CoFIDA-M: Concept-Aware Feature Modulation for Cross-Domain Adaptation with Image-Only Inference

Supplementary Materials

Nurjahan Sultana, Moi Hoon Yap, Xinqi Fan, Wenqi Lu
Department of Computing and Mathematics, Manchester Metropolitan University
Dalton Building, Chester Street, M1 5GD Manchester, UK
nurjahan.sultana@stu.mmu.ac.uk; {m.yap, x.fan, w.lu}@mmu.ac.uk

1. Extended Metrics and Statistical Analysis

We report balanced accuracy (BA) in Table 1, to provide an additional view of performance under the substantial class imbalance present in the datasets. AUROC and melanoma recall were presented in the main paper, as they are the most clinically relevant metrics for safety critical screening. BA is included here because it offers a simple and robust measure of performance across both classes without being dominated by the large number of other samples. This makes BA a suitable complementary metric for comparing methods under the highly imbalanced distribution observed across the eight evaluation datasets.

To assess whether differences in BA between methods were statistically significant, we ran a one way ANOVA on the full set of per dataset and per seed results. For each method, we collected BA values for all eight evaluation datasets (Md, MIDd, D7d, HAM, Mc, MIDc, D7c, Fitz) and all five random seeds, giving $8 \times 5 = 40$ observations per method. These 40 BA values were treated as the performance distribution for that method, and we fitted an ANOVA model with BA as the dependent variable and "method" as the factor.

The ANOVA showed a strong main effect of method on BA, $F(15, 624) = 26.87$, $p < 0.0001$. Across all methods, the best performance was obtained by our proposed CoFiDA-M, with a mean BA of 70.96 ± 8.24 over its 40 observations. Most baseline methods lay in a lower band of mean BA values (typically around 50–62), confirming that CoFiDA-M achieves a clear absolute gain in balanced accuracy across the eight datasets.

We then applied a post hoc Tukey HSD test to compare methods pairwise while controlling the familywise error rate. Tukey's test confirmed that CoFiDA-M's mean BA was significantly higher than that of every baseline method (all adjusted $p < 0.001$). The test also indicated that some baseline pairs differed significantly from one another, while others did not, which is consistent with the overall method effect observed in the ANOVA.

2. Detailed Ablation Studies

A natural question is whether the full MONET concept space is needed, or if a smaller subset is sufficient. To study this, we first trained a random forest on the MILK10k metadata and used its feature importances to rank MONET concepts for predicting melanoma. This probe suggested that 'MONET_pigmented' is the most informative concept, followed by 'erythema', 'vasculature', 'skin markings', 'gel water-drop artifacts', and 'ulceration', with 'hair' contributing least. We then trained the teacher under four configurations: the full MONET concept set; the top six concepts from this ranking; the top four concepts only; and a single–concept variant using 'MONET_pigmented' alone. Figure 1 shows that the full concept set gives the highest AUROC and melanoma recall. Using the top six concepts retains much of this performance but still falls short of the full configuration, while the top four and single–concept variants show clear drops. This suggests that the teacher benefits from the full, diverse set of MONET attributes, as they provide complementary clinical cues that are lost when the concept space is compressed to only the most important few.

2.1. Analysis of Distillation Loss Weights

A key hyperparameter in our distillation stage is the feature alignment weight λ_{feat}, which controls how strongly the student is encouraged to match the teacher's concept-edited features rather than only its final logits. To study this trade-off, we vary λ_{feat} from 0.0 (logit-only distillation) to 1.0 (dominant feature alignment).

As shown in Figure 2, performance peaks at $\lambda_{\text{feat}} = 0.1$. With no feature alignment ($\lambda_{\text{feat}} = 0$), the student does not fully inherit the teacher's clinical reasoning, leading to a 2.9% AUROC drop relative to the optimum. When feature alignment is too strong ($\lambda_{\text{feat}} \geq 0.5$), it severely over-

Table 1. Balanced Accuracy (BA) (%) mean ± std over **5** seeds. Left: **Dermoscopic** (A), macro-average across Dermoscopic and Right **Clinical** (B), macro-average across Clinical. MIDd, MIDc, D7d, D7c, HAM and Fitz are unseen test datasets.

	(A) Dermoscopic					(B) Clinical				
Setting / Method	**Md**	**MIDd***	**D7d***	**HAM***	**Avg**	**Mc**	**MIDc***	**D7c***	**Fitz***	**Avg**
Classical / Standard										
DANN	62.45 ± 0.12	60.80 ± 0.06	62.28 ± 0.09	51.51 ± 0.03	59.26	58.56 ± 0.05	49.86 ± 0.02	51.35 ± 0.07	41.21 ± 0.05	50.25
DeepCoral	65.10 ± 0.02	64.17 ± 0.06	69.31 ± 0.05	49.36 ± 0.09	61.99	61.49 ± 0.03	51.90 ± 0.01	51.41 ± 0.07	51.40 ± 0.02	54.05
MMD	68.38 ± 0.10	69.33 ± 0.11	58.87 ± 0.06	73.80 ± 0.17	67.60	61.60 ± 0.02	52.49 ± 0.04	45.19 ± 0.01	53.80 ± 0.03	53.27
MeanTeacher	55.76 ± 0.12	45.86 ± 0.09	46.78 ± 0.08	36.90 ± 0.14	46.33	42.58 ± 0.06	51.58 ± 0.08	44.86 ± 0.12	50.70 ± 0.06	47.93
IT–RUDA	76.37 ± 0.04	59.75 ± 0.06	62.48 ± 0.08	75.27 ± 0.07	68.47	53.64 ± 0.06	51.40 ± 0.07	53.63 ± 0.08	52.10 ± 0.03	52.69
Modern Source Free										
SHOT++	65.59 ± 0.10	54.47 ± 0.08	59.90 ± 0.08	61.36 ± 0.09	60.33	53.20 ± 0.05	55.80 ± 0.07	50.58 ± 0.06	45.28 ± 0.07	51.21
DINE	69.36 ± 0.05	56.85 ± 0.07	74.49 ± 0.02	85.32 ± 0.06	71.51	52.35 ± 0.04	51.10 ± 0.03	70.96 ± 0.08	77.26 ± 0.13	62.92
SFDA	50.00 ± 0.01	50.00 ± 0.02	49.47 ± 0.01	49.79 ± 0.04	49.82	50.00 ± 0.01	50.00 ± 0.01	53.45 ± 0.05	50.00 ± 0.02	50.86
CPD	51.90 ± 0.04	47.87 ± 0.02	52.49 ± 0.02	54.38 ± 0.03	51.66	51.79 ± 0.01	47.60 ± 0.02	53.40 ± 0.03	49.30 ± 0.09	50.52
Modern Test Time										
CoTTA	59.26 ± 0.09	48.35 ± 0.10	50.27 ± 0.04	54.67 ± 0.05	53.14	49.42 ± 0.02	47.24 ± 0.05	49.83 ± 0.02	46.40 ± 0.07	48.22
TENT	65.28 ± 0.04	61.10 ± 0.05	68.59 ± 0.10	72.38 ± 0.09	66.84	59.52 ± 0.08	55.40 ± 0.04	54.26 ± 0.09	58.62 ± 0.07	56.95
WoC	62.46 ± 0.08	58.30 ± 0.15	60.75 ± 0.06	72.06 ± 0.12	63.39	51.84 ± 0.03	47.52 ± 0.02	58.83 ± 0.07	50.28 ± 0.05	52.12
Multimodal										
DAMP	66.82 ± 0.04	56.37 ± 0.05	56.10 ± 0.03	71.34 ± 0.06	62.66	51.78 ± 0.03	49.57 ± 0.07	52.64 ± 0.03	49.82 ± 0.05	50.95
DALUPI	74.30 ± 0.01	58.72 ± 0.02	60.23 ± 0.05	72.32 ± 0.04	66.39	51.73 ± 0.02	50.34 ± 0.04	52.68 ± 0.05	50.24 ± 0.02	51.25
Source Only	73.82 ± 0.06	57.49 ± 0.03	47.75 ± 0.05	64.94 ± 0.01	61.00	69.37 ± 0.04	47.63 ± 0.02	48.87 ± 0.08	50.12 ± 0.01	54.00
Ours (CoFIDA-M)	86.73 ± 0.05	71.57 ± 0.02	63.57 ± 0.05	70.35 ± 0.05	**73.06**	80.42 ± 0.01	65.84 ± 0.06	68.44 ± 0.09	60.78 ± 0.07	**68.87**

Balanced Accuracy (BA) is computed as $\mathrm{BA} = \frac{1}{2}(\mathrm{TPR} + \mathrm{TNR})$, where TPR is the melanoma recall and TNR is the recall for other cases.

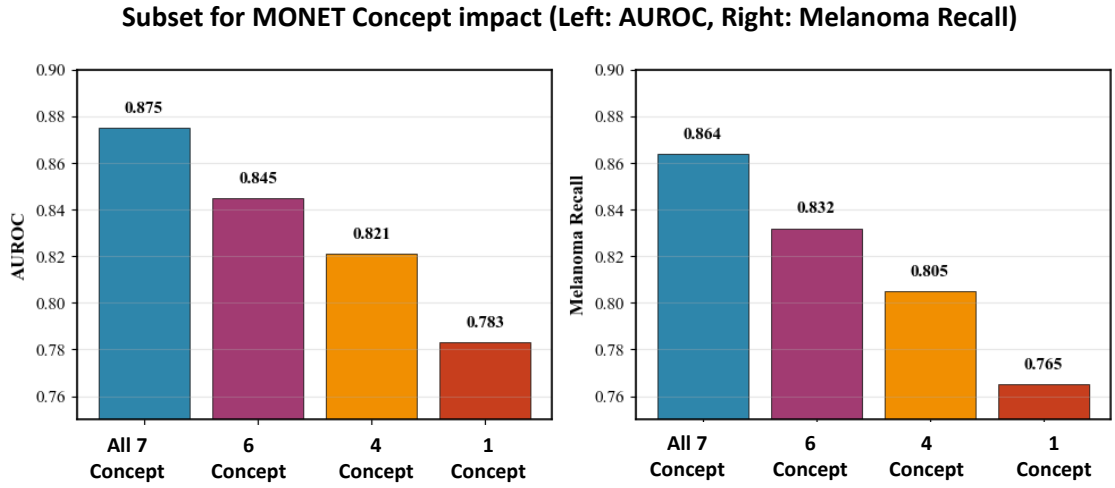


Figure 1. Ablation on MONET concept subsets. Using the full set of 7 MONET concepts produces the highest AUROC and melanoma recall. Performance systematically degrades with concept subset reduction, demonstrating that comprehensive clinical knowledge is essential for optimal cross-domain adaptation.

constrains the student, limiting its ability to form its own image-only representations and causing up to 66% performance degradation. This demonstrates that effective distillation requires balanced feature guidance. The setting $\lambda_{\text{feat}} = 0.1$ provides sufficient concept-aware feature supervision while preserving flexibility for efficient image-only learning.

2.2. Analysis of Predictive Confidence Separation

While calibration assesses the reliability of confidence scores overall, effective clinical deployment requires models to distinguish between correct and incorrect predictions. We therefore analyze the **Confidence Gap**, defined as the difference between mean confidence for correct versus incorrect predictions.

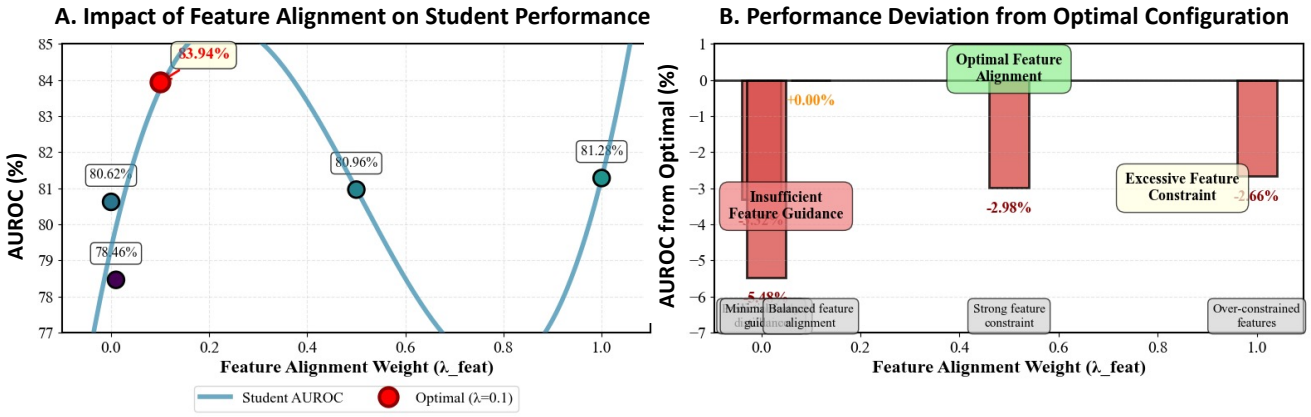


Figure 2. **Distillation feature alignment weight analysis.** (A) Student AUROC as a function of the feature alignment weight λ_{feat} shows a clear optimum at 0.1. (B) AUROC deviation from this optimum reveals that insufficient feature guidance ($\lambda_{\text{feat}} = 0$) reduces performance by 2.9%, while excessive feature constraints ($\lambda_{\text{feat}} \geq 0.5$) cause severe degradation up to 66%, confirming the critical need for balanced knowledge transfer.

As shown in Figure 3, CoFiDA-M achieves the largest confidence gap compared to DINE and DALUPI, indicating our image-only student provides more decisive predictions: it assigns higher confidence when correct and remains appropriately uncertain when wrong. This behavior is particularly desirable for clinical decision support, where confidence interpretability directly impacts usability.

3. Implementation and Training Details

3.1. MONET concept scores and interpretation

MONET is a medical image–text foundation model built from clinical literature and expert knowledge. It predicts human–readable clinical concepts as probabilities, which makes model behaviour easier to inspect and audit. For each lesion, MONET outputs a value in $[0, 1]$ for attributes such as ulceration, hair, vessels, erythema, pigmentation, gel artefacts and ink marks. These values come from a multi–label classifier trained to recognise the presence of each concept: values near 1 indicate presence, values near 0 indicate absence, and mid–range values indicate uncertainty. In our work, we use these MONET probabilities as structured clinical information during training. Table 2 shows example MILK10k metadata, including the MONET concept scores that we feed into our model.

3.2. Augmentation Policies

We use separate weak and strong augmentation pipelines for the target images to define the consistency loss. The weak view applies only a Random Horizontal Flip with probability 0.5. The strong view uses the same flip plus Color Jitter (brightness $= 0.2$, contrast $= 0.2$, saturation $= 0.1$, hue $= 0.02$) and Gaussian Blur with kernel size 3 and $\sigma \in [0.1, 2.0]$. For the labeled source branch we apply a milder Color Jitter (brightness $= 0.1$, contrast $= 0.1$, saturation $= 0.05$, hue $= 0.02$) together with Random Horizontal Flip. Standard tensor conversion and normalization are applied in all cases.

3.3. Compute and Time Requirements

CoFiDA-M demonstrates practical training efficiency and fast inference suitable for clinical deployment. The MONET-guided teacher trains in approximately **40 minutes** (50 epochs). The subsequent image-only student distillation is notably more efficient, requiring only **25–30 minutes** for the same number of epochs—representing a **25–37% reduction** in training time compared to the teacher stage.

Most importantly, Figure 4, the final student model achieves high-speed inference, processing a single image in just **0.53 ms** on average. This represents a **6.2%** speed improvement over the fastest baseline (DALUPI, 0.56 ms) and demonstrates that our method provides robust domain adaptation without sacrificing deployment efficiency. The combination of reasonable training times and fast inference makes CoFiDA-M practical for real-world clinical implementation and potential mobile deployment.

4. Qualitative Analysis and Interpretability

4.1. t-SNE of Image-Only Student Features

To check whether the image-only student inherits the teacher's structure, we project its features into 2D with t-SNE. Figure 5 compares the student's pre-edit features (left, v_S) with the post-edit features (right, u_S) for melanoma and other lesions drawn from both dermoscopic and clinical domains. Before the learned edit, the two classes overlap heavily. After the edit, melanoma and other lesions form much clearer clusters and the separation score im-

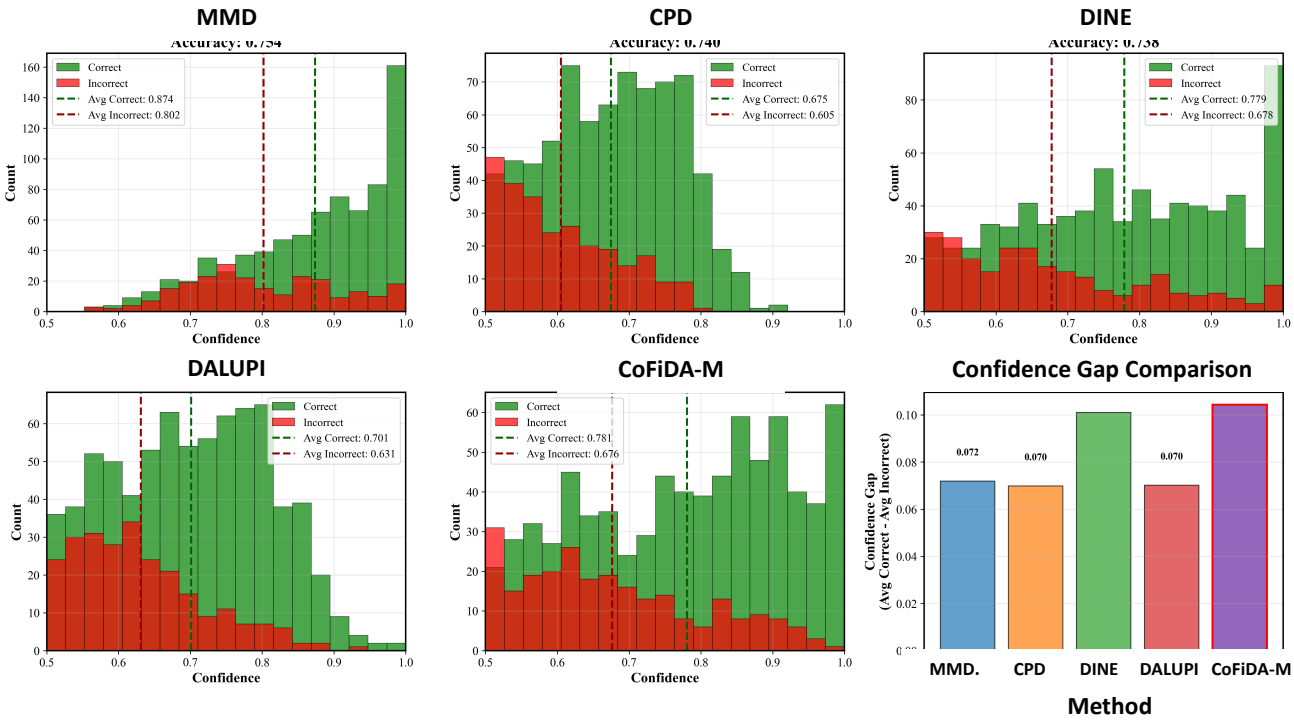


Figure 3. Confidence behavior across best baseline models. Each histogram shows the distribution of confidence for correct (green) and incorrect (red) predictions for five methods. The dashed vertical lines indicate the mean confidence for correct and incorrect cases. A larger separation between these lines reflects more reliable use of confidence. Although overall accuracy is similar across models, CoFiDA-M shows the widest gap between correct and incorrect confidence, indicating the most trustworthy confidence estimates. The bar chart summarizes the confidence gap for all methods.

Table 2. MONET concept probabilities for sample lesions.

Lesion_id	MONET ulceration_crust	MONET hair	MONET vasculature_vessels	MONET erythema	MONET pigmented	MONET gel_water_drop fluid_dermoscopy_liquid	MONET skin_markings pen_ink_purple_pen
IL_0000652	0.166 749 373	0.163 600 949	0.002 283 759	0.124 314 77	0.719 495 254	0.220 399 191	0.237 600 89
IL_0000652	0.659 858 767	0.156 477 614	0.016 397 272	0.032 356 863	0.847 014 362	0.138 121 264	0.148 776 498
IL_0003176	0.348 609 118	0.614 718 388	0.013 415 243	0.447 484 502	0.061 976 764	0.296 341 196	0.058 005 734
IL_0003176	0.392 950 492	0.897 667 690	0.367 881 561	0.645 776 203	0.122 107 737	0.719 937 352	0.329 811 751
IL_0004688	0.889 924 748	0.120 788 396	0.004 545 638	0.487 298 359	0.036 549 860	0.146 776 046	0.086 027 341
IL_0004688	0.548 514 651	0.208 268 517	0.196 452 934	0.519 807 548	0.058 424 306	0.319 812 227	0.262 882 598
IL_0005081	0.689 616 352	0.086 868 476	0.002 981 893	0.158 400 112	0.088 117 636	0.037 188 167	0.009 897 140
IL_0005081	0.820 477 552	0.612 078 543	0.444 256 644	0.272 835 535	0.259 830 049	0.384 243 353	0.220 156 278

proves, even though the student never receives MONET concepts directly. This shows that the student has absorbed the teacher's concept-aware feature editing purely through distillation.

4.2. Feature Editing Maps

To give an intuition for how the FiLM based edit operates in the teacher, we visualize its behavior on a melanoma example and a other example Figure 6. For each image we first compute a Grad-CAM map from the convolutional layer just before FiLM to obtain the base attention pattern on the raw backbone features v. We then pass the same features through the MONET conditioned FiLM layer to obtain the edited features u, and recompute Grad-CAM on this edited representation. The top row shows the original image, the Grad-CAM heatmap and the attention overlay, before and after the edit.

The bottom row reports three complementary views of

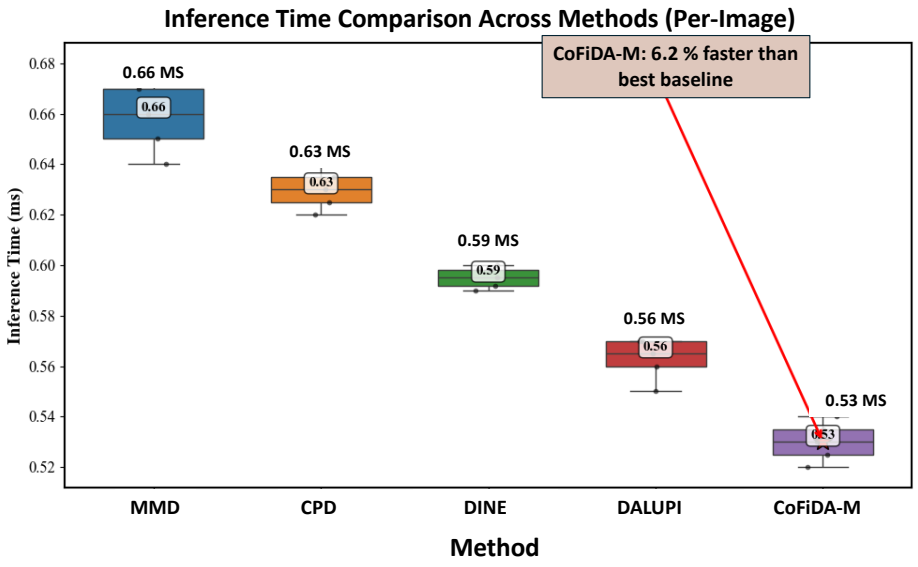


Figure 4. Inference speed benchmark. The image-only CoFiDA-M student achieves the fastest inference (0.53 ms/image), outperforming all baselines including DALUPI (0.56 ms). This efficiency gain, combined with its performance improvements, makes CoFiDA-M practical for clinical deployment.

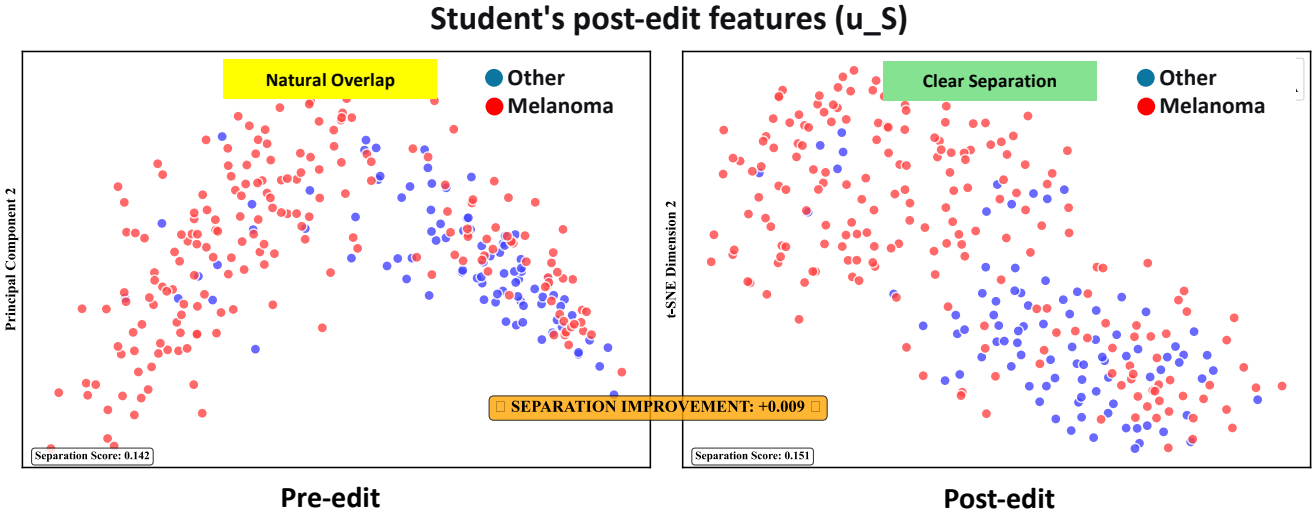
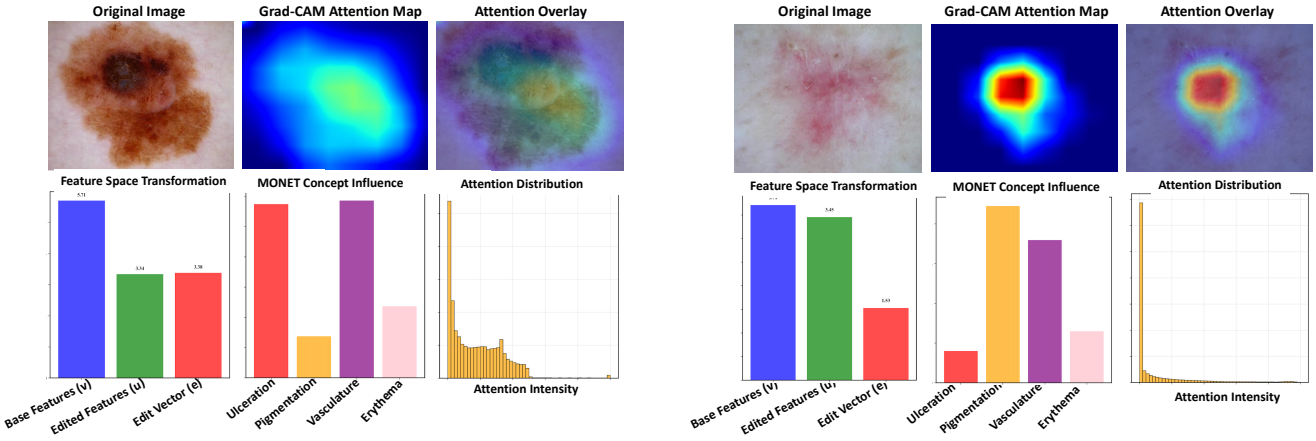


Figure 5. t-SNE of the image-only student's feature space. Left: pre-edit features (v_S) show substantial overlap between melanoma and other lesions. Right: post-edit features (u_S) learned through distillation produce a clearer class boundary and a higher separation score, indicating that the student has internalized the teacher's concept-aware editing while using images only at both train and test time.

the edit. The "Feature Space Transformation" bars summarize the norms of v, u and the edit vector $e = u - v$, showing how strongly the concept guided edit moves the representation. The "MONET Concept Influence" panel plots the MONET probabilities for key attributes such as pigment, ulceration, vessels and erythema that drive the FiLM parameters for that case. Finally, the "Attention Distribution" histogram shows how attention becomes more concentrated after editing. Together these examples illustrate that the teacher uses MONET concepts to steer attention toward clinically relevant parts of the lesion. The student model never sees MONET at test time, but is trained to imitate these edited features and decisions using image only inputs.



(a) Melanoma example.

(b) Other example.

Figure 6. Concept-guided feature editing visualisation. The top row shows original images, Grad-CAM attention before editing (base features v), and attention after MONET-guided FiLM editing (edited features u). The bottom row quantifies the transformation across feature norms (v, u, and edit vector $e = u - v$), concept probabilities driving the edit, and attention concentration changes. The teacher uses MONET concepts to guide attention towards clinically meaningful regions, and the student learns to reproduce this behaviour without concept inputs.