IT UNIVERSITY OF COPENHAGEN

Feature-Based Melanoma Classification

Samuel Joshual Martis
samj@itu.dk

Dimitra Filareti Tsairi
dimt@itu.dk

Ruben Coloma Garcia
ruco@itu.dk

Elena Linares Leal
elel@itu.dk

Project GitHub Repository:
https://github.com/SamuelJ-hub/2025-FYP-groupDolphins.git

Course: Projects in Data Science – Spring 2025
Course Code: BSPRDAS1KU

Course Manager: Veronika Cheplygina, Associate Professor
Teacher: Yucheng Lu, Postdoc

May 30, 2025

1 Introduction

Skin cancer is one of the most common cancers globally and also one of the most treatable when detected early. Dermatologists often rely on visual inspection and dermoscopy to evaluate suspicious skin lesions. However, access to dermatological care can be limited, particularly in low-income or remote settings. As mobile imaging devices become more accessible, there is growing interest in developing automated tools that support early skin cancer diagnosis. With rapid progress in machine learning and AI, researchers are achieving promising results in automated skin lesion classification.

Although a definitive diagnosis ultimately requires a biopsy, our aim is to explore whether visual features such as asymmetry, irregular borders, and atypical coloration — key components of the widely used ABC rule in dermatology — can provide a reliable indication to distinguish melanoma from non-melanoma skin lesions. Melanoma, the most aggressive form of skin cancer, often exhibits these distinct visual characteristics, making it particularly suitable for automated analysis based on these features.

In this project, we investigate whether such visual features, when extracted automatically, can be used to train a machine learning system capable of distinguishing melanoma from non-melanoma lesions. Our goal is to build a model that generalizes well to unseen images and could support low-cost, image-based pre-screening of melanoma in real-world settings.

As part of this project, we also explore an open question related to fairness and bias introduced by automated feature extraction. This is discussed in Section 7.

2 Dataset Analysis

This project is based on the PAD-UFES-20 dataset, a collection of skin lesion images developed by the Dermatological and Surgical Assistance Program at the Federal University of Espírito Santo. The dataset was created to support research in automated skin lesion analysis, especially for cases where dermoscopic imaging is not available. It contains over 2,200 clinical images of skin lesions, collected using smartphone cameras in real-world, low-resource settings. Unlike many datasets relying on high-resolution dermoscopic images, PAD-UFES-20 reflects the challenges of amateur image quality, variable lighting, and inconsistent backgrounds.

2.1 Dataset Description

The dataset consists of 2,298 images of skin lesions from 1,373 patients – meaning that some patients have multiple images of the same or different lesions. Each image is accompanied by a segmentation mask, which outlines the lesion area and a corresponding entry in a CSV file: `metadata.csv`. The metadata contains clinical information including patient age and sex, lesion location, Fitzpatrick skin type, lesion diameter, diagnosis.

Each lesion falls into one of six diagnostic categories:

Diagnosis	Abbreviation	Type
Basal Cell Carcinoma	BCC	Malignant
Squamous Cell Carcinoma	SCC	Malignant
Melanoma	MEL	Malignant
Actinic Keratosis	ACK	Benign
Seborrheic Keratosis	SEK	Benign
Nevus	NEV	Benign

Table 1: *Diagnostic categories in the PAD-UFES-20 dataset*

For this project, these diagnoses are grouped into two classes:

- **Melanoma (MEL)** — the primary malignant target class (47 samples)
- **Non-melanoma** — encompassing all other lesion types, both benign and non-melanoma malignant (2,056 samples)

In the dataset, all malignant cases (BCC, SCC and MEL) are biopsy-proven, while benign cases may be diagnosed clinically by a consensus of dermatologists. According to the metadata, approximately 58% of the samples are confirmed by biopsy, which adds credibility to the labels and supports reliable evaluation.

The segmentation masks play a key role in our analysis, as they should isolate the lesion from surrounding skin and visual artifacts such as hair or ink marks. This focused region makes feature extraction more accurate and consistent.

2.2 Preprocessing and Filtering

Before proceeding with feature extraction and classification, several data cleaning steps were necessary to ensure consistency and quality across the dataset:

- **Image and mask matching**: A data preparation pipeline was implemented. This script matched metadata entries with their corresponding image and segmentation masks.
- **Mask validation and filtering**: Samples with missing images or masks were removed, reducing the dataset from 2,298 to 2,103 complete and usable image-mask pairs.

While these steps successfully filtered out problematic samples, additional pre-processing techniques could have further improved the data quality. For example, automated removal of distracting features such as hair or ink marks would likely enhance the accuracy of visual feature extraction.

3 Comparison Between Manual and Automatic Hair Annotations

This section analyzes the consistency between manual hair annotations, conducted as part of a prior project and the automated hair feature extraction methodology developed for this project. This comparative assessment is crucial for evaluating the reliability of our automated hair detection system, a fundamental component for the extended classification methodology.

3.1 Manual Annotations as Ground Truth

The manual annotations, produced by multiple annotators in a previous project, are used as the ground truth. These annotations assign a hair presence level to each image, based on visual inspection. For this project, they provide a benchmark for evaluating the automatic method. To account for individual variability across annotators, ratings were averaged and rounded to the nearest discrete level (0, 1, or 2) to create a single reference label per image.

3.2 Automated Hair Detection Method

The software developed in this project processes each image to estimate the amount of hair coverage. This is done by highlighting thin dark patterns using morphological filters and quantifying their coverage over the total image area. The resulting percentage is then mapped to three levels of hair presence:

- **Level 0 (Absent)**: Hair coverage below a lower threshold (e.g., 3%).
- **Level 1 (Sparse)**: Coverage between the low and high thresholds.
- **Level 2 (Significant)**: Coverage above the upper threshold (e.g., 10%).

These thresholds were defined based on initial visual evaluation and empirical testing.

3.3 Distribution of Hair Annotations

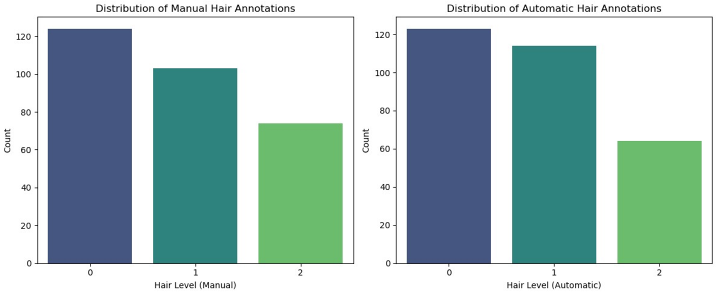

Figure 1: *Distribution of manual and automatic hair presence annotations*

3.4 Comparison Procedure

To evaluate the agreement between manual and automatic annotations, the software computes a confusion matrix and Cohen's Kappa score. The confusion matrix shows how often the automatic levels matched the manual ones. The Kappa score quantifies the agreement while accounting for chance. A higher score indicates better alignment between the two methods, with standard interpretation thresholds (e.g., > 0.6 = substantial agreement).

3.5 Results and Observations

The comparison between manual annotations and the automatic method was performed on 301 images. The agreement matrix shows that most inconsistencies occur between adjacent levels (for example, level 0 and level 1), which is reasonable given the subjective nature of visual hair evaluation.

Cohen's Kappa score was 0.32, which falls within the "fair agreement" range according to standard interpretation guidelines (0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial). This suggests that the automatic system captures hair presence to some extent, but there's still a noticeable gap compared to human annotations.

A higher agreement could likely be achieved by improving the hair detection method. In particular, the thresholds used to classify the amount of hair are currently chosen manually and based on visual inspection, which can introduce bias or inconsistency. Additionally, the criterion for detection — counting the number of pixels found by a black-hat filter — is sensitive to small changes in lighting and contrast. Adjusting the thresholds using data or more reliable pixel features could improve agreement with manual annotations.

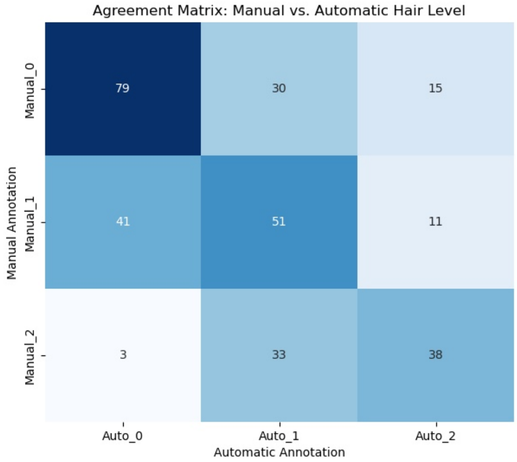

Figure 2: *Agreement matrix comparing manual and automatic hair level annotations*

4 Feature Extraction

Feature extraction is an important step in transforming raw image data into meaningful, quantifiable metrics that can be used for classification. In the context of skin lesion analysis, we aim to extract features that mimic what a dermatologist would visually assess when evaluating a suspicious lesion. Our feature set is based on the **ABC rule** used in dermatology — **A**symmetry, **B**order irregularity and **C**olor variation — as well as an analysis of **hair interference**, which can affect diagnostic accuracy.

4.1 Asymmetry

Asymmetry is a strong visual indicator of malignancy. Benign lesions are usually symmetrical, while malignant ones, like melanoma, tend to grow unevenly. To capture this, we computed a rotational asymmetry score. The lesion mask is first cropped and centered to isolate the relevant region. The cropped mask is then rotated iteratively by 30 degrees for a total of six rotations (covering 180 degrees). At each rotation, the lesion is compared to its horizontally flipped version using a logical XOR[1] operation to measure differences between the two sides. These differences are normalized by the lesion's total area to generate a score for each rotation. The final rotational asymmetry score is the average of the scores from all six rotations, with higher values indicating greater asymmetry. Asymmetry ranges from 0 (perfect symmetry) to 1 (high asymmetry).

4.2 Border Irregularity (Compactness)

Border irregularity is another visual indicator for melanoma detection. Benign lesions tend to have smooth, well-defined edges, while malignant lesions often display jagged or uneven borders. To capture this, we calculated a compactness score, which measures how closely the lesion's shape resembles a perfect circle. Using the lesion mask, we computed the area and estimated the perimeter by applying morphological erosion and subtracting the result from the original mask. The compactness is then computed using the formula:

$$\text{Compactness} = \frac{4\pi \times \text{area}}{\text{perimeter}^2}$$

A perfect circle yields a score close to 1, while irregular shapes result in lower values. We inverted this value to produce a compactness score where higher values indicate more irregular, potentially malignant borders.

4.3 Color Variation

Color variation is another useful indicator for detecting melanoma. Benign lesions often appear uniform in color, while malignant ones tend to show a mix of shades, such as darker browns, reds, or even bluish areas. To capture this, we extracted the pixel values within the lesion mask and calculated basic statistical measures. Specifically, we computed the mean and standard deviation of each color channel (Blue, Green, and Red). A higher standard deviation suggests more variation in color, which could be indicative of malignancy. These features provide a simple but effective representation of chromatic complexity within the lesion.

To ensure consistency and improve model training, these color features were standardized by sub-

[1] XOR: returns true only when the two inputs differ.

tracting the dataset mean and dividing by the standard deviation, transforming them to a common scale with a mean of 0 and standard deviation of 1. This process helps the model better interpret color variations across different lesions.

4.4 Hair Feature

Hair presence in lesion images can interfere with important visual features such as border clarity and color distribution, potentially reducing classification performance. Instead of attempting hair removal as a preprocessing step, we chose to quantify hair interference and use it as a feature in our model.

Our method detects hair-like structures using a black-hat morphological filter, which highlights dark, thin patterns against a lighter background. The process is as follows:

1. Convert the image to grayscale.
2. Apply black-hat morphological filtering with a 25×25 rectangular structuring element to enhance dark, hair-like features.
3. Threshold the filtered image to create a binary mask of potential hair regions.
4. Calculate the percentage of pixels affected by hair relative to the total image size.

This proportion, representing the hair coverage percentage, is then categorized into one of three discrete hair levels:

- Level 0: Low hair ($< 3\%$)
- Level 1: Moderate hair (3–10%)
- Level 2: High hair ($> 10\%$)

These two values — the coverage percentage and the hair level — are included as features in the classification pipeline.

In total, each image yields a set of 10 features: one asymmetry score, one border score, six color features (mean and standard deviation for B, G, R), and two hair interference values. These features serve as inputs to the classification models in the next section.

5 Classification and Evaluation

The goal of this section is to train and evaluate models for melanoma classification based on features extracted from dermoscopic images. Two Logistic Regression models were implemented:

- **Baseline Model**: trained using only standard ABC features.
- **Extended Model**: trained using ABC features plus an automatically derived hair coverage feature.

5.1 Baseline Model

The baseline model relied exclusively on the standard ABC features. After training, it was evaluated on a test set of 419 samples, including 10 melanomas.

Performance on Test Set:

- Accuracy: 85.9%
- Recall (Melanoma): 70.0%
- Precision (Melanoma): 11.1%
- ROC-AUC: 0.9254

The model achieved high recall[2], meaning it correctly identified most melanoma cases. However, precision[3] was low, reflecting a substantial number of false positives. The strong ROC-AUC[4] suggests the model has good class-separation capability, although a better decision threshold or additional features may be required for improved specificity.

5.2 Extended Model

The extended model incorporated the same ABC features as the baseline, augmented with a hair coverage feature automatically extracted from the images. This model was trained using a dedicated validation set for hyperparameter tuning, with recall used as the scoring metric to prioritize sensitivity to melanomas.

Performance on Validation Set:

- Accuracy: 63.9%
- Recall (Melanoma): 70.0%
- Precision (Melanoma): 4.5%
- ROC-AUC: 0.798

Performance on Test Set:

[2] Recall: proportion of actual positives correctly identified by the model.

[3] Precision: proportion of predicted positives that are true positives.

[4] ROC-AUC: overall measure of model performance based on true and false positive rates across thresholds.

- Accuracy: 66.8%
- Recall (Melanoma): 100%
- Precision (Melanoma): 6.7%
- ROC-AUC: 0.916

The extended model demonstrated perfect recall on the test set, successfully detecting all melanoma cases. However, the cost of this sensitivity was a low precision, with many non-melanomas misclassified as melanoma. This trade-off may be acceptable in clinical screening contexts, where minimizing false negatives is critical.

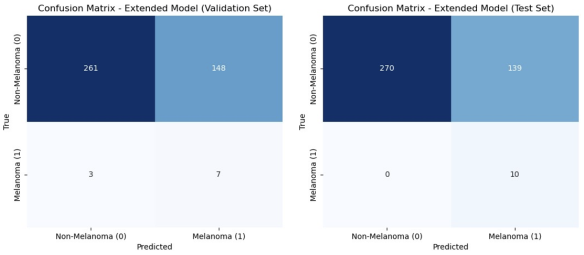

Figure 3: *Confusion matrices for the Extended Model on the validation and test sets*

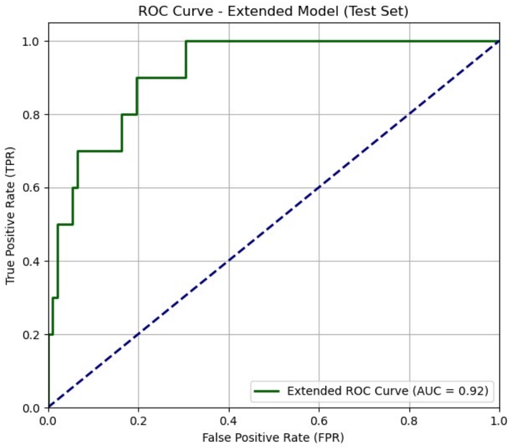

Figure 4: *ROC curve for the Extended Model on the test set*

6 Limitations

While this project aims to develop a model for automated skin lesion analysis, it is important to acknowledge several limitations inherent in the dataset, the chosen methodologies, and the scope of this work. Addressing these limitations in future research could lead to more generalizable and accurate diagnostic models.

6.1 Dataset Characteristics and Bias

Image Quality Variability: Images were collected using various smartphone devices, resulting in inconsistent lighting, color balance, and resolution. This variability poses a significant challenge for automated feature extraction methods, as they are highly sensitive to image quality. For instance, subtle color variations or border irregularities might be hidden by poor lighting or motion blur.

Dataset Imbalance: While the dataset includes six types of skin lesions, the distribution of cases—particularly the relatively low number of Melanoma samples (47) compared to Non-melanoma lesions (2,056)—can lead to class imbalance issues during model training. This imbalance can result in models that are biased towards the majority class and perform poorly on the minority (Melanoma) class, which is clinically the most critical to identify.

Segmentation Mask Quality: As noted in the project description, the provided segmentation masks, made by other students, can vary in quality. Some may over- or under-segment the lesion, especially in ambiguous cases. Inaccurate masks can lead to inaccurate feature extraction, as the features (like color statistics or compactness) are derived directly from the masked lesion area.

Hair Annotation Quality: The manual hair annotations used for evaluation were based on subjective judgment and may include inconsistencies. These limitations affect the reliability of both the training data and the performance evaluation.

6.2 Feature Extraction Methodology

Simplicity of Features: The extracted features (asymmetry, compactness, color variation, and hair interference) are relatively simple. They might not fully capture the complex visual characteristics that experienced dermatologists use for diagnosis.

Thresholding in Hair Detection: Our automatic hair detection method uses a black top-hat filter with fixed parameters (25×25 kernel, threshold value of 30). These settings were chosen empirically and may not generalize well across all images, particularly those with varying lighting conditions or hair textures. This limitation was reflected in the agreement analysis, where a moderate Cohen's Kappa score indicated inconsistencies between manual and automatic hair annotations.

Exclusion of “D” and “E” of ABCDE: This project focuses only on A, B, and C features. The “D” (Diameter) and “E” (Evolution) components of the ABCDE rule are not directly incorporated. Diameter is available in the metadata but was not used as a feature in the same way as visual features, and temporal “Evolution” data is not present in the dataset. This excludes important diagnostic information.

6.3 Model Scope

Absence of Clinical Context Integration: While patient metadata (age, location, Fitzpatrick type) is available, the primary classification method relies solely on extracted visual features. Incorporating clinical features (e.g., patient history, symptoms, lesion growth over time) could significantly enhance diagnostic accuracy, as dermatologists often consider a combination of visual and clinical information. This limits the real-world applicability of the model without clinical oversight.

Addressing these points could be the focus for future work to make better automatic systems for classifying skin lesions.

7 Fairness and Bias in Automated Feature Integration

We added an automatically computed hair-level feature to the extended classification model to improve its performance by accounting for image quality and occlusions. However, this brings up concerns about fairness, accuracy and possible bias.

7.1 Validity of the Hair Feature

As mentioned in Section 3.2-3.3, the match between the manual hair labels and the automatic hair-level estimates is only moderate. Using a feature that is not very accurate can cause the model to learn from random errors instead of useful information. This raises a question: is the model learning something real or is it reacting to noise in the data?

7.2 Dataset Imbalance and Feature Correlation

If images with more hair are more common in one class (e.g., non-melanoma), or if hair is linked to image quality, the model might wrongly connect high hair levels with a specific class. This could cause incorrect predictions. We did not control for this problem during training, so the model might be learning patterns that are not actually related to the lesion type.

7.3 Fairness in Real-World Scenarios

For the model to be useful in real medical situations, it needs to work well on all kinds of images. If the amount of hair in the image affects predictions, the model may not work as well for certain patients or in places where the images are not as clear. This could lead to unfair results and lower trust in the model.

7.4 Potential Improvements

To reduce bias and make the model more reliable, we could:

- Improve the hair detection method so it gives results closer to the manual labels.
- Use stratified sampling to balance the hair levels in the training data.
- Check model performance separately for images with different hair levels.
- Try fairness-aware training methods to avoid strong connections between predictions and hair level.

8 Conclusion and Discussion

This project explored how image-derived features can be used to classify skin lesions as melanoma or not, using Logistic Regression. We compared two models: one trained only on traditional ABC features, and another that also included a hair coverage feature automatically extracted from the images.

The baseline model demonstrated strong general performance. It showed a remarkable ability to differentiate between lesion types, achieving a ROC-AUC of 0.93. Furthermore, it successfully identified 70% of melanomas in our test set. While this is encouraging, in a critical medical setting, missing 3 out of 10 melanomas remains a significant concern.

Aiming for higher detection rates, our extended model was designed with increased sensitivity in mind. It achieved its goal, successfully detecting all melanomas (100% recall) in the test set. However, this came at a considerable cost: a high number of false positives. Its precision was only 6.7%, meaning that while no melanomas were missed, many benign lesions were incorrectly flagged as suspicious. This could lead to unnecessary patient anxiety or invasive procedures.

Our results also show that adding new features, even if automatically extracted, doesn't always lead to better overall performance. The hair coverage feature may have introduced bias or noise, so it's important to validate such features carefully before relying on them.

While our models are far from perfect, this project confirms the potential of image-based tools in aiding early melanoma detection. With more balanced datasets, richer features, and better tuning, such models could become useful assistants to doctors or even power mobile apps for early self-checks.

However, we must also be realistic about the limitations. Features extracted from images depend heavily on the quality and consistency of those images. Variations in lighting, zoom or focus can all affect results. It's essential to define clear standards for how images should be taken, especially if these tools are ever used in the hands of non-experts.

References

[Combalia et al.2019] Marc Combalia, Noel Codella, Veronica Rotemberg, Brian Helba, Verónica Vilaplana, Oron Reiter, Susana Puig, Allan Halpern, Harald Kittler, and Josep Malvehy. 2019. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

[Esteva et al.2017] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[González-Díaz et al.2023] Isaac González-Díaz, Gabriella Fabbrocini, and Luigi Cinque. 2023. Automatic hair detection in dermoscopy images: A survey and a new approach. *Journal of Biomedical Informatics*, 141:104365.

[Pacheco et al.2020] Andre G. C. Pacheco, Gabriel R. S. Lima, Luis F. S. Lima, Alexandre E. S. Salomao, Rodrigo B. de Oliveira, Wesley S. dos Santos, Fabiana M. Lima, and Renato A. Krohling. 2020. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221.

[Tschandl et al.2019] Philipp Tschandl, Noel Codella, Bengu Nisa Akay, Giuseppe Argenziano, Rainer P. Braun, Helena Cabo, David Gutman, Allan Halpern, Brian Helba, and Harald Kittler. 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7):938–947.