Fairness Analysis

1. Introduction

In this document, we investigate the performance and fairness of our melanoma classification model across different skin tones. Ensuring both high performance and equitable treatment in medical imaging models is essential, particularly given the historical and ongoing disparities in healthcare affecting marginalized communities. For a more detailed analysis, refer to the *'fairness_exlporation.ipynb'* notebook on this Google drive link.

While the incidence of melanoma is significantly lower among the Black population (1.0 per 100,000) compared to white individuals (23.5 per 100,000), the **10-year melanoma-specific survival rate** is also notably lower: **73% for Black patients**, compared to **88% for white patients** and **85% for other racial groups**. This disparity underscores the importance of developing models that accurately identify melanoma across all skin tones, particularly in underrepresented and underserved populations Groh et al., 2022.

In the context of melanoma detection, this challenge is compounded by several factors:

- Individuals with darker skin tones are statistically less likely to develop melanoma, but are also less likely to receive timely or adequate care, often resulting in later-stage diagnoses and worse clinical outcomes. Consequently, publicly available dermoscopic datasets contain disproportionately few images of darker skin tones
- Lesions on darker skin often exhibit lower contrast with the surrounding skin, making it more difficult to visually distinguish diagnostic features such as border irregularities and color variation — both critical indicators in melanoma detection.

The datasets used to train our model — **ISIC 2019 and 2020** — are heavily skewed toward **Fitzpatrick skin types I–III**. Moreover, these datasets lack expert-annotated Fitzpatrick scale labels, necessitating a custom skin tone estimation pipeline for fairness analysis, as described below.

1.1 Skin Tone Estimation Methodology

To estimate skin tone for each image, the following procedure was applied:

1. **Lesion segmentation** was performed using a pretrained model to isolate the lesion from the surrounding skin.
2. **ITA (Individual Typology Angle)** values were calculated based on pixel values outside the segmented lesion.
3. Each image was then assigned to one of six skin tone categories based on its ITA score: *Very Light*, *Light*, *Intermediate*, *Tan*, *Brown*, and *Dark*.

These ITA-based skin tones were subsequently mapped to Fitzpatrick skin types as follows:

- Very Light → Type I
- Light → Type II
- Intermediate → Type III
- Tan → Type IV
- Brown → Type V
- Dark → Type VI

This approach is robust to minor segmentation errors, as it only requires a reasonable estimate of the background skin tone to derive meaningful ITA values.

1.2 Fairness Evaluation Metrics

To assess both the general performance of the model and its behavior across different skin tone groups, we employ the following metrics:

General Performance Metrics

- Accuracy
- Precision
- Recall
- F1 Score
- AUC

Fairness Metrics (Across Skin Tones)

- Disaggregated Performance Metrics (per-skin tone accuracy, precision, recall, F1)
- Demographic Parity
- Accuracy Parity
- Equality of Odds
- Calibration Across Groups
- Expected Calibration Error

These metrics allow us to quantify both the overall effectiveness of the model and its fairness in treating individuals with varying skin tones equitably.

1.3 Note on Skin Tone Estimation

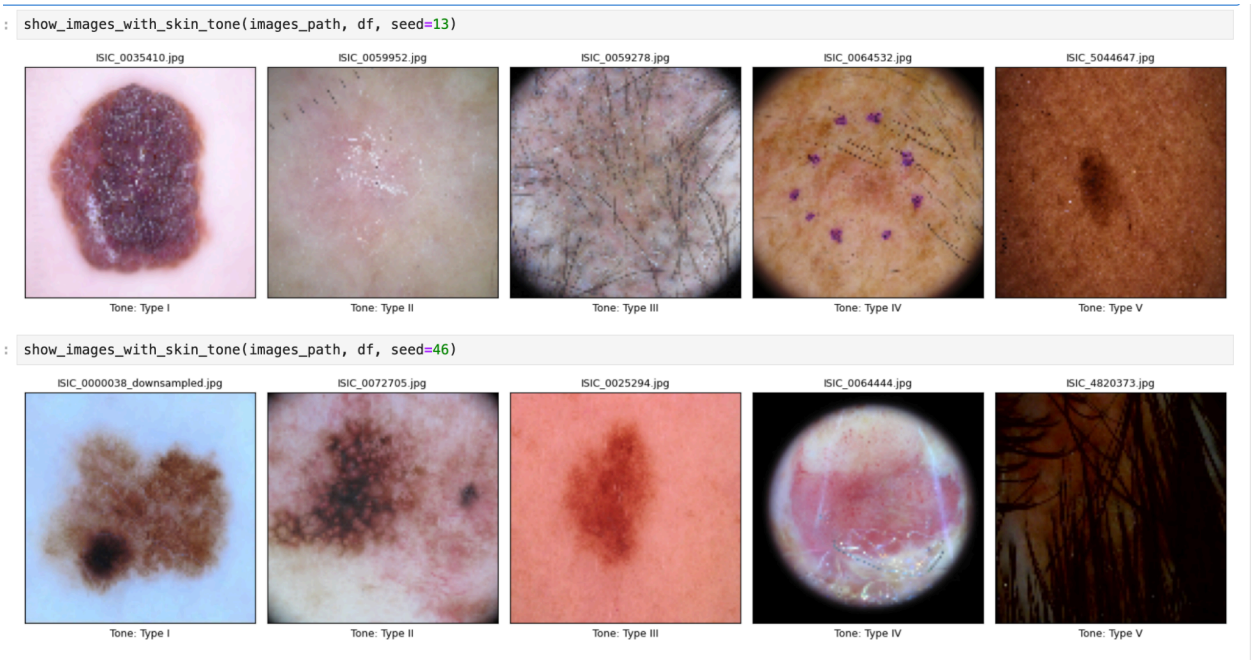
The performance of our skin tone estimation pipeline yields mixed results. The pipeline itself is based on lesion segmentation followed by ITA (Individual Typology Angle) value computation.

The model performs consistently well for lighter skin tones (Fitzpatrick Types I–III). However, several challenges arise when estimating darker skin tones (Types IV–VI), resulting in reduced reliability. These challenges include:

- **Lesions occupying the entire image**, leaving insufficient background skin for tone estimation.
- **Excessive presence of hair**, obscuring both the lesion and surrounding skin.
- **Black margins or padding** surrounding the image, which skew the ITA calculation toward darker tone estimates.

Due to these limitations, the estimated number of images classified as Fitzpatrick Types IV and V is likely **slightly inflated**. Notably, despite the model's tendency to over-classify darker tones, **no images were assigned to Type VI**, underscoring both the limitations of our estimation method and the deeper institutional bias present in existing skin cancer datasets.

Furthermore, the absence of expert-annotated skin tone labels in the metadata prevents precise validation of our estimation procedure. As such, while the method offers a useful approximation, its reliability for darker skin tones remains uncertain and warrants further investigation or expert calibration.



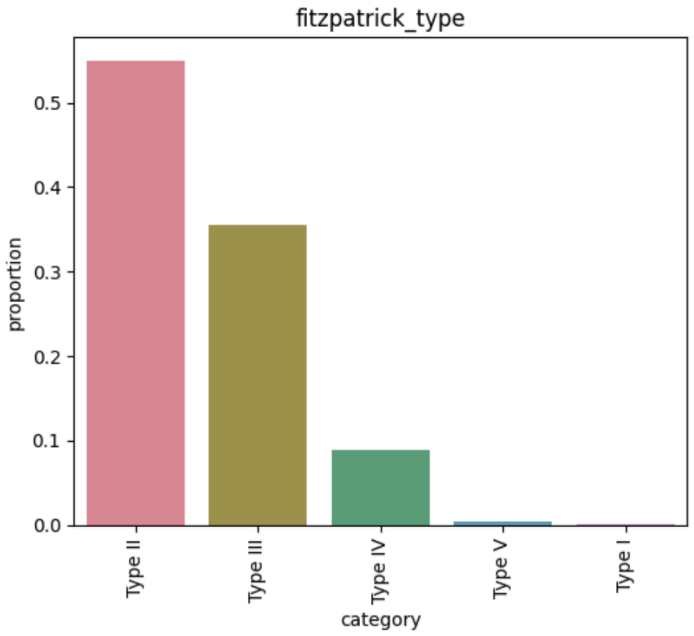
2. Exploratory Data Analysis

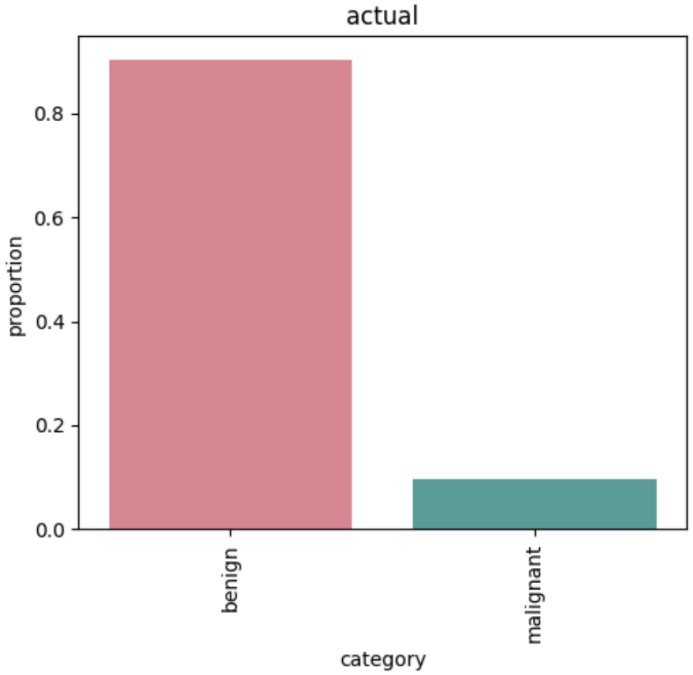
2.1 Value Distributions

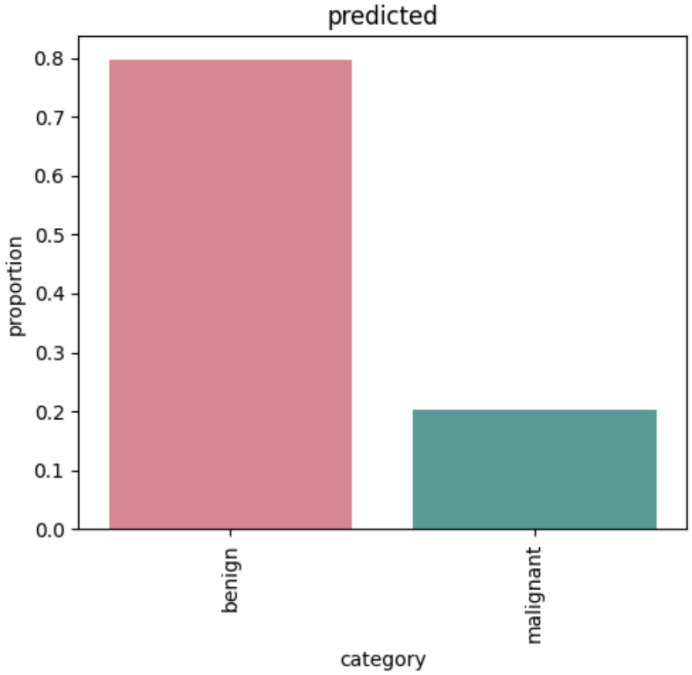
Below, we observe the following:

- The target variable is highly imbalanced.
- Predicted labels do not perfectly align with the ground truth.
- Lighter skin tones (Fitzpatrick Types I–III) account for approximately 91% of the estimated skin tone distribution.

```
actual
benign       0.903598
malignant    0.096402
Name: proportion, dtype: float64
-------------
predicted
benign       0.796702
malignant    0.203298
Name: proportion, dtype: float64
-------------
fitzpatrick_type
Type II     0.549625
Type III    0.355772
Type IV     0.088606
Type V      0.003898
Type I      0.002099
Name: proportion, dtype: float64
-------------
```

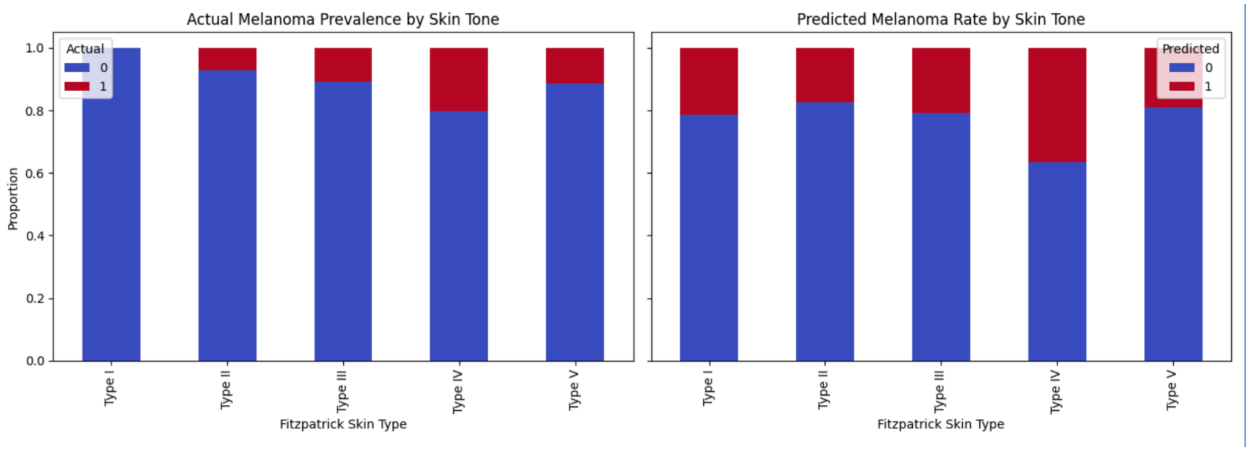






2.2. Observations on Melanoma Prevalence Across Skin Tones

The figures below illustrate melanoma prevalence across estimated skin types, both in the ground truth and the model's predictions.



Interestingly, **darker Fitzpatrick types exhibit a higher rate of melanoma** in this dataset, which contrasts with epidemiological data from real-world populations. This imbalance is most evident with **Type I skin that does not contain a single melanoma case.** One possible explanation is that individuals with darker skin may be **less likely to seek dermatological evaluation unless the lesion appears particularly suspicious**. This hypothesis, while plausible, remains speculative and would require rigorous clinical investigation to confirm.

Furthermore, the model appears to **amplify this trend**, as evidenced by an increased number of **false positive predictions** for darker skin tones. This suggests that the model may be **overcompensating** in the absence of balanced training data and highlights the importance of dataset diversity and fairness-aware evaluation

3. Performance and Fairness Evaluation

3.1 Across skin tones collectively

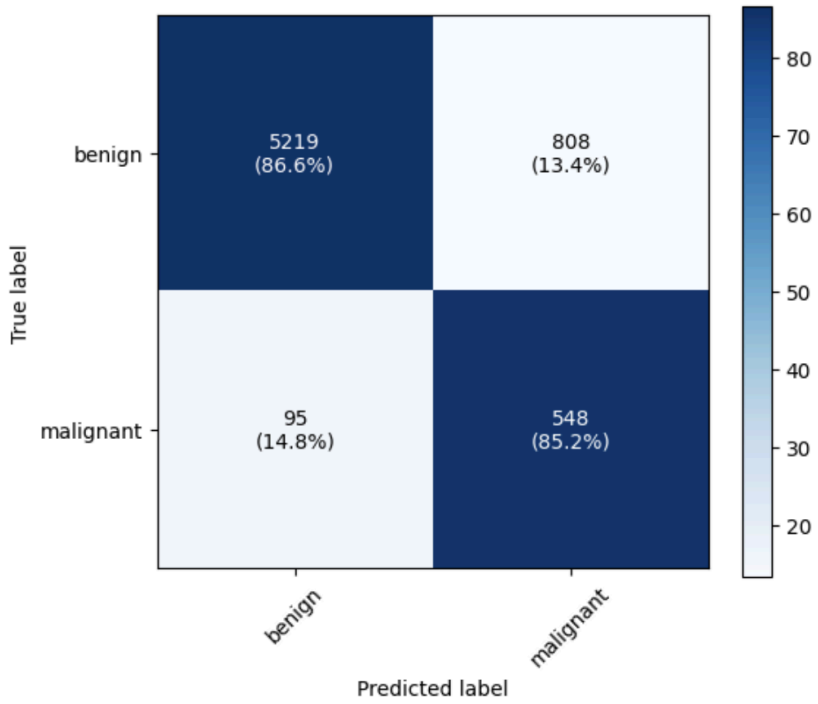
Despite significant class imbalance in the target variable and varying image quality, the model demonstrates promising performance. Given the imbalance, **accuracy is not an appropriate metric** for evaluation.

As the model's task is to distinguish between benign and malignant lesions, it is critical that **melanoma cases (positives) are correctly identified**, as false negatives could have serious clinical consequences. To prioritize sensitivity, the model allows for a higher number of false positives.

The model achieves a **recall of 85%**, successfully identifying the majority of melanoma cases, and an **AUC of 0.94**, indicating strong overall discriminative ability. However, the relatively **low precision** results in an **F1 score of 0.55**.

Overall, the model shows potential, but there remains **room for improvement**, particularly in reducing false positives and enhancing precision. Future work could focus on refining model calibration and incorporating additional clinical features to elevate performance to production-level standards.

	fitzpatrick_type	accuracy	precision	recall	f1	auc
0	all	0.8646	0.4041	0.8523	0.5483	0.9402



3.2 Across skin tones individually

3.2.1 Disaggregated Performance Metrics

Disaggregated performance metrics refer to the basic performance indicators (accuracy, precision, recall, F1 score, and AUC) computed **separately for each skin tone group**. This allows us to assess how model performance varies across different Fitzpatrick types, rather than relying solely on overall metrics.

Since there are **no melanoma cases for Type I skin**, the precision, recall, F1 and AUC couldn't be calculated for this subgroup.

	fitzpatrick_type	accuracy	precision	recall	f1	auc
0	all	0.8646	0.4041	0.8523	0.5483	0.9402
1	Type I	0.7857	0.0000	0.0000	0.0000	NaN
2	Type II	0.8740	0.3432	0.8429	0.4878	0.9404
3	Type III	0.8673	0.4431	0.8417	0.5806	0.9344
4	Type IV	0.7986	0.5023	0.9000	0.6448	0.9484
5	Type V	0.8462	0.4000	0.6667	0.5000	0.8116

Above, we observe that **performance differences across skin tones are notable, though not extreme**. Interestingly, **F1 scores are higher for slightly darker skin tones**, with the highest performance observed for **Fitzpatrick Type IV**, and the lowest for **Type II**. This is somewhat counterintuitive, given the model's training data is heavily skewed toward lighter skin tones.

One possible explanation for this trend is because on average, melanoma images are darker than benign images. We observed this fact in the initial EDA. The relatively **small sample size** for darker skin types may also inflate certain performance metrics, particularly if the model is highly sensitive within a narrow distribution of examples.

Interestingly enough, the **AUC values for Type II, III and IV skin are almost identical**. For Type I and Type V skin **AUC is notably lower**, but that too can be attributed to the small sample size of the test set, and the general underrepresentation of very light and very dark skin tones.

Further investigation is warranted to determine whether these trends persist with a more balanced dataset and expert-annotated skin tone labels.

3.2.2. Demographic Parity

Demographic parity evaluates whether a model treats different demographic groups equally - **regardless of actual outcomes**

In simple tearms, it asks the question: Does the model predict the positive class (melanoma) at the same rate for different groups (skin types)?

```
fitzpatrick_type
Type I      0.214286
Type II     0.174850
Type III    0.207332
Type IV     0.363790
Type V      0.192308
Name: predicted, dtype: float64
```

A **Demographic Parity Difference (DPD)** of **0.189** (Type IV - Type II) indicates a disparity in the model's prediction rates across demographic groups. Specifically, there is a **18.9 percentage point difference** in the proportion of positive predictions (melanoma diagnoses) between Type II and Type IV skin types.

Possible Causes of Demographic Parity Disparity

The observed disparity in **demographic parity** is likely the result of a combination of the following factors:

- **Higher observed melanoma rates for Fitzpatrick Type IV** in the dataset, which may lead the model to assign higher probabilities of melanoma to individuals with darker skin tones.
- **Model bias introduced by visual features** associated with darker skin or darker image regions, which may cause the model to **overpredict the positive class** in these cases.

Implications

This level of disparity is **concerning** in high-stakes settings such as medical diagnostics. While demographic parity does not account for differences in disease prevalence between groups, a value of **0.189** may reflect **potential bias** in model behavior In this context, the disparity may result in:

- **Overdiagnosis** (increased false positives) in darker skin tones,
- Or **underdiagnosis** (increased false negatives) in lighter skin tones.

Having considered the value of the disparity (**18.9**) and the underlying distributions in the dataset, we could say that the model is **slightly amplifing the higher melanoma trends** observed in the test set.

3.2.3 Accuracy Parity

Accuracy Parity is a fairness metric that checks whether a model is **equally accurate across different demographic groups** (Fitzpatrick skin types)

It asks: Is the overall classification accuracy consistent across groups?

An **Accuracy Parity Difference (APD)** of **0.08** indicates that the model's accuracy for Fitzpatrick **Type II** skin is **8 percentage points higher** than for **Type I** skin. While this level of disparity is generally considered acceptable in many applied machine learning contexts, it should still be interpreted with caution in a high-stakes medical context like melanoma detection

```
fitzpatrick_type
Type I      0.785714
Type II     0.873977
Type III    0.867257
Type IV     0.798646
Type V      0.846154
Name: correct, dtype: float64
```

One plausible explanation for this difference is the **unequal distribution of samples** across skin types and **small sample sizes**. Type II skin tones are typically overrepresented in public dermoscopic datasets, leading to better model generalization for that group. Conversely, the relatively small number of samples for Type I may result in **less reliable learning**, leading to lower accuracy.

3.2.4 Equality of Odds

Equality of Odds is a fairness metric that checks whether a model has **equal error rates across demographic groups** (Fitzpatrick skin types)

It asks: Are both the **True Positive Rate (sensitivity/recall)** and **False Positive Rate (fall-out)** consistent across groups?

This ensures that the model:

- Detects disease equally well across all skin tones (equal TPR)
- Does not overpredict disease for any group (equal FPR)

	group	TPR	FPR
0	Type II	0.842912	0.123642
1	Type III	0.841699	0.129612
2	Type I	NaN	0.214286
3	Type IV	0.900000	0.227176
4	Type V	0.666667	0.130435

TPR difference: 0.2333333333333334

FPR difference: 0.10353451742940786

The observed **True Positive Rate (TPR) difference** aligns with earlier findings from other fairness metrics: the model identifies melanoma cases **at a higher rate for Type IV skin tones**. This indicates that the model is more sensitive (i.e., has higher recall) for these groups.

This elevated TPR is accompanied by a corresponding increase in the **False Positive Rate (FPR)**, which is an expected trade-off. A lower decision threshold for classifying a lesion as malignant naturally increases the likelihood of capturing true positives, but also raises the risk of misclassifying benign lesions as malignant.

In summary, the model exhibits **unequal TPR and FPR across skin tone groups**, violating the criteria for **Equalized Odds**. This suggests that the model's decision boundary is not equally calibrated across demographic groups and may lead to differential clinical consequences depending on skin tone.

3.2.5 Calibration Across Groups and ECE

Calibration Across Groups evaluates whether a model's **predicted probabilities** are **equally reliable across different demographic groups.** It asks: When the model predicts a probability of 70%, does that reflect a 70% actual chance of melanoma — for every group?

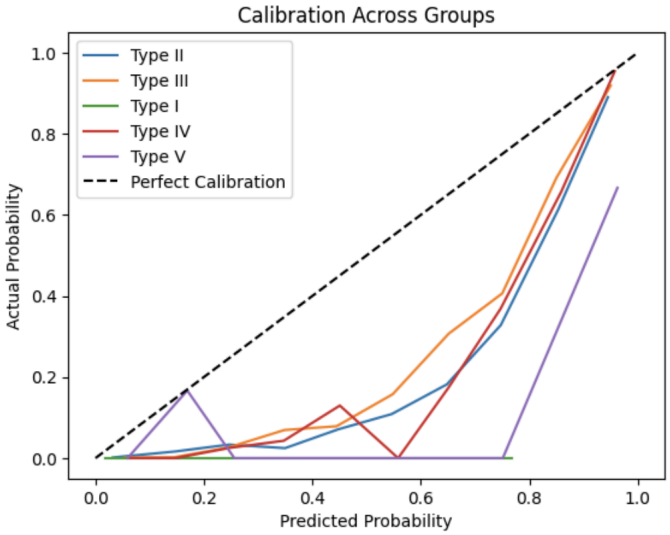
Calibration is especially important in disease detection because:

- Doctors may use the predicted probability to decide whether to biopsy a lesion.
- A poorly calibrated model may **overestimate risk** for one group and **underestimate it** for another, leading to **unequal treatment**.

The model is perfectly calibrated if the points for each group lie on the diagonal line**.**

ECE stands for **Expected Calibration Error** — it is a metric that tells you **how well a model's predicted probabilities match the actual outcomes**. In other words: When your model says "there's a 70% chance this lesion is melanoma," is it actually right 70% of the time?

ECE quantifies the **average mismatch** between what your model *thinks* is true (its confidence) and what's *actually* true (the ground truth), across **all probability levels**.



	group	ECE
0	Type II	0.1549
1	Type III	0.1874
2	Type I	0.2321
3	Type IV	0.2309
4	Type V	0.2201

As with the other fairness metrics, the **Expected Calibration Error (ECE)** values further support the observation that the model exhibits **bias toward false positives**, particularly for individuals with **darker skin tones**.

Since Type I skin has such a small sample size and has not malignant cases, the ECE values for this group are not reliable. With the exception of Type I skin, the ECE increases progressively from lighter to darker skin types, indicating that the model's **predicted probabilities become less reliable** as skin tone darkens. Specifically, for **Fitzpatrick Types IV and V**, the model is significantly **overconfident** in

its positive predictions — assigning high probabilities to many benign cases, resulting in a higher rate of **false positives**.

This pattern is consistent with earlier findings on **TPR** and **FPR disparities**, and confirms that the model does not maintain consistent calibration across demographic groups. This may reflect a combination of:

- **Data imbalance** across skin tones,
- **Overfitting to overrepresented groups**, and
- **Spurious correlations in image features associated with darker skin tones**.

Addressing these issues may require incorporating more diverse training data and applying **fairness-aware calibration techniques**.

4. Summary and Conclusion

The model shows strong **overall discriminative ability** (AUC = 0.9402) and prioritizes **sensitivity** (recall = 85%) in melanoma detection. However, it does not satisfy key fairness criteria such as **demographic parity**, **accuracy parity**, or **equalized odds** across skin tones. Moreover, its predicted probabilities are less reliably calibrated for darker skin types.

Disaggregated analysis revealed that, while **F1 scores were slightly higher for some darker skin tones** (e.g., Type IV), this trend appears to stem from **elevated recall at the cost of increased false positives**. A **Demographic Parity Difference (DPD) of 0.189**, higher **False Positive Rates (FPR)**, and larger **Expected Calibration Errors (ECE)** for darker tones all support the conclusion that the model is more likely to **overpredict melanoma** for these groups.

This pattern reflects both **dataset imbalances** (91% of samples from Types I–III) and **limitations in the skin tone estimation pipeline**. Importantly, while prioritizing sensitivity can be justified in clinical applications to minimize missed melanoma cases, the resulting disparities raise concerns about equitable treatment across populations.

To move toward a fairer and more clinically reliable system, future work should focus on:

- Expanding and balancing datasets,
- Improving skin tone estimation pipelines or obtaining expert-annotated skin tone labels,
- Exploring fairness-aware training, calibration, and threshold adjustment techniques,
- Investigating the controlled use of artificial data augmentation (e.g., skin tone darkening).

Such efforts are essential to ensure **equitable and trustworthy AI-assisted dermatology** for all patient populations.