Projects in Data Science – BSPRDAS1KU

Asta Trier Wang – awan@itu.dk
Bruno Alessandro Damian Modica Figueira – brmo@itu.dk
Jan Peter Cardell – japc@itu.dk
Maja Kalina Oska – mkos@itu.dk
Philip Münster-Hansen – pmyn@itu.dk

github.com/Damianmdcf/2025-FYP-Final-groupA

May 30, 2025

Abstract

This project investigates whether melanoma can be reliably detected with Machine Learning using three of the most common features to detect melanoma (Asymmetry, Border irregularity, and Color variation). Working with the PAD-UFES-20 dataset, we implemented a baseline model and tested multiple classifiers, with logistic regression performing best. We further evaluated the effects of preprocessing hair removal, image-level data augmentation, and resampling techniques on model performance. While hair removal modestly improved performance metrics, the extracted ABC features alone were insufficient for a high-performance model. The model was sensitive to data imbalance, and image-level augmentations degraded performance. However, SMOTE-based oversampling led to slight gains in F1 and AUC scores. This project finds that preprocessing, especially hair removal, is essential to improve model performance, and further concludes that image quality should be prioritized over sample quantity.

1 Introduction

Skin cancer is one of the most common cancer types worldwide. In 2022 alone, there were more than 331,000 reported melanoma cases and more than 58,000 deaths worldwide, making it one of the most dangerous types of skin cancer.[1]

Fortunately, more than 95% of skin cancer cases are treatable if detected early[2], however, biopsying every suspicious lesion is impractical and resource-intensive. Since melanoma lesions are often characterized by asymmetry, irregular border, and variation of color in their architectural features,[3] many dermatologists rely on Asymmetry, Border, Color, Diameter, and Evolving (ABCDE) features to detect and prioritize melanoma cases.[4]

[1] Victoria State Government, *Melanoma*

[2] International Agency for Research on Cancer, *Melanoma of Skin Fact Sheet*, Global Cancer Observatory, 2022.

[3] Tsao et al., *Early detection of melanoma: Reviewing the ABCDEs*

[4] Puckett et al., *Melanoma Pathology*

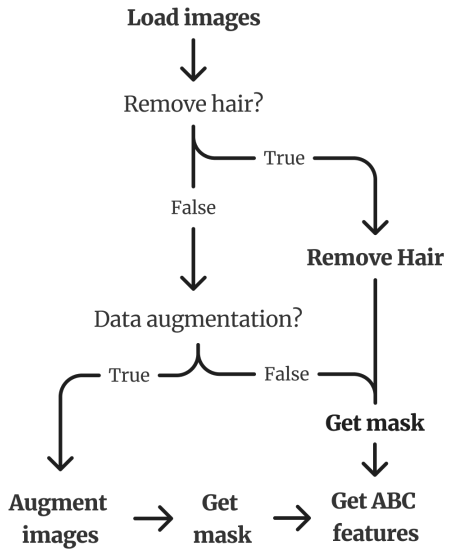

Figure 1: Flowchart of feature extraction

With the growing role of machine learning in medical diagnostics, it is realistic to consider a working algorithm based on the ABC-features to detect melanoma. A reliable model could reduce medical workload and increase the early-stage detection of cancerous lesions.

This report focuses on these ABC-features, aiming to asses whether they can reliably detect melanoma in images of lesions.

1.1 Research question

This study investigates the research question:

> ”Can melanoma be reliably detected using the ABC-features: Asymmetry, Border, and Color?”

Additionally, we explore the impact of data imbalance and investigate the open question:

> ”How does data augmentation affect machine learning models’ ability to detect melanoma given imbalanced datasets?”

2 Methods

This section outlines the overall methodology, including data preprocessing, lesion segmentation, feature extraction, model setup, data splitting, classifier training, and comparative analysis. Figure 1 visually displays the feature extraction process in a flowchart.

2.1 Lesion Segmentation

The images were fed into a mask generator, which applies a combination of morphological operations to produce a binary lesion mask. This mask isolates the lesion from the background, ensuring that future operations will focus only on the lesion.

2.2 Feature Extraction

For each image, the ABC-features are computed. These values and corresponding Z-scores are returned in a CSV file.

2.3 Data Splitting

The dataset was split 80/20 (train/test), producing two data files including all three features, to avoid data leakage.

2.4 Baseline, Extended, and Open-Question Dataset

To assess the performance of ABC-features, hair-removal, and data-augmentation on melanoma classification, we created different feature datasets for each model variant:

- Baseline dataset: Uses the features extracted from the original images and corresponding masks.
- Extended dataset: Applies a hair-removal step before masking, then extracts the same features.
- Open-Question datasets: Different datasets, either applying the oversampling method SMOTE, under-sampling the majority class, or applying image-level augmentation to actual melanoma images.

2.5 Classification and evaluation

For each of the datasets, four classifiers were trained (logistic regression, decision tree, random forest, and k-nearest neighbors) under different parameters. Using Stratified-K-folds with K=5, AUC, and F1-scores were estimated.

2.6 Comparative Analysis

To assess which classifiers performed best across the different datasets, we carried out a comparative analysis based on cross-validated performance metrics, including F1- and AUC-score. Having found the best classifiers for each of our different datasets, the test results were compared to conclude which performed best.

3 Dataset analysis

3.1 Data Source

The data used in this study, the PAD-UFES-20 dataset[5], comes from Dermatological and Surgical Assistance Program at the Federal University of Espírito Santo (UFES-Brazil) and includes 2,298 images of seven different types of skin lesions (Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen's disease (BOD), Melanoma (MEL), and Nevus (NEV)), where BCC, SCC and MEL are classified as cancerous.

3.2 Class Imbalance

The dataset consists of 52 melanoma cases and 2246 non-melanoma cases. This results in significant class imbalance, where melanoma (the minority class) represents only about 2.3% of the data. Such disproportion can be problematic for the machine learning process, as models can be biased toward the majority class during training. As a result, a model may achieve high overall performance while failing to correctly detect melanoma cases.

3.3 Duplicate Data

The dataset includes 2298 total images with 1641 unique skin lesions, meaning some lesions appear in more than one image. We chose to keep these duplicates, as the provided images vary in lighting, angle, and quality due to being captured by a smartphone. This natural variability reduces the risk of overfitting and helps the model generalize better to real-world conditions.

3.4 Data Cleaning

During preprocessing, we implemented a data cleaning step to exclude images of insufficient quality. Specifically, any image for which a segmentation mask could not be successfully extracted, resulting in an empty or fully black mask, was automatically discarded, unless a valid mask was available from a previous year's dataset. This approach ensures that only images with clearly defined lesion regions are included in feature extraction and model training. This decision helps prevent noisy or invalid data from negatively impacting model performance, particularly given the variable quality of smartphone-captured images in the dataset.

[5]Pacheco et al., *PAD-UFES-20 dataset*, Mendeley Data

4 Annotation of hair amount

The PAD-UFES-20 dataset presents a notable challenge: When hair overlaps a lesion, accurate feature extraction becomes significantly more difficult. To address this, each of the 5 group members manually annotated 200 randomly selected images from the dataset. Hair was scored on a scale from 0 to 2, where 0 means no visible hair, 1 means some hair, and 2 means a lot of hair.

Fleiss' Kappa is a statistical method used to measure inter-rater agreement between more than two annotators, while accounting for the agreement that could occur by chance. For all 5 human annotators, the Fleiss' Kappa score was 0.609. This score indicates very substantial agreement[6]

To establish a ground truth, we applied majority voting to determine the final hair score for each of the 200 images. A summary of the scores is shown below:

Hair score (0-2)	Image count
0	94
1	73
2	33

Table 1: Summary of human hair annotations

As seen in the sample, hair is a significant issue since 53% of the images contain at least some visible hair. To mitigate this, we developed a hair removal function.

4.1 Hair removal

The hair removal process works as follows:

- **Structuring element:** Choose a structuring element with a chosen kernel size and shape. The structuring element is used for morphological operations. For this project, we used a cross-shaped kernel.
- **Closing:** Apply the structuring element across the grayscale image. Perform dilation followed by erosion to produce a smoothed version of the image, where small dark hairs are removed.

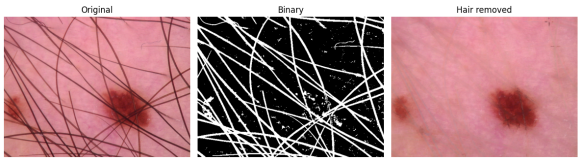

Figure 2: Example of hair removal

- **Blackhat:** Subtract the original image from the closing image to produce a blackhat image, highlighting only the dark hair.
- **Binary:** Apply a chosen threshold to convert the image to a binary mask.
- **Inpaint:** Use the binary mask to fill in the highlighted hair regions by interpolating from nearby pixels within a chosen radius. The result is an image where dark hairs have been smoothed away.

4.2 Optimizing parameters through annotations

The hair removal function includes three tunable parameters:

1. Kernel size
2. Threshold
3. Radius

The binary image produced during the process indicates the number of pixels flagged as hair. We rescale this count to a value of 0, 1, or 2 to match the scale of our human annotations. This allows a direct comparison to the ground truth.

To evaluate the performance of different parameter settings, we used Cohen's Kappa, which is a metric similar to Fleiss' Kappa, but designed for agreement between two annotators, in this case, the established ground truth and computer assessment. Through iterative experimentation, we found that the best-performing parameters were: kernel size = 5, threshold = 10, and radius = 3. With these settings, we achieved a Cohen's Kappa of 0.191. This score falls just at the border between what is considered slight and fair agreement[7]. This means that our hair removal function is not optimal, which we should be aware of when

[6] DATAtab, *Fleiss' Kappa*

[7] DATAtab, *Fleiss' Kappa*

evaluating the models. This is likely due to noise, such as shadows and non-dark hairs.

5 Feature extraction

For each of the ABC-features, we implemented an automated extraction method based on the image and/or the corresponding binary mask. Each image was then processed to obtain individual measurements of Asymmetry, Border, and Color. The following section describes the steps for computing masks and extracting features.

5.1 Mask

Due to multiple problems with the provided masks, we decided to make our own masks. This decision was based on the fact that both Asymmetry and Border rely only on the mask, so these two features would be the same for both the baseline and extended models. Further, not all images had provided masks, which would have forced us to drop these images, losing valuable data. This process is shown in Figure 3 and described below:

- Converts the input image to grayscale.
- The algorithm chooses a starting point (seed) for region growing based on the following scenarios:
  - If a prior binary mask was provided from the previous year's students (annotated mask), it chooses the middle point of the given mask as the seed.
  - Otherwise, it chooses the darkest pixel in the grayscale image as a seed.
- Region growing:
  - Without an annotated mask, the region growing algorithm will accept any neighbor whose absolute intensity difference from the seed is at most a set threshold.
  - With an annotated mask, the same procedure is applied, but only within the bounds of the annotated mask.
- This process is repeated until the mask covers all connected pixels within the intensity threshold.
- We iteratively tested different thresholds and selected the one that achieved the best performance on a sample of images evaluated by visual inspection.

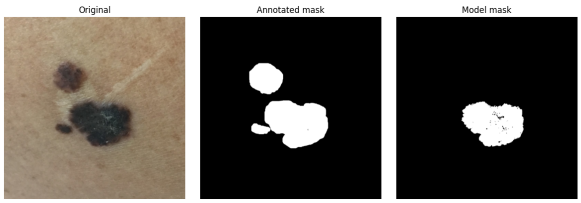

Figure 3: Example of masking

- The region growing produces a mask; if it is unable to produce one, it falls back to return the annotated mask. If none are available, we drop the image.

5.2 Asymmetry

- The binary mask is cropped to the lesion reducing the background.
- The cropped mask is then rotated 6 times, 30 degrees at a time, to cover 180 degrees.
- At each rotation, the cropped mask is flipped horizontally and vertically, and the non-overlapping pixels are counted (XOR) and used to compute asymmetry for that rotation:

$$A = \frac{\text{Non-overlapping pixels}}{\text{Total lesion pixels}}$$

- Finally, the asymmetry scores for all 6 rotations are averaged to ensure a more robust measurement of asymmetry.

5.3 Border

- The OpenCV function finds the contour by scanning the entire binary mask to find the curves that connect all continuous points along a boundary that has the same pixel intensity.
- Compute area in contour, A.
- Compute length of perimeter of contour, P.
- Compactness formula is applied:

$$\text{B} = \frac{P^2}{4\pi A}$$

5.4 Color

- The binary mask is applied to the image to ensure only measuring irregularity on the lesion.
- The image is divided into 200 superpixels using SLIC segmentation.
- The RGB image is converted to the HSV (Hue, Saturation, Value) color space to allow for more meaningful analysis of hue and saturation, which are important in detecting color irregularity. Unlike RGB, HSV separates color information from intensity (value), making it more suitable for identifying subtle color variations in skin lesions, since it is less sensitive to lighting differences.
- The mean of hue and saturation of each superpixel is computed.
- The hue and saturation variance of means across superpixels is returned, and the formula for irregularity score is applied to give hue and saturation equal weight. We omit the value component of HSV, since we don't want the lighting in the picture to affect the measurement of color irregularity.

$$C = 0.5 \cdot \text{Var}(\text{Hue}) + 0.5 \cdot \text{Var}(\text{Saturation})$$

5.5 Visualizations of the features

Figure 4 represents a correlation matrix with density plots along the diagonal. It shows the features calculated on the baseline model, plotted against each other to visualize how the melanoma and non-melanoma classes differ across them.

The density plot for the color feature shows that non-melanoma cases tend to have more consistent values, while melanoma displays higher variation. However, the plot also reveals a potential issue with this feature: most lesions have very low scores, clustered near zero. This may be related to how the lesion masks were computed. Since the mask function relies on differences in pixel intensity in a grayscale image, it may exclude regions with color irregularity (see Figure 3). As a result, parts of the lesion that contribute to color irregularity could be omitted, leading to artificially low feature values.

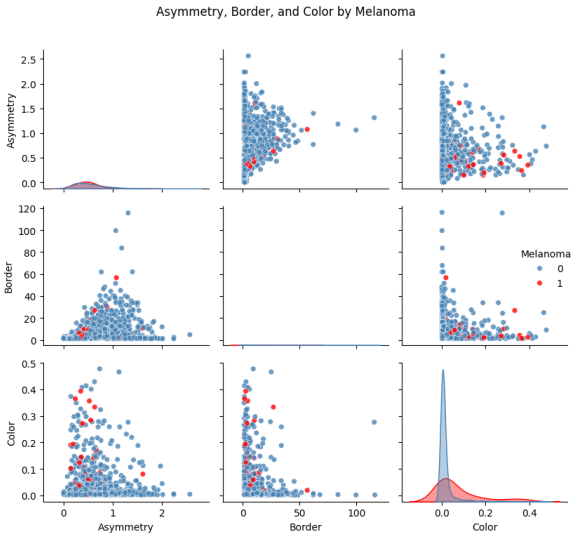

Figure 4: Feature correlation matrix of baseline

Additionally, the features appear on different scales, which will present a problem for model training. Therefore, some scaling is necessary before deriving conclusions.

5.6 Feature scaling

To address the features appearing on different scales, we standardized all feature values using Z-scores. These standardized values were used during classifier training to ensure comparable feature scales.

With the comparable Z-scores we can see in Figure 5 that both the border- and asymmetry feature have very similar distributions for melanoma and non-melanoma, which might introduce problems for correctly classifying. On the other hand, the color feature displays a difference in distribution between the two classes, making it a more useful feature for model training.

6 Classification and evaluation

To achieve the best possible classification result, we tested and compared the following classifiers:

K-Nearest Neighbors (KNN): A simple, untrained classifier that predicts the label of a data point based on the majority label of its K closest neighbors in the training data. Since KNN uses euclidean distance to find neighbors, the feature scales should be standardized. It is important to choose the right value for K: too small a K could lead to overfitting to noise, while a too large K

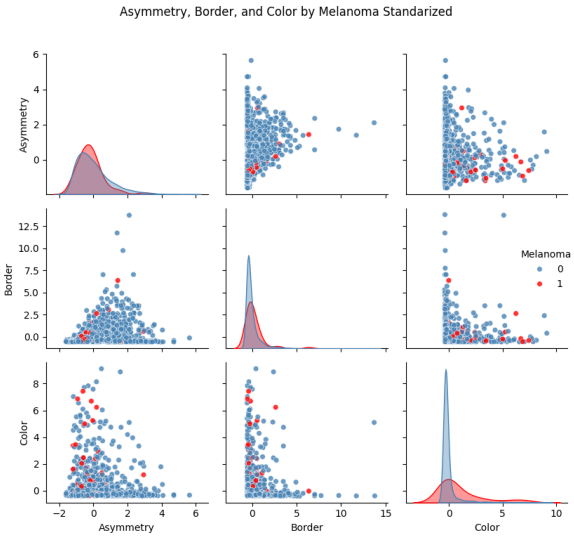

Figure 5: Feature correlation matrix of standardized baseline

can lead to bias towards the majority class. This is especially problematic in our imbalanced dataset with relatively few melanoma cases. To reduce the chance of ties in the binary classification, we chose to test odd values for K.

Logistic Regression: A linear model that estimates the probability of the possible outcomes, like a melanoma or non-melanoma lesion, by using a sigmoid function to produce a value between 0 and 1. The model learns weights for each feature from the training data and works best when the classes are linearly separable. The classification threshold can be adjusted to improve the performance on imbalanced datasets. In our case, we tested lower thresholds to catch more positive cases due to the imbalance in melanoma labels and to avoid false negatives.

Decision Tree: A model that splits the data based on feature thresholds, forming a tree-like structure where each leaf represents a predicted label. A key parameter is the tree's maximum depth: a very deep tree can overfit to noise in the training data, while a shallow tree may introduce bias.

Random Forest: Uses the same basic idea as a decision tree, but builds multiple trees instead of just one. Each tree is built slightly differently, and their predictions are combined through majority voting to classify new data points. This approach reduces the risk of overfitting compared to a single decision tree. A key parameter is the number of trees: more trees can improve the performance, but at the same time it increases the computational cost.

Measuring performance

In machine learning, various metrics are used to evaluate model performance. A basic metric that is commonly used is accuracy, which is calculated by dividing the number of correctly classified items by the total number of items:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

In the case of a highly imbalanced dataset, accuracy becomes misleading. For instance, if the model predicts all samples as the majority class (non-melanoma), it can still achieve a very high accuracy, even if it fails to detect a single melanoma case.

In the context of medical imaging, especially for cancer detection, minimizing false negatives is crucial. A missed melanoma case could be life-threatening, so we emphasize the metric **recall**, which quantifies the ability to detect positive cases:

$$\text{Recall} = \frac{TP}{TP + FN}$$

However, high recall may come at the cost of increased false positives. Since each false positive may lead to unnecessary and costly biopsies, and can concern the patient, we also track **precision**, which measures how many predicted positives are actually correct:

$$\text{Precision} = \frac{TP}{TP + FP}$$

To balance these two aspects, we use the **F1-score**, the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

To evaluate model performance across all possible thresholds, we also use the Receiver Operating Characteristic Area Under the Curve

(ROC AUC). The ROC curve plots the *true positive rate* (recall) against the *false positive rate* for various thresholds. The AUC value summarizes this curve into a single number between 0 and 1, representing the model's ability to distinguish between classes across all thresholds.

Based on these considerations, this project mainly focuses on F1 and AUC to evaluate model performance.

Train/test split: To avoid overfitting our classifiers to the training data and to get a clear idea of how the model performs on unseen data, we started by splitting our dataset into training and test sets. We used a common 80/20 split to have enough data for training while still keeping a portion for reliable evaluation. We used sklearn's `train_test_split` function and, because of the low frequency of melanoma cases, we used stratification based on the label. We created a separate train/test split for each model setup (baseline and extended).

Stratified K-fold method: During the training process, we used the stratified K-fold method, a resampling technique particularly useful for imbalanced datasets. Unlike standard K-fold, stratified K-fold ensures that each fold maintains the same class distribution as the full dataset, which is particularly important given the underrepresented melanoma class. By calculating the evaluation metrics across all folds and averaging them, we obtain a more stable estimate of the model's performance. [8]

6.1 Baseline model

To determine the most effective classifier, we tested different parameter setups for the four classifiers. We used 5-fold cross-validation to evaluate each model, focusing mainly on AUC and F1.

Figure 6 shows box-plots of the AUC and F1 scores across all five folds to visualize which combination of classifier and parameter performs best.

Based on these results, logistic regression with a threshold of 0.03 was selected, since it had the best trade-off between AUC and F1.

After selecting the model, we trained it on the full training set and evaluated it on the held-out test set. Performance was assessed using multi-

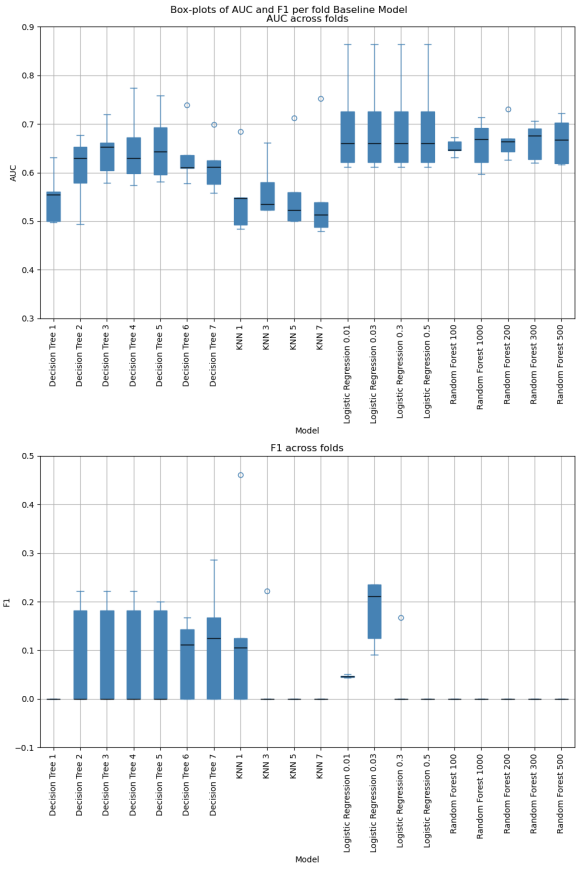

Figure 6: Box-plots of AUC and F1 across folds and classifiers

[8] scikit-learn developers, *StratifiedKFold*

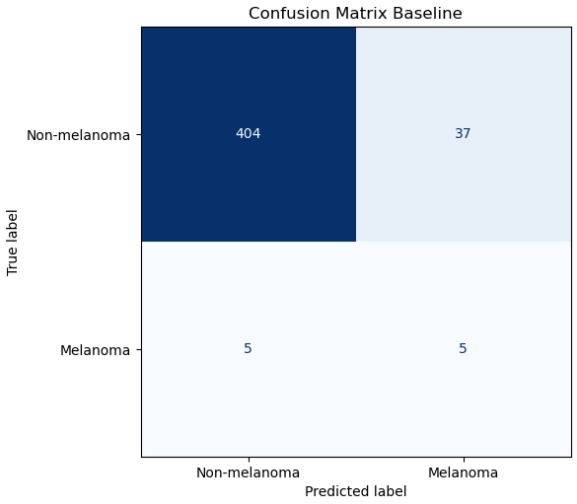

Figure 7: Confusion matrix Baseline model

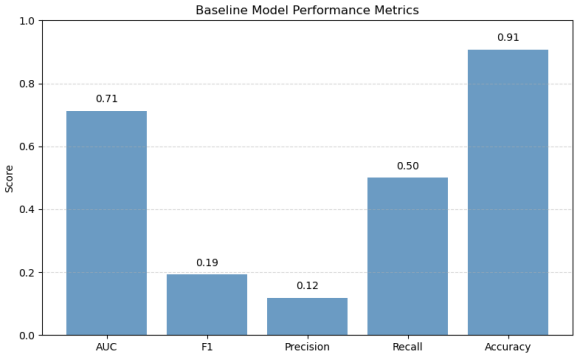

Figure 8: Performance metrics: Baseline model

ple metrics: Precision, Recall, F1-score, AUC, and the confusion matrix components (True Positives, False Positives, False Negatives, True Negatives).

Figure 7 shows the confusion matrix for the baseline model. The model correctly detected five true melanoma cases (True Positives) and missed five cases (False Negatives). It also produced 37 False Positives, misclassifying non-melanoma lesions as melanoma.

Figure 8 displays all the performance metrics of the baseline model. With a rather high AUC of 0.71, our model overall does a fairly good job of correctly classifying. However, the F1 of 0.19 and the Precision of 0.12 indicate that the model falsely classified many lesions as melanoma. The Recall of 0.5 means the model correctly identified half of the total melanoma cases.

6.2 Extended model

In order to assess whether masking out the hair in the images increases model performance, we trained an extended model on the images with hair removed. The same process as for the baseline model was followed. After training the classifiers, logistic regression with a threshold of 0.03 turned out to be the best model again. Figure 9 shows the box-plots, which were used in the process.

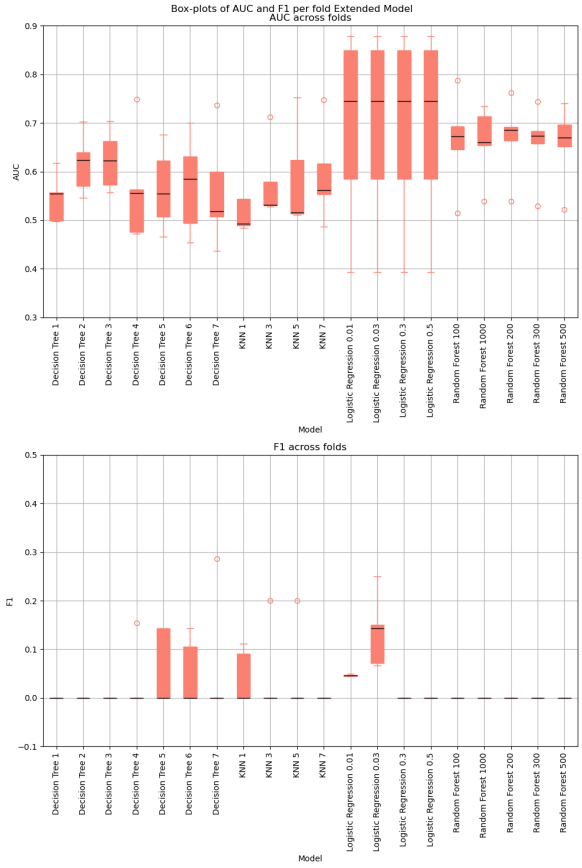

Figure 9: Box-plots of AUC and F1 across folds and classifiers

The results after running the extended model on the held-out test data are displayed in Figures 10 and 11. The confusion matrix (Figure 10) reveals that the extended model correctly classified one more melanoma case than the baseline model (5 vs. 6), and misclassified fewer cases as melanoma than the baseline model (34 vs. 37). This suggests that the extended model overall performs slightly better.

This improvement is further reflected in the performance metrics shown in Figure 11. The extended model achieves higher scores in AUC (0.77 vs. 0.71), F1-score (0.24 vs. 0.19), Precision (0.15

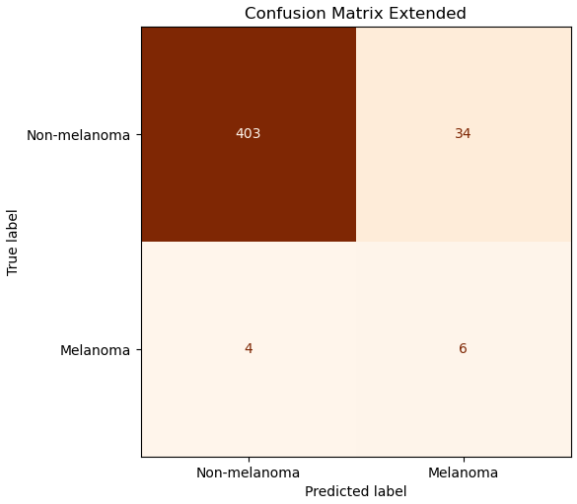

Figure 10: Confusion matrix Extended model

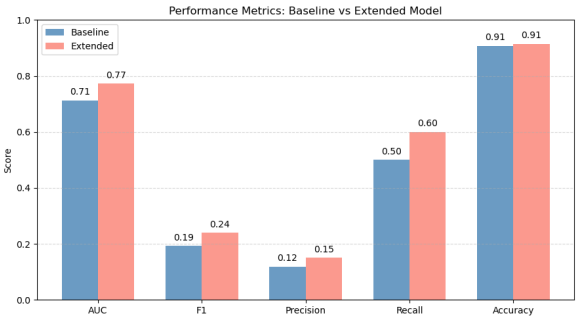

Figure 11: Performance metrics: Baseline and Extended model

vs. 0.12), and Recall (0.60 vs. 0.50). These gains indicate that the extended model is more sensitive to melanoma cases, making it more reliable for clinical use. However, it is worth mentioning that since we are dealing with such a low number of actual melanoma cases in the test set, it only takes a few more true positives to influence these measurements than it might have if we had a more balanced dataset, so the results might not be significantly different.

7 Open question

While our baseline model evaluated the feasibility of classifying melanoma using the ABC features (asymmetry, border, and color), it was trained on a highly imbalanced dataset. As previously discussed, the dataset contains only 52 melanoma cases out of 2,298 samples, which introduces a substantial risk of overfitting toward the majority class. This motivated our open research question:

> ”How does data augmentation affect machine learning models’ ability to detect melanoma given imbalanced datasets?”

To address this, we extended our original pipeline to incorporate several augmentation strategies aimed at improving class balance and feature strength. We decided to include synthetic sample generation using the following methods:

7.1 Synthetic Minority Oversampling TEchnique (SMOTE)

Synthetic Minority Oversampling TEchnique (SMOTE) over-samples the minority class of the dataset by creating synthetic samples. This is done by selecting a minority class data point at random and finding its K nearest minority class neighbors. Then, one of the neighbors is chosen at random, and a new synthetic instance is generated by averaging the two points. As SMOTE generates synthetic data points by interpolating between nearest neighbors in feature space, it is sensitive to the scale of features. Standardizing first ensures that all features contribute equally. We used imblearn’s `SMOTE` function, with a sampling strategy of 0.3, to generate synthetic samples until the minority class size reaches 30% of the majority class size.

7.2 Under-sampling

Another technique used to address class imbalance is under-sampling, also available as one of the imblearn’s `RandomUnderSampler` functions, which randomly selects a subset of the majority class and discards the rest. As before, the sampling strategy can be chosen, using sampling strategy = 0.5, the function decreases the size of the majority class to twice the size of the minority class. It successfully balances the classes in the dataset, ensuring that the model will not be biased towards the majority class. However, we risk losing information due to discarding a subset of the dataset.

7.3 Image-level augmentation

Another approach to tackle class-imbalanced datasets is to create more samples by modifying the original melanoma images. We decided to apply three different image-level augmentation methods to each melanoma image in our training

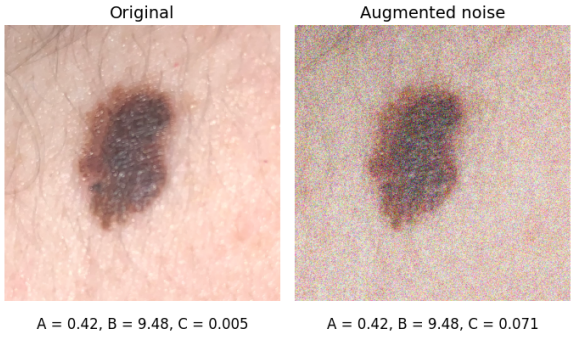

Figure 12: Example of noise added to image. Notice the change in the color feature value.

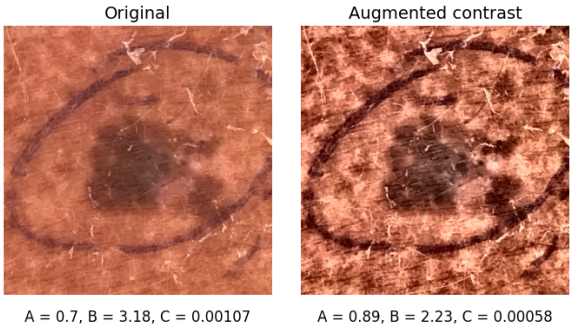

Figure 13: Example of CLAHE applied to image. Notice the change in asymmetry and border feature values.

data: Gaussian noise, CLAHE, and distortion of the border.

7.3.1 Gaussian noise

Gaussian noise is a type of distortion added to an image by slightly changing the brightness of pixels, where the size of the changes follows a Gaussian distribution. We used it to simulate real-world imperfections, like those caused by smartphone cameras. This ensures that the model does not overfit to certain image imperfections, but rather focuses on meaningful structural features of the lesion itself, making the model generalize better. [9] This is visualized in Figure 12.

7.3.2 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Contrast Limited Adaptive Histogram Equalization (CLAHE) is designed to enhance local contrast by equalizing small sections of the image independently. This helps reveal details in both darker and lighter regions. It is beneficial to all three features: It ensures that asymmetry calculations reflect actual shape differences rather than illumination artifacts, it also sharpens the edges of the lesion, which improves border extraction, and as CLAHE enhances contrast without distorting the natural color distribution, which helps us capture subtle differences in pigmentation. [10] This is visualized in Figure 13.

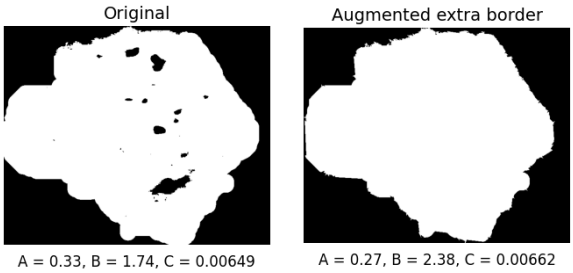

Figure 14: Example of extra border augmentation. Notice change in A and B features.

7.3.3 Distortion of border

Since melanoma lesions are characterized by asymmetry and irregular borders,[11] we applied an augmentation technique that randomly perturbs the lesion's boundary to increase the diversity of training samples. This technique introduces slight distortion to the lesion shape, which adds variability to the features A (Asymmetry) and B (Border), making the model better generalize to unseen and naturally irregular cases. Since melanoma lesions are often asymmetric with jagged borders, this is a way to create an artificial sample that represents features that are expected in melanoma lesions.

7.4 Method and applications

In order to successfully answer the open question, we tested the individual performance of the above-mentioned data augmentation types. We decided to also include a combination of the two

[9] Anonymous, *Data Augmentation in Training CNNs: Injecting Noise to Images*, ICLR 2020

[10] Li et al., *Deep Learning-Optimized CLAHE for Contrast and Color Enhancement in Suzhou Garden Images*

[11] *Anonymous. (2020). ABCDE's of melanoma: Asymmetry, Border, Color, Diameter, Evolving.*

sampling methods, since it allows us to create a balanced dataset with a trade-off between not having too many synthetic samples, and also not having to discard too many real samples.

A fourth dataset was created by combining the training baseline dataset with "new" augmented melanoma images (image-level augmentation). We applied noise, CLAHE and distortion of the border to all images, and then extracted new synthetic features. When creating melanoma samples using the Gaussian noise method, we used the mask from the original image, as region growing doesn't work on noisy images. We also noticed that Gaussian noise inflated the color feature. To control this, we applied Z-scores when training the classifiers. By applying these three image-level augmentation methods to the minority class, we increased its size from 42 to 168 cases.

Furthermore, we wanted to test a combination of all data augmentation methods, as previous studies have shown that applying multiple data augmentation techniques can lead to improved results.[12] Last dataset was created by first applying image-level augmentation, then SMOTE and finally under-sampling the majority class. This left with the following datasets to train and test:

- Oversampling (SMOTE) by 30 percent
- Undersampling the majority class to 50 percent
- SMOTE and Undersampling
- Image-level augmentation
- A combination of the above

7.4.1 Avoiding data leakage

When training, using traditional stratified K-folds, on the augmented data two main issues are introduced:

1. Augmented images should all be grouped together with the original image(s) that they originated from to avoid data leakage across folds.

[12] *Perez et al. (2018). Data Augmentation for Skin Lesion Analysis.*

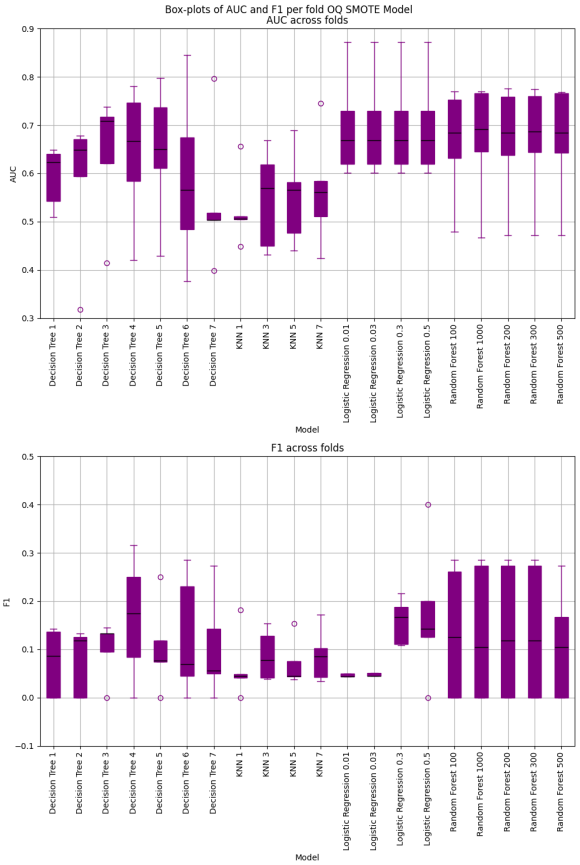

Figure 15: Box-plots of AUC and F1 across folds and classifiers

2. Synthetic data should not be present in the validation fold.

To address these issues, two solutions were applied. First, when training on the image-level augmented data, we used stratified grouped K-fold cross-validation to ensure that all augmented versions of the same image remained within the same fold. This reduced the risk of overfitting, although it may have resulted in a smaller test set. Then we removed synthetic data from the validation fold. Secondly, for the SMOTE training, we applied SMOTE isolated in each of the folds, and of course, not on the validation fold.

7.4.2 Results

To determine the most effective model, we followed the same process as in the baseline and extended model. However, this time not only did we have to compare the performance of different classifiers, but also check the performance of best classifiers across each of the augmented datasets. Figure 15 shows the box plots showing AUC and

F1 scores for the SMOTE dataset, similar figures were used to choose the best classifier for the rest of the augmented datasets.

After choosing the best classifier and parameter for each augmentation dataset on training data. All our open question models performed best with logistic regression, but at different thresholds. Their performance and threshold is summarized in Table 2.

Method	Threshold	Mean AUC	F1 Score
SMOTE	0.3	0.70	0.16
Under	0.5	0.72	0.12
SMOTE + Under	0.5	0.70	0.17
Image Aug.	0.03	0.72	0.04
All combined	0.5	0.62	0.12

Table 2: Performance of data augmentation methods on logistic regression on training data

7.5 Finding and Evaluating the best Data Augmentation

After having found the best classifier for each data augmentation dataset, we tested each of them on the held-out data to compare performance. Taking into consideration the best trade-off between AUC and F1, the SMOTE dataset with a logistic regression model and a threshold of 0.3, performed best. The comparison of performance across data-augmented models can be seen in Figure 16. The SMOTE model achieves a higher score than the baseline model AUC (0.72 vs 0.71), F1 score (0.22 vs 0.19), Precision (0.14 vs 0.12), and Recall (0.6 vs 0.5). The confusion matrix of the test results can be seen in Figure 17. This shows that we successfully improved the machine learning model's ability to detect melanoma through synthetic oversampling. As mentioned earlier, since we are dealing with a highly imbalanced dataset with few melanoma cases, we have to be aware that it takes very few true positives to inflate Recall and Precision.

Under-sampling removed useful non-melanoma data and limited the training set, resulting in moderate (Precision of 0.15 and F1 of 0.22) but less stable performance than SMOTE. Image-level augmentation, despite increasing sample diversity, led to poor performance. Models trained on this data tended to overfit and classified nearly all cases as melanoma, resulting in extremely

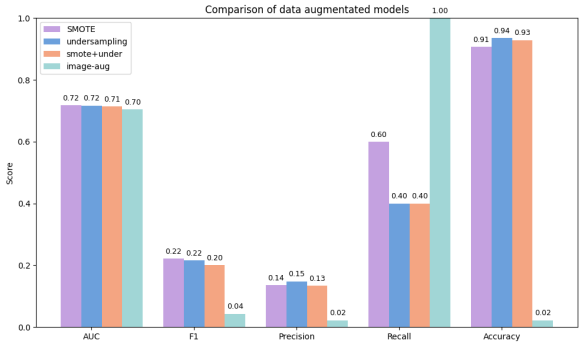

Figure 16: Performance metrics across data-augmented models

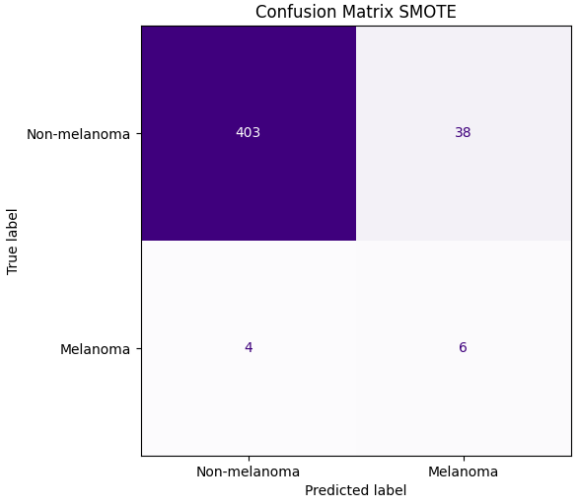

Figure 17: Confusion matrix OQ SMOTE model

low precision (0.02) and F1 scores (0.04). This suggests the augmented images significantly distorted the distribution of features, making them unrealistic or misleading for training.

Even though the SMOTE-augmented model outperformed the baseline by addressing class imbalance, it still performed slightly worse than the extended model, as Figure 18 shows. This comparison highlights the importance of feature strength over sample quantity. High-quality and well-preprocessed data enables more accurate and generalizable learning than artificially enlarged, but potentially noisy datasets.

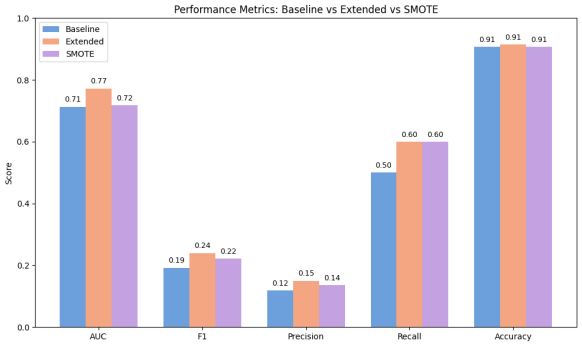

Figure 18: Performance metrics: Baseline, Extended and SMOTE model

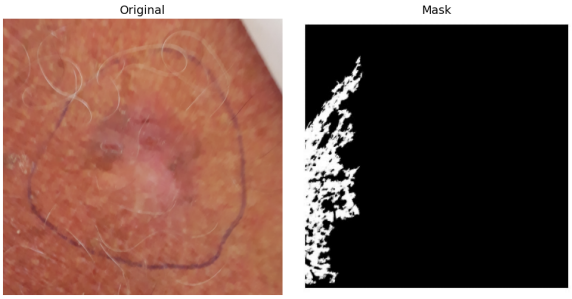

Figure 19: Example of inaccurate mask

8 Discussion and limitations

The main goal of this project was to evaluate whether a machine learning model can reliably detect melanoma based on ABC-features. We initially expected that the extended model with hair removal and the models with data augmentation would perform better than the baseline, due to the presence of visual noise (hair) and the severe class imbalance. Despite these improvements, all models could likely benefit from further refinement. Several limitations became apparent throughout the project.

Our masking method introduces several implications. First, it is easily affected by shadows, darker areas, or some data augmentation methods (e.g. CLAHE), leading to inaccurate masks. This affects the reliability of the extracted features. For example, as seen in Figure 19, an inaccurate mask can heavily inflate asymmetry and border feature values, creating heavy outliers and weakening the model.

Another problem arises when we use the annotated mask as a fallback when unable to produce a new one. This introduced multiple limitations, since the fallback masks sometimes have multiple lesions, and the feature extraction methods are based on the assumption of one lesion per image. The asymmetry feature is usually assumed to be in the range from 0-1. However, if multiple lesions are present in the same image, the range increases. This is due to the way asymmetry is computed by dividing the non-overlapping pixels by the total pixels in the mask. Nevertheless, since the data is standardized before training, the problem is mitigated. Border is computed only for the biggest ”true area” of the mask, meaning that it will completely ignore other lesions.

The region growing method for mask creation also has consequences for the color feature, since it stops as soon as the difference in color between pixels is more than a set threshold. Therefore, some of the color irregularity could be masked out.

An alternative approach could have been to use only the annotated masks. Since our new masks include only one lesion, we are dropping potentially valuable information from surrounding smaller lesions, such as color irregularity. However, this approach might lead to a poorer overall performance. As mentioned, the asymmetry and border irregularity methods depend on one lesion per image, meaning results could vary. Further, not all images have an annotated mask, which would force us to drop all images without an annotated mask. Since we have an annotated mask for $\tilde{9}0\%$ of the images, the problem is limited.

Furthermore, to evaluate the model’s performance before testing, we performed cross-validation with standard K-fold. However, due to duplicate lesions from the same patient, group K-fold could have been implemented instead (as it was in the open question models), to ensure we don’t obtain overly optimistic validation scores.

Finally, the dataset consists primarily of lighter skin tones, meaning our models likely don’t generalize to different skin tones, as it has only been trained on lighter skin tones. Furthermore, the ability of the region-growing method to create a mask is dependent on pixel intensity thresholds. This has been optimized for lighter skin tones,

where a difference in color between the lesion and skin tone is expected to be high. To adjust for this likely bias, the models could instead have been trained on a dataset with all skin tones equally represented. Additionally, we could have looked into the possibility of training individual models for each specific skin tone, where the region growing threshold is adjusted accordingly.

9 Conclusion

9.1 Key findings

The project has led to several key findings:

The ABC features, as implemented in this project, are not sufficient on their own to reliably predict melanoma.

We found that applying hair removal improved the model's performance slightly, increasing the AUC by 0.06 and the F1 by 0.05.

The model was also highly sensitive to class imbalance in the given dataset; without balancing, predictions skewed heavily towards the majority class. Across the models tested in the project, logistic regression appeared to be the best-performing classifier; however, the optimal classification threshold increased when using more balanced data (via data augmentation).

Data augmentation improved performance compared to the baseline model, however, not all augmentation methods were beneficial. The three image-level augmentation methods (Gaussian Noise, CLAHE, and distorting the border) seemed to decrease model performance, likely due to unrealistic transformations. When it comes to data resampling, random under-sampling also decreased the model's performance, predicting fewer melanoma cases accurately, possibly due to the loss of valuable data. SMOTE improved the model's F1 by 0.03 and AUC 0.01 compared to the baseline model. Overall, while SMOTE shows potential, it's effectiveness appears highly sensitive to implementation choices, underscoring the importance of further experimentation and model validation.

9.2 Future work

If more time and data were available, the model could be expanded to cover all ABCDE features, which could potentially improve the performance. Similarly, further investigation into the use of other variables from the dataset could be explored, such as medical history or demographics.

The low Cohen's Kappa for the hair removal function suggests that an improvement of this function is needed for better preprocessing. For example a method which also accounts for light-colored hairs would likely be an improvement.

As mentioned in section 8, due to the existence of multiple images of the same lesions, implementing Group-K-folds could be an alternative to the baseline evaluation method instead of the usual K-folds to improve model assessment.

As referred to throughout the project, the way the mask is created results in several problems, including bias towards skin tone and feature extraction complications. For future work, testing different mask creation methods is suggested.

In future work, it would be useful to test out even more combinations of data augmentation methods to showcase a clear overview of their interactions, or even add more augmentation methods to potentially improve performance.

References

Victoria State Government. 2024. *Melanoma*. Better Health Channel. Accessed May 20, 2025. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/melanoma

WHO. 2022. *Melanoma of Skin Fact Sheet*. Global Cancer Observatory, World Health Organization Accessed May 20, 2025. https://gco.iarc.who.int/media/globocan/factsheets/cancers/16-melanoma-of-skin-fact-sheet.pdf

Tsao et. al. 2015. *Early detection of melanoma: Reviewing the ABCDEs.*. American Academy of Dermatology Accessed May 20, 2025. https://www.sciencedirect.com/science/article/pii/S0190962215000900

Puckett et. al. 2024. *Melanoma Pathology.* StatPearls. Accessed May 20, 2025. https://www.sciencedirect.com/science/article/pii/S0190962215000900

Pacheco et. al. 2024. *PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones.* Mendeley Data. Accessed May 26, 2025. https://data.mendeley.com/datasets/zr7vgbcyr2/1

DATAtab *Fleiss' Kappa* Accessed May 23, 2025. https://datatab.net/tutorial/fleiss-kappa

scikit-learn developers. *StratifiedKFold* Accessed May 23, 2025. https://scikit-learn.org/

stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

Anonymous. 2020. *Data Augmentation in Training CNNs: Injecting Noise to Images* Conference submission at ICLR. Accessed May 26, 2025. https://openreview.net/pdf?id=SkeKtyHYPS

Li et. al. 2024. *Deep Learning-Optimized CLAHE for Contrast and Color Enhancement in Suzhou Garden Images* International Journal of Advanced Computer Science and Applications Accessed May 26, 2025. https://thesai.org/Downloads/Volume15No12/Paper_81-Deep_Learning_Optimized_CLAHE_for_Contrast_and_Color_Enhancement.pdf

Anonymous. (2020). ABCDE's of melanoma: Asymmetry, Border, Color, Diameter, Evolving. Accessed May 26, 2025. https://www.beaumont.org/conditions/melanoma/abcde's-of-melanoma#:~:text=Asymmetry%20%E2%80%93%20Melanoma%20is%20often%20asymmetrical,smooth%2C%20well%2Ddefined%20borders.

Perez et. al. 2018. *Data Augmentation for Skin Lesion Analysis* ResearchGate Accessed May 24, 2025. https://www.researchgate.net/publication/328010811_Data_Augmentation_for_Skin_Lesion_Analysis