Projects in Data Science Final Project

https://github.com/bart758/2025-FYP-GroupF-2

Mihai Botoc
mbot@itu.dk
Marius Salcutan
masal@itu.dk

Bartosz Kochanski
bako@itu.dk
Lukáš Trstenský
lutr@itu.dk

El Mars Sabourin
elsr@itu.dk

Abstract

This project uses automated or semi-automated feature extraction from images of skin lesions for the purpose of classifying said lesions as Melanoma or Non-Melanoma. We use two different models, each with varying levels of complexity and multiple classification methods, to examine the differences between their performance and learn valuable information regarding their different qualities.

1 Introduction

Melanoma is a type of skin cancer that arises from melanin-producing skin cells called melanocytes, whose development is typically triggered by exposure to ultraviolet light such as that from the sun (Mayo Clinic, 2023; NCI, 2025). While accounting for only around 1 percent of all skin cancers, Melanoma is the reason for most skin cancer related fatalities, and is becoming increasingly common, affecting 3.1 million people and causing 59,800 death as of 2015, giving it the title as the most dangerous skin cancer (Cust et al., 2015; American Cancer Society, 2023).

About 30 percent of cases begin from existing moles and the rest in normal skin, making it challenging to identify, but possible with the naked eye. A more reliable method is dermoscopy, which is a tool that captures magnified skin lesions. Technological advancement has allowed the use of artificial intelligence to provide a danger score to dermoscopic images and distinguish between other types of skin cancers , However, this method is limited by the lack of training data due to the small number of patients, and by the complexity of identifying Melanoma from other skin lesions.

This report aims to support early diagnosis using a feature extraction approach based on the ABC (Asymmetry, Borders, Color) method. We applied the program-implemented features to a large collection of skin lesion images, processing and quantifying them based on characteristics associated with melanoma.

2 Dataset analysis and annotation

2.1 Dataset Source and Content

The data set used for this project is the PAD-UFES-20 (Pacheco et al., 2020) dataset, which contains 2,298 images of 1,641 skin lesions with 6 different diagnoses, taken from 1,373 patients. For our work, we will mainly focus on detecting melanoma cases. The distribution of melanoma and non-melanoma lesions is as follows:

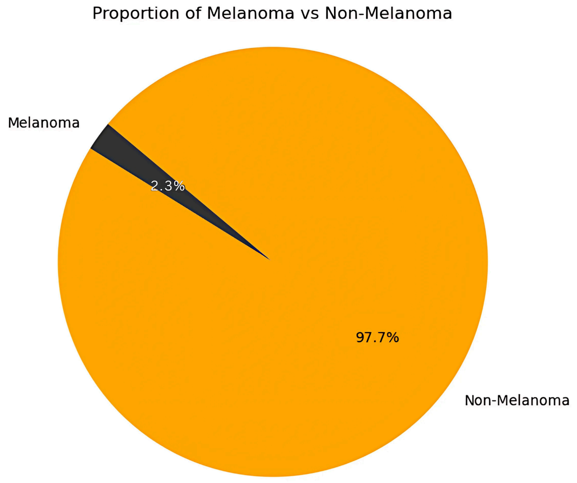

Figure 1: Distribution of data

In addition to the lesion images, there is also a metadata file with additional information about the patient as well as the lesion and the ground truth diagnostic of the lesion.

2.2 Hair Annotations

The presence of body hair on and around skin lesions is common, and that is also the case for melanoma. In fact, it may start in mutated hair follicles (NIH, 2019). The presence of hair around the lesion can block important visual features, which are key for diagnosis using ABC features, leading to worse accuracy in the results. To account for this, we manually annotated each of the images in the dataset based on the amount of hair visible in the image. For each image, we gave a rating of 0, 1, or 2 for no hair, some hair, and a lot of hair respectively. Our extended model incorporates a hair feature which computes the ratio of hair in the picture.

3 Baseline Model

The baseline model is our simplest approach to identifying melanoma in lesion images. It consists only of feature extraction and image classification, without any data preprocessing. The model follows this structure:

```
Load images from directory
For image in images:
    extract feature A;
    extract feature B;
    extract feature C;
With features and ground truth:
    Train model;
    Test model accuracy;
```

3.1 Feature extraction

Feature A measures the asymmetry of the lesion area, feature B measures the irregularity of the lesion border and feature C measures the color variation in the lesion area [Kraus and Muehlenbein, 2024].

3.1.1 Feature A

For feature A, we use an algorithm that determines the asymmetry of the lesion area in the binary mask of the lesion by following this structure:

```
With binary mask of image:
    Find middle row;
    Find  middle column;
    Split mask by middle row;
    Split mask by middle column;
    Compare halfs;
    Calculate total asymetry;
```

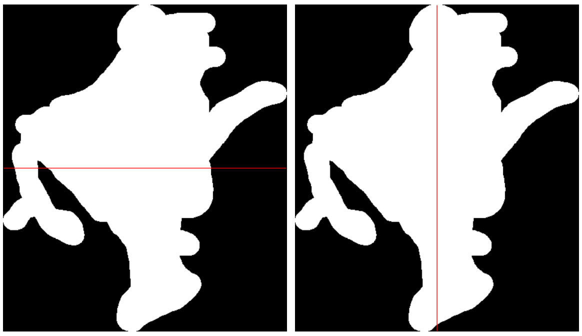

Figure 2: Mask split horizontally and vertically

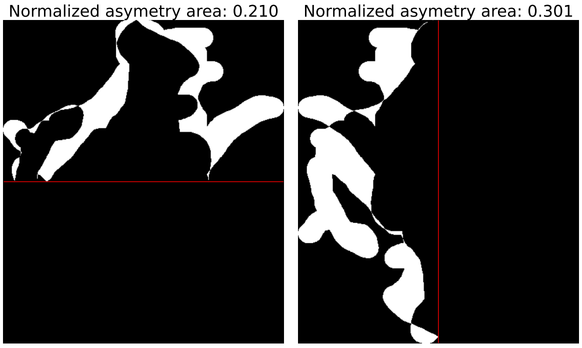
Normalized asymetry area: 0.210
Normalized asymetry area: 0.301

Figure 3: XOR area of halves

The final total asymmetry value is calculated as:

$$assymetry = \frac{A_{XORvert} + A_{XORhori}}{2 * A_{mask}}$$

where A is area, and ranges from 0, totally asymmetrical, to 1, perfectly symmetrical.

This method works well for a majority of cases; however, there are also some special circumstances where the method fails, for example, when there are multiple lesions in the image or when the lines of symmetry are not at the centers of the image.

3.1.2 Feature B

Feature B evaluates the compactness of the lesion to assess how close the shape of the lesion is to a perfect circle and recognize irregular borders. The algorithm is structured as such:

```
With binary mask:
    Compute area;
    Compute perimeter;
    Calculate compactness;
```

We obtain the area of the lesion by getting the sum of the binary mask, and the perimeter using the Crofton method. The compactness score is then calculated using the formula:

$$\text{compactness} = \frac{4\pi A}{l^2}$$

Where A is the lesion area and l is the lesion parameter which we estimate using the Crofton method. The compactness score is in the range [0,1], where lower scores indicate no compactness (irregular shape) and a higher score indicates a more circular shape.

Initially we calculated convexity score as the ratio of the lesion area to its convex hull area (the tightest boundary that encloses the lesion mask), and although it captured the irregularities of the lesion well, it was unreliable in smaller lesions. We replaced the convexity method with a compactness metric using Crofton perimeter, since it better estimated circularity and was more stable for differently sized lesions in our experiments.

3.1.3 Feature C

For the implementation of feature C, we use an algorithm that measures color diversity by clustering the pixels in the masked lesions and comparing distance between the unique colors. The model follows this structure:

```
With RGB image and mask:
    List all colors in lesion area;
    Cluster pixels using KMeans;
    Extract cluster centers ;
    Compute pairwise distances
    between cluster centers;
    Return max distance;
```

The method calculates the normalized color variation using:

$$\text{color\_variation} = \frac{\max_{i,j} \|c_i - c_j\|_2}{\sqrt{3}}$$

Where c_i and c_j are cluster centers in the range [0,1]. 0 indicates no color variation (only a single color is present in the masked lesion), while 1 indicates max color contrast (stark color differences).

We initially tested lesions for signs of blue veil only to identify melanoma. Blue veil refers to blue pigmentation in the lesion in an irregular area with a white overlay. The presence of blue veil heavily indicates melanoma, but not all melanomas show such signs. Processing the images through solely such test had led to unreliable results. Thus, we expanded the way we identified melanomas by color by analyzing the difference between colors groups in the lesion.

4 Extended Model

The extended model for melanoma detection builds upon the baseline model, adding more features to potentially increase accuracy.

4.1 Feature extraction

Our extended model incorporates Feature D to measure the lesion's diameter and Feature E to account for the lesion's growth (evolution). Furthermore, the model contains the Hair Feature to categorize the amount of hair on the lesion.

4.1.1 Feature D

Feature D extracts the maximum diameter of the lesion using the image's associated metadata. This allows the measurement of the lesion's size without the need of pixel calculations. The algorithm follows this structure:

```
Given image with metadata:
    Access ′diameter_1′ and
    ′diameter_2′ values;
    Return the larger of the two;
```

Where `image.metadata[′diameter_1′]` represents the horizontal diameter and `image.metadata[′diameter_2′]` represents the vertical diameter of the lesion. The function chooses the higher of the two:

$$D_{\text{max}} = \max(\text{diameter}_1, \text{diameter}_2)$$

Which provides a simple way of measuring the standardized maximum span of the lesion's size. However, one of the limitations of this method is that it relies solely on the images' external metadata, and therefore, it is dependent on the accuracy of the given measurements, which as we have noticed, contains some errors.

4.1.2 Feature E

Feature E determines whether a lesion has shown signs of growth over time by using the images' metadata (similarly to the implementation of feature D). The algorithm follows this structure:

```
Given image with metadata:
    Retrun grew from metadata;
```

This method does not involve analysis of the images and fully relies on external metadata and its accuracy, thus is subject to the same limitations as stated in 4.1.1.

4.1.3 Hair Feature

The hair feature (Mostame, 2023), determines the amount of hair on and around the lesion, relative to where it is placed on the body. It works along the lines of:

```
With Image:
    Isolate dark regions;
    Detect lines;
    Create mask from lines;
    Return ratio of non-zero
    pixels in mask over image
    size;

For all images:
    Compute hair ratio;
    Retrieve location of lesion
    from metadata;
    Compute mean and std of ratios
    for each location;
    Normalize ratios with
    region mean and std;
    Return normalized ratios;
```

One of the main problems we tackled when developing this feature, was that lesions in the dataset appear all around the body, in areas where the expected amount of hair varies, therefore it is unreliable to measure and compare hair around lesions from different parts of the body using the same scale *(since, for example, ears are expected to have less hair than arms).* In search of a solution that resolves this, we created a method that returns the relative amount of hair there is in / around a lesion, compared to other lesions in that same region of the body.

5 Significance of features

5.1 Features A-E

To determine whether our collected features are significant, we used the statistical t-test for comparison of means between the collected feature values for melanoma lesions and non-melanoma lesions. The results can be seen in *Table 1*:

	Mean			t-Test
	MEL	NO-MEL	Diff	p-value
feat_A	0.215	0.218	-0.002	0.939
feat_B	0.643	0.715	-0.072	0.081
feat_C	0.293	0.305	-0.011	0.384
feat_D	14.641	11.551	3.090	0.016
feat_E	0.923	0.610	0.313	0.000
feat_F	-0.172	0.030	-0.202	0.125

Table 1: Results from t-test for Mean Difference between Melanoma and Non-Melanoma lesions

As is very evident, not all of our features are statistically different between the Melanoma (MEL) and the Non-Melanoma (NO-MEL) groups. The most significant features are Feature D and Feature E taken directly from the dataset, and are impossible to extract from just images; however, in a real-world application, it would be very easy to collect.

As for our automated features, none have above the 95% significance to be classified as significant by convention, however their significance is still widely different among them and can serve as a guide for their impact on the classification. By this standard, Feature B is the most valuable, followed by Feature C, which might still have a positive impact on the accuracy of the model, with Feature A trailing far behind.

5.2 Hair feature

In addition to the t-test, see *Table 1:Feat_F*, in which we find that our hair feature does most likely have a positive contribution towards our model. We also examined the accuracy of the hair feature by comparing the results with our manual annotation for the amount of hair in a lesion picture from the mandatory exercise. This was done on a train-test data split basis to ensure that the accuracy value will translate to the images in the current dataset.

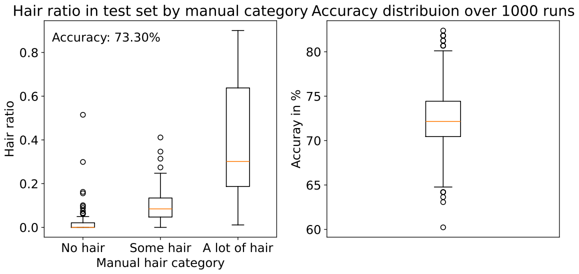

Figure 4: Comparison of automated hair extraction vs manual annotations

As visible in *Figure 4* the accuracy of our hair feature is around $72.5 \pm 7.5\%$ depending on the data split.

6 Classification and evaluation

6.1 Classifiers and Performance

There are many available classifiers to choose from when developing such models, each with their own advantages and limitations. In order to ensure that we get the full picture of the classifier performance, we wanted to have a way to compare them each time we implemented new features and extraction methods.

We developed an evaluator that automates this process. The metric we aimed to maximize was the recall of true melanoma cases, since in the context of dermatological screening, a high-recall model functions as a triage tool: it prioritizes identifying all potentially cancerous cases for further expert review, thus supporting early detection and intervention. The results are displayed in *Figure 5*.

We have achieved the best results using Logistic Regression (*Figure 6 and Figure 7*): to train and evaluate the classification model, the dataset was partitioned using an 80/20 split via the `train_test_split` function from the Scikit-learn library. The `stratify` parameter ensures that the class distribution remained consistent between both sets, which is crucial to avoid biased evaluation metrics in unbalanced datasets; we will elaborate on this subject later in the paper. The `random_state=42` parameter was set to ensure reproducibility of our results.

With an accuracy of approximately 63% and a recall of 88%, for the extended model, it correctly classified 7 of 8 melanoma lesions, but misclassified 76 non-melanoma lesions; which by our metric, recall, is substantially better than the baseline model, which only got a recall of 60%. For detailed performance differences between the models see *Figure 8*.

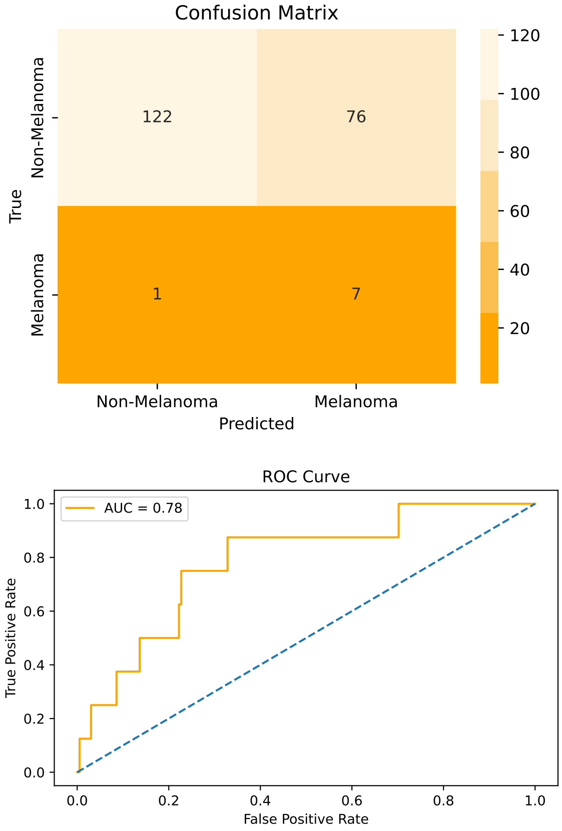

Figure 5: Performance of Logistic Regression using extended method.

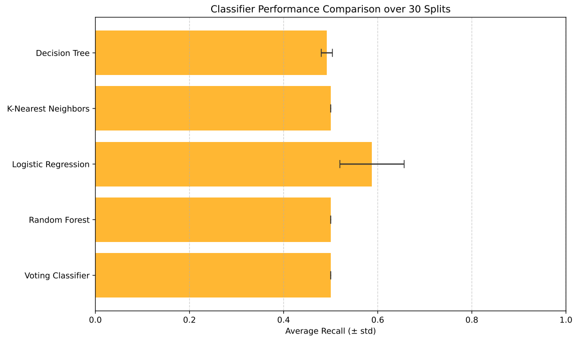

Figure 6: Baseline method (ABC).

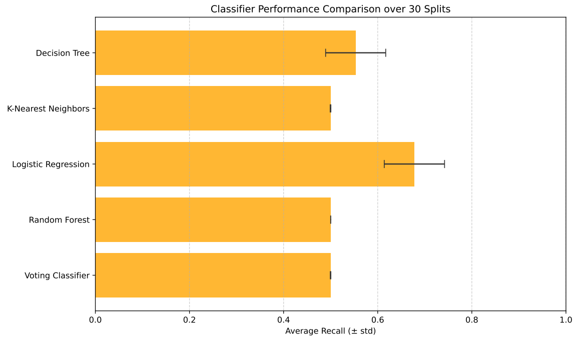

Figure 7: Extended method (ABCDE + Hair).

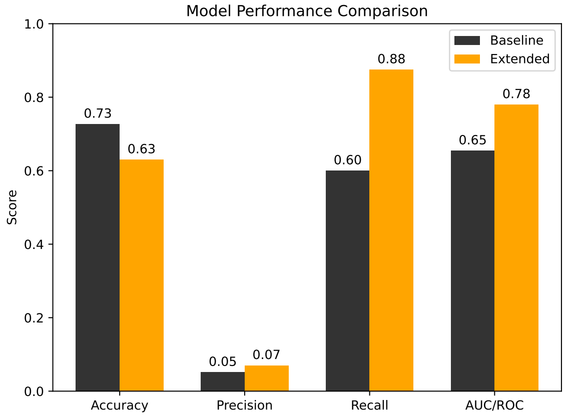

Figure 8: Performance of baseline and extended models (using Linear Regression).

Overall, the extended model is far more accurate and certain in its categorization; this can be seen in *Figure 9*, which shows the probability of the model choosing the label melanoma on the y-axis and UMAP projected feature space of the model on the x-axis. From the plots, we can see that the true melanoma cases are generally higher in the probability interval for the extended model, which means that it is more likely that if we add another melanoma image, the label of that image is likely to be predicted as melanoma. Furthermore, we see that the overall spread of the data points is much higher, allowing for easier separation of lesions into categories.

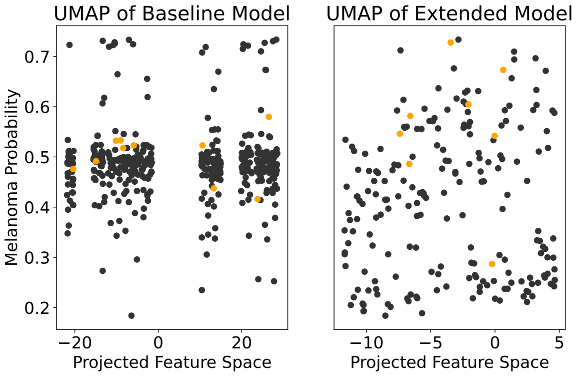

Figure 9: Probability of label melanoma compared to projected feature space.

6.2 Limitations

The implementations of the melanoma-determining features occasionally used the external data of the images (the metadata). For example, Feature D and Feature E in the extended model exclusively relied on on the values in the $'diameter'$ and $'grew'$ categories, respectively, to analyze the images. Although this makes features functionally easier (entirely removes the need to analyze image pixels), it means that the presence and accuracy of external data for the dataset are critical for appropriate results. After manual inspection, we have noticed discrepancies in the diameter values of the lesions in the metadata, as can be seen in *Figure 10*.

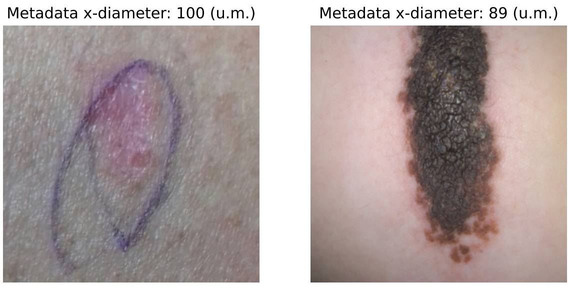
Metadata x-diameter: 100 (u.m.)
Metadata x-diameter: 89 (u.m.)

Figure 10: Diameter error in dataset.

7 Discussion and Results

Through this experiment, we learned that even with limited data and readily available tools from libraries like `sk-learn`, we can develop models that identify melanoma in lesions with enough precision to aid decision-making, however, far from enough to completely abolish the need for human input.

A revelation we came to during this experiment was that lesions appear in many different forms, visually vary in shape, size, color, symmetry, more than we could have anticipated. This extreme diversity set humble expectations to our capabilities to diagnose a particular type of lesion using the features.

However, in the end, our model did provide a solid basis which could be improved by reworking the way certain features are implemented. For example, pixel-based analysis for the evaluation of a lesion's diameter for Feature D, allowing for a more precise estimate and more certainty compared to a possibly flawed metadata.

7.1 Addressing class imbalance

During the initial development of our baseline classification model for skin lesions, we observed that the model predicted lesions exclusively as non-melanoma and, despite this, achieved deceptively high performance metrics (*Figure 11*). As we've mentioned before, it is due to the class imbalance within the dataset, where only approximately 2% of the lesions are labeled as melanoma.

Class imbalance is a common issue among datasets in several fields: Computer Vision (Sun et al., 2017), Machine Learning (Buda et al., 2018), particularly within the medical domain, where acquiring large and balanced datasets can be immensely expensive or practically infeasible. In our dataset, melanoma was severely underrepresented, and as a result, the model's sensitivity to such lesions was heavily reduced, since it was optimizing for the majority class to minimize loss; undermining its utility in clinical applications where early and accurate melanoma detection is critical.

We ask the question: How does class imbalance in the data affect the model's ability to detect melanoma? Moreover, can addressing the class imbalance lead to better performance without sacrificing overall accuracy?

To resolve this issue, we explored several balancing strategies: including the use of class-weighted loss functions and data resampling techniques such as oversampling of the minority class and undersampling of the majority class (SMOTE, ADASYN, standard random oversampling and undersampling).

The greatest increase in performance was after implementing weighted loss, which penalizes misclassifications of underrepresented classes more heavily, improving model performance on imbalanced datasets (*Figure 12*):

$$\mathcal{L}_{\text{weighted}} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}\left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right]$$

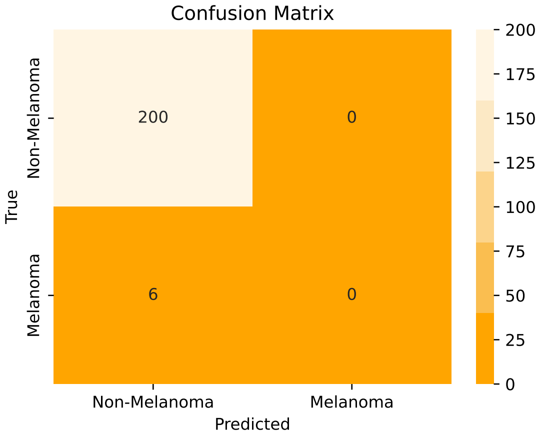

Figure 11: Confusion matrix before class balancing.

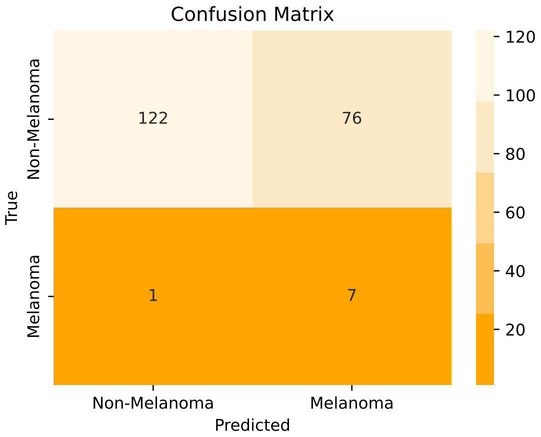

Figure 12: Confusion matrix with weighted loss implementation.

Luckily, a very efficient implementation of this is provided by the `sk_learn` Logistic Regression classifier by setting the `class_weight` parameter as `balanced`. Combining this with either: SMOTE, ADASYN and undersampling / oversampling gave us slightly better results, although statistically insignificant. Best results were achieved when oversampling the minority class (*Figure 13*), although almost identical to the results of exclusively using weighted loss:

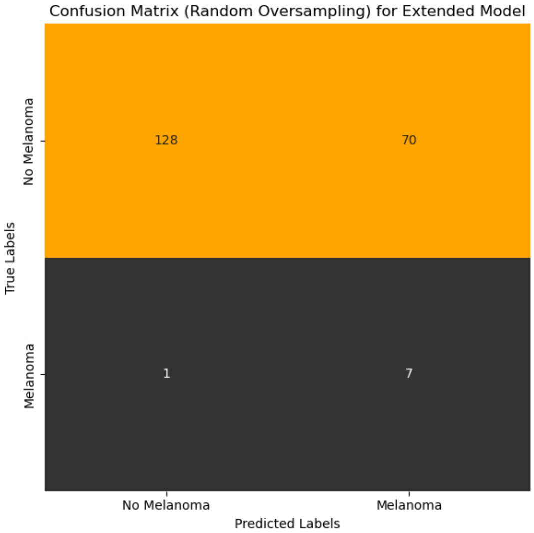

Figure 13: Confusion matrix using Weighted loss + Oversampling of the melanoma class.

Class balancing is a necessary step toward eliminating bias and ensuring that the model's diagnostic capabilities are distributed equally, this extends further than lesion diagnosis, it incorporates skin color, demographic representation, and other clinically relevant attributes to promote fairness and accuracy across diverse patient populations. In our opinion, future work should incorporate fairness metrics and subgroup analysis to further evaluate the impact of these interventions and to ensure that clinical AI systems serve all patients effectively and responsibly.

In conclusion, while a systematic evaluation of a lesion could be more accessible and easier to execute, it should not replace professional consultation and should only serve as a helping hand alongside the other existing techniques of diagnosis. However, once we overcome the current hurdles, highly developed methods of systematic lesion diagnosis hold great potential to benefit humanity by enabling early detection of skin conditions, improving access to care in underserved regions, and supporting clinicians with faster, more consistent assessments—ultimately contributing to better health outcomes and reduced global disease burden.

Acknowledgments

To the students before us, who made our lives easier by manually masking lesions, and preparing baseline feature extraction functions.

References

[Kraus and Muehlenbein 2024] S. Kraus and C. Muehlenbein. 2024. *Malignant Melanoma*. StatPearls [Internet]. StatPearls Publishing. `https://www.ncbi.nlm.nih.gov/books/NBK470409/`

[NIH2019] National Institutes of Health. 2019. *Some Melanomas May Start in Hair Follicles*. NIH Research Matters. `https://www.nih.gov/news-events/nih-research-matters/some-melanomas-may-start-hair-follicles`

[Mayo Clinic2023] Mayo Clinic Staff. 2023. *Melanoma: Symptoms and Causes*. Mayo Clinic. `https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884`

[NCI2025] National Cancer Institute. 2025. *Melanoma Treatment (PDQ®)–Patient Version*. U.S. Department of Health and Human Services. `https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq`

[American Cancer Society2023] American Cancer Society. 2023. *Skin Cancer Research Highlights*. American Cancer Society. Available at: `https://www.cancer.org/research/acs-research-highlights/skin-cancer-research-highlights.html`

[Mostame2023] Parham Mostame 2023. *Hair removal from skin images*, `https://www.kaggle.com/code/parhammostame/hair-removal-from-skin-images`.

[Cust et al.2015] Cust, A. E., Armstrong, B. K., Goumas, C., Jenkins, M. A., Schmid, H., Aitken, J. F., Hopper, J. L., Kefford, R. F., Giles, G.

G., Mann, G. J. 2015. *The risk of cutaneous melanoma in Australian families with a history of melanoma: A population-based study*. ResearchGate. Available at: https://www.researchgate.net/figure/Age-standardized-CMM-cases-per-100-000-people-by-age-of-Australian-US-European-and_fig1_311818237

[Wang et al.2024] Y. Wang, T. Yu, J. Cai, S. Kalia, H. Lui, Z. J. Wang, and T. K. Lee. 2024. *AI-Enhanced 7-Point Checklist for Melanoma Detection Using Clinical Knowledge Graphs and Data-Driven Quantification*. arXiv preprint arXiv:2407.16822. https://arxiv.org/abs/2407.16822

[Sun et al.2017] K. Sun, B. Xiao, D. Liu, and J. Wang. 2017. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1707.06642*. https://arxiv.org/abs/1707.06642.

[Buda et al.2018] M. Buda, A. Maki, and M. A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259. https://arxiv.org/abs/1710.05381.

[Pacheco et al.2020] Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C.R. Alves Jr, José G.M. Esgario, Alana C. Simora, Pedro B.C. Castro, Felipe B. Rodrigues, Patricia H.L. Frasson, Renato A. Krohling, Helder Knidel, Maria C.S. Santos, Rachel B. do Espírito Santo, Telma L.S.G. Macedo, Tania R.P. Canuto, Luíz F.S. de Barros, 2020, PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones, *Data in Brief*, Volume 32, 106221, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2020.106221