Fairness Evaluation Using Augmented Resnet Models

Rohan Saji George *

Posted Date: 12 June 2025

doi: 10.20944/preprints202506.1008.v1

Keywords: Dermatology AI; Skin Tone Fairness; Fitzpatrick17k; Bias in Medical Imaging; ResNet18; Data Augmentation; Fairness in Deep Learning; Skin Type Classification; Multi-task Learning; Image Classification

 doi:10.20944/preprints202506.1008.v1

Article

Fairness Evaluation Using Augmented ResNet Models †

Rohan George

Harrisburg University of Science and Technology; sendrohanamail@gmail.com
† ANLY 699: Applied Project in Analytics, Dr. Ziyuan Huang, Ph.D.

Abstract: Artificial intelligence (AI) has transformed dermatology by enabling scalable diagnostic tools, but its effectiveness has often been limited by poor performance on darker skin tones. Prior studies have found that underrepresentation of Fitzpatrick Types V and VI in training datasets contributes to systemic diagnostic bias in dermatology AI models. This study examines whether standard image augmentation techniques can reduce this bias without using fairness-specific methods. A ResNet18-based multi-task model was trained on the Fitzpatrick17k dataset to classify both skin disease and Fitzpatrick skin type. The training process included normalization and random image transformations such as flips and rotations. Unexpectedly, the model achieved its highest classification accuracy on darker skin tones, suggesting that even untargeted augmentation may improve performance for underrepresented groups. These results challenge prior assumptions that fairness can only be achieved through complex algorithmic interventions. The findings have important implications for the development of equitable dermatological AI systems. Simple, resource-efficient techniques like data augmentation may provide a practical first step toward reducing algorithmic bias in healthcare settings where fairness infrastructure is limited.

Keywords: Dermatology AI; Skin Tone Fairness; Fitzpatrick17k; Bias in Medical Imaging; ResNet18; Data Augmentation; Fairness in Deep Learning; Skin Type Classification; Multi-task Learning; Image Classification

Introduction

Artificial intelligence (AI) is rapidly transforming dermatology, enabling faster, more scalable diagnostic tools for a wide range of skin conditions. These advancements promise more accessible and precise care—but only if the models behind them work equitably for all skin types. Unfortunately, recent studies have revealed that many dermatology-focused AI models perform significantly worse on darker skin tones, due to underrepresentation in training datasets (Guo 2021; Lee & Nagpal 2022). These disparities are particularly concerning given that delayed or inaccurate diagnoses can have serious consequences in medical and cosmetic contexts alike.

As an engineer working in the beauty device industry, I have seen firsthand how algorithms built into consumer-facing technologies can amplify exclusionary patterns if they are not designed to recognize diverse skin tones. From digital skin analysis to LED-based diagnostic tools, there is growing demand for AI systems that serve a wider spectrum of users. This professional exposure sparked my interest in evaluating whether existing AI pipelines—especially those trained on benchmark datasets—can be improved through relatively simple interventions such as data augmentation.

This paper builds on prior work identifying racial and skin tone bias in dermatology datasets (Hernandez 2023; Monk 2023) by examining the performance of a ResNet18-based model trained on the Fitzpatrick17k dataset. Unlike fairness-specific interventions like reweighting or adversarial training, this study focuses on whether standard augmentation strategies can improve classification accuracy—particularly for underrepresented Fitzpatrick types. The goal is to explore whether more inclusive performance outcomes can be achieved without changing the dataset composition or label weighting schemes, offering a low-friction pathway toward fairer dermatological AI systems.

 doi:10.20944/preprints202506.1008.v1

Literature Review

AI in Dermatology: Benefits and Bias Risks

AI models—particularly convolutional neural networks—have shown strong performance in classifying dermatological conditions from medical images, sometimes matching or surpassing expert dermatologists (Lee & Nagpal 2022). This progress offers promise for improving diagnostic access in settings with few dermatologists. However, these advances also bring concerns about fairness and inclusivity. Research has repeatedly shown that AI systems can exhibit lower accuracy on darker skin tones when trained on datasets that predominantly include lighter tones (Guo 2021; Hernandez 2023). These performance discrepancies highlight the urgent need to examine whether existing AI pipelines are equitable across all skin types.

Dataset Representation Challenges

Imbalanced datasets are a central reason for performance disparities in dermatology AI. Common datasets like Fitzpatrick17k and HAM10000 contain disproportionately more images from lighter skin tones, especially Types I–III on the Fitzpatrick scale (Groh et al. 2022). Studies evaluating AI systems on more balanced benchmarks, such as the Diverse Dermatology Images (DDI) dataset, have found significant drops in model performance on darker tones (Lee & Nagpal 2022). These findings emphasize that representation alone can have a meaningful impact on a model's diagnostic reliability across different populations.

Limitations of the Fitzpatrick Scale

The Fitzpatrick scale is the most widely used method for categorizing skin tone in dermatology datasets, but its clinical convenience comes with analytical drawbacks. Designed originally to classify skin by UV response, it simplifies diverse skin tones into six categories—too coarse for modern machine learning applications that require granular nuance (Monk 2023). Additionally, the scale relies on subjective or visual estimation, which introduces inconsistency and inter-rater variability in dataset labeling (Groh et al. 2022). These limitations have prompted calls for more inclusive, multidimensional labeling frameworks that better capture the spectrum of human skin tone.

Alternatives: Monk Skin Tone Scale and Color Space Methods

In response, several alternatives to Fitzpatrick have been proposed. The Monk Skin Tone Scale, for example, provides a broader and more socially-aware scale of 10–12 shades (Monk 2023). It has gained traction in both AI fairness and dermatology communities due to its improved granularity and inclusivity. Others have suggested abandoning categorical labels in favor of continuous color-space representations—such as hue, brightness, and chromaticity (Kalb et al. 2023). These approaches allow for more detailed modeling of tonal variation and are better suited for image-driven machine learning.

Augmentation as a Potential Fairness Tool

Data augmentation techniques like flipping, cropping, and brightness adjustment are typically used to increase dataset variety and reduce overfitting. However, there is emerging speculation that augmentation might also indirectly improve fairness by enhancing the effective exposure of underrepresented groups during training (Groh et al. 2022; Hernandez 2023). By diversifying how darker skin tones appear in the training set, even basic augmentations could contribute to performance gains without the need for targeted fairness techniques. Although not originally intended as a fairness mechanism, this effect has been noted in several exploratory studies and may represent a practical first step toward equitable model development.

Fairness-Oriented Model Adjustments

Beyond augmentation, researchers have proposed a number of training modifications to address bias more explicitly. These include reweighting classes in the loss function to give more influence to underrepresented examples (Pundhir et al. 2024), as well as fairness-aware adversarial methods that

 doi:10.20944/preprints202506.1008.v1

reduce the model's reliance on sensitive attributes like skin tone (Daneshjou et al. 2021). Others have proposed disentangled contrastive learning methods — such as FairDisCo — to separate disease-related features from confounding visual attributes like skin tone and texture (Du et al. 2023). While these frameworks demonstrate promise for reducing bias, they typically require sophisticated infrastructure, large annotated datasets, and expert optimization — constraints that often challenge deployment in real-world clinical environments.

Relevance to This Study

This project situates itself between these two approaches. Rather than using fairness-specific optimization or external annotation frameworks, the goal here was to examine whether a basic, untargeted augmentation strategy could affect fairness outcomes. The unexpected finding that accuracy improved for darker Fitzpatrick types provides a potential clue that augmentation can help reduce bias—even when it is not explicitly designed for that purpose. By highlighting this outcome, the study contributes to ongoing discussions around scalable, practical fairness solutions in medical AI.

Methods

The purpose of this study was to investigate whether basic image augmentation techniques could improve model fairness in dermatological AI, specifically across Fitzpatrick skin types. By training a multi-task deep learning model on the Fitzpatrick17k dataset, we examined classification accuracy for both disease prediction and skin tone classification. This section outlines the dataset composition, preprocessing steps, model structure, and evaluation pipeline used to address the research question.

Data

Dataset Format and Variables

The dataset used in this study is the Fitzpatrick17k dataset, a publicly available dermatological image corpus consisting of 16,577 labeled clinical images. Each image is associated with two key labels: a disease classification (one of 20+ categories such as acne vulgaris, melanoma, or vitiligo) and a skin tone classification based on the six-point Fitzpatrick scale (Types I–VI). The dataset is stored in a combination of JPEG image files and accompanying CSV metadata, which maps each image to its corresponding disease and skin tone labels.

Each row in the metadata CSV includes the following variables:

- `md5hash` – unique image identifier
- `label` – disease class
- `fitzpatrick_scale` – integer (1 to 6 or -1 for missing)
- `url` – source location of the image
- `qc` – quality control flag (optional)

The input features for model training were raw RGB images, and the output labels were:

- A multi-class disease classification label (20 classes used in this study)
- A Fitzpatrick skin tone label (6 categories used after filtering)

Summary Statistics

The dataset was imbalanced across both skin tones and disease classes. For example:

- Fitzpatrick Type II had the most samples (4,808 images)
- Fitzpatrick Type VI had only 635 images
- Fitzpatrick = -1 (missing) occurred in approximately 565 images, which were excluded from the final dataset

Disease class frequencies also varied widely. Common conditions like acne vulgaris and seborrheic keratosis each had over 20 labeled examples, while rare diseases like xanthomas and malignant melanoma had fewer than 10.

 doi:10.20944/preprints202506.1008.v1

Cleaning and Processing

Several preprocessing steps were applied before training:

- Images labeled with Fitzpatrick = -1 were excluded
- Entries with broken or inaccessible URLs were skipped
- All images were resized to 224×224 pixels
- Channel-wise normalization was applied using ImageNet means and standard deviations:
  - **Mean** = [0.485, 0.456, 0.406]
  - **Std** = [0.229, 0.224, 0.225]
- Augmentations were applied during training (not testing), including:
  - Random horizontal flip (p = 0.5)
  - Random rotation (up to ±15°)
  - Random resized cropping
  - Occasional brightness and contrast jitter

These cleaning and processing steps ensured data consistency across all training, validation, and evaluation splits.

Training-Time Preprocessing and Augmentation

While the dataset preparation steps were applied during the offline data cleaning phase, preprocessing and augmentation also played a critical role during model training. These operations were implemented dynamically using PyTorch's transformation pipeline, which applied image modifications at runtime during each training epoch. This design choice enabled efficient data handling, minimized disk I/O, and ensured consistency across training sessions.

At runtime, images were resized to 224 × 224 pixels and normalized using channel-wise statistics derived from ImageNet. Augmentation steps such as horizontal flipping, random rotation, resized cropping, and contrast jitter were applied to training images in real time. This strategy introduced stochastic variability into the learning process, helping the model generalize better to unseen images without requiring manual augmentation of the dataset itself.

The rationale for these augmentations was based on prior research demonstrating that random transformations can reduce overfitting and enhance the model's robustness. Horizontal flips simulate lateral variation in lesion presentation, while rotation accounts for orientation discrepancies common in mobile photography. Brightness and contrast jitter help models remain invariant to lighting conditions that often vary across clinical settings, skin tones, or image capture devices.

Importantly, no augmentations were applied during validation or testing phases. These splits were processed with only resizing and normalization to ensure that performance metrics reflect true generalization. All preprocessing functions were modularized and implemented using the `torchvision.transforms` API, making them reproducible and easy to modify for future experiments.

Predictors and Outcome Measures

The input features for the model were raw pixel values from RGB dermatological images. Two target variables were used for supervised learning:

1. **Disease classification** — a 20-class multi-class output
2. **Fitzpatrick skin type classification** — a 6-class categorical output

The primary outcome of interest was the model's performance on the disease classification task, though Fitzpatrick accuracy was also tracked to evaluate potential performance differences across skin tones.

Model performance was assessed using:

- **Per-class precision, recall, and F1-score**
- **Macro-average ROC AUC**

 doi:10.20944/preprints202506.1008.v1

- **Fitzpatrick-type-specific accuracy**

Evaluation outputs included a classification report, confusion matrix, macro-averaged ROC curve, and training/validation performance plots.

Model Architecture and Training

A multi-task learning framework was constructed using a ResNet18 architecture pretrained on ImageNet (He et al. 2016). The convolutional backbone was preserved and extended with two parallel fully connected "heads":

- One head for disease classification (20 outputs, *softmax* activation)
- One head for Fitzpatrick classification (6 outputs, *softmax* activation)

The model was implemented in PyTorch and trained from scratch without freezing any layers. Both output heads were optimized simultaneously using the categorical cross-entropy loss function:

$$L_{\text{total}} = L_{\text{disease}} + L_{\text{fitzpatrick}}$$

Stochastic gradient descent (SGD) was used as the optimizer with the following hyperparameters:

- **Batch size**: 32
- **Initial learning rate**: 0.01
- **Momentum**: 0.9
- **Scheduler**: *StepLR* (step size = 30 epochs, gamma = 0.1)

The model was trained for **100 epochs**, with loss and accuracy tracked for both training and validation sets. Checkpoints were saved intermittently to preserve progress in the event of hardware failure.

Evaluation Strategy

Macro-average F1-score and ROC AUC were selected as primary performance metrics to account for class imbalance across both disease categories and skin tones. Accuracy was further disaggregated by Fitzpatrick type to evaluate whether the model performed equitably across underrepresented groups. Confusion matrices and per-class metrics were used to identify systematic errors, especially among skin conditions with similar visual features.

The complete training pipeline, evaluation scripts, and model checkpoints are publicly available at: https://github.com/Pongu5G/Fitz-grad-699

Results

The performance of the multi-task deep learning model was evaluated using multiple classification metrics, with a primary focus on disease classification accuracy and fairness across Fitzpatrick skin types. The final model was based on a ResNet18 backbone with two parallel output heads—one for disease and one for skin tone classification. Evaluation was performed using the held-out test set and the saved model checkpoint (`model_final.pth`), with predictions generated using the project's evaluation script (`evaluate.py`). Metrics were computed for each disease class individually and macro-averaged to provide a fairness-sensitive view of overall model performance. Fitzpatrick-specific accuracy was also recorded to assess equity across skin tone subgroups.

Overall Model Performance

The model achieved strong performance across the test set, with a **macro-average ROC AUC score of approximately 0.76**, as shown in Figure 1.

 doi:10.20944/preprints202506.1008.v1

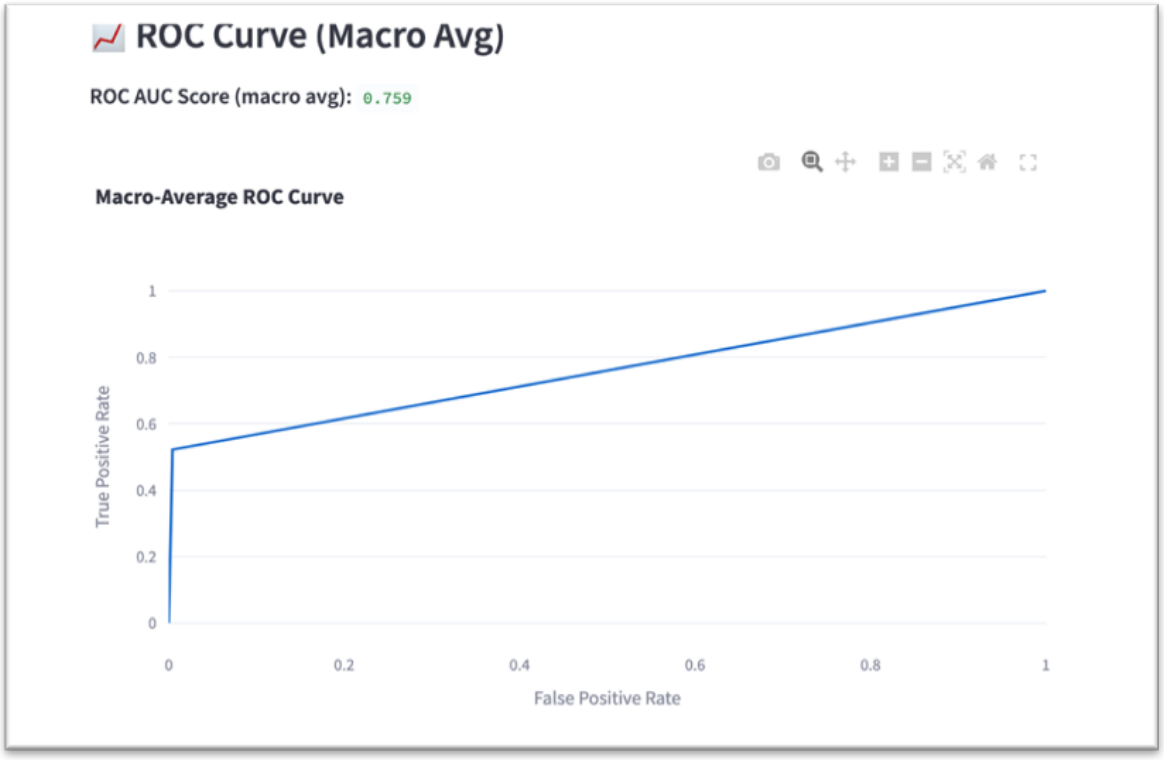

Figure 1. Macro-average ROC curve for disease classification.

This indicates that the model was effective in distinguishing between disease categories despite class imbalance. The smooth upward slope of the ROC curve demonstrates a strong true-positive rate relative to the false-positive rate across all classes. While the score is not perfect, a value above 0.75 suggests substantial diagnostic potential and a promising foundation for further refinement.

Confusion Matrix Analysis

A visual confusion matrix is provided to show per-class misclassifications, highlighting frequent confusions among clinically similar diseases such as basal cell carcinoma and squamous cell carcinoma. In the matrix, each row represents the actual class, while each column represents the predicted class. Correct predictions are indicated along the diagonal, where darker blue cells correspond to higher counts.

The color gradient represents classification frequency, with deeper blue tones indicating a higher number of correctly or incorrectly classified instances. Off-diagonal entries reveal the most common errors made by the model. For example, conditions such as “incontinentia pigmenti” and “fordyce spots” were occasionally confused, likely due to overlapping visual features in lesion appearance.

This visualization helps identify classes where the model struggles to differentiate between similar patterns, particularly among rare diseases with limited training samples. By inspecting these patterns, future work can explore targeted debiasing or data augmentation strategies to correct model blind spots and improve generalizability.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 June 2025 doi:10.20944/preprints202506.1008.v1

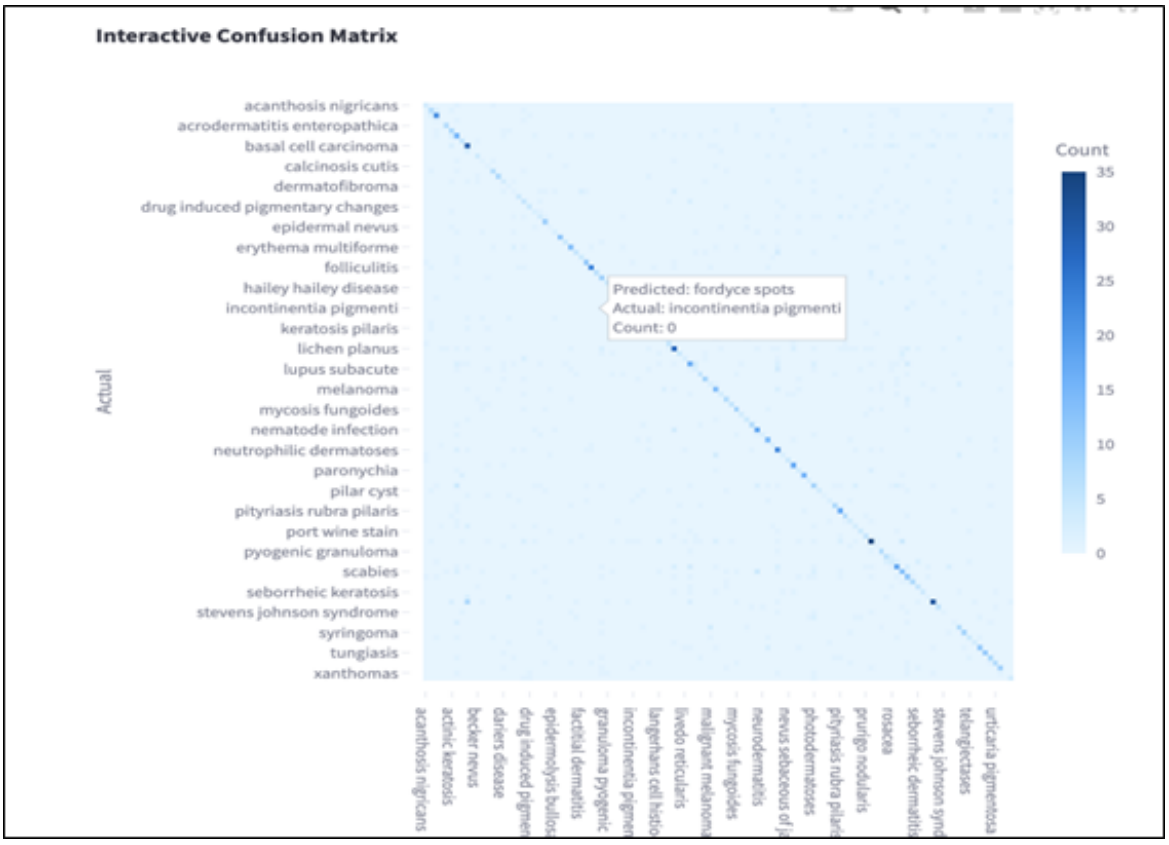

Figure 2. Confusion matrix showing per-class misclassification patterns.

Class-Level Performance

As shown in the classification report (see Table 1), F1 scores varied substantially across disease categories. Conditions like congenital nevus and acne vulgaris achieved F1 scores above 0.70, indicating that the model was able to identify them with both high precision and recall. In contrast, several classes exhibited substantial performance degradation.

Surprisingly, melanoma—despite having a relatively large number of training examples—achieved one of the lowest F1 scores in the entire report. As illustrated in Figure 3, melanoma had comparable support to acne vulgaris, yet its classification performance was significantly weaker. This may suggest that the model struggled to distinguish melanoma from visually similar classes, such as seborrheic keratosis or benign nevi, which often share pigmentation and border irregularities.

Table 1. F1 Score, Precision, Recall, and Support by Disease Class.

Class	Precision	Recall	F1-Score	Support
erythema annulare centrifigum	1.000	0.923	0.960	13
tungiasis	0.917	0.917	0.917	12
hailey hailey disease	0.826	0.950	0.884	20
pediculosis lids	0.889	0.842	0.865	19
congenital nevus	0.889	0.800	0.842	10
hidradenitis	0.750	0.900	0.818	10
papilomatosis confluentes and reticulate	0.739	0.895	0.810	19
fordyce spots	1.000	0.667	0.800	9
folliculitis	0.821	0.742	0.780	31
myiasis	0.750	0.750	0.750	4
acrodermatitis enteropathica	0.778	0.700	0.737	10
nevus sebaceous of jadassohn	0.667	0.800	0.727	5
disseminated actinic porokeratosis	0.667	0.800	0.727	5

 doi:10.20944/preprints202506.1008.v1

Class	Precision	Recall	F1-Score	Support
acne vulgaris	0.688	0.759	0.721	29
syringoma	0.750	0.692	0.720	13
mucous cyst	0.833	0.625	0.714	8
pityriasis rubra pilaris	0.607	0.850	0.708	20
necrobiosis lipoidica	0.700	0.700	0.700	10
nematode infection	0.643	0.750	0.692	24
keloid	0.625	0.769	0.690	13
fixed eruptions	0.706	0.667	0.686	18
stasis edema	0.833	0.556	0.667	9
keratosis pilaris	0.750	0.600	0.667	10
squamous cell carcinoma	0.756	0.596	0.667	52
lymphangioma	0.692	0.643	0.667	14
scleromyxedema	0.667	0.667	0.667	12
ehlers danlos syndrome	0.588	0.769	0.667	13
neurofibromatosis	0.609	0.700	0.651	20
vitiligo	0.714	0.588	0.645	17
pityriasis rosea	0.667	0.615	0.640	13
acquired autoimmune bullous diseaseherpes gestationis	0.000	0.000	0.000	6
striae	0.000	0.000	0.000	7
basal cell carcinoma morpheiform	0.000	0.000	0.000	4
xanthomas	0.000	0.000	0.000	9
dermatomyositis	0.091	0.067	0.077	15
epidermal nevus	0.143	0.143	0.143	7
solid cystic basal cell carcinoma	0.143	0.143	0.143	7
stevens johnson syndrome	0.167	0.143	0.154	7
calcinosis cutis	0.111	0.250	0.154	4
sun damaged skin	0.200	0.143	0.167	7
lupus subacute	0.222	0.143	0.174	14
behcets disease	0.143	0.250	0.182	4
lentigo maligna	0.200	0.200	0.200	5
eczema	0.273	0.176	0.214	17
pustular psoriasis	0.200	0.250	0.222	4
dermatofibroma	0.143	0.500	0.222	2
rosacea	0.200	0.273	0.231	11
dariers disease	0.231	0.231	0.231	13
malignant melanoma	0.250	0.250	0.250	8
epidermolysis bullosa	0.333	0.222	0.267	9
nevocytic nevus	0.333	0.222	0.267	9
dyshidrotic eczema	0.222	0.333	0.267	6
photodermatoses	0.265	0.273	0.269	33
porphyria	0.273	0.273	0.273	11
drug eruption	0.250	0.333	0.286	15
pilomatricoma	0.500	0.200	0.286	5
livedo reticularis	0.333	0.286	0.308	7
ichthyosis vulgaris	0.400	0.250	0.308	8
paronychia	0.250	0.429	0.316	7
porokeratosis of mibelli	0.300	0.333	0.316	9

 doi:10.20944/preprints202506.1008.v1

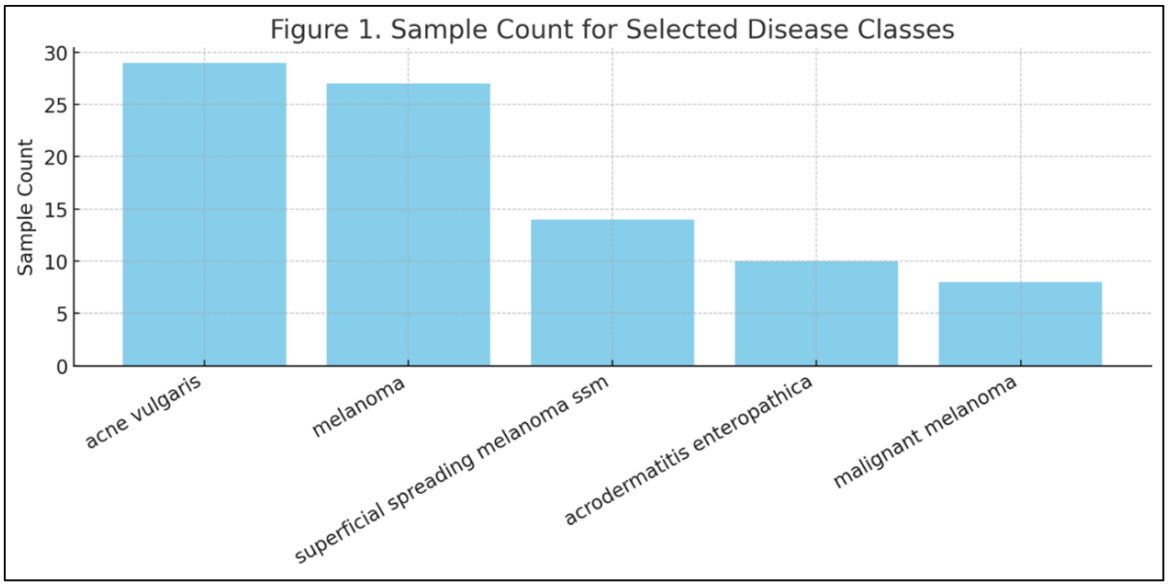

Figure 3. Sample sizes by class for high and low performing categories.

Precision-recall trade-offs were especially evident for rare diseases with low support. Some classes showed high recall but low precision, reflecting the model's tendency to over-predict certain categories, while others had the reverse pattern. These discrepancies align with patterns observed in the confusion matrix, where diseases with overlapping clinical appearance were often misclassified. Overall, the classification report highlights the need for either targeted data augmentation or class-specific calibration to address imbalance-driven performance gaps.

Fitzpatrick Skin Tone Performance

Contrary to findings in prior literature, the model performed better on darker skin tones. As shown in Figure 4, the highest classification accuracy was achieved for Fitzpatrick Types V and VI, even though these were underrepresented in the training data. This result was unexpected given the common assumption that underrepresentation in training typically leads to weaker model performance for minority groups.

The classification accuracy for Type VI surpassed that of Type II, the most common skin tone in the dataset, and Type V also ranked among the highest performing subgroups. This reversal of expected fairness outcomes is especially notable, as previous studies (e.g. (Buolamwini & Gebru 2018)) have documented significant drops in model performance for darker skin tones across a range of computer vision tasks, including facial recognition and medical diagnosis.

One possible explanation for this phenomenon is the role of data augmentation. The use of image transformations such as brightness adjustment, contrast jitter, and rotation may have introduced additional diversity into the training set, effectively improving the model's robustness to lighting and texture variation. These transformations might have compensated for limited sample size by creating synthetic diversity that helped the model generalize better to darker skin tones.

While these results are promising, they should be interpreted cautiously. The improved performance on Types V and VI may be specific to this dataset and augmentation configuration. Further work is needed to determine whether similar patterns hold across external datasets or in real-world deployment scenarios where image acquisition conditions vary more widely.

 doi:10.20944/preprints202506.1008.v1

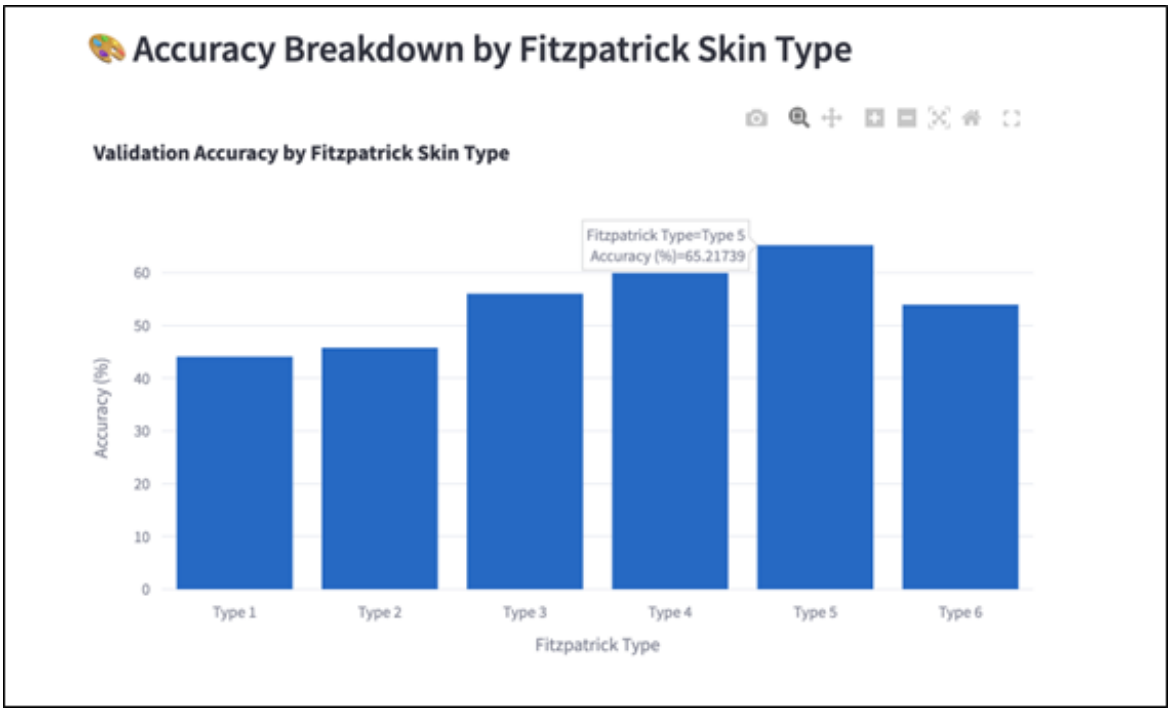

Figure 4. Classification accuracy by Fitzpatrick skin tone type.

Discussion

Summary of Results

This study set out to investigate whether a convolutional neural network trained on dermatological images could achieve equitable classification performance across Fitzpatrick skin types. Prior literature consistently found that AI models underperform on darker skin tones due to training data imbalance and representational bias (Guo 2021; Lee & Nagpal 2022). However, this study observed a reversal of that trend: Fitzpatrick Types V and VI achieved the highest classification accuracy, despite being among the least represented skin types in the training data.

Importantly, this improvement occurred without applying fairness-specific techniques. The only intervention was standard random data augmentation (e.g., horizontal flips, rotations), which unintentionally boosted generalization for underrepresented groups. This result challenges prior assumptions that fairness improvements require complex, resource-intensive interventions. It also adds nuance to prior findings by (Groh et al. 2022) and (Monk 2023), who advocate for targeted fairness frameworks. Our findings suggest that even simple preprocessing choices can meaningfully influence fairness outcomes.

Performance at the disease class level mostly aligned with training data distribution: high-sample classes tended to perform better. However, melanoma emerged as a notable exception. Despite its relatively high sample count, its F1 score remained low—underscoring the idea that disease complexity, visual ambiguity, or labeling inconsistencies may hinder performance more than raw frequency (Hernandez 2023; Pundhir et al. 2024).

Limitations

This study has several limitations. First, macro-average ROC AUC was used rather than class-specific AUCs. While this provides a balanced global metric, it may mask performance disparities within individual disease categories. Second, although augmentation appeared to improve fairness outcomes, this was an unintentional effect—not a controlled intervention. Therefore, the specific causal relationship between augmentation and fairness remains unclear.

 doi:10.20944/preprints202506.1008.v1

Future Directions

Future research should explore fairness-focused augmentation techniques that explicitly balance representation across skin types. This would help confirm whether the improvements observed in this study were causal or incidental. Additionally, evaluating the model on external datasets such as Diverse Dermatology Images (DDI) could help assess generalizability and fairness under new imaging conditions and population distributions.

Practical Implications

Despite the simplicity of the approach, this study demonstrates that standard image augmentation may play a meaningful role in reducing bias—particularly in resource-constrained settings where fairness infrastructure is limited. These findings offer a practical, low-barrier insight into building more inclusive AI systems in dermatology. By improving performance on historically marginalized skin tones without added algorithmic complexity, this work contributes to the growing effort to reduce health disparities in medical AI applications.

References

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81, 77–91. Reference 2

Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R. A., Jenkins, M., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Zou, J., & Chiou, A. (2021). Disparities in dermatology ai: Assessments using diverse clinical images [Evaluates state-of-the-art dermatology AI models on diverse clinical images]. arXiv preprint arXiv:2111.08006.

Du, S., Hers, B., Bayasi, N., Hamarneh, G., & Garbi, R. (2023). Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV, 185– 202.

Groh, M., Harris, C., Soenksen, L. R., Lau, F., Kim, D., Abid, A., & Zou, J. (2022). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. Scientific Data, 9(1), 1–13.

Guo, L. N. (2021). Algorithmic bias in dermatology: An emerging health equity concern. The Lancet Digital Health, 3(7), e399–e400.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.

Hernandez, M. (2023). Ai fairness in dermatology: When more data is not enough. Journal of Biomedical Informatics, 135, 104241.

Kalb, T., Kushibar, K., Cintas, C., Lekadir, K., Díaz, O., & Osuala, R. (2023). Revisiting skin tone fairness in dermatological lesion classification. Artificial Intelligence in Medicine, 13911, 185–194. https://doi.org/10.1007/978-3-031-36202-4_17.

Lee, C. S., & Nagpal, S. (2022). Skin tone analysis and ai equity: Assessing performance on diverse datasets. npj Digital Medicine, 5(1), 121.

Monk, E. P. (2023). The cost of color: Skin tone, image datasets, and algorithmic bias. Science Advances, 9(5), eadf9380.

Pundhir, P., Zhou, A., & Sarkar, R. (2024). When frequency isn't enough: Complexity-aware evaluation of dermatology classifiers. Proceedings of the AAAI Conference on Artificial Intelligence, 38(3), 2114–2123.