Project Report

Fairness in Skin disease prediction

Giorgia Pavani, Livia Del Gaudio

May 30, 2024

Abstract

This project makes use of a medical dataset comprising images depicting skin disease in pediatric patients. These images have been augmented via data augmentation techniques and diffusion models to compensate for limited data availability. The main objective is to examine bias within the dataset, particularly considering its inherent bias towards Caucasian skin tones. Additionally, we plan to conduct experiments by injecting bias, for example by generating images of specific skin tones, to observe model's behavior under varied conditions. We anticipate implications related to fairness when testing the model on other skin color, and to this extent we aim to propose possible strategies to mitigate the phenomenon.

1 Description

The aim of this project is to study and assess fairness in the prediction of skin diseases, based on images collected by the Pediatrics Unit of Ospedale Sant'Orsola. Due to the scarcity of medical data, data augmentation techniques were employed to generate additional realistic images.

1.1 Goals

Our work will address the following tasks:

1. perform skin disease predictions *as is* on images generated by the diffusion model;
2. bias injection by prompting the diffusion models on a specific skin tone and carrying out experiments to observe how the model behaves;
3. study fairness implications and try to suggest mitigation methods.

1.2 Deliverables

To ensure the reproducibility of our study, we will provide a detailed final report along with a GitHub repository containing the implementation of our project.

2 Dataset

We use *skin-disease-dataset* containing images grouped into 9 different diseases: *esantema-iarogeno-farmaco-indotta*, *esantema-maculo-papuloso*, *esantema-morbiliforme*, *esantema-polimorfo-like*, *esantema-virale*, *orticaria*, *pediculosi*, *scabbia* and *varicella*. This dataset is employed for multiclass classification tasks, with further details provided in the following sections.

2.1 Dataset inspection

By inspecting the dataset, we immediately observed that it is highly **unbalanced** both in **disease type and skin tone**. Indeed, as the majority of examples are pictures of caucasian children, the distribution of images is skewed towards white skin tones, which is compliant with the fact that the images were all collected from a single hospital located in Italy. As a matter of fact, we noticed that

there is a small amount of examples where the skin tone is brown or dark. Moreover, we observed that there is a skewed distribution of data also among classes (i.e. skin diseases), as some diseases such as *esantema morbilliforme* have very few examples compared to other labels. To further assess the class distribution among the whole dataset, we plotted it in Figure 1.

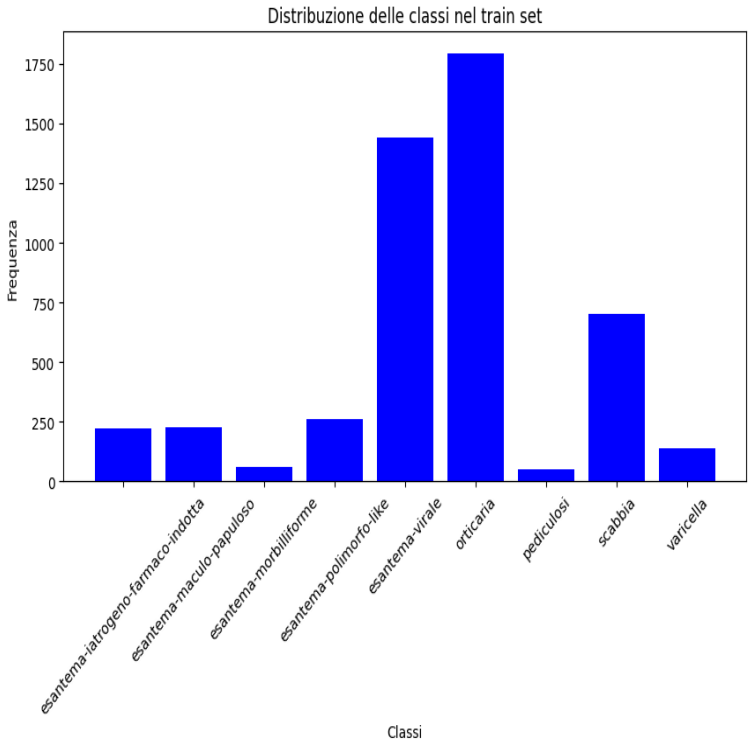


Figure 1: Class distribution in original Dataset

In addition to that, we have noticed that there are some classes where the imbalanced distribution of skin tones is more evident, as it happens in *pediculosi* or *esantema polimorfo-like* where there are no examples of images capturing darker skin, but even the opposite phenomenon can be observed in *scabbia*, where the majority of images show a darker skin tone.
Both the unbalanced disease class distribution and the dataset polarized toward caucasian skin tone may result in unfair predictions: diseases are more likely to be misclassified and harder to detect in dark skin tones. The issue of unbalanced skin tone resulting in poor performances for non-caucasian skins can be considered a fairness and non discrimination problem. To deal with it and to try to mitigate such issue, we decided to follow the approach described in [1] and to try to enhance those results.
In order to develop our project we created three datasets:

- **Mixed Dataset**: we incorporated in a single dataset both the original and darkened datasets following the pipeline presented in [2].
- **Balanced Mixed Dataset**: augmented the original Mixed Dataset using sliding window method definded in [3].
- **Extended Mixed Dataset**: we extended the mixed dataset by performing new skin color generation, in order to feed the model with images showing intermediate skin colors.

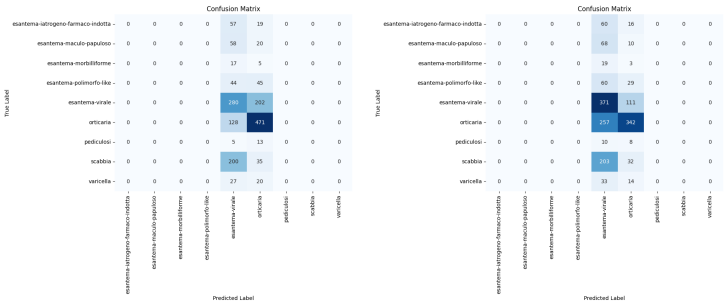
We then trained and tested the model with both dataset and with different approaches, and evaluated them according to *accuracy*, *F1-score* and *recall* metrics.

3 Implementation and results

In the following sections we are going to present and discuss our previously mentioned experiments. The backbone of the model used is *ResNet50* and the whole training pipeline and hyperparameters are the same for all experiments. We tried different approaches to handle the inherent biases discussed above, namely unbalanced class and skin tone distribution.

3.1 Baseline

The first experiment involved performing skin disease prediction *as is* by training the model on the original dataset. Then, we tested it on both light and dark images to obtain a benchmark for the comparison of further results. Results reflect the inherent biases already identified in the previous section: as we can see from the confusion matrices in Figure 2 that model predicts only the most frequent classes. Table 1 highlights model's poor performance; in this case, accuracy is not particularly relevant due to the high class imbalance.



(a) original (b) dark

Figure 2: Confusion matrices - Baseline

Dataset	*Accuracy*	*F1-score*	*Recall*
original	0.4563	0.1212	0.1519
dark	0.4332	0.1180	0.1490

Table 1: Result on test sets - Baseline

3.2 Mixed Dataset

In order to tackle the bias arising from the lack of data representing diseases in darker skin tone, we generated images mimicking the presence of dark skins by applying filters to the original images. Then, we built a new dataset by mixing the original images with the dark ones so that the model is trained on a higher variety of skin tones. However, given the highly unbalanced class distribution, reported in Figure 3, we still did not expect to obtain high performances from our model. As a matter of fact, training the model on extremely skewed data results in a model which tends to predict always the most frequent classes with the aim of minimizing the loss. This behavior was confirmed by the results we obtained for this experiment, which are shown in Figure 4.

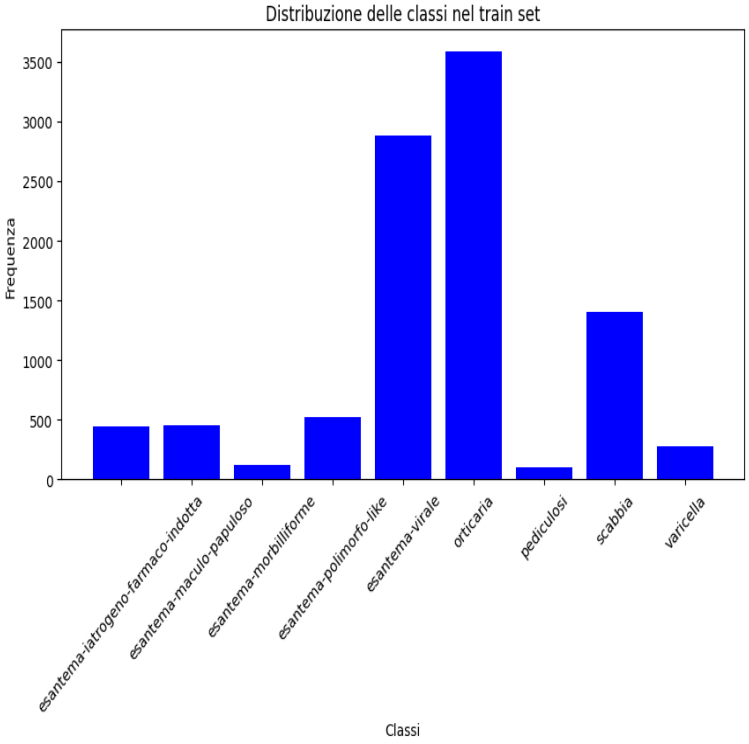
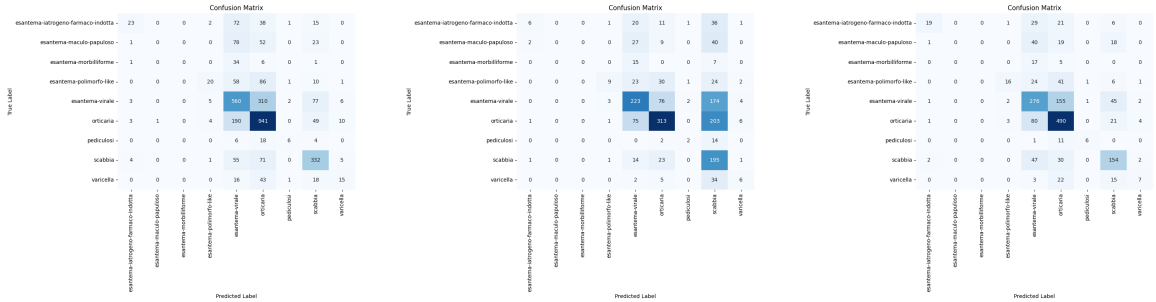


Figure 3: Class distribution in Mixed Dataset



(a) mixed (b) dark (c) original

Figure 4: Confusion matrices - Mixed Dataset

Dataset	***Accuracy***	***F1-score***	***Recall***
mixed	0.1906	0.1556	0.2316
original	0.2333	0.1803	0.2623
dark	0.1713	0.1553	0.2485

Table 2: Result on test sets - Mixed Dataset

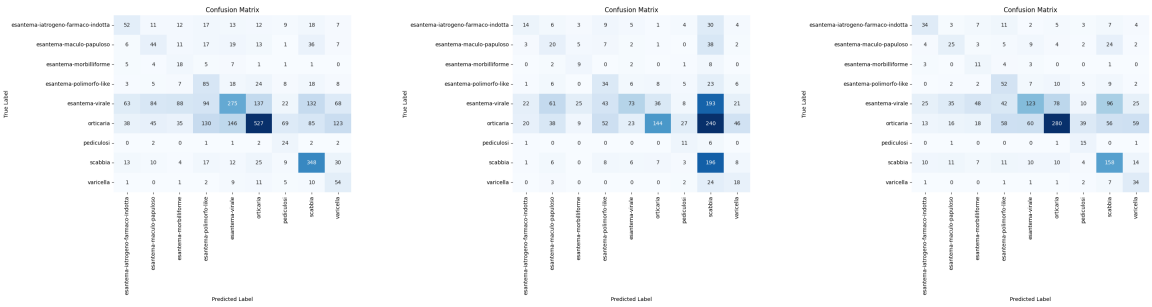
As expected, the model learns to correctly predict only the over represented classes, whilst struggles in predicting the others. By comparing the performances on original and dark test sets, we can observe slightly higher performances in the first one. Furthermore, we can observe from the confusion matrix of dark test set that, as expected, the model tends to predict dark images as belonging to *scabbia* disease, as a consequence of *scabbia* class containing mostly dark skin images.
Therefore, we tried to also mitigate this unwanted biased phenomenon by adopting the following two distinct approaches:

- Applying **class weighting** to the loss function;
- **Undersampling** the classes with significantly more data, namely *esantema-virale*, *orticaria* and *scabbia*.

The following subsections provide detailed explanations of these two strategies and of the results achieved.

3.2.1 Class weights

The first approach to mitigate the negative effects of unbalanced distribution of classes, involved adding weighting to the loss function. The weight was different for each class, depending on the total number of samples in the training set. Results are shown in Figure 5.



(a) mixed (b) dark (c) original

Figure 5: Confusion matrices - Mixed Dataset with class weights

Dataset	*Accuracy*	*F1-score*	*Recall*
mixed	0.4352	0.3460	0.4775
original	0.4447	0.3767	0.5338
dark	0.3153	0.2698	0.3835

Table 3: Result on test sets - Mixed Dataset with class weights

This approach proved to be beneficial: with respect to the previous results also the less represented classes are now likely to be predicted. Analysing the metrics reported in Table 3, we observe a great improvement in both original and dark test set. However, despite the enhanced results, there remains a substantial performance gap between the dark and original test sets.
As we were not completely satisfied, we also tried a second approach, which involved undersampling the classes with significantly more elements to balance the dataset.

3.2.2 Undersampling majority classes

Undersampling handles skewed distribution of classes by reducing the number of instances in the majority classes to create a more balanced dataset. We chose as undersampling strategy for this experiment *Random Undersampling*, which involves randomly selecting a subset of instances from the majority classes to match the number of instances in the minority classes. In our case-study, we identified *esantema-virale*, *orticaria* and *scabbia* as majority classes and we applied to them the undersampling pipeline. Figure 6 shows the distribution of data among classes after having applied this strategy.

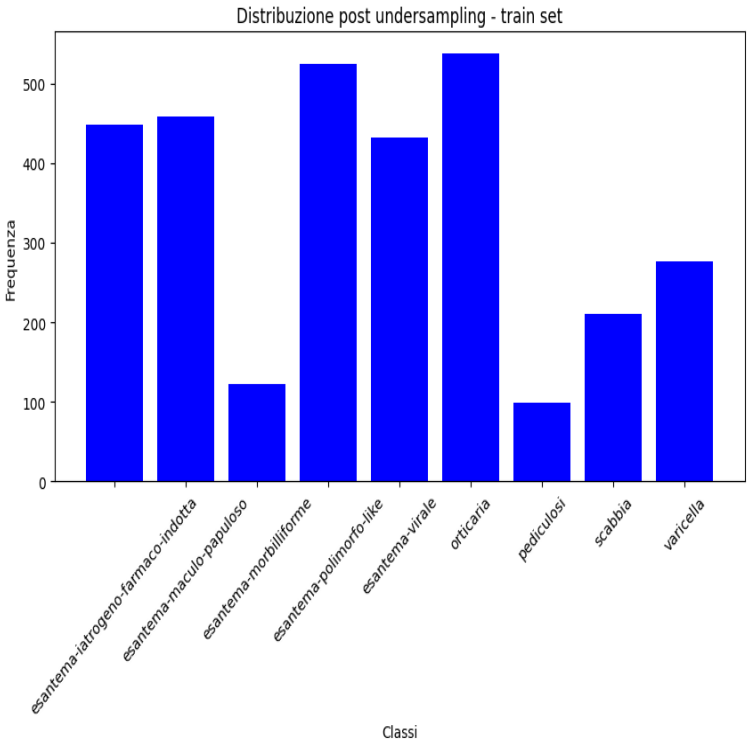
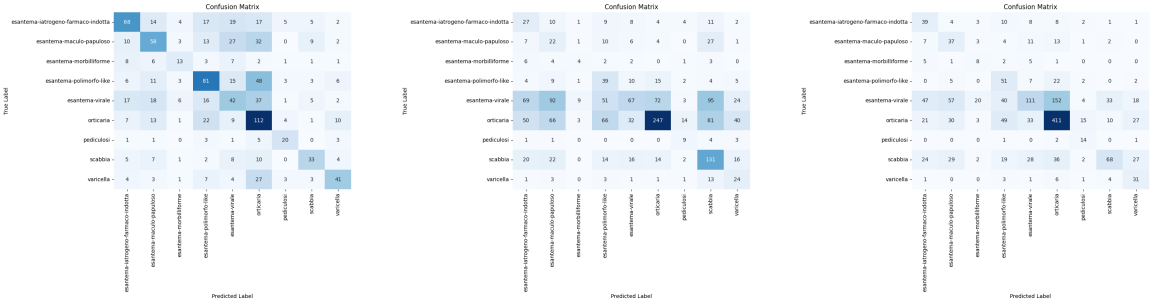


Figure 6: Class distribution in Mixed Dataset with undersampling

As it can be observed from the previous histogram, classes are now less skewed and have a more balanced distribution: the majority of them have a total number of samples between 400 and 500. However, there are classes (*esantema-morbilliforme*, *pediculosi* and *varicella*) that are still less represented in the dataset.
After mitigation of class imbalance with undersampling, we expected our model to improve its ability in classifying diseases more uniformly, and therefore to have a more unbiased distribution of predictions. As we can see from the confusion matrix in Figure 7, our hypothesis was indeed confirmed. However, the model still performs quite poorly in terms of results, especially when tested on the dark set, as we can notice from Table 4 that all the considered metrics are below 50%.



(a) mixed (b) dark (c) original

Figure 7: Confusion matrices - Mixed Dataset with undersampling

Dataset	*Accuracy*	*F1-score*	*Recall*
mixed	0.4487	0.4532	0.4461
original	0.4678	0.3942	0.5075
dark	0.3463	0.2906	0.3752

Table 4: Result on test sets - Mixed Dataset with undersampling

Due to the poor performance observed in the underrepresented classes, we attempted to balance the dataset by further reducing the number of instances in the highly represented classes. Our goal was to achieve approximately 100 to 300 samples per class. However, this approach led to poor performance because the significantly reduced dataset hindered the model's ability to generalize and effectively learn the intrinsic characteristics of the diseases.

3.3 Balanced Mixed Dataset

Applying techniques to handle unbalanced dataset distribution have a positive impact in the performances for both original and dark images. However, we wanted to furthermore improve performances and we augmented our Mixed Dataset by using sliding window approach cited above. The resulting dataset class distribution is depicted in Figure 8.

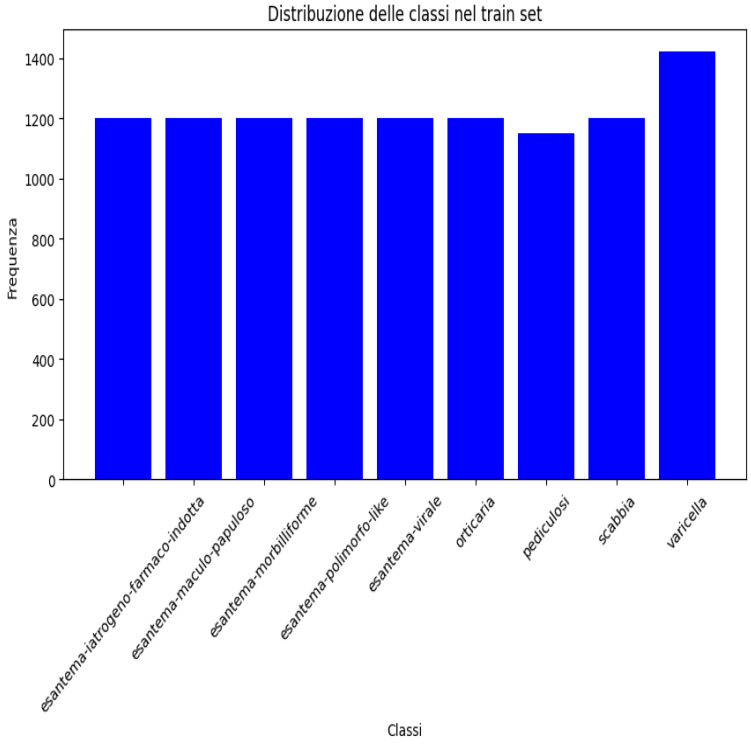
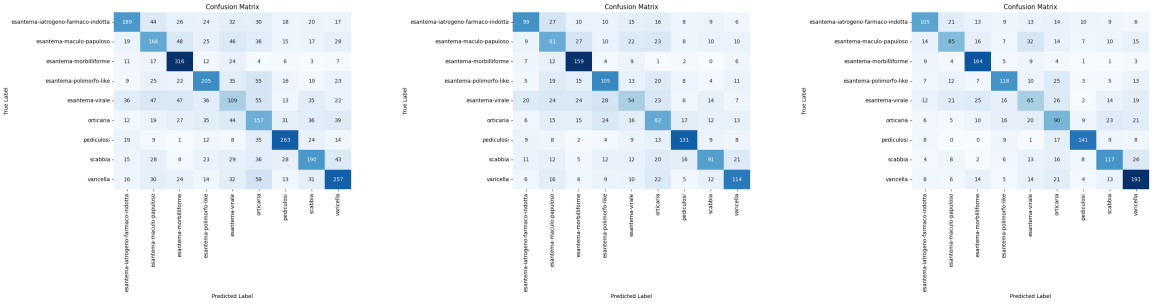


Figure 8: Class distribution in Balanced Mixed Dataset

Starting from this balanced dataset, we performed three separated experiments where we took different percentages of original and dark images and we trained the model. Then we tested the model on mixed, dark and original test sets in order to evaluate the effects on the performances of having dataset composed by different percentages of skin tone data.

3.3.1 50% original - 50% dark

We trained the model on an equal proportion of original and dark dataset, then tested it on the mixed dataset, on the original dataset, on the dark dataset. In the following Figure 9 we report the confusion matrix of all tests, and in Table 5 the results with the metrics.



(a) mixed (b) dark (c) original

Figure 9: Confusion matrices - Balanced Mixed Dataset (50% - 50%)

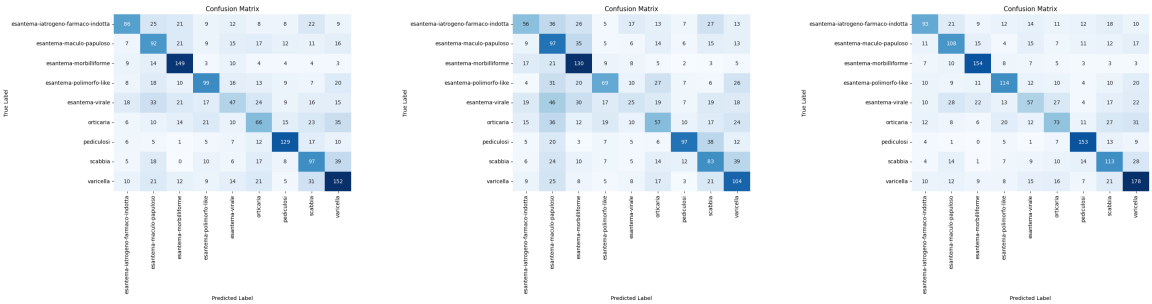
Dataset	*Accuracy*	*F1-score*	*Recall*
mixed	0.5059	0.5023	0.5059
original	0.5757	0.5689	0.5714
dark	0.5109	0.5070	0.5115

Table 5: Result on test sets - Balanced Mixed Dataset (50% - 50%)

Results shown in both Figure 9 and Table 5 are the best obtained so far. Indeed, it can be observed that the confusion matrices, for all the test sets are a way more balanced than before and the model's predictions are fairly distributed among classes. In terms of metrics, we still see a difference in the performances between dark and original dataset, since the first one yields worse performances. However, comparing this discrepancy with the previous results, we can still notice that the bias towards light skin is slightly reduced.
As a consequence, data augmentation with the aim of composing a uniformly distributed dataset seems to be an essential and important step in the mitigation of biases toward darker skin tones.

3.3.2 80% original - 20% dark

In this experiment we trained the model using different proportions of images with lighter (original ones) and darker skin tones. Specifically, the training set comprised 80% images with lighter skin tones and 20% with darker skin tones. Due to this dataset composition, we did not expect that the model to have outstanding performances on the dark test set since few examples in the training set are dark. As usual, the model was trained on this dataset and evaluated on mixed, original, and dark skin tone datasets. Figure 10 and Table 6 depict the results of those tests.



(a) mixed (b) dark (c) original

Figure 10: Confusion matrices - Balanced Mixed Dataset (80%-20%)

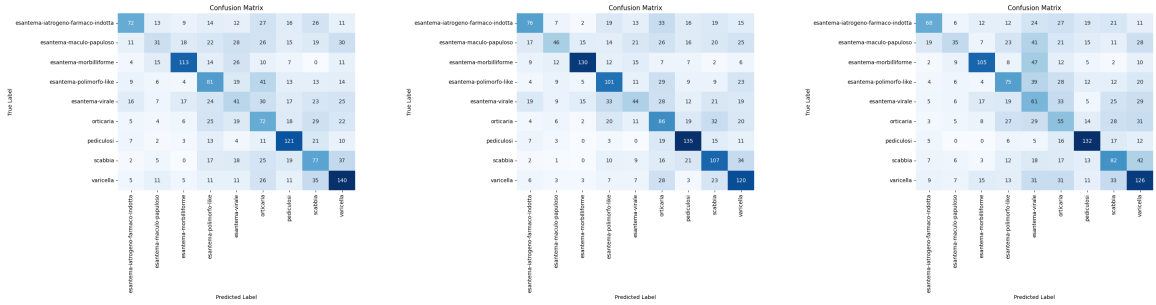
Dataset	*Accuracy*	*F1-score*	*Recall*
mixed	0.4912	0.4840	0.4894
original	0.5581	0.5489	0.5553
dark	0.4004	0.3920	0.4008

Table 6: Result on test sets - Balanced Mixed Dataset (80%-20%)

Results are slightly worse comparing with the previous experiment, but predictions are still equally distributed among all the classes. As expected, performances on dark dataset are lower and this is justified by the dataset composition: only 20% of the data belong to dark dataset and therefore the model will struggle in learning correlations with this skin tone.

3.3.3 20% original - 80% dark

We trained the model using different proportions of images with lighter (original ones) and darker skin tones. Specifically, the training set consisted of 20% images with lighter skin tones and 80% with darker skin tones. As a consequence of the broad dominance of dark examples over light ones in the mixed training set, we expect improved performances when the model is tested on dark images, while worse results when tested on light images. As usual, the model was trained on mixed dataset and evaluated on mixed, original, and dark skin tone datasets. In Figure 11 and Table 7 are shown the results of this experiment.



(a) mixed (b) dark (c) original

Figure 11: Confusion matrices - Balanced Mixed Dataset (20%-80%)

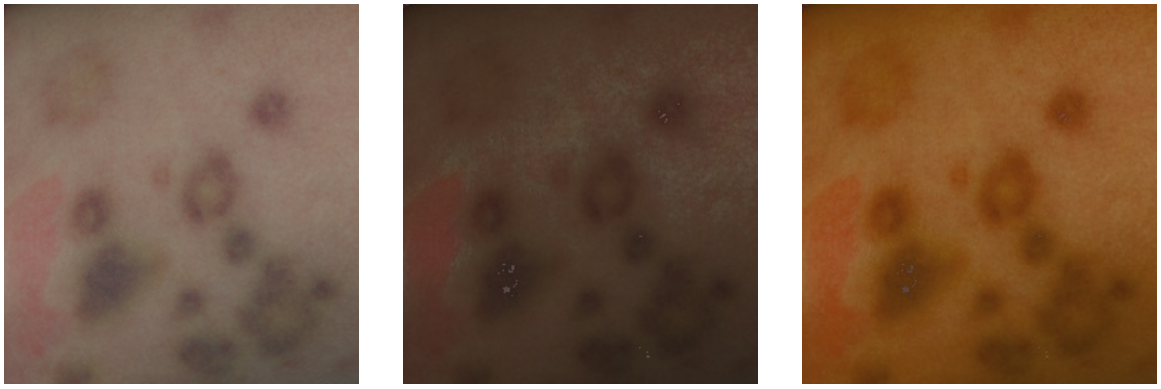
Dataset	*Accuracy*	*F1-score*	*Recall*
mixed	0.4050	0.3971	0.4016
original	0.3954	0.3951	0.3939
dark	0.4713	0.4639	0.4722

Table 7: Result on test sets - Balanced Mixed Dataset (20%-80%)

Results show a remarkable degradation of performances when the model is tested on the original dataset, as well as an enhancement on the dark dataset. This behavior is indeed compliant with our expectations, whereas we can notice from Table 7 that the model shows performances lowered in a range between 8-10% on the mixed test set. We hypothesize that the drop in performance is due to the nature of the data used in this test. Indeed, 80% of the data (representing darker skin tones) is artificially generated. As a result, these images may not accurately capture the real-image variability, leading to poorer model performance.
As a consequence, we decided to generate new images, always following the pipeline in [2], of an intermediate skin tone and perform experiments in order to assess whether increasing variability to the artificially generated dataset will lead to an enhancement of the performances. In the next section, we will provide a detailed description of this new experiment.

3.4 Mixed dataset extended through new skin color generation

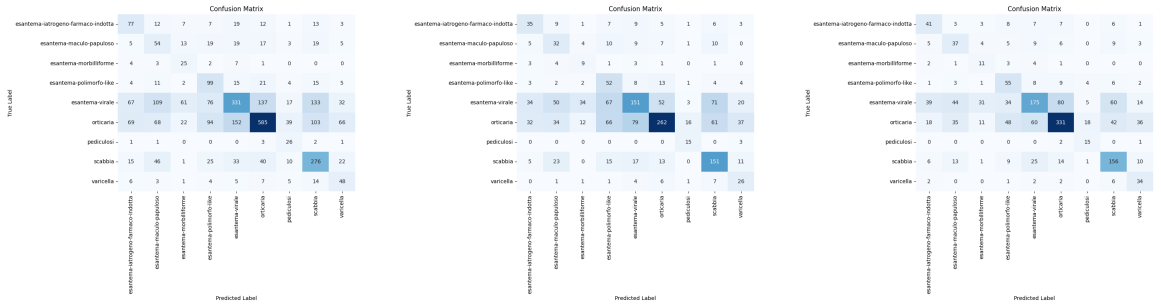
The dataset generated for darker skin tones sometimes results in unrealistic representations, occasionally producing grayish skin tones. To address this issue, we selected a more realistic intermediate skin tone that better represents a range of darker skin tones. An example of this adjustment is shown in Figure 12, where the original, darker, and intermediate skin tones are compared.



(a) original (b) dark (c) intermediate

Figure 12: Skin tones

As a further attempt to tackle the bias related to unfair distribution of skin tones, we proceeded to extend the mixed dataset presented in subsection 3.2 by incorporating images generated to represent intermediate skin tones, in order to add variability to dark dataset. We then trained the model adding class weights (to face unbalanced distribution of classes problem) and evaluated its performances, shown in Figure 13 and Table 8, to assess bias mitigation.



(a) mixed (b) dark (c) original

Figure 13: Confusion matrices - Extended Mixed Dataset with class weights

Dataset	*Accuracy*	*F1-score*	*Recall*
mixed	0.4639	0.3958	0.5245
original	0.5194	0.4567	0.5853
dark	0.4453	0.3959	0.5160

Table 8: Result on test sets - Extended Mixed Dataset with class weights

By observing these results, we notice that injecting more variance in the skin color distribution results in a large improvement of performances with respect to the experiment with Mixed Dataset and class weights, presented in Table 3. In particular, we observe a general improvement in performance, with a notable increase in all metrics studied, especially for the dark dataset. This suggests that this approach can be highly beneficial to deal with skin tone biases.

4 Results

Our experiments were mainly focused on the mitigation of biases due to the imbalanced distribution of classes and to the scarce variability of data concerning skin tones. In this section we briefly report the results we obtained by implementing the strategies presented and discussed above.
Our Baseline experiment addressed skin disease prediction *as is*, highlighting the model's tendency to predict only the most numerous classes as a consequence of the skewed distribution of classes. Introducing a Mixed Dataset with images artificially adjusted to represent darker skin tones led to a slight improvement. However, the model still favored the most represented classes, resulting in sub-optimal performance on both original and dark test sets and suggesting the persistence of the bias. Therefore, we further addressed this issue by applying Class weights to the loss function, which led to improved model's ability to predict less represented classes. Although there was a noticeable improvement in terms of metrics, a significant performance gap between dark and original test sets remained. We also tried to tackle class imbalance by Undersampling majority classes. This approach resulted in a more balanced dataset and improved the uniformity of predictions across classes. Despite this, the overall performance, especially on the dark test set, remained below 50%. Another strategy to handle the inherent biases consisted in augmenting the mixed dataset with a sliding window approach to create a more uniformly distributed Balanced Mixed Dataset. Training the model on equal proportions of original and dark images significantly balanced the predictions and reduced the bias towards light skin tones, as reflected by improved performance metrics. This technique was the most effective, as it yielded the best results.
We then carried out other experiments with the balanced dataset, by taking different proportions of dark and original data employed for training. Results showed that higher proportions of dark images led to better performance on dark test sets but worse performance on original test sets, and vice-versa. In general, we observed that the model tends to perform worse on synthetic data with respect to realistic data.
Lastly, building a Mixed dataset extended through new skin color generation introduced a more realistic intermediate skin tone to the mixed dataset. This led to a substantial improvement in performance across all metrics and test sets, particularly for the dark dataset, compared to the corresponding experiment on the mixed dataset, discussed in subsubsection 3.2.1.
To sum up, while we did not achieve exceptionally high results, our studies led to notable improvements. Our highest F1-score reached 0.57 on the original dataset and 0.51 on the dark dataset. This is a significant increase from our baseline scores of 0.12 and 0.11, respectively.

5 Conclusion

Our work achieves the goals listed in subsection 1.1 by exploring different strategies related to fairness in skin disease prediction. In particular, this study highlighted the trade-off and the impact of dataset composition on model performance, as well as a viable strategy for mitigating inherent biases related to fairness consisting in increasing variability in the training data. Even though we notably improved the performances of the model with respect to the baseline and partially addressed the bias-related issues, we didn't manage to achieve completely satisfying results.
All those considerations are useful for identifying possible directions for future improvements. First of all, it would be interesting to integrate the balanced dataset with intermediate skin color images, to assess the performances on the balanced dataset with greater skin tone variability. Then, it can be useful to adopt novel techniques for data augmentations, for example GANs or VAEs, to achieve more realistic images and to study the impact of realism on the performances of the model. Finally, extending the dataset using a larger number of samples, surely can have a positive impact on the results, but memory and hardware limitations must be taken into consideration as well.

References

[1] Skin disease classifier *https://github.com/khanenab/skin_disease_classifier*

[2] Skin tone changing *https://github.com/khanenab/skin-tone-changing*

[3] Data augmentation *https://github.com/aequitas-aod/synthesizer_image/blob/main/generate_images.ipynb*