Does Incorporating Additional Features Alongside Traditional ABC Criteria Significantly Improve The Precision of Skin Lesion Classification Compared To Using ABC Features Alone?

Patrycja Zdyb pazd@itu.dk
O'Neal Okutu onok@itu.dk
Marta Zuzanna Richert mazr@itu.dk
Aryan Suraj Anvekar aanv@itu.dk
Nicolai Alonso Kofoed niak@itu.dk

Abstract

Skin cancer is among the most common malignancies worldwide, with basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma being the primary types. To assess skin lesions, traditional dermoscopic evaluation relies heavily on the ABC criteria - Asymmetry, Border irregularity, and Color variation. In this project, we investigate whether incorporating additional image features beyond the ABC criteria can significantly enhance the precision of classifying skin lesions as benign or malignant. Our analysis with machine learning classifiers - including Random Forest and K-Nearest Neighbors - demonstrates that models leveraging both ABC and additional features achieve improved classification accuracy compared to models using ABC features alone. These findings highlight the potential benefits of integrating a broader feature set for more precise early screening of skin cancer, especially in resource-limited settings.

1 Introduction

This project aims to evaluate whether incorporating complementary features alongside the traditional ABC criteria can significantly enhance the performance of automated skin lesion classification. In particular, features such as Blue-White Veil Detection, Haralick Texture Features, Hair Count, and Hair Removal preprocessing are hypothesized to provide richer diagnostic information beyond the conventional asymmetry, border irregularity, and color variation metrics. Experiments were conducted using the PAD-UFES-20 dataset, which contains approximately 2,000 dermoscopic images representing a variety of lesion types, including malignant cases (basal cell carcinoma, squamous cell carcinoma, and melanoma) as well as benign or precancerous conditions (actinic keratosis, nevus, and seborrheic keratosis).

1.1 Dataset

For this project, we used a public dataset called PAD-UFES-20, which includes 2,298 images of six skin lesions. These images come from 1,373 patients, and there are a total of 1,641 unique lesions. Some patients have more than one image of the same lesion, and we made sure to take that into account when training our model. We decided to work with all the provided images. In addition to the images, the PAD-UFES-20 dataset includes a file called metadata.csv, which contains detailed information for each image. This includes 26 variables, such as patient background and specific measurements related to the lesion. The table below shows the types of lesions, their abbreviations, and how they are classified.

Lesion Type	Abbrevation	Classification
Basal Cell Carcinoma	BCC	Skin cancer
Squamous Cell Carcinoma	SCC	Skin cancer
Melanoma	MEL	Skin cancer
Actinic Keratosis	ACK	Skin disease
Seborrheic Keratosis	SEK	Skin disease
Nevus	NEV	Skin disease

Table 1: Types of skin lesions and their abbreviations

The PAD-UFES-20 dataset includes two main types of skin lesions: skin cancer (malignant: BCC, SCC, MEL) and skin diseases (benign: ACK, SEK, NEV). Skin cancers are often caused by UV exposure, with Melanoma (MEL) being the most aggressive. Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC) are less severe but still need treatment. Among the benign lesions, Actinic Keratosis (ACK) can progress

to SCC if untreated, while Seborrheic Keratosis (SEK) and Nevus (NEV) are generally harmless.

2 Application of Methods

2.1 Initial Process of Subset Selection

In order to begin the process of method testing, we deemed it necessary to choose an appropriate subset size of images to analyse. Using a small subset size would sacrifice the accuracy of our analysis and pose a significant risk of overfitting. It wouldn't allow us to fully analyse all the different patterns and variations in features, as the model would learn patterns specific to the sample data however, fail to recognize patterns in features in unseen cases or cases with anomalies. Hence, we decided to use all the 2,298 images and we decided to use the pre-existing material of lesion masks for the images already given to us by our teachers.

2.2 Feature Extraction

To investigate whether incorporating additional features alongside traditional ABC features enhances the precision of skin lesion classification, we structured our study around a comparative framework. The baseline model was constructed using the core ABC features - Asymmetry, Border, and Color - which are well-established indicators in dermatological diagnostics. These features served as our control group, enabling us to benchmark the performance of classifiers trained solely on the baseline features. To explore how adding additional features would affect the model's performance, we developed an extended baseline model with the following additional features:

- Blue white veil
- GLCM (Gray-Level Co-occurrence Matrix) and Haralick
- hair count
- hair removal preprocessing

These features were selected based on their relevance to skin lesion morphology and their capacity to capture additional structural and textural patterns that may not be fully represented by ABC features alone. We designed and implemented scripts to extract both the baseline and extended features from our 2,298 images. These feature sets were then used to train and test classifier models, to determine whether the inclusion of additional features would lead to measurable improvements in classification accuracy/precision in predicting malignant and benign lesions. This methodology enabled us to directly assess and compare how well the models performed in identifying malignant and benign lesions, based on different sets of features.

2.2.1 Asymmetry

Asymmetry is an important visual feature in dermatology, especially in the diagnosis of malignant melanomas, which often present with irregular and asymmetric shapes. To evaluate the asymmetry of skin lesions, we developed a method that assigns each lesion a discrete asymmetry score based on its binary mask.

The process begins by isolating the lesion from the mask image using a cropping function that removes excess background. Each cropped mask is then padded into a square shape so that it remains centered and fully visible during rotation.

The mask is rotated into 8 different angles (every 45 degrees), and at each angle, a comparison is made between the two halves of the lesion: top vs. bottom (horizontal axis) and left vs. right (vertical axis). These halves are flipped and then compared to one another using logical operations to assess shape similarity.

Specifically, we use XOR (exclusive OR) to highlight differences between the halves - asymmetric regions where one half has pixels the other doesn't. The more pixels that differ, the higher the asymmetry. These XOR comparisons are used internally to assign a score from 1 to 3:

- A score of 1 means the lesion is largely symmetric on both axes.
- A score of 2 indicates symmetry on one axis but not the other.
- A score of 3 means the lesion is asymmetric on both axes.

For each lesion, scores are calculated across all rotated angles. From these, we record two key metrics:

- Best Asymmetry Score: the lowest (most symmetric) score across all rotations.
- Mean Asymmetry Score: the average score across all rotations, providing a general assessment of the lesion's symmetry.

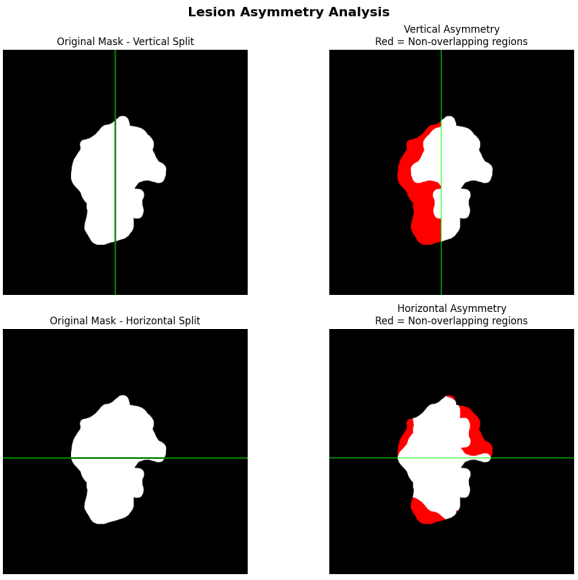


Figure 1: Lesion asymmetry analysis for mask PAT_20_29_850 (best asymmetry score = 3 and mean asymmetry score = 3)

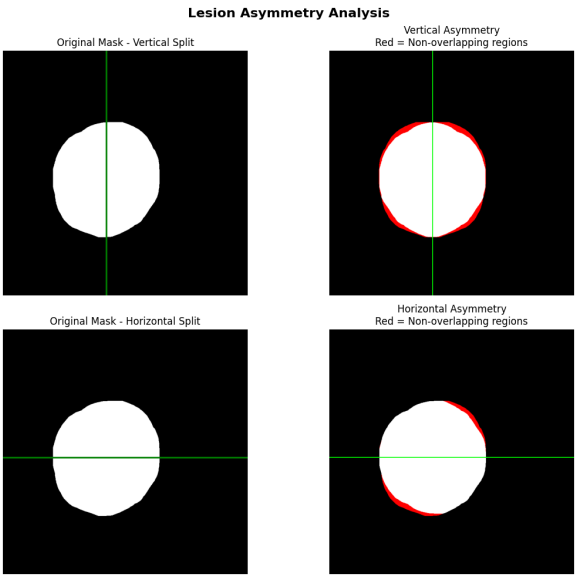


Figure 2: Lesion asymmetry analysis for mask PAT_39_55_233 (best asymmetry score = 3 and mean asymmetry score = 3)

These scores give us a quantifiable and rotation-independent measure of asymmetry, which is useful for distinguishing between benign and potentially malignant lesions in subsequent classification tasks. While asymmetry alone doesn't confirm cancer, it's a strong visual indicator when combined with other features like color.

2.2.2 Border

Feature B also played a central role in our approach to classifying skin lesions. In our project feature B represents the average circularity of segmented regions within lesion masks. Circularity is a shape descriptor that measures how close an object is to a perfect circle, mathematically defined as:

$$\text{Circularity} = \frac{4\pi \times \text{Area}}{\text{Perimeter}^2}$$

This feature was calculated for each lesion mask using region properties, and an average was taken across all valid regions . Importantly, we removed background noise by excluding very small regions (area ¡ 10 pixels), which enhanced the quality and removed the noise.

Why Circularity Matters

Clinically, malignant lesions often present with irregular, less circular shapes. We then quantified this irregularity numerically. Smooth, rounded borders yield higher circularity values, while irregular or jagged edges produce lower scores. By focusing on this geometric feature, we aimed to provide a complementary perspective to other shape and texture-based descriptors, enhancing the overall robustness of our classification approach.

2.2.3 Color

Color plays a crucial role in distinguishing malignant melanomas from benign or precancerous lesions, as specific colors (white, red, light brown, dark brown, blue-gray, and black) are associated with underlying biological markers like melanin distribution, vascular changes, and regression. The primary objective was to extract meaningful color features from within the lesion area only, excluding surrounding healthy skin, to support machine learning classification of skin lesion types.

Lesion Masking and Preprocessing

Each image was paired with a binary mask file that indicates the lesion area. We used these masks to isolate the lesion by applying a bitwise and operation, producing a masked image that retained only the pixels inside the lesion boundary. To ensure consistency, the masks were resized to match

their corresponding images when necessary.

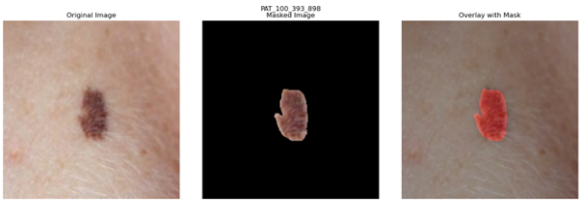


Figure 3: Image: PAT_39_55_233

Color Feature Extraction

After extracting the masked lesion region, several types of color features were computed:

1. Basic Statistical Features

   We computed the mean, median, and standard deviation of the pixel intensities across the RGB channels for all non-black pixels in the lesion. These statistical measures provide a compact summary of the lesion's overall color composition.

2. Color Variation and Diversity

   Using MiniBatchKMeans clustering (k=5), we identified the dominant colors within the lesion. The color variation was calculated as the average pairwise Euclidean distance between these cluster centers, indicating how diverse the color distribution is. Additionally, we computed a color diversity score, counting how many of the cluster centers were at least 30 units apart in color space - representing visibly distinct color groups.

3. Color Entropy

   To further quantify the complexity of the lesion's color composition, we computed a normalized 3D RGB histogram (8 bins per channel) and derived its Shannon entropy, reflecting how evenly color intensities are distributed within the lesion.

4. Color Asymmetry

   The lesion image was split vertically, and the mean RGB values of the left and right halves were compared using the Euclidean distance. This color asymmetry score captures visual irregularities that may suggest malignancy.

5. Blue Dominance Dark Ratio

   We calculated the proportion of pixels where the blue channel had the highest intensity (blue dominance), and the proportion of very dark pixels (intensity ¡ 50), both of which can indicate particular lesion types or depth of pigmentation.

6. Highly Saturated Regions

   The masked image was converted to HSV color space, and the proportion of pixels with saturation values greater than 150 was computed. This high saturation ratio can highlight the presence of intense color signals often associated with abnormal lesions.

7. Border Contrast

   The contrast between lesion edge colors and internal lesion colors was measured using the mask's morphological gradient to define the border. This border contrast score can indicate sharp pigmentation changes along lesion edges - common in melanoma.

Specific Color Presence

We also implemented a rule-based system to detect the presence of six clinically relevant colors within each lesion: white, red, light brown, dark brown, blue-green (proxy for blue-gray), and black. For each color, we defined a specific RGB range and applied a thresholding operation using cv2.inRange() to check whether that color was present within the lesion. Each detected color was assigned a binary value (1 if present, 0 otherwise), and their sum was recorded as a color count feature.

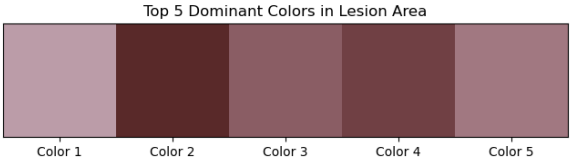


Figure 4: Dominant Color Palette Extracted from Lesion Area of PAT_39_55_233.png

According to established diagnostic criteria, melanoma lesions typically contain at least three distinct colors. Our system mirrors this by outputting a total color presence score (0–6), which can be optionally weighted by 0.5 as in clinical scoring systems, where a maximum score of 3 is possible.

Color	Clinical Interpretation
White	Regression structures (fibrosis, scarring)
Red	Inflammation, neovascularization
Light Brown	Superficial melanin deposits (epidermis)
Dark Brown	Deeper melanin (upper dermis)
Blue-Green	Dermal melanin (papillary dermis)
Black	Dense melanin in upper epidermal layers

Table 2: Color Metrics Summary

2.2.4 White Blue Veil (BWV)

What is white blue veil?

The white blue veil, commonly referred to as blue white veil in dermatology, is a distinctive dermoscopic feature observed during the examination of pigmented skin lesions. It is characterized by a confluent blue pigmentation overlaid with a white, “ground glass” haze, which is a visual phenomenon seen on certain pigmented lesions where the appearance is described as a whitish translucent area that overlays a blue or bluish background.

Why use BWV and its clinical significance?

We chose to add the blue white veil as a feature part of our extended baseline as MEL lesions can often come in shades of blue. Using BWV will help enhance the accuracy of the classifier models by enabling the model to better distinguish malignant lesions from benign ones based on characteristic pigmentation patterns. The clinical significance of the blue-white veil (BWV) in dermoscopic images lies in its strong association with malignant melanoma, making it a valuable diagnostic marker in the early detection and differentiation of skin cancers. Strong presence of BWV percentage is associated with invasive/aggressive melanomas (one that has penetrated deeper into the skin) and signals vertical growth of malignant cells into the dermis and it is a warning sign that the lesion might be a serious melanoma.

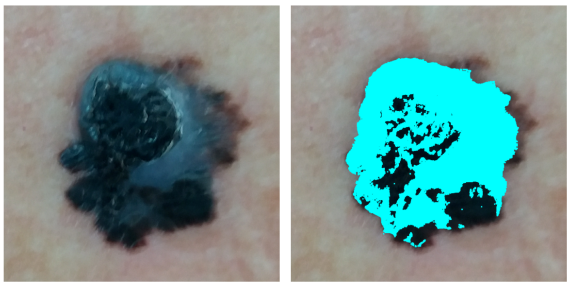


Figure 5: Applying white blue veil method for image PAT_340_714_314

How was BWV applied?

1. Image Loading and Preprocessing:

   Each lesion image is loaded and converted from BGR to RGB format. It is then flattened into a one-dimensional pixel array for analysis. To isolate the lesion, completely black pixels (background or non-lesion areas) are excluded.

2. Blue Dominance Calculation:

   To detect the hallmark blue hue of BWV, the proportion of pixels where the blue channel intensity exceeds both red and green is calculated. This ”blue dominance” ratio quantifies the extent of blue coloration within the lesion.

3. Color Range Definition and Detection:

   Predefined RGB ranges for clinically relevant colors—especially white and blue-green associated with BWV—are used to create binary masks. Each mask flags pixels that fall within its specific color range.

4. Application of Color Masks and Color Presence Determination:

   These masks determine the presence or absence of each color by checking for any matching pixels. The presence of BWV-related colors like white and blue-green is recorded as binary indicators for analysis.

5. Feature Aggregation and Quantification:

   A feature vector is generated for each lesion, combining the blue dominance ratio, binary color presence flags (e.g., white, blue-green), and additional statistical metrics such as color variation, diversity, and asymmetry.

6. Quantitative Veil Feature Output:

   The blue dominance ratio (ranging from 0 to 1) serves as a continuous metric of BWV presence. A higher ratio indicates a stronger presence of BWV, which correlates with an increased risk of malignancy.

2.2.5 GLCM And Haralick

In our extended baseline model, we incorporated Haralick texture features, which are derived from the Gray-Level Co-Occurrence Matrix (GLCM). These features allowed us to capture subtle textural properties of skin lesions, particularly relevant when color or shape alone are insufficient for accurate classification.

Why We Focused on Haralick, Not GLCM Directly

Although the GLCM is a foundational component in texture analysis, we did not include the raw GLCM matrices or their direct statistics as features in our model. Instead, we computed higher-level statistical summaries from the GLCM, the Haralick features. This decision was both practical and easier to interpret, as raw GLCMs are high-dimensional and difficult to use directly, while Haralick features provide meaningful numerical summaries. While GLCM captures the frequency of pixel intensity pairs (i,j)at a specific distance d and angle , it results in a high-dimensional matrix:

GLCM(i,j) = count of pixel pairs with intensities i and j

Instead of using these matrices directly, we computed Haralick features, these are:

- **Contrast** (measures local intensity variation)
- **Correlation** (measures linear dependence of intensity values)
- **Energy** (also known as Angular Second Moment)
- **Homogeneity** (measures closeness of distribution to the diagonal)
- **Entropy** (measures texture randomness)
- **Dissimilarity** (absolute intensity difference)
- **ASM** (Angular Second Moment) – equivalent to **Energy** but calculated separately in some implementations

Each of these properties was computed over four angles =0,45,90,135 and then averaged to ensure rotation invariance. Haralick features proved especially powerful in distinguishing fine-grained irregular textures, often seen in malignant lesions.

2.2.6 Haircount

In dermoscopic imaging, body hair often interferes with the accurate assessment of skin lesions by obscuring critical diagnostic features such as lesion borders, color patterns, and texture. These elements are vital for detecting melanoma and other skin cancers. Hair not only complicates visual evaluation by dermatologists but also introduces noise that can reduce the accuracy of automated algorithms used in computer-aided diagnosis. To address this issue, our project includes a dedicated hair detection step within the preprocessing pipeline. Rather than removing hair directly, this process is designed to quantify its presence. A morphological Blackhat filter is applied to highlight dark, thin structures (like hair) within the lesion area, which has been isolated using a segmentation mask. Detected regions are then filtered by shape and size to count only those structures likely to be hairs. This count helps us flag images that may require additional cleaning before undergoing further feature extraction or modeling.

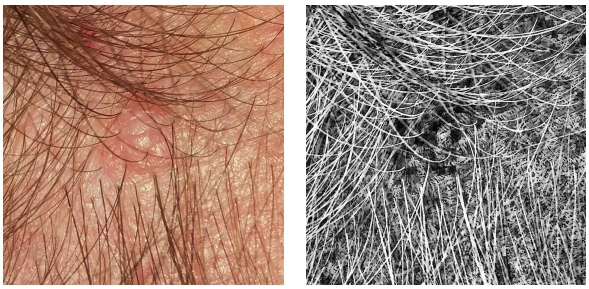


Figure 6: Applied Blackhat filter for image PAT_21_37_965. Left: original lesion area; Right: hair-enhanced visualization.

This step plays an important role in maintaining data quality and improving model performance. Analysis of our dataset revealed that most images had a low hair count (typically 0–3 hairs), indicating minimal interference. However, a subset contained significantly more, with some images showing over 10 hairs. By identifying these cases early, we can apply targeted preprocessing only where needed - ensuring that lesion features are accurately captured and not distorted by hair interference. In summary, integrating hair detection into the preprocessing workflow enhances the

overall reliability of the skin cancer classification pipeline. By proactively identifying hair-heavy images, we are able to safeguard against one of the most common visual artifacts in dermoscopy, thereby improving the accuracy and robustness of both manual assessments and machine learning–based analyses.

3 Classifiers

3.1 Performances for Baseline and Extended

In this project, multiple machine learning classifiers were considered to tackle the classification task effectively in both our Normal and Extended baseline as we aim to measure the importance of additional features to lesion classification. Among the various algorithms explored, two models were selected for detailed evaluation and comparison: **Random Forest** and **K-Nearest Neighbors (KNN)**. The performance of both classifiers was evaluated using the following metrics:

Precision: Precision tells us how many of the instances that the model predicted as positive were correct.

Recall: Recall measures how well the model captures all the actual positive cases.

F1 Score: The F1 score is a combined measure that balances both precision and recall.

3.2 Baseline Model Performance: Random Forest vs. KNN:

In the baseline classification task, two machine learning models were evaluated: Random Forest and K-Nearest Neighbors (KNN). After tuning the Random Forest model, it showed significantly better performance than KNN across all key evaluation metrics.

Random Forest – Test Set Performance:

- **Accuracy**: 65.00%
- **Precision**: 64.86%
- **Recall**: 65.00%
- **F1 Score**: 64.89%

These metrics indicate that the Random Forest model maintains a balanced trade-off between correctly identifying positive and negative cases. The weighted average F1 score of **64.89%** highlights its consistent performance across both classes.

K-Nearest Neighbors– Test Set Performance:

- **Accuracy**: 57.89%
- **Precision**: 57.29%
- **Recall**: 57.89%
- **F1 Score**: 57.06%

Compared to Random Forest, KNN performed worse in all metrics, suggesting it may not capture the underlying patterns of the data as effectively in this baseline setup.

3.3 Extended Baseline Model Performance: Random Forest vs. KNN:

In the extended baseline classification task, the two machine learning models Random Forest and K-Nearest Neighbors (KNN) were again evaluated. With additional data refinements and an expanded feature set, the tuned Random Forest model delivered notably stronger performance across all key evaluation metrics compared to KNN once again.

Random Forest – Overall Test Set Performance

- **Accuracy**: 68.50%
- **Precision**: 68.53%
- **Recall**: 68.50%
- **F1 Score**: 68.14%

These results reflect a well-balanced model with consistent accuracy and generalization across both classes.

K-Nearest Neighbors– Overall Test Set Performance:

- **Accuracy**: 57.48%
- **Precision**: 57.26%
- **Recall**: 57.48%
- **F1 Score**: 57.29%

KNN continued to underperform relative to the Random Forest model, with lower scores in every evaluation metric. This suggests that KNN may struggle with the extended feature space or is less capable of capturing complex relationships that occur in the dataset. It even performed worse with the extended baseline but this could be due to reasons like: The new features do not strongly increase separation between classes and may instead just introduce noise. Also KNN has no built-in way to prioritize useful features, unlike models like Random Forest which do that inherently and could therefore be the reason for these values. parskip

3.4 3.4 Stratified K-Fold Cross Validation:

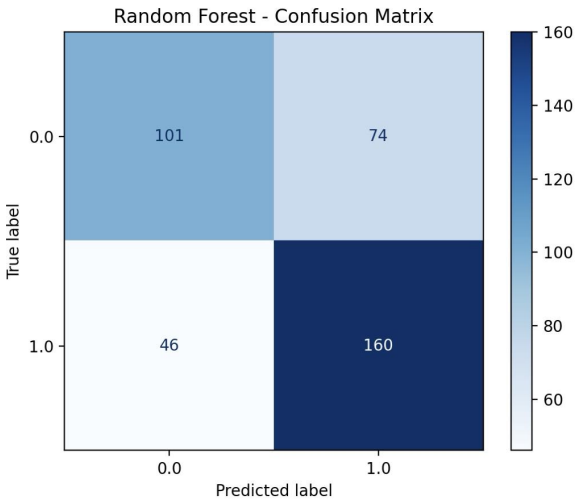
Since our dataset is imbalanced (approx: **Benign** = 0.45, **Malignant** = 0.55), we used Stratified K-Fold Cross-Validation with 5 folds to ensure a fair evaluation of our models. Unlike standard K-Fold cross-validation, which splits the data into K equal parts randomly, Stratified K-Fold maintains the original class distribution within each fold. This means that every subset used for training and validation contained roughly the same proportion of each class as the full dataset. This method splits the data into 5 equal folds, and in each iteration, 4 folds are used for training while the remaining fold is used for validation; this process is repeated 5 times so that each fold serves as the validation set once, and the performance metrics are then averaged across all iterations to provide a stronger estimate.

4 Results

4.1 Interpretation and Analysis:

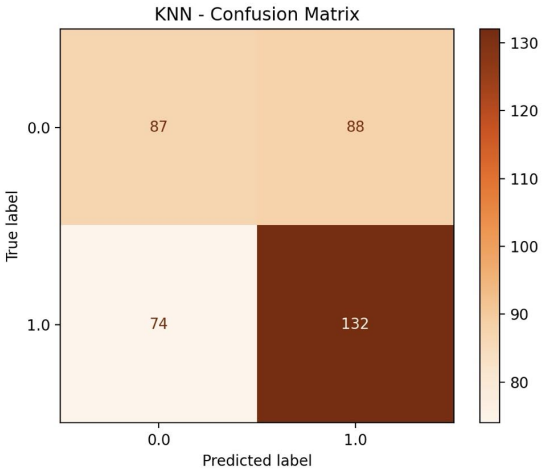
With the help of the classifiers, we were able to generate **Confusion matrices**, **ROC Curves**, **Precision Recall graphs** and **Decision Trees** for the baseline to further aid in our interpretation of the results given to us and the accuracy of the models we used.

1. Random Forest Confusion Matrix (class=0)



The Random Forest classifier correctly identified **101** instances of Benign (0) TP, **160** cases of Malignant (1) TN. The model misinterpreted **74** instances as Malignant (FN) and **46** as Benign (FP) as seen in the confusion matrix above.

2. KNN Confusion Matrix



For the KNN model, **87** TP, **132** TN, **88** FN and **74** FP were identified. Comparing its ratio to that of Random Forest's, it is visible that it's not as effective as RF in classifying the data, hence our reason to go with Random Forest as our Main Classifier.

3. Decision Tree

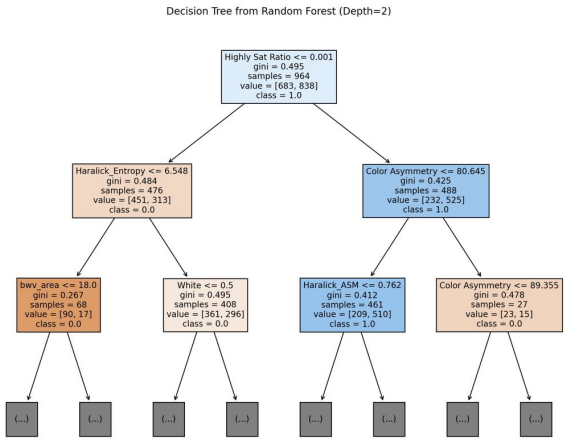


Figure 7: Decision tree visualization limited to depth 2.

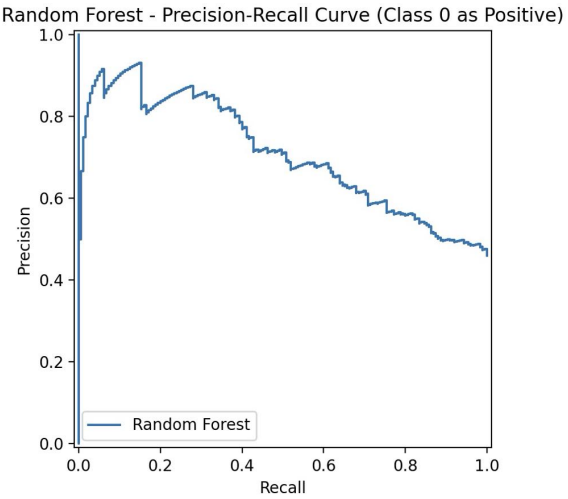
This visualization represents a single decision tree extracted from the Random Forest model, limited to a depth of 2 for interpretability. The root node splits on the Highly Sat Ratio (which detects highly saturated areas potentially linked to abnormal pigmentation), highlighting its relevance as a first-level indicator in classification.

The tree then continues to split based on features that help distinguish between classes: Haralick_Entropy and bwv_area appear more associated with class 0 predictions, suggesting that lower texture entropy and specific area characteristics are useful for identifying this class.

Color Asymmetry, Haralick_ASM, and White ratio guide predictions toward class 1, indicating that differences in color distribution, texture regularity, and lightness are informative for identifying this class.

The Gini impurity values at each node represent how mixed the classes are - lower values show higher node purity, meaning the model is more confident in those splits.

4. Precision Recall (class = 0)

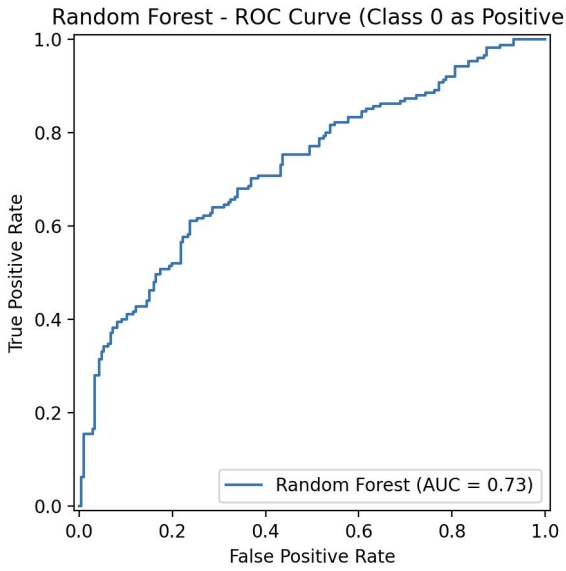


This visualization shows the Precision-Recall Curve for the Random Forest model, with class 0 treated as the positive class. The curve begins with high precision at very low recall, meaning the model's earliest predictions for class 0 are highly reliable.

As recall increases, precision gradually declines, indicating that as the model tries to identify more positives, it also introduces more false positives. This reflects a typical trade-off between being conservative and precise versus being more inclusive but riskier.

The overall shape of the curve suggests that the model maintains strong performance across a range of thresholds. It starts off very confident in a small number of predictions and gradually broadens its detection while still keeping a relatively high precision — a desirable trait in imbalanced or high-stakes classification settings.

5. ROC Curve



This visualization shows the ROC Curve for the Random Forest model, where class 0 is treated as the positive class. The curve plots the True Positive Rate against the False Positive Rate across different threshold settings, revealing how well the model distinguishes between the two classes. The AUC (Area Under the Curve) is 0.73, meaning the model has a **73%** probability of correctly distinguishing a class 0 sample from a class 1 sample. Since a value of 0.5 indicates random guessing and 1.0 represents perfect separation, this result suggests the model performs better than chance and shows solid classification ability. The upward curve and AUC indicate that the Random Forest is able to capture patterns useful for separating the classes, although further improvement may be possible through feature enhancement, threshold tuning, or trying alternative models.

5 Evaluation and Limitations

5.1 Evaluation of colour detection

While the automated color analysis pipeline provided a systematic approach for quantifying lesion color characteristics, several limitations were observed throughout the process. One of the primary challenges was the discrepancy between detected colors and the colors visually perceived in dermoscopic images.

The detection of specific colors such as red or blue-gray was sometimes inaccurate due to overlapping RGB values with other colors or subtle shades that fell outside the predefined thresholds. In such cases, certain clinically relevant colors were either missed or misclassified. For example, lesions with reddish inflammation might not be flagged as "red" if the intensity did not fall within the specified bounds, despite being visually apparent. Some limitations in the color analysis stemmed from variations in brightness and reflections within the dermoscopic images, which sometimes led to inaccurate color detection - especially around lesion borders. Although lesion masks helped isolate the area of interest, they did not fully eliminate these inconsistencies, occasionally distorting features such as color entropy and border contrast. Furthermore, the use of MiniBatchK-Means to identify five dominant colors introduced simplification. This method assumes lesions can be represented by a limited number of color clusters, which may not reflect the complexity of heterogeneous or subtly shaded lesions, leading to less accurate representation of true pigmentation patterns. To improve the color detection, future work could use better methods to fix brightness and reflection problems in the images, like adjusting the lighting or removing shadows. Using color models that are closer to how humans see color, such as CIELAB or HSV, could help pick up on subtle color differences more accurately. Instead of always looking for exactly five main colors, it would be better to let the number of colors change depending on how complex the lesion is. Also, using smart computer programs like deep learning to find features and detect image problems could make the analysis more reliable and closer to what doctors see.

5.2 Evaluation of Blue White Veil

While the code excelled in efficiently extracting a rich set of color features and demonstrated clear strengths in identifying relevant attributes such as blue channel dominance, color diversity, and border contrast - key indicators for Blue-White Veil detection. After further evaluation, we found a few areas where the feature extraction could be slightly improved. For example, we could make the detection of blue-white veil patterns more precise by including texture or shape-related features, not just color. Instead of using fixed RGB color ranges, we could use more flexible thresholds that adjust to different lighting conditions, in order to improve resilience against lighting variability.

6 Analysis of our Classifier Models Results

This section of the report consists of an in depth analysis of how the two classifier models performed.

Random Forest

Random Forest improved from **65.00%** to **68.50%** in accuracy, a moderate increase. This result shows that some of the added features in the extended baseline helped the model make better decisions. Features like Haralick texture (e.g., entropy), Blue-White Veil presence, and color asymmetry likely added value by capturing visual or structural lesion patterns not fully described by asymmetry, border, and color alone.
However, the improvement was not dramatic because random forest already performed well on the baseline. The initial baseline already captures most of the useful information, adding new features only helped a little as they maybe didn't fully provide an entirely new insight to our previous baseline. Additionally, random forest automatically filters out irrelevant or redundant features through its internal splitting process, so any added features that were noisy or weak were likely ignored. Some added features, like Haralick entropy or Blue-White Veil ratio, probably had real predictive power; they added structural or pigmentation patterns that ABC criteria missed. But then there were others, like hair count and lesion area , which may have had weaker links or overlapped with existing features (like asymmetry and color diversity which were already included).

These weak or redundant features mentioned were probably filtered out automatically by the random forest algorithm. That's why our accuracy increased only modestly because only a subset of the added features was genuinely useful, and the rest were safely ignored (this may be a contributing factor as we perceive it, though it does not imply certainty). In short, the model benefited from a few valuable new features but was not transformed by them, hence the small but meaningful increase in accuracy, recall and precision.
This selective behavior is actually one of Random Forest's biggest strengths. It protects the model from overfitting to irrelevant features and ensures that adding more features also doesn't easily hurt performance. That's why even with the extended baseline features, our model didn't degrade like KNN did.

KNN (k-nearest neighbor algorithm)

KNN performed slightly worse on the extended baseline, dropping from **57.89%** to **57.48%** in accuracy. This may seem surprising at first, but it's a known issue with KNN when more features are added without careful preprocessing. KNN calculates distances between data points to find "neighbours." If some features are irrelevant or noisy, they distort those distances, especially when all features are treated equally. Unlike Random Forest, KNN doesn't know which features are useful; it assumes they all are. So if some of the new features incorporated in the extended baseline didn't clearly help separate benign from malignant lesions, we believe it potentially confused the model.
Furthermore, another reason the performance got slighlty worse is due to the "curse of dimensionality." (Geeksforgeeks.org K-Nearest Neighbors and Curse of Dimensionality). This refers to the issue where the performance of the algorithm deteriorates as the number of features (dimensions) in a dataset increases. This is because, in high-dimensional spaces, data points tend to spread out, making it harder for KNN to find relevant neighbors and leading to increased test error and a reduction in effective sample size. Impact of Dimensionality on KNN Performance:

- **Increased Sparsity**: As dimensions increase, the volume of the space grows exponentially. Consequently, the data becomes more spread out, making it harder to find meaningful neighbors.
- **Equal Distances**: In higher-dimensional spaces, distances between points become less meaningful because the distance between any two points tends to become more uniform or equidistant. This phenomenon is known as the "curse of dimensionality."
- **Degraded Performance**: KNN relies on the assumption that nearby points in the feature space are likely to have similar labels (i.e., belong to the same class or have similar output values). However, in high-dimensional

spaces, this assumption may no longer hold true due to increased sparsity and equalization of distances. As a result, KNN may struggle to accurately classify data points, leading to degraded performance.

So, while Random Forest could ignore or filter out unhelpful features, KNN could not filter or prioritize features, so it struggled with the added complexity and with the "curse of dimensionality", it led to slightly worse performance in the extended baseline.

7 Conclusion

This study examined whether incorporating additional features alongside the traditional ABC features significantly improves the precision of skin lesion classification compared to using ABC features alone. We developed and tested various models, including Random Forest and KNN and evaluated their performance first using only the ABC features and then with the additional features included. We found results (Stated earlier) that clearly indicated that Random Forest outperformed KNN using only the ABC features.

However, our research question was not limited to identifying the better-performing model but also whether the adding extra features would improve classification performance. In the extended baseline, Random Forest had better true positive and true negative rates (accuracy increased **3.5%**) than in the baseline, confirming that it learned from the extra features. KNN, on the other hand, showed more misclassifications across the board, with slightly higher false positives and false negatives (accuracy decreased **0.41%**). This confirms that the added features made the feature space harder to navigate for KNN, while Random Forest was able to use them effectively.

This led to our conclusion that incorporating these additional features alongside the traditional ABC features improves classification precision modestly in 1'st model, Random Forest. Therefore, we determined that the added features enhance precision depending on the model as KNN didn't get enhanced but also did not decrease significantly as it was only by **0.41%** which is very close to the original and could be due to added noise because the new features were added. We also concluded that Random Forest is the most effective model overall performing better at baseline and benefiting from the added features.

These results both highlight the significant potential of algorithmic and machine learning approaches in early skin cancer detection. Future research should focus on refining the Random Forest model and explore more advanced techniques as well as incorporating a wider range of relevant and new features to further improve diagnostic accuracy. Collaboration between this model and dermatology professionals may also further enhance the clinical effectiveness of these classifications systems.

References

M. Emre Celebi, Hitoshi Iyatomi, William V. Stoecker, Randy H. Moss, Harold S. Rabinovitz, Giuseppe Argenziano, and H. Peter Soyer. 2008. Automatic detection of blue-white veil and related structures in dermoscopy images. *Computerized Medical Imaging and Graphics*. https://pmc.ncbi.nlm.nih.gov/articles/PMC3160648/

Ping Hu and Tie-jun Yang. 2016. Wavelet-based texture analysis for pigmented skin lesion detection. *Proceedings of SPIE, the International Society for Optical Engineering*. SPIE. https://ui.adsabs.harvard.edu/abs/2016SPIE10024E..1XH/abstract

Adwait Laud, Shruti Borkar, Shrijanya Rai, and Dhirendra Mishra. 2023. Efficacy check of Haralick and symmetry features for skin lesions classification. *International Journal of Computer Applications*.

Old Authors. 2016. Ten top tips - distinguishing benign and malignant skin lesions. *Pulse Today*. https://www.pulsetoday.co.uk/clinical-feature/uncategorised/ten-top-tips-distinguishing-benign-and-malignant-skin-lesions/

Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P. Rodrigues, Rajeswari Chengoden, Sweta Bhattacharya, and Kuruva Lakshmanna. 2022. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*. https://www.nature.com/articles/s41598-022-22644-9

Kavita Sultanpure, Bhairavi Shirsath, Bhakti Bhande, Harshada Sawai, Srushti Gawade, and Suraj Samgir. 2024. Hair and scalp disease detection using deep learning. Department of Information Technology, Vishwakarma Institute of Technology, Pune, India. https://arxiv.org/pdf/2403.07940

GeeksforGeeks. 2024. K-Nearest Neighbors and Curse of Dimensionality. *GeeksforGeeks*. https://www.geeksforgeeks.org/k-nearest-neighbors-and-curse-of-dimensionality/#what-is-the-curse-of-dimensionality