Supplementary Information for FSD-Net: A Unified Skin-Tone-Aware Framework for Fair Skin Lesion Segmentation and Classification

Ubaid Khan[1], Usman Habib[2], Sheheryar Mohsin Qureshi[3*], Muhammad Ehtisham Hassan[4], Adeel Ashraf Cheema[5], Sameed Tariq[6]

[1]Software Engineering Department, FAST School of Computing, National University of Computer and Emerging Sciences, H-11, Islamabad, 44000, Pakistan.
[3]Sciences and Engineering, University of the West of Scotland, Renfrewshire, PA1 2BE, Scotland.
[4]Department of Computer and Software Engineering, Air University Aerospace and Aviation Campus, Kamra, 43570, Pakistan.
[5]Department of AI and DS, National University of Computer and Emerging Sciences, Chiniot-Faisalabad, 35400, Pakistan.
[6]Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, 23460, Pakistan.

*Corresponding author(s). E-mail(s): sheheryar.qureshi@uws.ac.uk;

Supplementary Information

We provide additional architectural specifications, implementation details, robustness analyses, statistical results, and diagnostic figures that support the main manuscript. We use the prefix S for every supplementary note, equation, table, and figure. The main manuscript cites each item at the point where the additional detail is needed.

Table S1 Encoder tap specification.

Tap	Source block	Output shape	Role
F_{low}	2	$14 \times 14 \times 768$	Low level edges and texture
F_{mid}	5	$14 \times 14 \times 768$	Intermediate lesion structure
F_{high}	11	$14 \times 14 \times 768$	High level semantics conditioned by FiLM

Supplementary Note S1: Detailed architecture specification

ViT-B/16 encoder and feature taps

We use a ViT-B/16 encoder with 12 transformer blocks, a patch size of 16, an input resolution of 224×224 pixels, and a hidden dimension of 768. We extract intermediate features from blocks 2, 5, and 11. These taps represent low level appearance, intermediate lesion structure, and high level semantics. Because the encoder keeps the same token grid across depth, each tap has a spatial size of 14×14 and 768 channels. The decoder creates the spatial hierarchy through learned upsampling.

Hierarchical decoder formulation

We apply FiLM to F_{high} before decoding. At the first decoder stage, a transposed convolution maps the conditioned feature from 768 channels to 512 channels and doubles its spatial size. A parallel transposed convolution projects F_{mid} to the same shape. We concatenate both tensors to form the first fused decoder feature.

$$D_1 = \text{ConvT}_{768\to512}(F'_{\text{high}}) \in \mathbb{R}^{28\times28\times512}, \tag{S1}$$

$$S_1 = \text{ConvT}_{768\to512}(F_{\text{mid}}) \in \mathbb{R}^{28\times28\times512}, \tag{S2}$$

$$D'_1 = \text{Concat}(D_1, S_1) \in \mathbb{R}^{28\times28\times1024}. \tag{S3}$$

At the second stage, the decoder upsamples the accumulated feature to 56×56 pixels and 256 channels. In parallel, it projects F_{low} to the same dimensions before concatenation.

$$D_2 = \text{ConvT}_{1024\to256}(D'_1) \in \mathbb{R}^{56\times56\times256}, \tag{S4}$$

$$S_2 = \text{ConvT}_{768\to256}(F_{\text{low}}) \in \mathbb{R}^{56\times56\times256}, \tag{S5}$$

$$D'_2 = \text{Concat}(D_2, S_2) \in \mathbb{R}^{56\times56\times512}. \tag{S6}$$

The final learned upsampling stage maps the fused feature to 112×112 pixels and 128 channels. A 1×1 convolution produces the mask logits. Bilinear interpolation restores the original 224×224 resolution, and a sigmoid function converts the logits into lesion

probabilities.

$$D_3 = \mathrm{ConvT}_{512\to128}(D_2') \in \mathbb{R}^{112\times112\times128}, \tag{S7}$$

$$M_{\mathrm{logit}} = \mathrm{Conv}_{1\times1}(D_3) \in \mathbb{R}^{112\times112\times1}, \tag{S8}$$

$$\hat{M} = \sigma\big(\mathrm{Up}_{2\times}(M_{\mathrm{logit}})\big) \in [0,1]^{224\times224}. \tag{S9}$$

Every transposed convolution uses a kernel size of 2 and a stride of 2 without additional padding. This choice doubles the spatial resolution at each decoder stage.

ITA conditioned feature modulation

We pass the normalized ITA score s through two independent multilayer perceptrons. Each network follows Linear$(1, 64)$, ReLU, and Linear$(64, 768)$. The first network predicts the channel scale $\gamma(s)$, while the second predicts the channel shift $\beta(s)$. We broadcast both vectors over the 14×14 feature grid and calculate

$$F'_{\mathrm{high}} = \gamma(s) \odot F_{\mathrm{high}} + \beta(s). \tag{S10}$$

This module adds about 0.10 million parameters to the 97.50 million parameter segmentation baseline. We apply the modulation only to the high level representation because this tensor carries the semantic information that is most sensitive to changes in lesion contrast.

Supplementary Note S2: Implementation details

Preprocessing

We remove hair artifacts before resizing the image. First, we apply morphological black hat filtering with a 17×17 cross shaped structuring element. We then use Navier Stokes inpainting with a radius of 3 pixels to replace the detected hair pixels. A 5×5 median filter suppresses residual noise. For configuration C6, we convert the image to CIELAB and apply CLAHE only to the luminance channel. We use a clip limit of 2.0 and an 8×8 tile grid. Configurations C1 to C5 omit CLAHE so that the ablation separates its effect from the FiLM contribution.

ROI generation and padding selection

We threshold the predicted lesion probability map at 0.5 and retain the largest connected component. Next, we calculate its bounding box and expand each side by 25 percent. We convert the expanded region to a square crop, clip it to the image boundaries, and resize it to 224×224 pixels. We selected the padding value on the validation set before evaluating the test set. The 25 percent setting produced the best validation result. We therefore fixed it before test set evaluation.

Table S2 ROI padding comparison on the HAM10000 validation split.

Padding	Macro F1 difference	Observed limitation
0%	−2.1 percentage points	Excluded useful perilesional structure
25% selected	Reference	Balanced lesion and context coverage
40%	−1.4 percentage points	Included more background artifacts

Table S3 Segmentation network training configuration.

Parameter	Value
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Loss	Dice plus binary cross entropy
Batch size	8
Precision	Mixed precision
Early stopping	Patience 10, maximum 50 epochs
Input resolution	224×224
Encoder initialization	ImageNet pretrained ViT-B/16

Supplementary Note S3: Training and reproducibility settings

We trained the segmentation and classification networks separately. We selected checkpoints using validation performance and applied early stopping to limit overfitting. Tables S3 and S4 report the settings used for the final experiments.

All classification configurations use the same offline augmented training pool. Therefore, changes across C1 to C6 reflect the interventions assigned to each configuration rather than different augmented samples.

Supplementary Note S4: Robustness and auxiliary analyses

Skin tone threshold sensitivity

The main analysis uses $\tau = 0.73$, which corresponds to a raw ITA value of 41.4 degrees. We repeated the evaluation across nearby thresholds and also included the Otsu threshold. The model retained its qualitative ranking across the tested range. However, the absolute fairness gaps changed as the subgroup sizes changed.

Table S4 Classification network training configuration.

Parameter	**Value**
Optimizer	AdamW with differential learning rates
Learning rate	2×10^{-6} early layers, 2×10^{-5} late layers, 2×10^{-4} head
Weight decay	0.05 backbone, 0.01 head
Warmup and schedule	10 epochs followed by cosine decay
Batch size	24, effective 48 with gradient accumulation
Gradient clipping	5.0
Early stopping	Patience 20 using validation macro F1
Loss for C6	Focal loss with $\gamma = 2$ and inverse frequency weights
Loss for C1 to C5	Cross entropy with label smoothing 0.05
Mixup	$\alpha = 0.8$
CutMix	$\alpha = 1.0$ with probability 0.60
Tone aware sampling	WeightedRandomSampler with threefold dark group weight
Offline augmentation	Elastic deformation, HSV shifts, and Gaussian noise with up to 25 copies per image

Table S5 Threshold sensitivity for FSD-Net C6 on HAM10000.

τ	**Dark proportion**	**Light accuracy**	**Dark accuracy**	**EO gap**
0.65	3.0%	0.842	0.848	0.153
0.68	5.3%	0.838	0.912	0.044
0.70	6.6%	0.838	0.900	0.040
0.73 selected	8.9%	0.835	0.911	0.025
0.75	11.7%	0.831	0.927	0.009
0.834 Otsu	28.2%	0.816	0.909	0.098

Supplementary Note S5: Computational analysis

We measured single image inference using mixed precision on an NVIDIA RTX 2070 Super Max-Q with 8 GB of memory. FiLM adds about 0.10 million parameters to the segmentation network. The displayed GFLOP count remains unchanged because the two small conditioning networks contribute little at the reported precision. The full pipeline includes preprocessing, segmentation, ROI generation, and classification.

Supplementary Note S6: Statistical analysis

Bootstrap confidence intervals

We used 1,000 bootstrap resamples to estimate uncertainty on the patient disjoint HAM10000 test set. The observed estimate comes from the original test set. The

Table S6 Computational cost breakdown.

Component	Parameters (M)	GFLOPs	Latency (ms)	Memory (MB)
Baseline ViT-U-Net	97.50	61.53	294	534
ViT-U-Net with FiLM	97.60	61.53	322	908
Swin-V2-Tiny classifier	27.88	8.76	94	889
EfficientNet-B2 reference	8.25	2.01	40	1,036
Complete FSD-Net	125.48	70.29	441	1,535

bootstrap mean averages the metric across the resampled test sets. These quantities can differ, especially for fairness metrics calculated from the smaller dark group. The

Table S7 Bootstrap results for FSD-Net C6 on HAM10000.

Metric	Bootstrap mean	95% confidence interval	n
Overall accuracy	0.8422	[0.8234, 0.8611]	1,512
Macro F1	0.7379	[0.6957, 0.7769]	1,512
Balanced accuracy	0.7200	[0.6715, 0.7645]	1,512
Light accuracy	0.8351	[0.8148, 0.8533]	1,377
Dark accuracy	0.9134	[0.8667, 0.9556]	135
Performance gap	0.0751	[0.0234, 0.1236]	–
EO gap	0.0917	[0.0022, 0.2531]	–
EOD difference	0.0937	[0.0159, 0.2531]	–

dark group interval is wider than the light group interval because the dark group contains 135 images, while the light group contains 1,377 images. The EO gap also has a wide interval because it combines class specific true positive rates estimated from the smaller group.

McNemar pairwise tests

We compared the six trained configurations through all 15 pairwise McNemar tests on the same 1,512 test images. We applied the continuity correction and used a Bonferroni threshold of 0.0033. None of the comparisons crossed the corrected threshold.

Table S8: McNemar pairwise tests on the HAM10000 test set.

Comparison	b	c	χ^2	p	Significant at 0.0033
C1 vs C2	72	64	0.360	0.548	No
C1 vs C3	51	58	0.330	0.566	No
C1 vs C4	83	57	4.464	0.035	No

Comparison	b	c	χ^2	p	Significant at 0.0033
C1 vs C5	67	74	0.255	0.613	No
C1 vs C6	84	76	0.306	0.580	No
C2 vs C3	66	81	1.333	0.248	No
C2 vs C4	63	45	2.676	0.102	No
C2 vs C5	49	64	1.735	0.188	No
C2 vs C6	63	63	0.008	0.929	No
C3 vs C4	98	65	6.282	0.012	No
C3 vs C5	66	66	0.008	0.931	No
C3 vs C6	82	67	1.315	0.251	No
C4 vs C5	46	79	8.192	0.004	No
C4 vs C6	58	76	2.157	0.142	No
C5 vs C6	56	41	2.021	0.155	No

The C4 versus C5 comparison produced the smallest p value at 0.004, although it remained above the corrected threshold. We therefore treat it as suggestive rather than confirmatory evidence. The baseline versus proposed comparison gave $\chi^2 = 0.306$ and $p = 0.580$. This result indicates that the paired classification errors did not differ significantly, even though the EO gap decreased from 0.0512 to 0.0249.

Supplementary Note S7: Additional diagnostic figures

The following figures provide the uncertainty and error analyses cited in the main manuscript.

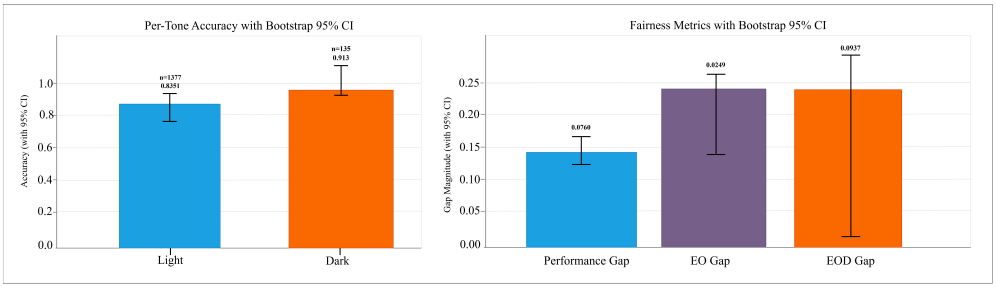


Fig. S1 Per tone accuracy and fairness metrics with bootstrap 95 percent confidence intervals. The dark group interval is wider because it contains 135 samples, while the light group contains 1,377 samples. Supplementary Table S7 reports the numerical bounds.

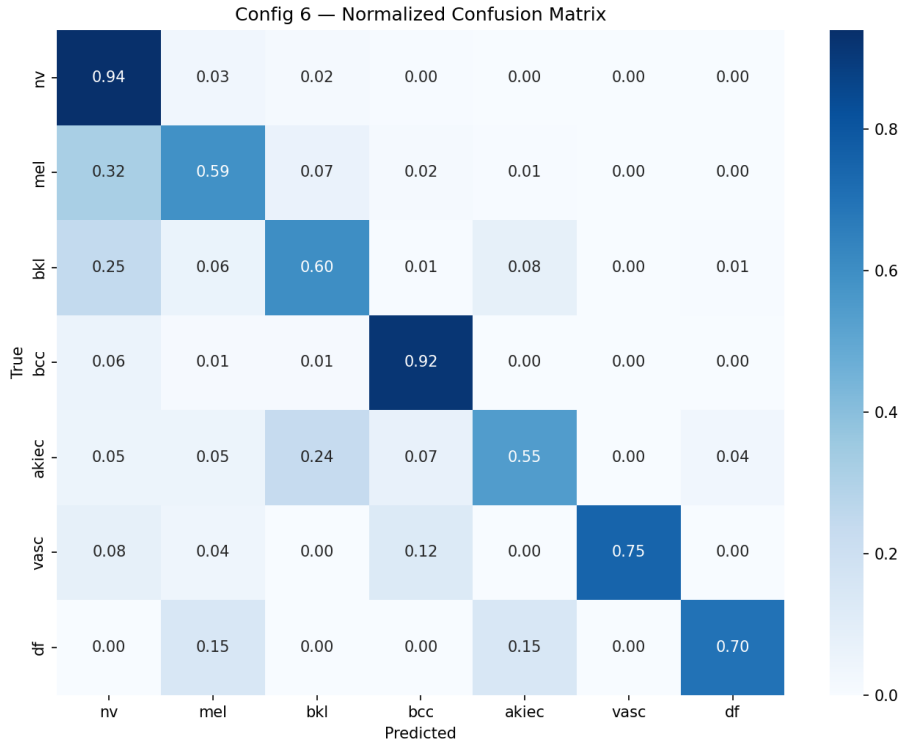


Fig. S2 Confusion matrix for FSD-Net C6 on the patient disjoint HAM10000 test set with 1,512 images. The displayed cells show row-normalized class proportions. Errors occur most often between melanoma and melanocytic nevus and between benign keratosis like lesions and melanocytic nevus.

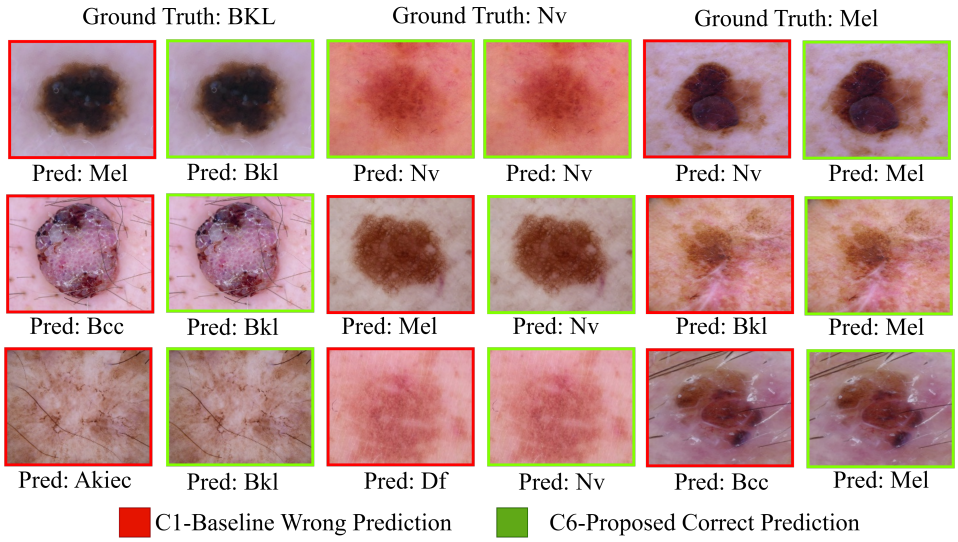


Fig. S3 Representative qualitative classification results showing corrected baseline errors, agreements, and failure cases across tone groups.

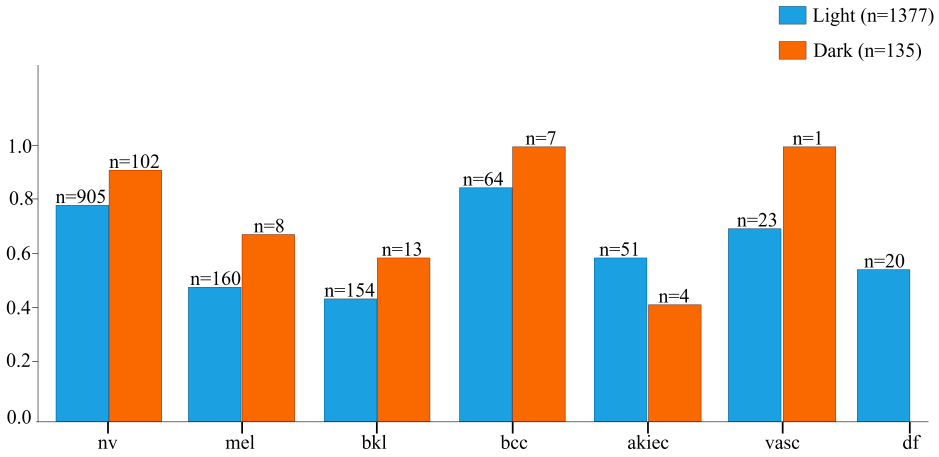


Fig. S4 Per-class accuracy stratified by skin tone on the HAM10000 test set. Numbers above bars indicate sample count per group. The dark subgroup is dominated by nv (102/135, 75.6%), the highest-accuracy class, which partially explains the higher aggregate dark accuracy.