Exploration of Haralick Features and the Effects on Established Skin Lesion Classification Methods

https://github.com/BossThePro/2025-FYP-groupKangaroo

Malik Jensen
IT-University of Copenhagen
majen@itu.dk

Nikolaj Pahus Pedersen
IT-University of Copenhagen
nppe@itu.dk

Peter Henrik George Mitchell
IT-University of Copenhagen
pemi@itu.dk

Simon Bruun-Simonsen
IT-University of Copenhagen
simbr@itu.dk

Victor Bloch Daugaard Hansen
IT-University of Copenhagen
vibh@itu.dk

1 Introduction

1.1 Context

On a global scale Melanoma had 331,722 new cases of skin cancer in 2022. Globally 58,667 died of skin cancer in 2022 [1]. Currently Melanoma stands as the 17th most common cancer globally [2]. If the average 2020 rates remain unchanged, melanoma cases will continue to rise. Predictions forecast a rise of 50% to approximately 510,000, with deaths increasing to approximately 96,000 by 2040. This forecast is based upon population growth and aging. Death rates must fall by greater than 2% per year worldwide to keep forecasted 2040 case counts below 2020 levels [3].

In Denmark the aged population is quite high. Roughly 20% of the population is over the age of 65 [4]. The increase in melanoma and skin cancers in general is related to the increased aged population. Factors of lifetime UV exposure increases the risk of mutations, older skin repairs less effectively, and immune surveillance is weaker. Thus lesions have greater likelihood of progressing to tumors [5].

1.2 Purpose

Known methods to help diagnose melanoma include the ABCDE method: Asymmetry, Border, Colour, Diameter, and Evolving [6]. The ABCDE criteria are primarily used to identify features associated with melanoma but can also help spot other types of skin cancer, such as basal cell carcinoma or squamous cell carcinoma, which might also exhibit irregular features [7]. However, not all non-melanoma skin cancers fit the ABCDE criteria, and not all lesions that meet ABCDE criteria are cancerous [8].

Haralick score is another feature that can help in early detection of skin cancers [9][10]. Introduced in 1973 by Robert M. Haralick. A Haralick score refers to a set of texture features derived from a Gray Level Co-occurrence Matrix (GLCM). It is widely applicable in image analysis. Its precedence in medical imaging and pattern recognition has seen it explored in various papers from 2016 to 2023 [9][10]. Thus our report will investigate the following:

> *To what extent can Haralick score be used alongside ABCDE methods in predictive modelling of skin cancers?*

2 Data Cleaning and Processing

We used the PAD-UFES-20 dataset, which consists of 2298 skin lesion images from 1373 patients [11].

2.1 Data Cleaning

In regards to the data cleaning process we wanted to reduce the amount of columns to remove some of the metrics that were not very useful to us in terms of model building such that we only kept focus on the most model appropriate columns [12].

When looking through our data we realised that there were multiple images of the same lesion in different stages. This causes a problem because

we would have to group these lesions together when dividing the data into test and training data. Therefore these images were reduced down to one image of each lesion in the data.

We also stepped into a problem with the mask as most of the features used relies upon a mask for the image. We realised that for some of the images a mask was missing and could subsequently not extract useful features from it. This led to the removal of images without masks [13].

Furthermore some of the masks were completely black rendering them useless as they would remove everything in the image including the lesion. This would lead to problems with several features since they rely on the masks, so these completely black masks were also removed [13].

We ended up with 1519 different lesions after this process with 719 of them being cancerous and 800 being non-cancerous.

The datafile also contained a lot of information about each patient. Some of the information consisted of their age, whether they had smoked or not, background of their father and mother etc. We removed most of these columns as we deemed them irrelevant for our research, while keeping some of the information like gender, age, image id etc. as they could be used for analysis in later stages [14].

3 Segmentation

3.1 Hair Feature

There are two hair features, one for the amount of black hairs and one for the amount of white hairs, both of the features are computed in the same way. The feature is calculated by doing the blackhat operation on the image, taking the sum of the blackhat image by summing up all pixel values. Afterwards, it needs to be normalized, which is done by dividing by the resolution multiplied by 255, as that is the upper limit of each pixel value.

This will give a minimum value of 0 and a theoretical maximum value of 1, but that is a very unlikely value from the blackhat operation. A value of 1 would require an entirely white result from the blackhat. The highest value on our training data was about 0.09, so we multiply that by 10 to get closer to the 0-1 range, and clamp it to the range make sure it does not go out of bounds.

The feature, "blackhatScoreBlack" takes the sum of the blackhat run on the original grayscale image, and "blackhatScoreWhite" uses the blackhat on a color-flipped image, with all of the color values of the image reversed, making the black pixels white and white pixels black.

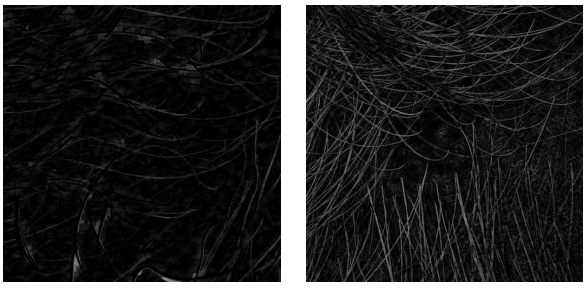
3.2 Hair Removal

We improved the hair removal function in 3 ways, scaling the kernel size of the blackhat based on resolution, scaling the threshold on the blackhat output based on the blackhatScore, and determining which color of hair to remove based on which blackhatScore is highest.

The kernel size scaling was done, as the images are of different resolution, but do not seem to show smaller or larger areas depending on it. Instead, the level of detail and "pixel density" are improved. Imagine that on an image of lower resolution, a hair's width is 3 pixels. On a higher resolution image, e.g. double for convenience, a hair would be 6 pixels wide. The kernel size should scale off of this to minimize the effect that resolution has on the blackhat.

The blackhat threshold scaling was done to try and control how aggressive the inpainting operation should be on an image. Our first thought was that it should be aggressive (low threshold) on stronger blackhat responses (high score), and more careful (high threshold) on images with lower response (low score). Our reasoning was that the strong response images would need to have more hair removed, so it should be more aggressive, and weaker responses wouldn't need to change much, so it should be careful.

An example of the difference between weak hair response and strong hair response can be seen in Figure 1



(a) Weak Hair Response (b) Strong Hair Response

Figure 1: Differences between weak and strong hair response

What we found was that the strong response images got destroyed by the inpainting operation, as

the low threshold made it inpaint on almost the entire image. The weaker response would also often not be able to reach over the high threshold, which would result in almost no inpainting. This was changed to make the algorithm more aggressive on weaker responses, and more careful on stronger responses.

4 Feature Extraction

4.1 Asymmetry

Asymmetry is when two halves of a shape, image or object are not the same. Either in form, size, or position or a mixture of all three [15]. An asymmetry score is a quantifiable measure of this.

An asymmetry score is often used in dermatology and computer-aided diagnosis systems to assess how symmetrical or asymmetrical a skin lesion is. It is an important part of the "A" in the ABCD rule for melanoma detection [16][17].

Axis-based reflection was conducted in two methods: Asymmetry Index (ASI) and Flip-based Asymmetry Scores.

4.1.1 Asymmetry Index

Asymmetry Index (ASI) is a score which compares two halves of a lesion mask by splitting the image in half, flipping one part horizontally and then overlaying each lesion on top of each other. It then checks how much both sides of the lesions overlap, and gives a percentage based score as output [18].

The formula used to compute the Asymmetry Index is given by the following:

$$ASI = \frac{\Delta AK}{AL} \times 100 \tag{1}$$

Where

- ΔAK is the difference between the two sides of the lesion
- AL is the total size of the lesion

4.1.2 Naive Flip-Based Asymmetry Scores

Based on ASI, a horizontal flip was also considered an improvement, and as such an adjusted formula was introduced

$$\begin{aligned} \text{Score}_X &= \frac{\Delta AK_Y}{\sum AL} \\ \text{Score}_Y &= \frac{\Delta AK_X}{\sum AL} \\ \text{Mean Score} &= \frac{\text{Score}_Y + \text{Score}_X}{2} \end{aligned} \tag{2}$$

Where

- ΔAK_Y is the difference between the horizontal flip
- ΔAK_X is the difference between the vertical flip

4.2 Border

Border irregularity is a key indicator in the diagnosis of skin cancers like melanoma [19]. It refers to lesion borders that are jagged, blurry, or poorly defined, where the transition from lesion to surrounding skin lacks clear delineation. To quantitatively assess this irregularity, we analyze the lesion's mask using shape descriptors such as Compactness, Convexity, Solidity, and Fractal Dimension. Each of these metrics captures different aspects of border complexity, helping to characterize the irregularity in a more objective and reproducible way [20].

4.2.1 Compactness

The Compactness Index quantifies how tightly a skin lesion's area is packed within its perimeter [21]. It is defined as:

$$compactness = \frac{4\pi * A}{perimeter} \tag{3}$$

It ranges from 0.0 to 1.0, where 1.0 represents a perfect circle, the most compact shape [22]. While the compactness metric is valuable for identifying circular or regularly shaped lesions, it is sensitive to noise along the lesion boundary. Small jagged edges or segmentation artifacts can significantly inflate the perimeter, reducing the compactness score even if the lesion's true shape is fairly regular. Accurate perimeter estimation is therefore critical. Using the skimage library, one can compute a more precise perimeter by estimating transitions between object and background pixels in a 4-neighbourhood, where H and V represent horizontal and vertical edge transitions, respectively.

Using the skimage library, one can compute a more precise perimeter by estimating transitions between object and background pixels in a 4-neighbourhood, where H and V represent horizontal and vertical edge transitions, respectively.

$$P \approx \frac{\pi}{2}(H + V) \tag{4}$$

Since compactness depends heavily on both area and perimeter, the quality of the segmentation

mask plays a central role. An inaccurate mask that fails to tightly follow the lesion border or contains noisy, irregular edges can lead to misleading compactness values and distort shape analysis. Masks with multiple lesions, may also cause some influence on the compactness as it wont correctly estimate the malignant lesion.

One solution to this would be locating each island(lesion) on the mask, but again we wont know which one the img_id correlates to, so another solution would be locating the largest of the island and compute its compactness. But then again, what if that isn't the one we should focus on. This is why the quality of the mask is important with the border features.

4.2.2 Convexity

Convexity is a morphological feature that quantifies how closely the shape of a lesion approaches a convex shape, that is the tightest bound. It is defined as the ratio between the perimeter of the lesion's convex hull and the actual perimeter of the lesion [23]

$$Convexity = \frac{\text{Perimeter of lesion}}{\text{Perimeter of convex hull}} \quad (5)$$

The convex hull is the smallest convex shape that completely encloses the lesion. It "fills in" any concave indentations along the border, essentially wrapping tightly around the lesion [24].

Because the convex hull perimeter ignores small indentations, the actual lesion perimeter is usually equal to or longer than the convex hull perimeter. Therefore:

- A convexity value close to 1 indicates that the lesion border is smooth and nearly convex, with few or no indentations.
- A convexity value significantly higher than 1 suggests that the lesion has a more indented or concave border, causing the actual perimeter to be much longer than that of its convex hull.

While convexity is effective at identifying irregularities in the form of indentations along the lesion border, it is less sensitive to sharp outward protrusions (peaks). In some cases, these protrusions are incorporated into the convex hull, increasing its perimeter and potentially skewing the convexity value.

As a result, a lesion with sharp peaks but minimal indentations may appear to have a high convexity score, suggesting a more regular shape than is actually present. This can introduce outliers or misleading data, where the lesion seems well-formed simply because the convex hull perimeter closely matches the lesion's perimeter.

An example of a convex hull can be seen in Figure 2

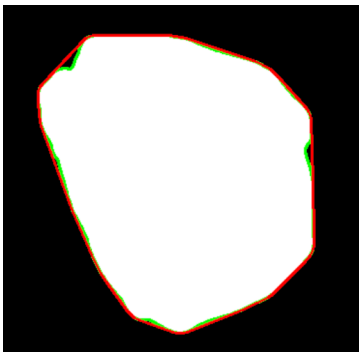


Figure 2: An example of the convex hull on a skin lesion mask

4.2.3 Solidity

Solidity [25] is a shape descriptor used to quantify how convex or filled a lesion is [26]. In the context of skin lesion analysis, solidity is calculated as the ratio between the lesion's area and the area of the smallest convex polygon that fully encloses it [27]:

$$Solidity = \frac{\text{Lesion Area}}{\text{Convex Hull Area}} \quad (6)$$

This metric provides important insights into the lesion's border characteristics

- Values close to 1: indicate that the lesion's shape is compact and convex, with smooth and regular borders.
- Values significantly less than 1: suggest that the lesion has an irregular, concave, or highly indented border.

Irregular and non-compact lesion borders are often associated with malignancy [28], such as melanoma, since malignant lesions tend to exhibit asymmetry and uneven growth patterns. Thus, solidity serves as a useful morphological feature in computer-aided diagnosis systems by quantifying border irregularity, which is a critical factor in clinical evaluation of skin lesions.

By incorporating solidity alongside other border features such as border irregularity indices, compactness, and fractal dimension automated skin lesion classification models can improve their accuracy in distinguishing benign from malignant lesions.

4.2.4 Fractal Dimension

Fractal dimension is a measure of complexity for geometric patterns that do not conform to traditional definitions of ideal mathematical shapes. In image analysis, it provides a way to quantify irregularity and self-similarity in structures that cannot be fully described using simple features [29]. We use fractal dimension to complement shape-based features such as compactness, solidity, and convexity:

- Compactness quantifies how circular or compact a shape is.
- Convexity compares the shape to its convex hull, highlighting concavities.
- Solidity relates the area of the shape to the area of its convex hull.

However, these features may miss more subtle or complex irregularities, especially those present at multiple scales. This is where fractal dimension becomes useful. It captures complexity across scales, revealing fine-grained irregularities that other shape metrics might overlook.

We use an intensity-based fractal dimension, computed with a modified box-counting method. The key idea is to not only consider how the object fills space but also how intensity varies within local regions.

The algorithm works as follows:

1. Divide the grayscale image into grid of square cells, each with side length ϵ
2. For each i, compute $\delta I_i = max(cell_i) - min(cell_i)$ (this captures the intensity range of the cell)
3. Compute $I(\epsilon) = \sum_i^M (\delta I_i + 1)$ (where M is total number of cells, and +1 ensures that uniform cells contribute positively)
4. Repeat for several values of ϵ (e.g. 2, 4, 8, 16, 32)
5. Plot $\log I(\epsilon)$ vs $\log(\frac{1}{\epsilon})$ and fit a line using linear regression
6. The resulting slope of the line is the fractal dimension

Using this in python, it may not be perfect, if the image quality is not on par. Using imagej might yield better results using the fraclac plugin, which uses this method and is optimised for it.

An example of fractal dimension can be seen in Figure 3

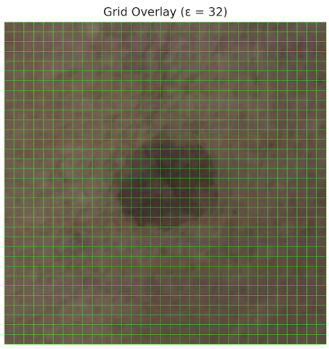


Figure 3: An example of fractal dimension on a skin lesion image

4.3 Color

For the color feature, our primary focus was based on previous research, where Simple Linear Iterative Clustering and finding important colors were used as deciding factors when diagnosing skin cancer. KMeans was also used to find the mean and standard deviation of RGB values for each image [30]

4.3.1 Simple Linear Iterative Clustering

Simple Linear Iterative Clustering (SLIC) is an algorithm created in order to group many pixels into few pixels, called superpixels[31]. SLIC is based on K-means clustering along with a modified euclidean distance based on the 5D LABXY color space, it then allocates the amount of superpixels specified, based on the modified euclidean distance, essentially checking for how close a color in a given pixel is compared to the pixels around it, as well as how close the pixels are by location[31].

Since we were mainly interested in the lesion itself, we applied the mask to the images before using SLIC, such that we only did the superpixel segmentation on the lesion itself.

This does have the potential drawback of missing important information outside of the lesion, and within the lesion if the ground truth (mask)

is constructed poorly, but this approach was preferred since we wanted the lesion to be in focus. The segmentation itself is handled by scikit-image, which has a built-in SLIC function.

An example of the superpixel segmentation can be seen in Figure 4

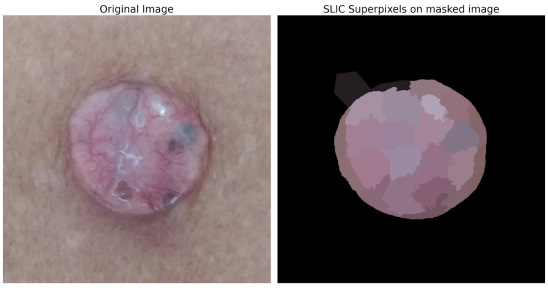


Figure 4: An example of superpixel segmentation on a skin lesion

4.3.2 Color Ratio

We then used decision rules based on previous research [30] to decide on the color of each cluster, based on the shortest euclidean distance to each color, as seen in Equation 7.

$$Dist(i) = \sqrt{(R_i - R_{ref})^2 + (G_i - G_{ref})^2 + (B_i - B_{ref})^2} \tag{7}$$

This was accompanied by a threshold, which discards any colors that are too far away from any of the known skin cancer colors. The table of colors, and thresholds can be seen in Table 1

We also decided to discard any colors that did not cover at least 5% of the lesion, in order to make sure that a significant portion of the lesion was covered by the color found, such that it does not discover things that are not present in the image.

4.3.3 K-means

We also applied K-means segmentation to the images, this is also part of what SLIC makes use of, but in this case, we limited the K-means segmentation to the amount of colors found in the color ratio, to determine the amount of clusters (different colors) that the segmentation will create for a given lesion. We then took the mean of the RGB values, and the standard deviation of the RGB values, as our features based on K-means. This was done to train the model on the potential differences between the RGB values of the different lesions

An example of the K-means segmentation for a given lesion can be seen in Figure 5

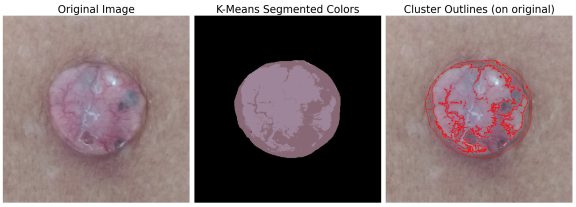


Figure 5: Image showing the K-Means segmentation based on amount of important colors found

4.4 Haralick Texture

Haralick texture is a method of image classification. It uses mathematical methods of explaining spectral, textural and contextual features on an image. These three ways to conceptualise an image have been inspired by the way humans interpret images. In short it quantifies the spatial characteristics of an image by looking at repeating patterns [32].

The reasoning of choosing it as a feature is because it makes it possible to look at the textural characteristics of a lesion and determine if it is cancerous or not. It would add another layer for the model to interpret as it would be different from looking at the asymmetry, border or color. Studies on using this method have also been done in a similar context of finding benign and malignant lesions. From this study they found their LGBM classifier which was trained on the Haralick features had an accuracy of 93.6% [33].

When looking to get the Haralick features we look at an image, which is grayscaled, and divide the pixels into pairs. These pairs are defined by picking a pixel and then picking one of its neighbours. A distance can also be defined, so if the distance is 1 then the neighbours will be touching the core pixel. If it were 2 then the pixel would be 2 pixels away and so on. The direction of where the neighbour is found could be to the right or left of it, which would correspond to 0 degrees. It could also be the one 45 degrees diagonal of the pixel, the pixel directly above or below (90 degrees) or the other diagonal pixel 135 degrees of it [32]. See Figure 6 for an example of the directions

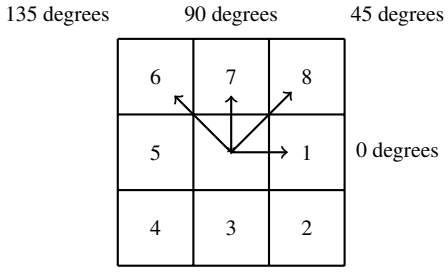


Figure 6: Grid of neighbour directionality [32]

For each of these pixel pairs there is an associated grayscale value, which ranges from 0 (black) to 255 (white).

For these pixel pairs we create a matrix, where we look at how many times two specific grayscale values are beside each other.

We then create several gray-level co-occurrence matrices (GLCM), with the amount corresponding to the different possible directions (0, 45, 90, 135), so 4 matrices. The values in the matrices then consists of the amount of gray scale neighbors for each directionality.

In each matrix this corresponds to the number of times that a given pixel with a specific grayscale value occurs next to another pixel with the same value. These matrices are used to capture local texture patterns and spatial relationships. The matrix can also be normalised by dividing each entry by the total number pixel pairs [32].

For each of these GLCMs we can extract the 14 features from them. They all have their different ways of being calculated. A feature like contrast looks at the local variation so if there are a lot of pairs where the brightness value for them differ. Another one is the correlation which looks for linear correlation in each direction [32].

Of these 14 features we used only 13 of them as the 14th feature is said to be unstable and computationally heavy [34][35].

After extracting the 13 features for each of the four directions for one image we take the mean of each feature for each GLCM, so that we get down from $13 * 4 = 52$ features back to 13. It is from these final features we interpret the image and its characteristics [32]

5 Model

In terms of model training, our pipeline consisted of cross validating over our features, attempting to tweak, add and remove different features and feature groups in order to see how it impacted the model. We also attempted a few different models, such as SVC, Logistic Regression and Random Forest to see how each model would perform on our features.

5.1 Training-test split ratio

In order to attempt to find the optimal split of training and testing data, we made use of previous research into the optimal ratio, which was done on medical imaging of brain tumors [36].

By taking the results found in the research, and finding the mean of each split ratio across the four different models, we ended up choosing a split of 73% training and 27% testing data, as it had the highest overall mean of all the ratios within our specified range. The mean was chosen as the deciding factor as to not introduce any bias towards certain modelling procedures.

5.2 Cross-Validation

For cross validation, we used the KFold function from scikit-learn, which splits the training data into n-1 clusters of training and one cluster of validation data. This was done in order to see how well our features worked on testing data, without using the test data itself. It allowed us to improve our features and see the impact of the individual feature groups by running cross-validation on specific groups of features

We attempted a few different values for the amount of folds, but ended up settling for 6 folds as it appeared to be a good split between training and validation data

5.3 Comparing Different Models

' From Table 2 we see that logistic regression performed better than the other types of models that we used in the cross-validation tests.

At first, it could potentially be contributed to overfitting through pure amount of features, but this should not be the case, since although we have 31 features, we are not breaking the one in ten rule, which states that for logistical regression models, there should be at least 10 events for each feature in order to not introduce bias [37]. Here an event is defined as the binary label with the least amount of occurrences.

Looking at our training data, we have 587 non-cancerous and 522 cancerous labels. Since the cancerous labels are less, we define the events as the cancerous labels. Assuming a similar split to

the overall training data for each split, we would end up having $522 * \frac{1}{6} = 87$ cancerous labels for validation and $522 * \frac{5}{6} = 435$ cancerous labels for training data

Since $\frac{435}{31} \approx 14$ we have not violated the one in ten rule during cross validation, and hence we did not violate this rule during the final model either, which means that overfitting based on the amount of features should be less of a worry. There are harsher versions of the rule though, such as the 1 in 20 rule and 1 in 50 rule [38], which we do violate, so there are still certain worries present in terms of overfitting. Other research has also shown that the one in ten rule may be a conservative estimate depending on what research is being conducted [39], so we have reasons to believe we are on the safe side of overfitting

Another reason might be that the scikit-learn implementation handles multicollinearity better for logistical regression models compared to the other types, since it uses regularization [40], this also appears in support vector machine [41] but does not appear in random forest [42] and K-Nearest Neighbours [43], which would explain the gap between the Logistic Regression and Support Vector Machine compared to Random Forest and K-Nearest Neighbours.

Based on these findings and our results, we ended up using the logistic regression for the final model.

6 Results

The results of the final model's performance with and without haralick features can be seen in Table 3

6.1 Logistic Regression With Haralick

We can see that the model accuracy is 70.68%, while the precision is 69.5%. The accuracy is the model's ability to predict true positives and true negatives, while precision is the probability of a positive prediction actually being cancerous. Though the difference between them is very small, only by 1.18%.

In context to the lesions, it means that the model's prediction is correct 70.68% of the time. Though when the model predicts a lesion to be cancerous it misdiagnoses 30.5% of the time. We would rather misdiagnose someone for having a cancerous lesion than missing people who have a cancerous lesion. This means we would like to have a high recall, because this is the ratio for the amount of times we guess correctly on a cancerous lesion. So the higher the recall, the less cancerous lesions we misdiagnose. In our case, our recall is 71.28%, which is higher than both accuracy and precision. This is good compared to the other ratios as I want to be sure of catching a cancerous lesion when there is one, compared to being overall accurate. F1, which is the harmonic mean of precision and recall, is 70.38%, it is typically a good metric when there is a large difference between occurences of different classes. In our case, the occurences are roughly equal, which explains why the F1 value is close to both precision and recall.

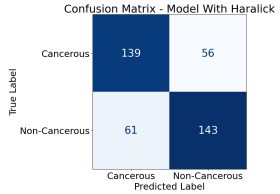
6.2 Logistic Regression Without Haralick

Compared to the logistic regression with haralick features, we observed no significant difference in the accuracy between both models, while precision increased by 2.75%. The biggest change is recall which decreased from 71.28% to 64.1%, which is a fall of 7.18%.

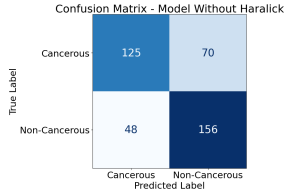
The decrease in the F1 score is also caused by the drop in recall. F1 may not say much about our model in our case, since the classifications have a similar amount of occurrences.

These results imply that the model got worse at predicting a cancerous lesion as cancerous.

The confusion matrices of both models can be seen in Figure 7



(a) Confusion Matrix - Model With Haralick Features



(b) Confusion Matrix - Model Without Haralick Features

Figure 7: Visualization of confusion matrices over models with and without haralick features

6.3 ROC Curves for Final Models

The ROC curve describes the relation between true positive rate and false positive rate. The more true positives we want, the more false positives we will also get. When the ROC curve (blue line) is above the straight dotted line then the model performs better than a random guess. A perfect model would be able to predict correctly every time while not getting a single false positive.

The ROC curves can be seen in Figure 8, Figure 9

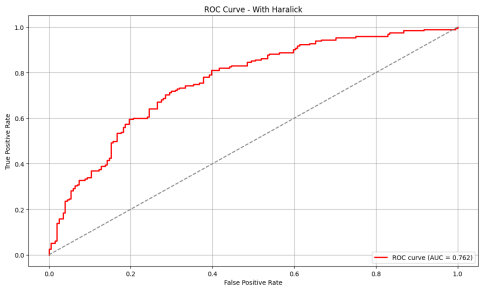


Figure 8: Figure showing the ROC curve over the final model with haralick features included. AUC = 0.762

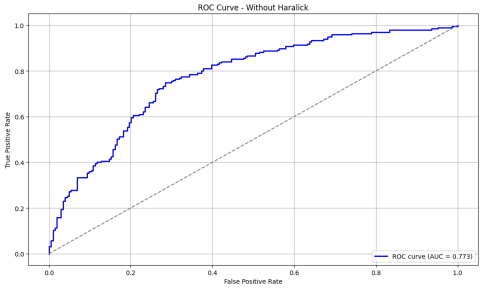


Figure 9: Figure showing the ROC curve over the final model without haralick features included. AUC = 0.773

For our two models the curve looks very similar with minor differences. The area under the curve for the model without Haralick is 0.011 higher.

Overall the model without the Haralick has a higher precision, accuracy, and AUC score, while the model with Haralick has a higher recall.

When evaluating the performance of our 2 models, we should look at our model performance metrics. AUC is most relevant when there is no difference in how important false positives and false negatives are. In our case, false negatives are way more important to minimize, as we'd much rather misdiagnose someone with cancer than we would miss a cancer case. So AUC might not be the ultimate metric for model performance here. When diagnosing cancer, the most important model statistic is recall, which the model with haralick does better in. Thus it could still be argued that the model including haralick is superior, even though the model without haralick scored higher on the AUC.

7 Discussion

The result of this study highlights several important insights, many of which are how our model could be improved.

7.1 Potential Improvements

Our model may have introduced multicollinearity, as some features describe similar attributes, which could weight that specific characteristic more in a biased manner. A potential way to mitigate this could be to check for the variance inflation factor (VIF), which provides a measure of how much the variance is increased due to multicollinearity [44].

After finalizing the model we also found out that we made a mistake in the Haralick normalization, where we accidentally normalized over all of the data, first for the training data and then for the test data, which essentially created differently scaled features for the training and test data. This does mean that we cannot necessarily conclude anything in terms of Haralick, since the features could be scaled very differently from each other.

There have been succesful implementations of Haralick in other literature, where accuracies of up to 95% have been observed [45][10].

The border and asymmetry features were also not normalized, which could potentially impact the model. The actual impact is a lot harder to measure though, since it depends on the weighting that the logistic regression assigns each feature.

The choice of looking for cancerous vs. non-cancerous instead of melanoma/non-melanoma also poses an obstacle as ABC features are mostly used to classify melanoma lesions. Our reasoning for picking cancerous vs. non-cancerous was that the amount of melanoma cases in our dataset was deemed too small, although this does potentially impact the final results, due to features not matching up with the types of cancer in our skin lesions.

7.2 Limitations

The mask for some of the images contained noise/errors in the form of white edges which didn't mark the lesion and would cause complications in our features. Either these masks could have been corrected or simply removed to improve the quality of the data.

Another problem with masks was that some consisted of multiple lesions or lesions of very unusual shape. The mask should be as precise according to the image as possible, which we found wasn't always the case. These tendencies would cause problems in our calculations as they add another level of complexity too hard to correct for.

Examples of the masks with errors can be seen in Figure 10, Figure 11

Figure 10: Example of lesion mask that does not fit to any lesion

Figure 11: Example of multiple lesions within same mask

Another limitation was in terms of the hair removal, where images with very high amounts of hair of one color will give a very strong response in both blackhats, as it starts counting the spaces between the hairs as the opposite color. This is not really an issue on images with little hair.

This leads to it often guessing wrong on images with a lot of hair when deciding which color to inpaint.

An example of this response can be seen in Figure 12

(a) Skin Lesion With Black Hair

(b) Blackhat on black hair

(c) Blackhat on white hair

Figure 12: Lesion showing strong response on both hair colors, although only one appears in the original image

We tried solving this by running a Canny edge detector, and finding the average color of the edge pixels. This turned out to work decently, though it sometimes had trouble with not all images having the same brightness. This would sometimes trick the algorithm into thinking that white hairs are darker than they really are, counting them as black instead. To fix this, we would have to threshold the black/white decision based on the brightness of the image. We would also need to take into account the size and color of the lesion, as well as skin color, which would all have an effect on the average brightness of the image. This could have also been fixed by always taking the pictures with a light on the lesion, mitigating the difference in brightness.

8 Conclusion

We can conclude from the two models that the model with the Haralick features has a higher recall, which is more appropriate in a medical sense because we would rather predict correctly on cancerous lesions more than on non-cancerous. With the mistakes in mind our models may not actually have a significance at all due to mistakes such as normalisation on the test data itself, other fea-

tures not being normalised, and multicollinearity between features. As such, little can be concluded directly from the model with haralick features if anything.

The model without haralick features still turned out decent with a recall of 64.1%. Though still having some mistakes, it is still more valid than the model with haralick features, due to not having the same normalization mistake with normalizing over training data, followed by normalizing over test data.

9 Appendix

Color	Appearance	Color Reference (RGB)	Threshold (Euclidean Distance, Non-normalized RGB Values)
Light Brown		(200, 155, 130)	30.6
Mid-brown		(160, 100, 67)	63.75
Dark Brown		(126, 67, 48)	51
White		(230, 230, 230)	63.75
Black		(31, 26, 26)	63.75
Blue-gray		(75, 112, 137)	127.5

Table 1: Table showing the different important colors and their associated thresholds [30]

	Logistic Regression	Random Forest	Support Vector Machine	K-Nearest Neighbours
Accuracy	0.675	0.6066	0.6622	0.5811
Precision	0.6609	0.5963	0.6488	0.5655
Recall	0.6729	0.5805	0.6634	0.5757
F1 Score	0.6660	0.5866	0.6544	0.5677

Table 2: Table showing the results from cross-validation across different models

	Logistic Regression With Haralick	Logistic Regression Without Haralick
Accuracy	0.7068	0.7043
Precision	0.6950	0.7225
Recall	0.7128	0.6410
F1 Score	0.7038	0.6793

Table 3: Table showing the results from final model with and without Haralick features

References

[1] World Cancer Research Fund. *Skin Cancer Statistics*. URL: https://www.wcrf.org/preventing-cancer/cancer-statistics/skin-cancer-statistics/.

[2] Freddie Bray et al. "Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: A Cancer Journal for Clinicians* 74.3 (May 2024), pp. 229–263. ISSN: 0007-9235, 1542-4863. DOI: 10.3322/caac.21834. URL: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834.

[3] Melina Arnold et al. "Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040". In: *JAMA Dermatology* 158.5 (May 2022), p. 495. ISSN: 2168-6068. DOI: 10.1001/jamadermatol.2022.0160. URL: https://jamanetwork.com/journals/jamadermatology/fullarticle/2790344.

[4] Danmarks Statistik. *Statistikbanken*. 2025. URL: https://www.statbank.dk/FOLK1A.

[5] Simone Garcovich et al. "Skin Cancer Epidemics in the Elderly as an Emerging Issue in Geriatric Oncology". In: *Aging and disease* 8.5 (2017), p. 643. ISSN: 2152-5250. DOI: 10.14336/AD.2017.0503. URL: http://www.aginganddisease.org/EN/10.14336/AD.2017.0503.

[6] Naheed R. Abbasi et al. "Early Diagnosis of Cutaneous Melanoma: Revisiting the Abcd Criteria". In: *JAMA* 292.22 (Dec. 2004), p. 2771. ISSN: 0098-7484. DOI: 10.1001/jama.292.22.2771. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.292.22.2771.

[7] R. J. Friedman, D. S. Rigel, and A. W. Kopf. "Early Detection of Malignant Melanoma: The Role of Physician Examination and Self-Examination of the Skin". In: *CA: A Cancer Journal for Clinicians* 35.3 (May 1985), pp. 130–151. ISSN: 0007-9235. DOI: 10.3322/canjclin.35.3.130. URL: http://doi.wiley.com/10.3322/canjclin.35.3.130.

[8] Ashfaq A Marghoob et al. "Instruments and New Technologies for the in Vivo Diagnosis of Melanoma". In: *Journal of the American Academy of Dermatology* 49.5 (Nov. 2003), pp. 777–797. ISSN: 01909622. DOI: 10.1016/S0190-9622(03)02470-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0190962203024708.

[9] Adwait Laud et al. "Efficacy Check of Haralick and Symmetry Features for Skin Lesions Classification". In: *International Journal of Computer Applications* 185.3 (Apr. 2023), pp. 1–8. ISSN: 09758887. DOI: 10.5120/ijca2023922679. URL: http://www.ijcaonline.org/archives/volume185/number3/laud-2023-ijca-922679.pdf.

[10] Dmitry S. Raupov et al. "Skin Cancer Texture Analysis of OCT Images Based on Haralick, Fractal Dimension, Markov Random Field Features, and the Complex Directional Field Features". In: ed. by Qingming Luo et al. Beijing, China, Oct. 2016, p. 100244I. DOI: 10.1117/12.2246446. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2246446.

[11] Andre G.C. Pacheco et al. "PAD-UFES-20: A Skin Lesion Dataset Composed of Patient Data and Clinical Images Collected from Smartphones". In: *Data in Brief* 32 (Oct. 2020), p. 106221. ISSN: 23523409. DOI: 10.1016/j.dib.2020.106221. URL: https://linkinghub.elsevier.com/retrieve/pii/S235234092031115X.

[12] GeeksForGeeks. *ML — Overview of Data Cleaning*. URL: https://www.geeksforgeeks.org/data-cleansing-introduction/.

[13] Center For Biomedical Informatics & Information Technology. *Cleaning Data: The Basics — CBIIT*. URL: https://datascience.cancer.gov/training/learn-data-science/clean-data-basics.

[14] Donald Farmer. *Clean Data Is the Foundation of Machine Learning — TechTarget*. URL: https://www.techtarget.com/searchenterpriseai/tip/Clean-data-is-the-foundation-of-machine-learning.

[15] American Academy of Dermatology Association. *What to Look for: ABCDEs of Melanoma*. URL: https://www.aad.org/public/diseases/skin-cancer/find/at-risk/abcdes.

[16] Franz Nachbar et al. "The ABCD Rule of Dermatoscopy". In: *Journal of the American Academy of Dermatology* 30.4 (Apr. 1994), pp. 551–559. ISSN: 01909622. DOI: 10.1016/S0190-9622(94)70061-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S0190962294700613.

[17] M. Emre Celebi et al. "A Methodological Approach to the Classification of Dermoscopy Images". In: *Computerized Medical Imaging and Graphics* 31.6 (Sept. 2007), pp. 362–373. ISSN: 08956111. DOI: 10.1016/j.compmedimag.2007.01.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895611107000146.

[18] Nourelhoda M. Mahmoud and Ahmed M. Soliman. "Early Automated Detection System for Skin Cancer Diagnosis Using Artificial Intelligent Techniques". In: *Scientific Reports* 14.1 (Apr. 2024), p. 9749. ISSN: 2045-2322. DOI: 10.1038/s41598-024-59783-0. URL: https://www.nature.com/articles/s41598-024-59783-0.

[19] DermNet. *ABCDEFG of Melanoma*. Oct. 2023. URL: https://dermnetnz.org/topics/abcdes-of-melanoma.

[20] Maryam Ramezani, Alireza Karimian, and Payman Moallem. "Automatic Detection of Malignant Melanoma Using Macroscopic Images". In: *Journal of Medical Signals and Sensors* 4.4 (2014), pp. 281–290. ISSN: 2228-7477. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236807/.

[21] Ihab Zaqout. "Diagnosis of Skin Lesions Based on Dermoscopic Images Using Image Processing Techniques". In: *Pattern Recognition - Selected Methods and Applications*. Ed. by Andrzej Zak. IntechOpen, July 2019. ISBN: 978-1-78985-499-2 978-1-78985-500-5. DOI: 10.5772/intechopen.88065. URL: https://www.intechopen.com/books/pattern-recognition-selected-methods-and-applications/diagnosis-of-skin-lesions-based-on-dermoscopic-images-using-image-processing-techniques.

[22] Environmental Systems Research Institute. *Compact Shape Definition — GIS Dictionary*. URL: https://support.esri.com/en-us/gis-dictionary/compact-shape.

[23] Afsah Saleem et al. "Segmentation and Classification of Consumer-Grade and Dermoscopic Skin Cancer Images Using Hybrid Textural Analysis". In: *Journal of Medical Imaging* 6.03 (Aug. 2019), p. 1. ISSN: 2329-4302. DOI: 10.1117/1.JMI.6.3.034501. URL: https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-6/issue-03/034501/Segmentation-and-classification-of-consumer-grade-and-dermoscopic-skin-cancer/10.1117/1.JMI.6.3.034501.full.

[24] GeeksForGeeks. *Convex Hull Algorithm*. URL: https://www.geeksforgeeks.org/convex-hull-algorithm/.

[25] Md Kamrul Hasan et al. "A Survey, Review, and Future Trends of Skin Lesion Segmentation and Classification". In: *Computers in Biology and Medicine* 155 (Mar. 2023), p. 106624. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2023.106624. arXiv: 2208.12232 [eess]. URL: http://arxiv.org/abs/2208.12232.

[26] Mutlu Mete and Nikolay Metodiev Sirakov. "Optimal Set of Features for Accurate Skin Cancer Diagnosis". In: *2014 IEEE International Conference on Image Processing (ICIP)*. Paris, France: IEEE, Oct. 2014, pp. 2256–2260. ISBN: 978-1-4799-5751-4. DOI: 10.1109/ICIP.2014.7025457. URL: http://ieeexplore.ieee.org/document/7025457/.

[27] GeeksForGeeks. *Find the Solidity and Equivalent Diameter of an Image Object Using OpenCV Python*. URL: https://www.geeksforgeeks.org/find-the-solidity-and-equivalent-diameter-of-an-image-object-using-opencv-python/.

[28] Mayo Clinic. *Melanoma Pictures to Help Identify Skin Cancer*. URL: https://www.mayoclinic.org/diseases-conditions/melanoma/in-depth/melanoma/art-20546856.

[29] Vladyslav Nikitin and Valery Danylov. "FRACTAL DIMENSION CALCULATION TECHNIQUES FOR SKIN LESION CHARACTERISATION". In: *FRACTAL DIMENSION CALCULATION TECHNIQUES FOR SKIN LESION CHARACTERISATION*. July 2024. ISBN: 978-617-8312-39-8. DOI: 10.62731/mcnd-26.07.2024.005. URL: https://archives.mcnd.org.ua/index.php/conference-proceeding/article/view/137.

[30] S. Oukil et al. "Automatic Segmentation and Melanoma Detection Based on Color and Texture Features in Dermoscopic Images". In: *Skin Research and Technology* 28.2 (Mar. 2022), pp. 203–211. ISSN: 0909-752X, 1600-0846. DOI: 10.1111/srt.13111. URL: https://onlinelibrary.wiley.com/doi/10.1111/srt.13111.

[31] R. Achanta et al. "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (Nov. 2012), pp. 2274–2282. ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2012.120. URL: http://ieeexplore.ieee.org/document/6205760/.

[32] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. "Textural Features for Image Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (Nov. 1973), pp. 610–621. ISSN: 0018-9472, 2168-2909. DOI: 10.1109/TSMC.1973.4309314. URL: http://ieeexplore.ieee.org/document/4309314/.

[33] Jyoti Madake et al. "Vision-Based Skin Lesion Characterization Using GLCM and Haralick Features". In: *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. Gwalior, India: IEEE, Dec. 2022, pp. 1–6. ISBN: 978-1-6654-7719-2. DOI: 10.1109/IATMSI56455.2022.10119457. URL: https://ieeexplore.ieee.org/document/10119457/.

[34] Michael V. Boland. *Haralick Texture Features*. URL: https://murphylab.web.cmu.edu/publications/boland/boland_node26.html.

[35] Leonardo Portes et al. "Feature Fusion-Enhanced t-SNE Image Atlas for Geophysical Features Discovery". In: *Scientific Reports* 15.1 (May 2025), p. 17152. ISSN: 2045-2322. DOI: 10.1038/s41598-025-01333-3. URL: https://www.nature.com/articles/s41598-025-01333-3.

[36] Muthuramalingam Sivakumar, Sudhaman Parthasarathy, and Thiyagarajan Padmapriya. "Trade-off between Training and Testing Ratio in Machine Learning for Medical Image Processing". In: *PeerJ Computer Science* 10 (Sept. 2024), e2245. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.2245. URL: https://peerj.com/articles/cs-2245.

[37] Peter Peduzzi et al. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis". In: *Journal of Clinical Epidemiology* 49.12 (Dec. 1996), pp. 1373–1379. ISSN: 08954356. DOI: 10.1016/S0895-4356(96)00236-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895435696002363.

[38] *Chapter 8: Statistical Models for Prognostication: Problems with Regression Models*. URL: https://web.archive.org/web/20041031140843/http://painconsortium.nih.gov/symptomresearch/chapter_8/sec8/cess8pg2.htm.

[39] E. Vittinghoff and C. E. McCulloch. "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression". In: *American Journal of Epidemiology* 165.6 (Jan. 2007), pp. 710–718. ISSN: 0002-9262, 1476-6256. DOI: 10.1093/aje/kwk052. URL: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwk052.

[40] scikit-learn. *LogisticRegression*. URL: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[41] scikit-learn. *SVC*. URL: https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html.

[42] scikit-learn. *RandomForestClassifier*. URL: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[43] scikit-learn. *KNeighborsClassifier*. URL: https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.

[44] Michael H. Kutner, ed. *Applied Linear Statistical Models*. 5th ed. The McGraw-Hill/Irwin Series Operations and Decision Sciences. Boston: McGraw-Hill Irwin, 2005. ISBN: 978-0-07-238688-2.

[45] Bhuvaneshwari Shetty et al. "Skin Lesion Classification of Dermoscopic Images Using Machine Learning and Convolutional Neural Network". In: *Scientific Reports* 12.1 (Oct. 2022), p. 18134. ISSN: 2045-2322. DOI: 10.1038/s41598-022-22644-9. URL: https://www.nature.com/articles/s41598-022-22644-9.