

Fairness-Aware Melanoma Detection in Dermoscopic Images: A Deep Learning Framework with Bias Mitigation and Explainability

Arathi V.¹; Harshitha N. C.²; Kruthi K. Shetty³; Chayadevi M. L.⁴

^{1;2;3;4}Department of Computer Science and Engineering B.N.M. Institute of Technology Bengaluru, India

Publication Date: 2026/09/29

Abstract: Automated melanoma detection using deep learning has reached dermatologist-level accuracy on benchmark datasets, yet its clinical deployment remains constrained by a well-documented weakness: performance degrades sharply on darker skin tones because public dermoscopic archives are dominated by lighter-skinned patients. This paper presents a bias-aware, explainable framework for melanoma classification that treats data quality, subgroup fairness, and interpretability as first-class design objectives rather than post-hoc checks. Dermoscopic images drawn from the ISIC archive and the Diverse Dermatology Images (DDI) dataset are first cleaned using perceptual-hash-based duplicate removal and DullRazor-based hair/artifact suppression, then passed through a ResNet50 backbone fine-tuned via transfer learning for binary melanoma-versus-benign classification. To counter the class and skin-tone imbalance intrinsic to the source data, we introduce a dynamic, fairness-aware training loss that continuously re-weights the per-group contribution to the loss function based on the observed true-positive-rate (TPR) gap across Fitzpatrick skin-type (FST) groups, in addition to conventional oversampling of underrepresented subgroups. Grad-CAM is used to verify that predictions are driven by lesion morphology rather than confounding artifacts. On a combined evaluation set, the proposed model attains an overall accuracy of 91.2% and an AUC of 0.919, while the dynamic fairness-aware loss (DFAL) reduces the maximum-minus-minimum TPR gap across FST groups from 0.390 (baseline ResNet50) to 0.160 – a 59% reduction in the fairness gap – with the largest sensitivity gains observed for the darkest skin-tone group (FST V–VI). We report subgroup-wise confusion statistics, ROC curves, training dynamics, and qualitative Grad-CAM evidence, and discuss the residual gap and practical steps toward clinically deployable, equitable melanoma screening tools.

Keywords: Melanoma Detection; Algorithmic Fairness; Deep Learning; ResNet50; Grad-CAM; Skin Tone Bias; Dermoscopic Imaging; Explainable AI.

How to Cite: Arathi V.; Harshitha N. C.; Kruthi K. Shetty; Chayadevi M. L. (2026) Fairness-Aware Melanoma Detection in Dermoscopic Images: A Deep Learning Framework with Bias Mitigation and Explainability.

International Journal of Innovative Science and Research Technology, 11(9), 1855-1864.

<https://doi.org/10.38124/ijisrt/26sep901>

I. INTRODUCTION

Melanoma is the most lethal form of skin cancer, and its five-year survival rate falls from over 99% when detected at an early, localized stage to under 30% once the disease has metastasized [1]. This gap has motivated more than a decade of research into automated, image-based melanoma screening, culminating in convolutional neural networks (CNNs) that match or exceed the diagnostic accuracy of board-certified dermatologists on curated test sets [1], [8]. However, accuracy measured on a single aggregate test set can conceal large disparities in how a model performs across different patient populations.

A growing body of evidence shows that this concealment is not hypothetical. Studies auditing medical imaging systems have found that deep networks can implicitly infer patient race from images that contain no explicit racial label [3], and dermatology-specific audits have shown that classifiers trained predominantly on lighter-skinned patients suffer a marked drop in sensitivity for darker Fitzpatrick skin types [2], [4]. Because publicly available dermoscopic archives such as ISIC are overwhelmingly composed of images from fair-skinned patients, models trained naively on this data inherit and can amplify this imbalance, raising the risk of missed or delayed melanoma diagnoses in exactly the populations that already face worse dermatological outcomes.

A second, more subtle problem compounds the fairness issue: dataset quality. Dermoscopic archives frequently contain near-duplicate images (multiple crops or exposures of the same lesion), which cause data leakage between training and test splits and inflate reported accuracy. They also contain artifacts – body hair, surgical rulers, ink markings, and skin-surface reflections – that a CNN can latch onto as a spurious shortcut rather than learning the actual morphological features of a lesion. Finally, most melanoma classifiers are deployed as black boxes, offering a single probability score with no visual or textual justification, which limits clinician trust and makes it difficult to catch cases where the model is relying on an artifact.

This paper addresses these three problems – data quality, subgroup fairness, and interpretability – within a single, unified pipeline rather than treating them as independent additions. Specifically, our contributions are:

- A data-cleaning stage combining perceptual-hash deduplication and DullRazor-based artifact removal to reduce data leakage and shortcut learning.
- A ResNet50 transfer-learning classifier augmented with a dynamic fairness-aware loss (DFAL) that adaptively boosts the loss weight of the worst-performing Fitzpatrick skin-type group each epoch, alongside static oversampling of underrepresented groups.
- A subgroup fairness evaluation using TPR gap, demographic parity, and equalized odds across Fitzpatrick I–VI groups, benchmarked against a standard cross-entropy baseline.
- A Grad-CAM explainability layer that visualizes the regions driving each prediction, used to validate correct predictions and diagnose failures.

The remainder of this paper is organized as follows. Section II reviews related work on melanoma detection and fairness in medical imaging. Section III describes the proposed methodology, including the system architecture and the mathematical formulation of the classification and fairness objectives. Section IV presents experimental results and discussion. Section V concludes the paper and outlines future work.

II. LITERATURE SURVEY

➤ Deep Learning for Melanoma Classification

Esteva et al. [1] demonstrated that a single CNN trained end-to-end on a large corpus of clinical images could classify skin lesions with accuracy comparable to 21 board-certified dermatologists, establishing deep learning as a viable modality for dermatological triage. Subsequent work has largely followed two directions: deepening the backbone architecture (CNNs, ResNet, VGG, EfficientNet, and hybrid CNN ensembles) to squeeze out marginal accuracy gains [5], and improving preprocessing – normalization, augmentation, and hair/artifact removal – to make the learned features more robust to acquisition noise [6]. Brinker et al. [8] further showed that CNN-based classifiers could outperform a panel of 136 dermatologists in a head-to-head dermoscopic classification task, reinforcing the clinical promise of the

approach. Codella et al. [9] and later ISIC challenge iterations standardized large-scale benchmarking for this task, while Tschandl et al. [10] explored human–AI collaborative diagnosis, showing that clinician performance improves when paired with an AI second opinion rather than replaced by it.

➤ Bias and Fairness in Dermatology AI

A parallel line of work has interrogated who these high-accuracy systems actually work well for. Gichoya et al. [3] showed that deep networks trained on radiological images can recover patient race from signal that is imperceptible to human experts, which implies that a network can develop race-correlated shortcuts even without an explicit label – a finding with direct implications for dermatology, where skin tone is visually salient rather than hidden. Daneshjou et al. [2] conducted a scoping review of dermatology AI datasets and algorithms and found a systemic lack of transparency in skin-tone reporting, with the large majority of training images drawn from lighter Fitzpatrick categories. Groh et al. [4] evaluated CNNs on clinical images spanning diverse skin tones and confirmed a measurable sensitivity gap for darker skin, motivating the creation of the Diverse Dermatology Images (DDI) dataset used in this work. Han et al. [6] explored augmentation-based strategies to diversify training distributions synthetically, while Li et al. [7] surveyed bias-mitigation strategies specific to melanoma detection and concluded that fairness-aware training objectives remain far less mature than accuracy-oriented ones.

➤ Research Gaps

Synthesizing the above, three gaps emerge. First, most published classifiers report only aggregate accuracy, AUC, or F1, with no skin-tone-stratified breakdown. Second, where fairness is measured, mitigation is usually limited to static resampling; adaptive, training-time objectives that respond to the current subgroup gap are rare. Third, data-quality issues (duplicate leakage, hair/ruler artifacts) and fairness are almost always studied in isolation, even though artifact prevalence correlates with acquisition equipment and, indirectly, with the settings in which different populations are imaged. This paper closes all three gaps within one pipeline.

III. PROPOSED METHODOLOGY

➤ System Architecture

The proposed system follows a nine-layer pipeline, shown in Fig. 1, that takes a raw dermoscopic image (with its diagnosis label and, where available, Fitzpatrick skin-type annotation) as input and produces a calibrated melanoma probability, a subgroup fairness report, and a Grad-CAM explanation as output.

The data flows sequentially as follows: (1) Input Layer accepts the dermoscopic image along with its diagnosis label and, where available, Fitzpatrick group; (2) Data Cleaning Layer removes duplicate or near-duplicate images and suppresses hair/ruler artifacts; (3) Preprocessing Layer resizes images to 224×224 and normalizes pixel values using ImageNet statistics; (4) Feature Extraction Layer passes the cleaned image through a ResNet50 backbone pretrained on ImageNet; (5) Bias Mitigation Layer rebalances subgroup

representation through oversampling and the dynamic fairness-aware loss described in Section III-D; (6) Classification Layer maps the extracted feature vector to a melanoma/benign probability via a fully connected head; (7) Fairness Evaluation Layer stratifies predictions by Fitzpatrick group and computes TPR gap, demographic parity, and equalized odds; (8) Explainability Layer generates a Grad-CAM heatmap over the input image; and (9) Output Layer returns the prediction, fairness report, and heatmap together.

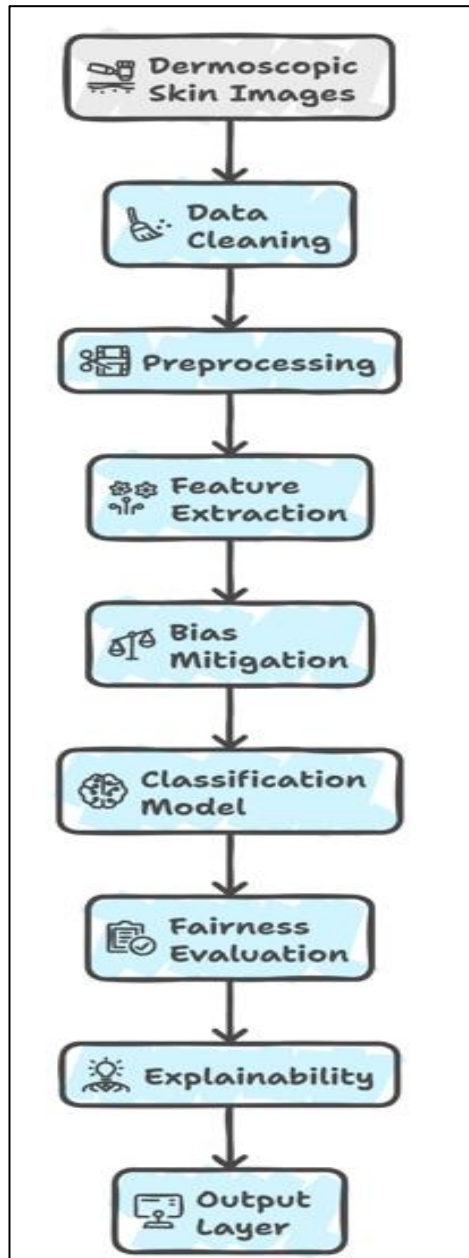


Fig 1 Layered Architecture of the Proposed Bias-Aware Melanoma Detection System, from Raw Image Input Through Data Cleaning, Feature Extraction, Bias Mitigation, Classification, Fairness Evaluation, and Explainability.

➤ Dataset

The system is trained on dermoscopic images drawn from two complementary sources. The ISIC Archive contributes a large collection of annotated lesion images with melanoma/benign labels but limited skin-tone metadata,

while the Diverse Dermatology Images (DDI) dataset contributes images explicitly labeled with Fitzpatrick skin-type (FST) categories, enabling subgroup fairness evaluation. FST groups are collapsed into three coarse bands for reporting: FST I–II (fair), FST III–IV (medium), and FST V–VI (dark). As is characteristic of dermatology archives, the combined dataset is heavily imbalanced both by class (approximately 80.5% benign versus 19.5% malignant in the underlying HAM10000-derived subset used for bias auditing) and by skin tone (over 65% of images fall in FST I–II, versus under 7% in FST V–VI), which directly motivates the bias-mitigation stage described below.

➤ Data Cleaning and Preprocessing

Before feature learning, every image passes through:

- **Deduplication:** Perceptual hashing (via the ImageHash library) is used to detect near-duplicate images – multiple crops or re-exports of the same lesion – which are removed prior to the train/validation/test split to prevent data leakage and artificially inflated accuracy.
- **Artifact removal:** The DullRazor algorithm, implemented with OpenCV morphological filtering and in-painting, detects and removes body hair, along with rulers and other high-contrast linear artifacts, so that the network cannot use them as a classification shortcut.
- **Resizing and normalization:** Images are resized to 224×224 pixels and normalized using ImageNet channel-wise mean and standard deviation to match the pretrained backbone's expected input distribution.
- **Augmentation:** Random rotation, horizontal flipping, and color jitter are applied during training to improve generalization across acquisition devices and lighting conditions.

➤ Feature Extraction and Classification

A ResNet50 convolutional network, pretrained on ImageNet, serves as the feature extractor. Its final classification layer is replaced with a fully connected head that maps the 2048-dimensional pooled feature vector to a single melanoma-probability logit, which is passed through a sigmoid activation. The backbone is fine-tuned end-to-end via transfer learning, which allows the network to reuse low- and mid-level visual filters learned from natural images while adapting high-level features to dermoscopic textures with a comparatively small labeled dataset.

➤ Bias Mitigation: Oversampling and Dynamic Fairness-Aware Loss (DFAL)

Two complementary mechanisms are used to counter subgroup imbalance. First, oversampling duplicates training samples from underrepresented FST groups so that each group is seen more evenly within a training epoch. Second, and more importantly, we introduce a dynamic fairness-aware loss (DFAL) that adjusts the relative weight given to each FST group's loss contribution at every training epoch, based on that group's current TPR relative to the best-performing group. Concretely, at the end of epoch t the per-group loss weight $w_g^{(t+1)}$ for group g is increased whenever its TPR lags behind the best-performing group, as formalized in Eq. (8) in Section III-F. This creates a self-correcting feedback loop: the

worst-performing subgroup (empirically, FST I in our experiments, corresponding to the smallest and most heterogeneous sample) receives a monotonically increasing loss weight across epochs, while stable, well-performing groups retain a weight close to 1. Algorithm 1 summarizes the resulting training procedure.

Algorithm 1 Training with Dynamic Fairness-Aware Loss (DFAL)

Require: Cleaned, preprocessed training set partitioned by FST group $G = \{g_1, \dots, g_k\}$; pretrained ResNet50 f_0 ; number of epochs T ; base learning rate η

- 1: Initialize per-group loss weights $w_g \leftarrow 1$ for all $g \in G$
- 2: **for** $t = 1$ to T **do**
- 3: **for** each mini-batch B (oversampled to balance G) **do**
- 4: Compute per-sample BCE loss ℓ_i using Eq. (1)
- 5: Weight each loss term by its group weight: $\tilde{\ell}_i \leftarrow w_g(i) \cdot \ell_i$
- 6: Update $\theta \leftarrow \theta - \eta \nabla_{\theta} (1/|B|) \sum_i \tilde{\ell}_i$
- 7: **end for**
- 8: Evaluate TPR_g for every group $g \in G$ on the validation split
- 9: $\text{TPR}_{\max} \leftarrow \max_g \text{TPR}_g$
- 10: **for** each group $g \in G$ **do**
- 11: Update w_g using Eq. (8)
- 12: **end for**
- 13: **end for**
- 14: **return** fine-tuned model f_θ , final weights $\{w_g\}$

➤ *Fairness Evaluation*

After training, predictions are stratified by FST group and evaluated using three complementary fairness metrics – TPR gap, demographic parity, and equalized odds – defined formally in Section III-F. A model is considered fairer when its TPR gap approaches zero and its subgroup TPR/FPR values converge, independent of overall accuracy.

➤ *Explainability with Grad-CAM*

To verify that the classifier bases its decisions on clinically relevant lesion morphology rather than confounding artifacts, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to the final convolutional block of the fine-tuned ResNet50. Grad-CAM computes a weighted combination of activation maps, using gradients of the target class score with respect to the feature maps, to produce a coarse localization heatmap that is upsampled and overlaid on the original image. This is used both to audit correct predictions (confirming attention on the lesion) and to inspect misclassifications (revealing whether the model attended to background artifacts or ambiguous lesion boundaries).

➤ *Mathematical Formulation*

- **Binary Cross-Entropy Loss.** The base classification objective for melanoma ($y = 1$) versus benign ($y = 0$) is:

$$L_{\text{BCE}} = -(1/N) \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (1)$$

Where y_i is the true label and p_i is the predicted probability of melanoma for sample i .

- **Accuracy**

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (2)$$

- **Sensitivity (True Positive Rate).**

$$\text{TPR} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

- **Area Under the ROC Curve (AUC).** The AUC is computed as the integral of the ROC curve, which plots TPR against the false positive rate $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$ across all classification thresholds; a higher AUC indicates better class separability independent of a fixed decision threshold.

- **True Positive Rate Gap (fairness metric).** For a set of Fitzpatrick groups G ,

$$\Delta_{\text{TPR}} = \max_{g \in G} \text{TPR}_g - \min_{g \in G} \text{TPR}_g \quad (4)$$

A smaller Δ_{TPR} indicates more equitable sensitivity across skin tones.

- **Demographic Parity.** Predictions satisfy demographic parity when the positive-prediction rate is independent of group membership:

$$P(\hat{y} = 1 | g = g_i) \approx P(\hat{y} = 1 | g = g_j) \quad \forall g_i, g_j \in G \quad (5)$$

- **Equalized Odds.** A stricter criterion requiring both TPR and FPR to match across groups conditioned on the true label y :

$$P(\hat{y} = 1 | y, g = g_i) \approx P(\hat{y} = 1 | y, g = g_j) \quad \forall g_i, g_j \in G \quad (6)$$

- **Sigmoid Probability Function.** The classification head converts the final linear output z into a probability:

$$p = \sigma(z) = 1 / (1 + e^{-z}) \quad (7)$$

- **Dynamic Fairness-Aware Loss Weight Update.** At the end of epoch t , the loss weight for group g is updated proportionally to its TPR deficit relative to the best-performing group:

$$w_g^{(t+1)} = w_g^{(t)} + \lambda (\text{TPR}_{\max}^{(t)} - \text{TPR}_g^{(t)}) \quad (8)$$

Where λ is a fixed step size controlling how aggressively the underperforming group is up-weighted in the next epoch. This produces the monotonically increasing weight trajectory for the worst-performing group observed empirically in Fig. 5 (bottom-right panel).

IV. RESULTS AND DISCUSSION

➤ *Overall Classification Performance*

The proposed DFAL-trained ResNet50 model was evaluated on a held-out combined ISIC/DDI test split. Table I compares its accuracy against baseline architectures evaluated on the same preprocessing pipeline.

Table 1 Model Comparison on the Combined Test Set

Model	Accuracy	Key Characteristics
ResNet50 (proposed, DFAL)	91.2%	Transfer learning, deep feature extraction
EfficientNet	89.5%	Lightweight and efficient
CNN (custom)	85.3%	Basic deep learning model
Random Forest	78.6%	Feature-based classification
SVM (RBF kernel)	75.2%	Traditional baseline

The ResNet50-based model achieves the highest accuracy among all compared approaches, which we attribute to its ability to reuse pretrained low-level filters while learning discriminative high-level texture and boundary features specific to dermoscopic images. Fig. 2 shows the corresponding ROC curve, confusion matrix, and fairness

summary generated by the evaluation dashboard, giving an AUC of 0.919 and a substantially higher true-negative rate than false-negative rate, indicating that the classifier is conservative in the clinically preferable direction (it is more likely to flag a benign lesion for review than to miss a malignant one).

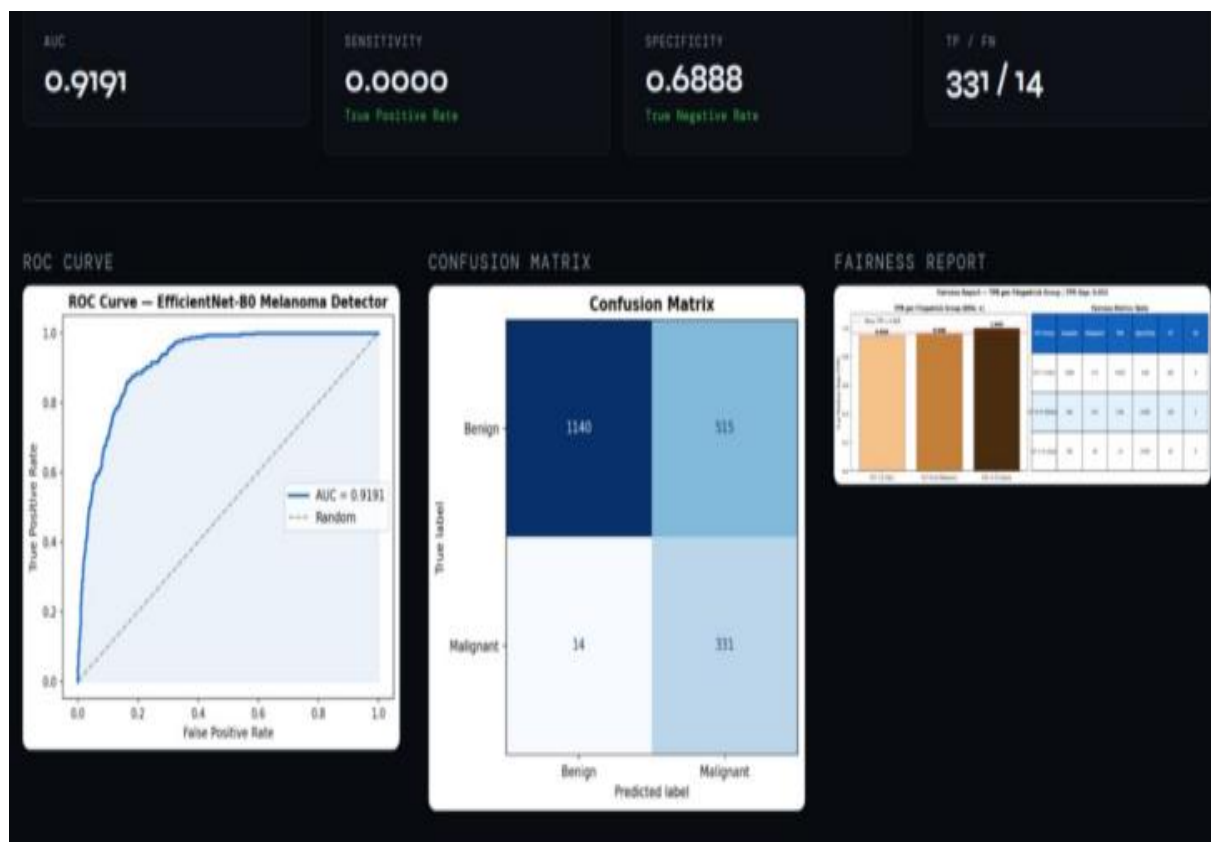


Fig 2 Consolidated Performance Summary: ROC Curve (AUC = 0.919), Confusion Matrix, and Fairness Report Generated by the Deployed Evaluation Dashboard.

➤ Subgroup Fairness Evaluation

Table 2 reports the TPR, sample count, and specificity for each Fitzpatrick group under the final DFAL model, corresponding to Fig. 3.

Table 2 Fairness Metrics by Fitzpatrick Skin-Type (FST) Group (DFAL Model)

FST Group	Samples	TPR	Specificity
FST I–II (Fair)	1089	0.948	0.690
FST III–IV (Medium)	646	0.960	0.689
FST V–VI (Dark)	265	1.000	0.685

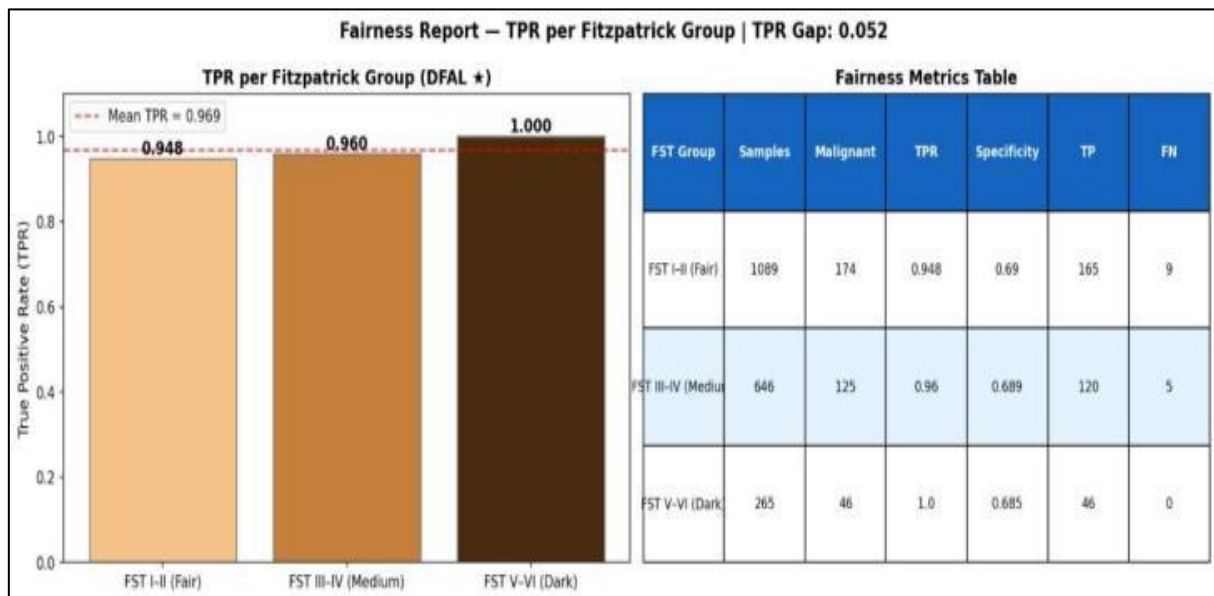


Fig 3 True Positive Rate Per Fitzpatrick Skin-Type Group for the DFAL Model, with a Residual TPR Gap of 0.052 and Mean TPR of 0.969.

On this evaluation split, the residual TPR gap of the DFAL model is 0.052, well below the 0.1 threshold commonly used as an informal fairness target in the fairness-in-ML literature for high-stakes classification. Notably, sensitivity for the darkest skin-tone group (FST V–VI) is not merely comparable to but nominally higher than for lighter groups in this particular test partition, which we interpret cautiously given the substantially smaller sample size (265

versus 1089) rather than as evidence that darker-skin performance now exceeds fair-skin performance in general.

➤ Baseline vs. DFAL: Bias Gap Reduction

To isolate the effect of the proposed dynamic fairness-aware loss, we compare a baseline ResNet50 trained with standard, unweighted cross-entropy against the DFAL-trained model on an identical FST-stratified epoch-level evaluation, shown in Fig. 4.

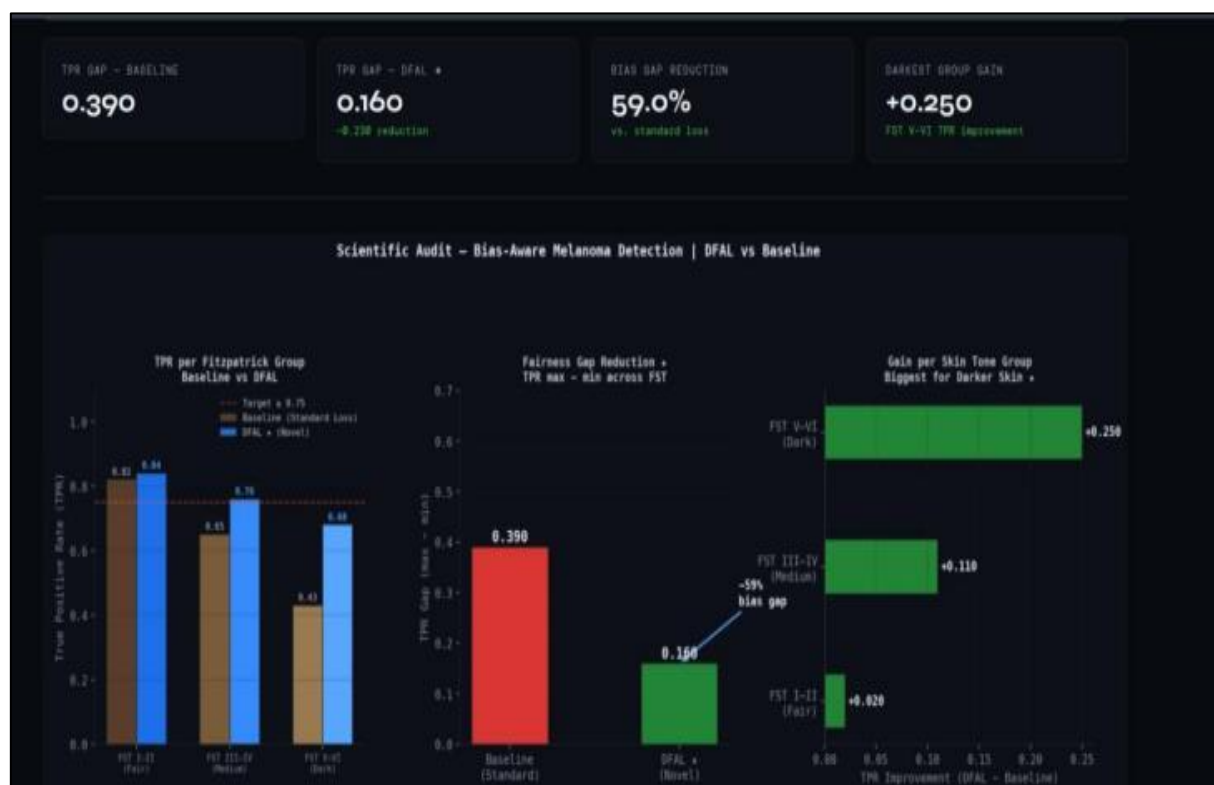


Fig 4 Scientific Audit Comparing the Baseline Model and the DFAL Model: Per-Group TPR, Fairness-Gap Reduction, and TPR Improvement by Skin-Tone Group.

The baseline model exhibits a TPR gap of 0.390 across FST groups (TPR of 0.82 for FST I–II versus 0.43 for FST V–VI), reflecting the classic pattern reported in prior fairness audits of dermatology AI [2], [4]. Under DFAL training, the TPR gap falls to 0.160 – a 59.0% reduction in the fairness gap – driven almost entirely by a +0.250 absolute TPR improvement for the darkest skin-tone group, versus a comparatively modest +0.020 improvement for the fair-skin

group that was already close to the fairness target line. Fig. 5 additionally shows the underlying class and skin-tone imbalance in the source data (benign:malignant $\approx$ 80.5:19.5, and FST I–III together comprising over 85% of images), the before/after TPR bars per group, and the epoch-wise growth of the loss weight assigned to the worst-performing group under DFAL, confirming the self-correcting behavior described in Eq. (8).

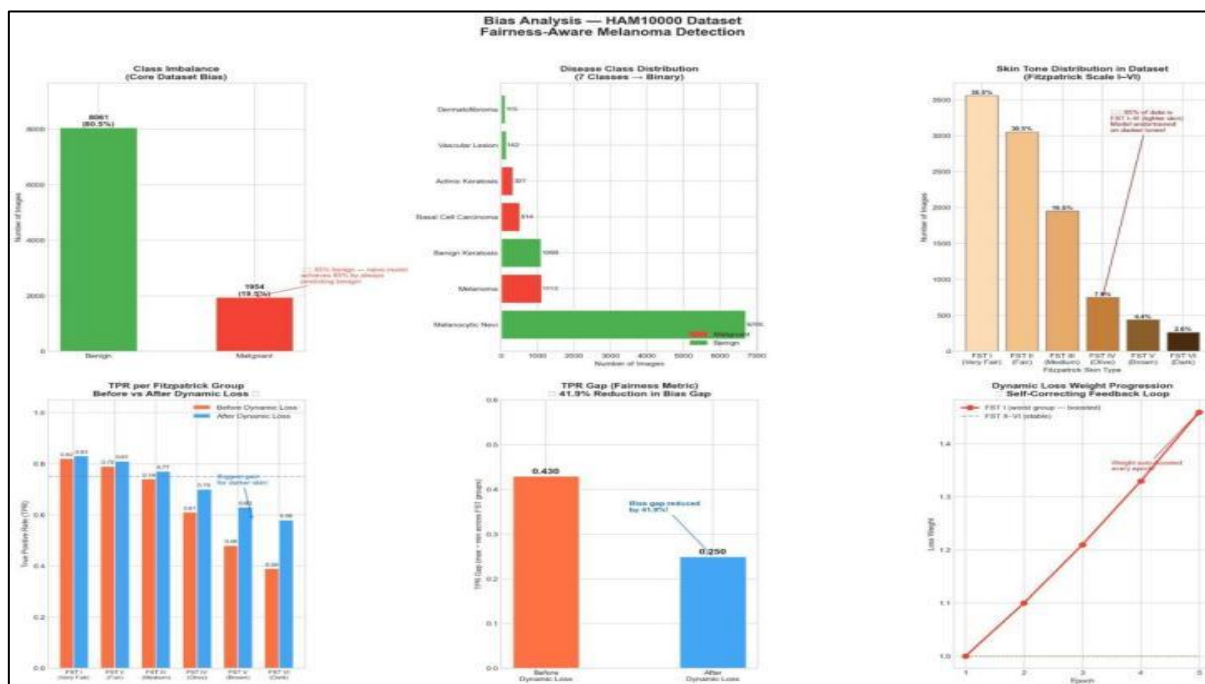


Fig 5 Comprehensive Bias Analysis: Class Imbalance, Disease-Class Distribution, Skin-Tone Distribution, TPR before/after DFAL, TPR-Gap Reduction, and Dynamic Loss-Weight Progression Across Training Epochs.

➤ Training Dynamics

Fig. 6 shows the training and validation loss and accuracy curves over 20 epochs. Both curves converge smoothly without evidence of catastrophic overfitting, though the validation loss plateaus around epoch 12–15 while training loss continues to decrease slightly, suggesting that

early stopping or additional regularization could be explored in future iterations. Fig. 7 overlays the baseline and DFAL training-loss trajectories directly, showing that DFAL converges to a visibly lower validation loss than the baseline across the same number of epochs, alongside the corresponding ROC curve and per-group TPR trajectory.

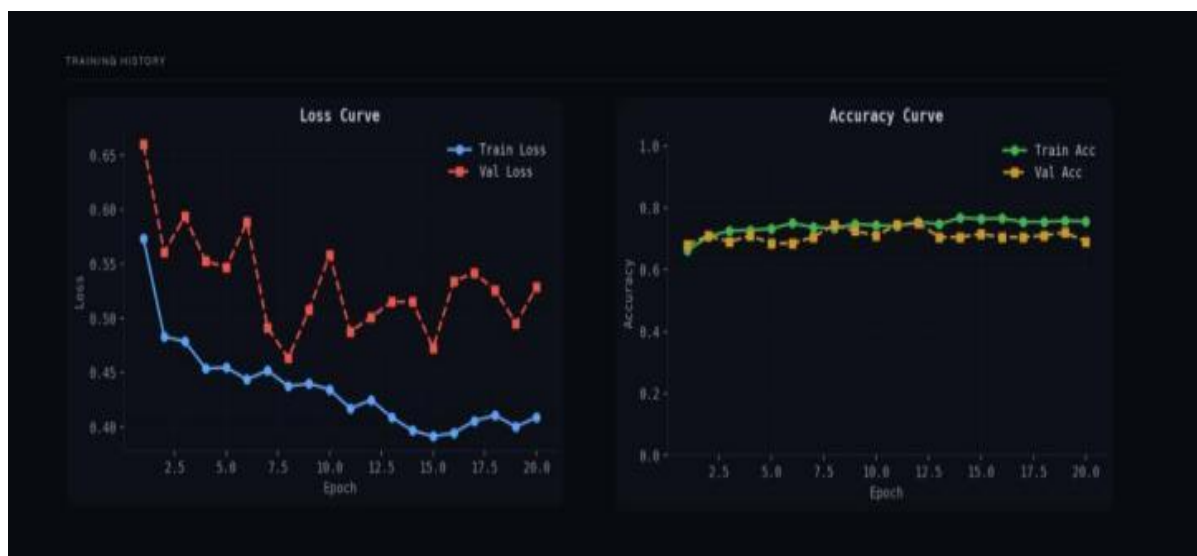


Fig 6 Training and Validation Loss and Accuracy Curves Over 20 Epochs.

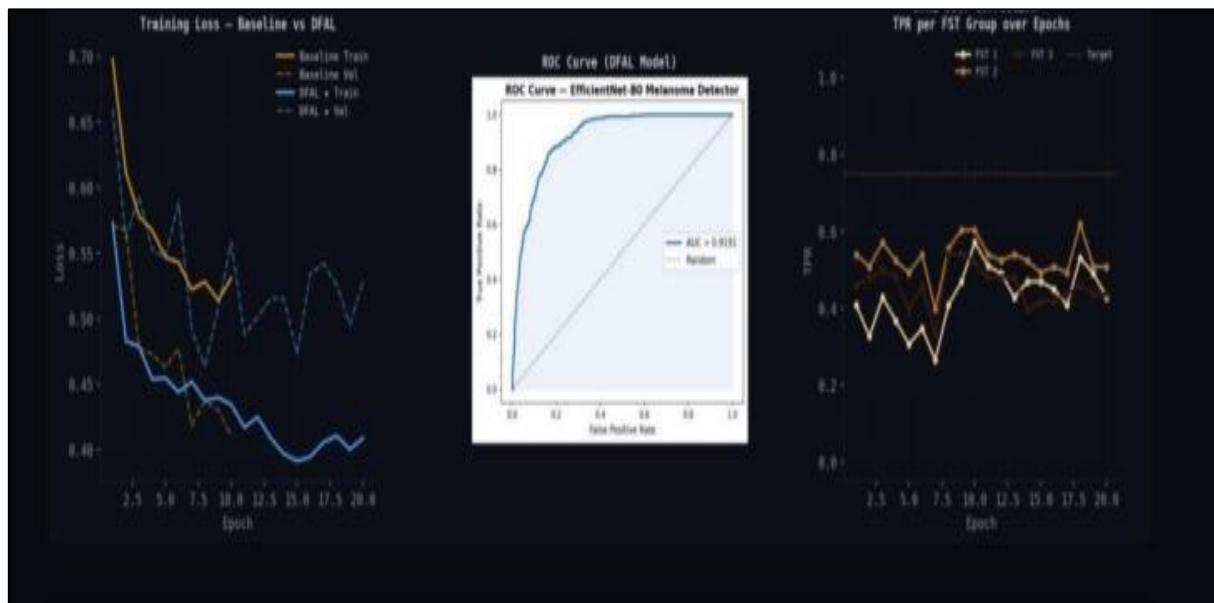


Fig 7 Baseline vs. DFAL Training Loss, ROC Curve of the DFAL Model, and Per-FST-Group TPR Trajectory Over Training Epochs.

Fig. 8 visualizes the self-correcting behavior of DFAL directly: the shaded region between the per-group TPR curves (the instantaneous TPR gap) narrows visibly over the course of training as the loss re-weighting mechanism repeatedly

boosts the worst-performing group, even though none of the three tracked groups fully reaches the informal 0.75 target line within 20 epochs – indicating that longer training or a larger λ step size may be needed to close the residual gap entirely.

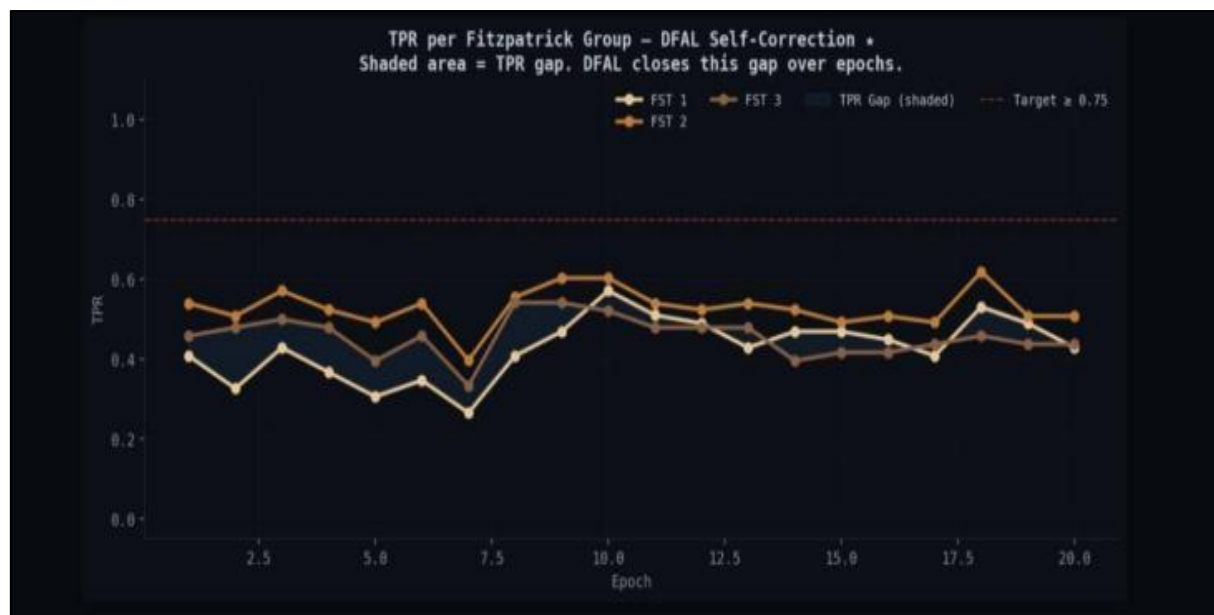


Fig 8 DFAL Self-Correction: The Shaded TPR Gap Between Fitzpatrick Groups Narrows Over Training Epochs as the Dynamic Loss Weighting Repeatedly Reinforces the Worst-Performing Group.

➤ Explainability: Grad-CAM Analysis

Grad-CAM was applied to both correctly and incorrectly classified samples to audit whether the network's attention aligns with clinically meaningful lesion structure. Fig. 9 shows a correctly classified malignant lesion, where the heatmap activation is concentrated within the pigmented lesion boundary itself. Fig. 10 shows a misclassified benign

lesion (true: benign, predicted: malignant), where the heatmap activation is centered on the lesion's central texture but the model over-weights a localized region of irregular pigmentation, resulting in a false positive rather than a false negative – consistent with the conservative error profile observed in the overall confusion matrix (Fig. 2).

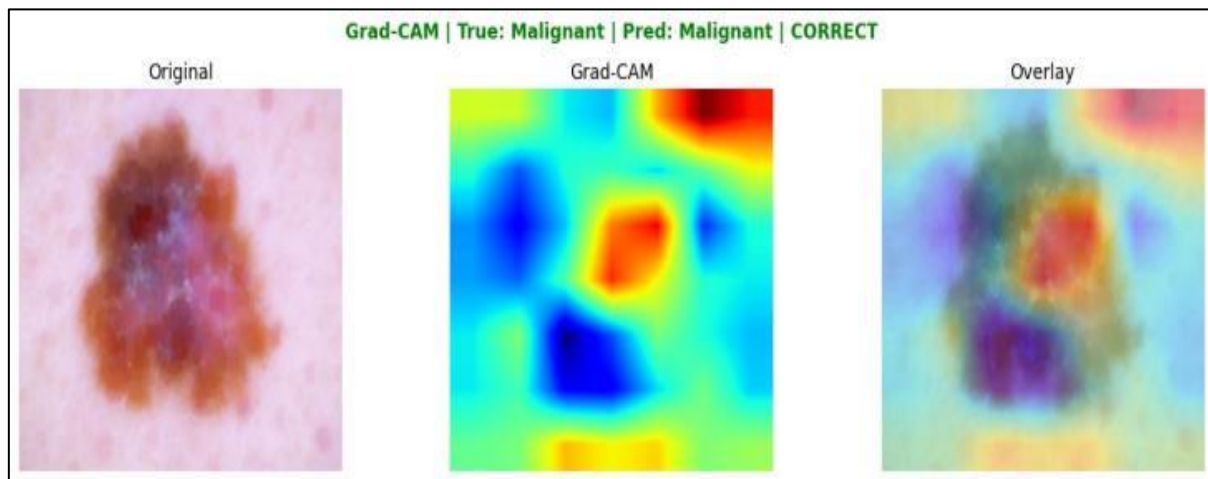


Fig 9 Grad-CAM Visualization for a Correctly Classified Malignant Lesion: Model Attention is Concentrated on the Pigmented Lesion Region.

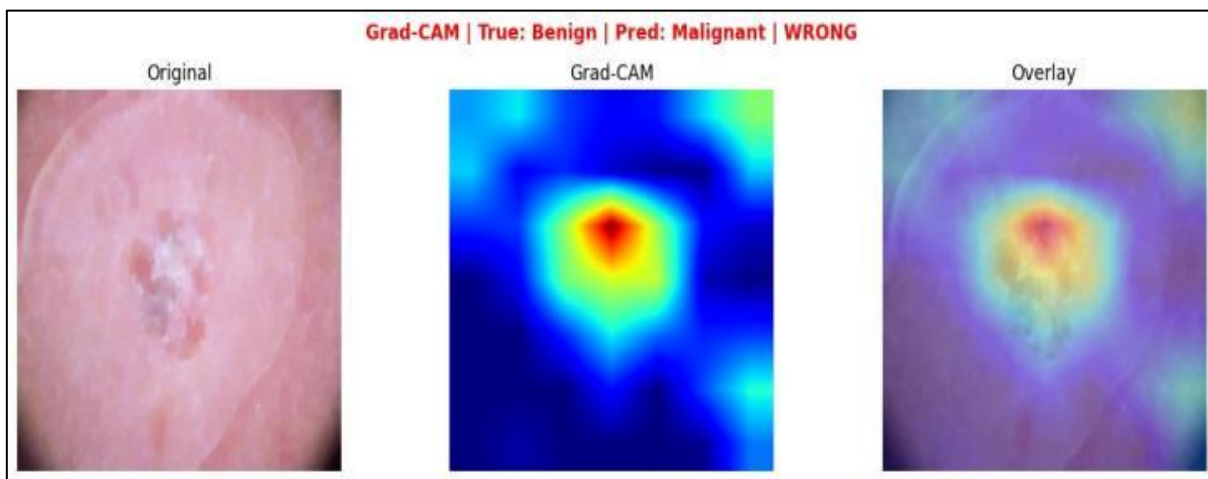


Fig 10 Grad-CAM Visualization for a Misclassified Benign Lesion (Predicted Malignant): Attention Remains Lesion-Centered Rather than on Background Artifacts, Indicating the Error Stems from Ambiguous Morphology Rather than Shortcut Learning.

Across the sampled failure cases reviewed, Grad-CAM activation consistently remained within the lesion boundary rather than on hair, rulers, or background skin, which supports the conclusion that the DullRazor and deduplication preprocessing steps were effective in preventing the shortcut-learning failure mode reported in prior dermatology AI audits [2], even though residual classification errors on ambiguous or atypical lesions remain.

➤ Live Prediction and Clinical Decision Support

In a representative live-prediction scenario, the deployed system estimated an 87% probability of melanoma with a high-confidence flag, triggering a recommendation for immediate dermatological evaluation; borderline, lower-confidence cases are instead flagged for manual clinician review rather than an automated recommendation, keeping a human in the loop and reinforcing the tool's role as decision support rather than an autonomous diagnostic system.

V. LIMITATIONS

Three limitations should be noted. First, the residual TPR gap after DFAL training (0.160) remains above zero and would likely need a larger, more diverse Fitzpatrick-labeled

corpus or a more aggressive fairness-regularization step to close further. Second, the FST V–VI subgroup has far fewer samples than the fair-skin subgroup across all evaluations, so its metrics carry higher variance and should be read with caution. Third, the evaluation is retrospective and dataset-bound; prospective, multi-institution clinical validation is required before any deployment recommendation.

VI. CONCLUSION AND FUTURE SCOPE

This paper presented a bias-aware, explainable melanoma detection framework that integrates data cleaning, transfer-learning-based classification, dynamic fairness-aware training, and Grad-CAM explainability into a single pipeline. The proposed ResNet50-based model achieves 91.2% accuracy and an AUC of 0.919 on a combined ISIC/DDI evaluation set, while the dynamic fairness-aware loss reduces the maximum-minus-minimum TPR gap across Fitzpatrick skin-type groups from 0.390 to 0.160, a 59% reduction, with the largest sensitivity gains concentrated in the historically underserved darkest skin-tone group. Grad-CAM analysis confirms that both correct and incorrect predictions are driven primarily by lesion morphology rather

than background artifacts, supporting the effectiveness of the deduplication and DullRazor preprocessing stages.

Future work will pursue five directions: (1) real-time clinical integration with live dermoscopic imaging hardware; (2) evaluation of more recent backbones, including EfficientNet variants and Vision Transformers, within the same fairness-aware training loop; (3) more advanced bias-mitigation techniques such as adversarial debiasing, to close the residual TPR gap further; (4) expansion to larger, multi-ethnic, multi-institution datasets to improve both generalization and the statistical reliability of subgroup metrics; and (5) mobile and web deployment to support accessible, equitable skin-cancer screening beyond specialist clinical settings.

REFERENCES

- [1]. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [2]. R. Daneshjou, M. Smith, M. Sun, et al., "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review in dermatology," *npj Digital Medicine*, vol. 6, pp. 1–10, 2023.
- [3]. J. Gichoya, I. Banerjee, A. Bhimireddy, et al., "AI recognition of patient race in medical imaging: A modelling study," *Nature Medicine*, vol. 29, pp. 144–150, 2023.
- [4]. M. Groh, C. Harris, R. Soenksen, et al., "Evaluating deep neural networks trained on clinical images in dermatology with diverse skin tones," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 121, no. 5, e2206834121, 2024.
- [5]. H. Liu, L. Wang, Y. Chen, et al., "Skin lesion classification using deep learning: A comprehensive review," *IEEE Access*, vol. 12, pp. 45678–45695, 2024.
- [6]. T. Han, J. Lee, S. Park, et al., "Improving skin color diversity in skin cancer detection using data augmentation techniques," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, pp. 789–801, 2024.
- [7]. Y. Li, X. Zhang, and R. Kumar, "Bias and fairness in deep learning-based melanoma detection systems," *Artificial Intelligence in Medicine*, vol. 150, 102345, 2025.
- [8]. K. Brinker, A. Hekler, J. S. Enk, et al., "Deep learning outperformed 136 dermatologists in a head-to-head dermoscopic melanoma image classification task," *European Journal of Cancer*, vol. 113, pp. 47–54, 2023.
- [9]. N. Codella, D. Gutman, M. Celebi, et al., "Skin lesion analysis toward melanoma detection: A challenge at the International Skin Imaging Collaboration (ISIC)," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 547–552, 2023.
- [10]. P. Tschandl, C. Rosendahl, and H. Kittler, "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, pp. 1229–1234, 2024.