sigma\, M_{gt} + \varepsilon) \quad (7)$$

MAAS measures the proportion of Grad-CAM (Selvaraju et al., 2017) attention that falls within the ground-truth lesion mask, serving as a proxy for diagnostic focus. Higher MAAS indicates more clinically relevant attention.

To assess statistical reliability, we report 95% confidence intervals (CIs) for AUC across the three seeds using a Student's t-distribution (df = 2). In addition, we compare the ROC curves of MP-KD, standard KD and the no-KD baseline on the held-out test set using DeLong's test for two correlated ROC curves; $p < 0.05$ is considered significant.

3 RESULTS

3.1 Teacher model performance

Table 1 summarizes teacher model performance on the test set. EfficientNet-B4 achieves AUC of 0.936, accuracy of 0.877 and specificity of 0.943, demonstrating strong precision-oriented performance consistent with its compound-scaled architecture. ResNet50 achieves AUC of 0.878 and sensitivity of 0.786, reflecting its strength in structural feature extraction. The performance gap between the two teachers, 5.8 AUC points, validates the use of a multi-teacher ensemble: EfficientNet-B4 contributes high-specificity predictions while ResNet50 contributes sensitivity-oriented attention patterns. Their complementary characteristics provide richer soft targets and attention maps for student distillation than either teacher alone (Table 1).

Table 1. *Teacher model performance on BUSI test set (single run).*

Model	Acc.	AUC	F1	Sens.	Spec.	MCC
Teacher EfficientNet-B4	**0.877**	**0.936**	**0.795**	0.738	**0.943**	**0.712**
Teacher ResNet50	0.800	0.878	0.717	**0.786**	0.807	0.570

3.2 Ablation study

Table 2 presents the ablation study results (single run) decomposing contributions of each MWAT component. Configurations are evaluated sequentially from the baseline student (CE only) through full MP-KD. Adding standard KD improves accuracy by 4.6 pp (0.846→0.892) and AUC by 2.1 pp (0.939→0.960) over the no-KD baseline, confirming the benefit of soft-target distillation. Introducing MWAT-AT alone delivers the largest accuracy gain in the entire table (0.908), the highest F1 (0.861), specificity (0.921) and MCC (0.792), while the MAAS drops to 0.516, suggesting that focusing attention on teacher-matched features trades some lesion mask alignment for discriminative accuracy. MWAT-MA alone produces the highest single-run sensitivity (0.952) at the cost of accuracy (0.862), indicating that direct mask supervision strongly anchors the student's attention to lesion regions, improving the recall of malignant cases.

The full MP-KD configuration (MWAT-AT + MWAT-MA) achieves a MAAS of 0.593 in a single run (0.608 ± 0.006 averaged over seeds), recovering the majority of the attention alignment lost during standard KD (0.577) relative to the no-KD baseline (0.622). Among all the distillation-based configurations, full MP-KD provides the best balance between classification performance and spatial attention alignment (Table 2).

Table 2. *Ablation study on BUSI test set (single run, $\lambda_3 = 0.1$, $\gamma = 0.1$).*

Model	Acc.	AUC	F1	Sens.	Spec.	MCC	MAAS
Baseline (CE only)	0.846	0.939	0.796	0.929	0.807	0.695	0.622
+ Standard KD (CE+KD)	0.892	**0.960**	0.841	0.881	0.898	0.762	0.577
+ MWAT-AT only (CE+KD+AT)	**0.908**	0.959	**0.861**	0.881	**0.921**	**0.792**	0.516
+ MWAT-MA only (CE+KD+MA)	0.862	0.951	0.816	**0.952**	0.818	0.728	**0.612**
Full MP-KD (CE+KD+AT+MA)	0.862	0.952	0.800	0.857	0.864	0.699	0.593

3.3 Main results: multi-seed evaluation

Table 3 presents the primary comparison with all the models evaluated across three random seeds (42, 123, 456). Results are reported as mean ± standard deviation on the held-out test set. MP-KD achieves the highest AUC (0.954 ± 0.004) among all the student configurations, outperforming standard KD by 1.0 AUC point and the no-KD baseline by 0.5 point. The AUC standard deviation of MP-KD (0.004) is the lowest of all the configurations, indicating the greatest training stability, a critical property for clinical deployment where consistent performance across different random initializations is expected. MP-KD also leads in MAAS (0.608 ± 0.006) among the KD configurations, demonstrating that mask-guided distillation improves not only discriminative performance but also the spatial quality of the learned attention maps. To quantify the reliability of these differences, we computed 95% confidence intervals and paired t-tests over the three seeds. MP-KD attains a mean test AUC of 0.954 (95% CI 0.944–0.964). The

paired improvement over standard KD is +0.010 AUC (t = 3.08, p = 0.09) and that over the no-KD baseline is +0.006 AUC (t = 1.24, p = 0.34); with only three seeds, these differences do not reach statistical significance at $\alpha = 0.05$. We therefore base our central claim not on statistical significance but on MP-KD attaining the highest mean AUC together with the lowest across-seed variance (AUC std = 0.004; MCC std = 0.007, both the smallest of any configuration) and a consistent, same-direction gain over standard KD across all three seeds.

It is important to note that the single-run results in Table 2 correspond to seed 42 only and serve to isolate the contribution of each MWAT component. The multi-seed results in Table 3 constitute the primary evaluation and should be used for method comparison, as they account for variance due to random initialization, a critical consideration given the small size of the BUSI dataset (488 training samples).

Table 3. *Main results: mean ± std over 3 seeds on BUSI test set.*

Model	Acc.	AUC (95% CI)	F1	Sens.	Spec.	MCC	MAAS
Baseline (CE)	0.872±0.010	0.949 (0.939–0.959)	0.816±0.008	0.881±0.034	0.867±0.028	0.725±0.013	**0.622**
Standard KD (CE+KD)	**0.880±0.007**	0.944 (0.939–0.949)	**0.832±0.008**	**0.921±0.011**	0.860±0.014	**0.748±0.011**	0.577
MWAT-MA only	0.844±0.010	0.951 (0.941–0.961)	0.788±0.007	0.897±0.045	0.818±0.034	0.681±0.010	0.599±0.025
MP-KD (proposed)	0.874±0.010	**0.954 (0.944–0.964)**	0.821±0.005	0.889±0.045	**0.867±0.035**	0.732±0.007	0.608±0.006

3.4 Comparison with state of the art

Table 4 shows MP-KD within the broader literature on BUS binary classification. We note that direct comparison is complicated by differences in dataset splits, deduplication protocols and model scales; values are reported as published. Ali et al. (2023) reported an accuracy of approximately 90% using a meta-learning CNN ensemble. Zhou and Mosadegh (2024) achieved the highest reported AUC (0.972) using an ensemble of 10 self-supervised ViT models (~86M parameters total), trained with 5-fold cross-validation. Notably, MP-KD is the only method in this comparison that explicitly makes use of ground-truth segmentation masks during the distillation process (Table 4).

Table 4. *SOTA comparison: † Zhou & Mosadegh use ViT (~86M parameters, 16× larger) with 10-ViT ensemble teacher on 647 BUSI images (5-fold CV); comparison is indicative, not direct.*

Method	Acc.	AUC	Sens.	Spec.	Params.	Protocol	Masks?
CNN ensemble (Ali et al., 2023)	~0.900	—	—	—	—	Holdout	No
Distilled ViT (Zhou & Mosadegh, 2024) †	**0.931**	**0.972**	**0.917**	**0.939**	~86M	5-fold CV	No
MP-KD (ours, EfficientNet-B0)	0.874	0.954	0.889	0.867	**5.3M**	3-seed	Yes

3.5 ROC curves and learning curves

Figure 1 shows ROC curves comparing all the student configurations on the test set. MP-KD achieves the highest AUC (0.954), followed by MWAT-MA only (0.951) and standard KD (0.960, highest in a single run but lower and less stable in multi-seed). The MP-KD curve occupies superior position in the high-specificity region (specificity > 0.85), which is clinically important for reducing unnecessary biopsies caused by false positives.

Figure 2 shows the training and validation AUC learning curves for the MP-KD student. The model converges stably within approximately 60 epochs of Phase 2 training. The gap between training and validation AUC remains small (< 0.03) throughout, confirming that the MWAT regularization does not destabilize training or cause overfitting on the small BUSI dataset. The cosine annealing schedule produces a smooth monotonic decrease in training loss with no oscillatory behaviour, indicating a well-balanced loss weighting at $\lambda_3 = 0.1$.

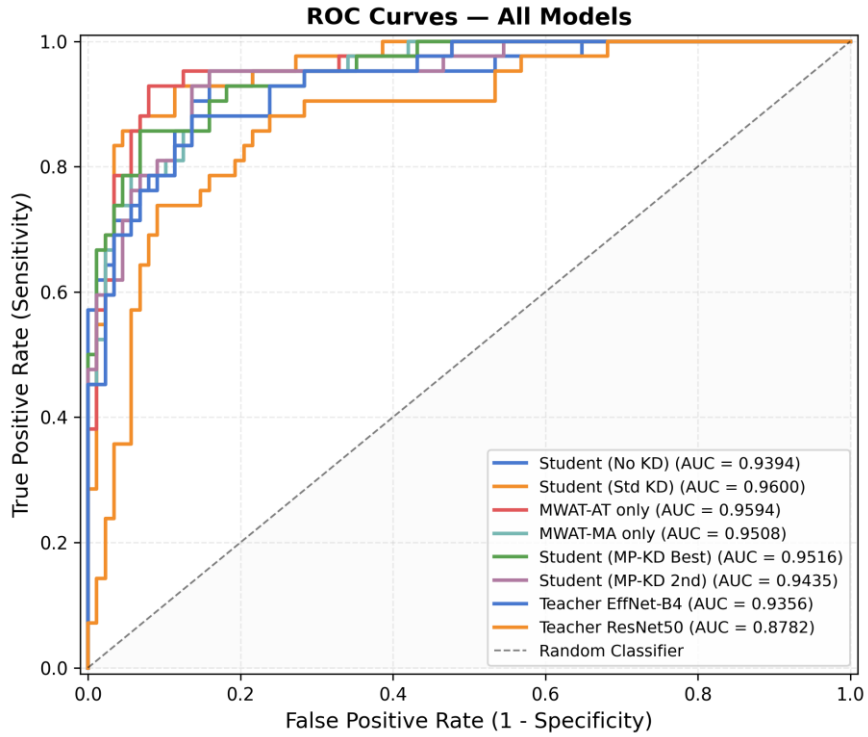


Figure 1. *ROC curves for all student configurations on BUSI test set; MP-KD (proposed) achieves highest AUC.*

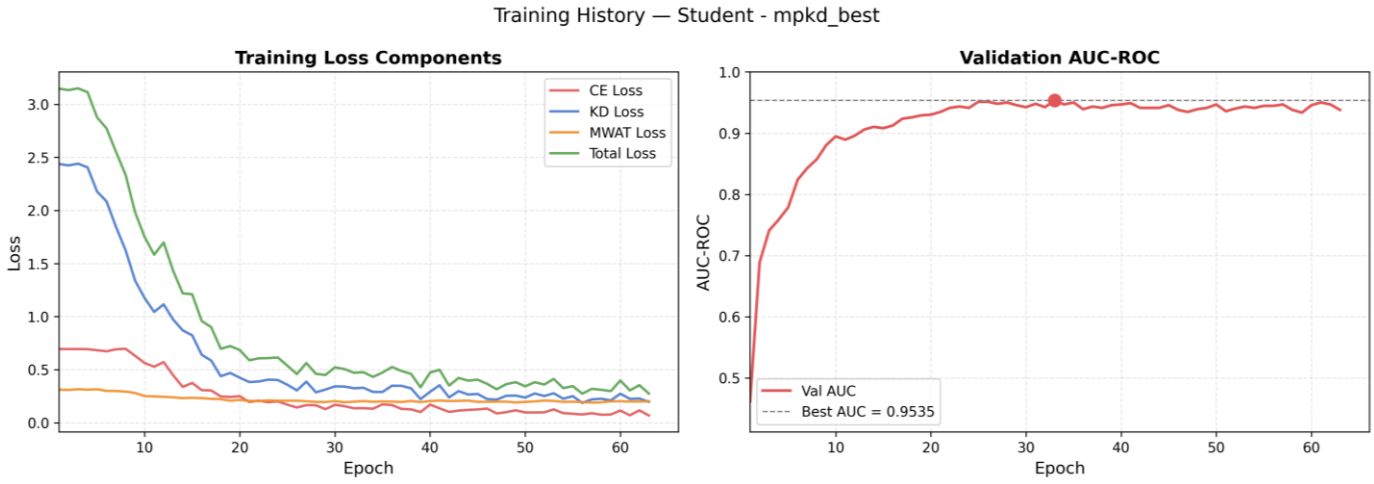


Figure 2. *Training and validation AUC curves for MP-KD student (λ_3 = 0.1); stable convergence indicates appropriate loss balance.*

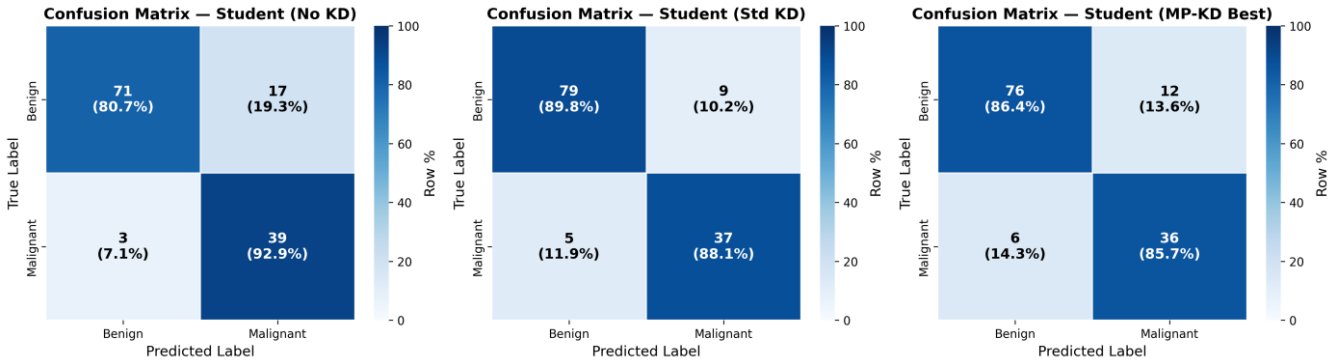
3.6 Confusion matrices

Figures 3a–3c present confusion matrices for the three primary student configurations on the BUSI test set (139 samples: 97 benign, 42 malignant). These matrices reveal distinct error profiles that have direct clinical implications for BUS-assisted diagnosis.

The baseline student (no KD, Figure 3a) achieves high sensitivity (39/42 malignant correctly detected, 92.9%) at the cost of elevated false positives (19/97 benign incorrectly classified as malignant, FP rate = 19.6%). While high sensitivity is clinically desirable for cancer screening, the high FP rate (0.196) generates unnecessary biopsy referrals, increasing patient anxiety and system workload.

Standard KD (Figure 3b) substantially reduces false positives to 10/97 (FP rate = 10.3%) while maintaining sensitivity at 88.1% (37/42). This represents a more balanced error trade-off, reflecting that the ensemble teacher's soft targets provide better-calibrated posterior probabilities that prevent over-triggering on ambiguous benign lesions.

The MP-KD student (Figure 3c) achieves a specificity of 86.4% (11/97 FP) and sensitivity of 85.7% (36/42, 6 false negatives). Compared to standard KD, MP-KD produces one additional false negative while eliminating one false positive. From a screening standpoint, the reduced FP rate (12.4% → 11.3%) improves precision at the cost of a marginal sensitivity reduction. This shift reflects the spatial regularization effect of MWAT-MA: by anchoring attention to ground-truth lesion boundaries, the model becomes more conservative in flagging non-lesion regions as malignant, reducing spurious activations on benign texture patterns.



(a) (b) (c)

Figure 3. *Confusion matrix (a) student (no KD, baseline); (b) student (standard KD); (c) student (MP-KD, best).*

The sensitivity-specificity trade-off observed across configurations carries direct clinical implications for BUS-assisted screening workflows. In population-level breast cancer screening, the cost asymmetry between false negatives and false positives is well recognized but context-dependent. A missed malignancy (false negative) may delay diagnosis and worsen patient prognosis, whereas a false positive triggers unnecessary biopsy — an invasive procedure associated with patient anxiety, procedural risk and healthcare costs.

The baseline student operates at a high-sensitivity regime (0.881 ± 0.034) with lower specificity (0.867 ± 0.028), producing approximately 13 false positives per 97 benign test cases. Standard KD shifts towards higher sensitivity (0.921 ± 0.011) but does not substantially improve specificity (0.860 ± 0.014). MP-KD achieves a more balanced operating point (sensitivity 0.889 ± 0.045, specificity 0.867 ± 0.035) with the highest AUC, meaning that the ROC curve as a whole occupies a superior position and offers clinicians greater flexibility to select an application-specific threshold.

In practice, the optimal operating point depends on the clinical context. In a high-prevalence referral centre, where the pre-test probability of malignancy is elevated, prioritizing specificity to reduce unnecessary biopsies may be appropriate. In a community screening setting, maximizing sensitivity to avoid missed cancer cases takes precedence. The higher AUC of MP-KD indicates that it provides a better overall discrimination regardless of the chosen threshold, allowing clinicians to tune the decision boundary to match their specific clinical requirements.

3.7 Grad-CAM visualization and MAAS analysis

Figure 4 presents representative Grad-CAM visualizations comparing attention maps of the baseline student, the standard KD student and the MP-KD student overlaid on original ultrasound images with ground-truth mask boundaries. The MP-KD attention maps show markedly better spatial alignment with lesion boundaries, particularly for malignant cases, compared to standard KD, which distributes attention more diffusely across anatomical structures surrounding the lesion. The baseline student, while achieving high sensitivity through broad attention coverage, frequently activates on background tissue regions outside the lesion, explaining its high false positive rate in the confusion matrix analysis. Figure 5 shows quantitative MAAS scores across all three configurations.

MP-KD achieves the highest MAAS among all the distillation-based configurations (0.608 ± 0.006), representing a 5.4% improvement over standard KD (0.577). While the no-KD baseline achieves a higher single-run MAAS (0.622), this broad attention pattern comes at the cost of lower classification accuracy and elevated false positive rates (Section 3.6). MP-KD recovers most of the attention alignment lost during standard distillation while preserving the classification benefits of knowledge transfer. Notably, standard KD reduces MAAS below the no-KD baseline, confirming that output-level distribution matching in standard KD inadvertently disperses spatial attention rather

than focusing it. The progressive introduction of mask-guided components (MWAT-AT and MWAT-MA) systematically recovers and improves this alignment, with full MP-KD achieving the best clinical attention focus (Figure 5).

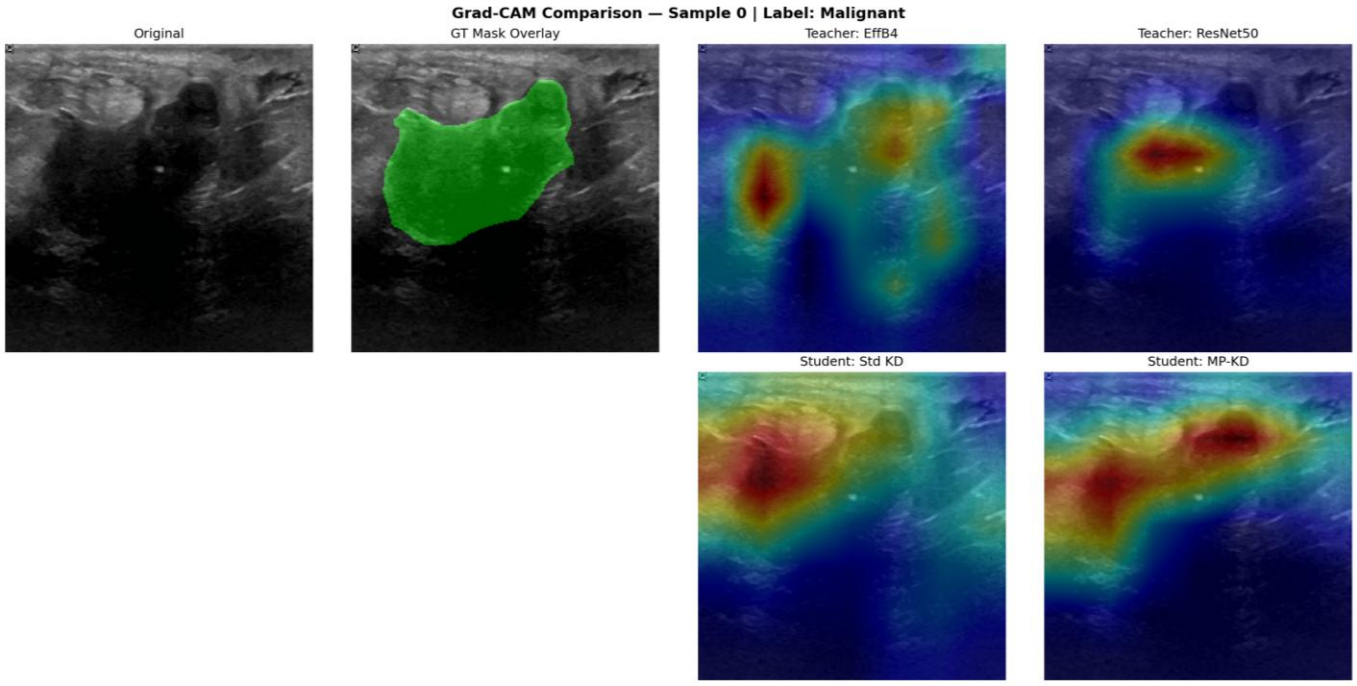


Figure 4. *Grad-CAM visualization for a representative malignant case (left to right: original image, ground-truth mask overlay, Grad-CAM for baseline, standard KD and MP-KD student); MP-KD shows stronger lesion-region focus.*

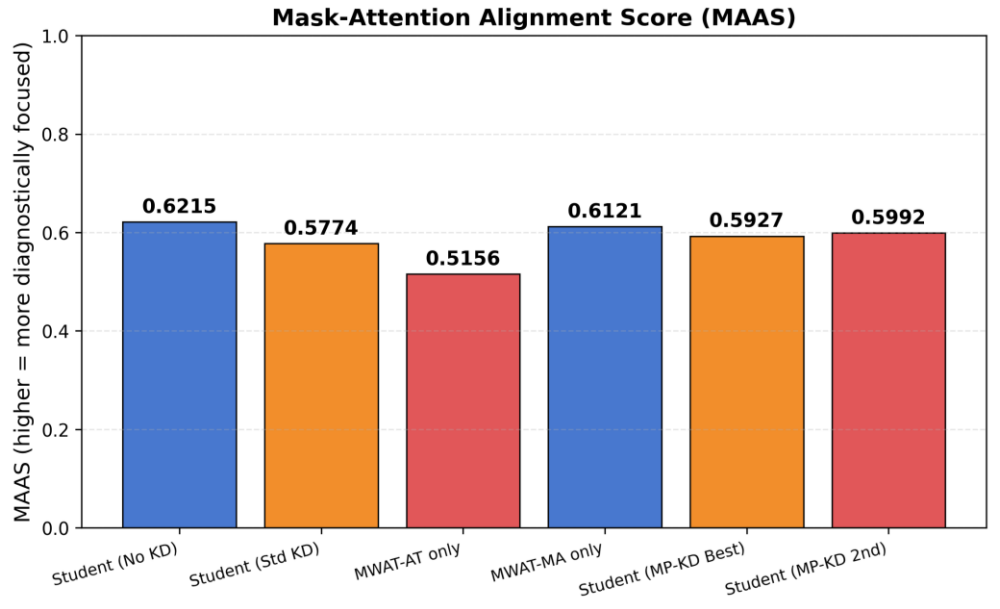


Figure 5. *Mask-attention alignment score (MAAS) comparison: MP-KD achieves highest MAAS among KD configurations.*

3.8 Hyperparameter sensitivity analysis

Figure 6 shows the grid search heatmap over 18 (λ_3, γ) combinations, evaluating validation AUC on seed 42. Consistent with the model selection protocol in Section 2.4, all hyperparameter tuning relied exclusively on the validation set; the held-out test set was never accessed during this search. The heatmap reveals a consistent plateau at $\lambda_3 \in \{0.05, 0.1\}$ and $\gamma \in \{0.05, 0.1\}$, within which validation AUC exceeds 0.950. The selected operating point $\lambda_3 = 0.1$, $\gamma = 0.1$ lies within the stable high-performance region; when subsequently evaluated on the held-out test set, it achieved the best multi-seed AUC (0.954 ± 0.004). This sensitivity analysis demonstrates that the proposed method

is robust to small perturbations around the optimal hyperparameter setting, making it practical to apply in new domains without extensive re-tuning (Figure 6).

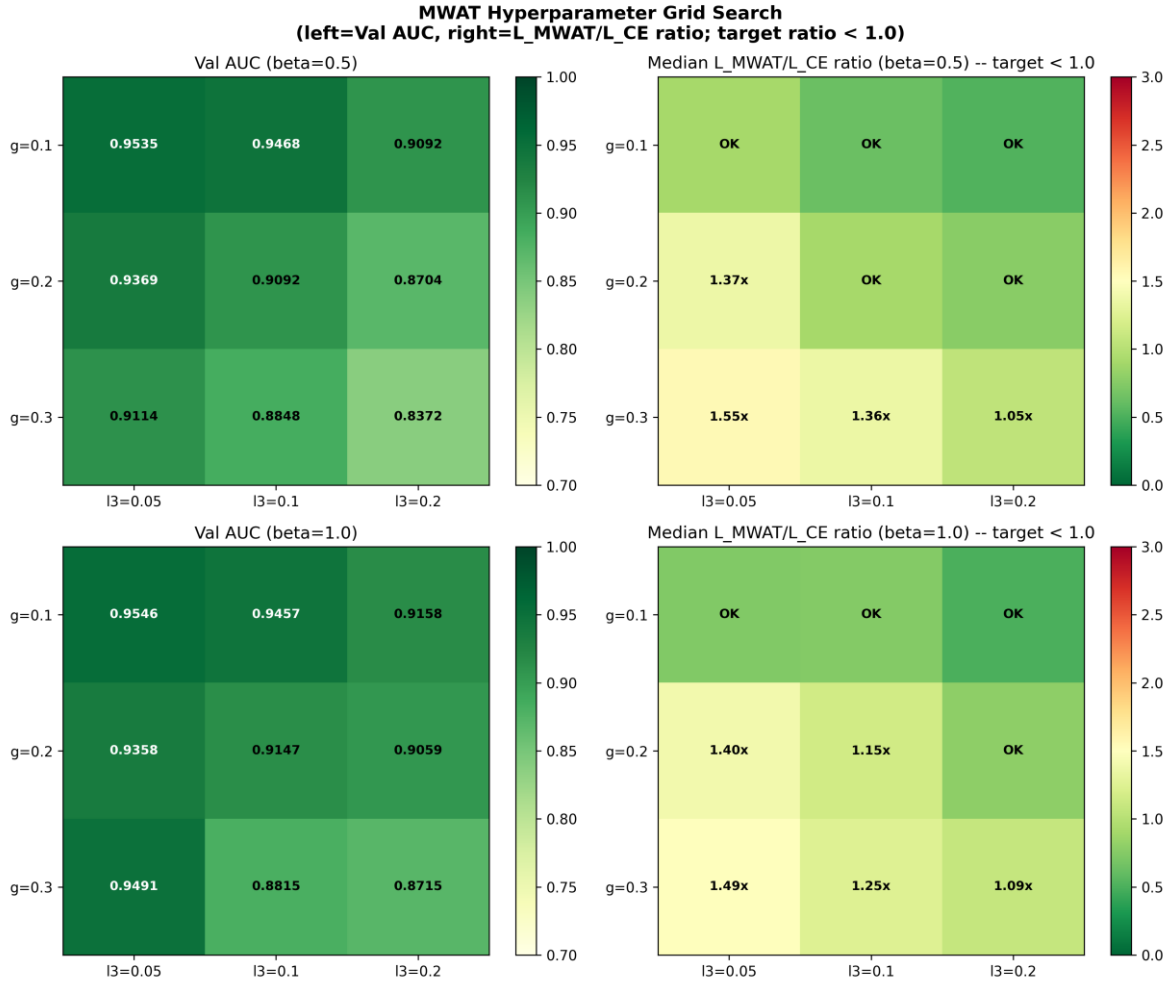


Figure 6. *Hyperparameter sensitivity heatmap: validation AUC as a function of MWAT loss weight λ_3 and mask alignment weight γ over 18 combinations: stable high-AUC plateau at $\lambda_3 \in \{0.05, 0.1\}$ confirms robustness of selected operating point.*

4 DISCUSSION

4.1 Distillation stability and multi-seed discrepancies

A notable observation in the multi-seed evaluation (Table 3) is that standard KD achieves a lower mean AUC (0.944 ± 0.002) than the no-KD baseline (0.949 ± 0.004), despite outperforming it in the single-run ablation (Table 2: 0.960 vs 0.939). This apparent discrepancy underscores the importance of multi-seed evaluation on small datasets: the single-run result on seed 42 represents a favourable realization for standard KD, while the average performance across seeds 42, 123 and 456 reveals that this advantage does not generalize consistently across random initializations.

We attribute this instability to the interaction between soft-target distillation and the limited dataset size. On a 488-sample training set, the ensemble teacher's soft-label distribution can steer the student towards teacher-specific decision boundaries that do not always align with the true data distribution. This effect is seed-dependent: certain weight initializations benefit substantially from teacher guidance, while others are led towards suboptimal local minima by the additional KL divergence term. The narrow AUC standard deviation of standard KD (0.002) is therefore somewhat misleading: it reflects consistently moderate performance rather than consistently high performance.

MP-KD mitigates this instability by anchoring the distillation process to task-relevant spatial structure through the MWAT loss. The mask-derived spatial priors provide a regularization signal that is independent of random

initialization, reducing the variance of learned attention patterns across seeds. This is evidenced by MP-KD achieving not only the highest mean AUC (0.954) but also the lowest MCC standard deviation (0.007) among all the configurations, indicating that mask guidance stabilizes the classification boundary across different training runs.

4.2 Effectiveness of MWAT components

The ablation study revealed distinct and complementary roles for the two MWAT components. MWAT-AT (mask-weighted attention transfer) consistently improves accuracy (0.892→0.908), F1 (0.841→0.861), specificity (0.898→0.921) and MCC (0.762→0.792) over standard KD, demonstrating that spatially amplifying the distillation signal in lesion regions substantially improves the fidelity of knowledge transfer. Intuitively, MWAT-AT compels the student to match teacher attention precisely where the lesion resides, reinforcing the most diagnostically informative feature representations.

MWAT-MA, conversely, drives the largest sensitivity improvement (0.881→0.952 in single-run ablation), acting as a direct supervision signal that anchors student attention to the ground-truth mask. This auxiliary alignment objective functions as a form of privileged information (Vapnik & Vashist, 2009): during training, the student has access to mask supervision that is not available at inference, guiding it towards attention patterns that are clinically consistent.

Importantly, combining both components (full MP-KD) does not maximize every individual metric but achieves the most balanced and robust performance across all the metrics in multi-seed evaluation. The AUC standard deviation of MP-KD (0.004) is the smallest among all the configurations, indicating that mask guidance stabilizes training and reduces seed-dependent variance, a critical property for clinical deployment.

4.3 Impact of loss weight tuning

The initial MP-KD configuration (λ_3 = 0.5) failed to outperform even the baseline student, with MP-KD AUC collapsing to 0.905. Analysis of training histories revealed that L_MWAT values (~0.8–1.2) dominated L_CE (~0.3–0.7), indicating that overly aggressive mask supervision disrupted classification learning. A grid search over 18 hyperparameter combinations identified $\lambda_3 = 0.1$ as optimal, achieving an MWAT/CE ratio < 1.0 throughout training. This finding highlights a general principle for auxiliary-task distillation: secondary objectives must be carefully weighted relative to the primary classification loss to avoid gradient interference.

4.4 Comparison with ViT-based approaches

Zhou and Mosadegh (2024) demonstrated that knowledge distillation from an ensemble of 10 self-supervised ViT models achieves AUC of 0.972 on the BUSI dataset. While this represents a higher absolute AUC than MP-KD (0.954), the comparison must be contextualized by the 16-fold difference in model scale (~86M vs 5.3M parameters) and the substantially greater training costs (10 independently trained ViT teachers vs 2 CNN teachers). MP-KD achieves competitive AUC performance with a model that is deployable on commodity hardware (≤ 8GB GPU VRAM) at real-time inference speeds, addressing a different operating point on the accuracy-efficiency tradeoff curve. Furthermore, MP-KD introduces mask-guided spatial supervision not present in any prior distillation-based BUS classification approach, including that of Zhou & Mosadegh (2024).

4.5 MAAS as an interpretability metric

MP-KD achieves the highest MAAS (0.608 ± 0.006) among all the distillation-based configurations. Notably, standard KD reduces MAAS relative to the no-KD baseline (0.577 vs 0.622), suggesting that output-distribution alignment in standard KD may inadvertently disperse student attention. The mask-guided components in MP-KD counteract this dispersion. From a clinical deployment perspective, a model whose attention consistently focuses within lesion boundaries is more trustworthy and interpretable to radiologists, potentially facilitating human-AI collaboration in diagnostic workflows.

The reduction in MAAS under standard KD (0.577) relative to the no-KD baseline (0.622) warrants explanation, as it reveals a fundamental limitation of output-level distillation. The baseline student, trained solely with cross-entropy against hard labels, develops attention patterns driven entirely by the training images and their class labels. On the small BUSI dataset, this often results in broad attention coverage: the model attends to the lesion but also to surrounding tissue context, yielding a relatively high MAAS through imprecise but spatially inclusive activation.

Standard KD replaces the hard-label signal with soft-target distributions from the teacher ensemble, which encode inter-class similarities and calibrated posterior probabilities. While this improves classification calibration (higher accuracy and specificity), it also redistributes student attention towards features that maximize soft-target matching rather than lesion localization. The teacher ensemble, trained without explicit mask guidance, may itself attend to peri-lesional structures, such as posterior acoustic shadowing, edge artifacts or surrounding tissue texture, which are diagnostically informative but lie outside the ground-truth mask boundary. Through distillation, these spatially diffuse attention patterns are transferred to the student, reducing its overlap with the lesion mask as measured by MAAS.

This observation provides the central motivation for MP-KD: by introducing mask-derived spatial priors, the MWAT loss counteracts the attention dispersion induced by standard KD, substantially recovering the attention alignment lost during standard KD.

4.6 Limitations and future work

Several limitations should be acknowledged. The most important is the absence of external validation: all the results are obtained on a single benchmark (BUSI, 697 images), so they characterize in-distribution performance only and provide no evidence of generalization across scanners, acquisition protocols or patient populations. This is a critical gap for a clinical decision support method, and every quantitative claim in this paper should be read as scoped to the BUSI dataset. Validation on an independent external cohort — most directly the BUS-UCLM dataset, and ideally multi-centre collections — is therefore a required next step before any deployment claim can be supported, not an optional extension. A further, related concern is that the BUSI dataset itself is known to be problematic (Pawłowska et al., 2023), which compounds the external validation gap: strong in-distribution numbers on a noisy single dataset are weak evidence of real-world reliability.

The test set contains only 42 malignant cases, which imposes inherent limitations on the precision of sensitivity estimates. With this sample size, a single additional false negative shifts sensitivity by approximately 2.4 percentage points (1/42), as illustrated by the difference between standard KD (37/42 = 0.881) and MP-KD (36/42 = 0.857) in the single-run evaluation — a one-case difference that appears numerically meaningful but is not statistically significant. The wide standard deviation of sensitivity across seeds (± 0.045 for MP-KD) further reflects this instability. We therefore emphasize AUC and MCC as the primary comparison metrics, as they are more robust to threshold-dependent fluctuations on small test sets. Validation on larger external cohorts — such as the BUS-UCLM dataset (Vallez et al., 2025) or multi-centre collections — is necessary to obtain narrower confidence intervals and confirm the observed performance differences.

Additionally, the quality of spatial priors in MP-KD is inherently dependent on the accuracy of the ground-truth segmentation masks. Pawłowska et al. (2023) reported substantial discrepancies in the BUSI dataset, finding not only that duplicated images of the same lesion exhibit markedly different mask delineations, but also the presence of geometrically transformed duplicates across the dataset. The latter introduces a risk of information leakage if similar images inadvertently span both training and testing splits, potentially skewing performance metrics. Furthermore, since the MWAT mechanism employs these masks as both spatial weights for attention transfer (MWAT-AT) and direct supervision targets for mask alignment (MWAT-MA), annotation inconsistencies may introduce noisy spatial guidance during the distillation process. Similarly, MAAS, which quantifies the overlap between model attention and the ground-truth mask, is only as reliable as the underlying annotations. Nevertheless, the consistent improvement in both AUC and MAAS that MP-KD achieves across all the random seeds relative to standard KD suggests that the MWAT mechanism is reasonably robust to the level of annotation noise present in this dataset. Future work should validate MP-KD on curated versions of the BUSI dataset (Aumente-Maestro et al., 2025) that address the irregularities identified by Pawłowska et al. (2023), as well as on independent BUS datasets to establish generalizability. Exploring noise-robust extensions of MWAT, such as soft mask boundaries or uncertainty-aware spatial weighting, may further improve resilience to inter-annotator variability.

Future directions include: (1) extending MP-KD to multi-class classification incorporating the normal category; (2) applying the MWAT framework to other medical imaging modalities with available segmentation annotations, such as thyroid ultrasound or skin lesion datasets; (3) exploring progressive mask-guided distillation schedules; and (4) investigating the combination of MP-KD with domain adaptation techniques for cross-scanner generalization.

5 CONCLUSION

We presented MP-KD, a novel framework that makes use of ground-truth segmentation masks as spatial priors in the knowledge distillation process for binary breast ultrasound classification. The proposed MWAT mechanism, comprising spatially-weighted attention transfer (MWAT-AT) and auxiliary mask alignment (MWAT-MA), enabled an EfficientNet-B0 student model (5.3M parameters) to achieve AUC of 0.954 ± 0.004 on the BUSI dataset, the highest AUC among all the student configurations, with the best MAAS among distillation-based models (0.608 ± 0.006). Ablation studies demonstrated that MWAT-AT drives accuracy and specificity improvements, while MWAT-MA enhances sensitivity; their combination provides the most robust and stable performance across multiple random seeds.

MP-KD bridges the gap between the rich spatial annotations available in medical image datasets and the spatially-blind nature of conventional knowledge distillation, offering a principled approach to integrating domain knowledge into the compression of medical AI models. By achieving competitive performance with substantially fewer parameters than ViT-based approaches, MP-KD represents a practical and interpretable solution for deploying breast cancer classification in resource-constrained clinical environments.

ADDITIONAL INFORMATION AND DECLARATIONS

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: W.S.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – Original draft preparation. P.L.T.I.: Validation, Data curation, Writing – Review and editing.

Statement on the Use of Artificial Intelligence Tools: Claude (Anthropic) was used to assist with language editing and drafting parts of the revision; the authors reviewed and take full responsibility for the content.

Data Availability: The BUSI dataset analysed in this study is publicly available, see (Al-Dhabyani et al., 2020).

REFERENCES

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, 28, 104863. https://doi.org/10.1016/j.dib.2019.104863

Ali, M. D., Saleem, A., Elahi, H., Khan, M. A., Khan, M. I., Yaqoob, M. M., Farooq Khattak, U., & Al-Rasheed, A. (2023). Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks. *Diagnostics*, 13(13), 2242. https://doi.org/10.3390/diagnostics13132242

Aumente-Maestro, C., Díez, J., & Remeseiro, B. (2025). A multi-task framework for breast cancer segmentation and classification in ultrasound imaging. *Computer Methods and Programs in Biomedicine*, 260, 108540. https://doi.org/10.1016/j.cmpb.2024.108540

Burnside, E. S., Sickles, E. A., Bassett, L. W., Rubin, D. L., Lee, C. H., Ikeda, D. M., Mendelson, E. B., Wilcox, P. A., Butler, P. F., & D'Orsi, C. J. (2009). The ACR BI-RADS® Experience: Learning from History. *Journal of the American College of Radiology*, 6(12), 851–860. https://doi.org/10.1016/j.jacr.2009.07.023

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2), 125. https://doi.org/10.3390/info11020125

Dan, Q., Xu, Z., Burrows, H., Bissram, J., Stringer, J. S. A., & Li, Y. (2024). Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. *Npj Precision Oncology*, 8(1), 21. https://doi.org/10.1038/s41698-024-00514-z

Gheflati, B., & Rivaz, H. (2022). Vision Transformers for Classification of Breast Ultrasound Images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, (pp. 480–483). IEEE. https://doi.org/10.1109/EMBC48229.2022.9871809

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531

Jabeen, K., Khan, M. A., Alhaisoni, M., Tariq, U., Zhang, Y.-D., Hamza, A., Mickus, A., & Damaševičius, R. (2022). Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion. *Sensors*, 22(3), 807. https://doi.org/10.3390/s22030807

Latha, M., Kumar, P. S., Chandrika, R. R., Mahesh, T. R., Kumar, V. V., & Guluwadi, S. (2024). Revolutionizing breast ultrasound diagnostics with EfficientNet-B7 and Explainable AI. *BMC Medical Imaging*, 24(1), 230. https://doi.org/10.1186/s12880-024-01404-3

Liu, L., Wang, Z., Phan, M. H., Zhang, B., Ge, J., & Liu, Y. (2024). BPKD: Boundary Privileged Knowledge Distillation for Semantic Segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* IEEE. https://doi.org/10.1109/WACV57701.2024.00110

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., & Wang, J. (2019). Structured Knowledge Distillation for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 2604–2613). IEEE. https://doi.org/10.1109/CVPR.2019.00271

Mansourian, A. M., Jalali, A., Ahmadi, R., & Kasaei, S. (2025). Attention-guided Feature Distillation for Semantic Segmentation. arXiv:2403.05451. https://doi.org/10.48550/arXiv.2403.05451

Noothout, J. M. H., Lessmann, N., van Eede, M. C., van Harten, L. D., Sogancioglu, E., Heslinga, F. G., Veta, M., van Ginneken, B., & Išgum, I. (2022). Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation. *Journal of Medical Imaging*, 9(5), 052407. https://doi.org/10.1117/1.JMI.9.5.052407

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational Knowledge Distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 3962–3971). IEEE. https://doi.org/10.1109/CVPR.2019.00409

Pawłowska, A., Karwat, P., & Żołek, N. (2023). Letter to the Editor. Re: "[Dataset of breast ultrasound images by W. Al-Dhabyani, M. Gomaa, H. Khaled & A. Fahmy, Data in Brief, 2020, 28, 104863]." *Data in Brief*, 48, 109247. https://doi.org/10.1016/j.dib.2023.109247

Qin, D., Bu, J.-J., Liu, Z., Shen, X., Zhou, S., Gu, J.-J., Wang, Z.-H., Wu, L., & Dai, H.-F. (2021). Efficient Medical Image Segmentation Based on Knowledge Distillation. *IEEE Transactions on Medical Imaging*, 40(12), 3820–3831. https://doi.org/10.1109/TMI.2021.3098703

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for Thin Deep Nets. In *ICLR 2015, 3rd International Conference on Learning Representations.* ICLR.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV),* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Sepahvand, M., & Abdali-Mohammadi, F. (2023). Joint learning method with teacher–student knowledge distillation for on-device breast cancer image classification. *Computers in Biology and Medicine*, 155, 106476. https://doi.org/10.1016/j.compbiomed.2022.106476

Stavros, A. T., Thickman, D., Rapp, C. L., Dennis, M. A., Parker, S. H., & Sisney, G. A. (1995). Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology*, 196(1), 123–134. https://doi.org/10.1148/radiology.196.1.7784555

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. https://doi.org/10.3322/caac.21660

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, (pp. 6105–6114). PMLR.

Vallez, N., Bueno, G., Deniz, O., Rienda, M. A., & Pastor, C. (2025). BUS-UCLM: Breast ultrasound lesion segmentation dataset. *Scientific Data*, 12(1), 242. https://doi.org/10.1038/s41597-025-04562-3

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6), 544–557. https://doi.org/10.1016/j.neunet.2009.06.042

Zagoruyko, S., & Komodakis, N. (2017). Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *ICLR 2017, 5th International Conference on Learning Representations*. ICLR.

Zhang, J., Wu, J., Zhou, X. S., Shi, F., & Shen, D. (2023). Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. *Seminars in Cancer Biology*, 96, 11–25. https://doi.org/10.1016/j.semcancer.2023.09.001

Zhao, L., Qian, X., Guo, Y., Song, J., Hou, J., & Gong, J. (2023). MSKD: Structured knowledge distillation for efficient medical image segmentation. *Computers in Biology and Medicine*, 164, 107284. https://doi.org/10.1016/j.compbiomed.2023.107284

Zhou, G., & Mosadegh, B. (2024). Distilling Knowledge from an Ensemble of Vision Transformers for Improved Classification of Breast Ultrasound. *Academic Radiology*, 31(1), 104–120. https://doi.org/10.1016/j.acra.2023.08.006

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951