Knowledge Distillation from Medical Vision Foundation Models to Lightweight Networks for Edge Deployment: A Comparative Study

Shuwei Liu
College of Biomedical Engineering, Fudan University, 220 Handan Road, Shanghai 200433, China
E-mail: liusw23@m.fudan.edu.cn

Keywords: Knowledge distillation, medical foundation models, edge deployment, model compression, quantization, medical image classification

Received: June 19, 2026

Medical vision foundation models pretrained on large-scale domain-specific data achieve strong clinical performance, but their computational cost precludes deployment on resource-constrained edge devices. We investigate whether the domain-specific representations of a medical foundation model transfer more effectively through knowledge distillation (KD) than generic ImageNet-pretrained features. We compare a domain-specific foundation model (RETFound, ViT-Large, 307M parameters) against a general-purpose pretrained model (ConvNeXt-Base, 89M) as KD teachers for two lightweight students (MobileNetV3-Small, 2.5M; EfficientNet-Lite0, 4.7M) on two clinical benchmarks – HAM10000 (dermoscopy, 7 classes) and APTOS-2019 (diabetic retinopathy, 5 classes) – over a grid of distillation temperature and loss weight, with INT8 quantization and Grad-CAM analysis. Because RETFound is pretrained on retinal images, APTOS-2019 is in-domain for it whereas HAM10000 (dermoscopy) constitutes an out-of-domain test. On balanced accuracy, the appropriate metric for these class-imbalanced datasets, ConvNeXt-Base yields the stronger student in all four settings (e.g., HAM10000 MobileNetV3: 0.826 vs. 0.763) – including on APTOS-2019, where RETFound's retinal pretraining is in-domain – despite fewer parameters and no domain-specific pretraining; RETFound distillation can even fall below the no-distillation baseline. RETFound requires higher temperatures for effective transfer, whereas ConvNeXt is robust across a broad range. INT8 quantization preserves accuracy within 0.3% while compressing MobileNetV3-Small to 4.2 MB. These results indicate that a teacher's fine-tuned quality, rather than domain-specific pretraining, predicts distillation success, and offer practical guidance for edge deployment of medical image classifiers.

Povzetek:

1 Introduction

Deep learning has demonstrated transformative potential across medical imaging tasks, including dermatological lesion classification [8], diabetic retinopathy screening [11], and pathological diagnosis [5]. However, a persistent gap separates the GPU-equipped servers where models are developed from the resource-constrained edge devices where clinical inference must occur. Point-of-care diagnostics in primary-care clinics, mobile screening in underserved regions, and embedded systems in portable imaging devices demand models that deliver reliable predictions within strict latency and memory budgets [21, 28].

Concurrently, medical vision foundation models have established a new performance frontier. RETFound [32], pretrained on over 900 000 retinal images via a masked autoencoder objective [12], and UNI [5], pretrained on more than 100 million pathology tiles, show that domain-specific self-supervised pretraining yields representations that substantially outperform ImageNet-pretrained counterparts on in-domain tasks. Such models capture fine-grained, domain-specific patterns – microaneurysms in retinal images, subtle textural differences between benign nevi and melanoma – that generic features may not encode. Yet they are typically large Vision Transformers [7] with hundreds of millions of parameters (RETFound: ViT-Large, 307M), rendering direct edge deployment infeasible.

Knowledge distillation (KD) [13] offers a principled route to model compression: a compact student network is trained to mimic the softened output distribution of a larger teacher, with the "dark knowledge" in those probabilities providing richer supervision than one-hot labels alone. KD has been applied in medical imaging for skin-lesion classification [15, 31] and diabetic retinopathy screening. However, prior medical-imaging KD studies employ conventional CNN teachers pretrained on ImageNet. A critical question remains open: *does the domain-specific knowledge captured by a medical foundation model transfer more effectively through distillation than generic ImageNet features?*

This question carries practical weight. If foundation-model teachers produce superior students, the large investment in domain pretraining yields compounding returns, raising the performance ceiling of compressed models for the edge. Conversely, if distillation fails to pre-

serve domain-specific representations, the advantage of such teachers may be confined to settings where full-scale inference is feasible.

To address this gap, we present a systematic comparison of a medical foundation model (RETFound, ViT-Large) against a general-purpose pretrained model (ConvNeXt-Base) as KD teachers for lightweight edge deployment. Our contributions are: (i) **a head-to-head comparison of a medical foundation model and a general-purpose model as KD teachers**: we distil RETFound (307M, domain-specific MAE pretraining) and ConvNeXt-Base (89M, ImageNet pretraining) into MobileNetV3-Small (2.5M) and EfficientNet-Lite0 (4.7M) on HAM10000 and APTOS-2019 with three random seeds, finding that, on balanced accuracy, ConvNeXt-Base produces the stronger student in all four settings; (ii) **a hyperparameter sensitivity analysis**: we sweep temperature τ and loss weight α, finding that RETFound requires higher temperatures ($\tau \geq 4$) for effective transfer whereas ConvNeXt is robust across a broad range; and (iii) **an end-to-end edge pipeline with explainability**: we quantify INT8 post-training quantization (accuracy within $\pm 0.3\%$, MobileNetV3-Small compressed to 4.2 MB) and use Grad-CAM to compare teacher attention.

2 Related work

2.1 Medical vision foundation models

The foundation-model paradigm – large-scale self-supervised pretraining followed by fine-tuning – has reshaped medical image analysis [4, 17]. RETFound [32] used MAE pretraining on 904 170 retinal images, outperforming ImageNet-pretrained baselines on ophthalmic tasks. UNI [5] extended this to pathology with over 100M tiles, BiomedCLIP [30] adopted contrastive language–image pretraining, and MedSAM [19] adapted SAM [16] for segmentation. These models share a deployment challenge: their large parameter counts demand substantial computation. While foundation models have been benchmarked in clinical domains [1], their utility *as knowledge-distillation teachers* has not been studied.

2.2 Knowledge distillation in medical imaging

Knowledge distillation [13] transfers learned representations through softened probability distributions. Extensions include feature-based methods [22, 29], relation-based approaches [20, 26], and methodological surveys [10, 3]. In medical imaging, Islam et al. [15] applied KD to skin-cancer classification with conventional CNN teachers, and Zheng et al. [31] proposed self-knowledge distillation for skin-lesion classification. Critically, all of this prior work uses traditional CNN teachers; none examines whether medical foundation-model representations distil more effectively than generic features.

2.3 Edge deployment and model compression

Lightweight architectures such as MobileNetV3 [14] and EfficientNet [25] achieve favourable accuracy–efficiency trade-offs through depthwise-separable convolutions and squeeze-and-excitation blocks. Post-training quantization [9], particularly INT8 dynamic quantization, reduces model size and latency with minimal accuracy loss, enabling point-of-care diagnostics on hardware without GPU support [21]. We evaluate the full pipeline from foundation-model teacher to quantized edge student.

2.4 Explainability in medical AI

Grad-CAM [24] produces visual explanations by weighting feature maps with gradient information. In medical imaging, such explanations are critical for clinical trust: practitioners must verify that models attend to diagnostically relevant structures rather than spurious correlations [23]. We use Grad-CAM to compare the spatial attention of the two teacher models.

3 Materials and methods

Figure 1 illustrates the experimental pipeline.

3.1 Datasets

HAM10000 [27] contains 10 015 dermoscopic images across seven diagnostic classes: melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses (akiec), vascular lesions (vasc), and dermatofibroma (df). The dataset is heavily imbalanced (nevi $\approx$67%). We use an 80/10/10 stratified train/validation/test split. **APTOS-2019** [2] contains 3 662 retinal fundus photographs graded into five diabetic-retinopathy severity levels: 0 (none), 1 (mild), 2 (moderate), 3 (severe), 4 (proliferative). We use an 80/10/10 stratified split. Images are resized to 224×224 pixels. Training augmentations include random horizontal/vertical flips, rotation ($\pm 30^\circ$), colour jitter (± 0.2), and random resized cropping; validation and test images undergo centre cropping and ImageNet [6] normalization.

3.2 Teacher and student models

RETFound (ViT-Large, 307M) [32] was pretrained via MAE on 904 170 retinal fundus and OCT images. We use the publicly released weights and fine-tune the full model with a linear classification head ($1024 \to n_c$). For HAM10000 (dermoscopy, not ophthalmic), this tests cross-domain transfer of medical foundation-model knowledge. **ConvNeXt-Base (ImageNet-1K, 89M)** [18] is a modernized CNN that incorporates ViT design principles within a convolutional framework, representing conventional transfer learning without domain-specific pretraining. Both

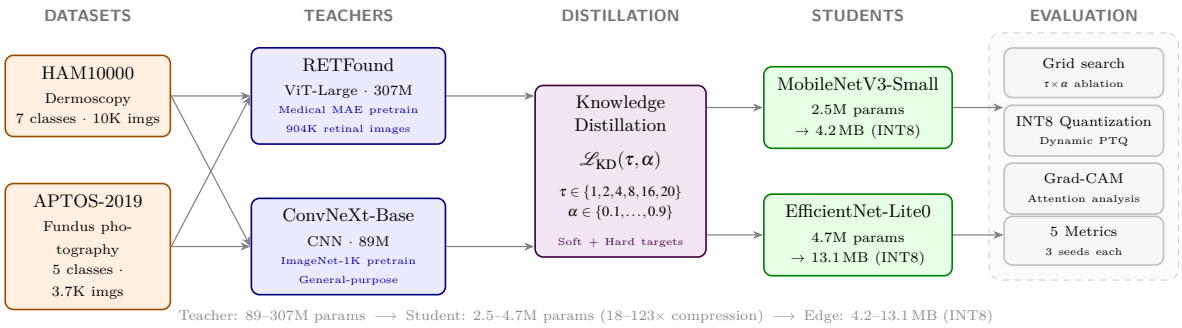

Figure 1: Experimental framework. Two teachers – a domain-specific medical foundation model (RETFound, ViT-large, 307m) and a general-purpose pretrained model (ConvNeXt-base, 89m) – are compared as KD teachers for two lightweight students (MobileNetV3-small, 2.5m; EfficientNet-Lite0, 4.7m) across two clinical benchmarks. Distilled students undergo a grid search over temperature τ and loss weight α, INT8 post-training quantization, and grad-CAM attention analysis.

teachers are fine-tuned with AdamW (lr $= 1 \times 10^{-4}$, cosine annealing, weight decay 0.01, batch size 32, 50 epochs, early stopping patience 10 on validation loss).

MobileNetV3-Small (2.5M) [14] is a hardware-aware, NAS-optimized architecture with inverted residuals, squeeze-and-excitation, and hard-swish activations. **EfficientNet-Lite0 (4.7M)** [25] is a compound-scaled MBConv architecture without squeeze-and-excitation (removed in the Lite variant for edge compatibility). Both students are initialized with ImageNet-pretrained weights.

3.3 Knowledge distillation framework

We adopt logit-based KD [13]:

$$\mathcal{L}_{\text{KD}} = \alpha\,\tau^2\,\text{KL}\Big(\sigma\big(\tfrac{\mathbf{z}_S}{\tau}\big)\,\Big\|\,\sigma\big(\tfrac{\mathbf{z}_T}{\tau}\big)\Big) + (1-\alpha)\,\mathcal{L}_{\text{CE}}(\mathbf{y}, \sigma(\mathbf{z}_S)), \tag{1}$$

where $\mathbf{z}_S$ and $\mathbf{z}_T$ are the student and teacher logits, τ is the temperature, α balances soft and hard targets, and the τ^2 factor compensates for the reduced gradient magnitude at higher temperatures.

Grid search. For ConvNeXt teachers we sweep $\tau \in \{1, 2, 4, 8, 16, 20\}$ and $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ (30 combinations on HAM10000; 22 on APTOS-2019). For RETFound teachers we sweep $\tau \in \{2, 4, 8\}$, $\alpha \in \{0.3, 0.5, 0.7\}$ on HAM10000 (9 combinations) and $\tau \in \{4, 8\}$, $\alpha \in \{0.5, 0.7\}$ on APTOS-2019 (4 combinations). The reduced RETFound space reflects preliminary runs in which extreme configurations ($\tau \leq 1$ or $\tau \geq 16$; $\alpha = 0.1$ or $\alpha = 0.9$) consistently fell below the no-KD baseline. This asymmetry in search effort is a limitation; however, the best RETFound configurations remain below ConvNeXt results obtained even at non-optimal ConvNeXt settings (e.g., $\tau = 2, \alpha = 0.7$ already exceeds the best RETFound result on HAM10000), so the gap is unlikely to be an artefact of search budget. No-KD baselines are trained with standard cross-entropy. Students are trained with SGD (momentum 0.9, weight decay 5×10^{-4}, lr 0.01 with cosine annealing, batch size 64, 100 epochs, early stopping patience 8 on validation accuracy).

3.4 Post-training quantization

We apply INT8 dynamic quantization (PyTorch) to `nn.Linear` layers, quantizing weights to INT8 for storage while activations remain in FP32. We report FP32 accuracy (GPU), INT8 accuracy (CPU), model size, and CPU inference latency, the last measured on an Intel Core i7-12700H CPU (single-threaded, eager mode, averaged over 50 forward passes). Quantization is evaluated on the best ConvNeXt-KD checkpoint (seed 42).

3.5 Grad-CAM visualization

We generate Grad-CAM heatmaps [24] from the last convolutional block of ConvNeXt-Base and the last transformer block of RETFound. Two representative test samples per dataset are selected, yielding eight maps that compare the two teachers across both datasets.

3.6 Evaluation metrics

We report accuracy, balanced accuracy (macro-averaged recall), macro F1-score, Cohen's κ (quadratic-weighted for the ordinal APTOS-2019 grades), and macro-averaged one-vs-rest AUC. Because both datasets are class-imbalanced, we treat *balanced accuracy* as the primary metric; raw accuracy is dominated by the majority class and is reported for completeness. All experiments use three random seeds (42, 123, 456); we report mean ± standard deviation.

4 Results

4.1 Teacher model performance

Table 1 reports teacher performance after fine-tuning. On HAM10000 – a dermoscopy benchmark that lies outside RETFound's retinal pretraining domain – ConvNeXt-Base outperforms RETFound on every metric despite having

3.4× fewer parameters, reaching 0.890 balanced accuracy versus 0.802 (an 8.8-point gap). On APTOS-2019, where RETFound's pretraining is in-domain, ConvNeXt still leads on balanced accuracy (0.692 vs. 0.645), though the gap narrows. This ordering – a general-purpose CNN matching or beating a larger domain-specific foundation model after fine-tuning on limited data (APTOS: 3 662 images) – is itself notable and frames the distillation results that follow.

4.2 Student performance by teacher type

Table 2 presents the core comparison. On balanced accuracy – our primary metric – ConvNeXt-KD produces the stronger student in all four teacher–student–dataset settings, and exceeds the no-KD baseline in all four. On raw accuracy the picture is noisier: the two teachers are statistically indistinguishable for MobileNetV3 on APTOS-2019 (0.759 vs. 0.763), the single cell in which RETFound edges ConvNeXt, and KD slightly trails no-KD for MobileNetV3 on HAM10000 – both differences fall within one standard deviation. We therefore anchor our conclusions on balanced accuracy, F1 and κ, which are appropriate for imbalanced data and on which the ConvNeXt advantage is uniform. Importantly, this advantage is not merely an artefact of RETFound being evaluated out-of-domain on HAM10000: it also holds on APTOS-2019, whose retinal images are in-domain for RETFound's pretraining. That the general-purpose teacher prevails even in RETFound's home domain indicates that domain-specific pretraining does not by itself guarantee more effective distillation.

Several observations emerge. (1) **ConvNeXt-KD produces stronger students on balanced accuracy.** On HAM10000 it improves MobileNetV3 balanced accuracy by 4.8 points over no-KD and 6.3 points over RETFound-KD; on APTOS-2019 the margins over RETFound-KD are 1.3 (MobileNetV3) and 4.8 (EfficientNet) points. The advantage holds across F1, κ and AUC. (2) **RETFound-KD can degrade students.** On HAM10000, RETFound-distilled MobileNetV3 reaches 0.763 balanced accuracy, 1.5 points *below* the no-KD baseline (0.778). (3) **Larger students benefit more.** EfficientNet-Lite0 (4.7M) consistently outperforms MobileNetV3-Small (2.5M) and gains more from distillation. (4) **KD helps more on smaller data.** On APTOS-2019 (2 929 training images) ConvNeXt-KD improves EfficientNet balanced accuracy by 6.6 points over no-KD, versus 2.2 points on HAM10000 (8 012 images).

4.3 Temperature and loss-weight ablation

ConvNeXt teacher. Table 3 reports representative ConvNeXt grid-search rows on HAM10000 (the full 30-row grid is available from the author on request). The best balanced accuracy is at $\tau = 8, \alpha = 0.7$ (0.856), with $\tau = 20, \alpha = 0.1$ giving the highest accuracy (0.887) and κ (0.783). ConvNeXt-KD is robust: any $\tau \geq 2$ with $\alpha \geq 0.5$ exceeds the no-KD baseline (balanced accuracy 0.789 at seed 42), except $\tau = 16$. On APTOS-2019 (Table 4), ConvNeXt-KD reaches its best balanced accuracy and κ at $\tau = 8, \alpha = 0.3$ (0.632 and 0.894).

RETFound teacher. Table 5 reports the targeted RETFound ablation. On HAM10000, accuracy rises from 0.818 ($\tau = 2$) to 0.836 ($\tau = 8, \alpha = 0.3$), confirming a need for higher temperatures; the best balanced accuracy (0.808) is at $\tau = 4, \alpha = 0.3$. On APTOS-2019, $\tau = 4, \alpha = 0.5$ is best (accuracy 0.809, balanced accuracy 0.647). Across both datasets, RETFound favours $\tau = 4$–8 with moderate $\alpha \approx 0.3$.

4.4 INT8 quantization impact

Table 6 summarizes quantization effects. INT8 dynamic quantization preserves accuracy within $\pm 0.3\%$ for all four students; the HAM10000 models even improve marginally (+0.1%). MobileNetV3-Small gains a 29% size reduction (5.9→4.2 MB) thanks to its higher proportion of `nn.Linear` parameters, whereas EfficientNet-Lite0 – dominated by MBConv blocks not targeted by dynamic linear quantization – barely shrinks. On CPU, MobileNetV3-Small is roughly 5.5× faster than EfficientNet-Lite0 (∼650 vs. ∼3800 ms per image in single-threaded eager mode), making it the preferred edge architecture when latency is the binding constraint.

4.5 Grad-CAM visualization

Figure 2 compares teacher attention. On HAM10000, ConvNeXt produces sharp, localized heatmaps over lesion boundaries and internal texture, whereas RETFound's maps are markedly more diffuse. On APTOS-2019 both teachers highlight vascular and exudate regions for DR-positive samples and attend broadly for DR0. The diffuseness of the RETFound maps is partly an artefact of applying gradient-based attribution to a Vision Transformer, whose global self-attention spreads gradients across patches; the comparison is therefore qualitative and illustrative rather than a quantitative measure of where each model "looks". With that caveat, both teachers base their predictions on clinically meaningful regions.

5 Discussion

A stronger, cheaper teacher wins. The central finding is that ConvNeXt-Base, with 3.4× fewer parameters and no domain-specific pretraining, distils into stronger students than RETFound on balanced accuracy across both datasets and both students. The most economical explanation is also visible in Table 1: ConvNeXt is simply the better fine-tuned teacher. We regard this not as a confound to be explained away but as the practical message – *a teacher's fine-tuned quality predicts student quality more reliably than its domain-specific pretraining*. Two observations support that the effect is not merely teacher accuracy. First, on APTOS-2019 the teacher balanced-accuracy gap

Table 1: Teacher model performance after fine-tuning (mean ± std, 3 seeds). κ is cohen's κ (HAM10000) and quadratic-weighted κ (aptos-2019). Best per dataset in **bold**.

Dataset	Teacher	Acc.	BAcc.	F1	κ	AUC
HAM10000	**ConvNeXt-Base**	**.902±.023**	**.890±.016**	**.867±.033**	**.816±.041**	**.988±.004**
	RETFound (ViT-L)	.847±.012	.802±.017	.786±.016	.713±.016	.969±.003
APTOS-2019	**ConvNeXt-Base**	**.811±.007**	**.692±.011**	**.670±.010**	**.897±.006**	**.945±.002**
	RETFound (ViT-L)	.785±.021	.645±.049	.634±.031	.895±.009	.931±.001

Table 2: Student performance by teacher type at the best (τ, α) per configuration (mean ± std, 3 seeds). "No KD" is standard cross-entropy training. κ is cohen's κ (HAM10000) and quadratic-weighted κ (aptos-2019). Best per student–dataset block in **bold** (balanced accuracy as primary criterion).

Dataset	Student	Method	Acc.	BAcc.	F1	κ	AUC
HAM10000	MobileNetV3-Small	No KD	.866±.008	.778±.025	.786±.015	.742±.016	.974±.004
		ConvNeXt-KD	.863±.016	**.826±.018**	**.801±.024**	.740±.031	.972±.005
		RETFound-KD	.834±.008	.763±.047	.757±.029	.679±.030	.968±.003
	EfficientNet-Lite0	No KD	.876±.007	.821±.005	.820±.010	.760±.011	.979±.002
		ConvNeXt-KD	**.881±.011**	**.843±.006**	**.838±.022**	**.774±.018**	**.982±.001**
		RETFound-KD	.857±.005	.811±.019	.803±.005	.728±.004	.976±.002
APTOS-2019	MobileNetV3-Small	No KD	.730±.040	.599±.010	.575±.025	.840±.020	.906±.008
		ConvNeXt-KD	.759±.042	**.635±.022**	**.607±.042**	**.862±.024**	**.924±.012**
		RETFound-KD	.763±.041	.622±.041	.604±.050	.855±.032	.921±.011
	EfficientNet-Lite0	No KD	.748±.018	.599±.008	.584±.020	.848±.008	.906±.006
		ConvNeXt-KD	**.770±.015**	**.665±.027**	**.627±.021**	**.875±.008**	**.936±.003**
		RETFound-KD	.768±.006	.617±.006	.603±.012	.852±.015	.918±.005

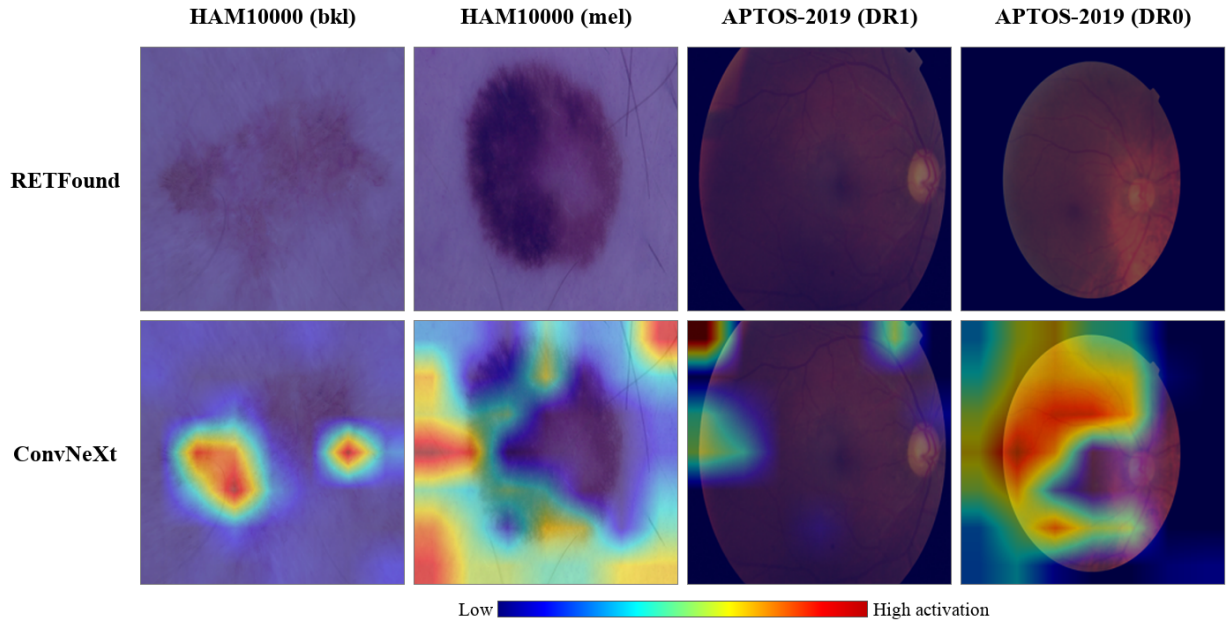

Figure 2: Grad-CAM comparison of RETFound (top row) and ConvNeXt-base (bottom row) on representative test samples: HAM10000 (bkl, mel) and aptos-2019 (DR1, DR0). Warmer colours indicate higher activation. ConvNeXt yields sharper, more localized maps; retfound's ViT-based maps are more diffuse.

(4.7 points) is smaller than the resulting EfficientNet student gap (4.8 points) yet the ordering is preserved; second, RETFound-KD can fall *below* the no-KD baseline, which a uniformly "weaker but still useful" teacher would not cause. Beyond accuracy, ConvNeXt's convolutional inductive biases may yield smoother, better-calibrated soft labels than RETFound's patch-based attention, and RETFound's MAE pretraining – optimized for reconstruction, not discrimination – may not produce logits that distil well through soft-label matching alone.

Hyperparameter sensitivity differs by teacher. ConvNeXt-KD is robust across a broad range (on

Table 3: ConvNeXt→MobileNetV3 grid search on HAM10000 (seed 42, representative rows). Best values in **bold**.

τ	α	Acc.	BAcc.	κ	AUC
1	0.1	0.871	0.821	0.750	0.980
2	0.7	0.869	0.810	0.743	0.980
4	0.1	0.881	0.835	0.772	0.977
8	0.5	0.879	0.811	0.763	0.978
8	0.7	0.857	**0.856**	0.727	**0.982**
16	0.3	0.823	0.754	0.652	0.954
20	0.1	**0.887**	0.830	**0.783**	0.978
20	0.3	0.887	0.846	0.782	0.980

Table 4: ConvNeXt→MobileNetV3 ablation on aptos-2019 (seed 42, best per τ). Best values in **bold**.

τ	α	Acc.	BAcc.	κ	AUC
1	0.7	0.804	0.600	0.847	0.921
2	0.7	0.793	0.617	0.875	0.935
4	0.7	0.763	0.609	0.855	0.923
8	0.3	0.796	**0.632**	**0.894**	0.933
16	0.5	0.771	0.587	0.867	0.920
20	0.5	0.785	0.645	0.890	0.927
20	0.7	**0.801**	0.627	0.890	0.930

HAM10000, any $\tau \geq 2$ with $\alpha \geq 0.5$ beats no-KD), whereas RETFound-KD has a narrow effective range, needing $\tau \geq 4$ with moderate $\alpha \approx 0.3$. RETFound's ViT logits appear over-confident at low temperature, requiring stronger softening to expose useful inter-class structure. Practitioners distilling from a ViT-based foundation model should therefore start at $\tau \geq 4$.

Edge feasibility. INT8 dynamic quantization preserves accuracy within $\pm$0.3% for all students, consistent with post-training quantization being safe for medical image classifiers when degradation stays below 1% [9]. ConvNeXt-distilled MobileNetV3-Small, after INT8 quantization, is a 4.2 MB model with HAM10000 accuracy 0.855 and APTOS-2019 accuracy 0.804 – comfortably within the storage and latency budgets of mobile applications. At $\sim$650 ms per image on a single CPU thread it processes roughly 1.5 images/s without a GPU, sufficient for point-of-care screening; we note these are eager-mode figures and that a compiled or NPU-accelerated runtime would reduce them substantially.

Limitations. Several constraints bound our conclusions. (i) We study two datasets and two modalities; chest radiography and pathology would strengthen generalization, and our claims about "medical foundation models" rest on a *single* model (RETFound) – others (UNI [5], BiomedCLIP [30]) may behave differently. (ii) We use only logit-based KD; feature-based methods [22, 29] that transfer intermediate representations may better exploit a ViT's patch-level features and could change the ranking – the most important direction for future work. (iii) With three seeds, statistical power is limited; we mitigate this by anchoring claims on balanced accuracy and on the *consistent direction* across four independent settings rather than on any single cell, but formal significance testing with more seeds is warranted. (iv) The grid search is asymmetric (30 vs. 9 configurations on HAM10000), which may understate RETFound's optimum, although the margin argument in the Methods bounds this risk. (v) Quantization uses dynamic INT8 on CPU; static quantization with calibration on dedicated edge hardware (Jetson, smartphone NPUs) would give more realistic benchmarks. (vi) RETFound is pretrained on retinal images, so HAM10000 is an out-of-domain test by design; this isolates cross-domain transfer but limits in-domain conclusions. (vii) Our Grad-CAM analysis is qualitative and limited to teachers; quantitative teacher–student attention-agreement metrics would strengthen claims about attention transfer.

6 Conclusion

We presented a systematic comparison of a medical vision foundation model (RETFound, ViT-Large, 307M) against a general-purpose pretrained model (ConvNeXt-Base, 89M) as knowledge-distillation teachers for edge deployment of medical image classifiers. Across HAM10000 and APTOS-2019 with two lightweight students and three seeds, ConvNeXt-Base distils into stronger students on balanced accuracy in all four settings, despite 3.4$\times$ fewer parameters and no domain-specific pretraining; on HAM10000 it improves MobileNetV3 balanced accuracy by 6.3 points over RETFound-KD (0.826 vs. 0.763), and RETFound-KD can even fall below the no-distillation baseline. The two teacher types differ in hyperparameter sensitivity – ConvNeXt robust, RETFound needing $\tau \geq 4$ – and INT8 quantization compresses MobileNetV3-Small to 4.2 MB within $\pm$0.3% accuracy, making it a practical point-of-care solution. These findings indicate that a teacher's fine-tuned quality, not domain-specific pretraining, is the better predictor of distillation success, and that modern CNNs with strong ImageNet initialization remain competitive – often preferable – KD teachers for medical imaging. Future work should extend the comparison to feature-based distillation, additional modalities and foundation models, and hardware-specific edge benchmarks.

Acknowledgement

The author thanks the providers of the publicly available HAM10000 and APTOS-2019 datasets and the authors of RETFound for releasing their pretrained weights. This research received no external funding. The HAM10000 dataset is available at `https://doi.org/10.7910/DVN/DBW86T`, the APTOS-2019 dataset at `https://www.kaggle.com/c/aptos2019-blindness-detection`, and the RETFound

Table 5: RETFound→MobileNetV3 ablation (seed 42). Best per dataset in **bold**.

Dataset	τ	α	Acc.	BAcc.	κ	AUC
HAM10000	2	0.5	0.818	0.778	0.794	0.962
	4	0.3	0.831	**0.808**	**0.831**	0.973
	8	0.3	**0.836**	0.806	0.811	**0.970**
APTOS-2019	4	0.5	**0.809**	**0.647**	**0.892**	**0.923**
	8	0.5	0.768	0.607	0.881	0.913

Table 6: INT8 dynamic quantization on the best ConvNeXt-KD checkpoint (seed 42). FP32 latency is measured on GPU and INT8 latency on CPU, reflecting the target edge scenario.

Dataset	Student	FP32 Acc.	INT8 Acc.	ΔAcc.	Size (MB)	CPU (ms)
HAM10000	MobileNetV3-Small	0.854	0.855	+0.001	5.9→4.2	688
	EfficientNet-Lite0	0.889	0.890	+0.001	13.2→13.1	3602
APTOS-2019	MobileNetV3-Small	0.807	0.804	−0.003	5.9→4.2	642
	EfficientNet-Lite0	0.766	0.763	−0.003	13.1→13.1	3960

weights at `https://github.com/rmaphoh/RETFound_MAE`. The author declares no conflict of interest.

References

[1] M. A. Al-Masud, J. M. Lopez Alcaraz, N. Strodthoff (2025) Benchmarking ECG foundational models: A reality check across clinical tasks, *arXiv preprint arXiv:2509.25095*. `https://doi.org/10.48550/arXiv.2509.25095`

[2] Asia Pacific Tele-Ophthalmology Society (2019) APTOS 2019 blindness detection, *Kaggle Competition*. `https://www.kaggle.com/competitions/aptos2019-blindness-detection`

[3] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, A. Kolesnikov (2022) Knowledge distillation: A good teacher is patient and consistent, *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10915–10924. `https://doi.org/10.1109/CVPR52688.2022.01065`

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, et al. (2021) On the opportunities and risks of foundation models, *arXiv preprint arXiv:2108.07258*. `https://doi.org/10.48550/arXiv.2108.07258`

[5] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, et al. (2024) Towards a general-purpose foundation model for computational pathology, *Nature Medicine*, 30, pp. 850–862. `https://doi.org/10.1038/s41591-024-02857-3`

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei (2009) ImageNet: A large-scale hierarchical image database, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. `https://doi.org/10.1109/CVPR.2009.5206848`

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale, *Proc. Int. Conf. on Learning Representations (ICLR)*.

[8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun (2017) Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, 542(7639), pp. 115–118. `https://doi.org/10.1038/nature21056`

[9] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. Keutzer (2022) A survey of quantization methods for efficient neural network inference, *Low-Power Computer Vision*, Chapman and Hall/CRC, pp. 291–326. `https://doi.org/10.1201/9781003162810-13`

[10] J. Gou, B. Yu, S. J. Maybank, D. Tao (2021) Knowledge distillation: A survey, *International Journal of Computer Vision*, 129, pp. 1789–1819. `https://doi.org/10.1007/s11263-021-01453-z`

[11] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *JAMA*, 316(22), pp. 2402–2410. `https://doi.org/10.1001/jama.2016.17216`

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick (2022) Masked autoencoders are scalable vision learners, *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988. `https://doi.org/10.1109/CVPR52688.2022.01553`

[13] G. Hinton, O. Vinyals, J. Dean (2015) Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531*. https://doi.org/10.48550/arXiv.1503.02531

[14] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, et al. (2019) Searching for MobileNetV3, *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

[15] N. Islam, K. M. Hasib, F. A. Joti, A. Karim, S. Azam (2024) Leveraging knowledge distillation for lightweight skin cancer classification: Balancing accuracy and computational efficiency, *arXiv preprint arXiv:2406.17051*. https://doi.org/10.48550/arXiv.2406.17051

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, et al. (2023) Segment anything, *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

[17] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, et al. (2017) A survey on deep learning in medical image analysis, *Medical Image Analysis*, 42, pp. 60–88. https://doi.org/10.1016/j.media.2017.07.005

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie (2022) A ConvNet for the 2020s, *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

[19] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang (2024) Segment anything in medical images, *Nature Communications*, 15, 654. https://doi.org/10.1038/s41467-024-44824-z

[20] W. Park, D. Kim, Y. Lu, M. Cho (2019) Relational knowledge distillation, *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. https://doi.org/10.1109/CVPR.2019.00409

[21] A. Rancea, I. Anghel, T. Cioara (2024) Edge computing in healthcare: Innovations, opportunities, and challenges, *Future Internet*, 16(9), 329. https://doi.org/10.3390/fi16090329

[22] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio (2015) FitNets: Hints for thin deep nets, *Proc. Int. Conf. on Learning Representations (ICLR)*.

[23] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin (2022) Transparency of deep neural networks for medical image analysis: A review of interpretability methods, *Computers in Biology and Medicine*, 140, 105111. https://doi.org/10.1016/j.compbiomed.2021.105111

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization, *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74

[25] M. Tan, Q. V. Le (2019) EfficientNet: Rethinking model scaling for convolutional neural networks, *Proc. 36th Int. Conf. on Machine Learning (ICML)*, PMLR 97, pp. 6105–6114.

[26] Y. Tian, D. Krishnan, P. Isola (2020) Contrastive representation distillation, *Proc. Int. Conf. on Learning Representations (ICLR)*.

[27] P. Tschandl, C. Rosendahl, H. Kittler (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Scientific Data*, 5, 180161. https://doi.org/10.1038/sdata.2018.161

[28] Y. Xu, T. M. Khan, Y. Song, E. Meijering (2025) Edge deep learning in computer vision and medical diagnostics: A comprehensive survey, *Artificial Intelligence Review*, 58(3), 93. https://doi.org/10.1007/s10462-024-11033-5

[29] S. Zagoruyko, N. Komodakis (2017) Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, *Proc. Int. Conf. on Learning Representations (ICLR)*.

[30] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, et al. (2025) A multimodal biomedical foundation model trained from fifteen million image–text pairs, *NEJM AI*, 2(1). https://doi.org/10.1056/AIoa2400640

[31] J. Zheng, K. Xie, D. Zhang, Z. Lv, X. Yu (2025) Stepwise self-knowledge distillation for skin lesion image classification, *Scientific Reports*, 15. https://doi.org/10.1038/s41598-025-10717-4

[32] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, et al. (2023) A foundation model for generalizable disease detection from retinal images, *Nature*, 622(7981), pp. 156–163. https://doi.org/10.1038/s41586-023-06555-x