Fair Skin Lesion Classification

Kedarnath P
MTech Computer Science & Automation
(SR NO: 20902)
kedarnathp@iisc.ac.in

Abstract

In recent years, deep learning models have shown significant promise in the field of skin lesion classification. However, biases present in the training data can lead to unfair and inaccurate performance across different demographic groups and body regions. In this study, we investigate the impact of class bias, age bias, gender bias, and localization bias on skin lesion classification performance and propose various mitigation techniques to address these biases. We employ data augmentation(oversampling), adversarial training, and class balance loss to improve the fairness and accuracy of our model. Our results demonstrate that these techniques effectively mitigate class bias, age bias, and gender bias, leading to improved performance across all demographic subgroups, localizing subgroups, and among classes.

1 Introduction

Skin lesions are a common medical concern, and accurate diagnosis is crucial for effective treatment. In the field of dermatology, early detection of skin lesions can lead to prompt diagnosis and treatment of skin diseases such as skin cancer. Skin cancer is the most prevalent type of cancer, with melanoma being one of its deadliest forms. Early detection of melanoma is essential for effective treatment, but accurate diagnosis can be challenging.

Deep learning algorithms have shown promise in the computer-aided diagnosis of skin lesions. However, skin lesion datasets used for training these algorithms are often biased towards certain skin types, ages, and ethnicities. This can result in poor performance of lesion detection algorithms on underrepresented populations, leading to disparities in health outcomes. To address this issue, there is a need for a reliable, fair, and accurate system for skin lesion classification that avoids biases and discrimination based on demographic and other hidden features.

2 Problem Statement

Despite advances in computer-aided diagnosis of skin lesions, existing systems may not be free from bias and can lead to unequal outcomes for different population groups. Discrimination based on demographic factors such as gender, age, skin color, and ethnicity can occur, negatively affecting diagnostic accuracy and fairness. Addressing these biases and ensuring fairness in skin lesion classification is critical for achieving equitable health outcomes for all individuals.

3 Project approach

1. **Bias Investigation**
    - Perform data analysis and investigate the presence of bias in the skin lesion dataset.
    - Explore the impact of various biases, such as class imbalance and feature imbalance, on model performance.
2. **Bias Mitigation**
    - Implement various bias mitigation techniques, such as, pre-processing (Oversampling) the data to balance class distributions, and using algorithmic techniques (adversarial learning) to reduce hidden feature biases in the model.
    - Evaluate the effectiveness of bias mitigation techniques.

Our project approach for "Fair Skin Lesion Classification" involves using a pre-trained Convolutional Neural Network (CNN) model, fine-tuning it for the skin lesion classification task, and evaluating the model's performance on a balanced test dataset. This approach aims to minimize biases and ensure fairness in the classification of skin lesions. The following steps outline our methodology:

3.1 Experiment Setting

Task: Multi-class (7) classification problem. Given an Image of Skin lesion, We need to predict the skin lesion type (class) for the diagnosis.

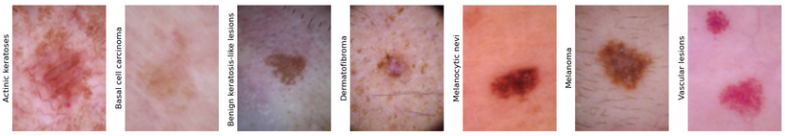

Figure 1: A sample image from Dataset

We will use the DenseNet121 model, which is a popular CNN model pre-trained on the ImageNet dataset. DenseNet121 has been widely used for skin lesion classification tasks in various research papers, demonstrating its suitability for our experiment.

We will use the following Dataset: HAM10000
The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, will be used for this project. This dataset will serve as the basis for both investigating biases and developing a fair skin lesion classification model that avoids discrimination based on demographic factors.
Note that, the metadata is also provided along with the dataset which contains the information like image_id, skin_lesion_type_id, age, gender, localization, etc.

3.2 Model Fine-tuning

We will fine-tune the DenseNet121 model on our training dataset, adapting the model to the specific task of skin lesion classification. This process will involve updating the model's parameters using the training data, allowing the model to learn the features and patterns unique to skin lesions.

3.3 Model Evaluation

To ensure fairness in our model's performance, we will test the fine-tuned DenseNet121 on a balanced test dataset and perform a demographic disparity check across various factors. These factors can be considered hidden features causing bias in the model. We test the model on the following:

- Overall test data: Evaluating the model's performance on the entire balanced test dataset.
- Gender: Assessing the model's performance separately for male and female test data.
- Age groups: Analyzing the model's performance across four age groups: below 20 years, 20-40 years, 40-60 years, and above 60 years on the same test set.

- Localization: Examining the model's performance on skin lesions located in different body regions, including upper, mid, and lower body areas.
- By following this approach, we aim to develop a fair and accurate skin lesion classification model that minimizes biases and ensures equitable performance across various demographic factors.

Implementation (Github repository) : LINK

4 Bias Investigation

4.1 Class Bias Investigation

The training dataset is imbalanced, with a long-tailed distribution of class labels. This imbalance may lead the model to be biased towards specific classes, especially those with a higher number of samples.

Figure 2: Imbalanced Training Dataset

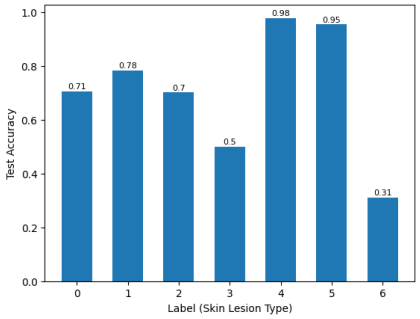

Figure 3: Model performance on balanced Test dataset

From the test accuracy results, we can observe that the model is biased towards specific classes, especially class label 4 (Melanocytic nevi), which is the majority class in the training dataset. The model has low test accuracy for class label 6 (Melanoma), the deadliest form of skin cancer, indicating that the class imbalance has negatively impacted the model's performance.

Analysis: The class imbalance in the training dataset has led to biases in the model's performance, resulting in higher test accuracy for majority classes and lower test accuracy for minority classes. This is particularly concerning for class label 6 (Melanoma), as accurate detection is crucial for early diagnosis and treatment of this deadly skin cancer.

Several factors could be contributing to the observed biases:

- The model may be overfitting to the majority class, class label 4 (Melanocytic nevi), due to the high number of samples compared to other classes. This can cause the model to become overly specialized in detecting this specific class, resulting in poor performance on minority classes.
- The minority classes, especially class label 6 (Melanoma), may have insufficient training samples to allow the model to learn the distinguishing features necessary for accurate classification. This can lead to poor generalization when encountering new samples from these classes.

4.2 Gender Bias Investigation

To analyze the presence of gender bias in the model's performance, we first evaluate the accuracy of the model on separate male and female test sets. This evaluation helps us understand whether the model has difficulty learning features specific to one gender over the other, despite partial gender imbalances in the training dataset.

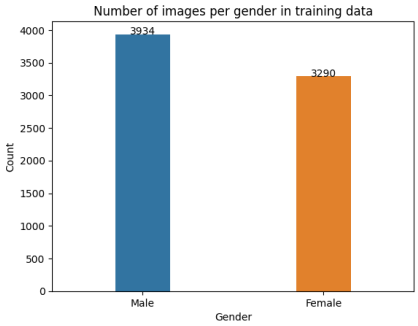

Figure 4: Gender samples in Train dataset

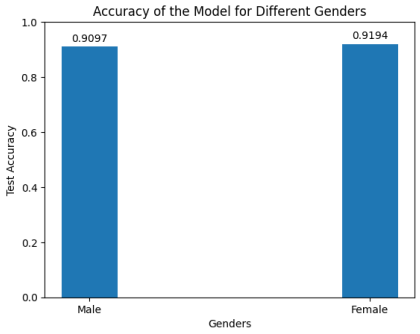

Figure 5: Model performance on Gender (Test data)

From the test accuracy results, we can observe that there is no significant gender bias in the model's performance. Despite the partial imbalance in the training dataset, the model performs similarly on both male and female test sets. This suggests that the model has effectively learned features from both genders, without a strong preference for one over the other.

Furthermore, even after creating an imbalance between male and female samples, there is no substantial difference in the model's accuracies on the male and female test sets. This indicates that the model is resilient to gender bias and performs well across both genders.

It is worth noting that, despite having more male samples in the training dataset, the model does not have difficulty learning female features. This observation counters the potential concern that the model might struggle to learn male-specific features, such as body hair or beards. Instead, the model's performance remains consistent across both genders.

4.3 Age Bias Investigation

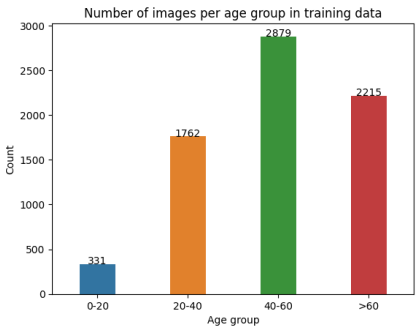

Figure 6: Age subgroup samples in Train dataset

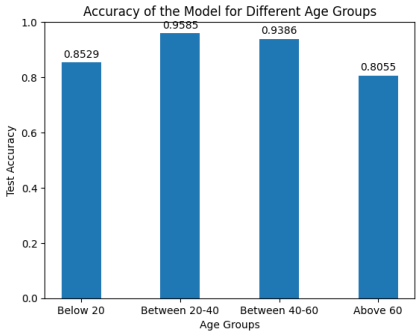

Figure 7: Model performance on Age subgroup (Test data)s

We have created four age subgroups and added a new column 'age_subgroup' in the metadata, using the information from the 'age' column.

To analyze the presence of age bias in the model's performance, we first evaluate the accuracy of the model on separate age subgroups. This evaluation helps us understand whether the model has difficulty learning features specific to certain age groups, despite potential imbalances in the training dataset.

From the test accuracy results, we can observe that there is a bias towards younger and mid-age groups. The model's performance is relatively higher for patients in the 20-40 and 40-60 age groups. In contrast, the model is underperforming for the old-age group (above 60 years). This may be attributed to the fact that older individuals have more complex skin structures, such as aging skin, which might be more challenging for the model to learn and classify.

There is no direct relationship between the number of samples per age group and test accuracy. This suggests that other factors, such as the complexity of skin features and the quality of the images, might be influencing the model's performance across different age groups.

4.4 Localization Bias Investigation

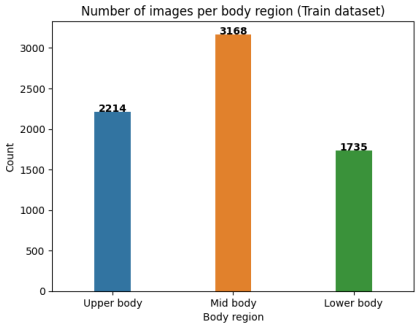

Figure 8: Body region subgroup samples in Train dataset

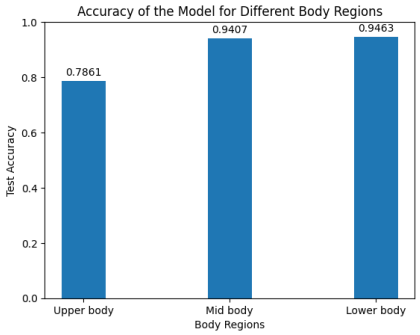

Figure 9: Model performance on Body region subgroups (Test data)

We have created three subgroups and added a new column 'body_region' in the metadata, using the information from the 'localization' column.

To analyze the presence of localization bias in the model's performance, we first evaluate the accuracy of the model on separate body region subgroups. This evaluation helps us understand whether the model has difficulty learning features specific to certain body regions, despite potential imbalances in the training dataset.

From the test accuracy results, we can observe that there is a bias towards specific localization, especially the mid body and lower body regions. The model's performance is relatively lower for patients with skin lesions on the upper body regions. This underperformance may be attributed to the presence of hair on the scalp, neck, and facial hair, which might be more challenging for the model to learn and classify.

Interestingly, there is no direct relationship between the number of samples per body region and test accuracy. This suggests that other factors, such as the complexity of skin features, quality of the images, and the presence of occlusions or artifacts, might be influencing the model's performance across different body regions.

5 Bias Mitigation

5.1 Class Bias Mitigation

5.1.1 With pre-processing technique: Oversampling minority classes

We analyze the model's performance by comparing the oversampling (pre-processing) technique, the class balance loss (in-processing) technique, and a combination of both techniques. We first apply the oversampling technique to minority classes as a pre-processing step before training the model.

While the test accuracies of a few minority classes have improved, there is still a bias towards the majority classes. This indicates that the oversampling technique alone may not be sufficient to mitigate class bias effectively.

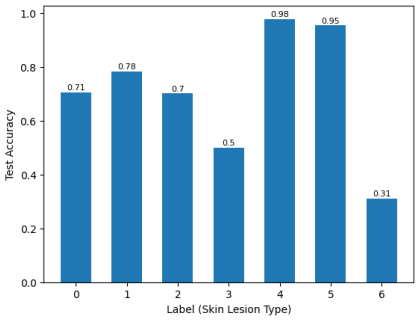

Figure 10: Without oversampling

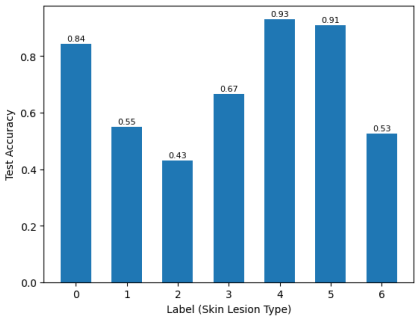

Figure 11: With Oversampling.

5.1.2 With In-processing technique: Class balance loss

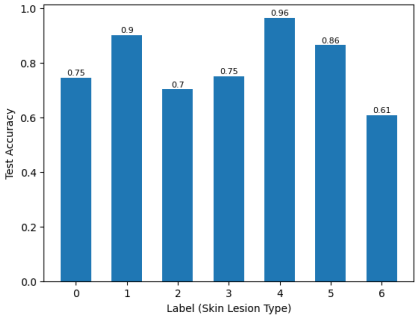

Figure 12: With Class balance loss

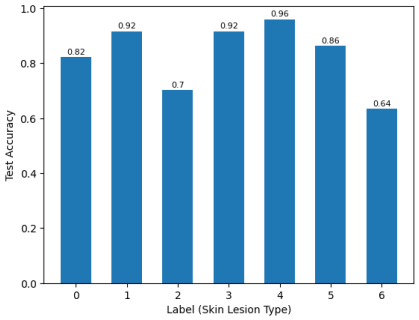

Figure 13: With Oversampling and Class balance loss

Next, we apply the class balance loss [3] technique as an in-processing step during model training. And Finally, we combine both the oversampling and class balance loss techniques and retrain the model.

We can observe that by combining both techniques, the test accuracies of a few minority classes have increased, and class bias is reduced to some extent. This indicates that using a combination of pre-processing and in-processing techniques can provide a more effective solution for mitigating class bias in the model.

The class bias mitigation techniques applied in this study, particularly the combination of oversampling and class balance loss, have shown promising results in reducing class bias and improving the model's performance across all classes. However, some classes still exhibit lower accuracies compared to others, suggesting that further research and fine-tuning of mitigation techniques may be required.

5.2 Age Bias Mitigation

5.2.1 With pre-processing technique: Oversampling minority age groups

In an attempt to mitigate age bias, we employed the oversampling technique on minority age groups to balance the representation of different age groups in the dataset. Upon analyzing the results, we observe that the oversampling technique has led to an increase in test accuracy for the younger age group (0-20), showing a positive impact on the model's performance for this group. However, the improvements for other age groups are not as significant. In particular, the test accuracy for the older age group (60+) has not improved considerably.

There could be multiple reasons for the limited success of oversampling in mitigating age bias for older age groups:

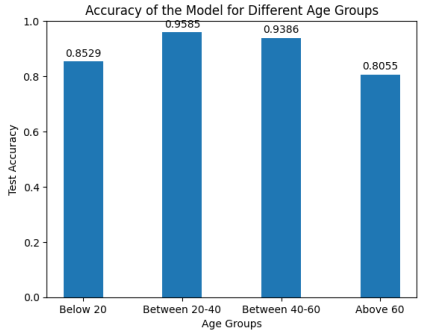

Figure 14: Without oversampling

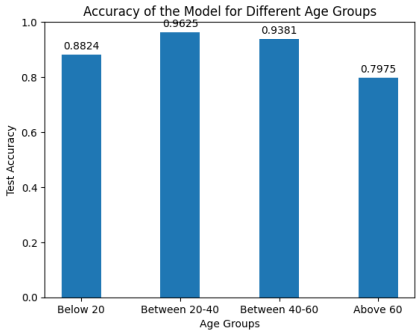

Figure 15: With Oversampling minority age groups

- Oversampling might not be sufficient on its own to address the age bias, especially if the model struggles to learn complex features and patterns specific to older age groups, such as aging skin and other age-related skin changes.
- The oversampling technique can sometimes lead to overfitting, where the model becomes too focused on the oversampled instances, leading to limited generalization capabilities.

.

5.2.2 With In-processing technique: Adversarial training

In our approach, we aim to mitigate age bias in skin lesion classification by implementing adversarial training. We start by defining a custom GradientReversalLayer that reverses the gradient during backpropagation. We then create an AgeClassifier class, which is a neural network designed to predict age groups.

During training, we use a combination of the main classification loss (cross-entropy or class-balance loss) and an adversarial loss (cross-entropy for age classification) to update the model parameters. We control the trade-off between the classification loss and the adversarial loss with a hyperparameter α (0.1 in our case).

The goal of this adversarial training is to force the model to learn features that are useful for the main task (skin lesion classification) while being less informative about the protected attribute (age). By doing so, we expect to mitigate age bias in the classification results

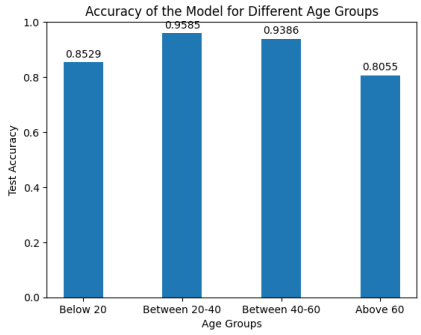

Figure 16: with Age bias

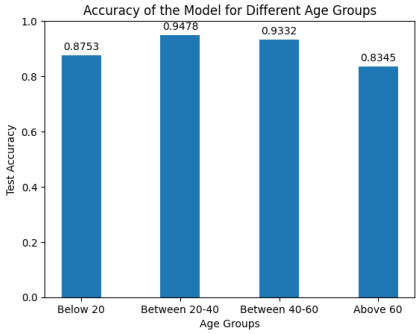

Figure 17: with oversampling + adversarial training

Analysis: Upon analyzing the results, we observe that adversarial training led to an improvement in test accuracy for the older age group (60+), which was the primary target for mitigation. However, there was a slight decrease in performance for the other age groups.

There could be several reasons for the mixed success of adversarial training in mitigating age bias:

Adversarial training aims to reduce the model's dependence on age-related features while maintaining performance on the main task. This balance may not be perfect, leading to a trade-off between fairness and overall model performance.

The *hyperparameter* α, which controls the trade-off between classification loss and adversarial loss, may need further tuning to achieve better results. Increasing alpha can enhance fairness and privacy properties but may also decrease the model's accuracy on the main task.

In conclusion, while adversarial training has shown effectiveness in improving the model's performance for the older age group (60+), it has led to a slight decrease in performance for the other age groups. Further experimentation with different mitigation techniques, model architectures, and hyperparameter tuning might be necessary to effectively address age bias across all age groups. Tuning the hyperparameter alpha can help strike a balance between fairness and overall model performance, leading to more satisfactory results.

5.3 Localization Bias Mitigation

5.3.1 With pre-processing technique: Oversampling minority body region groups

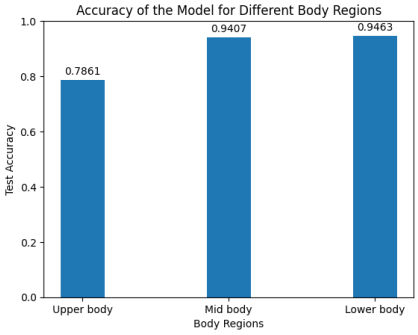

Figure 18: Without oversampling

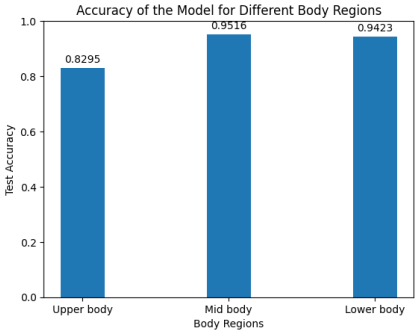

Figure 19: With Oversampling minority body region groups

To mitigate the localization bias, we employed the oversampling technique on minority body regions to balance the representation of different body regions in the dataset.

Upon analyzing the results, we observe that the oversampling technique has led to an increase in test accuracy for the minority class (Upper body), indicating a positive impact on the model's performance for this body region.

There could be several reasons for the limited success of oversampling in mitigating localization bias:

- Oversampling might not be sufficient on its own to address the localization bias, particularly if the model struggles to learn complex features and patterns specific to certain body regions, such as the presence of hair or other unique characteristics.
- The oversampling technique can sometimes lead to overfitting, where the model becomes too focused on the oversampled instances, resulting in limited generalization capabilities.
- The underlying model architecture might not be well-suited to capturing the localization-specific features and patterns in the data, necessitating further adjustments or a different architecture altogether.

5.3.2 With In-processing technique: Adversarial training

We observed an improvement in test accuracy for the Upper body region, indicating a reduction in localization bias. However, there was a slight decrease in performance for the Mid and Lower body regions.

Analysis: The adversarial training effectively mitigated the localization bias, particularly for the Upper body region. The mitigation technique forces the model to learn more generalized features,

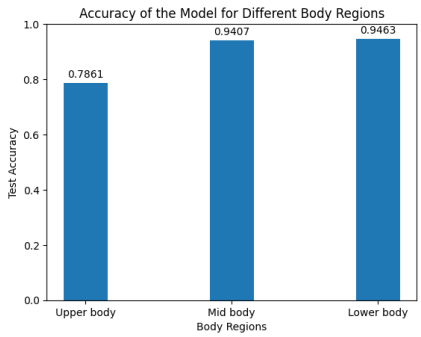

Figure 20: With localization bias

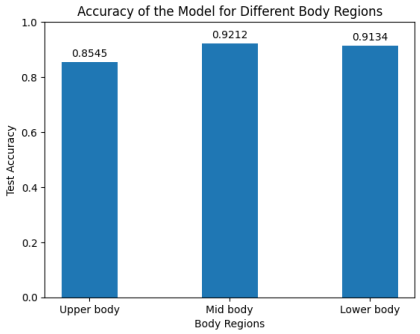

Figure 21: with oversampling + adversarial training

which could have contributed to the improvement in performance for the Upper body region. However, this generalization may also result in a slight decrease in performance for other body regions.

One key observation is that the hyperparameter alpha, which controls the trade-off between classification loss and adversarial loss, plays a crucial role in the effectiveness of the mitigation technique. By tuning alpha, we can find an optimal balance between fairness and accuracy. A higher value of alpha puts more emphasis on minimizing adversarial loss, potentially improving fairness but decreasing model accuracy for the main task. Conversely, a lower value of alpha focuses more on minimizing classification loss, improving model accuracy for the main task but possibly increasing localization bias.

6 Conclusions

In this study, we investigated the impact of various biases, including class bias, age bias, gender bias, and localization bias, on skin lesion classification. We explored several mitigation techniques to address these biases, such as data augmentation, oversampling, adversarial training, and class balance loss. Our findings highlighted the importance of addressing these biases to achieve more accurate and fair model performance across different demographic groups and body regions.

Data augmentation and oversampling proved beneficial in mitigating class bias, age bias, and gender bias, leading to improvements in classification accuracy across all demographic groups. The class balance loss technique helped to address class imbalance issues and contributed to the overall improvement of the model's performance. Adversarial training, an in-processing technique, led to a more balanced performance across different age groups and body regions, with notable improvement in test accuracy for the older age groups and Upper body region group. Nevertheless, this technique resulted in a slight decrease in performance for the other age and body region groups. The hyperparameter alpha played a crucial role in balancing fairness and accuracy in adversarial training, with further tuning offering potential improvements in model performance.

7 Future Work

Future work in this area should focus on refining the mitigation techniques to achieve a better balance between fairness and accuracy across different biases. This could involve developing new data augmentation strategies, exploring alternative oversampling methods, fine-tuning adversarial training with additional hyperparameters, or experimenting with novel class balance loss functions. Moreover, investigating the interplay between different biases and their combined impact on model performance could provide valuable insights into the development of more robust and fair skin lesion classification systems. Finally, validating the effectiveness of the proposed mitigation techniques on larger and more diverse datasets will be essential to ensure the generalizability of the results and the applicability of the methods in real-world clinical settings.

References

[1] A Survey of Fairness in Medical Image Analysis: Concepts, Algorithms, Evaluations, and Challenges

[2] Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey

[3] Class-Balanced Loss Based on Effective Number of Samples