IT UNIVERSITY OF COPENHAGEN

Second Semester Project

Medical Image Analysis

May 30, 2025

Contributors:
Anis Kadem, Olivér Gyimóthy, Rudra Kaushik, Etele Kovács, Péter Ónadi

Github: `https://github.com/Oliverdron/2025-FYP-groupOrca`

Bachelor Degree
Data Science

Contents

Introduction

Skin cancer affects millions of people worldwide, and finding it early makes treatment much more successful. While doctors use tools like dermoscopy to spot dangerous lesions, these methods are not always available to everyone, and non-specialists can miss subtle signs. At the same time, nearly everyone carries a smartphone with a decent camera, which opens up the possibility of simple self-checks at home. The challenge is that phone photos vary a lot in lighting, focus, and size, and things like hair or shadows can throw off automated analysis.

In this project, we zoom in on three straightforward visual cues—Asymmetry, Border , and Color from the classic "ABCDE" rule. First we keep our focus on these clear, easy-to-understand features, while avoiding the complexity of deep neural networks and large amounts of extra data. After that we try to include other measures as well. We test our approaches on the PAD-UFES-20 dataset, which includes over 2100 smartphone-taken lesion images with matching masks, so we know we are working under real-world conditions.

Annotation Analysis of the Preliminary Dataset

2.1 Purpose

As part of this project, we had to analyze manual annotations about hair density in images made by our preliminary groups. Our goal was to assess annotation quality, measure inter-rater agreement, and discover patterns, inconsistencies or bias in the annotations. This was a foundational task for our group before moving on to the final dataset.

2.2 Dataset

During the mandatory assignment phase, each group manually annotated hair density on a subset of a dataset of images. As all 5 members of our team were in different groups, we could use all annotations from the 5 different groups. Overall, we had approximately 600 photos annotated over 5 groups. The annotators each assigned a discrete rating for an image on the following scale:

Table 2.1: Hair Coverage Classification

Score	Description
0	No hair
1	Some hair
2	A lot of hair

In each group there were 4 to 5 annotators.

2.3 Methods

To assess the quality of the annotations, we employed several analytical methods:

- **Fleiss' Kappa Score**: Used to evaluate inter-rater agreement within each annotation group.
- **Majority Label Calculation**: Determined the majority label for each data point and analyzed its distribution.

- **Conflict Analysis**: Identified and examined common label conflicts in cases where no majority agreement was present.
- **Per-Rater Distribution**: Visualized the frequency of each label selected by individual annotators to assess labeling tendencies.
- **Pairwise Rater Agreement (Confusion Matrix)**: Assessed pairwise agreements and disagreements between annotators using confusion matrices.

2.4 Results

We experienced a fair amount of agreement, most annotators agreed on hair amount. However, there are groups which are less accurate and there are many images which caused confusion among annotators.

To evaluate inter-rater agreement across annotation groups, we computed Fleiss' Kappa scores. The results, shown in Table 2.2, range from moderate to substantial agreement. Based on these values, we selected Group G (lowest agreement) and Group B (highest agreement) for further analysis.

Table 2.2: Fleiss' Kappa Scores by Group

Group	Kappa Score
O	0.708
G	0.699
J	0.586
B	0.881
N	0.709

We interpreted the kappa scores using the following guidelines: values below 0.20 indicate slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; and values above 0.60 suggest substantial agreement.

2.4.1 Disagreement Patterns

Most images had a clear majority rating. In instances without a majority label, the most frequent conflicts were between ratings 1 and 2, followed by 0 and 1. These patterns suggest consistent disagreement boundaries rather than random variability.

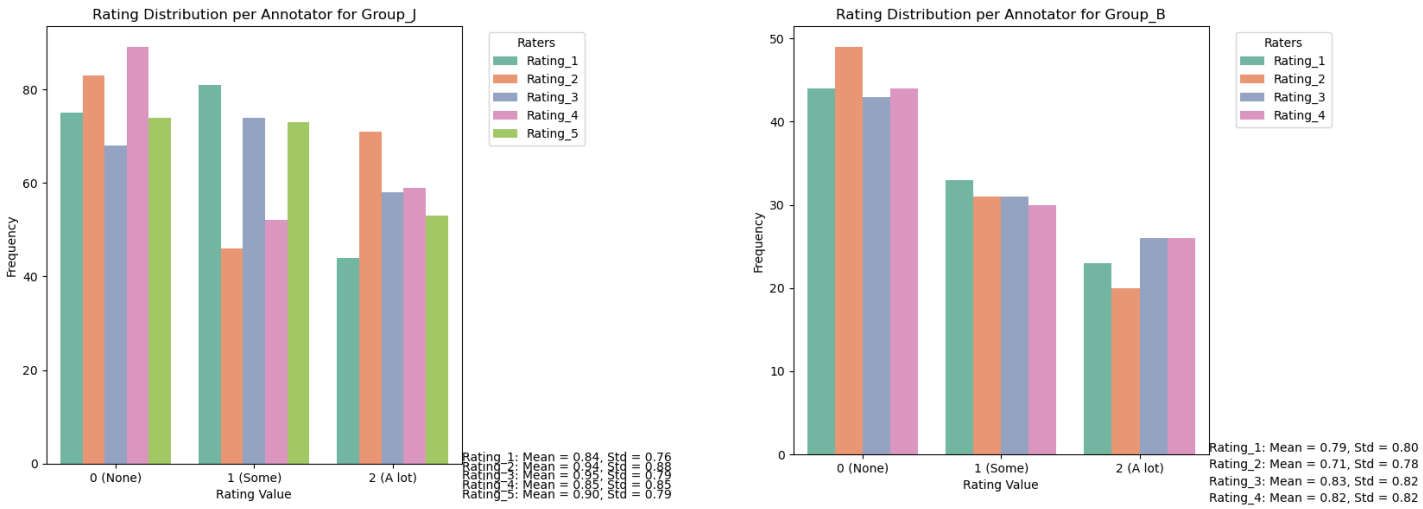
2.4.2 Per-Rater Distributions

Annotators had individual preferences in their use of labels. Some appeared to favor specific ratings, potentially indicating personal bias or differing interpretations of the annotation criteria.

2.4.3 Pairwise Confusion Matrices

Pairwise confusion matrices revealed high agreement between most annotators, typically exceeding 70%. Disagreements were predominantly mild, with label differences of one point ($\Delta = 1$). Rarely stronger disagreements ($\Delta = 2$) also occured. These findings suggest group-level variation may come from annotator bias, subjective judgment, or inconsistent attention to annotation guidelines.

Full heatmaps and label distributions are available upon request (Or see Appendix A).



(a) Group J (b) Group B

Figure 2.1: Per-annotator label distribution for Group J (lower agreement) and Group B (higher agreement).

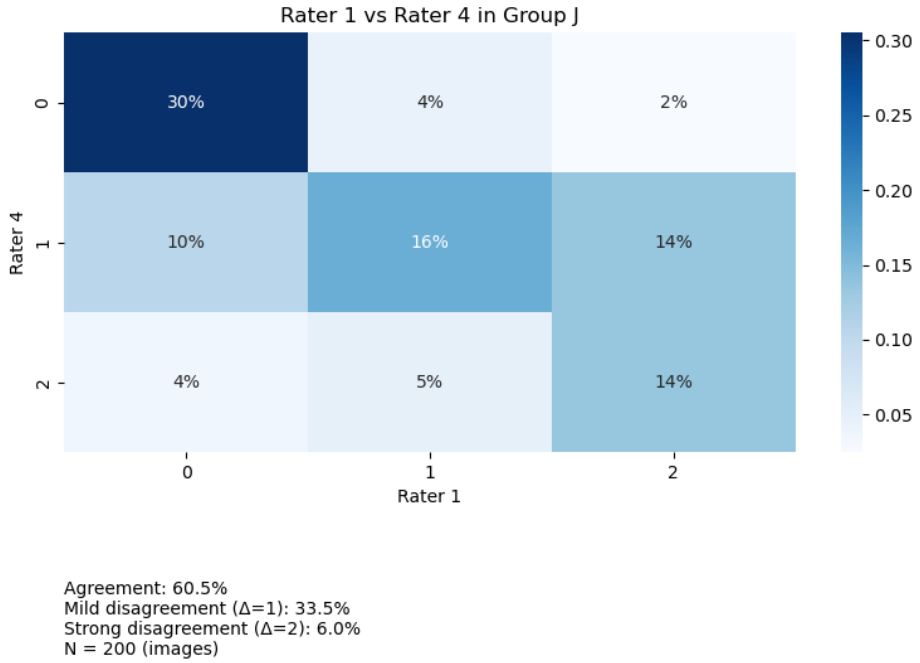


Figure 2.2: Pairwise confusion matrix for annotators in Group J, showing the agreement between a pair of annotators. The darker colors represent higher agreement (number of identical labels).

Skin lesion dataset

3.1 Data Source

The images and associated metadata employed in this study are drawn from the PAD-UFES-20 public repository (Pacheco et al., 2020). This dataset was acquired as part of a low-cost teledermatology initiative at the Federal University of Espírito Santo, Brazil. All clinical photographs were captured with consumer-grade smartphones under routine outpatient conditions, thereby providing a realistic spectrum of image quality and lesion presentations.

3.2 Data Composition

The dataset contains 2298 dermoscopic images drawn from 1373 patients, covering six different lesion types. Each record includes a high-resolution RGB image (in PNG format) that shows the lesion itself, alongside a corresponding binary segmentation mask (also PNG) that precisely outlines the lesion boundary.

Along with these images, a CSV file provides up to 26 metadata fields per record. For our analysis, we focus exclusively on three columns: the image file path, the mask file path, and the ground-truth diagnostic label. All other metadata—such as patient smoking and drinking habits or family history of cancer—is noted, but set aside, leaving those factors as an opportunity for open questions as follow-up studies.

3.3 Preliminary Characterization

An initial examination of the dataset's metadata highlights three important characteristics. First, the lesions fall into six diagnostic categories, three of which are malignant—basal cell carcinoma (BCC), squamous cell carcinoma (SCC, including Bowen's disease), and melanoma (MEL)—and three of which are benign—actinic keratosis (ACK), nevus (NEV), and seborrheic keratosis (SEK).

Second, the dataset exhibits a notable class imbalance: roughly 30% of the images depict cancerous lesions, while the remaining 70% represent benign conditions. This disparity must be taken into account during model training and evaluation to avoid bias toward the majority class.

Finally, there is considerable variability in image resolution. Some photos are as small as 100 × 100 pixels, whereas others exceed 3000×3000 pixels. This wide range reflects the use of different smartphone cameras and capture settings, and it poses challenges for pre-processing, where fixed-size operations may underperform.

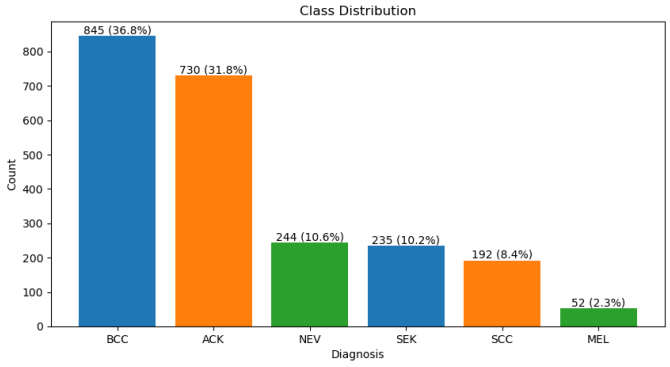


Figure 3.1: Class distribution of lesions in the dataset: 30% of the images depict malignant lesions, while 70% represent benign conditions. This class imbalance needs to be considered during model training to avoid biased performance.

3.4 Data Cleaning

Any image whose mask was entirely black (no lesion delineation), absent, or non-existent was deemed unusable and removed from the analysis. These invalid masks preclude calculation of morphological features — for example, a zero-area mask yields undefined values for perimeter, convexity, and fractal dimension — and would propagate NaNs or infinities through the downstream pipeline. In total, 141 images (6.14 % of the original repository) were excluded on these grounds. We also had to consider faulty masks, which we also excluded from the dataset. We detected these from visualizing our feature's values distribution.

Workflow of `img_preprocess_util.py`

4.1 Image Pre-processing

- The image pre-processing routine implemented in `img_preprocess_util.py` begins with **Bilateral denoising**, a technique intended to suppress high-frequency noise while preserving lesion boundaries. By weighting both spatial proximity and intensity similarity, the filter smooths pixel values without blurring edges. In practice, however, the fixed parameters chosen for the bilateral filter often prove suboptimal: under-smoothing leaves speckle noise that can corrupt texture metrics, whereas over-smoothing attenuates fine lesion details essential for accurate asymmetry and border irregularity measurements.

- Subsequent quality assessment employs both the **Structural Similarity Index (SSIM)** and **Peak Signal-to-Noise Ratio (PSNR)** to decide whether a more aggressive Non-Local Means denoising step is warranted. Images registering an SSIM below 0.80 or a PSNR under 20 dB are passed through the NLMeans algorithm. While this two-stage approach seeks to adaptively address heterogeneous noise levels across smartphone devices, its reliance on rigid thresholds fails to accommodate the wide variability in skin phototypes and lighting conditions.

- **Contrast Limited Adaptive Histogram Equalization (CLAHE)** in the LAB color space is then applied to enhance local contrast and reveal subtle textural features. Although CLAHE can unearth faint pigment networks or vascular patterns, its tile-based nature sometimes introduces boundary artifacts—small "nubbly" regions where adjacent tiles meet.

- Hair removal is performed via a **morphological black-hat operation** that highlights hair structures against the background, and neighbouring pixel values are used to fill the detected regions. Unfor- tunately, this heuristic approach struggles whenever hair density is high or when hairs are very light (e.g., blond or white), producing blurred or incomplete removal that leaves residual artifacts.

- Finally, lesion boundary is extracted by cascading **Sobel** gradient computation with **Canny edge detection** using global thresholds (low = 50, high = 150). These fixed thresholds are often bad for images with uneven illumination and wrinkles are frequently misclassified as lesion edges, contaminating subsequent shape and perimeter-based feature calculations.

- To address these shortcomings, more adaptive or learning-based methods should be considered. A small convolutional network (e.g., a U-Net variant) trained specifically to segment hair artifacts could replace morphological heuristics, avoiding the pitfalls of inpainting.

Feature Extraction

The feature-extraction stage begins once each lesion image has been pre-processed and its binary mask paired. A suite of descriptor functions—covering shape (e.g. area, perimeter, fractal dimension), color (e.g. mean hue, variance), and texture (e.g. local binary patterns)—is applied in turn to the 'cleaned' image and mask. Each function delivers a single numeric value, which is recorded alongside the lesion's metadata. By writing out these values immediately to the output CSV, the pipeline produces a table of features that is ready for further action, analysis and classifier training, while it is still possible to add new records without changing the existing data.

After we extracted features, we could also analyze their correlation and their distribution. This helped us spot outlier images or masks, and decide whether features influence one another or not.

5.1 Asymmetry

First, the mask is labeled and the k largest connected components are retained to focus on the principal lesion. Each component is then rotated through n evenly spaced angles between $0°$ and $180°$, tightly cropped, and bisected horizontally and vertically. The pixel-wise exclusive-OR between each half and its flipped counterpart yields a mismatch count, which is normalized by the blob's area to produce a per-rotation asymmetry score. Averaging these values across rotations and blobs yields a final asymmetry metric in $[0, 1]$.

Because this method relies on precise mask delineation, small errors in contour extraction or extreme aspect-ratios (e.g., very elongated lesions) can inflate the score, producing outliers.

5.2 Border Irregularity

Border irregularity combines two classical shape descriptors into a single index, $M = I \times C$, where $I = P^2/(4\pi A)$ measures "spikiness" (with perimeter P and area A), and $C = A_h/A$ quantifies convexity relative to the convex hull area A_h. After isolating the top k blobs, each blob's perimeter and area are computed, and convex-hull area is estimated via the Quickhull algorithm. Multiplying I and C amplifies deviations from circular, convex shapes characteristic of malignant growth.

Perimeter estimation can be sensitive to resolution and mask noise: small artifacts or "jagged" boundaries may exaggerate P, while oversmoothed masks under-estimate it. Very small lesions produce unstable ratios.

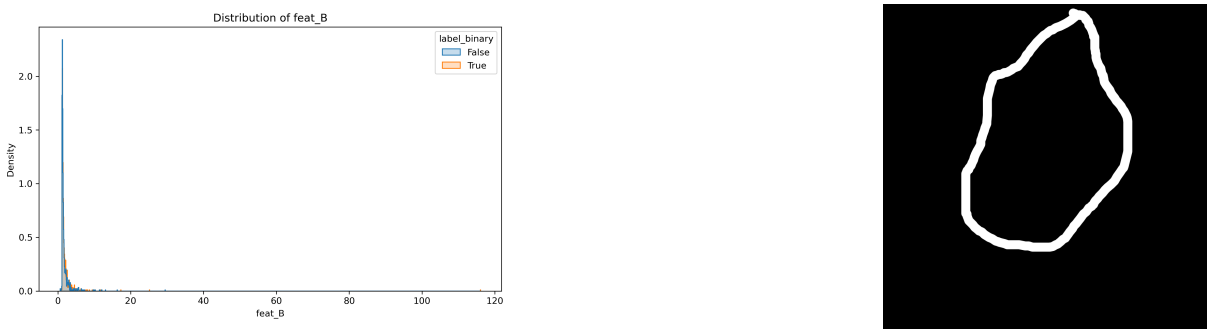


Figure 5.1: Analysis of Border Irregularity index (B). (a) The distribution of B exhibits a peak with outliers, suggesting most lesions have regular shapes, but some exhibit irregularities that are indicative of outliers. (b) The second image illustrates the outlier mask where the segmentation is incomplete, leading to an artificially inflated B value due to an exaggerated perimeter. This highlights the sensitivity of the Border Irregularity index to resolution, mask noise, and segmentation artifacts.

5.3 Color Heterogeneity

Color heterogeneity measures the maximal perceptual distance between dominant lesion hues in CIE-Lab space . Each of the k largest blobs is cropped from the original image, optionally downsampled, and converted from RGB to Lab. *K-means clustering* partitions pixel colors into n clusters; the maximum Euclidean distance among cluster centroids represents the blob's heterogeneity. Averaging across blobs yields a final value in $[0, \infty)$.

This feature can be confounded by uneven illumination or specular highlights, which introduce artificial color clusters.

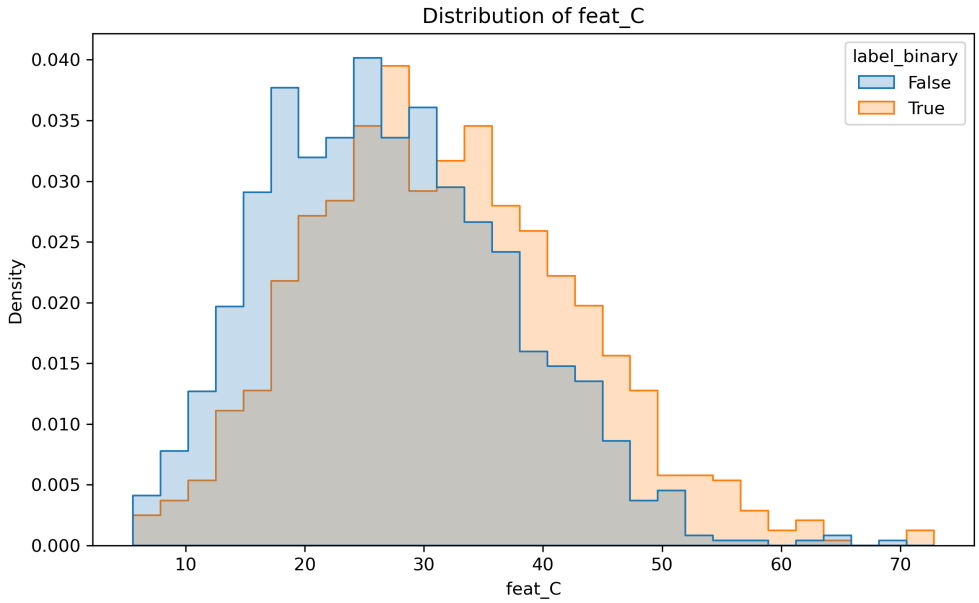


Figure 5.2: Distribution of cancerous and non-cancerous lesions with respect to their color heterogeneity (C value). While the distributions are similar, they are not identical, suggesting that the color heterogeneity feature could still help in distinguishing between the two classes, though it may require further refinement due to potential overlap.

5.4 Hair Coverage

Hair coverage assigns a discrete density category by thresholding a hair-enhanced mask, filtering out small components, and computing the proportion of pixels occupied by hair. Coverage fractions above 70% map to "high", between 40%-70% to "medium", and below 40% to "low". Dense hair can obscure lesion borders, reducing the reliability of shape and color descriptors. However, hair coverage itself is a poor predictor of malignancy and correlates negligibly with diagnostic labels.

5.5 Vascular Score

The vascular score isolates red-channel enhancement and Frangi vesselness to estimate the prominence of blood-vessel–like structures . After gamma-correcting the red channel, the image is converted to HSV and thresholded to produce a red-region mask. Frangi filtering on the grayscale image highlights tubular structures; multiplying vesselness by the red mask and summing yields a raw score

Bright illumination and erythematous backgrounds can produce false positives, while poorly contrasted vessels in dark skin types may be under-detected.

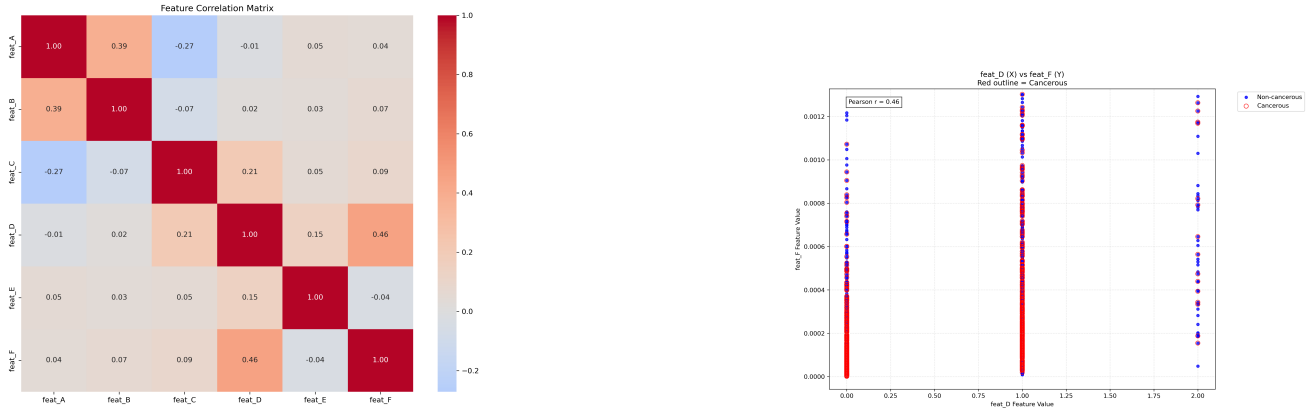


Figure 5.3: Analysis of correlations and scatterplot of specific features. (a) The first image shows all feature correlations, revealing that most features are not strongly correlated with each other, indicating independent relationships between hair coverage and other characteristics. (b) The second image focuses on the correlation between features D (hair score) and F (streak score). The moderate correlation between the two (Pearson = 0.46) suggests a mild association, likely due to both being line-based features, where the method may confuse hair with streaks."

5.6 Streak Score

Streaks are assessed by extracting a border ring around the lesion—via dilation minus the original mask—and computing the mean normalized Frangi response within that ring . Contours identify the largest blob, whose area must exceed a minimal threshold to avoid noise. The Frangi output is normalized to [0, 1], masked by the ring, and averaged to yield the streak score.

Small lesions or those with indistinct margins produce spurious low scores, while skin folds or wrinkles can mimic streak-like signals, generating false positives.

Classifier

6.1 Architecture and Workflow

Classification is built upon a base Classifier class that sums up shared evaluation and prediction. The TrainClassifier subclass implements the processing of the cleaned feature CSV file, performs a patient-wise split to training, validation, and test sets using stratified grouping to ensure balanced label distribution across the sets, conducts randomized hyperparameter search on user-specified pipelines, and optimizes decision thresholds based on a chosen scoring metric (e.g., F1). The LoadClassifier subclass simply loads previously saved models (.pkl files) and exposes identical evaluation and prediction methods, enabling rapid deployment once training is complete. Classifier uses Pipeline object which is a sequence of data transformers with a final predictor, the classifier model. We used StandardScaler scaler for all of our classifiers as a data transformer, in order to standardize features by removing the mean and scaling to unit variance so we ensure equal weighting of features.

6.2 Objective

Our goal is to assign a binary label—cancerous versus non-cancerous—to each dermoscopic record, leveraging the six handcrafted features (A–F) as inputs. In classical supervised fashion, labels are derived from the `label_binary` field (1 = cancerous, 0 = benign), and model performance is assessed on held-out patients to ensure generalizability across individuals.

6.3 Model Selection Strategy

Multiple candidate pipelines are compared by their mean cross-validation ROC AUC and by the stability of their learning curves. The training hyperparameter tuning method records the best parameters and CV scores for each model; we select the pipeline with the highest validation AUC, subject to minimal overfitting (i.e., small train–validation gap).

Choosing the appropriate features for the model is also a challenging task, as different subsets of features can lead to varying results. To address this, we conducted feature importance analysis using methods such as permutation importance and feature selection techniques. Feature importance plots were generated to visualize which features had the most impact on model performance, helping us identify the most relevant features for predicting malignancy(See Appendix A).

We also visualized several model evaluation metrics to assess performance comprehensively. These included learning curves to monitor convergence, decision boundaries to understand how the model classifies data, and PAC-ROC (Partial Area Under the ROC Curve) for more detailed performance analysis across different decision thresholds. Additionally, calibration curves were used to assess how well the predicted probabilities align with the true outcomes. Post-threshold optimization, we also inspect precision–recall trade-offs to choose the most clinically appropriate cutoff (balancing sensitivity and specificity).

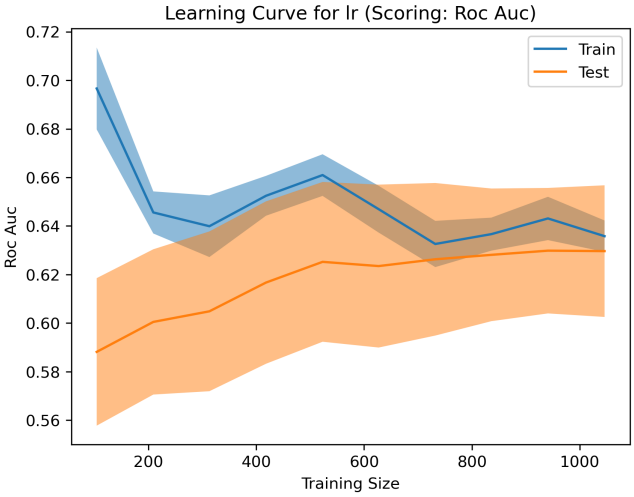


Figure 6.1: Learning curve for Logistic Regression classifier. As the number of training samples increases, the accuracy on both training and test sets stabilizes at approximately 0.63.

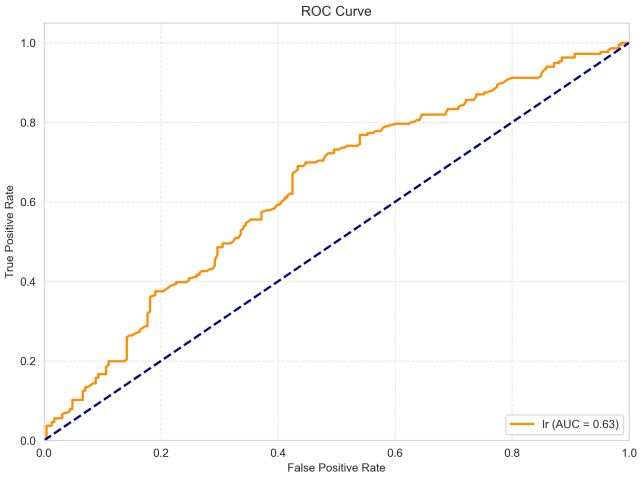


Figure 6.2: ROC curve for Logistic Regression classifier. The AUC of 0.63 suggests moderate ability to separate the classes.

6.4 Performance Analysis

6.4.1 Baseline Model

We built he baseline model using features A to C and employed two classifiers: *Logistic Regression* and *Random Forest*. On the held-out test set, Logistic Regression emerged as the top-performing model, achieving an ROC AUC of 0.573 and an overall accuracy of 0.573. The sensitivity (recall) for detecting cancerous lesions was 0.548.

The learning curve shows convergence after approximately 60% of the training data, indicating that the sample size was likely sufficient for this feature set. Additionally, the ROC and calibration plots reveal a moderate level of discriminative performance and probability calibration(See Appendix A). The Random Forest classifier and other baseline models performed worse across almost all metrics.

6.4.2 Extended Model

The extended model used all six extracted features and evaluated six classifiers: *Logistic Regression*, *Random Forest*, *MLP*, *K-Nearest Neighbors*, *Gradient Boosting*, and *XGBoost*. Following model training, hyperparameter tuning, and validation, the best-performing model was MLP, which achieved an ROC AUC of 0.61, an overall accuracy of 0.60, and a recall of 0.64.

The remaining models underperformed to varying degrees—some came close to the MLP's performance, while others fell significantly behind.

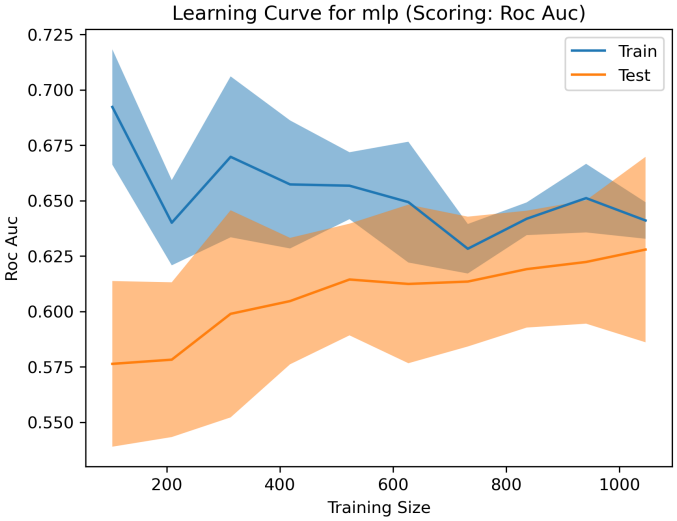


Figure 6.3: Learning curve of the MLP classifier.

Figure 6.4: Cross-validation performance comparison of all classifiers.

6.5 Sources of Error and Overfitting

Inaccurate predictions most often arise from small or ambiguous lesions (low mask-to-image ratios) and from images with heavy hair or uneven illumination that distort feature values. Overfitting manifests when high-capacity models (e.g. deep trees) capture noise in the training subset—evidenced by a gap between training and validation ROC AUC in the learning curves. Mitigation strategies include limiting tree depth, applying stronger regularization in logistic or SVM pipelines, incorporating PCA for dimensionality reduction, and enforcing early stopping or reduced `n_iter` in hyperparameter searches.

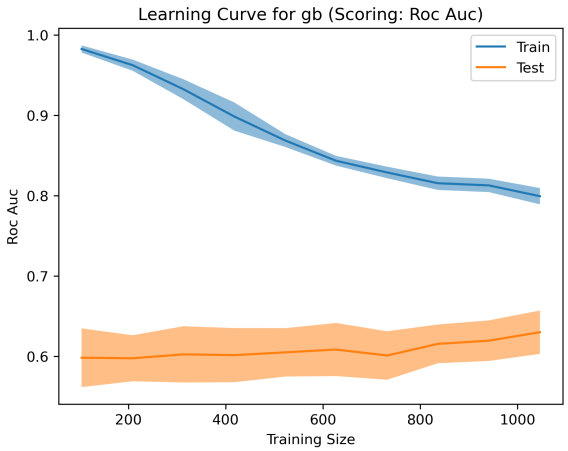


Figure 6.5: Example of overfitting: The learning curve shows a persistent gap between training and validation ROC AUC, indicating the model is capturing noise rather than general patterns.

Discussion and conclusions

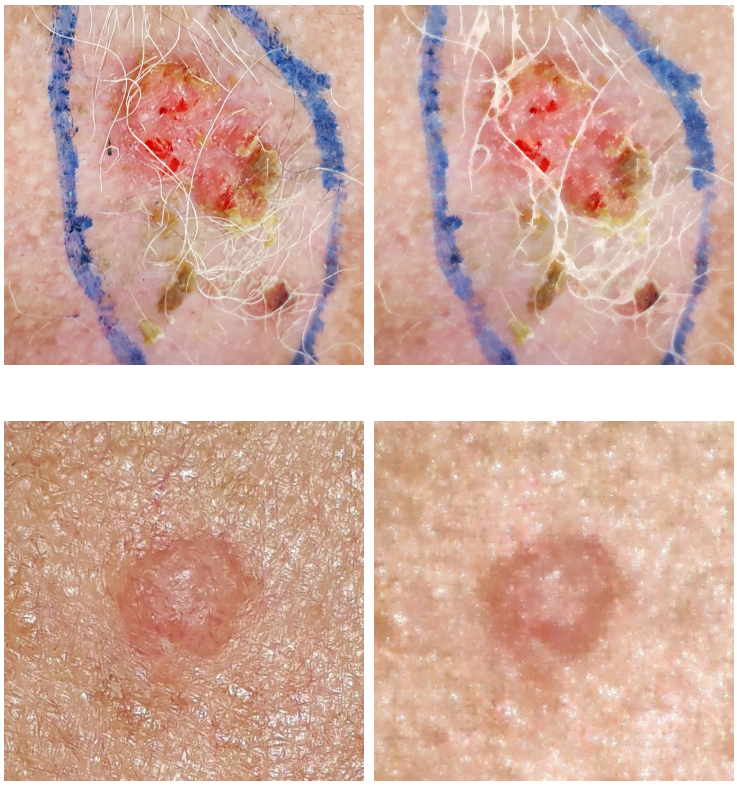
7.1 Limitations

Firstly, because our entire study is grounded in the PAD-UFES-20 dataset—clinical images captured on consumer-grade smartphones under routine outpatient conditions—several intrinsic constraints must be acknowledged. First, image quality varies dramatically: some frames exceed 3 000×3 000 px, offering rich detail for edge-based features but incurring high storage and compute costs, while others are as small as 100×100 px, which expedites processing but yields noisy, quantized measurements that destabilize perimeter and area calculations.

Secondly, dense hair occlusion—especially bright (blond or white) strands—cannot be reliably inpainted and often leaves blurred residues, corrupting both shape and color features. Severe uneven illumination (e.g. one-sided shadows or specular highlights) forces global thresholds in edge detectors to rise, causing loss of subtle vessel or lesion boundary information. Blurry or out-of-focus images further degrade texture and streak metrics, while pronounced skin wrinkles can be mistaken for vascular or streak signals. Add to this medical markings drawn around lesions, and it becomes clear that spurious contours may dominate the extracted descriptors.

Not to mention, segmentation masks themselves exhibit heterogeneous quality: some contain multiple disconnected blobs, others fail to fully cover the lesion, and a minority are entirely absent or misaligned. Such inconsistencies can produce outliers in convexity and mask-to-image ratios, necessitating manual thresholding rather than an end-to-end adaptive solution.

Lastly, all pre-processing steps—hair removal, denoising, edge detection—use fixed parameters, rendering them brittle across the dataset's spectrum of skin phototypes and acquisition conditions. The lack of automated, data-driven parameter tuning thus limits generalizability and demands careful manual review whenever deploying this pipeline on a new dataset.



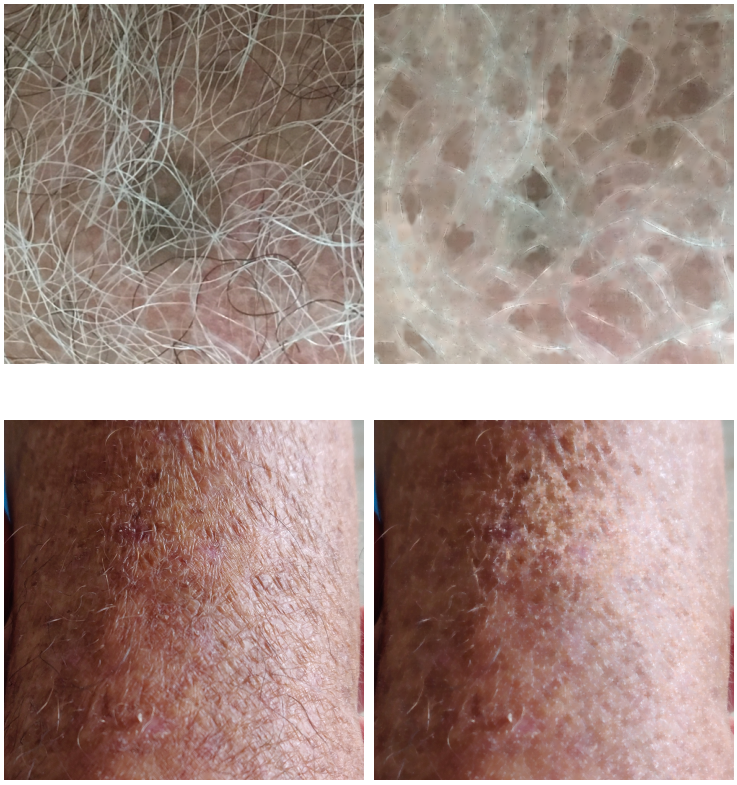


Figure 7.2: (original image on the left, inpainted on the right)
Examples of the mentioned disadvantages in order:
bright hair, wrinkled skin, too much hair, uneven illumination

7.2 Concluding Remarks

Medical image processing has become an indispensable component of contemporary healthcare, extending well beyond dermatology into domains such as neuroimaging, radiology, and pathology. Techniques for automated tumour detection in brain MRI, pulmonary nodule classification in CT scans, and histopathological analysis of biopsy slides illustrate the broad potential of computer-aided diagnosis to augment clinician expertise, reduce diagnostic delays, and improve patient outcomes.

Identification of cancerous lesions is key to prevention. Developing an application for this cause would save many lives across the world. Making it easy to check a lesion periodically, from home. This would also reduce the burden on healthcare systems, or help people how can not get professional help.

Realizing this vision, however, depends critically on the quality of input images. We recommend the following best practices for consumer acquisition: ensure uniform, diffuse lighting (avoid strong directional shadows), maintain a stable camera-to-skin distance (ideally 10–15 cm), use a neutral background free of glare or distracting patterns, and gently clear hair away from the lesion without stretching the skin. Adherence to these guidelines will produce the sharp, evenly illuminated photographs necessary for reliable pre-processing and feature extraction, thus maximizing the utility of automated lesion-detection tools in real-world settings.

7.3 Future Work

While our current approach is good, there are plenty of space for improvements. First, taking multiple photos of the same lesion over time could contribute to a growing feature, changing shape, or color shifts can signal a developing cancer. Second, adding simple patient details like smoking and drinking habits could help the model spot lifestyle-related risks; studies hint at a link between these behaviors and certain skin cancers, so it makes sense to feed that information in. Finally, asking about family history of cancer could flag inherited risk factors. By combining image features with personal and genetic background, we

could build a more accurate, personalized early-detection system—especially for people already at higher risk.

7.4 Open question

To what extent could improving imaging conditions for example consistent lighting, image size and resolution enhance the performance and reliability of our skin lesion classification pipeline?

The quality of the performance is heavily dependent on the quality and consistency of the input data.

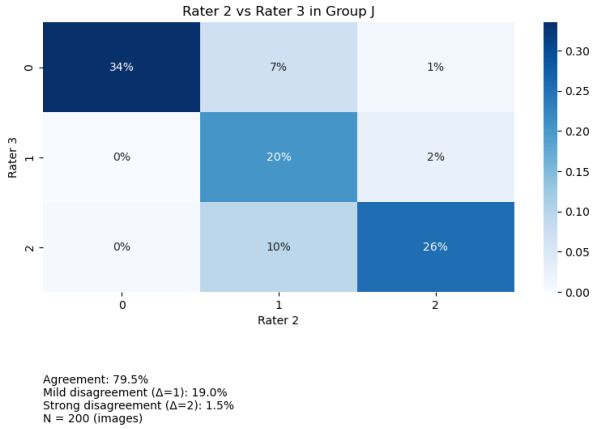
In the current input, images from the PAD-UFES-20 dataset present a varying conditions, for example inconsistent lighting and different resolutions. These variations are misleading into wrong feature values even producing outliers. In the end the model's accuracy and precision is dropping.

Consistent lighting could remove shadows and glare, that would improve the quality of the features like finding borders, vessels or color homogeneity. Similarly, standardized image size and resolution could make features including diameter and area viable.

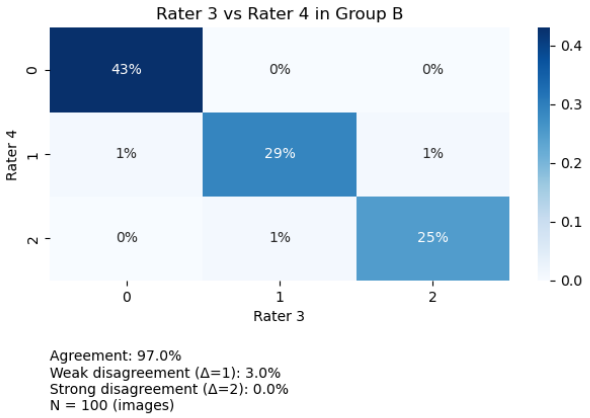
We could make experiments where we preprocess a subset of the dataset to simulate improved conditions. For example, applying histogram equalization, resizing, etc., and then comparing the performance metrics before and after.

Ultimately, this question is about that we have to deal with real world problems, while trying to reduce them and much as possible, in order to get quality feature values. Taking these points into consideration for future data collections could improve the pipeline's performance.

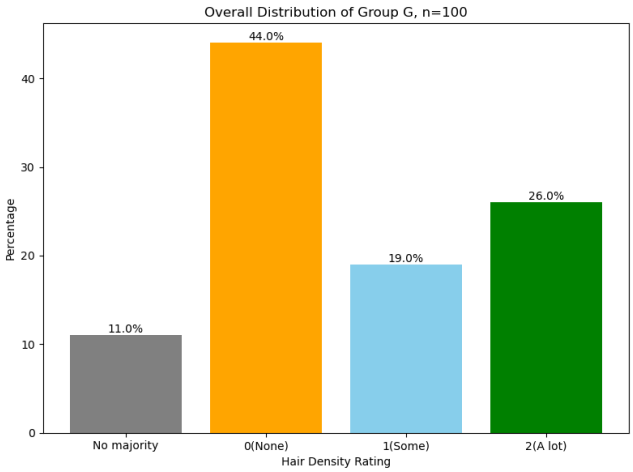
Appendix



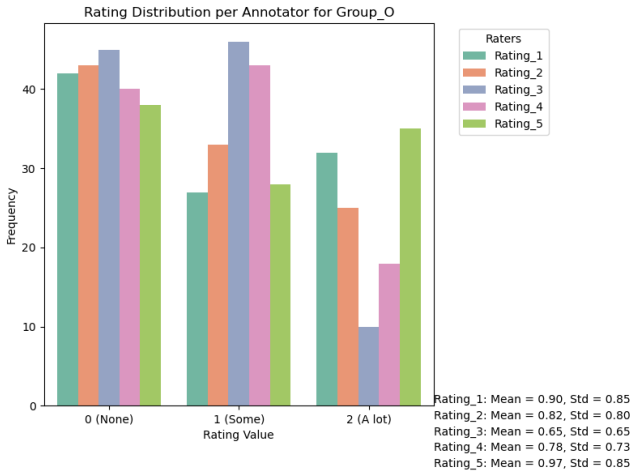
(a) Heatmap in group J for two raters



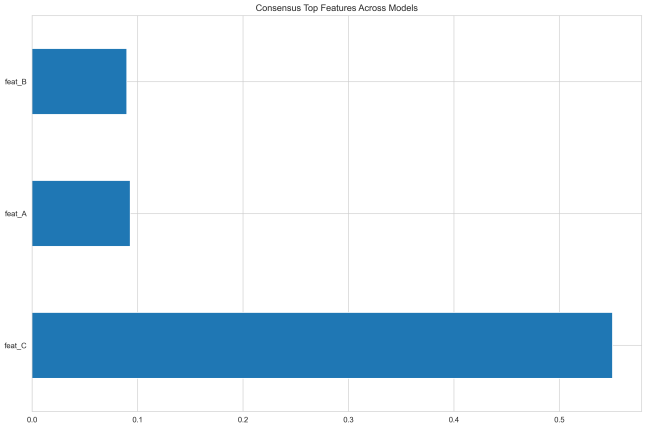
(b) Heatmap in group B for two raters



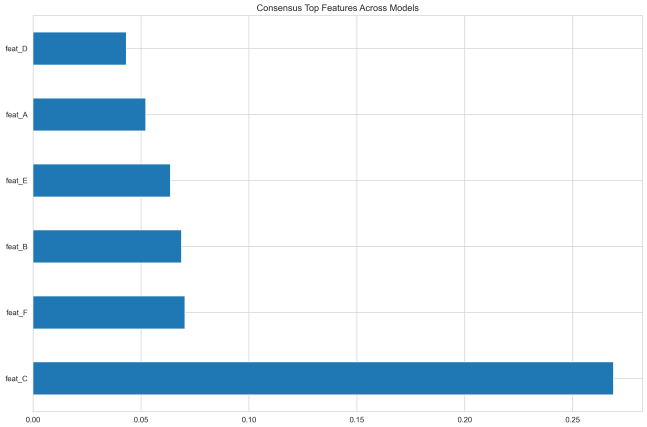
(c) Label distribution for group G



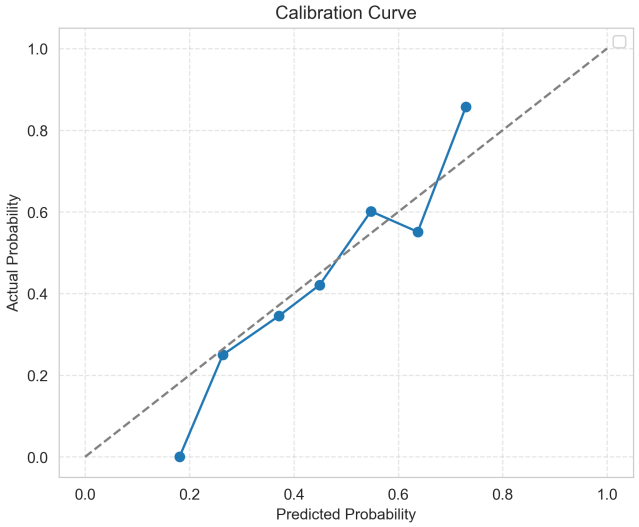
(d) Rating distribution for group O



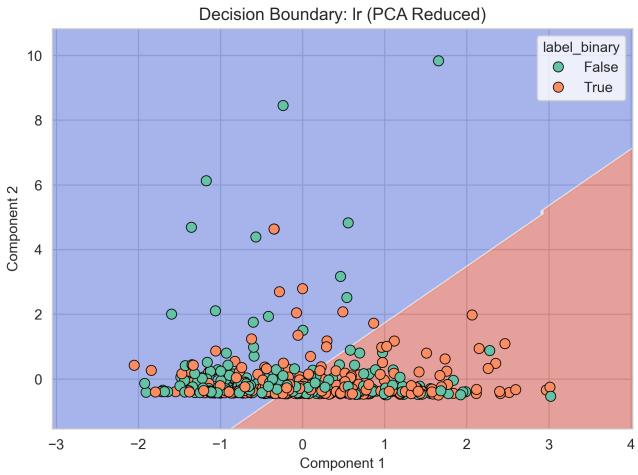
(e) Feature importance scores for baseline features



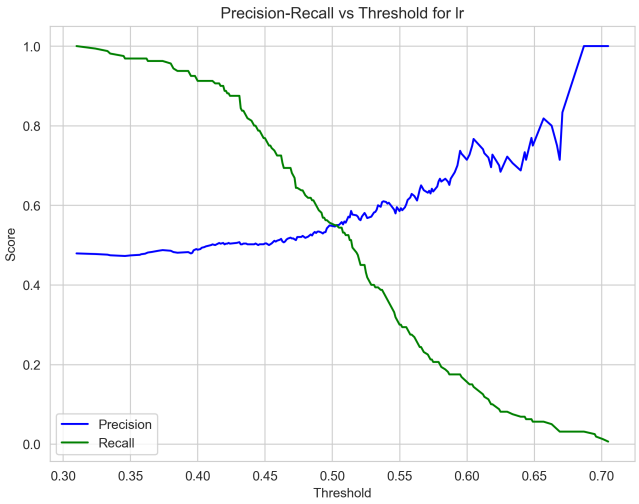
(f) Feature importance scores for extended features



(a) Calibration curve for baseline best-model



(b) Decision boundary for baseline best-model



(c) Precision recall vs threshold figure

Upon request all plots and figures are available.

References

[1] Pacheco, A. G. C., & Krohling, R. A. (2020). *PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones*. Data in Brief, 32, 106221.

[2] Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S. W., Gutman, D., ... & Halpern, A. C. (2019). *Skin lesion analysis toward melanoma detection: A challenge at the 2018 International Skin Imaging Collaboration (ISIC)*. Medical Image Analysis, 59, 101555.

[3] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). *The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions*. Scientific Data, 5, 180161.

[4] Rocha, A., Silva, D. F., Lima, L. S., de Souza, R. M., & Papa, J. P. (2020). *A new approach for skin lesion diagnosis based on hybrid deep learning systems*. Electronics, 9(9), 1503.

[5] Nazari, S., & Garcia, R. (2023). *Automatic skin cancer detection using clinical images: A comprehensive review*. Life, 13(11), 2123.

[6] Oda, J., & Takemoto, K. (2025). *Mobile applications for skin cancer detection are vulnerable to physical camera-based adversarial attacks*. Scientific Reports, 15, Article 3546.

[7] Scikit-learn developers. *Scikit-learn User Guide*. Archived at: https://web.archive.org/web/20250528201846/https://scikit-learn.org/stable/user_guide.html [Accessed May 2025].