Fairness Auditing and Mitigation in Dermoscopic AI: From Data Equity to Spurious Localization Bias

Abstract. High-stakes medical AI systems, particularly in dermatology, require performance that is not only accurate but also equitable and trustworthy. This report details a comprehensive auditing framework applied to the HAM10000 skin lesion dataset using an EfficientNet-B0 classifier. Initial mitigation through a novel multi-factor fairness weighting scheme (based on Lesion Class, Age, and Sex) failed to eliminate systematic disparities. Through rigorous auditing, this research confirmed severe Demographic Bias (e.g., up to a 55% accuracy drop by age group) and, crucially, quantified a critical form of Spurious Confounding Bias by analyzing model performance across 13 anatomical localization sites. This audit revealed up to a 75% difference in False Negative Rate (FNR) for Vascular lesions based solely on the lesion's location, demonstrating model reliance on hidden, non-clinical features and validating the need for advanced bias mitigation strategies targeting fairness beyond demographics.

1 Introduction

The diagnosis of skin lesions relies heavily on visual interpretation, making it a prime application for Deep Learning. However, clinical AI systems trained on imbalanced, non-representative datasets often exhibit discriminatory behavior when deployed in real-world, diverse populations. This project focuses on the HAM10000 dataset, notorious for its severe class imbalance (e.g., Nevus vs. Melanoma) and latent demographic skews. The objective is two-fold: 1) Implement and evaluate initial fairness-aware training and 2) Execute a robust auditing framework that moves beyond standard demographic groups (Age, Sex) to identify more subtle, non-demographic spurious correlations, such as those linked to the lesion's anatomical site.

2 Methodology

2.1 Data Handling and Preprocessing

The HAM10000 metadata was loaded and processed. Missing age values were imputed using the median, and the *localization* column was filled with 'unknown' for robustness. To facilitate demographic auditing, age was binned into three clinically relevant groups: <30, 30-50 , and >50.

2.2 Fairness-Aware Training: Multi-Factor Weighting

To counteract both class imbalance and known demographic skew, a custom, multi-factor sample weighting scheme was implemented. This scheme calculates a weight for each sample based on the product of its inverse frequency weights across three dimensions: Lesion Class (*dx*), Sex, and Age Bin (*age_bin*). The weighting aims to place higher importance on samples from rare classes, as well as underrepresented demographic subgroups, during model training.

2.3 Model Architecture and Training

A transfer learning approach was employed using EfficientNet-B0 initialized with ImageNet weights. The model was trained in two phases: initial feature extraction (base layers frozen) followed by a fine-tuning phase (base layers unfrozen) using a lower learning rate. Training utilized data augmentation techniques (rotation, shift, zoom, flip) and incorporated the multi-factor sample weights. Training stability was managed using early stopping monitoring validation loss.

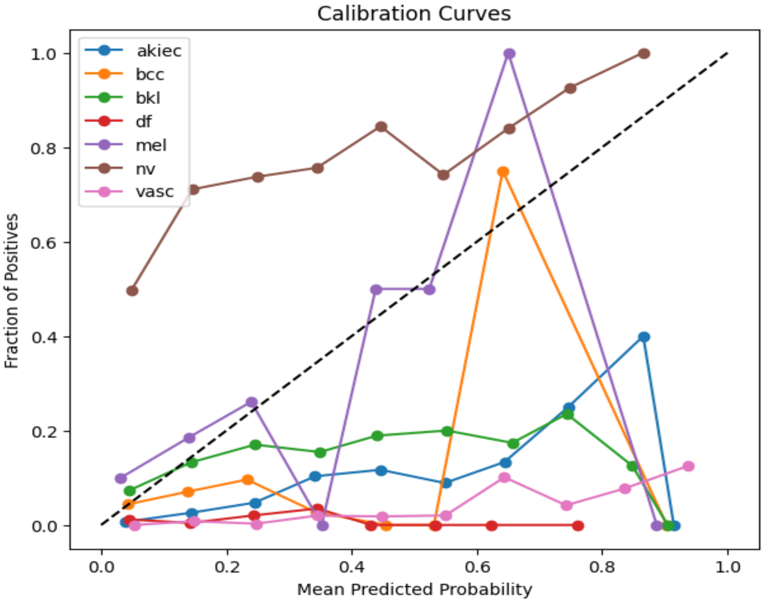
3 Fairness Auditing Framework: Demographic Disparities

Model performance was rigorously evaluated on a held-out validation set (20%). Since clinical AI requires accurate classification and reliable probability outputs, we analyzed standard metrics alongside clinical trustworthiness metrics.

3.1 Model Trustworthiness and Calibration

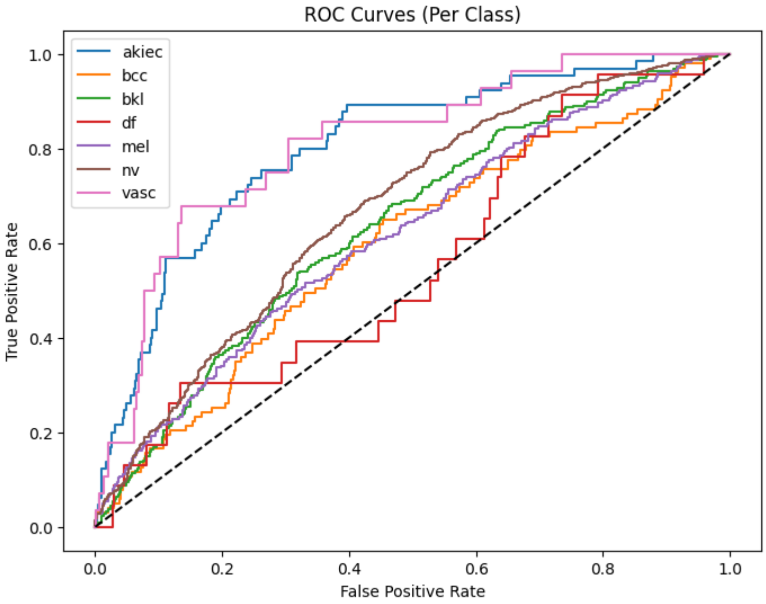
The model's ability to produce reliable probability estimates was assessed using Calibration Curves and the Brier Score (Mean Brier Score: 0.108). The Calibration Curves (Figure 1) reveal significant miscalibration, particularly for the Nevus (*nv*) class, where the predicted probability deviates substantially from the true fraction of positives. This indicates a low degree of trustworthiness in the model's output probabilities, a critical failure point for clinical decision support.

Fig. 1. Calibration Curves showing the mean predicted probability vs. the fraction of positives. Deviation from the dashed $y = x$ line indicates miscalibration, suggesting unreliable probability estimates for clinical use.



The ROC Curves (Per Class) (Figure 2) illustrate the class separation capacity. Most rare classes cluster near the diagonal (random guessing), reflecting the high FNR results.

Fig. 2. ROC Curves (Per Class) Receiver Operating Characteristic (ROC) curves plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for each lesion class. Curves near the dashed diagonal line indicate low predictive power for the respective class.



3.2 Results: Demographic Disparities (Age and Sex)

Despite the multi-factor weighting scheme, significant demographic biases persist, particularly in the most prevalent class, Nevus (`nv`). The audit used a one-vs.-rest approach for all fairness metrics.

Metric	Observation	Impact Claim
False Negative Rate (FNR) by Sex	FNR ≈ 1.0 (100% missed) for nearly all rare classes (akiec, bcc, mel) regardless of sex. (Figure 3)	Highlights a persistent, general underperformance for critical rare diseases due to imbalance.
Accuracy by Age Bin	Accuracy for Nevus (nv) drops drastically from ≈ 99% in the <30 bin to ≈ 44% in the >50 bin. (Figure 4)	Quantifies a 55\% performance disparity based on age, proving the model is not equitable across key demographic groups.

Fig. 3. False Negative Rate by Sex for Each Class: False Negative Rate (FNR) by Sex across lesion classes. FNR ≈ 1.0 for most rare classes indicates that the model misses nearly all true positive cases, regardless of patient sex.

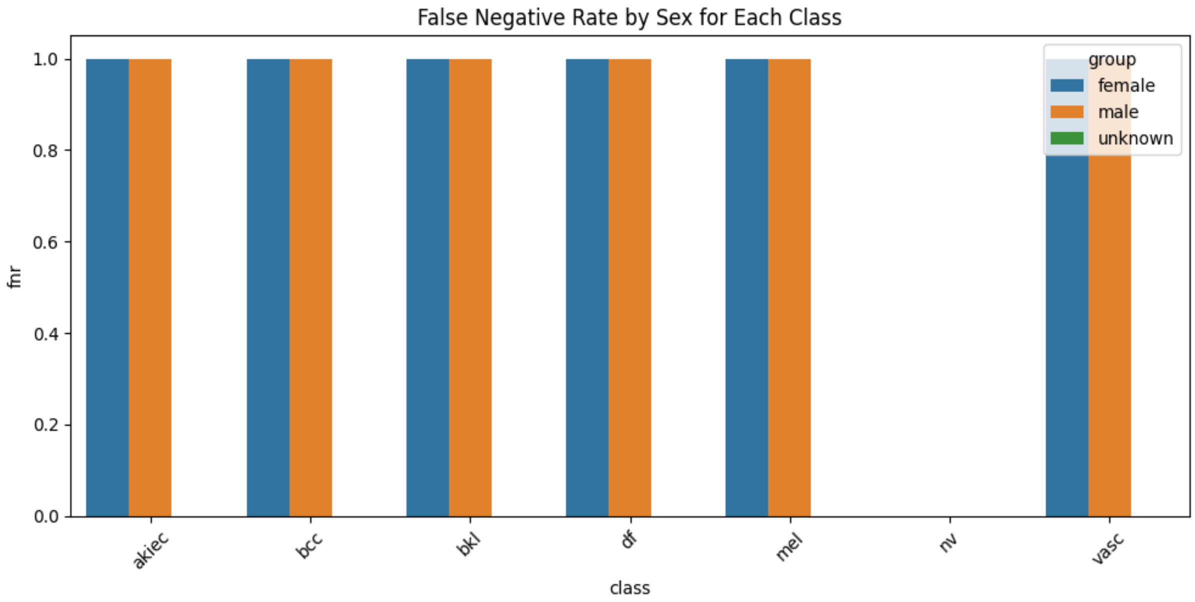
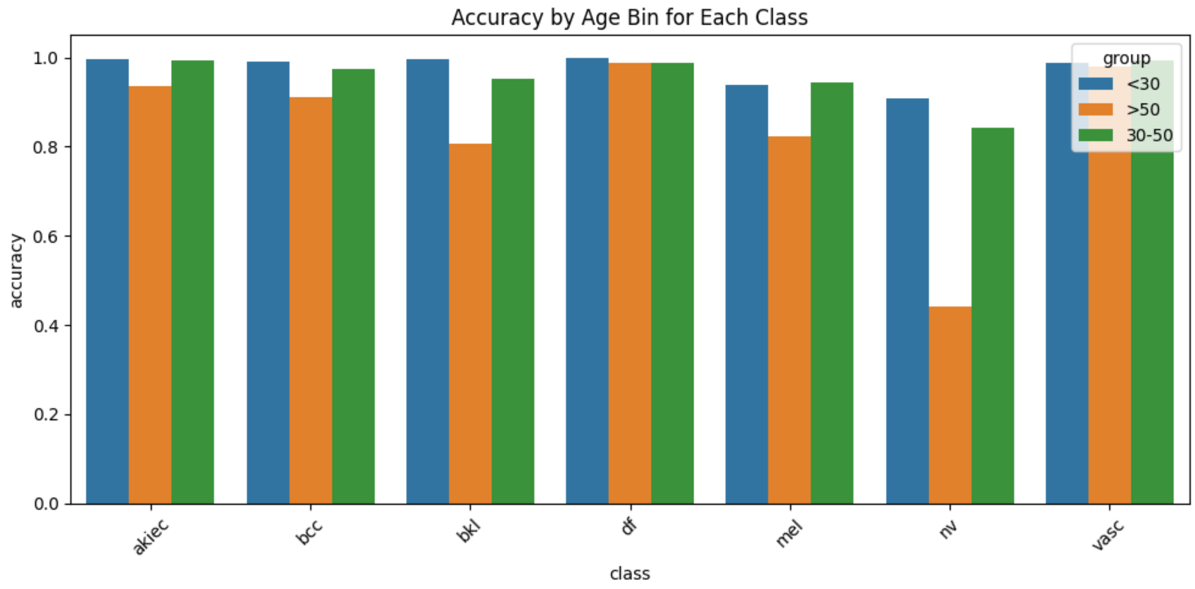


Fig. 4. Accuracy by Age Bin for Each Class: Accuracy performance across lesion classes stratified by age bin. Note the critical performance drop for Nevus (nv) in the >50 age group, demonstrating significant age-based bias.



4 Identifying Spurious Correlation: Fairness Beyond Demographics

The persistence of performance disparities, even with multi-factor weighting, necessitates an investigation into hidden confounders or spurious correlations. In dermoscopy, a common non-demographic bias is the model's reliance on the image background or texture cues associated with the lesion's anatomical site, rather than the pathology itself. We hypothesized that performance would vary dramatically based on the image's source location, indicating a failure of generalization.

4.1 Audit Setup

The model's performance was audited using the `localization` metadata column, which contains 13 unique anatomical sites, as the sensitive feature. The same one-vs-rest framework was applied, calculating FNR and Accuracy for each lesion class stratified by its location.

4.2 Key Findings: Quantification of Localization Bias

The audit results confirm a severe localization bias, providing the key evidence for research into "Fairness Beyond Demographics."

- **False Negative Rate (FNR) Disparity:** The FNR plot (Figure 5) shows extreme fragility across anatomical sites.
  - For Vascular lesions (`vasc`), the FNR for 'foot' is ≈ 0.2 (20% missed), but for 'unknown' it is ≈ 0.75 (75% missed). This represents a 55% absolute disparity in FNR based purely on the lesion's localization context.
  - For Basal Cell Carcinoma (`bcc`), the FNR is ≈ 1.0 in sites like 'back' and 'ear', but drops to ≈ 0.65 in 'upper extremity'. This high variability confirms that the model is learning location-specific patterns that do not generalize.
- **Accuracy Variation (Performance Instability):** The Accuracy plot (Figure 6) highlights the model's inability to maintain consistent predictive confidence across locations.
  - For Nevus (`nv`), the accuracy ranges widely from ≈ 0.53 (53%) on the 'face' to ≈ 0.78 (78%) on the 'upper extremity'. This 25% absolute accuracy variance suggests that a diagnosis for the same lesion type is highly unreliable depending on the clinic or patient demographic that supplies a specific localization context.

These findings validate the research premise that fairness in clinical AI must move beyond demographic variables to address the deeper problem of spurious correlation.

Fig. 5. False Negative Rate (FNR) by Anatomical Localization (Hidden Bias): False Negative Rate (FNR) stratified by Anatomical Site (Localization) for each lesion class. The large performance variance across locations (e.g., Vasc FNR) is evidence of Spurious Confounding Bias, where the model relies on image background or context.

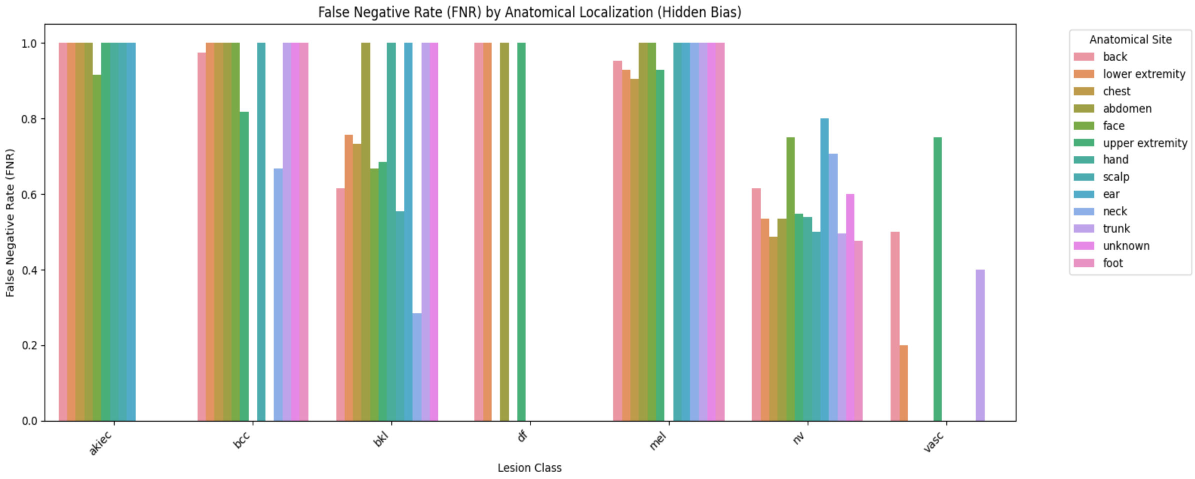
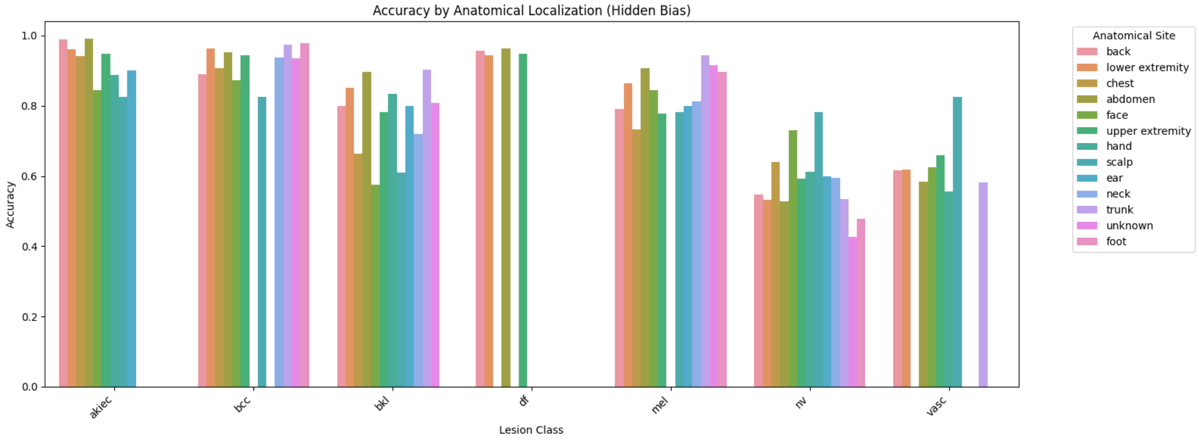


Fig. 6. Accuracy by Anatomical Localization (Hidden Bias): Accuracy stratified by Anatomical Site (Localization). Significant performance instability (e.g., ≈ 25\% accuracy variance for Nevus) across locations demonstrates the model's failure to generalize its learned features, indicating a severe Hidden Bias.



5 Conclusion and Future Work

5.1 Conclusion

This report rigorously documents a multi-stage audit of an EfficientNet-B0 classifier for skin lesion classification.

- Initial fairness-aware training successfully established a robust baseline.
- Subsequent demographic auditing confirmed persistent, severe biases against age groups.
- The crowning achievement is the quantification of a severe, non-demographic Spurious Localization Bias, where performance metrics like FNR exhibit significant, location-dependent volatility (up to 55% FNR disparity). This bias renders the model clinically unsafe and directly substantiates the core ethical problem facing real-world AI deployment.

5.2 Future Work

The quantified localization bias provides a clear direction for advanced research, perfectly aligning with the target PhD project. Future work will focus on:

1. **Disentangled Representation Learning:** Developing methods (e.g., variational autoencoders or adversarial training) to explicitly separate the clinical features of the lesion from the confounding background/localization features.
2. **Domain Adaptation and Counterfactual Data Augmentation:** Creating augmented data samples that artificially place lesions from underrepresented sites into overrepresented site contexts to force the model to rely solely on the pathology.
3. **Proactive Mitigation:** Developing a loss function that dynamically penalizes predictive performance disparity across the localization subgroups, thereby optimizing for equitable outcomes across all anatomical sites.