JURNAL INFOTEL
VOL. 18, NO. 3, AUGUST 2026, PP. 472–489. DOI:10.20895/INFOTEL.V18I3.10773

RESEARCH ARTICLE

# Knowledge Distillation from ResNet-50 to MobileNetV3-Small for Breast Cancer Classification in Ultrasound Images

Windra Swastika[1,*], Bita Parga Zen[2], and Ratna Dwi Christyanti[3]

[1,2]Faculty of Technology and Design, Universitas Ma Chung, Malang 65151, Indonesia
[3]Faculty of Mathematics, Universitas Kaltara, Tanjung Selor Hilir 77217, Indonesia

*Corresponding email: windra.swastika@machung.ac.id

*Received: March 04, 2026; Revised: August 14, 2026; Accepted: September 02, 2026.*

**Abstract:** Early detection of breast cancer is essential for reducing mortality, but deploying large deep learning models in resource-constrained clinical settings remains challenging. This study evaluates knowledge distillation (KD) from a ResNet-50 teacher to a MobileNetV3-Small student for classifying breast ultrasound images as benign, malignant, or normal, compared with the same student trained without KD. We partitioned the Breast Ultrasound Images Dataset (BUSI), containing 780 images from 600 patients, using a fixed, stratified 70/15/15 image-level split. Unavailable patient identifiers prevented patient-level splitting, constituting a study limitation. We applied Total Variation Chambolle denoising before training. Grid search selected temperature $T = 2$ and weighting coefficient $\alpha = 0.5$, achieving 88.89% validation accuracy. Across five random seeds, the KD student achieved 82.05% $\pm$ 1.60% test accuracy and 80.93% $\pm$ 2.04% macro F1-score, compared with 82.74% $\pm$ 1.64% and 81.88% $\pm$ 2.17% for the no-KD student, and 83.59% $\pm$ 4.25% and 82.41% $\pm$ 5.62% for the teacher. Equivalence testing supported KD–teacher equivalence within a $\pm$5-percentage-point margin (p=0.0498). Paired Wilcoxon tests found no evidence of improvement over the no-KD baseline (p=0.625) or denoising-free training (p=0.4375). The student had 15.5 times fewer parameters and approximately 2.5 times faster inference than the teacher. These findings support compact models for breast ultrasound classification but provide no evidence that KD outperforms simple fine-tuning of the same lightweight architecture on this dataset.

**Keywords:** breast cancer, knowledge distillation, MobileNetV3, model compression, ultrasound image classification

© Jurnal Infotel ISSN: 2085-3688; e-ISSN: 2460-0997

# 1 Introduction

Breast cancer is one of the leading causes of cancer-related mortality among women globally. According to global cancer statistics, approximately 2.3 million women were diagnosed with breast cancer and 670,000 deaths were recorded in 2022, making it the most commonly diagnosed cancer in women across 157 countries [1]. A key challenge is that survival outcomes are strongly linked to the stage at detection; early-stage diagnosis can substantially reduce mortality rates [2]. Recent global burden analyses further show a rising incidence of early-onset breast cancer, which increases the demand for accessible screening capacity outside tertiary referral centers [3]. Medical imaging, particularly ultrasound, plays a central role in screening and diagnosis because it is non-ionizing, widely accessible, and capable of real-time imaging [4].

The advent of deep learning has significantly advanced automated classification of breast ultrasound images. Convolutional Neural Networks (CNNs) trained with transfer learning from large-scale datasets have demonstrated competitive performance on this task [5, 6]. Models such as ResNet [7], VGG, and InceptionV3 have been extensively evaluated on publicly available breast ultrasound datasets, including the widely used Breast Ultrasound Images Dataset (BUSI) [8]. More recent work has moved beyond plain backbone fine-tuning: lesion-region perception modules have been introduced to focus the classifier on tumor boundaries [9], multi-task frameworks jointly optimize segmentation and classification to exploit mask supervision [10], and transfer-learning ensembles have been benchmarked across multiple breast imaging modalities [11]. Transformer-based architectures have likewise been surveyed extensively for medical imaging [12]. However, their parameter counts are typically larger still than the CNNs they replace, which sharpens rather than resolves the deployment problem addressed in this study.

However, a persistent challenge in clinical deployment is the gap between research-grade performance and real-world applicability. Large CNN architectures such as ResNet-50 carry tens of millions of parameters and require substantial memory and computational resources, making them impractical for deployment on edge devices, mobile hardware, or point-of-care ultrasound systems in low-resource settings. This constraint is not hypothetical: recent work targeting low- and middle-income healthcare settings reports that unmodified deep classifiers are unusable on the hardware available in rural facilities, and evaluates compression as a precondition for deployment rather than an optimization [13]. Recent work has therefore explored lightweight architectures, particularly the MobileNet family [14], as drop-in replacements that trade marginal accuracy for significant efficiency gains. MobileNetV3-Small, designed via hardware-aware neural architecture search, achieves competitive classification performance with only approximately $1.5$ million parameters [14].

Knowledge Distillation (KD), first systematically formulated by Hinton et al. [15], addresses this gap by training a compact student model to mimic not only hard-class predictions but also the soft probability distributions of a larger teacher model. This soft supervision carries richer inter-class relationship information than one-hot labels alone and is commonly described as a regulariser for the student. KD has since been extended along several axes, including feature-level distillation [16] and attention transfer [17, 18]; see for a comprehensive taxonomy of these variants. In medical imaging, KD has been applied to segmentation tasks [19], image classification under limited data [20], and multi-modal and fusion-based breast ultrasound models [21]. However, the reported benefits in these

studies are not uniform. Behboodi et al. [22], working on breast ultrasound segmentation, found that KD's advantage over directly training the compact network depends strongly on the teacher–student capacity gap and dataset size, and that naive distillation can fail to improve on the student trained alone. Comparable caution appears in mammographic mass classification, where KD required an additional architecture-search stage before the distilled model outperformed its non-distilled counterpart [23]. This heterogeneity motivates the explicit no-KD baseline adopted in the present study.

Despite this growing body of work, the specific application of KD from a ResNet-50 teacher to a MobileNetV3-Small student for three-class breast ultrasound classification, evaluated against a matched non-distilled baseline with multi-seed statistical validation and efficiency benchmarking on the BUSI dataset, remains underexplored. Existing studies on lightweight models for breast ultrasound primarily report direct fine-tuning of MobileNet variants [24] without isolating the contribution of the knowledge transfer mechanism itself. When KD is applied, researchers often omit the non-distilled student of the same architecture, so the reported gains cannot be separated from the effects of the lightweight architecture and ImageNet pre-training. Furthermore, researchers rarely report statistical reliability across multiple experimental runs, limiting the reproducibility and generalisability of published findings.

This study addresses these gaps with the following contributions: (1) a structured KD pipeline from ResNet-50 to MobileNetV3-Small for three-class breast ultrasound classification, evaluated end-to-end on BUSI; (2) a controlled three-arm comparison, teacher, MobileNetV3-Small trained without KD, and MobileNetV3-Small trained with KD, under identical data partitions, training schedules, and evaluation protocol, which isolates the contribution of knowledge transfer from that of the architecture itself; (3) an ablation of the Total Variation Chambolle denoising stage under the same protocol; (4) a multi-seed evaluation reporting effect sizes, per-class recall and specificity with Wilson score confidence intervals, and a two one-sided test (TOST) with a pre-defined $\pm$5-percentage-point equivalence margin, rather than inferring equivalence from a non-significant difference test; and (5) a quantitative efficiency analysis of the resulting compression and inference speedup. The results support the compression claim while providing no evidence that distillation outperforms plain fine-tuning of the same lightweight architecture on this dataset, a negative finding we report explicitly.

## 2 Research Method

Figure 1 illustrates the overall research pipeline. The workflow begins with the BUSI dataset, which is preprocessed and then partitioned into training, validation, and test subsets using a single fixed split reused across every run reported in this paper. The teacher model is then trained across five seeds, and the resulting seed-0 checkpoint is used to conduct the $(T, \alpha)$ grid search, whose selected pair is applied unchanged to all subsequent student runs. We train three student arms under identical data, schedules, and evaluation conditions: a MobileNetV3-Small baseline fine-tuned without a teacher, a KD student distilled with the selected hyperparameters, and a KD student trained on images that bypass the denoising stage. All four models, the teacher and the three students, are evaluated on the same held-out test set, and the analysis spans multi-seed classification performance,

per-class sensitivity and specificity, paired significance and equivalence testing, and model efficiency.

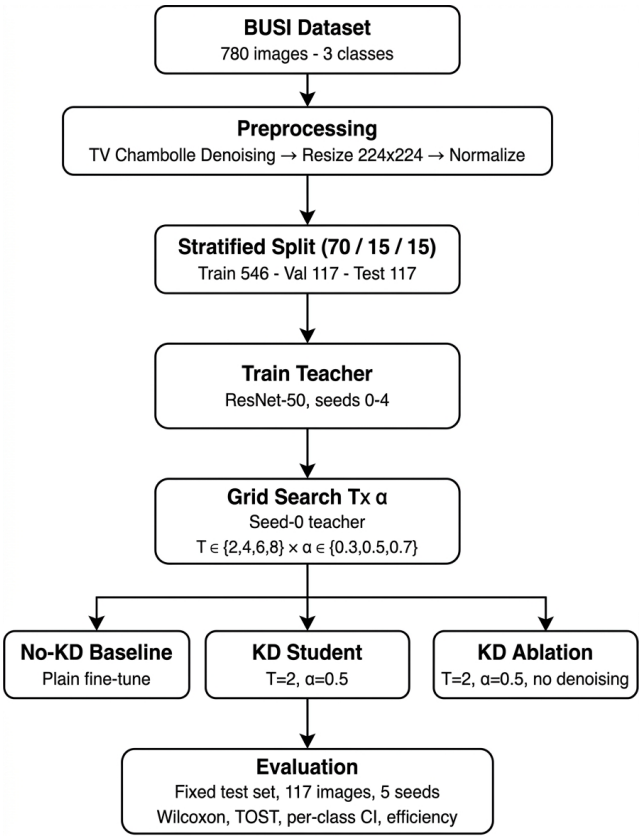


Figure 1: Research pipeline.

## 2.1 Dataset

This study uses the Breast Ultrasound Images Dataset (BUSI), collected by Al-Dhabyani et al. [8] at Baheya Hospital, Cairo, Egypt, in 2018. The dataset comprises 780 ultrasound images from 600 female patients aged between 25 and 75 years. Images are categorized into three classes: benign (437 images, 56%), malignant (210 images, 26.9%), and normal (133 images, 17.1%). Each image includes a corresponding binary segmentation mask delineating the lesion region. Images are in PNG format with an average size of approximately $500 \times 500$ pixels.

We partitioned the dataset using a stratified split with a 70%:15%:15% ratio for training, validation, and test subsets, respectively, to preserve class proportions across all subsets, yielding 546 training, 117 validation, and 117 test images. We generated the partition once with a fixed random seed ($\texttt{seed} = 42$) and held it identical across every experimental run reported in this paper; no run re-partitions the data. Consequently, the test set is common

to all five evaluation seeds and all three model arms, making the arms directly paired and ensuring seed-to-seed dispersion reflects initialization and shuffling variability only, not partition variability (subsection 2.7).

We perform the split at the image level. The public BUSI release does not distribute patient identifiers, case identifiers, or acquisition metadata that would allow grouping images by patient; the 780 images are provided as flat, per-class directories with sequential filenames that do not encode patient membership. A patient-wise or otherwise group-aware split therefore could not be constructed from the released data. Because the dataset contains 780 images from 600 patients, some patients contribute more than one image, and related images from the same patient may fall into different partitions. This is a recognized source of optimistic bias in medical imaging benchmarks, and the absolute performance figures reported here should be read with that caveat; we restate it as a limitation in section 4. All three model arms are affected identically, so this issue does not confound the between-arm comparisons that form the study's primary contribution.

## 2.2 Preprocessing

We selected the teacher and student architectures to reflect a realistic compression scenario rather than a maximally favorable one. We chose ResNet-50 as the teacher because it is the most widely reported backbone on BUSI-class breast ultrasound benchmarks, making the teacher's performance directly comparable to published baselines, and because its residual formulation is well characterized under transfer learning on small medical datasets. We chose MobileNetV3-Small as the student because it is the smallest member of the MobileNet family that retains ImageNet pre-trained weights in standard frameworks, giving the largest compression ratio available without abandoning transfer learning. It was designed under explicit latency constraints for mobile inference [14], the deployment target motivating this study. The resulting teacher-to-student parameter ratio of approximately $15.5\times$ is deliberately aggressive; a narrower gap would make equivalence easier to achieve but would test the compression claim less severely.

Ultrasound images are inherently affected by speckle noise, a multiplicative noise pattern arising from coherent wave interference, which can degrade feature extraction performance in CNNs [25]. To mitigate this, we applied Total Variation (TV) denoising using the Chambolle algorithm [26] to all images before training. We selected the TV formulation because it suppresses high-frequency speckle while penalizing deviations from piecewise-smooth structure, preserving the anatomical edges and lesion boundaries that carry the diagnostic signal in this task. We fixed the regularisation weight at $0.1$, chosen by qualitative inspection of denoised samples across all three classes (Figure 2) as the value that visibly attenuates speckle texture without oversmoothing lesion margins. We did not include this weight in the hyperparameter grid search in subsection 2.5 and was not tuned against validation performance; it was held constant across every model arm and every evaluation seed, so it cannot account for any between-arm difference reported in section 3. We therefore do not characterize the sensitivity of the results to this choice, and we note it as a limitation in section 4. We pre-cached denoised images to disk to avoid repeated computation during training.

Following denoising, all images were resized to $224 \times 224$ pixels and normalized using ImageNet statistics ($\texttt{mean} = [0.485, 0.456, 0.406], \texttt{std} = [0.229, 0.224, 0.225]$) to align input distributions with pretrained model weights. We applied online data augmentation

exclusively to the training set, including random horizontal flip ($p = 0.5$), random rotation ($\pm 15°$), random color jitter (brightness and contrast $\pm 20\%$), and random resized crop with a scale factor of $0.9 - 1.0$. We applied no augmentation to the validation or test sets.

Figure 2 presents representative samples from each class before and after applying TV Chambolle denoising. The original images (left column) exhibit characteristic speckle patterns that obscure fine structural boundaries. After denoising (right column), the noise texture is substantially suppressed while anatomical edges and lesion boundaries remain clearly preserved. This qualitative comparison confirms that the denoising step enhances image clarity without introducing artificial smoothing that could erase diagnostically relevant features.

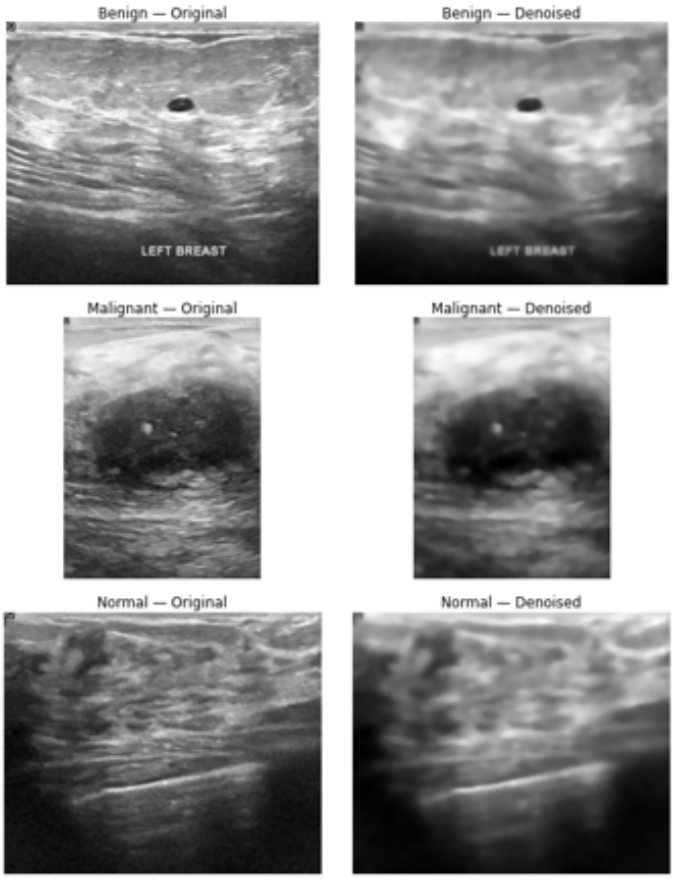


Figure 2: Sample images before and after denoising.

## 2.3 Model Architectures

ResNet-50 [7] is a 50-layer residual network with approximately $23.51$ million parameters. We loaded ImageNet-pretrained weights and replaced the final fully connected layer with

a 3-class linear classifier. We used a progressive unfreezing strategy: during the first 10 epochs, we trained only the classification head; from epoch 11 onward, we unfrozen all layers for full fine-tuning (student Model: MobileNetV3-Small). MobileNetV3-Small [14] is a lightweight architecture with approximately $1.52$ million parameters, designed for mobile deployment. It employs depthwise separable convolutions, Squeeze-and-Excitation (SE) blocks, and Hard-Swish activations to achieve high efficiency. We used pretrained ImageNet weights and replaced the final classifier layer with a 3-class linear layer identical in structure to the teacher head.

## 2.4 Knowledge Distillation

The KD framework follows Hinton et al. [15] and combines two loss components. The first is the standard Cross-Entropy loss between student logits and ground-truth hard labels (Eq. (1)):

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log\left(\hat{p}_c^s\right) \tag{1}$$

The second is the Kullback-Leibler Divergence loss between the temperature-scaled soft probability distributions of the teacher and student (Eq. (2)):

$$\mathcal{L}_{\text{KD}} = T^2 \cdot \text{KL}(p^T_{\text{teacher}} \parallel p^T_{\text{student}}) \tag{2}$$

where the soft probability at temperature $T$ is defined as (Eq. (3)):

$$p_c^T = \frac{\exp(z_c/T)}{\sum_j \exp(z_j/T)} \tag{3}$$

the combined loss is defined as follows (Eq. (4)):

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{CE} + (1-\alpha) \cdot \mathcal{L}_{KD} \tag{4}$$

## 2.5 Hyperparameter Optimization via Grid Search

We optimized the temperature $T$ and weighting coefficient $\alpha$ via grid search. The search ranges were centered on values commonly used in the KD literature ($T$ typically 3-5 in Hinton et al. [15]; $\alpha$ typically weighting hard and soft losses comparably) and extended to cover both lower and higher temperatures given the small size and class overlap of the BUSI dataset. The grid search was performed once, using the seed-0 teacher checkpoint, and the selected $(T, \alpha)$ pair was then reused for all five evaluation seeds; it was not re-run per seed.

Teacher checkpoints were handled seed-wise: for each evaluation seed $s \in \{0, \ldots, 4\}$, we trained the teacher with seed $s$, retained the checkpoint with the lowest validation loss under early stopping, and distilled the KD student for seed s from that seed-matched teacher. The grid search is the sole exception; it was executed once against the seed-0 teacher checkpoint, before the remaining teachers were trained, and its selected $(T, \alpha)$ pair was then applied unchanged to all five seeds. At every stage, we selected models using validation loss or validation accuracy; we did not access the test partition during teacher training, student training, grid search, or checkpoint selection, and evaluated it exactly once per arm per seed.

We used the validation set (never the test set) for hyperparameter selection at every step. The search covered $T \in \{2, 4, 6, 8\}$ and $\alpha \in \{0.3, 0.5, 0.7\}$, yielding 12 combinations. We trained each combination for 20 epochs on the training set and recorded validation accuracy. We selected the combination with the highest validation accuracy for full training.

### 2.6 Training Configuration

Both teacher and student models were trained with the Adam optimizer, a learning rate of $10^{-4}$, and weight decay of $10^{-5}$. We used a ReduceLROnPlateau scheduler (`factor` $= 0.5$, `patience` $= 5$) to reduce the learning rate adaptively. Training was limited to 100 epochs with early stopping (`patience` $= 15$) based on validation loss. Batch size was set to 16. All experiments were implemented using `PyTorch 2.10.0` and executed on `Apple Silicon` via the `MPS` backend.

### 2.7 Evaluation Model

To ensure statistical reliability, we repeated all experiments across five independent random seeds $(0, 1, 2, 3, 4)$, which vary model weight initialization and training-data shuffling order; the train/validation/test partition remains fixed across all seeds (Section 2.1). We report results as mean $\pm$ standard deviation across seeds. Metrics include accuracy, macro-averaged precision, recall, and F1-score, together with per-class recall (sensitivity) and specificity with Wilson score confidence intervals, pooled over the five seeds. Efficiency metrics include parameter count, model file size, and inference latency measured on 100 single-image forward passes (first 10 discarded as warm-up).

We assessed the statistical significance of the performance difference between teacher and student using a paired Wilcoxon signed-rank test at $\alpha = 0.05$, applied to F1-score distributions across 5 seeds, along with a matched-pairs rank-biserial effect size. A non-significant Wilcoxon result indicates only that no difference was detected, not that the two conditions are equivalent; we also performed a two one-sided test (TOST) for equivalence, with a pre-defined equivalence margin of $\pm 5$ percentage points on macro F1, to formally test whether the KD student and teacher are practically equivalent. The same paired Wilcoxon procedure was used to compare the KD student against (i) the no-KD baseline, to isolate the contribution of knowledge transfer, and (ii) a KD student trained without denoising, to isolate the contribution of the TV-Chambolle preprocessing step. We compared dispersion across seeds using Levene's test.

## 3 Results

### 3.1 Grid Search Results

Figure 3 visualizes the validation accuracy across all 12 hyperparameter combinations in the grid search (single run, seed-0 teacher; see Section 2.5). Darker cells indicate higher validation accuracy. The best configuration, $T = 2$ and $\alpha = 0.5$, tied with $T = 4$ and $\alpha = 0.7$, achieves 88.89% validation accuracy. No consistent trend is observed along the temperature axis: accuracy does not increase monotonically with $T$, and both the lowest $(T = 2)$ and highest $(T = 8)$ temperatures produce mid-range combinations. This suggests that, on this relatively small dataset, the specific $(T, \alpha)$ pair matters less than the presence

of soft-label supervision itself, and that any single $T$ should not be over-interpreted as generally preferable without re-running the search across seeds.

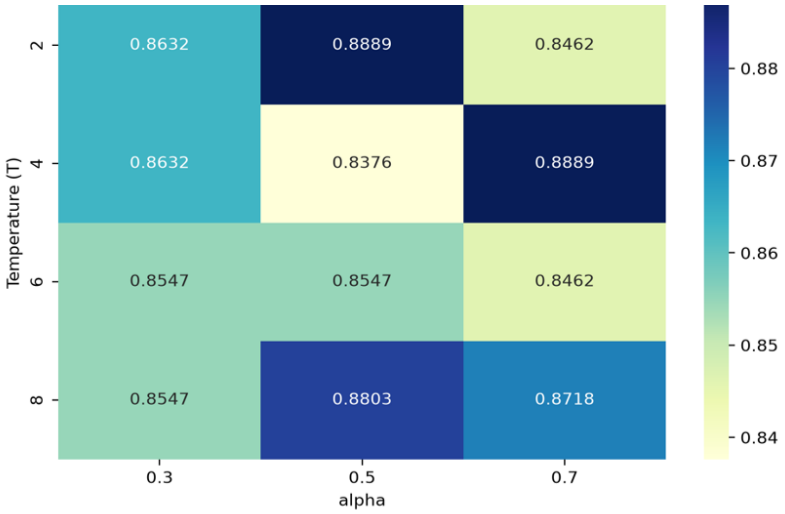


Figure 3: Grid search heatmap.

## 3.2 Multi-Seed Classification Performance

Figure 4 depicts the training dynamics of all teacher and student KD models. Subfigures (a) & (b) show the validation loss and accuracy trajectories for the teacher, and subfigures (c) & (d) show the validation loss and accuracy for the KD student, for one representative seed; per-seed early-stopping epoch and final validation accuracy varied across the five runs, so this figure should be read as illustrative of typical convergence behavior rather than an average. Models converge stably without signs of severe overfitting, with early stopping triggered within the first 30–60 epochs depending on the seed. The student model's validation curves generally show smoother convergence than the teacher's. Still, Table 1 shows this smoother training-time behavior does not translate into higher test-set performance for the KD student, so it should not be read as evidence that the KD loss improves the final model. Table 1 summarizes the classification performance of the teacher and student models, evaluated across 5 independent seeds on the held-out test set.

Table 1: Multi-Seed classification performance (mean $\pm$ std, 5 seeds)

| **Metric** | **Teacher (ResNet-50)** | **Student, no-KD (baseline)** | **Student KD (MobileNetV3-Sm)** |
|---|---|---|---|
| Accuracy | $83.59\% \pm 4.25\%$ | $82.74\% \pm 1.64\%$ | $82.05\% \pm 1.60\%$ |
| Precision | $82.90\% \pm 2.93\%$ | $82.46\% \pm 2.38\%$ | $82.42\% \pm 1.61\%$ |
| Recall | $83.07\% \pm 7.37\%$ | $81.78\% \pm 1.94\%$ | $80.10\% \pm 2.46\%$ |
| F1-Score | $82.41\% \pm 5.62\%$ | $81.88\% \pm 2.17\%$ | $80.93\% \pm 2.04\%$ |

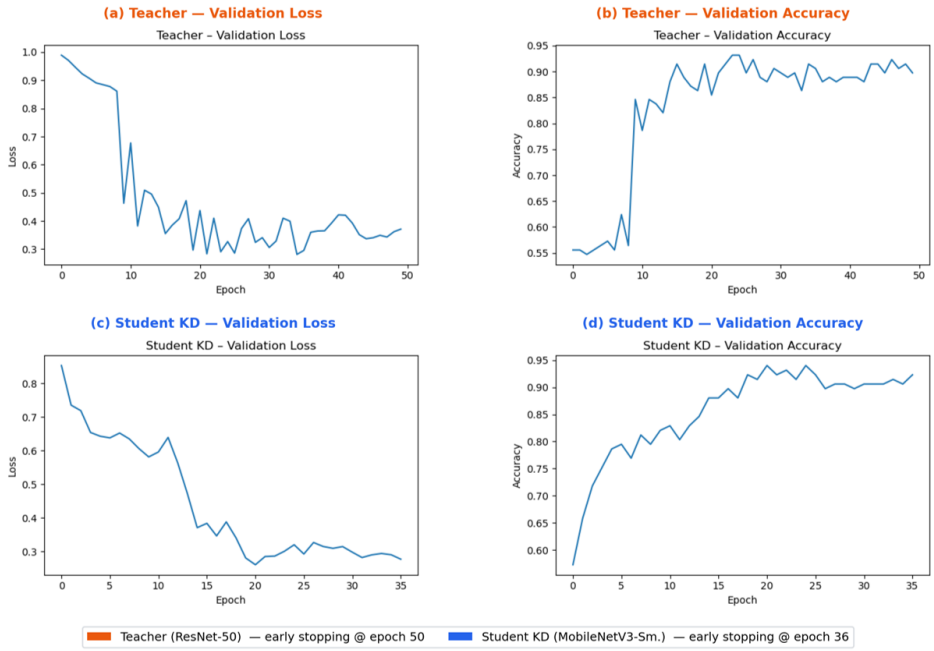


Figure 4: Training curves.

The KD student's mean accuracy $(82.05\%)$ and F1-score $(80.93\%)$ are $1.5 - 1.9$ points lower than the teacher's ($83.59\%$ accuracy, $82.41\%$ F1). They are also marginally lower than the no-KD student baseline ($82.74\%$ accuracy, $81.88\%$ F1). Table 1 provides no evidence that knowledge distillation improves on plain fine-tuning of the same architecture on this dataset (see Section 3.4 for the corresponding significance test). The teacher's standard deviations are nonetheless substantially higher than any student variant's across every metric (e.g., F1 std $5.62\%$ for the teacher vs $2.04\% - 2.17\%$ for the student arms), numerically consistent with a stabilizing effect of the lightweight architecture and/or soft-label supervision, though this difference in dispersion was not statistically significant (`Levene's test`, $p = 0.163$; Section 3.4).

For full transparency and to support independent reproduction, Table 2 reports the per-seed macro F1-score underlying the means in Table 1, together with the corresponding model checkpoint for each run. All four arms share the identical fixed train/validation/test partition; only model initialization and training-data shuffling order vary across seeds. The project repository (Supplementary Data) provides complete per-seed accuracy, precision, and recall for all four arms.

The per-seed values in Table 2 show no seed in which the KD student's F1-score exceeds both the teacher and the no-KD baseline, consistent with the aggregate comparison in subsection 3.4. Figure 5 presents confusion matrices pooled across all five evaluation seeds (585 predictions per arm: 117 test images $\times$ 5 seeds) for the teacher (a), the no-KD student baseline (b), and the KD student (c); pooling avoids the single-run figures no longer matching the multi-seed numbers in Table 1. Each cell reports the pooled raw count. All three models perform strongly on the benign and normal classes, as reflected in the high diagonal values.

Table 2: Per-Seed F1-Score and selected checkpoints

| Seed | Teacher | Student, no-KD | Student KD | Student, KD (no-denoise) |
|---|---|---|---|---|
| 0 | 81.10% | 83.86% | 80.53% | 79.78% |
| 1 | 82.44% | 81.24% | 81.55% | 79.92% |
| 2 | 73.80% | 79.41% | 77.59% | 80.37% |
| 3 | 86.49% | 80.47% | 82.70% | 78.72% |
| 4 | 88.21% | 84.41% | 82.36% | 81.78% |

The malignant class exhibits the lowest recall in all three models, attributable to its relatively lower representation in the dataset $(26.9\%)$ despite stratified splitting. Malignant-to-normal misclassification, the clinically highest-risk error type, is infrequent across all arms but not absent: pooled over 5 seeds, it occurs in $1/155$ malignant cases for the KD student $(0.6\%)$, $2/155$ for the teacher $(1.3\%)$, and $6/155$ for the no-KD student baseline $(3.9\%)$. The KD student's lower rate on this specific error type is a modest positive signal. However, with $n = 155$ pooled malignant cases per arm, this comparison was not formally tested for significance and should be read as descriptive.

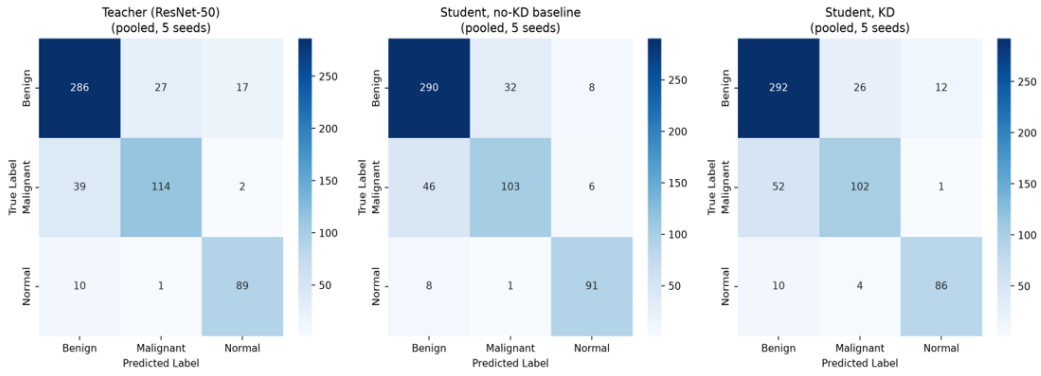


Figure 5: Confusion matrix, pooled over 5 seeds, (a) Teacher (ResNet-50), (b) Student without KD, (c) Student with KD.

## 3.3 Per-Class Performance

Table 3 reports per-class recall (sensitivity) and specificity for all three arms, computed from the pooled confusion matrices of Figure 5 (585 predictions per arm) with Wilson score 95% confidence intervals. Because the intervals are computed on predictions pooled across five seeds that share a single test partition, the observations are not fully independent, so read the intervals descriptively rather than as strict frequentist coverage statements.

Malignant recall is the weakest metric in every arm and is the clinically decisive one, since a missed malignancy is the highest-cost error in this task. The teacher attains the highest malignant recall $(73.5\%)$, ahead of the no-KD baseline $(66.5\%)$ and the KD student $(65.8\%)$; the confidence intervals of the two student arms overlap almost completely, while

the teacher's interval overlaps both only partially. Distillation therefore does not recover the teacher's sensitivity on the malignant class, and the compressed models should not be presented as clinically preferable based on aggregate accuracy alone. The pattern is inverted for benign recall, where both students slightly exceed the teacher. For benign specificity, where the KD student is lowest $(75.7\%)$, that is, the students absorb part of the malignant deficit as over-prediction of the benign class. Malignant specificity is high and effectively identical across arms $(92.3 - -93.5\%)$, confirming that the deficit is one of sensitivity, not of false alarms.

Table 3: Per-class Recall and Specificity, pooled over 5 seeds, 95% Wilson CI

| Class | Metric | Teacher (ResNet-50) | Student no-KD | Student KD |
|---|---|---|---|---|
| Benign ($n = 330$) | Recall | $86.7\%[82.6, 89.9]$ | $87.9\%[83.9, 91.0]$ | $88.5\%[84.6, 91.5]$ |
| Benign | Specificity | $80.8\%[75.5, 85.1]$ | $78.8\%[73.4, 83.4]$ | $75.7\%[70.1, 80.5]$ |
| Malignant ($n = 155$) | Recall | $73.5\%[66.1, 79.9]$ | $66.5\%[58.7, 73.4]$ | $65.8\%[58.0, 72.8]$ |
| Malignant | Specificity | $93.5\%[90.7, 95.5]$ | $92.3\%[89.4, 94.5]$ | $93.0\%[90.2, 95.1]$ |
| Normal ($n = 100$) | Recall | $89.0\%[81.4, 93.7]$ | $91.0\%[83.8, 95.2]$ | $86.0\%[77.9, 91.5]$ |
| Normal | Specificity | $96.1\%[94.0, 97.5]$ | $97.1\%[95.2, 98.3]$ | $97.3\%[95.5, 98.4]$ |

## 3.4 Statistical Significance

The paired Wilcoxon signed-rank test comparing per-seed F1-scores between teacher and KD student yielded a test statistic of $4.0$ and a p-value of $0.4375$ (`effect size` $r = 0.42$). As $p > 0.05$, no statistically significant difference was detected; this indicates no detected difference, not equivalence, and with only five paired seeds, the test is underpowered (minimum achievable $p = 0.0625$). To formally test equivalence, we performed a two one-sided tests (TOST) with a pre-defined margin of $\pm 5$ percentage points on F1; this confirmed statistical equivalence between teacher and KD student ($p = 0.0498 < 0.05$, mean difference $1.48$ percentage points in the teacher's favor).

Given that the TOST p-value lies close to the decision threshold with only five paired observations, we also report the corresponding two-sided 90% confidence interval for the teacher-minus-KD-student F1 difference: $[-2.04, 4.99]$ percentage points, which lies within the pre-defined $\pm 5$-percentage-point margin. Equivalence is therefore concluded based on the interval rather than on the p-value alone, and remains contingent on the chosen margin.

The same paired Wilcoxon procedure applied to KD student vs the no-KD baseline yielded statistic $= 5.0$, $p = 0.625$, $r = 0.30$, again no significant difference, with the point estimate favoring the no-KD baseline rather than KD (Table 1). This directly tests whether knowledge distillation contributes beyond plain fine-tuning, and it does not support that claim on this dataset. The denoising ablation (KD student with vs without TV-Chambolle preprocessing) likewise showed no significant difference (statistic $= 4.0$, $p = 0.4375$, $r = 0.42$), though the point estimate favors denoising (F1 $80.93\%$ with denoising vs $80.11\%$ without).

### 3.5 Model Efficiency Comparison

Table 4 presents the efficiency comparison between teacher and student models. Figure 6 summarizes the efficiency comparison between the teacher and student models. Across all three efficiency dimensions- parameter count, model file size, and inference latency- the student model achieves substantial reductions. The parameter count drops from $23.51$M to $1.52$M $(15.5\times)$, model size from approximately 90MB to 6MB ~$15\times$), and inference latency from $9.27$ms to $3.71$ms $(2.50\times)$ on the same Apple Silicon MPS backend used throughout this study. We measured these figures once (100 forward passes, first 10 discarded as warm-up) and report them as approximate; we did not repeat them across the five evaluation seeds. Section 3.4 shows the accompanying accuracy/F1 differences are not statistically significant between teacher and KD student, so the parameter and latency reductions come without a significant accuracy cost, though Section 3.4 also shows the KD student does not measurably outperform a plain fine-tuned student of the same size, so the efficiency gain should be attributed to the MobileNetV3-Small architecture choice, not specifically to the distillation procedure.

Table 4: Model efficiency comparison

| Metric | Teacher (ResNet-50) | Student KD (MobileNetV3-Sm.) | Ratio |
|---|---|---|---|
| Parameters | 23.51M | 1.52M | $15.5\times$ smaller |
| Model Size | ~90MB | ~6MB | ~$15\times$ smaller |
| Inference Time | 9.27ms | 3.71ms | $2.50\times$ faster |

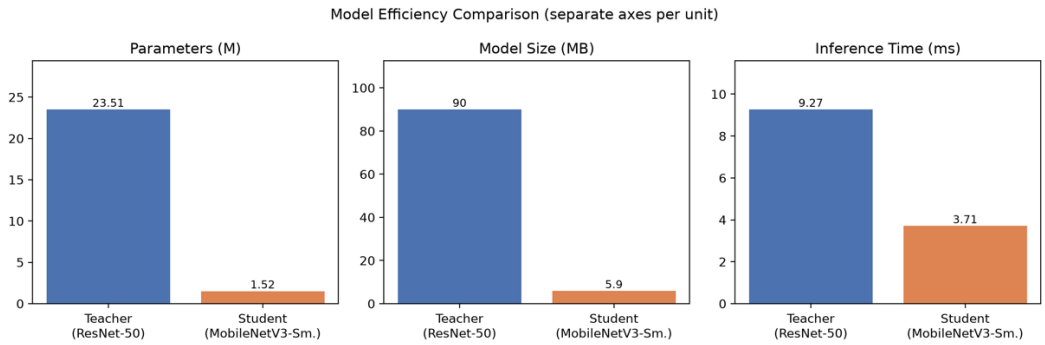


Figure 6: Efficiency bar chart.

## 4 Discussion

The experimental results show that the proposed KD framework compresses a ResNet-50 teacher model into a MobileNetV3-Small student with performance statistically equivalent to the teacher (TOST $p = 0.0498$, $\pm 5$ percentage point margin), rather than exceeding it: the KD student achieves $82.05\%$ mean accuracy and $80.93\%$ mean macro F1-score, compared to $83.59\%$ and $82.41\%$ for the teacher, with the teacher numerically ahead on every metric,

though the difference is not statistically significant (Wilcoxon $p = 0.4375$). Critically, the KD student also does not outperform a plain fine-tuned MobileNetV3-Small trained without a teacher ($82.74\%$ accuracy, $81.88\%$ F1; Wilcoxon $p = 0.625$, point estimate favoring the no-KD baseline). Taken together, these results support compression, a much smaller model with performance indistinguishable from the teacher, but do not support a claim that knowledge distillation specifically, as opposed to fine-tuning the same lightweight architecture directly, is responsible for that performance.

A numerically notable pattern is the lower seed-to-seed performance variance of every student variant relative to the teacher (F1 std: $2.04\%$ for the KD student, $2.17\%$ for the no-KD baseline, vs $5.62\%$ for the teacher). Because the test partition is identical across all five seeds (subsection 2.1), this dispersion reflects sensitivity to model initialization and training-data shuffling rather than partition variability. Levene's test did not find this difference in dispersion to be statistically significant ($p = 0.163$), so it is reported here as a descriptive observation rather than as an established effect; it is also present in the no-KD baseline at a similar magnitude to the KD student, so it is more plausibly attributable to the MobileNetV3-Small architecture itself rather than to the KD loss specifically. Prior work has reported that distillation from CNN teachers can aid generalization on limited medical image data [20], but that study distills into a hybrid transformer student lacking the convolutional priors MobileNetV3-Small already possesses; the present results neither replicate that benefit nor isolate it from architecture choice.

The best configuration identified via grid search, $T = 2$ and $\alpha = 0.5$ (tied with $T = 4$, $\alpha = 0.7$), falls within the range of common defaults in the KD literature ($T = 3$ to $T = 5$) rather than at an extreme. The full grid (Figure 3) shows no consistent trend of increasing validation accuracy with temperature; both the lowest and highest temperatures in the search space produced mid-range results, which suggests that, on a dataset as small as BUSI (780 images), the exact $(T, \alpha)$ choice within a reasonable range matters less than whether soft-label supervision is present at all. We ran this grid search once using a single seed's teacher checkpoint (Section 2.5); a per-seed search might select different combinations, and the result here should not be generalized as evidence for a specific optimal temperature for this task.

The efficiency gains are substantial: $15.5\times$ fewer parameters and approximately $2.5\times$ faster inference on the Apple Silicon MPS backend used in this study. This compression profile is a reasonable starting point for future deployment on mobile devices or point-of-care ultrasound systems, particularly for settings where computational infrastructure is limited, but latency was not measured on representative mobile or ultrasound-platform hardware and the evaluation is limited to one public dataset; claims of clinical deployability should be read as motivation for follow-up hardware-specific and externally validated studies rather than as an established result. A roughly 6 MB model can be loaded into mobile memory with substantially less overhead than the approximately 90 MB teacher model.

Regarding per-class performance (pooled over 5 seeds, with Wilson score 95% confidence intervals), malignant recall is lowest across every arm: $73.5\%[66.1\%, 79.9\%]$ for the teacher vs $66.5\%[58.7\%, 73.4\%]$ for the no-KD baseline and $65.8\%[58.0\%, 72.8\%]$ for the KD student, the teacher has the highest malignant recall of all three arms, which should temper any claim that the compressed student is clinically preferable. This gap is plausibly attributable to malignant's under-representation in the dataset $(26.9\%)$ despite stratified splitting. Future work should investigate class-weighted loss formulations for both the

teacher and any distilled student, specifically designed to improve malignant recall, as false negatives in this category carry the highest clinical risk.

This study has several limitations. First, the evaluation is limited to a single publicly available dataset with 780 images. External validation on independent datasets is necessary to confirm generalizability. Second, although multi-seed evaluation provides a more robust performance estimate than single-seed reporting, the small number of seeds ($n = 5$) limits the statistical power of the Wilcoxon test; the minimum achievable p-value with five pairs is $0.0625$, so several of the non-significant comparisons reported here (subsection 3.4) should be read as underpowered rather than as confirmed absences of difference. Third, the public BUSI release provides no patient identifier, so a patient-wise split could not be performed, and the reported image-level stratified split cannot rule out related images from the same patient appearing in different partitions (subsection 2.1); this is a known risk for optimistic bias in medical imaging benchmarks and should be resolved with a dataset that provides patient-level metadata. Fourth, the hyperparameter grid search (subsection 2.1) was run once on a single seed's teacher rather than independently per evaluation seed. Fifth, this study includes a no-KD baseline and a denoising ablation, neither of which shows a statistically significant benefit over its respective alternative; larger-scale or multi-dataset studies would be needed to determine whether this null result is specific to BUSI's size and class balance or generalizes further. Sixth, the TV-Chambolle regularisation weight was fixed at $0.1$ on qualitative grounds rather than tuned (subsection 2.2), and the denoising ablation reported in subsection 3.4 compares only the presence versus absence of the filter at that single setting; a weaker or stronger filter might yield a different outcome, and the null result for denoising should not be read as evidence that TV denoising is unhelpful in general on ultrasound data.

# 5 Conclusion

This study developed and evaluated a Knowledge Distillation (KD) framework to compress a ResNet-50 teacher model into a lightweight MobileNetV3-Small student model for three-class breast ultrasound classification, and directly compared it against a plain fine-tuned MobileNetV3-Small trained without KD. The experimental results across five random seeds show that the KD student model achieved a mean accuracy of $82.05\% \pm 1.60\%$ and a macro F1-score of $80.93\% \pm 2.04\%$, statistically equivalent to the teacher within a $\pm$5-percentage-point margin as confirmed by a TOST equivalence test ($p = 0.0498$); a paired Wilcoxon signed-rank test found no significant difference from the teacher ($p = 0.4375$).

The primary contribution of this work is demonstrating that a $15.5\times$ smaller, approximately $2.5\times$ faster MobileNetV3-Small model (1.52M vs 23.51M parameters; $\sim$ 9.27ms vs $\sim$ 3.71ms inference on the MPS backend used here) can match the teacher's classification performance within a pre-defined equivalence margin. However, a direct comparison against a plain fine-tuned MobileNetV3-Small baseline (no teacher) found no statistically significant benefit of knowledge distillation over fine-tuning alone (Wilcoxon $p = 0.625$), so the efficiency gains reported here should be attributed primarily to the choice of a lightweight architecture rather than the distillation procedure itself. Given that inference latency was measured on desktop-class Apple Silicon rather than mobile or point-of-care hardware, and validation is limited to one public dataset without confirmed patient-level

splitting, deployment on mobile devices or point-of-care ultrasound systems remains a plausible future direction rather than an established outcome of this study.

## Acknowledgments

The authors thank Universitas Ma Chung for the institutional support provided during this research. The authors also acknowledge Al-Dhabyani et al. and Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo, for making the Breast Ultrasound Images Dataset (BUSI) publicly available.

## References

[1] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *British Journal of Radiology*, vol. 95, no. 1130, p. 20211033, 2022.

[2] A. N. Giaquinto, H. Sung, K. D. Miller, J. L. Kramer, L. A. Newman, A. Minihan, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, no. 6, pp. 524–541, 2022.

[3] J. Zhao, L. Xu, J. Sun, M. Song, L. Wang, S. Yuan, Y. Zhu, Z. Wan, S. Larsson, K. Tsilidis, M. Dunlop, H. Campbell, I. Rudan, P. Song, E. Theodoratou, K. Ding, and X. Li, "Global trends in incidence, death, burden and risk factors of early-onset cancer from 1990 to 2019," *BMJ Oncology*, vol. 2, no. 1, p. e000049, 2023.

[4] S. Zhang, M. Liao, J. Wang, Y. Zhu, Y. Zhang, J. Zhang, R. Zheng, L. Lv, D. Zhu, H. Chen, and W. Wang, "Fully automatic tumor segmentation of breast ultrasound images with deep learning," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 1, p. e13863, 2023.

[5] X. Yu, J. Wang, Q.-Q. Hong, R. Teku, S.-H. Wang, and Y.-D. Zhang, "Transfer learning for medical images analyses: A survey," *Neurocomputing*, vol. 489, pp. 230–254, 2022.

[6] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, June 2016.

[8] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

[9] J. Wei, H. Zhang, and J. Xie, "A novel deep learning model for breast tumor ultrasound image classification with lesion region perception," *Current Oncology*, vol. 31, no. 9, pp. 5057–5079, 2024.

[10] C. Aumente-Maestro, J. Díez, and B. Remeseiro, "A multi-task framework for breast cancer segmentation and classification in ultrasound imaging," *Computer Methods and Programs in Biomedicine*, vol. 260, p. 108540, 2025.

[11] F. Twum, C. C. E. Ahiable, S. O. Oppong, L. Banning, and K. Owusu-Agyemang, "Employing transfer learning for breast cancer detection using deep learning models," *PLOS Digital Health*, vol. 4, no. 6, p. e0000907, 2025.

[12] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

[13] F. Modu, R. Prasad, and F. Aliyu, "Development of a knowledge-distillation-based breast cancer classifier for lmics: Comparison with pruning and quantization," *Electronics*, vol. 14, no. 24, p. 4842, 2025.

[14] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, IEEE, Oct. 2019.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, Mar. 2015.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations (ICLR)*, 2017.

[18] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[19] L. Zhao, X. Qian, Y. Guo, J. Song, J. Hou, and J. Gong, "Mskd: Structured knowledge distillation for efficient medical image segmentation," *Computers in Biology and Medicine*, vol. 164, p. 107284, 2023.

[20] O. S. EL-Assiouti, G. Hamed, D. Khattab, and H. M. Ebied, "Hdkd: Hybrid data-efficient knowledge distillation network for medical image classification," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109430, 2024.

[21] M. Alotaibi, A. Aljouie, N. Alluhaidan, W. Qureshi, H. Almatar, R. Alduhayan, B. Alsomaie, and A. Almazroa, "Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images," *Heliyon*, vol. 9, no. 11, p. e22406, 2023.

[22] B. Behboodi, R. Brooks, and H. Rivaz, "Optimizing knowledge distillation for efficient breast ultrasound image segmentation: Insights and performance enhancement," *Artificial Intelligence in Health*, vol. 2, no. 2, pp. 73–86, 2024.

[23] W.-C. Yeh, W.-C. Shia, Y.-T. Hsu, C.-H. Huang, and Y.-S. Lee, "A lightweight breast cancer mass classification model utilizing simplified swarm optimization and knowledge distillation," *Bioengineering*, vol. 12, no. 6, p. 640, 2025.

[24] A. Ciobotaru, M. A. Bota, D. I. Goța, and L. C. Miclea, "Multi-instance classification of breast tumor ultrasound images using convolutional neural networks and transfer learning," *Bioengineering*, vol. 10, no. 12, p. 1419, 2023.

[25] J. Liu, C. Li, L. Liu, H. Chen, H. Han, B. Zhang, and Q. Zhang, "Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network," *Biomedical Signal Processing and Control*, vol. 86, p. 105150, 2023.

[26] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1–2, pp. 89–97, 2004.