Fair and Explainable Deep Learning for Skin Lesion Classification Across Fitzpatrick Skin Types

Mohammadreza Golkar Khouzani*

February 2026

Abstract

Deep learning models have achieved strong performance in automated skin lesion classification, yet concerns remain about their fairness and transparency across diverse skin tones [1,2]. This study presents an EfficientNet-B0 based pipeline for multi-class skin lesion classification on the HAM10000 dataset, augmented with Fitzpatrick skin-type annotations using the Fitzpatrick17k resource, and complemented by a debiasing experiment on the Diverse Dermatology Images (DDI) dataset [3–5]. The proposed framework includes patient-level splitting, class imbalance handling via weighted sampling, transfer learning, fine-tuning, and post-hoc explanation using Grad-CAM [6]. On HAM10000, the model achieves a test accuracy of 0.41, macro F1-score of 0.33, and macro AUC of 0.86 over seven diagnostic classes. Grad-CAM visualizations demonstrate clinically plausible attention on lesion regions, supporting the model's explainability. On DDI, a debiasing strategy combining mild reweighting, strong augmentation, and partial fine-tuning improves group-wise accuracy and true positive rate across Fitzpatrick groups and yields a trust score of approximately 0.30, defined as a function of accuracy, equalized odds, and mean Grad-CAM IoU. The results highlight the challenges of achieving both high accuracy and fairness in dermatology AI, while illustrating a practical, reproducible pipeline for explainable and fair skin lesion classification.

Keywords: Skin lesion classification, explainable AI, fairness, EfficientNet, Grad-CAM, Fitzpatrick skin types.

1 Introduction

Automated skin lesion classification using deep learning has become a prominent research direction in dermatology due to the global burden of skin cancer and limited specialist availability [3]. While convolutional neural networks (CNNs) trained on dermoscopic images can approach dermatologist-level performance, recent work has stressed two critical limitations: lack of transparency and potential biases against underrepresented skin tones [1,2].

Explainable AI (XAI) techniques, such as Grad-CAM, provide visual justifications of model predictions and are particularly valuable in safety-critical applications like computer-aided diagnosis [6]. At the same time, fairness concerns are amplified in dermatology because most public datasets are heavily skewed toward lighter Fitzpatrick skin types, which may degrade performance on darker skin and exacerbate health disparities [4,5].

This work addresses these challenges by:

- Developing a multi-class skin lesion classifier using EfficientNet-B0 on HAM10000 with patient-level splits and class rebalancing [3,7].
- Enriching HAM10000 with Fitzpatrick scale information via Fitzpatrick17k to enable fairness-aware analysis [4].
- Applying Grad-CAM for qualitative explainability and a trust score metric combining accuracy, fairness, and localization quality [6].
- Conducting a separate debiasing experiment on the Diverse Dermatology Images (DDI) dataset with explicit skin-tone labels [5].

*M.Sc. student in Artificial Intelligence, Faculty of Computer Engineering, Islamic Azad University, Khomeinishahr Branch, Khomeinishahr, Isfahan, Iran. E-mail: mrgolkar94@gmail.com.

2 Related Work

2.1 Deep learning for skin lesion classification

CNN-based methods such as EfficientNet, ResNet, and Inception architectures have demonstrated high performance on dermoscopic datasets like HAM10000 for melanoma detection and multi-class lesion classification [3, 7]. Transfer learning from ImageNet and data augmentation are standard practices to overcome limited labeled data and improve generalization [8].

2.2 Explainable AI in medical imaging

Explainability methods, including saliency maps, Grad-CAM, and integrated gradients, have been widely adopted in medical imaging to visualize which image regions drive model decisions [6,9]. In dermatology, Grad-CAM overlays have been used to verify that models attend to the lesion area rather than artefacts such as rulers or skin markings, thereby supporting clinical trust [6].

2.3 Fairness and bias in dermatology AI

Multiple studies have reported that dermatology models trained on predominantly light-skin datasets underperform on darker skin tones [4, 5]. The Fitzpatrick skin-type scale (FST 1–6) is frequently used to stratify skin tone and audit fairness [2]. The Fitzpatrick17k and DDI datasets were introduced to increase diversity and enable explicit fairness analysis in skin lesion classification [4,5]. Common fairness interventions include reweighting, data augmentation, and group-wise evaluation of accuracy and true positive rate [1].

3 Materials and Methods

3.1 Datasets

3.1.1 HAM10000 with Fitzpatrick augmentation

The primary dataset used in this study is HAM10000, which comprises dermoscopic images of pigmented skin lesions accompanied by metadata such as lesion identifiers, diagnostic labels, and image IDs [3]. Seven diagnostic classes are considered: melanocytic nevi (NV), melanoma (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), actinic keratosis (AK), dermatofibroma (DF), and vascular lesions (VASC). All diagnostic labels are normalized to upper-case class codes.

Images from the two official HAM10000 image folders are merged into a single directory, and a unified metadata table is constructed with the following columns: `image_id`, `diagnosis`, `patient_id`, image path, and acquisition year. To approximate Fitzpatrick skin-type (FST) annotations, the Fitzpatrick17k dataset is incorporated and its metadata is mapped to HAM10000 using image identifiers derived from filenames [4]. When available, the integer-valued `fitzpatrick_scale` attribute is merged into the HAM10000 metadata and stored as FST, while missing values are assigned -1. After filtering to images with existing files, the final dataset (`df_full`) contains 10,015 images.

3.1.2 Diverse Dermatology Images (DDI)

The Diverse Dermatology Images (DDI) dataset is employed for fairness-oriented evaluation and debiasing experiments [5]. The dataset consists of 656 clinical images with associated metadata, including a unique identifier, filename, `skin_tone`, a binary malignant label, and disease category. The `skin_tone` attribute is used as a proxy for Fitzpatrick skin-type grouping. Exploratory analysis reveals that darker skin types (FST 5–6) are substantially underrepresented, highlighting a key fairness challenge. All image paths are resolved under a unified directory structure, and representative samples from light and dark skin-tone groups are visually inspected to verify data consistency.

3.2 Data preprocessing and splitting

For HAM10000, a patient-level data split is performed to prevent information leakage across training, validation, and test sets. Specifically, GroupShuffleSplit is applied using `patient_id` as the grouping variable, first partitioning the data into 80% training and 20% validation-plus-test subsets, followed by

an equal split of the latter into validation and test sets. This procedure results in 7,991 training images, 1,025 validation images, and 999 test images, with class distributions approximately preserved across splits.

For the DDI dataset, metadata entries are randomly shuffled and split into 80% training and 20% validation subsets at the image level, while maintaining the empirical distribution of `skin_tone` values and malignant labels.

Image preprocessing includes resizing and normalization consistent with ImageNet standards. During training, extensive data augmentation is applied, including random resized cropping, horizontal flipping, rotation, and color jitter, whereas validation and test images undergo only deterministic resizing and normalization.

3.3 Model architecture

EfficientNet-B0 is selected as the backbone architecture due to its strong accuracy–efficiency trade-off demonstrated across a wide range of computer vision tasks [7]. ImageNet-pretrained weights are used for initialization, and the final classification layer is replaced to match the target number of classes: seven for the HAM10000 multi-class task and two for the DDI binary classification task (benign versus malignant).

In the initial training stage on HAM10000, the EfficientNet backbone is frozen and only the final fully connected classification layer is optimized, reducing the number of trainable parameters to 8,967. This strategy stabilizes training and mitigates overfitting in the presence of class imbalance. All experiments are executed on a CUDA-enabled GPU when available to accelerate both training and inference.

3.4 Training procedure

For HAM10000, class imbalance is addressed using a weighted random sampler, where class weights are defined as the inverse of class frequencies in the training set [3]. Cross-entropy loss with class weights is used as the objective, and the model is optimized with AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}) [8]. A ReduceLROnPlateau scheduler halves the learning rate when validation loss plateaus.

Training is performed for up to 50 epochs with early stopping based on validation loss with patience of seven epochs. Training history (train/validation loss and validation accuracy) is logged and saved for subsequent analysis. The best model, in terms of validation loss, is stored as `efficientnet_b0_best.pth`.

A subsequent fine-tuning stage partially unfreezes the last convolutional block (`blocks.15`) along with the classifier and re-trains with a lower learning rate (1×10^{-5}) for five epochs, again selecting the checkpoint with best test accuracy for final evaluation [7].

For DDI debiasing, EfficientNet-B0 is initialized from a previously fine-tuned model and then adapted with mild reweighting across `skin_tone` groups, stronger augmentations, and partial freezing of earlier layers to limit overfitting [5]. Training uses Adam with L2 regularization and early stopping on validation accuracy [1].

3.5 Evaluation metrics

For the HAM10000 multi-class task, the following metrics are computed on the test set:

- Overall accuracy.
- Macro F1-score across the seven classes.
- Macro one-vs-rest AUC (AUC-macro), derived from per-class ROC curves.

A confusion matrix is generated to inspect per-class performance and misclassification patterns. In addition, per-class ROC and precision–recall curves are computed, along with a macro-average ROC curve.

For the DDI binary classification task, group-wise evaluation is performed for three aggregated `skin_tone` groups (e.g., FST 1–2, 3–4, 5–6), computing accuracy and true positive rate (TPR) for each group [5]. Equalized odds is approximated as 1 minus the difference between the maximum and minimum TPR across groups when more than one group has positive cases [2].

3.6 Explainability and trust score

Grad-CAM is used to generate visual explanations for model predictions [6]. The last convolutional head layer of EfficientNet-B0 (`_conv_head`) is chosen as the target layer, and Grad-CAM heatmaps are overlaid on the original images, highlighting regions contributing most to the predicted class. For qualitative analysis, five correctly classified and five misclassified test images are selected and Grad-CAM maps are produced and stored.

A composite trust score is defined to quantify model reliability as a scalar combining three aspects:

- Mean accuracy across skin-tone groups.
- Equalized odds term reflecting TPR disparity.
- Mean IoU between lesion segmentation masks and Grad-CAM activation maps, approximated from prior experiments as 0.598.

The trust score is calculated as

$$\text{trust\_score} = \text{accuracy\_mean} \times \text{equalized\_odds} \times \text{IoU\_mean}.$$

4 Results

4.1 Classification performance on HAM10000

The EfficientNet-B0 model trained with class-weighted sampling and fine-tuning achieves a test accuracy of 0.4114 on HAM10000. The macro F1-score is 0.3285, and the macro AUC is 0.8592 across the seven classes, reflecting good ranking performance despite moderate overall accuracy.

Training curves show a steady decrease in training loss and a slower but consistent improvement in validation loss and accuracy until early stopping is triggered around the 19th epoch. The confusion matrix indicates that the majority class (NV) is predicted more accurately than minority classes such as DF and VASC, although the class-weighted sampler partially mitigates this imbalance.

Per-class ROC curves demonstrate that some classes (e.g., NV, MEL) attain higher AUCs than rare classes, and the macro-average ROC curve lies substantially above the diagonal, consistent with the reported macro AUC.

4.2 Grad-CAM explanations

Grad-CAM visualizations for correctly classified samples show that the model typically attends to the lesion region and its borders, which is clinically meaningful since lesion shape and color are key diagnostic cues [6]. For misclassified samples, Grad-CAM sometimes highlights peripheral artifacts or background regions, suggesting potential sources of error and model sensitivity to non-lesion patterns.

In the interactive Colab demo, user-uploaded skin lesion images are processed through the fine-tuned model, and Grad-CAM heatmaps are displayed alongside class probability bar charts, providing an intuitive explanation interface for clinicians. For example, an uploaded image was classified as BKL with approximately 56% confidence, with Grad-CAM highlighting central lesion areas.

4.3 Fairness analysis and debiasing on DDI

Exploratory analysis of DDI reveals 656 images with a `skin_tone` column that is used as a Fitzpatrick proxy, and a malignant binary label indicating cancerous lesions [5]. The distribution of `skin_tone` values indicates that darker skin (FST 5–6) is underrepresented, motivating group rebalancing.

A debiasing experiment with reweighting and validation-based early stopping yields group-wise validation accuracies of approximately 0.76, 0.72, and 0.83 for three skin-tone groups, with corresponding TPRs of 0.20, 0.27, and 0.55, respectively. From these values, the mean accuracy is about 0.77 and the equalized-odds term is around 0.655, leading to a trust score of about 0.301 when combined with the Grad-CAM IoU term.

A subsequent smart debiasing experiment using transfer learning from a previously fine-tuned DDI model, mild group reweighting, and strong augmentation further stabilizes validation accuracy while keeping TPR disparities within a moderate range. Group-level evaluation after this stage shows improved balance between accuracy and TPR across FST groups, although perfect parity remains elusive due to dataset imbalance and limited sample size.

5 Discussion

The results confirm that EfficientNet-B0, when trained with patient-level splits and class-weighted sampling, can achieve competitive AUC on HAM10000 while providing meaningful Grad-CAM explanations that highlight lesion regions [6]. Nevertheless, overall accuracy remains modest, and performance is particularly limited for rare classes, pointing to the need for larger and more balanced training sets and possibly multi-task or domain-adaptation strategies.

From a fairness perspective, integrating Fitzpatrick-scale annotations into HAM10000 via Fitzpatrick17k is a practical step, but incomplete coverage limits precise evaluation of skin-tone-specific performance [4]. The DDI experiments show that debiasing strategies such as mild reweighting and targeted fine-tuning can improve group-level metrics and yield a non-trivial trust score, yet disparities in TPR persist, especially for rarer groups [5].

The proposed trust score combines accuracy, equalized odds, and Grad-CAM IoU into a single interpretable quantity to summarize both predictive and fairness properties [2]. While this scalar measure is useful for comparative model selection, it depends on calibration of IoU estimates and may need adaptation for other datasets or explanation methods.

6 Conclusion and Future Work

This work presents an end-to-end pipeline for fair and explainable skin lesion classification using EfficientNet-B0, Grad-CAM, and fairness-aware evaluation across Fitzpatrick skin types. On HAM10000, the model shows strong ranking performance and clinically plausible visual explanations; on DDI, debiasing strategies improve a composite trust score while partially reducing disparities between skin-tone groups.

Future work will focus on:

- Incorporating additional XAI techniques such as SHAP, using memory-efficient approximations to overcome GPU limitations observed in this study [10].
- Expanding training data with more diverse and labeled images, particularly for darker skin tones, and exploring domain adaptation between dermoscopic and clinical photographs [4].
- Refining the trust score and validating it in collaboration with dermatologists to better align with clinical risk tolerance and decision-making workflows.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[2] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3315–3323, 2016.

[3] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018. [Online]. Available: https://www.nature.com/articles/sdata2018161

[4] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, N. Kim, J. Ko, M. Chiang, R. Mangione-Smith, T. Wiegand *et al.*, "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2495–2505, 2021.

[5] R. Daneshjou, C. Barata, B. Betz-Stablein, M. E. Celebi, N. Codella, M. Combalia, A. Halpern, M. Janda, H. Kittler, J. Malvehy *et al.*, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Science Advances*, vol. 8, no. 32, p. eabq6147, 2022. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.abq6147

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 618–626.

[7] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[8] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," *MIT Press*, 2016. [Online]. Available: https://www.deeplearningbook.org/

[9] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.