Skin Lession Detection

Sergio Bueso Domínguez serb@itu.dk
Martin Badstue Jørgensen mjoer@itu.dk
Jan Vivo Vivo jviv@itu.dk

Abstract

Skin cancer poses a significant health challenge globally, affecting millions of individuals. Early detection of cancerous lesions is crucial for successful treatment outcomes. This research aims to develop an application that employs image analysis techniques to detect the most common skin cancer type by far: Basal Cell Carcinoma (BCC). By evaluating critical features such as asymmetry, the presence of significant specific colors and the identification of white blue veil, our goal is to design a robust system capable of accurately identifying BCC in images.

As the final product of our study, we developed a single model to classify and detect such images, through rigorous testing and evaluation. We found that our classifier yielded relatively satisfactory results in classifying the previously mentioned lesions, with an accuracy above 50 percent. Given the limitations, this approach demonstrates substantial potential. For example, with future advancements, or the collaboration of professionals in the field of dermatology, a further enhancement of the accuracy and reliability of this model could take place

1 Introduction

This project focuses on developing a method to detect Basal Cell Carcinoma by analyzing images of skin lesions from the PAD-UFES-20 dataset. Utilizing image analysis techniques and machine learning classifiers, our objective is to create an efficient tool that can work along dermatologists and different medical experts to improve early skin cancer detection.

Upon first sight, it might seem too specific to design a model that predicts only one type of skin cancer. However, various reasons have led us to choose BCC as our main focus. Apart from being, by far, the most common type of skin cancer, it is also one that can only be diagnosed reliably when a skin biopsy is performed on it. That is, the area around the lesion numbed and the whole (or part) of the lesion removed, to be inspected more closely under a microscope. The suggested use of the model constructed in this project could correctly identify when the suspicion of BCC is high enough to perform a skin biopsy on it, and to avoid such a procedure when the probability of the lesion being BCC is too low.

Furthermore, being BCC the most common diagnose also in the dataset, with a relatively balanced presence of 61 percent against 39 percent in the feature file used to train the classifier, it guaranteed that the model would not as biased as it would have been, if it had been trained to find cancerous lesions in general (including melanoma and squamous cell carcinoma) which were 76 percent of the 139 images used.

In our study, we will experiment with various models, including K-Nearest Neighbors (KNN), Random Forest and logistic regression. Although we considered employing a Convolutional Neural Network (CNN), we opted against it due to its complexity, finding that a KNN approach was more suitable for our project's needs. Through this research, we aim to identify the most effective model for skin cancer detection, thereby enhancing early detection and treatment methods.

1.1 Databases

Our research utilizes the publicly available dataset PAD-UFES-20, which comprises 2,298 samples of six distinct types of skin lesions. The dataset includes data from 1,373 patients, featuring a total of 1,641 skin lesions. Table 1 below summarizes the

different types of skin lesions, along with their abbreviations and classifications. The fact that several lesion images correspond to same patients has been taken into account for the classifier.

Type of Skin Lesions	Abbreviation	Type
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

As shown in Table 1, the skin lesions are categorized into either skin cancer or skin diseases. The skin cancer types, BCC, SCC and MEL, generally result from excessive ultraviolet (UV) light exposure and uncontrolled melanocyte growth. Among these, Melanoma (MEL) is particularly aggressive, while BCC and SCC are less dangerous but still require prompt attention.

The skin disease types; ACK, SEK, and NEV, are typically less severe. Actinic Keratosis (ACK), if left untreated, can progress to Squamous Cell Carcinoma (SCC). Seborrheic Keratosis (SEK) often develops due to prolonged sun exposure and aging but does not usually pose a significant health risk.

Within the PAD-UFES-20 dataset, a metadata.csv file provides detailed information about the lesions and patients. This file includes 26 variables for each image, covering background information about the patient and measurements of the lesion.

2 Methods

2.1 Initial Subset Selection and Manual Analysis

To begin our method testing, we selected an initial subset of 80 images from the dataset. This number was manageable for manual analysis and sufficient for preliminary testing. We manually created masks that delineated the lesion areas in each image using the software Label Studio.

After masking the initial subset, we manually examined each image to identify the features we planned to extract algorithmically. This step allowed us to gauge the effectiveness of our code and adjust parameters accordingly. For each lesion, we assessed asymmetry, color variation, and the presence of a white-blue veil. Our observations were recorded in a CSV file, with each feature examined by at least two-thirds of our group to ensure accuracy.

2.2 Feature extraction

In order to create a model, we needed to first extract different features. The features we chose, as aforementioned, are; Asymmetry, color and white blue veil. We defined a script which would run our algorithms and extract the features for 150 images (randomly chosen from the Dataset), so as to obtain a precise and extensive dataset that we would later use to train our classifier.

The first 2 features were required as the bases of this study, and we chose to add the white blue veil as BCC lesions frequently come in shades of glossy black, which in most cases counts as a blue white veil. Therefore, we thought it would be relevant.

2.2.1 Asymmetry

1. **Mask Preprocessing:**

   - The mask is first cropped to remove any unnecessary empty space, retaining only the smallest possible area that contains all the active pixels.

2. **Midpoint Calculation:**

   - As the mask has been cut, the midpoint of the new cut image is now also the midpoint of the mask.

3. **Asymmetry Calculation:**

   - The mask is divided into upper, lower, left, and right halves. These halves are flipped and compared to the original halves to determine asymmetric and symmetric areas. An asymmetry score is calculated based on the ratio of asymmetric (non-overlapping) to symmetric pixels (overlapping).

     We established a threshold of 0.18, as it is vey complicated to find something completely symmetric in nature.This threshold interprets a fold as symmetric if less than 18 percent of the pixels are not overlapping (residues).

4. **Padding the Mask:**

- To prevent loss of content during rotation, the mask is padded to form a square with sides equal to the mask's diagonal length.

5. **Rotation and Asymmetry Evaluation:**
    - The mask is rotated at regular intervals, and asymmetry scores are calculated for each rotated position. This helps in understanding how the mask's asymmetry changes with different orientations.
6. **Best and Mean Asymmetry Scores:**
    - From the rotation results, the lowest (best) asymmetry score and the average (mean) asymmetry score are determined, providing insights into the mask's symmetry.
7. **Final Asymmetry Analysis:**
    - The overall best and mean asymmetry scores are computed, offering a comprehensive assessment of the mask's symmetry.

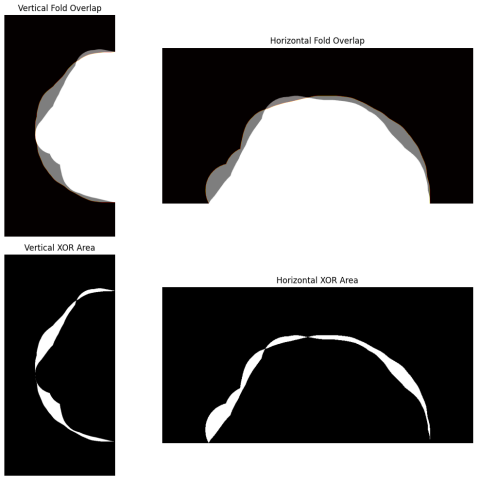

Figure 1: Mask fold visualization

2.2.2 Color

Our approach for extracting color features involves segmenting the lesion area and analyzing the color characteristics within these segments. Basing ourselves again on the color analysis of the ABCD rule from dermoscopedia, we set out to identify the presence of the following six colors in the skin lesions: red, white, black, light brown, dark brown, and blue gray, which are considered

Figure 2: Outputting Red, dark brown, black

the most relevant by professionals. The following outlines the general methodology used for this process:

1. **Segmentation with SLIC:**
    - The image is divided into superpixels using the Simple Linear Iterative Clustering (SLIC) algorithm. This ensures that the segmentation is focused on the lesion area by using a mask that excludes the skin outside the lesion. After trying different amounts, we decided to go with 50 segments, as the colors output aligned with our visual interpretations the most.
2. **Mean RGB Calculation:**
    - For each segment, we calculate the mean RGB values. This involves iterating through each segment and computing the average color values for the pixels within that segment.
3. **Color Classification:**
    - In essence an RGB-value is simply just a point in a 3 dimensional space. The mean RGB values for each segment are compared against a set of RGB-color values defined by ourselves (see below) for the six colors, and, when a specific distance is achieved from a given color, that color is given a binary value of 1 for the image.

These distances are calculated using the Manhattan distance, which is defined as the sum of absolute differences between points across all the dimensions. When the mean RGB value of a single segment is less than 90 units away from one of the colors we defined, that color is given a binary value of 1 for the image.

4. **RGB-values**

   - Red1: RGB (100, 34, 50)
   - Red2: RGB (173, 132, 132)
   - Dark brown: RGB (92, 64, 51)
   - Light brown: RGB (160, 120, 90)
   - Blue gray: RGB (112, 119, 163)
   - White: RGB (200, 200, 200)
   - Black: RGB (50, 50, 50)

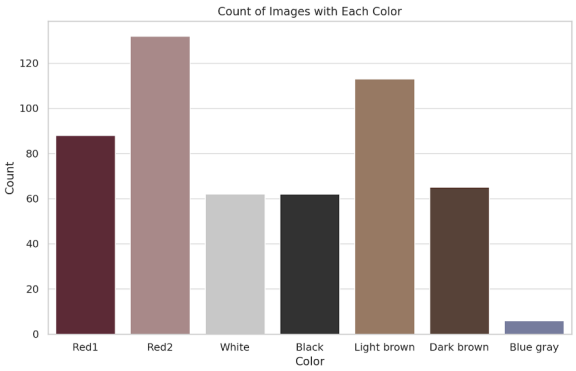

Figure 3: Count of images with each color

5. **Considerations:**

   - As red varied from pinkish red to very dark red in the lesions, we defined two reds, red1 and red2, to prevent misclassifications with this color.
   - As the SLIC segments took RGB means, purer colors such as red, black, or white where never really present in the segment averages, so we tilted the predefined values for these accordingly.
   - For variations of a single color, such as the case of dark and light brown, and blue gray, the distance for segments to be classified as such colors was increased to 90 units. This avoids cases where a slightly light brown also counts as dark brown, for example.

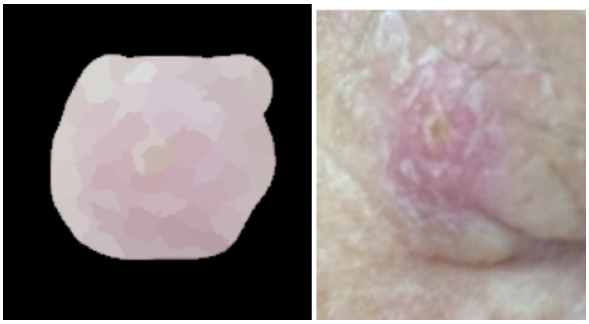

Figure 4: SLIC mask segmentation

6. **Color Extraction:**

   - The identified colors within the lesion are recorded, creating a profile of the lesion's color composition. Each color's presence is marked as a binary value, indicating whether it is detected in the lesion.

2.2.3 White blue veil

The white-blue veil is a significant feature in dermoscopic images, often associated with malignant melanoma. Our extraction method for detecting this feature involves converting the image to a specific color space and identifying color ranges that correspond to the white-blue veil. Image Conversion:

1. **LAB conversion:**

   - The image is converted from RGB to LAB color space. This conversion is chosen because the LAB color space separates luminance from color information, making it more suitable for color analysis.

2. **Finding LAB ranges:**

   - We define specific LAB color ranges that correspond to the white-blue veil. This range was found with an algorithm that gave the LAB value of a pixel when clicked on. Thus, we grouped several images from the web that clearly had blue white veil, and by repeatedly clicking on many pixels where the veil was present, we trained an appropriate range.

3. **Color Matching:**

- The image is scanned to create a binary mask where the pixels falling within these color ranges are marked.

4. **Mask Application:**
   - The predefined masks, which outline the lesion areas, are applied to the color-matched image. This ensures that only the regions within the lesion are considered for further analysis.
5. **Area Calculation:**
   - We calculate the area of the detected veil within the masked region. This involves determining the ratio of the veil area to the total masked area.
6. **Threshold Evaluation:**
   - The ratio of the detected veil area to the total lesion area is compared against a predefined threshold. If the ratio exceeds this threshold, the image is flagged as having the white-blue veil. In our case, found 8% yielded the results most according to our annotations.

2.3 Cross validation and classification model

There are a lot of different classification models that could be used in this kind of data. We tried to build and run our data through a random forest and a logistic regression, but the performance metrics turned out to be worse, so we settled for a knn(K-nearest-neighbor).

Scaling

We used MinMax Scaling to scale the ordinal data (both asymmetry values) to the binary scale of the rest of features. This was important since KNN classification is sensitive to standardization

Cross validation

We start with a ”GroupKFold” setup to ensure that samples from the same patient are not split between training and validation sets. This method divides the dataset into (k) folds. For each fold, (k-1) folds are used to train the model, while the remaining fold is used for validation. This process is repeated (k) times, and in our case, (k = 5).

Next, we test different numbers of neighbors to find the optimal parameters for the KNN classifier. We select the number of neighbors that achieves the highest accuracy. In our case, the best number of neighbors was found to be 5.

After identifying the best parameters, the model is trained on the entire dataset and saved. This allows us to later apply the model to a larger number of images.

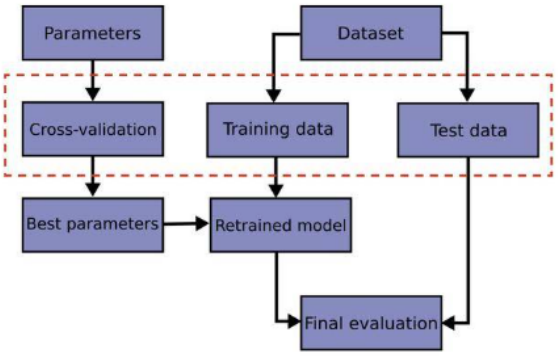

Figure 5: Flow diagram of the training process

3 Results

3.1 Extraction

We ran the image processor (which extracts the features and outputs them to the CSV file used for training) on 150 random images for the imgs_part_1 file in the PAD-UFES-20 dataset. The corresponding output file, 150features.csv, is attached with the project.

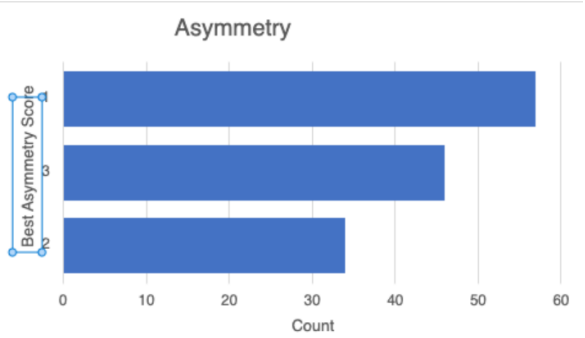

Figure 6: Count of different asymmetry values

3.2 Classifier

3.2.1 Performance

The performance of the KNN classifier was evaluated using the following metrics:

- Accuracy: The proportion of correctly classified instances.

- Precision: The proportion of true positive predictions among all positive predictions.
- Recall: The proportion of true positive predictions among all actual positives.
- F1-score: The harmonic mean of precision and recall, providing a single metric that balances both.

The average performance of the KNN classifier across multiple runs is summarized below:
Average Accuracy: 0.561
Average Precision: 0.639
Average Recall: 0.689
Average F1 Score: 0.656

	Predicted Negative	**Predicted Positive**
Actual Negative	3.8	6.8
Actual Positive	5.4	11.8

Table 2: Confusion Matrix for Final Classifier

3.2.2 Analysis and Interpretation

The classifier correctly identified 3.8 instances as negative and 11.8 instances as positive on average. It incorrectly classified 6.8 negative instances as positive and 5.4 positive instances as negative on average.

The KNN classifier exhibited moderate performance with an average accuracy of 56.1%. The precision, recall, and F1-score values indicate that the classifier has a reasonably good balance between identifying positive instances and minimizing false positives, which is something we were intentionally looking for. However, there is still great room for improvement in its accuracy and precision. To continue with the analysis, we plotted a Precision Recall Cruve, which can be a very useful representation of the usefulness of the classifier.
Figure 6 suggests the following:
At very low recall values (near 0), precision is very high (near 1). This suggests that for thresholds where very few positives are predicted, those predictions are highly accurate and reliable. However, as recall increases, precision drops significantly, indicating a trade-off between identifying more true positives and maintaining precision. After an initial drop, precision stabilizes around the 0.6-0.7 range as recall increases beyond 0.4, indicating that the classifier maintains moderate precision while identifying more true positives.

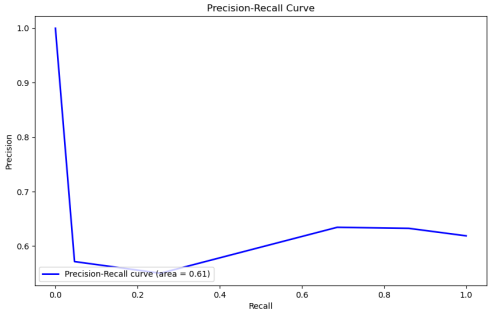

Figure 7: Precision Recall Curve

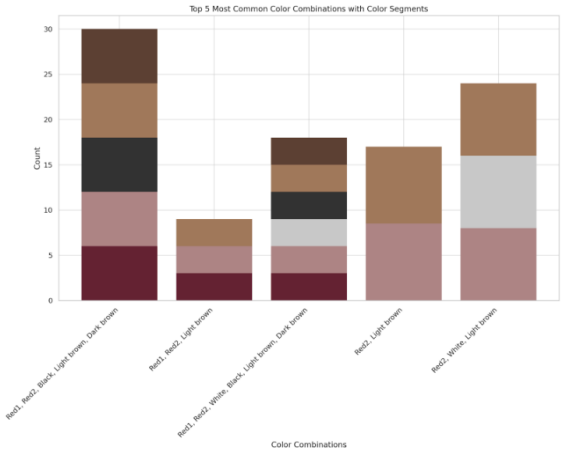

Figure 8: Most common color combinations

An AP (Average Precision) of 0.61 indicates moderate performance. While the classifier is better than random guessing (which would result in an AP closer to 0.5), there is still significant room for improvement.

4 Discussion

4.1 Comparing Manual vs Algorithmic Annotations

As aforementioned, before extracting the features algorithmically, we extracted them manually so we could compare and better adjust the parameters while coding. Now that our algorithms are complete, we can compare the manual against the automatic feature extraction and get some conclusion on how good said algorithms work.

4.1.1 Asymmetry

The ‘best asymmetry scores’ extracted from our images using the algorithm were in accordance

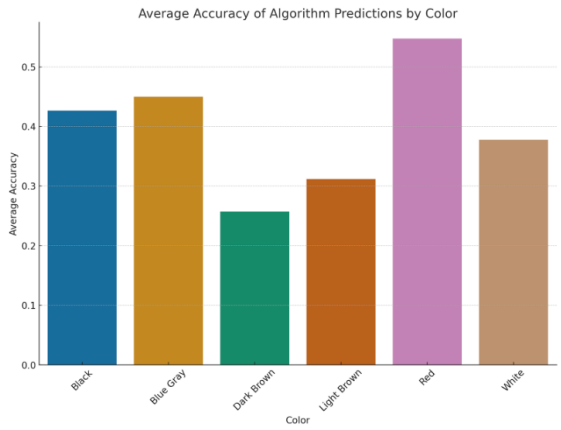

Figure 9: Average accuracy of each color

with our annotations more than 60% of the time, which arguably speaks in favor of the algorithm's performance

4.1.2 Color

To measure the color, we went over the images and noted which were the colors we could observe on the images. We had then the algorithm do the same thing.

We can see in figure 9 that compares the accuracy of the algorithm guessing colors against the ones we observed. We can see that the accuracy doesn't seem to be so great. We can attribute that to the fact that when we noted manually, we only noted down the most obvious colors, contrarily to the algorithm that noted down every single color it observed in the image segments.

Speaking in our favor, the accuracy is close to a hundred. Meaning that the algorithm was able to identify all the colors we noted, plus more colors that we did not note. That's what brought down the accuracy.

4.1.3 White Blue Veil

White blue veil is a very complicated one. There are some shades of blue that are very difficult to observe for the naked eye so when we were getting down we barely noticed any. The images we took for reference showed a lot more blue than what we saw on our images so our annotations on blue veil were close to 0. That's why comparing our annotations and the algorithm is not a proper way to evaluate how precise the later one is.

4.2 Limitations Observed

As the study was carried out and the model developed, several limitations were encountered, mostly related to the complicated nature of image processing. While the feature extraction provided somewhat similar results to those seen by eye, the methods used were not optimal. Here are some observed obstacles for each of the features:

For asymmetry, the masks were analyzed. However, the precision of these masks was not correct, as they were done by hand using Label Studio. This means that for some images, the mask is not truly representative of the shape of the lesion. Also, the decision of what is lesion and what is skin and what is lesion is subjective. This can affect the accuracy of the masks as well.

In color detection, the RGB values measured were means of different segments, meaning true colors were always mixed with other factors such as the shining of light on the lesions. Although we changed the default RGB values we were looking for to account for this, the extraction was still not optimal in this sense.

Lastly, although the algorithm used for blue white veil was very effective, some masks that were incorrectly including blue pen annotations around the lesions found blue white veil.

5 Conclusion

This study has examined the potential of algorithmic models for diagnosing skin lesions, specifically targeting Basal Cell Carcinoma (BCC). We developed and tested various models, finding that the K-Nearest Neighbors (KNN) approach delivered the best performance in classifying skin lesion data.

Despite moderate evaluation scores, our KNN model demonstrated an accuracy over 55% in predicting diagnoses using relatively simple methods. These results highlight the significant potential for algorithmic approaches in early skin cancer detection, if the limitations discussed above are tackled in more detail.

Future research should aim to refine the KNN model and explore more advanced techniques to improve diagnostic accuracy. Collaboration with dermatology professionals could further enhance the model's reliability and effectiveness.

References

International Business Machines Corporation 2024. *What is the KNN algorithm?* IBM/topics/knn

Paul E. Black 2019. *Manhattan distance* National Institute of Standards and Technology

Kumar, N. & Verma, R. 2022. *Skin lesion classification system using a K-nearest neighbor algorithm*,

Alquran, H., & Alqudah, A. M. 2018. *Blue whitish veil, atypical vascular pattern, and regression structures in dermoscopic images for melanoma diagnosis*. ResearchGate