Fair Melanoma Detection
Project Documentation

Table of Contents

Table of Figures

1. Introduction

Melanoma is an aggressive form of skin cancer that can progress rapidly if not detected early. Timely and accurate diagnosis is critical, especially given that clinical assessment can vary depending on a range of visual features and patient demographics. To address this diagnostic challenge, the project focuses on building a reliable and interpretable melanoma classification pipeline using dermoscopic images and patient metadata.

This solution was developed for the LUMEN Data Science 2024/25 competition which emphasizes not just accuracy, but also fairness and robustness – particularly across diverse skin tones. The following approach is grounded in medical relevance and data science reliable practices, integrating techniques to support generalization, explainability and bias mitigation.

At the core of the pipeline is a deep learning model based on a modified ResNet101 architecture, trained to classify lesions as malignant or benign. Prior to model training, extensive preprocessing is applied to improve image quality and standardize inputs. This includes lesion-focused cropping, hair artifact removal, and image resizing, which help the model focus on clinically relevant features. Alongside image data, metadata is used to link images with their corresponding masks, manage dataset splits, and provide file paths for loading images and masks during training and validation.

To enhance model performance and reduce overfitting, the training process incorporates data augmentation, dropout, and other regularization strategies. Beyond performance metrics, we also prioritize explainability using segmentation visualizations, allowing us to inspect model attention and confirm that predictions are based on medically meaningful regions of the image.

Importantly, fairness remains a central concern. The solution is evaluated not only on overall accuracy but also on its performance across different skin tones and demographics. This aligns with the broader goal of creating AI tools that are trustworthy and equitable in real-world clinical settings.

The following documentation provides a detailed breakdown of the methodology, including preprocessing steps, model architecture, training strategy, evaluation metrics, and interpretability tools. Provided solution aims to balance clinical intuition with technical rigor – offering a step forward in responsible, effective melanoma detection.

2. Problem Statement

The primary objective of this project is to classify dermoscopic images into **malignant** or **benign** categories.

To meet the challenge's requirements, the solution must:

- Leverage dermatologically informed preprocessing and feature engineering.
- Integrate both visual and associated non-visual (metadata) inputs in a consistent and meaningful way.
- Employ techniques to prevent overfitting, manage class imbalance, and avoid data leakage.
- Demonstrate robustness and fairness across skin tones and demographic groups.

The challenge permits the use of dermoscopic image sets collected between 2016 and 2020 and allows the inclusion of additional public datasets for model development, as long as their usage is well-documented and reproducible. In line with this, we incorporated selected samples from the Fitzpatrick17k dataset to address skin tone imbalance and improve model fairness across demographic groups. These publicly available images were integrated into the training process to support generalization and reduce bias in melanoma classification.

Evaluation criteria extend beyond overall classification accuracy, incorporating fairness metrics to ensure balanced performance across diverse skin tones. The final solution should prioritize clinical reliability, interpretability, and ethical responsibility in its predictions.

3. Exploratory Dataset Analysis

The project utilizes dermoscopic image datasets from 2016 to 2020. Each dataset includes:

- **Images**: High-resolution dermoscopic images of skin lesions.
- **Metadata**: Patient information, lesion type, and other attributes (depending on the year).

Building a high-performing and fair melanoma classifier using this dataset presents several real-world challenges:

3.1 Class Imbalance

Malignant cases are significantly underrepresented compared to benign cases, which can lead to biased predictions favoring the majority class. To address this, the training pipeline incorporates methods such as class reweighting and generates additional training examples through augmentation techniques applied to existing images.

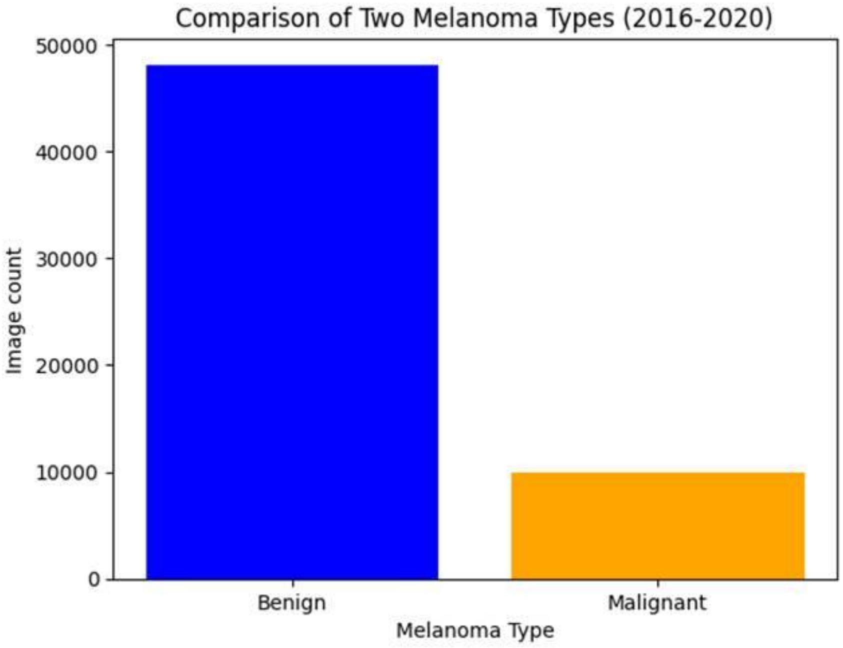


Figure 1. Comparison of two melanoma types (2016-2020)

3.2 Imbalance per Patient

The melanoma ISIC 2020 dataset exhibits an imbalance in the number of images per patient, where some patients contribute multiple images (some even more than 100) while others may only have a single image. To address this, we ensured that images from the same patient were not included in both the training and test sets, maintaining their independence.

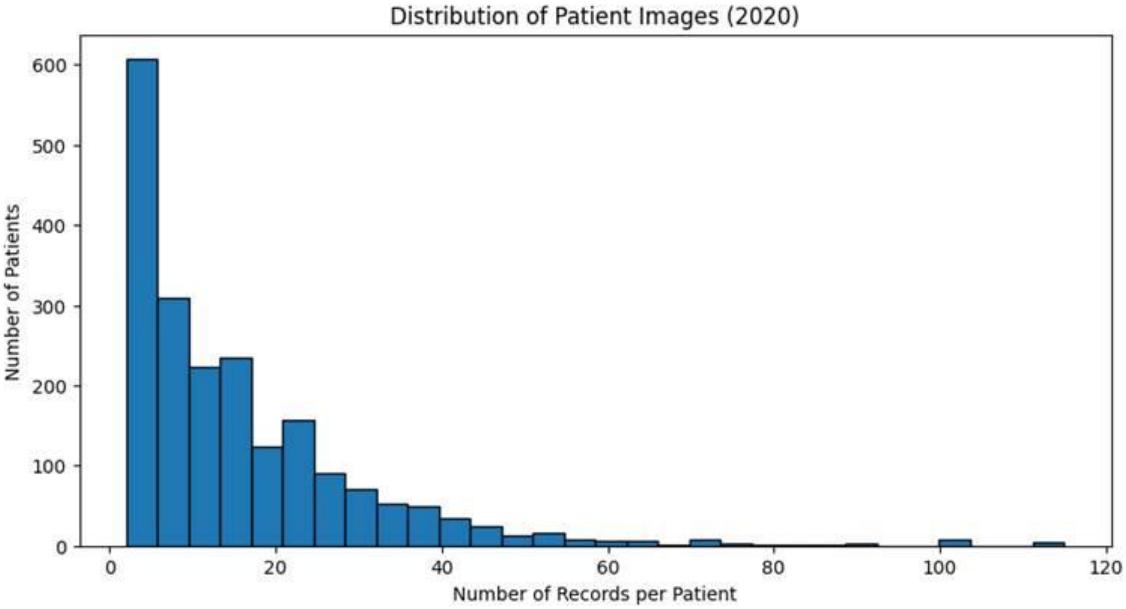


Figure 2. Distribution of patient images

3.3 Skin Tone Variability

Skin color can significantly affect the visual appearance of skin lesions, potentially leading to biased model performance across demographic groups. To ensure fairness and reduce disparities in clinical outcomes, it is important to evaluate model behavior across subgroups and consider skin tone-aware preprocessing or post-hoc fairness checks.

Given the underrepresentation of darker skin tones in the original dataset, we incorporated 127 additional images from the Fitzpatrick17k dataset, focusing on skin types IV, V, and VI. After filtering for relevant diagnoses – grouping melanoma, basal and squamous cell carcinoma, and actinic keratosis as malignant, and common benign types such as nevi and seborrheic keratosis – we manually selected images to ensure clinical relevance.

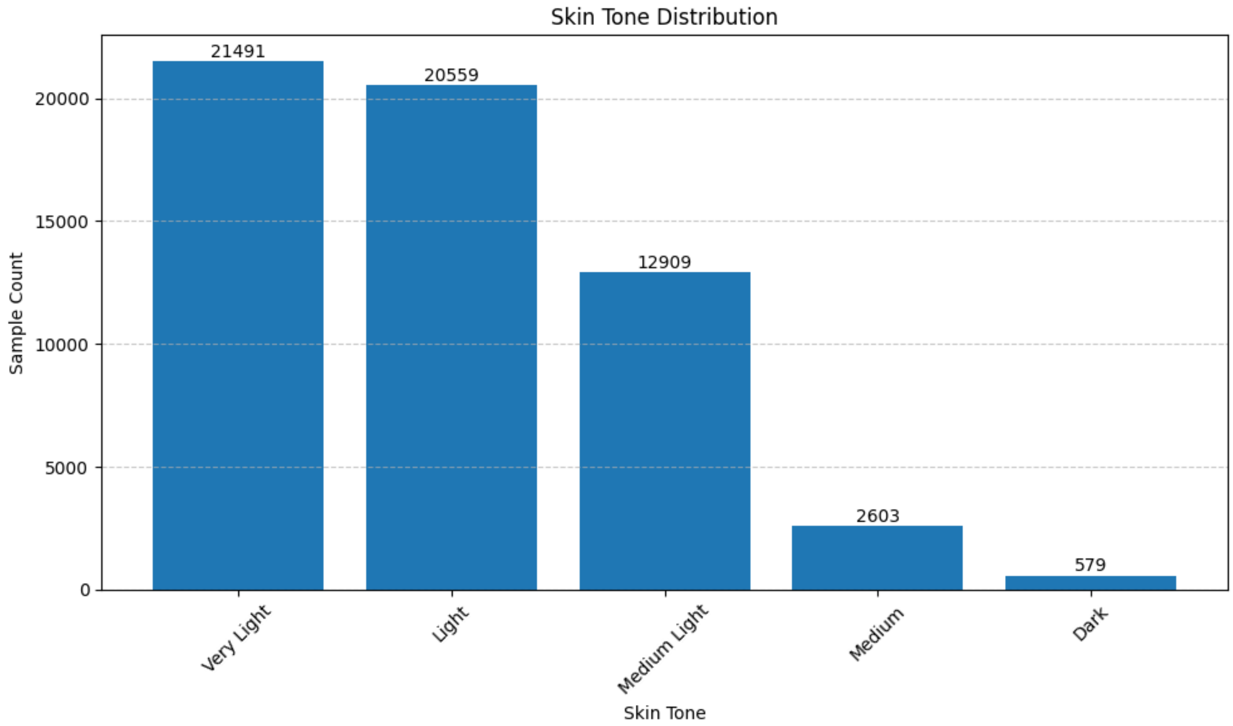


Figure 3. Skin tone distribution (2016-2020 and Fitzpatrick17k)

3.4 Image Artifacts

Many images contain visual noise – such as hair, shadows or glare, ink marks and similar – that can obscure lesion features and mislead the model. The preprocessing pipeline includes artifact reduction techniques like hair removal, contrast enhancement, and lesion-focused cropping to improve input quality.

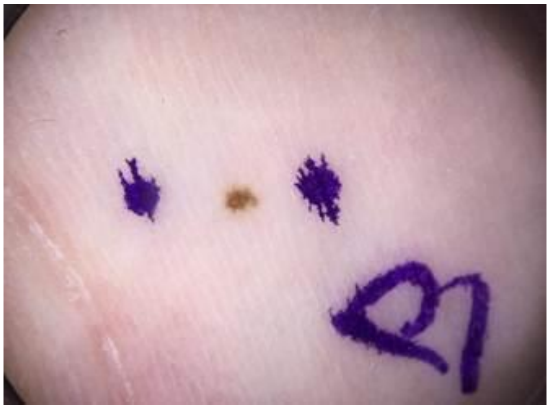


Figure 4. Presence of clinical markings

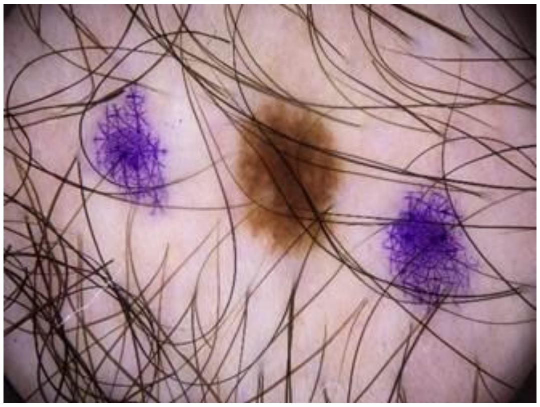


Figure 5. Presence of clinical markings and hair

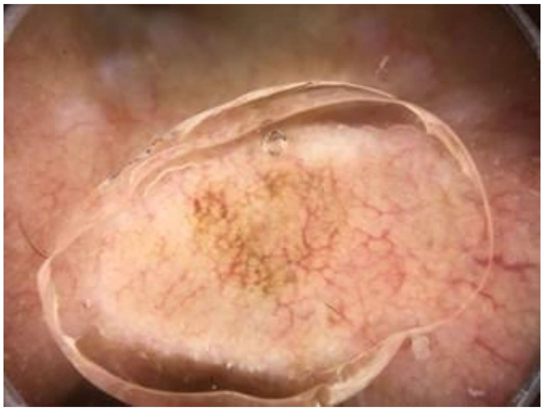


Figure 6. Presence of immersion fluid and air bubbles distorting the lesion view

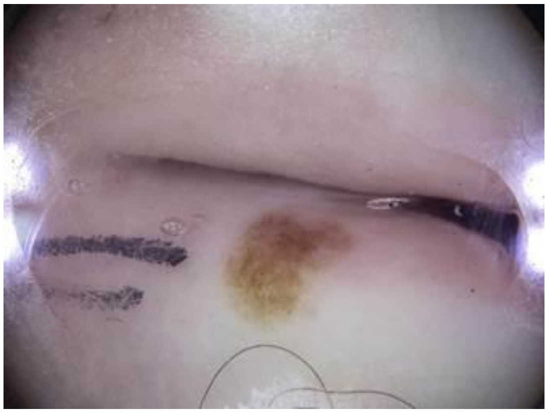


Figure 7. Presence of immersion fluid causing distortion of lesion

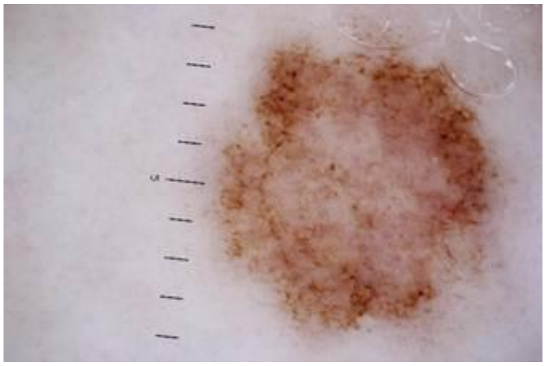


Figure 8. Presence of measurement overlay and immersion fluid air pocket.

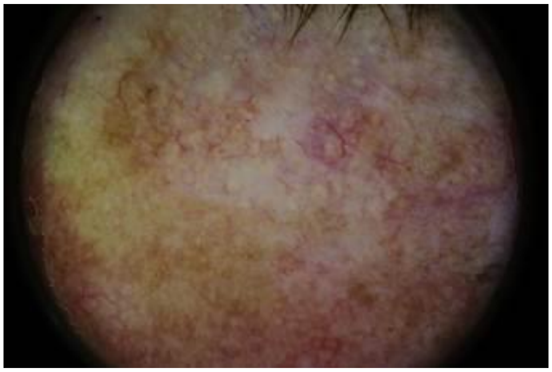


Figure 9. Obfuscation by a dermoscope

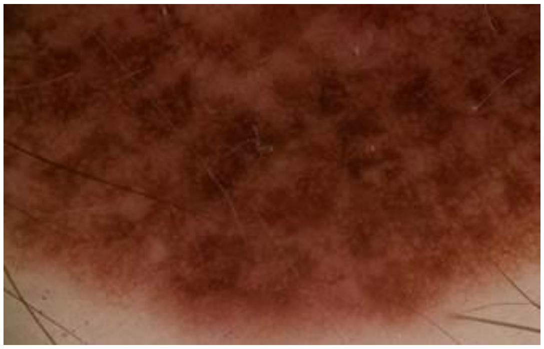


Figure 10. Cropped lession

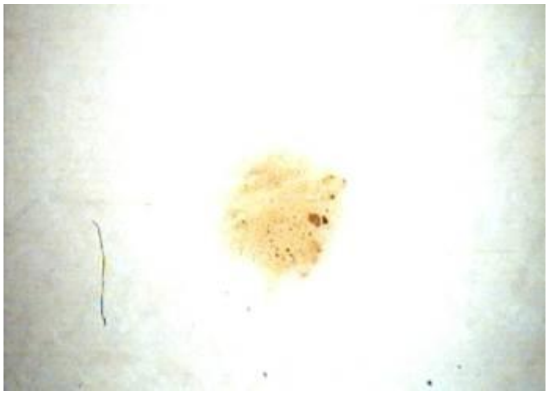


Figure 11. Presence of glare

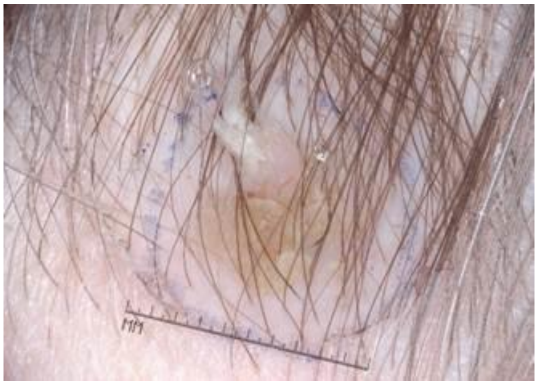


Figure 12. Lesion obfuscated by hair, measurement overlay and immersion fluid

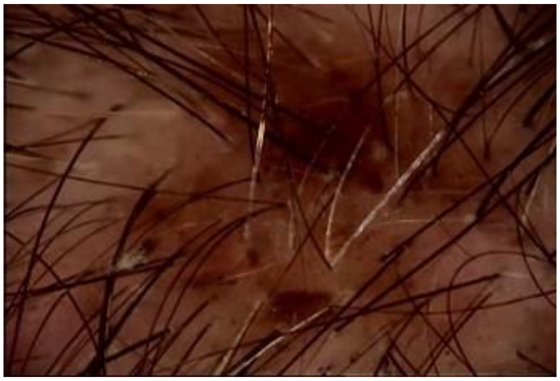


Figure 13. Lesion obfuscated by hair

Images and metadata collected over multiple years vary in quality, resolution, and format. This heterogeneity introduces distribution shifts that can degrade model performance if not properly addressed. Applying preprocessing and normalization techniques across datasets helps to mitigate these issues. The following section will provide a brief explanation of the image preprocessing methods, the model development, and how the processed images were integrated into the final pipeline.

Dataset	Benign samples	Malignant samples	Total
2016	727	173	900
2017	1626	374	2000
2018	8061	1954	10015
2019	15991	9340	25331
2020	32542	584	33126
Split total	58947	12325	71372

Table 1. Display of the number of images in the ISIC datasets by year and number of images in total

4. Methodology and Approach

The development of a reliable and fair melanoma classification system requires a carefully structured methodology. This pipeline integrates dermatologically adjusted preprocessing, robust model design and fairness-aware evaluation – ensuring clinical relevance, technical trust and reproducibility of results.

4.1 Data Organization and Metadata Handling

To ensure robust model training and mitigate potential biases, we implemented an extensive data organization strategy.

Duplicate Removal: Acknowledging the presence of duplicate images within and across ISIC datasets from 2016 to 2020 – as discussed in Analysis of the ISIC image datasets, duplicate removal procedure is applied to curate a non-redundant dataset. We identified and eliminated duplicate images to prevent data leakage between training and validation sets, thereby enhancing the model's generalization capabilities.

Dataset	Benign samples	Malignant samples	Total
2016	70	6	76
2017	1126	233	1359
2018	0	0	0
2019	14874	9088	23962
2020	32112	581	32693
Split total	48182	9908	58090

Table 2. Display of the number of images in the ISIC datasets by year and number of images in total after using duplicate removal techniques

Data Splitting Strategy:

- **ISIC 2020 Dataset:** Utilizing available patient IDs, we performed patient-level splitting to avoid data leakage. All patients with at least one malignant lesion were included, while the number of benign samples per patient was limited to maintain balance.
- **ISIC 2016–2019 Datasets:** In the absence of patient IDs, we conducted stratified splitting based on combined target class and estimated skin tone ({target}_{skin_tone}), ensuring representation of minority subgroups and balanced sampling of majority classes.

Class Imbalance Handling:

- **Malignant Lesions:** Due to their underrepresentation, malignant samples were duplicated within the training set. Accordingly, appropriate augmentation techniques, such as rotations and noise addition, were applied to minimize the risk of overfitting.
- **Skin Tone Representation:** Underrepresented skin tone groups were upsampled to enhance model fairness and reduce bias towards dominant groups.

4.2 Preprocessing Pipeline

Preprocessing was designed to enhance the medical quality of dermoscopic images and standardize inputs across multiple years of data collection. The following steps were implemented:

1. **Skin Tone Classification:**
   As the pipeline evolved, we identified the need for a more even approach to skin tone estimation. To address this, we developed a custom classification pipeline based on image analysis techniques, which estimates tone by analyzing the brightness of lesion surrounding skin pixels. Additionally, due to the complete underrepresentation of darker skin tones in the dataset, we expanded it by manually selecting 127 relevant images from the Fitzpatrick17k dataset (110 malignant, 17 benign). This targeted addition supports fairness-aware evaluation and enables more balanced training and validation splits across skin tone groups.

2. **Segmentation and Lesion Cropping:**
   We initially used the manually annotated lesion masks available in the ISIC 2017 and 2018 datasets to train a segmentation model. These ground truth masks enabled the model to learn to identify lesion borders with better precision which was then applied to all images - including those without provided masks - to generate lesion segmentations necessary for further preprocessing. Once the lesion mask was predicted, each image was cropped around the lesion area, effectively removing irrelevant background. This focuses the model's attention on relevant features and reduces visual noise.
3. **Hair Removal:**
   Visual obstructions such as hair strands were removed using classical image processing filters. This step ensures that lesion borders remain clear and undistorted, that way hair artifacts do not interfere with model performance. The hair removal step preprocesses skin lesion images by detecting and removing hair artifacts using morphological blackhat operations, thresholding and inpainting, similar to the approach demonstrated in Melanoma Hair Remove.
4. **Image Resizing:**
   In given preprocessing pipeline, image resizing was a critical step to standardize input dimensions for deep learning models. After lesion segmentation and cropping, we experimented with various resolutions (224, 256, 378, 512, 1024) to balance detail preservation and computational efficiency. Ultimately, we standardized images to 512×512 pixels, a dimension that maintained sufficient detail for accurate classification while ensuring manageable training times. This decision aligns with findings in the literature, where resizing images is shown to impact model performance and training time.
5. **Storing:**
   Once all steps are complete, the image is saved to disk to avoid repeating these computationally intensive operations and speed up the training process.

These preprocessing techniques reflected as critical methods for improving both model performance and fairness.

4.3 Model Development

The system includes two core deep learning models: a segmentation model for lesion detection and cropping, and a classification model for melanoma prediction. Both models were developed with attention to generalization, class imbalance, and training efficiency.

A U-Net model was trained to segment lesion areas using labeled masks from earlier datasets. For classification, we used a fine-tuned deep learning model optimized for both accuracy and fairness.

To improve generalization, we applied a range of augmentations: random flips, rotations, zooming, contrast adjustments, and varying input sizes. These techniques simulate diverse imaging conditions and simultaneously reduce the risk of overfitting. Special attention was given to the classes, and stratified sampling was employed across melanoma types and skin tone categories to ensure balanced representation in both training and validation datasets.

The models were trained using advanced combination of regularization techniques, including dropout and batch normalization, to stabilize learning and prevent overfitting, and the AdamW optimizer, which adapts the learning rate dynamically for efficient convergence. Additionally, early stopping was employed to halt training if validation performance did not improve, ensuring optimal model performance.

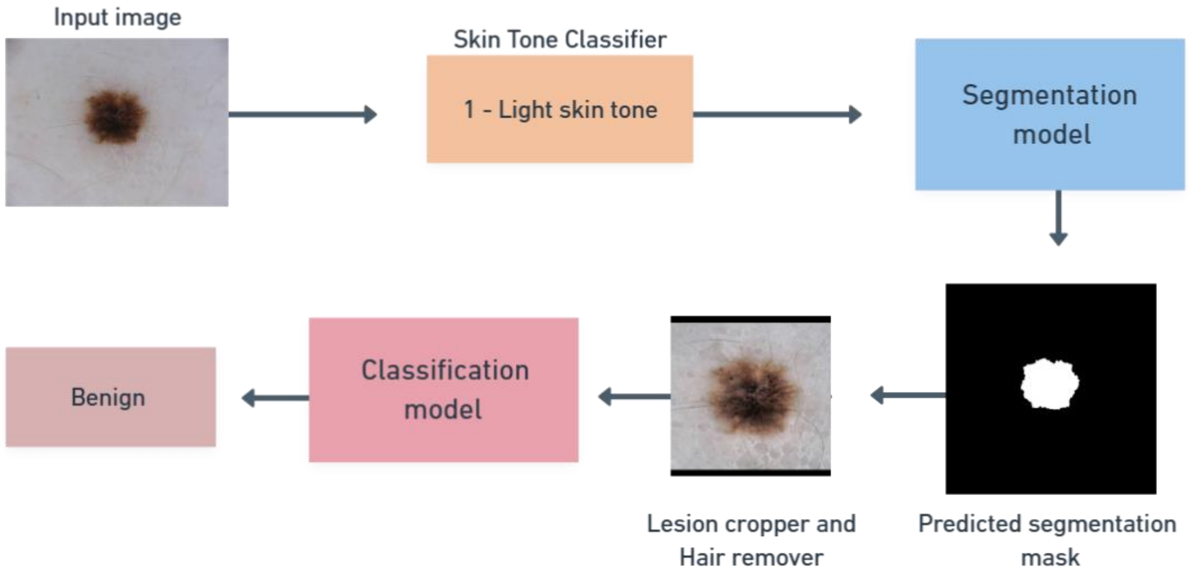


Figure 5. Illustration of model pipeline

4.4 Evaluation and Fairness

The model was evaluated using a comprehensive set of metrics:

- **Performance Metrics:**
  Sensitivity, specificity and precision-recall curves were used to assess classification accuracy and balance. Their detailed analysis and interpretation are presented in the Results section.
- **Fairness Assessment:**
  Model performance was analyzed across different skin tone categories to identify disparities and ensure equitable diagnostic outcomes.
- **Explainability:**
  To enhance reliability and interpretability, the classifier is trained exclusively on segmented lesion images. This approach removes background noise and ensures the model focuses on clinically relevant features such as borders, color, and asymmetry - consistent with the ABCDE dermatological criteria. As a result, the model's predictions are based on image regions that align with expert medical evaluation.

This methodology integrates technical rigor with clinical awareness. From robust preprocessing to fairness-centered evaluation, the pipeline is designed to deliver a melanoma classification tool that is accurate, interpretable, and ethically responsible – ready for real-world application in diverse clinical settings.

5. Results

The melanoma classification model achieved strong results, with an overall accuracy of **93.8%** and a **weighted F1 score of 0.94**. Sensitivity to malignant cases (recall: **89%**) and specificity for benign cases (recall: **95%**) reflect a well-calibrated balance between detecting critical cases and minimizing false positives. However, the risk associated with false negatives highlights the importance of integrating such a model into clinical decision support rather than using it as a standalone diagnostic tool.

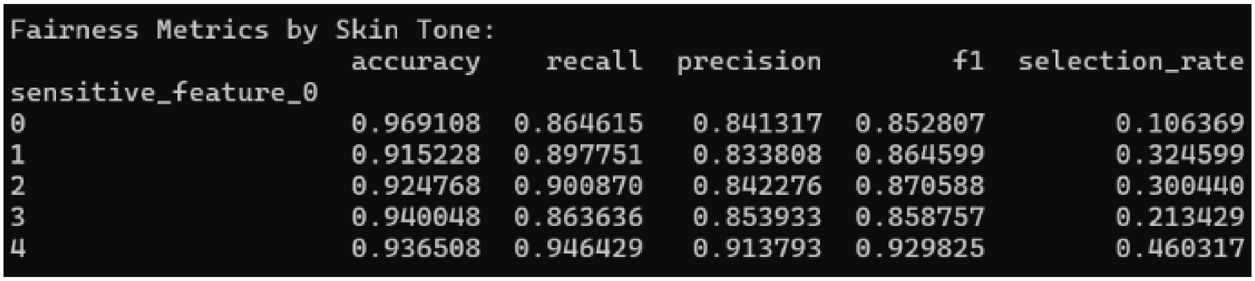
The dataset exhibits a class imbalance with a benign-to-malignant ratio of approximately 3.4:1, which was taken into account during model training and evaluation to ensure balanced performance across both classes. On the validation set, the final model achieved a loss of 0.1977 and an accuracy of 93.77%, indicating strong overall performance and good generalization to unseen data.

Metric	**Value**	**Notes**
Accuracy	93.8%	Overall correctness across both classes (calculated from recall + support)
Recall (Malignant)	89%	High sensitivity to malignant cases; the model effectively identifies most malignant lesions with very few false negatives
Precision (Malignant)	84%	Some false positives present; slightly lower confidence in positives
F1 Score (Malignant)	87%	Balanced measure for malignant detection performance

Recall (Benign)	95%	High specificity for malignant class; avoiding false positives
Precision (Benign)	97%	Very few benign cases misclassified as malignant, minimizing false positives
F1 Score (Benign)	96%	Strong consistency in benign classification

Table 3. Summary of model evaluation metrics

To support fairness, performance across five skin tone groups is also monitored. While disparities were initially observed – especially for darker skin tones – these were partially mitigated through targeted dataset augmentation. Although performance gaps have narrowed, some fairness challenges remain, particularly around selection rate.



```
Fairness Metrics by Skin Tone:
                    accuracy    recall  precision        f1  selection_rate
sensitive_feature_0
0                   0.969108  0.864615   0.841317  0.852807        0.106369
1                   0.915228  0.897751   0.833808  0.864599        0.324599
2                   0.924768  0.900870   0.842276  0.870588        0.300440
3                   0.940048  0.863636   0.853933  0.858757        0.213429
4                   0.936508  0.946429   0.913793  0.929825        0.460317
```

Figure 5. Display of fairness metrics by skin tone

6. Conclusion

This project demonstrates a comprehensive and ethical approach to melanoma detection using machine learning. With alignment to dermatological principles and leveraging advanced machine learning techniques, the solution achieves high accuracy while maintaining fairness, transparency, and explainability. The methodology and results provide a solid foundation for both further academic exploration and real-world clinical adaptation.

During the development process, we encountered several challenges, particularly related to handling the large volume of data and addressing dataset imbalances. The underrepresentation of certain skin tones and lesion types posed difficulties in ensuring fairness and generalization across different demographic groups. A major downside of the manually added images from Fitzpatrick17k dataset is the poor quality of the images, which could affect model generalization. Despite implementing techniques such as data augmentation and sampling to address the class imbalance, the malignant class remained significantly smaller than the benign class, which impacted the model’s ability to detect rare cases. Additionally, the inconsistency in metadata across different years and institutions added complexity to preprocessing, requiring extra attention to ensure accuracy and fairness in evaluations. These hurdles highlighted the importance of careful data management and continuous adjustments to the model to achieve robust and fair results.

According to faced disadvantages, special attention was given to clinical alignment through lesion-centered preprocessing steps, addressing dataset biases such as skin tone imbalance and promoting model explainability through visual interpretability techniques. Strengths of the solution include its clinical grounding, fairness-focused design, and modular structure supporting reproducibility.

For future work, several directions are proposed to further improve the model and align it with real-world clinical deployment. Collaboration with dermatologists will be crucial to iteratively refine annotations, evaluate model predictions, and validate interpretability methods, thus improving trust and usability. Expanding the dataset to include more images from underrepresented skin tones and rare lesion types, as well as sourcing data from different institutions, would enhance the model’s robustness and generalization. High-resolution images, like those in the ISIC dataset, would be best for both major and auxiliary models. Additionally, incorporating advanced feature engineering – such as textural analysis, lesion shape descriptors or integration of clinical metadata like patient history – could further boost classification performance. External validation on independent datasets is necessary to evaluate the model's adaptability to different imaging conditions and institutions. Finally, exploring tone-aware augmentation strategies or domain adaptation techniques could help equalize model performance across diverse skin tones, ensuring a fairer and more clinically reliable solution.

7. Sources

1. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations - ScienceDirect

2. Melanoma Hair Remove

3. Dark Corner on Skin Lesion Image Dataset: Does It Matter?

4. Dermoscopic dark corner artifacts removal: Friend or foe? - ScienceDirect

5. 2010.05351v1

6. (PDF) Statistical Analysis of Hair Detection and Removal Techniques Using Dermoscopic Images