Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Ubaid Khan
National University of Computer and Emerging Sciences
Usman Habib
National University of Computer and Emerging Sciences
Sheheryar Mohsin Qureshi
sheheryar.qureshi@uws.ac.uk

University of the West of Scotland
Muhammad Ehtisham Hassan
Air University
Adeel Ashraf Cheema
National University of Computer and Emerging Sciences
Sameed Tariq
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology

Article

Keywords: Fairness-Aware Skin Cancer, Skin Cancer Classification and Segmentation, FSD-Net, Skin Tone, Dermoscopy

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11319017/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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 & Aviation Campus, Kamra, 43570, Pakistan.
[5]Department of AI & 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;
Contributing authors: i248216@isb.nu.edu.pk; usman.habib@nu.edu.pk; ehtisham.eave@aack.au.edu.pk; Adeel.cheema@nu.edu.pk; sameedtariq@gmail.com;

Abstract

Deep learning models for skin lesion segmentation and classification achieve strong aggregate accuracy on dermoscopic benchmarks but generalise poorly across skin tones owing to the underrepresentation of darker Fitzpatrick types in public training datasets. Existing fairness-aware methods address this disparity exclusively at the classification stage, implicitly assuming that upstream segmentation is unbiased; an assumption increasingly contradicted by the literature. We propose FSD-Net (Fairness-aware Skin Disease Network), which to our knowledge is the first framework to enforce and evaluate fairness simultaneously at both the

segmentation and classification stages. FSD-Net couples a ViT-B/16 segmentation backbone conditioned on a continuous skin tone score through Feature-wise Linear Modulation with a Swin-V2-Tiny classifier operating on segmentation-guided region-of-interest crops. Evaluated on ISIC-2017, ISIC-2018, PH2, and the patient-disjoint HAM10000 benchmark, FSD-Net achieves Dice coefficients of 0.851, 0.892, and 0.927 on the three segmentation datasets and 84.2% classification accuracy with an Equalized Opportunity gap of 0.025, reducing to 0.004 with test-time augmentation. A six-configuration ablation study attributes a 3.68 percentage-point Dice improvement on ISIC-2017 directly to the FiLM module. Bootstrap confidence intervals and McNemar pairwise tests confirm that fairness improvements are statistically reliable.

Keywords: Fairness-Aware Skin Cancer, Skin Cancer Classification and Segmentation, FSD-Net, Skin Tone, Dermoscopy

1 Introduction

Skin cancer is one of the fastest-growing cancer types worldwide, and early detection plays a critical role in clinical outcomes: lesions detected at an early stage are generally more treatable, while late-stage disease demands aggressive systemic therapy [1–3]. Dermoscopy allows non-invasive observation of subsurface pigmentation patterns that are not detectable by the naked eye [4]. However, its interpretation relies heavily on practitioner expertise and remains subject to inter-observer variability, motivating the development of automated diagnostic assistance [5].

Deep learning has been increasingly applied to automated skin lesion segmentation and classification [6, 7]. Automated skin lesion segmentation and classification have attracted significant research interest, with recent approaches employing convolutional neural networks (CNNs), transformers, and foundation models [8–14] that have made meaningful improvements. Automated dermoscopic analysis has come a long way, tackling some issues that were left unanswered by previous techniques. In what follows, we summarize key research directions—segmentation-guided diagnostic pipelines, demographic fairness and fairness-aware classification, and unified segmentation–classification pipelines—while highlighting the shortcomings that motivate FSD-Net. We limit ourselves to describing some of the limitations of current approaches rather than providing a comprehensive survey.

These gains, however, are not uniformly distributed. Most public dermoscopic datasets contain mainly lighter skin types (Fitzpatrick I–IV) and the darker ones cover less than 10% of samples [15, 16]. Models that are trained on such data reflect this imbalance. Daneshjou *et al.* [17] reported AUC disparities exceeding fifteen percentage points between light and dark Fitzpatrick groups, diminishing diagnostic reliability for populations that already face worse outcomes from delayed presentation and reduced specialist access [18]. Fig. 1 illustrates this disparity and the gap addressed by the proposed framework.

This bias manifests differently at each stage: in segmentation, reduced lesion-to-skin contrast in darker images makes boundary delineation intrinsically harder.

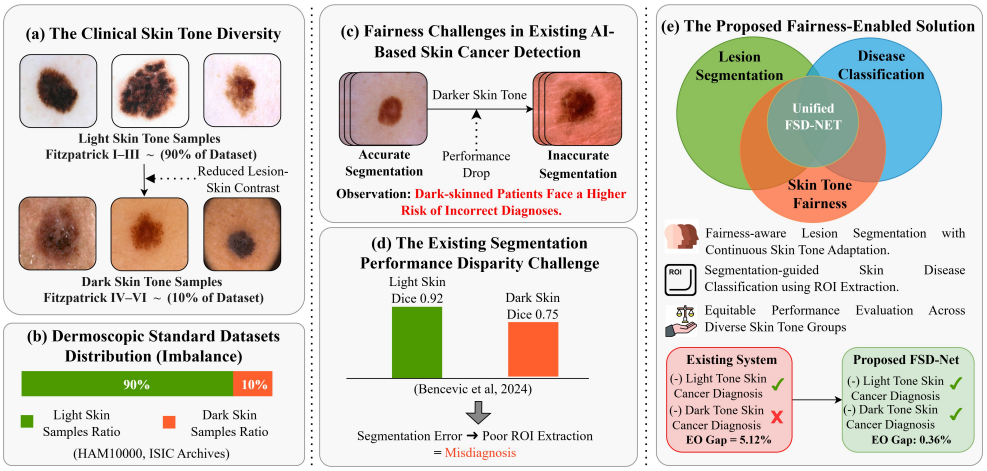


Fig. 1 Motivation and overview of the proposed FSD-Net framework. (a) Skin-tone diversity. (b) Dataset imbalance in public dermoscopic archives (90% light vs. 10% dark Fitzpatrick tones) [15, 16]. (c) Segmentation accuracy degrades on darker tones, raising misdiagnosis risk. (d) Quantitatively, Dice falls from 0.92 (light) to 0.75 (dark) for a representative baseline [19]. (e) FSD-Net unifies fairness-aware segmentation, classification, and skin-tone-stratified evaluation, cutting the Equalized Opportunity gap from 5.12% to 0.36% on patient-disjoint HAM10000.

Bencević *et al.* [19] showed that this degradation resists augmentation, reweighting, and balanced sampling. Their follow-up study [20] attributed much of this degradation to reduced lesion–skin contrast rather than skin tone alone. For classification, Corbin and Marques [21] identified systematic differences in model attention and prediction confidence across skin tones. Shah *et al.* [22] highlighted limitations of discrete Fitzpatrick categories [23], motivating continuous skin-tone representations such as the ITA-based formulation adopted in this work.

Existing fairness interventions target almost exclusively the classification stage: reweighting and sampling methods adjust loss contributions by demographic distribution [24, 25]; alignment-based methods enforce representational consistency across tone groups [26]; feature fusion and group-specific expert architectures address per-group performance directly [27, 28]. All of these methods [24–28] assume that the lesion representation delivered to the classifier is unbiased. Segmentation yields poorer masks on darker tones; however, every downstream operation inherits a degraded substrate [21, 22, 25]. Classification corrections can compensate partially, but cannot recover information lost during segmentation. Imran *et al.* [29] demonstrated empirically that lesion isolation before classification reduces background-induced tone bias, reinforcing that fairness interventions must reach the segmentation stage.

There has been a growing research effort that has suggested various approaches to classification that are designed to minimize disparities in performance across demographic groups. Moreover, many fairness-aware classifiers in the literature have been

proposed to address the issue of disparities. These methods include representation-based and augmentation-based approaches, such as color-invariant representation learning [30], debiasing [31], EdgeMixup [32], reweighting and sampling [24], feature fusion [27], optimal-transport alignment [26], group-specific experts [28], and diffusion-based augmentation [33]. Collectively, these approaches enhance fairness during the classification phase while maintaining diagnostic accuracy. More recently, Paxton *et al.* [25] presented two fairness strategies relevant to this work: distribution-aware reweighting (DRW), estimating continuous sample weights from individual typology angle (ITA) values using kernel density estimation; and SkewPrune, which prunes away feature channels that are most skewed by skin tone through skewness-based pruning [34].

Joint segmentation–classification frameworks have meanwhile emerged as the dominant diagnostic paradigm. Abu Owida *et al.* [35] proposed a dual-branch design using DenseNet-121 features with a morphological mask branch, while Deepa and Madhavan [36] coupled a Transformer-U-Net segmentation backbone with a CNN classifier. Binzagr and Hariri [14] currently represent the state of the art with a SAM-Adapter segmentation module and Vision Transformer classifier, achieving a Dice score of 94.27% on ISIC 2018 and an accuracy of 96.37% on HAM10000. However, none of these frameworks evaluate equitable diagnostic performance across skin tones.

Automated skin lesion segmentation has seen significant progress in deep learning, from CNN-based encoder–decoder models to transformer-based [37] and foundation models. The basic idea of dense medical image segmentation was set by U-Net [38] and then enhanced by adding attention mechanisms [39], adversarial learning, multi-scale feature fusion, and boundary-aware refinement to better delineate lesions [40–42]. More recent transformer-based approaches have further integrated global contextual modeling via self-attention and hierarchical feature representations [43]. Foundation models such as the Segment Anything Model (SAM) [13] reach state-of-the-art performance after domain-specific fine-tuning [14]. These advances, in spite of most current methods, are assessed by only overall scores, like Dice and Intersection over Union (IoU), and seldom by demographic subgroups.

Growing evidence shows that strong aggregate performance in dermatology AI can conceal substantial disparities across skin tones. Early work by Kinyanjui *et al.* [44] highlighted the severe underrepresentation of darker skin tones in ISIC datasets, while the release of Fitzpatrick17k [16] enabled standardized fairness evaluation across the full Fitzpatrick scale. Using this dataset, subsequent studies demonstrated clinically significant performance degradation on darker skin tones across both commercial and research models [17], reinforcing concerns that biased AI may exacerbate existing healthcare inequalities in dermatology [18]. Bencević *et al.* [19] investigated the origin of these disparities further, reporting statistically significant segmentation performance degradation on darker skin tones and showing that conventional mitigation strategies such as augmentation, reweighting, and balanced sampling were insufficient to eliminate the gap.

This defines the gap we address: segmentation-stage bias exists and propagates, fairness-aware methods target only downstream stages, and joint pipelines achieve

strong aggregate accuracy while omitting demographic disaggregation. To our knowledge, no published framework simultaneously enforces fairness at the segmentation stage, propagates fair representations to a classification module, and reports skin-tone-stratified evaluation across both stages. We introduce FSD-Net (Fairness-Aware Skin Disease Network) to close this gap; our main contributions are summarized as follows:

- We propose FSD-Net, a unified framework that integrates continuous ITA-conditioned FiLM segmentation with ROI-guided classification to improve skin-tone fairness.
- We introduce a six-configuration ablation that isolates the contribution of FiLM and demonstrates measurable improvements in segmentation fairness while preserving classification performance.
- We provide a comprehensive evaluation of segmentation, classification, cross-dataset generalization, and skin-tone fairness under patient-disjoint protocols, complemented by bootstrap confidence intervals and McNemar's test to quantify statistical uncertainty and assess the robustness of the observed differences.

Table 1 summarizes representative prior work across the four dimensions relevant to this study and the research gap addressed by our framework.

2 Results

2.1 Overview

We evaluate FSD-Net on segmentation, classification, fairness, and cross-dataset generalization using ISIC-2017 [7], ISIC-2018 [46], PH2 [47], and HAM10000 [6]. Segmentation is trained on the combined ISIC-2017 and ISIC-2018 [7, 46] training sets and evaluated on their respective test sets, with PH2 [47] for external validation. Classification is evaluated using the patient-disjoint HAM10000 split [6]. We compare FSD-Net with state-of-the-art methods and perform a six-configuration ablation study.

2.2 Datasets and Skin Tone Stratification

We use ISIC-2017 [7] and ISIC-2018 [46] for lesion segmentation training, with ISIC-2018 being one of the most widely adopted dermoscopic segmentation benchmarks [14, 41]. We evaluate segmentation generalization on two datasets not used for the training PH2 [47] and HAM10000 [6] using dermatologist-verified segmentation masks released by Tschandl *et al.* [48] for the latter. For HAM10000 [6], we report the segmentation results on the same patient-disjoint test partition used for classification, avoiding any lesion overlap between the images used to train the segmentation network and those used to evaluate it.

2.2.1 Patient-Disjoint Splitting Protocol

HAM10000 [6] contains 10,015 images derived from approximately 7,000 unique lesions identified through the `lesion_id` field, as multiple images of the same lesion are captured across follow-up visits. Random splitting therefore allows the same lesion to appear in both training and test, increasing the reported accuracy by approximately

Table 1 Comparative analysis of representative prior work across four dimensions relevant to this study. ✓ indicates that the dimension is addressed; × indicates that it is not.

Study	Venue/Yr	Focus	Segmentation	Classification	Fairness Eval.	Cross-Dataset
Xin et al. [45]	CBM, 2024	Segmentation only	✓	×	×	✓
Bencević et al.[19]	CMPB, 2024	Segmentation fairness	✓	×	✓	×
Ansari et al. [24]	MICCAI, 2024	Classification fairness	×	✓	✓	×
Corbin & Marques [21]	IEEE Access, 2023	Classification fairness	×	✓	✓	×
Abdulredah et al. [27]	BMC, 2025	Classification fairness	×	✓	✓	×
Paxton et al. [34]	IEEE JBHI, 2026	SkewPrune pruning	×	✓	✓	×
Deepa & Madhavan [36]	Sci. Rep., 2025	Joint pipeline	✓	✓	×	✓
Binzagr & Hariri [14]	Diagnostics, 2026	Joint pipeline (SAM)	✓	✓	×	✓
Imran et al. [29]	Sci. Rep., 2026	Joint with bias mitigation	✓	✓	✓	×
FSD-Net (Ours)	**This work**	**Unified fairness pipeline**	✓	✓	✓	✓

six to ten percentage points. We partition by `lesion_id` to produce 6,982 training, 1,521 validation, and 1,512 test images with no lesion crossing splits, which partially explains why our accuracy is lower than published values that use random splits.

2.2.2 Skin Tone Stratification

We quantify skin tone using the Individual Typology Angle (ITA) computed in CIELAB color space from non-lesion skin pixels and normalize it to $[0, 1]$. The continuous score s is used as a per-image conditioning signal during training without discretization. For evaluation, we stratify ITA into three groups for segmentation (light $s \geq 0.65$, medium $0.33 \leq s < 0.65$, dark $s < 0.33$), following prior fairness studies [19, 44]. For classification on HAM10000, we adopt a binary threshold of $\tau = 0.73$ (raw ITA 41.4°), corresponding to the Fitzpatrick II/III clinical boundary [49–51]. This partition yields 1,377 light and 135 dark test samples, motivating the bootstrap confidence intervals reported in Section 2.5.7.

Table 2 Datasets used in this study and their respective roles in segmentation and classification experiments.

Dataset	Instances	Modality	Classes	Role
ISIC-2017 [7]	2,750	Dermoscopic	3	Segmentation training and evaluation
ISIC-2018 [46]	3,694	Dermoscopic	7	Segmentation training and evaluation
PH2 [47]	200	Dermoscopic	3	External segmentation evaluation
HAM10000 [6]	10,015	Dermoscopic	7	Classification training and evaluation

2.3 Training Configuration

Both networks are trained with Adaptive Moment Estimation with Decoupled Weight Decay (AdamW) optimizer under the standard schedules with early stopping. The training configurations are given in Supplementary Tables (S3, S4).

2.4 Evaluation Metrics

We evaluate segmentation using Dice, IoU, sensitivity, and specificity, and classification using accuracy, macro F1, macro AUC, and balanced accuracy. Fairness is assessed through per-tone accuracy, performance gap, worst-group accuracy, equalized opportunity (EO) gap, and equalized odds difference (EOD) [52]. Statistical reliability is evaluated using 1,000-sample bootstrap 95% confidence intervals [53] and McNemar's test [54] with continuity correction. We also report the computational cost in terms of parameters, GFLOPs, inference latency, and peak GPU memory.

2.5 Performance Evaluation of the Proposed FSD-Net

We combine an ITA-conditioned FiLM module in a ViT-UNet segmentation backbone with CLAHE preprocessing, focal loss, tone-aware sampling, and Mixup/CutMix regularization in a Swin-V2-Tiny classifier to form the proposed FSD-Net configuration (C6). We report its performance across all four benchmarks in Sections 2.5.1 through 2.5.7, and defer comparative positioning and ablation decomposition to Sections 2.6 and 2.7, respectively.

2.5.1 Segmentation Performance

We train FSD-Net on ISIC-2017 [7] and ISIC-2018 [46] and evaluate on PH2 [47], HAM10000 [6]. As summarized in Table 3, FSD-Net demonstrates strong cross-dataset generalization on the unseen PH2 [47] and HAM10000 [6] sets, attributable to the ITA-conditioned feature modulation adapting high-level representations to skin tone while preserving the encoder's semantic features. Qualitative examples in Fig. 2 show accurate boundary delineation across all skin-tone groups, including low-contrast dark-tone lesions.

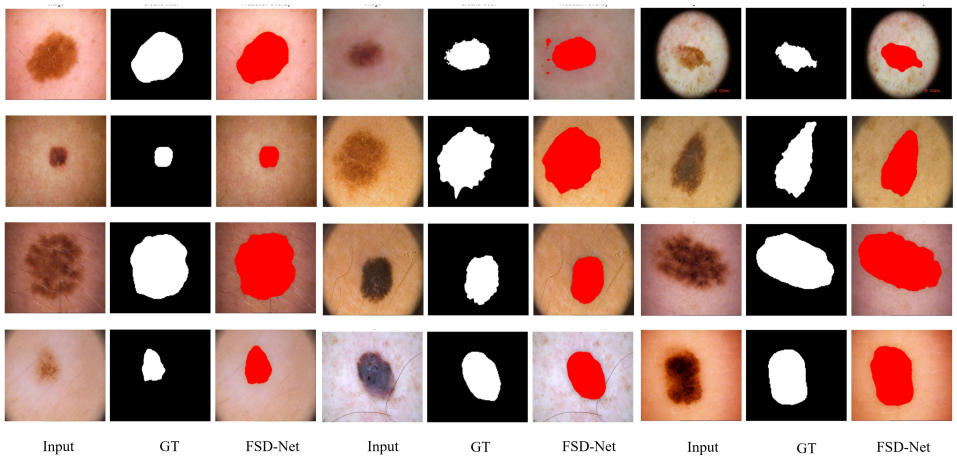


Fig. 2 Qualitative segmentation results across skin tones. Each row shows the original dermoscopic image, the ground-truth mask, and the predicted mask.

2.5.2 Per-Tone Segmentation Fairness

To evaluate fairness, we stratify segmentation performance by skin tone in Table 3. FSD-Net's dark-tone Dice exceeds its light-tone Dice on ISIC-2017, consistent with the ITA-conditioned modulation compensating for reduced lesion-to-skin contrast; ISIC-2018 remains the most challenging dataset for dark-tone samples, likely reflecting limited representation and greater imaging variability. PH2 performance is comparatively consistent across tone groups.

Table 3 Cross-dataset segmentation performance and skin-tone stratified fairness of the proposed FSD-Net (C6). The fairness gap is defined as the difference between the maximum and minimum group Dice scores. Dashes indicate unavailable stratified results.

Dataset	Overall Segmentation Performance				Skin-Tone Based Performance				
	Dice	IoU	Sensitivity	Specificity	Light	Medium	Dark	Worst	Gap
HAM10000 [6]	0.9028	0.8372	0.8681	0.9798	–	–	–	–	–
ISIC-2017 [7]	0.8512	0.7640	0.8024	0.9850	0.8464	0.8600	0.9251	0.8464	0.0786
ISIC-2018 [46]	0.8920	0.8218	0.9262	0.9462	0.9018	0.8632	0.7529	0.7529	0.1488
PH2 [47]	0.9272	0.8710	0.9581	0.9554	0.9241	0.9357	0.8497	0.8497	0.0860

2.5.3 Classification and Fairness Performance

We evaluated FSD-Net on the patient-disjoint HAM10000 [6] test set, with results summarized in Table 6. Test-time augmentation improves accuracy, macro AUC, and reduces the EO Gap to 0.0036, the strongest fairness profile under patient-disjoint evaluation. The Dark-tone accuracy exceeds the light-tone accuracy, a difference driven

by class composition rather than systematic bias, discussed further in Section 4. Supplementary Figure (S4) confirms that the per-class performance remains consistent across skin tones once class composition is controlled.

2.5.4 Qualitative Classification Analysis

Supplementary Figure (S3) shows representative cases in which the proposed method produces the correct prediction while the baseline fails, as well as cases in which both configurations agree. The comparison highlights darker-tone examples where the baseline misclassifies melanoma (*mel*) and benign keratosis-like lesions (*bkl*) as melanocytic nevi (*nv*). In these examples, FSD-Net produces more clearly delineated lesion ROIs, providing the downstream classifier with lesion-focused visual information that supports correct discrimination among these categories.

2.5.5 Error Analysis and Confusion Matrix

Of the 1,512 patient-disjoint test images, 239 (15.8%) are misclassified. The aggregate error rate is higher for light (16.5%) than dark (8.9%) skin tones; however, as discussed earlier, this difference reflects class composition rather than systematic per-class bias. The full normalized confusion matrix is provided in Supplementary Figure (S2), which further shows that errors are concentrated between the `mel`–`nv` and `bkl`–`nv` classes.

2.5.6 Computational Efficiency

We achieve an end-to-end inference time of 441 ms with 125.48 M total parameters (97.60 M segmentation and 27.88 M classification). The baseline ViT-U-Net contains 97.50 M parameters, while incorporating FiLM increases the segmentation model size to 97.60 M, corresponding to an additional 0.10 M parameters (approximately 0.1%). The reported GFLOPs remain unchanged at the displayed precision, while the measured latency and peak memory increase by 9.5% and 70%, respectively Supplementary Table (S6). These increases are primarily associated with the conditioning operations and intermediate activation storage introduced by FiLM. The full pipeline remains lighter than the SAM-ViT-based approach [14], indicating comparatively lower model complexity for deployment on commodity GPU hardware.

2.5.7 Bootstrap Confidence Intervals

Given the limited dark subgroup size (n=135), we accompany every fairness metric with bootstrap 95% confidence intervals computed from 1,000 resamples. The complete interval data are given in the Supplementary Table (S7). As expected, the dark-tone interval (n=135) is markedly wider than the light-tone interval (n=1,377), and the EO Gap's point estimate of 0.025 remains substantially below published HAM10000 prior studies [17, 25]. Despite this uncertainty, Supplementary Figure S1 visualizes the per-tone accuracy distributions in the supplementary material.

Table 4 State-of-the-art comparison on HAM10000 [6]. PD = patient-disjoint; R = random splits. EO marked † uses TTA.

Method	Year	Architecture	Split	Acc. (%)	F1	EO Gap	Fairness
SAM-ViT [14]	2026	SAM-Adapter + ViT-B/16	R	96.37	0.958	-	No
VGG19-ViT hybrid [55]	2025	VGG19 + ViT	R	97.10	0.970	-	No
Fuzzy Ensemble [56]	2025	Xception + IncRes + MobNet	R	95.14	-	-	No
Paxton DRW [25]	2025	ResNet50 + DRW	R	∼83	-	Yes	Yes
Imran et al. [29]	2026	CNN-ViT + unsup. seg.	R	81.0	-	-	Partial
Paxton SkewPrune [34]	2026	VGG11 + SkewPrune	R	79.0	0.650	0.0624	Yes
FSD-Net (Ours)	**2026**	**ViT-UNet + FM & Swin-V2**	**PD**	**84.2**	**0.7394**	**0.025 (0.004†)**	**Yes (seg+cls)**

2.6 Comparative Analysis

We contextualize FSD-Net against published state-of-the-art methods on HAM10000; the effect of input type and the robustness of our fairness conclusions under threshold variation are examined as part of the ablation study in Section 2.7.

2.6.1 State-of-the-Art Comparison

We compare FSD-Net with representative methods on HAM10000 [6], with the results summarized in Table 4. FSD-Net achieves (84.2%) classification accuracy among methods evaluated under patient-disjoint or fairness-aware protocols, while also reporting the lowest EO gap (0.025). Compared with Paxton DRW [25] and SkewPrune [34], we improve the classification accuracy and extend the fairness intervention beyond classification by incorporating fairness-aware lesion segmentation. Although SAM-ViT reports 96.4% accuracy [14], its results are obtained using random splits without fairness evaluation.

2.7 Ablation Study

To quantify each component's contribution, we perform a six-configuration ablation that progressively introduces data augmentation, tone-aware sampling, ITA-conditioned FiLM, CLAHE, and focal loss. Configuration C5 isolates the FiLM's effect while holding all other interventions fixed, and C6 is the complete FSD-Net. All configurations use their corresponding segmentation model for ROI extraction and the same offline-balanced training set.

2.7.1 Segmentation Ablation

We observe that the effects of the data-level interventions in C1-C4 vary across datasets and configurations. Moreover, when we introduce FiLM in the transition from C4 to C5 while retaining the preceding configuration components, the Dice score on ISIC-2017 increases from 0.8272 to 0.8640, corresponding to a 3.68 percentage-point improvement. With the proposed C6 configuration, we observe marginally lower aggregate Dice than C5, but more stable dark-tone performance on ISIC-2018. We summarize the segmentation ablation results in Table 5.

Table 5 Complete segmentation ablation (Dice / IoU).

Method	**ISIC-2017**		**ISIC-2018**		**PH2**	
	Dice	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
Baseline (C1)	0.8348	0.7864	0.8873	0.8138	0.9194	0.8579
+Aug (C2)	0.8505	0.7643	0.8916	0.8215	0.9206	0.8598
+Sampling (C3)	0.8619	0.7836	0.8921	0.8191	0.9214	0.8604
+Aug+Samp (C4)	0.8272	0.7349	0.8907	0.8243	0.9300	0.8754
+FiLM Iso. (C5)	**0.8640**	0.7804	**0.8997**	**0.8348**	0.9293	0.8748
Proposed (C6)	0.8512	0.7640	0.8920	0.8218	**0.9272**	0.8710

2.7.2 Classification Ablation

We find that aggregate accuracy remains relatively stable across all six configurations including the additional C5+TTA and C6+TTA rows which report test-time-augmentation variants of the corresponding trained models in Table 6. McNemar's testing [54] in Section 2.7.5 indicates that the observed differences in classification errors are not statistically significant after multiple comparison correction, whereas the EO Gap varies by more than an order of magnitude. Data-level interventions alone narrow this gap at a cost to accuracy; introducing FiLM recovers this accuracy while modestly widening the gap, indicating complementary fairness mechanisms at distinct pipeline stages. CLAHE and focal loss then refine the gap further, reaching its lowest value with test-time augmentation.

2.7.3 Effect of Input Type on Classification

We isolate the contribution of segmentation-guided ROI extraction by comparing three input regimes using the same classifier architecture and report the results in Table 7. We find that the full-image input yields accuracy 0.7900 and dark-tone accuracy 0.8520. Applying ground-truth ROI crops raises overall accuracy to 0.7970 and dark-tone accuracy to 0.8960. However, using the predicted ROI from our tone-conditioned

Table 6 Complete Classification Ablation (Patient-Disjoint HAM10000, n=1,512).

Method	Acc.	F1	AUC	Light	Dark	Gap	EO
Baseline (C1)	**0.8472**	0.7383	0.9611	0.8381	0.9407	0.1027	0.0512
+Aug (C2)	0.8419	0.7266	0.9675	0.8388	0.8741	0.0353	0.0412
+Sampling (C3)	0.8519	0.7417	0.9455	0.8453	0.9185	0.0732	0.0369
+Aug+Samp (C4)	0.8300	0.6867	0.9677	0.8250	0.8815	0.0565	0.0135
+FiLM Iso. (C5)	0.8519	**0.7446**	0.9562	0.8468	0.9037	0.0569	0.0337
Proposed (C6)	0.8419	0.7394	0.9575	0.8351	0.9111	0.0760	0.0249
C5+TTA	0.8479	0.7359	0.9548	0.8424	0.9037	0.0613	0.0206
C6+TTA	0.8433	0.7384	**0.9590**	0.8351	0.9259	0.0908	**0.0036**

segmentation produces accuracy 0.8419 and dark-tone accuracy 0.9111, outperforming both alternatives. The implications of these findings are discussed further in Section 4.

Table 7 Effect of classifier input type under the same Swin-V2-Tiny protocol.

Input	Accuracy	Macro F1	Macro AUC	Light Acc	Dark Acc
Full image	0.7900	0.6670	0.9220	0.7840	0.8520
Ground Truth ROI	0.7970	0.6810	0.9400	0.7870	0.8960
Predicted ROI	0.8419	0.7394	0.9575	0.8351	0.9111

2.7.4 Threshold Sensitivity Analysis

We use $\tau = 0.73$ throughout this paper, which corresponds to the Fitzpatrick II/III clinical boundary. To verify that our fairness conclusions are not artifacts of this specific choice, we evaluate fairness metrics across a range of thresholds, including Otsu's automatic threshold [57] ($\tau = 0.834$). Across $\tau \in [0.65, 0.75]$, dark-tone accuracy remains between 0.848 and 0.927 and the EO Gap stays below 0.16. Moreover, FSD-Net's ranking relative to the baseline is preserved at every tested value; full results are given in Supplementary Table (S5).

2.7.5 Statistical Significance of Ablation Comparisons

To complement the bootstrap analysis in Section 2.5.7, we perform McNemar's test with continuity correction across all fifteen pairwise comparisons on the patient-disjoint test set (n=1,512), using a Bonferroni-adjusted significance threshold of α=0.0033 [58]. The complete pairwise results are reported in Supplementary Table (S8). None of the fifteen pairwise comparisons reaches statistical significance after Bonferroni correction at the adjusted threshold of α=0.0033. The C4-C5 comparison (χ^2=8.192, p=0.004) is the closest to the adjusted significance threshold and is therefore treated as suggestive rather than confirmatory evidence for the contribution of FiLM. The segmentation-level Dice improvement reported in Section 2.7.1 provides

the corresponding segmentation-level evidence. For the Baseline–Proposed comparison (χ^2=0.306, p=0.580), McNemar's test does not indicate a statistically significant difference in paired classification errors, while the corresponding EO Gap decreases from 0.0512 to 0.0249.

3 Methods

Dermoscopic diagnosis is inherently influenced by skin tone, as reduced lesion-to-skin contrast in darker skin makes accurate boundary delineation more challenging and can propagate errors to subsequent classification. To address this problem, we propose FSD-Net, a unified fairness-aware framework that integrates skin-tone-aware lesion segmentation with downstream lesion classification in a unified inference pipeline. FSD-Net first estimates a continuous Individual Typology Angle (ITA) score from each input image to guide segmentation through ITA-conditioned Feature-wise Linear Modulation (FiLM). The predicted lesion mask is then used to generate a lesion-centered ROI for Swin-V2-Tiny classification, enabling fairness-aware learning across both stages of the diagnostic pipeline. The complete framework, including preprocessing, segmentation, ROI generation, classification, and evaluation components, is illustrated in Fig. 3.

3.1 Preprocessing

Preprocessing stabilizes downstream processing while preserving intrinsic chromaticity, which is required for reliable ITA estimation. Hair artifacts are removed by black-hat morphological filtering followed by inpainting and a median filter which suppresses residual noise. Contrast limited adaptive histogram equalization (CLAHE) is applied to the Commission Internationale de l'Éclairage $L^*a^*b^*$ color space (CIELAB) luminance channel to enhance low-contrast lesion boundaries. CLAHE is activated only in configuration 6, which represents the full proposed method. Configurations C1-C5 omit CLAHE to isolate the FiLM modulation contribution from input-level normalization effects.

3.2 Continuous Skin Tone Estimation

Skin tone is treated as a continuous image-derived ITA score rather than discrete Fitzpatrick categories [23] , which introduce artificial decision boundaries and may obscure subtle appearance variations that affect the lesion visibility [22, 25]. We quantify skin tone using the Individual Typology Angle (ITA), computed from non-lesion skin pixels in CIELAB space as

$$\mathrm{ITA} = \arctan\left(\frac{L^* - 50}{b^*}\right) \times \frac{180}{\pi}, \tag{1}$$

where L^* and b^* are the luminance and yellow–blue chromaticity components. Lesion pixels are excluded using the ground-truth mask during training or the predicted mask at inference; ITA is averaged over the remaining skin pixels and normalized to

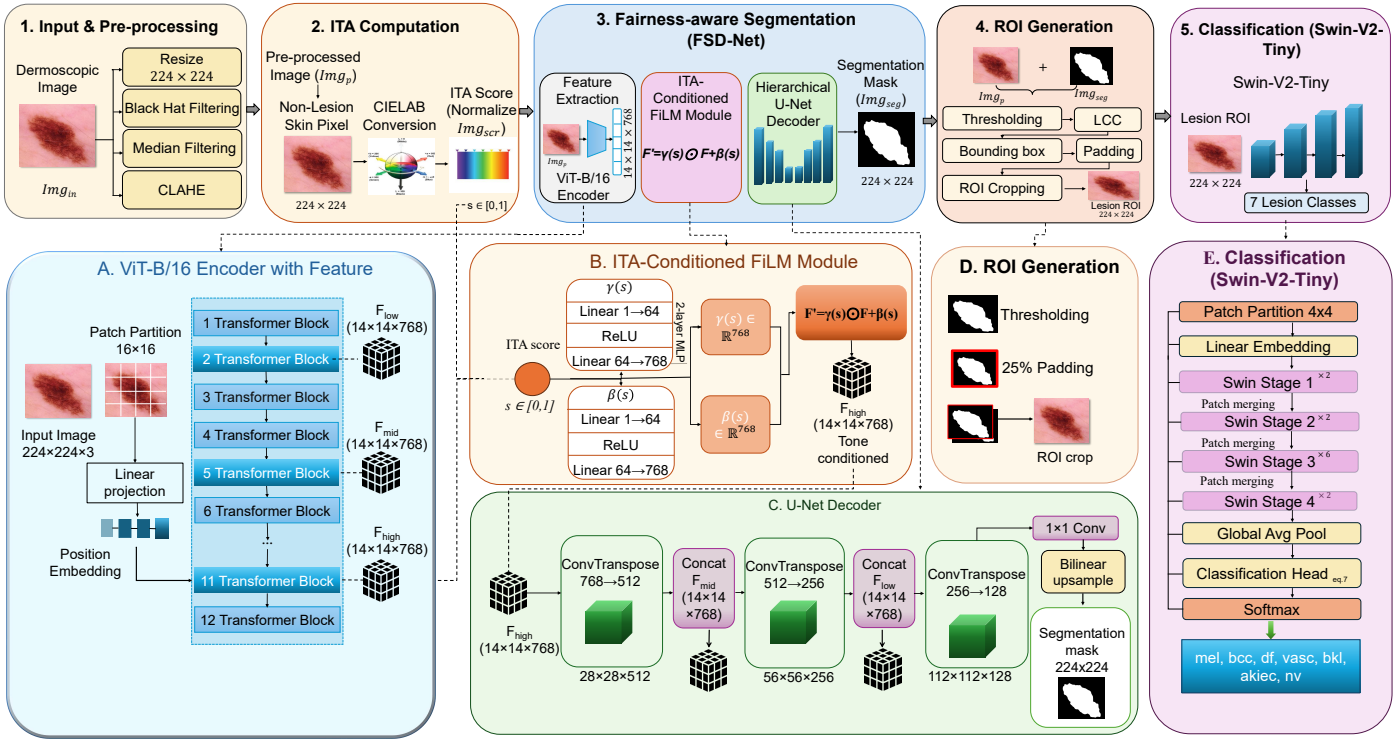


Fig. 3 Overview of the proposed FSD-Net framework. The pipeline consists of image preprocessing and continuous ITA estimation, fairness-aware lesion segmentation using a ViT-B/16 encoder with ITA-conditioned FiLM and a hierarchical U-Net decoder, segmentation-guided ROI generation, Swin-V2-Tiny classification, and evaluation using diagnostic and fairness metrics. The lower panels illustrate the detailed architectures of the ViT encoder, FiLM module, hierarchical decoder, ROI generation process, and classification network employed in the proposed framework.

a continuous score $s \in [0, 1]$ via

$$s = \frac{\text{ITA} - \text{ITA}_{\text{min}}}{\text{ITA}_{\text{max}} - \text{ITA}_{\text{min}}}, \tag{2}$$

where ITA_{min}=−90° and ITA_{max}=90°. These limits correspond to the theoretical angular range of the arctangent-based ITA formulation and provide a fixed, dataset-independent mapping of ITA values to the normalized interval $[0, 1]$. The normalized score is used directly to condition the feature modulation module and is not discretized during training or inference. Continuous conditioning preserves fine-grained skin-tone variation and avoids imposing artificial group boundaries on the feature modulation process. In contrast, discrete groups are used during evaluation to enable interpretable subgroup comparisons and the calculation of group-based fairness metrics using predefined thresholds. Segmentation performance is reported using three groups following prior work [19, 44]: light ($s \geq 0.65$), medium ($0.33 \leq s < 0.65$), and dark ($s < 0.33$). For HAM10000 [6] classification, where ∼90-91% of images are lighter-toned, we adopt a binary threshold of τ=0.73 (raw ITA=41.4°, approximately the Fitzpatrick II/III boundary [49–51]), yielding 135 dark samples in the 1,512-image test set. Threshold sensitivity is evaluated in Section 2.7.4. The score s conditions only the segmentation modulation module; the classifier receives no tone input at either training or inference time. Tone labels are used indirectly through a weighted sampler and offline augmentation during classification training, as described in Section 3.5.2.

3.3 Fairness-Aware Lesion Segmentation

3.3.1 Encoder and Decoder

We use a ViT-B/16 encoder [11] pretrained on ImageNet. Intermediate features are extracted from transformer layers 2, 5, and 11, yielding $\mathbf{F}_{\text{low}}$, $\mathbf{F}_{\text{mid}}$, and $\mathbf{F}_{\text{high}} \in \mathbb{R}^{C \times H \times W}$ (C=768), which encode progressively abstract lesion representations. A U-Net decoder with skip connections from the corresponding encoder layers upsamples the bottleneck features and produces a predicted mask $\hat{\mathbf{M}} \in [0, 1]^{H_0 \times W_0}$.

Because all three feature taps share the same 14×14 spatial resolution, the multi-scale structure is constructed within the decoder. At each decoder stage, a transposed convolution upsamples the accumulated decoder feature. In parallel, a matching transposed convolution projects the corresponding encoder feature to the same spatial resolution and channel width. The resulting feature maps are then concatenated, as illustrated in Fig. 3. The complete per-stage formulation is provided in the Supplementary Table (S1).

3.3.2 Feature-wise Linear Modulation

To mitigate skin-tone-dependent bias during lesion localization, we introduce a FiLM [59] module at the segmentation bottleneck. The module adapts the high-level semantic features $\mathbf{F}_{\text{high}}$ as a function of s:

$$\mathbf{F}'_{\text{high}} = \gamma(s) \odot \mathbf{F}_{\text{high}} + \beta(s), \tag{3}$$

where $\gamma(s), \beta(s) \in \mathbb{R}^C$ are the channel-wise scale and shift parameters generated by two multilayer perceptrons (MLPs), and $\odot$ denotes element-wise multiplication broadcast over spatial dimensions. Each MLP follows $\text{Linear}(1, 64) \rightarrow \text{ReLU} \rightarrow \text{Linear}(64, 768)$, introducing approximately 0.1 M additional parameters relative to the 97.5 M segmentation baseline. We apply modulation only to $\mathbf{F}_{\text{high}}$ because low-level features primarily capture tone-invariant edges and textures, whereas the bottleneck encodes the semantic representations where skin-tone-dependent contrast most directly influences the lesion boundary delineation. This design preserves generic ImageNet features in the early encoder, while adapting only the high-level semantic representation. We further motivate this design by the relationship between ITA and lesion-to-skin contrast. As darker skin tones exhibit lower reflectance, the lesion boundaries become less distinct, making segmentation more challenging. By supplying the continuous ITA score s to the bottleneck, we enable the network to adapt its semantic features to the expected contrast of each image. Consequently, the learned modulation enhances low-contrast boundary representations for darker images while remaining close to the identity for lighter images. The effectiveness of this design is confirmed by the C4-to-C5 transition in Table 5, where enabling FiLM while keeping all other components unchanged improves the Dice score on ISIC-2017 by 3.68 percentage points.

3.3.3 Segmentation Loss

We minimize a combined Dice and binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}, \tag{4}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{x \in \Omega} \hat{M}(x) \cdot M(x)}{\sum_{x \in \Omega} \hat{M}(x) + \sum_{x \in \Omega} M(x)}, \tag{5}$$

$$\mathcal{L}_{\text{BCE}} = -\sum_{x \in \Omega} \big[M(x) \log \hat{M}(x) + (1 - M(x)) \log(1 - \hat{M}(x)) \big], \tag{6}$$

where $\mathcal{L}_{\text{seg}}$ denotes the overall segmentation loss, $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{BCE}}$ denote the Dice and BCE loss terms, respectively, and x denotes a pixel location within the spatial domain Ω. Here, $M(x) \in \{0, 1\}$ represents the ground-truth binary mask value at pixel x, while $\hat{M}(x) \in [0, 1]$ represents the corresponding predicted lesion probability. We use the Dice term to optimize spatial overlap between the predicted and ground-truth masks, while the BCE term penalizes pixel-wise prediction errors. This combination is standard in medical image segmentation [19].

3.4 Region of Interest Extraction

We keep the threshold $\hat{\mathbf{M}}$ at 0.5 and extract the largest connected component to suppress spurious activations. A bounding box around this component is expanded by 25% on each side, squared, and resized to 224×224 pixels to form the classifier input. ROI extraction removes background confounders that drive shortcut learning in dermoscopic classifiers [29] and standardizes lesion extent across images. The 25%

padding fraction was selected empirically on the validation split prior to test-set evaluation Supplementary Table (S2). The complete architecture is illustrated in Fig. 3.

3.5 Classification Network

The classifier maps each ROI crop to a posterior over the seven HAM10000 [6] diagnostic classes.

3.5.1 Architecture

We use Swin-V2-Tiny [60] as the classification backbone. Swin-V2 extends the hierarchical shifted-window attention of Swin-V1 [12] with continuous relative position bias and residual post-normalization, improving cross-resolution generalization. The backbone produces a 768-dimensional feature vector $\mathbf{z}$, which is mapped to class logits by a custom head: Let $\mathbf{h} = \mathrm{Linear}_{768\to384}(\mathrm{LN}(\mathbf{z}))$; then

$$\hat{\mathbf{y}} = \mathrm{Softmax}\big(\mathrm{Linear}_{384\to7}(\mathrm{Dropout}_{0.4}(\mathrm{GELU}(\mathbf{h})))\big). \tag{7}$$

The complete classification model contains 27.88 M parameters. Its hierarchical window attention captures dermoscopic features at multiple scales, and its parameter count fits within the 8 GB GPU budget alongside the 97.5 M segmentation backbone.

3.5.2 Loss Function

To address the class imbalance in HAM10000 [6], where melanocytic nevus constitutes approximately 67% of images, we train the proposed configuration with focal loss [61] weighted by the inverse class frequency (Eq. 8). Configurations C1–C5 use standard cross-entropy with label smoothing to isolate this contribution in the ablation.
To mitigate the severe underrepresentation of dark-tone samples, we further apply tone-aware sampling and offline augmentation. Full augmentation and sampling parameters are given in Supplementary Tables (S3, S4); exact hyperparameters are given in Section 2.3.

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} w_c\,(1-\hat{y}_{i,c})^{\gamma}\,y_{i,c}\log\hat{y}_{i,c}, \tag{8}$$

where w_c is the per-class inverse-frequency weight and γ=2. A Label smoothing factor of 0.05 is applied throughout the experiments. Configurations C1-C5 use standard cross-entropy with label smoothing; focal loss is introduced only in C6 to isolate its contribution in the ablation.

3.6 Fairness Design Rationale

We promote fairness at the representation level rather than at the prediction level. Skin-tone conditioning is applied to the segmentation features to account for tone-related appearance variation during lesion delineation. The resulting lesion masks are

then used to generate ROI crops for the downstream classifier, without providing skin-tone information directly to the classifier. The classifier does not receive skin tone input, avoiding the complications of demographic-attribute prediction features. Tone-aware sampling and augmentation are applied during training only and introduce no tone information at inference. This design aligns with the clinical dermoscopic reasoning, which is driven by lesion morphology [62] rather than the surrounding skin tone. Moreover, it addresses the core limitation of classification-level fairness: such interventions cannot recover information lost during biased upstream segmentation. Among the fairness metrics reported in Section 2.4, we treat the equalized opportunity gap (EO Gap) as an important fairness criterion alongside diagnostic performance and the segmentation objective, since it directly measures parity in true-positive (sensitivity) rate across tone groups, which is clinically relevant for a diagnostic aid.

4 Discussion

Our results show a higher aggregate accuracy on dark-tone samples than light-tone samples, which is contrary to prior dermatology fairness studies [17, 25]. We attribute this to dataset composition rather than to bias: Supplementary Figure (S4) show that the dark subgroup is dominated by `nv` (75.6%), the easiest class to classify, while the light subgroup spans a harder mix of classes. Supplementary Table (S7) and Figure (S3) shows the dark subgroup has higher observed accuracy, but the subgroup sizes are highly unequal (135 vs. 1,377), and the dark-group uncertainty interval is substantially wider. Aggregate accuracy should therefore not be interpreted as a standalone fairness measure. We therefore treat aggregate accuracy as an unreliable fairness measure on its own, and instead take the EO Gap's reduction from baseline to test time augmentation (TTA) as stronger evidence of equity once class composition is accounted for.

This same effect helps explain why FSD-Net's 84.2% HAM10000 accuracy trails the 96% reported under random splits [14]. That gap reflects design choices rather than architectural limitations: a patient-disjoint split by `lesion_id` that eliminates known leakage [6], ROI-based rather than full-image classification [29], and tone-aware sampling with focal loss, which trades a small amount of accuracy for a lower EO gap. Among the fairness studies with patient-disjoint or fairness-aware protocols listed in Table 4, FSD-Net still reports the highest accuracy of the three (84.2% vs. 83.0 %) for Paxton DRW [25] and 79.0% for SkewPrune [34]), showing that rigorous fairness evaluation does not have come at substantial cost to accuracy.
Predicted-ROI inputs from the tone-conditioned segmentation also outperform ground-truth ROIs on dark-tone classification (91.1% vs. 89.6%), Table 7. These results suggest that the proposed modulation improves boundary localization under low contrast and produces more informative ROIs as a result. Supplementary Figure (S3) shows the method correctly classifying dark-tone `mel`/`bkl` cases where the baseline fails; further validation on independent datasets is left to future work.

5 Conclusion

FSD-Net provides a unified framework for skin-tone-aware lesion segmentation and classification by incorporating a continuous image-derived ITA representation into feature modulation and downstream prediction. Under the patient-disjoint HAM10000 evaluation, the proposed configuration achieves competitive overall classification performance while reducing the reported subgroup fairness gap, while ablation results indicate that ITA-conditioned feature modulation contributes to segmentation performance. The predicted-ROI analysis further supports the potential benefit of segmentation-guided inputs for downstream classification, although these findings should be interpreted within the limitations of the evaluated datasets and protocols. This study is limited by subgroup imbalance, sensitivity of fairness estimates to ITA threshold selection, and the absence of independent multi-center clinical validation. Future work will therefore focus on larger and more balanced datasets, threshold-robust fairness strategies, and prospective evaluation across diverse imaging conditions and clinical settings. Prospective clinical and reader-based studies will also be needed to determine whether the observed performance and fairness benefits translate into reliable clinical decision support.

Data availability

The datasets analysed in this study are publicly available from their respective repositories. The ISIC 2017 and ISIC 2018 challenge datasets are available through the International Skin Imaging Collaboration at https://challenge.isic-archive.com/data/. The HAM10000 dataset is available through the ISIC Archive at https://api.isic-archive.com/collections/66/ and Harvard Dataverse at https://doi.org/10.7910/DVN/DBW86T. The PH2 dataset is available from the PH2 Database repository at https://www.fc.up.pt/addi/ph2%20database.html. The processed data splits and additional implementation details used in this study are available from the corresponding author upon reasonable request.

References

[1] Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA: a cancer journal for clinicians* **69**, 7–34 (2019).

[2] Bray, F. *et al.* Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229–263 (2024).

[3] Saginala, K., Barsouk, A., Aluru, J. S., Rawla, P. & Barsouk, A. Epidemiology of melanoma. *Medical sciences* **9**, 63 (2021).

[4] Errichetti, E. & Stinco, G. Dermoscopy in general dermatology: a practical overview. *Dermatology and therapy* **6**, 471–507 (2016).

[5] Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *nature* **542**, 115–118 (2017).

[6] Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**, 180161 (2018).

[7] Codella, N. C. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic) (2018).

[8] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition (2016).

[9] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks (2017).

[10] Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks (2019).

[11] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale (2021). URL https://openreview.net/forum?id=YicbFdNTTy.

[12] Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows (2021).

[13] Kirillov, A. *et al.* Segment anything (2023).

[14] Binzagr, F. & Hariri, M. Foundation-model-driven skin lesion segmentation and classification using SAM-adapters and vision transformers. *Diagnostics* **16**, 468 (2026).

[15] Wen, D. *et al.* Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health* **4**, e64–e74 (2022).

[16] Groh, M. *et al.* Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset (2021).

[17] Daneshjou, R. *et al.* Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances* **8**, eabq6147 (2022).

[18] Adamson, A. S. & Smith, A. Machine learning and health care disparities in dermatology. *JAMA dermatology* **154**, 1247–1248 (2018).

[19] Benčević, M., Habijan, M., Galić, I., Babin, D. & Pižurica, A. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer methods and programs in biomedicine* **245**, 108044 (2024).

[20] Paxton, K. *et al.* Exploring the impact of skin color on skin lesion segmentation (2026).

[21] Corbin, A. & Marques, O. Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques. *IEEE Access* **11**, 78339–78352 (2023).

[22] Shah, P. *et al.* The impact of skin tone label granularity on the performance and fairness of AI based dermatology image classification models (2025).

[23] Gupta, V. & Sharma, V. K. Skin typing: Fitzpatrick grading and others. *Clinics in dermatology* **37**, 430–436 (2019).

[24] Ansari, F., Chakraborti, T. & Das, S. Algorithmic fairness in lesion classification by mitigating class imbalance and skin tone bias (2024).

[25] Paxton, K., Dehghani, Z., Aslansefat, K., Thakker, D. & Papadopoulos, Y. Mitigating individual skin tone bias in skin lesion classification through distribution-aware reweighting (2025).

[26] Aayushman, Gaddey, H., Mittal, V., Chawla, M. & Gupta, G. R. Fair and accurate skin disease image classification by alignment with clinical labels (2024). URL https://papers.miccai.org/miccai-2024/308-Paper2609.html.

[27] Abdulredah, A. A. *et al.* Towards unbiased skin cancer classification using deep feature fusion. *BMC Medical Informatics and Decision Making* **25**, 48 (2025).

[28] Xu, G. *et al.* Incorporating rather than eliminating: Achieving fairness for skin disease diagnosis through group-specific experts (2025).

[29] Imran, M. H. *et al.* Dermnet: integrative cnn-vit architecture for bias mitigation in dermatological diagnostics using advanced unsupervised lesion segmentation. *Scientific Reports* **16**, 5291 (2026).

[30] Pakzad, A., Abhishek, K. & Hamarneh, G. Circle: Color invariant representation learning for unbiased classification of skin lesions (2022).

[31] Bevan, P. J. & Atapour-Abarghouei, A. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification (2022).

[32] Yuan, H. *et al.* Edgemixup: improving fairness for skin disease classification and segmentation (2022).

[33] Munia, N. & Imran, A.-A.-Z. Dermdiff: generative diffusion model for mitigating racial biases in dermatology diagnosis (2025).

[34] Paxton, K., Aslansefat, K., Thakker, D., Papadopoulos, Y. & Maslekar, T. Enhancing fairness in skin lesion classification for medical diagnosis using prune

learning. *IEEE Journal of Biomedical and Health Informatics* **30**, 6605–6614 (2026).

[35] Owida, H. A. *et al.* A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic images. *Scientific Reports* **15**, 37823 (2025).

[36] Deepa, J. & Madhavan, P. An advanced skin lesion segmentation and classification framework using deep learning strategies. *Scientific Reports* **15**, 33926 (2025).

[37] Gomathi, S. & Arunachalam, N. Modified lstm based skin lesion segmentation and classification using optimization concept. *Discover Applied Sciences* **7**, 880 (2025).

[38] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation (2015).

[39] Söylemez, Ö. F. SkinAttn-Net: A multi-level attention-based network for skin lesion segmentation. *Scientific Reports* **16**, 3700 (2026).

[40] Amin, J., Azhar, M., Arshad, H., Zafar, A. & Kim, S.-H. Skin-lesion segmentation using boundary-aware segmentation network and classification based on a mixture of convolutional and transformer neural networks. *Frontiers in Medicine* **12**, 1524146 (2025).

[41] Wu, H. *et al.* Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical image analysis* **76**, 102327 (2022).

[42] Liu, Z., Hu, J., Gong, X. & Li, F. Skin lesion segmentation with a multiscale input fusion u-net incorporating res2-se and pyramid dilated convolution. *Scientific Reports* **15**, 7975 (2025).

[43] Xu, R. *et al.* Skinformer: Learning statistical texture representation with transformer for skin lesion segmentation. *IEEE Journal of Biomedical and Health Informatics* **28**, 6008–6018 (2024).

[44] Kinyanjui, N. M. *et al.* Fairness of classifiers across skin tones in dermatology (2020).

[45] Xin, C. *et al.* Transformer guided self-adaptive network for multi-scale skin lesion image segmentation. *Computers in biology and medicine* **169**, 107846 (2024).

[46] Codella, N. *et al.* Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019). arXiv:1902.03368.

[47] Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. & Rozeira, J. Ph 2-a dermoscopic image database for research and benchmarking (2013).

[48] Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nature medicine* **26**, 1229–1234 (2020).

[49] Wilkes, M., Wright, C. Y., du Plessis, J. L. & Reeder, A. Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology* **151**, 902–903 (2015).

[50] Del Bino, S., Sok, J., Bessac, E. & Bernerd, F. Relationship between skin response to ultraviolet exposure and skin color type. *Pigment cell research* **19**, 606–614 (2006).

[51] Eilers, S. *et al.* Accuracy of self-report in assessing fitzpatrick skin phototypes i through vi. *JAMA dermatology* **149**, 1289–1294 (2013).

[52] Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning (2016).

[53] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1993).

[54] McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).

[55] Najjar, F. H. *et al.* Transformer-aided skin cancer classification using vgg19-based feature encoding. *Scientific Reports* **15**, 40204 (2025).

[56] Halder, A. *et al.* A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification. *Scientific Reports* **15**, 6268 (2025).

[57] Otsu, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* **9**, 62–66 (1979).

[58] Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze* **8**, 3–62 (1936).

[59] Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: Visual reasoning with a general conditioning layer (2018). URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

[60] Liu, Z. *et al.* Swin transformer v2: Scaling up capacity and resolution (2022).

[61] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection (2017).

[62] Zhu, K. *et al.* Em-net: Effective and morphology-aware network for skin lesion segmentation. *Expert Systems with Applications* **285**, 127668 (2025).

Acknowledgements

The authors would like to acknowledge the institutions and researchers who made the publicly available datasets used in this study accessible.

Author contributions

U.K. developed the methodology, implemented the framework, conducted the experiments, analysed the results, and drafted the manuscript. U.H. contributed to study conception and design, supervised the research, interpreted the results, and critically revised the manuscript. S.M.Q. contributed to the study design, literature synthesis, and critical revision of the manuscript. M.E.H. contributed to visualization, experimental validation, and interpretation of results. A.A.C. contributed to methodological review and critical revision of the manuscript. S.T. contributed to literature review and writing. All authors reviewed and approved the final manuscript.

Funding

The authors received no specific funding for this research.

Ethics approval and consent to participate

This study used publicly available, de-identified datasets and involved no direct recruitment of human participants or collection of new patient data. The datasets were used in accordance with their respective access conditions and published protocols. Therefore, ethics approval and informed consent were not applicable to this secondary analysis.

Consent for publication

Not applicable. This manuscript does not contain identifiable personal information or images of individual participants.

Materials availability

No physical materials were generated or used in this study. The datasets used in the experiments are available from their respective public repositories.

Code availability

The source code and implementation details supporting the findings of this study are available from the corresponding author upon reasonable request.

Additional information

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationFSDNetAUnifiedSkinToneAwareFrameworkforFairSkinLesionSegmentationandClassification.pdf