Skin Lesion Classification

Anna Pólya
apol@itu.dk

Bogdan Alexandru Vanghelie
bova@itu.dk

Filip Madzinski
fili@itu.dk

Matej Paldan
pald@itu.dk

Miroslava Gemelová
mirg@itu.dk

Link to Github repository:
https://github.com/mirkagem/2025-FYP-groupG

May 30, 2025

Abstract

Early detection of malignancy in skin lesions is vital for improving the survival rate, which can be enhanced by accurate classification methods for at-home detection. This study evaluates classification methods using a data set of images and corresponding metadata with lesion-related information, to identify malignant and benign lesions. Using visually apparent features, K-Nearest Neighbours (KNN) and Random Forest (RF) models achieved high recall for cancerous lesions but exhibited high false positive rates. A metadata-driven model incorporating clinically relevant, non-visual features improved benign lesion detection while maintaining high sensitivity for malignancy. The final combined model, integrating selected metadata with ABC image features, achieved the best balance.

1 Introduction

Skin cancer is among the most prevalent forms of cancer worldwide, and melanoma represents one of its most aggressive and potentially lethal types. One in every three cancer diagnoses is skin cancer[1], and early detection remains critical for improving patient survival rates. When diagnosed in an early stage, the five-year survival rate for melanoma exceeds 97%, but this drops significantly, to approximately 30%, once the disease has progressed to later stages. This stark contrast underscores the urgency to develop reliable methods for the early detection and diagnosis of malignant skin lesions.

Visual assessment remains the primary method used by dermatologists to evaluate suspicious lesions. The ABC rule, assessing Asymmetry, Border, and Colour, provides a heuristic framework for identifying features commonly associated with skin cancer. Although this method has shown utility in clinical settings, it is limited by its subjectivity and variability between observers.

Access to timely dermatological evaluation can be limited due to long waiting times, geographic barriers, or a shortage of specialists in certain regions. These challenges have motivated the development of automated tools to assist in the early detection of skin cancer, particularly for individuals who may not have immediate access to professional care. Despite advances in computer-aided diagnosis, many existing systems focus solely on image-based features and overlook additional lesion-specific information that can be clinically relevant. In this study, we incorporate detailed lesion-related indicators, such as whether the lesion has itched, hurt, grown, bled, changed in appearance, or is elevated. These descriptors, often available through patient self-report or clinical observation, can provide valuable insight into the lesion's progression and severity. Furthermore, while image-based mobile diagnostic tools are becoming increasingly accessible to the general public, they are frequently used in uncontrolled environments and have shown a tendency to yield false-negative results. This poses a significant risk, as undetected malignant lesions may lead to delays in clinical diagnosis and treatment.

We developed and evaluated four models for binary and multiclass classification tasks: (1) a baseline model using ABC features, (2) an extended model including an additional hair-coverage feature, (3) a metadata-only model, and (4) a combined model integrating both image-derived and contextual data. Throughout the study, we focused on capturing as many true cases of skin cancer, reflecting the clinical imperative to minimise false negatives and ensure that as many malignant lesions as possible are correctly identified, while also aiming to avoid simply labelling all lesions as cancerous.

By comparing the predictive performance of different feature combinations, this research explores the role of both visual and non-visual data in the early detection of potentially cancerous skin lesions. The broader aim is to support the development of safer and more effective tools for at-home use, helping individuals identify suspicious lesions earlier and seek timely medical evaluation.

2 Related work

Dermatologists use various procedures to distinguish malignant melanomas from benign lesions, such as the ABCD rule (where D stands for diameter) [2], and the seven-point checklist. According to R. H. Johr [3], the features extracted using the ABCD rule are computationally less expensive compared to those of the seven-point checklist. It has also been observed that the ABCD rule yields a higher consistency in clinical diagnosis. In another study [4], the author investigated the possibility of automatically detecting dermoscopic patterns based on the ABCD rule using deep convolutional neural networks. The experimental results demonstrated an 88% accuracy in skin cancer classification. In a related study, Moussa et al.[5] employed the ABD rule—excluding colour from the traditional ABCD rule, as colour processing demands significant computational resources. They used the KNN classifier and achieved 89% accuracy.

From literature, it is observed that different features extraction techniques have been used to classify the skin lesion. However, the result of testing these combinations may be even better if we add up the metadata features. That is why we would like to include two other models; the metadata model, focusing on predicting cancer based on metadata only, and the combined model, which explores ABC features and metadata.

3 Data

3.1 Source

The dataset[6], PAD-UFES-20, used in this project was provided by Mendeley Data and collected using the Dermatological and Surgical Assistance Program. It includes 2298 images and metadata related to said images in a CSV file. All images have up to 24 features, including patient and lesion ID, diagnostic, age of the patient and symptoms, such as the lesion growing, hurting, bleeding or itching. There are 3 types of skin diseases and 3 types of skin cancer present in the dataset, the latter being proven by biopsy. Along the dataset, we were provided with masks for most but not all images.

Skin lesion	Percentage	Diagnosis
Basal Cell Carcinoma	36.8%	Cancerous
Squamous Cell Carcinoma	8.4%	Cancerous
Actinic Keratosis	31.8%	Non-cancerous
Seborrheic Keratosis	10.2%	Non-cancerous
Melanoma	2.3%	Cancerous
Nevus	10.6%	Non-cancerous

3.2 Data binarisation

The initial metadata, provided in CSV format (later referred to as metadata), presented two key challenges: the presence of missing values and inconsistencies in how binary features were represented across different columns. Although these features were intended to encode binary information (such as whether a lesion bled, hurt, or changed) some columns used Boolean values, while others stored equivalent information as string literals, creating inconsistencies that could hinder model interpretability and performance.

As a first step, we addressed missing data by dropping rows containing missing values. This allowed us to focus solely on the formatting inconsistencies across binary columns. Upon inspection, we observed that although formats varied across columns, the formatting within each individual column was consistent. This observation enabled a systematic approach to binarisation.

To standardise the representation of binary features, we implemented a general-purpose function that could process all relevant columns, converting string or Boolean values to a uniform binary format (0 or 1). This preprocessing step ensured that the dataset was clean, consistent, and fully numerical; ready for use in subsequent machine learning models.

3.3 Data cleaning

Following a thorough inspection of the metadata, 1474 images and their corresponding metadata entries were selected for use in the study. This decision was informed by the observation that a substantial portion of the dataset lacked complete metadata, which is essential for some feature extraction and modelling tasks. The selected subset consists exclusively of images for which full metadata was available.

4 Baseline model

The baseline model was developed as an initial step to evaluate the performance of the ABC features. This model served two primary purposes; first, to assess how well these features, when combined, could effectively distinguish between cancerous and non-cancerous lesions; and second, to identify the most suitable classification method.

Selecting the appropriate classifier was essential to ensure that the extracted features could be used to their highest potential, allowing for meaningful performance tuning in later stages. This phase of the study, therefore, focused not only on validating the utility of the ABC features but also on determining which classifier yielded the most reliable results in terms of both overall accuracy and the model's ability to correctly identify malignant lesions.

4.1 Image loading

The images are loaded using an ImageDataLoader class, fetching the paths of the image and mask directories. Matching images and masks are loaded, the mask is applied to the image and the masked image, mask and ID are stored in a list. This list can be later used for feature extraction, allowing easy access to image data relevant to a patient. In the rare case the image does not have a mask, a mask is created by Otsu's thresholding. It should be noted that these masks are typically of lower quality than the provided masks, however this is only applied to 5% of the used data.

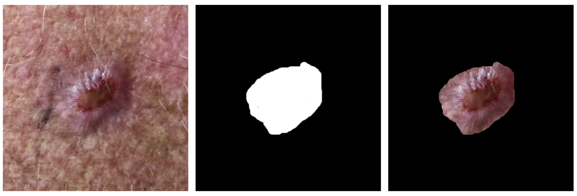

Figure 1: Masked image

With the dataset cleaned and the images successfully loaded, we proceeded to the next stage of the pipeline: feature extraction.

4.2 Features extraction

An essential early step in the development of the baseline model involved the evaluation and selection of suitable feature extraction methods for each component of the ABC framework. The contenders for feature extraction methods were taken from past projects examining skin cancer. Each function was applied to a representative subset of 200 images drawn from a separate skin lesion image dataset to assess its reliability and consistency. A research was then conducted, examining the strengths and limitations of each method within the specific context of skin cancer detection. This analysis considered both the theoretical basis of each function and its empirical performance on the test images.

Based on the outcomes of this evaluation, a single feature extraction function was selected for each of the three ABC criteria (asymmetry, border, and colour variation), ensuring that the final model incorporated the most robust and diagnostically meaningful representation of each feature.

4.2.1 Asymmetry

For the asymmetry component, several functions were evaluated, each quantifying asymmetry in a different manner. These were by calculating the mean, worst and best asymmetry. To justify this step we refer to the general ABC rule - if a lesion is asymmetric, it is often indicative of cancer[7].

Based on our testing, we have concluded that the function measuring the worst asymmetry (on multiple rotations of the image) was performing the best, and therefore we chose it for our ABC model later.

4.2.2 Border

To represent border irregularity, we considered three potential feature extraction methods: compactness, convexity, and the detection of streaks. Preliminary testing revealed that the streak-based method was susceptible to the presence of hair in the images, which significantly interfered with its reliability and reduced its diagnostic utility. The convexity measure, while robust to such interference, produced minimal variability across samples, showing little distinction between cancerous and non-cancerous lesions. Consequently, we selected compactness as the most informative and reliable indicator of border irregularity. Compactness quantifies the extent to which a lesion's shape deviates from a perfect circle, a property which has been clinically associated with malignancy.

4.2.3 Colour

For the colour component, we evaluated combinations of two colour spaces, RGB and HSV; and two measures, mean and variance. While HSV provides an alternative representation based on

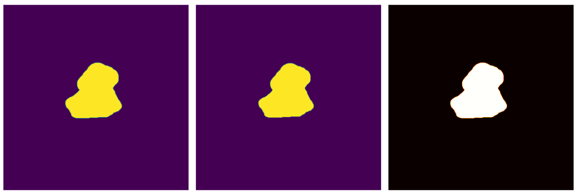

Figure 2: Compactness - in order: original mask, eroded mask and their difference shown in orange

hue, saturation, and value, we found the RGB model to be more diagnostically informative in this context. In particular, the variance within the red, green, and blue channels more effectively captured irregularities in pigmentation, which are often indicative of malignancy. In contrast, the HSV components, particularly brightness (value) and saturation are more susceptible to variations in lighting conditions and image quality, making them less reliable for lesion analysis of data sets with varying lighting conditions and quality.

Furthermore, between mean and variance, we selected variance as our feature of choice. Lesions exhibiting high colour heterogeneity, reflected in greater variance across different regions, are more likely to be malignant.[8] By quantifying this variation, the model is better equipped to detect abnormal pigmentation patterns, which are a recognised warning sign in the clinical assessment of skin cancer.

4.3 Details of implemented features

Worst asymmetry

This feature quantifies the worst-case asymmetry of a skin lesion based on the premise that asymmetry is a key visual indicator of malignancy. A higher asymmetry score is indicative of irregular structure, which is often associated with cancerous lesions.

The process begins by rotating the mask and cutting it, which essentially just removes the blank parts of the mask and ‘zooms in on the lesion’. This step reduces computational overhead by limiting analysis to the relevant area of the image.

Subsequently, each lesion is rotated 30 times, and at each rotation, an asymmetry score is computed by taking the amount of asymmetric pixels for both vertical and horizontal planes.

Specifically, the number of asymmetric pixels is calculated for each axis and then averaged. This value is normalised by dividing it by twice the total number of lesion pixels.

The final feature value is taken as the maximum asymmetry score across all rotations, representing the worst-case asymmetry observed in the lesion.

Compactness

This function calculates the compactness of a skin lesion. The more compact it is the less likely it is for it to be cancerous. The computation begins by determining the area of the lesion from its binary mask. A structuring element, functioning as a morphological operator, is then applied to erode the mask. This operation removes boundary pixels from the lesion, effectively shrinking it and isolating the core region. The difference between the original and eroded masks delineates the perimeter of the lesion. Once the area of the perimeter and the total area are obtained, compactness is calculated using the following formula:

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

This evaluates how closely the shape of the lesion approximates a perfect circle, which is the most compact shape. Lower score means that the lesion is more irregular, and potentially more concerning, lesion boundaries.

RGB variance

This feature calculates the variance of each colour channel of the picture within an image of the lesion. Notably, high variance in the colour blue usually indicates cancerous tissue.

Initially the image is segmented using SLIC (Simple Linear Iterative Clustering) algorithm that clusters pixels in the combined five-dimensional colour and image plane space, thereby efficiently generate compact, nearly uniform superpixels. These superpixels are more efficient to compute compared to analysing each individual pixel, drastically improving performance.

Following segmentation, the mean intensity of each colour channel (red, green, and blue) is computed for each superpixel and stored in a list. These means are subsequently grouped into separate lists corresponding to their respective colour channels.

Finally, the variance of each colour channel is calculated by dividing the sum of the mean values for that particular colour by the total number of segments. This metric serves as an indicator of intra-channel colour variability across the lesion image.

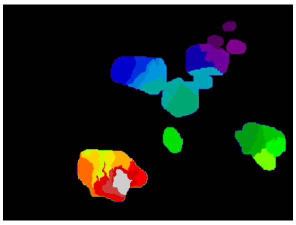

Figure 3: SLIC segmentation of a lesion

4.4 Splitting

An important part of classification is having an equal ratio of cancerous and non-cancerous samples in the training, validation and testing data. Imbalanced data can in the worst case result in a model being trained on either cancerous or non-cancerous images and be unable to recognize the difference. To minimize this, we utilized splitting methods that stratify the data.

4.5 Classifiers

An essential part of this study is the selection of an appropriate classifier. Several models and corresponding parameter configurations were evaluated, including K-Nearest Neighbours (KNN), Random Forest (RF), and Gaussian Process (GP) classifiers.

K-Nearest Neighbours: This model demonstrated stable performance and outputted strong results during the validation phase. However, performance during testing suggested signs of overfitting, as indicated by a decline in generalisation capability.

Random Forest: This classifier provided the best balance between recall and overall accuracy. Although minor overfitting was observed during testing, it was effectively mitigated through parameter optimisation.

Gaussian Process: This model displayed high recall at the expense of overall accuracy, often classifying above 98% of the images as cancerous. For these reasons, we decided to stop using it.

To verify whether our models are overfit, we dedicated 20% of the images to testing and used the remaining 80% for training and validation, again using a stratified 80/20 split, making use of cross-validation in the process. Parameter tuning using GridSearchCV was used to reduce overfitting of our model for Random Forest. For the KNN classifier, minimal variation was observed between validation and test performance; thus, no parameter adjustments were deemed necessary, as the model did not exhibit signs of overfitting.

In order to evaluate our models, we tested multiple metrics, such as overall accuracy, recall, ROC AUC, and F1 score. Among these, recall was considered the most critical, as the primary objective of the system is to minimise false negatives while preserving a reasonable level of overall accuracy, thereby avoiding the misclassification of all images as cancerous.

4.6 Results

The performance of the K-Nearest Neighbours (KNN) and Random Forest (RF) classifiers was assessed using the cleaned dataset[6] of 1474 images. Cross-validation results showed high recall values for true positives (cancerous lesions) for both models, achieving scores above 0.90. The two classifiers performed similarly also on the corresponding accuracy values and F1-scores.

Validation	Recall	Accuracy	F1
KNN:	0.91	0.71	0.82
Random Forest:	0.93	0.72	0.83
Test	Recall	Accuracy	F1
KNN:	0.94	0.73	0.83
Random Forest:	0.96	0.74	0.84

In the final results on the test set both models demonstrated a slight improvement in performance. The marginal increase in performance between cross-validation and test evaluation indicates little to no evidence of overfitting.

KNN	precision	recall	f1-score	support
0	0.42	0.13	0.20	77
1	0.75	0.94	0.83	218
accuracy			0.73	295
macro avg	0.58	0.53	0.52	295
weighted avg	0.67	0.73	0.67	295

Figure 4: Classification reports from KNN on the test data without the hair feature

RFC	precision	recall	f1-score	support
0	0.50	0.12	0.19	77
1	0.75	0.96	0.84	218
accuracy			0.74	295
macro avg	0.63	0.54	0.5	2 295
weighted avg	0.69	0.74	0.67	295

Figure 5: Classification reports from RF on the test data without the hair feature

Further inspection of the classification reports highlights a significant class imbalance in model performance. Both classifiers demonstrated high recall for class 1 (cancerous lesions) with values above 0.90 for both classifiers. However, this comes at the expense of poor performance on class 0 (non-cancerous lesions), with recall values only slightly above 0.10 respectively. This indicates that the model favours classifying lesions as cancerous, resulting in a substantial number of false positives. The macro average metrics further highlights this, with F1-scores of just 0.52, signalling significant disparity in class-wise predictive accuracy.

5 Extended model

The extended model builds upon the baseline by incorporating an additional feature calculating the amount of hair in the image. This augmentation was motivated by the observation that hair can obstruct key visual features of the lesion, potentially affecting the model's classification performance.

5.1 Hair extraction

This function extracts the amount of hair, counted as a ratio of pixels counted as hair compared to the total amount of pixels. A higher ratio indicates greater hair coverage, which can impact the performance of the model. This can be used in the model to account for the loss of accuracy due to hair obstructing the lesion.

The method begins by computing the black-hat transformation of the image, which is the difference between the morphological closing and opening of the image. The closing is the image that is first deteriorated, using some type of noise for example, and then smoothened, the opening is the opposite, where the image is first smoothened and then deteriorated. This is then used to calculate a threshold, which intensifies the hair contours, which when binarized gives the hair pixels as 1 and the rest as 0. Then the binarized threshold is used to measure the area obstructed by hair.

Finally, the hair ratio is computed by dividing the number of detected hair pixels by the total pixel count of the image.

5.2 Annotations

As part of the data preparation process, all 1,474 images were manually annotated to assess the presence of hair. Each image was independently evaluated by multiple annotators using a three-point ordinal scale, where a score of 0 indicated no visible hair, 1 denoted a small amount of hair, and 2 represented a large amount of hair. The final label for each image was set by averaging the individual ratings and rounding the result to the nearest integer, ensuring a consistent and interpretable representation of hair presence.

For interpretability and consistency with the manual annotations, the ratio of hair, extracted by the modal, was transformed to 0, 1 and 2 using thresholds. Images with less than 1% hair coverage were encoded as 0, those with 1–10% coverage as 1, and those with greater than 10% as 2. The resulting discrete values were subsequently compared with the manual annotations to evaluate the agreement between automated and human assessments and to validate the reliability of the automated hair detection method.

5.3 Cohen's kappa

To assess the level of agreement between manual and automated annotations, Cohen's Kappa statistic was employed. Cohen's Kappa is a measure of inter-rater reliability that quantifies the extent to which agreement between two sets of categorical ratings exceeds what would be expected by chance alone[9]. The cohen_kappa_score function from the sklearn library was used for this calculation.

The resulting Kappa value was -0.00497, suggesting no meaningful agreement between the automated and manual annotations. This value implies that the hair extraction function does not perform better than random guessing.

Inspection of individual cases revealed that the algorithm often failed to distinguish between hair and other visual elements such as skin folds or wrinkles. Furthermore, it was notably ineffective at detecting light-coloured hair, which contributed to the poor agreement.

5.4 Hair removal

Hair removal techniques were avoided in this study due to their potential to degrade the integrity of lesion imagery. Previous work demonstrated that such methods can remove parts of the lesion when the hair shares similar colour characteristics. Furthermore, the smoothening process, conducted after the removal, degrades the details of the image, making the process of extracting features harder.

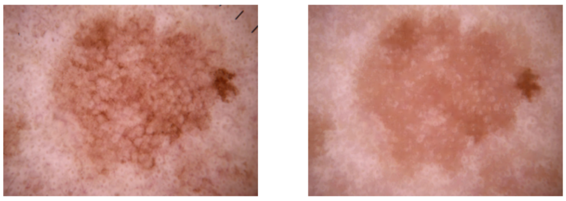

Figure 6: Image before and after hair removal: smoothening damaging the lesion

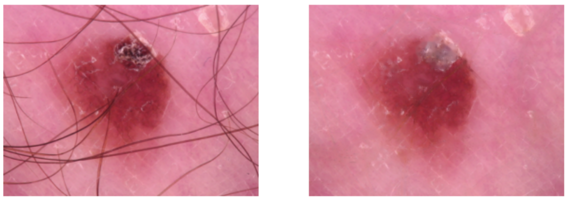

Figure 7: Image before and after hair removal: scar tissue removed

In contrast, the hair extraction method adopted here preserves the lesion in its entirety while estimating the degree to which hair may obstruct the view. By applying this extraction function after lesion masking, we ensured that only hair within the lesion area was considered. Although in many cases hair coverage is minimal, in instances where obstruction is substantial, this feature enables the model to anticipate reduced visibility and adapt its interpretation accordingly.

5.5 Results

The hair feature was added to both classification models (KNN and Random Forest). While KNN exhibited fairly consistent results across different rounds of evaluations, the Random Forest notable instability as illustrated in Figure 8.

The variability in Random Forest, particularly in recall, which ranged from 0.65 to 0.99 for class 1 (cancerous lesion), is of great concern. Moreover, it is inconsistent in classifying class 0 (non-cancerous lesions): in the first figure 40% of benign lesions were correctly identified, while in the second figure this dropped to merely 1%. Such volatility in performance undermines the reliability of the model in clinical settings.

These findings suggest that the inclusion of the hair feature does not improve but in fact compromise the model performance. In comparison to the results of the baseline model, where the only distinction is the absence of the hair feature, both KNN and Random Forest models exhibit diminished classification stability and reduced accuracy. This indicates that the hair feature, as currently extracted and utilised, introduces noise and does not improve model efficiency.

5.6 Classifier Choice

To reduce computational load, we chose to proceed with a single classifier from this point forward. While KNN showed solid results in the baseline model and has been commonly used in previous studies, we deliberately chose to explore a different approach by using the Random Forest classifier. Our goal was to move beyond the conventional methods used in related work and test a model that supports probability threshold tuning. This flexibility is especially important in medical applications, where the consequences of false positives and false negatives are not equal. By adjusting the decision threshold, we can better align the model's predictions with clinical priorities.

KNN	precision	recall	f1-score	support
0	0.20	0.13	0.16	77
1	0.73	0.82	0.77	218
accuracy			0.64	295
macro avg	0.47	0.48	0.47	295
weighted avg	0.59	0.64	0.61	295

RFC	precision	recall	f1-score	support
0	0.29	0.40	0.34	77
1	0.75	0.65	0.70	218
accuracy			0.58	295
macro avg	0.52	0.52	0.52	295
weighted avg	0.63	0.58	0.60	295

KNN	precision	recall	f1-score	support
0	0.36	0.17	0.23	77
1	0.75	0.89	0.82	218
accuracy			0.71	295
macro avg	0.56	0.53	0.52	295
weighted avg	0.65	0.71	0.66	295

RFC	precision	recall	f1-score	support
0	0.25	0.01	0.02	77
1	0.74	0.99	0.84	218
accuracy		0.73	295	
macro avg	0.49	0.50	0.43	295
weighted	0.61	0.73	0.63	295

Figure 8: 2 classification reports from KNN and RF on test data with the hair feature

6 Open question models

6.1 Metadata model

In this model the primary focus was to investigate the extent to which a reliable prediction of cancerous and non-cancerous lesions could be achieved using characteristics and symptomatic information. These parameters are not visually observed on an image, yet provide clinically relevant descriptions of the lesions (e.g. if it hurts). In this version, the algorithm does not have access to the image.

6.1.1 Feature selection

All the metadata information for this project was derived from the PAD-UFES-20 dataset [6]. For this model specifically, we used the metadata.csv provided in the dataset zipfile, which contains additional pieces of information about the patient and their skin lesion (such as if it bled, itched. . .).

To decide on which specific lesion-related features to use, a correlation matrix of the features was created using Spearman correlation. Subsequently, we examined the strength of the association between the features and the diagnosis, as well as their correlations with other relevant parameters. If two features were highly correlated with each other (such as diameter_1 and diameter_2), the one which was more correlated with the diagnoses was used. Value for 'high correlation' was established to be anything above 0.4 correlation. Additionally, we have selected a minimum correlation between diagnosis and feature for it to be used to be 10%. As a result, we proceeded with the following parameters: elevation, bleed, hurt, grew, itch and diameter_2.

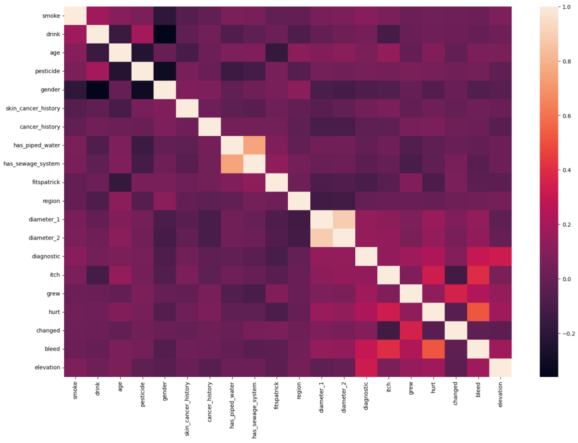

Figure 9: Correlation matrix using Spearman correlation

6.1.2 Building and fine-tuning the model

Subsequently, the selected features were used to train a Random Forest classifier. In order to achieve this efficiently and optimise the percentage of cancerous cases correctly identified (recall), we have decided to fine-tune the parameters using the sklearn function GridSearchCV.

To find the best combination of hyperparameters, we defined a parameter grid and then the algorithm systematically evaluated all possible combinations of them, while incorporating cross-validation in the process. This method allowed us to identify the parameter configuration that delivered the best performance across 5 stratified validation folds, as well as extract the best performance of the predefined scoring metric. For the validation aspect, we have used a Stratified K Fold with 5 splits, to make sure the class ratios of cancerous vs non-cancerous lesions remain the same across all training and validation subsets.
Once the best hyperparameters were found, we retrained the model on the full training set using those settings and evaluated it on the test data.

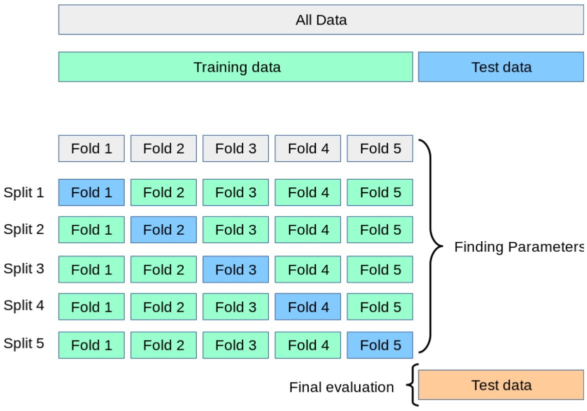

Figure 10: K-Fold cross-validation

6.1.3 Results

Comparing the performance on training and testing data, the Random Forest model produces consistent results, implicating that it did not overfit on the training data. After further inspection, we can observe a high recall for the cancer class (class 1), with over 90% of malignant cases being correctly identified in the screening process. This, however, comes with the cost of a relatively high false positive rate. Only 39% of benign lesions are correctly identified, meaning 61% are misclassified as malignant.

This trade-off is largely driven by class imbalance, where the model favours class 1 (cancerous). Nevertheless, the metadata model still presents a substantial improvement over the baseline model, where the false positive rate was significantly higher and benign cases were almost entirely overlooked.

Validation	precision	recall	f1-score	support
0	0.72	0.37	0.49	308
1	0.81	0.95	0.87	871
accuracy			0.80	1179
macro avg	0.76	0.66	0.68	1179
weighted avg	0.79	0.80	0.77	1179

Test	precision	recall	f1-score	support
0	0.61	0.39	0.48	77
1	0.81	0.91	0.86	218
accuracy			0.78	295
macro avg	0.71	0.65	0.67	295
weighted avg	0.76	0.78	0.76	295

Figure 11: Classification report from metadata model

6.2 Combined model

In the final stage, we explored how a combination of the ABC features and features from the metadata[6] model performed when combined.

The reason behind this approach was that the two feature sets capture complementary types of information - visual characteristics, patient-reported symptoms and physical observations that cannot be captured through image data alone. By combining both types of features, the aim was to create a more comprehensive model.

We began with the ABC model and incrementally added metadata[6] features, prioritising them by predictive value. At each step, we assessed model performance to identify the optimal number and combination of metadata features that would enhance accuracy without introducing unnecessary complexity or noise.

As with previous models, we have used the Random Forest classifier. To ensure consistency in evaluation, we used stratified 5-fold cross-validation to maintain class balance and performed hyperparameter tuning using GridSearchCV. Our primary evaluation metric remained recall, reflecting our emphasis on maximising the detection of cancerous cases.

Once the optimal set of metadata features and hyperparameters was determined, we trained the final model on the combined dataset and conducted a full evaluation.

6.2.1 Results

The incorporation of metadata[6]-derived features contributed significantly to improving the performance of the baseline model and slightly improving the metadata model, particularly regarding the improvement of the correct identification of benign lesions. After an iterative process of testing various combinations of the metadata features, the most optimal results were achieved using the following subset: ABC features (worst asymmetry, compactness and rgb variance), grew, bleed, elevation and itch.

Using this feature set, the Random Forest model resulted in maintaining a high recall of 96% for malignant lesions. Whilst some misclassification of benign lesions as cancerous remained, the frequency of false positives substantially reduced in comparison to the baseline model.

Validation	precision	recall	f1-score	support
0	0.79	0.34	0.47	307
1	0.81	0.97	0.88	871
accuracy			0.80	1178
macro avg	0.80	0.65	0.67	1178
weighted avg	0.80	0.80	0.77	1178

Test	precision	recall	f1-score	support
0	0.80	0.43	0.56	77
1	0.83	0.96	0.89	218
accuracy			0.82	295
macro avg	0.82	0.70	0.72	295
weighted avg	0.82	0.82	0.80	295

Figure 12: Classification reports from combined model

7 Limitations

The dataset used to train and evaluate our models (PAD-UFES-20 - source) presents multiple limitations, impacting different aspects of the research, including feature extraction, segmentation, model training and generalisability of the model.

To start, images of the skin lesions were captured using smartphones under uncontrolled conditions. This variability presents significantly in image quality, resulting in significant differences in resolution and lighting conditions, which negatively affect the extraction of the ABC features in the base model. Namely, the border irregularity extraction is not as precise on low-resolution images, especially if the image does not come with a precise image mask, and lighting differences distort colour perception. Similarly, SLIC segmentation used to map colour variance in different segments of the feature can have a problem segmenting the lesion properly due to noise, and if the colouring of images is strongly affected by the lighting tone.

Moreover, the dataset exhibits a notable imbalance in skin tone representation, with individuals of lighter skin tones being disproportionately represented. Not including all skin tones in both training and testing results in a mode, which is only optimised for people with similar features to the patients included in the training data. As mentioned before, skin colour is also linked to the performance of the hair extraction feature. To be exact, if there is a low contrast between the skin and the hair (for example, white skin and white/blond hair), the algorithm has a harder time differentiating between the two and might have an incorrect rating of how much hair there is on the lesion.

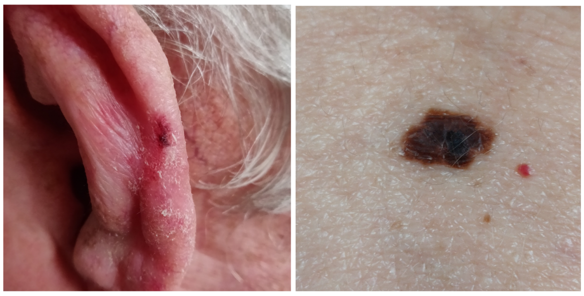

Figure 13: Comparison of quality of images

Lastly, there are some limitations on feature extraction which stem from the automatic feature retrieval. In an ideal world, each picture would be treated individually (eg. fine-tuning thresholds and extraction criteria), but for effectiveness, this is not possible. For example, we use only 1 threshold for creating a mask for images that do not have a mask provided by default, which can mean that there is a potential for lower accuracy due to the generated masks not being of the same quality as the given ones.

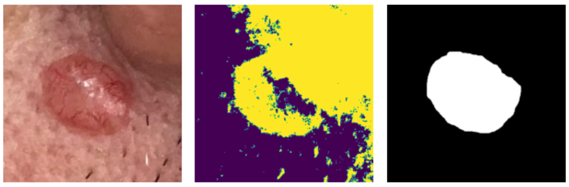

Figure 14: Automatic vs. provided mask

For future work/model optimisation, it would be beneficial to have the images made in a consistent method, with the same camera, in the same lighting and from the same distance from the skin. Moreover, in the case of more age and skin tone representative data would be possible to gather, it would be important to optimise the model to perform consistently across all races and genders.

8 Conclusion

This study aimed to explore the effectiveness of different classification methods for skin lesions, with the primary focus on identifying true positive (cancerous) lesions correctly. Through systematic experimentation with K-Nearest Neighbours (KNN) and Random Forest (RF) classifiers on a cleaned dataset of 1474 images of skin lesions, the baseline model exhibited high sensitivity in identifying class 1 (malignant lesions) achieving recall scores exceeding 90%. However, this came at the cost of a high false positive rate (benign lesions classified as cancerous). This raises concerns regarding the clinical applicability of the baseline model.

Thereafter, an experiment incorporating an additional feature representing the presence of hair in the images was conducted. This resulted in diminished performance both regarding accuracy and classification stability. The hair feature particularly impacted the Random Forest classifier's predictions, leading to unreliable outputs, highlighting the feature's insufficiency to contribute to worthwhile diagnostic information. These results highlight the importance of the feature relevance and quality for model generalisability.

To optimise computational resources and leverage prior findings, the Random Forest classifier was selected for further development.

A metadata-driven model was developed by integrating clinically relevant, but visually undetectable features, to potentially overcome some limitations of the baseline model. This model maintained the high recall values, similarly to the baseline model, over 90% on class 1 (cancerous lesions), meanwhile improved the performance in negative recall, with about 39% of non-cancerous lesions being diagnosed as malignant. This shows that the metadata model represented an improvement compared to the baseline model. Next step was to try combining the two to inspect the results.

The final, combined model, integrated a selected subset of the metadata[6] features, selected by manual testing, alongside the ABC features. This model achieved the most balanced performance, maintaining a high recall for true positives (malignant lesions) whilst further reducing false positive rates compared to all other models. The improved preciseness without compromising sensitivity improves the model's suitability for early, at home screening.

In summary, this study underscores the importance of feature selection and the integration of relevant clinical data, other than images of the lesion, in improving machine learning classification for skin lesions. Overall, these findings highlight the importance of combining image-derived features with patient metadata to improve diagnostic accuracy. While challenges related to class-imbalance are not overcome, the combined model presents a promising step towards clinical implementation, which can be explored in future work.

9 Discussion and Future Work

In the future, to develop this model further, it would be interesting to investigate the importance of lifestyle-related metadata, such as smoking habits, alcohol consumption, and living conditions. Including and analysing these elements may help reveal indirect risk indicators and improve early detection, especially in borderline or ambiguous cases.

Besides that, it would be interesting to assess whether the models developed in this study perform equally well across different types of skin cancer. In the current version, we grouped Melanoma (MEL), Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC) into a category which we called cancer. While this helped simplify the classification task and increase the number of positive examples, it also masked subtle distinctions between these cancer types. But since they have different symptoms, maybe for a specific type it would not be optimised enough.

Another important direction would be to assess whether the models developed in this study perform equally well across different types of skin cancer. In the current version, we grouped Melanoma (MEL), Basal Cell Carcinoma (BCC), and Squamous Cell Carcinoma (SCC) under a single "cancer" label. While this helped simplify the classification task and increase the number of positive examples, it also masked subtle distinctions between these cancer types. Since these cancers can differ significantly in appearance, progression, and symptoms, future work could develop models that classify them into separate categories.

References

[1] J.B. Heistein, U. Acharya, and S.K.R. Mukkamalla. *Malignant Melanoma*. Updated February 17, 2024. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan–. 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK470409/ (visited on 05/28/2025).

[2] F. Nachbar et al. "The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions". In: *Journal of the American Academy of Dermatology* 30.4 (1994), pp. 551–559. DOI: 10.1016/s0190-9622(94)70061-3.

[3] Robert H Johr. "Dermoscopy: alternative melanocytic algorithms—the ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist". In: *Clinics in Dermatology* 20.3 (2002), pp. 240–247. DOI: 10.1016/S0738-081X(02)00236-5. URL: https://doi.org/10.1016/S0738-081X(02)00236-5.

[4] Sergey Demyanov et al. "Classification of dermoscopy patterns using deep convolutional neural networks". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. 2016, pp. 364–368. DOI: 10.1109/ISBI.2016.7493284.

[5] Rebecca Moussa et al. "Computer-aided detection of Melanoma using geometric features". In: *2016 3rd Middle East Conference on Biomedical Engineering (MECBME)*. 2016, pp. 125–128. DOI: 10.1109/MECBME.2016.7745423.

[6] Andre G. C. Pacheco et al. "PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones". Version 1. In: (2020). CC BY 4.0. DOI: 10.17632/zr7vgbcyr2.1. URL: https://data.mendeley.com/datasets/zr7vgbcyr2/1.

[7] NHS. *Symptoms - Non-melanoma skin cancer*. Accessed: 2025-05-28. 2023. URL: https://www.nhs.uk/conditions/non-melanoma-skin-cancer/symptoms/.

[8] Cleveland Clinic. *Skin Self-Exam*. Accessed: 2025-05-28. 2023. URL: https://my.clevelandclinic.org/health/diagnostics/8648-skin-self-exam.

[9] Mary L. McHugh. "Interrater Reliability: The Kappa Statistic". In: *Biochemia Medica (Zagreb)* 22.3 (2012), pp. 276–282.