

## **First year project**

### **Authors**

András Fekete  
Dimitar Kochev  
Maciej Jałocha  
Mihael Stoyanov  
Sune Kjær

### **Email address**

afek@itu.dk  
diko@itu.dk  
macja@itu.dk  
svst@itu.dk  
sunc@itu.dk

## Table of Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Dataset</b>	<b>3</b>
<b>3</b>	<b>Feature extraction</b>	<b>3</b>
3.1	Asymmetry . . . . .	3
3.2	Color Variability . . . . .	4
3.2.1	HSV variation . . . . .	4
3.2.2	Relative presence of important colors . . . . .	4
3.3	General-purpose features . . . . .	5
3.3.1	Blue-white veil method . . . . .	5
3.3.2	Atypical pigmentation network method . . . . .	5
3.4	Automatic segmentation . . . . .	6
<b>4</b>	<b>Classification</b>	<b>7</b>
4.1	Logistic regression - Maciej . . . . .	7
4.2	KNN . . . . .	8
4.2.1	Parameter Search and Evaluation . . . . .	8
4.2.2	Final Model of KNN . . . . .	8
4.3	Decision trees . . . . .	8
4.3.1	Parameter Search Space . . . . .	8
4.3.2	Model Optimization and Evaluation . . . . .	9
4.3.3	Final Model of Random Forest Classifier . . . . .	9
4.4	Results . . . . .	9
<b>5</b>	<b>Discussion and conclusion</b>	<b>10</b>
5.1	Limitations . . . . .	10
5.2	Conclusion . . . . .	10

## Abstract

document.

## 1 Introduction

Skin cancer is a major health concern that is hard to detect without professional expertise in dermatology. For accurate detection doctors use systems like the ABCDE (Asymmetry, Border, Color, Diameter, Evolution) [1] and the Seven-Point Checklist [2]. While effective, these methods vary in interpretation and you still need prior knowledge for the background of the possible diagnosis.

Luckily, advancements in artificial intelligence (AI) and machine learning (ML) have shown promising results in enhancing skin cancer detection. At the European Academy of Dermatology and Venereology (EADV) Congress 2023, an AI system demonstrated near-perfect accuracy in identifying skin cancers. However, researchers agree that AI should support, not replace, human experts.

Our report aims to develop a classification model inspired by the Seven-Point Checklist to predict cancerous lesions only by analysing lesion images. We extracted features such as asymmetry, HSV color variance, and key colors, incorporating elements like the blue-white veil and atypical pigmentation networks. We built and combined logistic regression, K-Nearest Neighbors (KNN) and random forest models to optimize detection. Our goal is to create the prototype of an analysing tool which aims to assist patients and dermatologists in early detection via pictures taken with normal smartphones.

## 2 Dataset

Our dataset is based on 978 photos from [3] dataset. There are photos of all 6 diagnoses - BCC, ACK, SEK, MEL, NEV, SCC for which we have used manually created masks. Due to the fact that the dataset sometimes contains more than one photo of the same patient, and sometimes even of the same lesion, we have made sure that all the photos of single patient are treated as a cohesive group. After partitioning the data into training and testing subsets, these groups remain intact within the same split. We have achieved that using sklearn model selection method called Stratified-GroupKFold, which also maintains the proportion

of the classes within the splits. In our case, that would mean the proportion of cancerous and non-cancerous lesions. At the end, the dataset is split in the following manner:

Dataset information			
Diagnosis	Full	Train	Test
Cancerous	574	430	144
Non-cancerous	404	304	100
Total	978	734 (75%)	244 (25%)

## 3 Feature extraction

For each picture of our dataset we perform two manipulations - we resize them to a length of 250px and width of 250px; we perform SLIC (Simple Linear Iterative Clustering) algorithm from 'skimage'. The SLIC is used to segment the image into 250 superpixels with a compactness parameter of 100, effectively shaping them into squares. Post-segmentation, we convert RGB outputs to HSV for a more nuanced color analysis.

### 3.1 Asymmetry

Current protocols to assess lesion's dangerousness very often include asymmetry of the lesion, though they define them differently. In ABCD rule color, shape and structure symmetry is taken into account [1] whereas in Menzies method [4] only symmetry of the pigmentation pattern is used. In addition, there are two different types of symmetry - rotational and line one, but in our research we have not been able to identify a paper which would focus on a rotational. For a review of main methods we recommend [5].

In our paper we focus on shape line asymmetry only, because it offered a good performance-time trade-off. In [6] researchers developed a method basing only on a shape line symmetry and it in 93.5% agreed with a dermatologist's annotations.

Assessing line symmetry comprises two main steps:

- Finding axis which could be called an axis of symmetry
- Folding the picture alongside that axis and computing the overlap.

We have approached the problem by first proposing three methods ourselves based on pure mathematical properties of line symmetry and then by trying out minimum bounding box method [7] The main challenge and difference between these methods is the first step.

At the end we performed tests to see how well do each of these methods can differentiate the data. We randomly sampled one hundred annotated masks out of which 57 occurred to be cancerous:

Function	Score
Fully rotated aggregated symmetry	0.658
Max symmetry axis	0.646
Max Major + perpendicular minor symmetry axes	0.681
Minimum Bounding Box Method	0.670

## 3.2 Color Variability

### 3.2.1 HSV variation

For measuring color variability, we employ a static method to measure the color variability of images by calculating the variance, which indicates the dispersion of colors. This method is particularly valuable in highlighting the color diversity within a single skin lesion. We utilize the superpixels from the already performed SLIC function to simplify the image while preserving critical color data. We also use the converted HSV (Hue, Saturation, Value) version to facilitate easier analysis of color relationships. The numbers for each of the three channels vary between 0 and 1 as the conversion function supports a set of specific formulas. For each superpixel, we compute the mean and variance of HSV values, which serve as features (Hue, Saturation, Value) in our training and test dataset. The variance calculation enhances our diagnostic accuracy by providing a quantifiable measure of color distribution.

Additionally, our observations indicate that the mean Hue often aligns with the red spectrum, ranging from 350 to 10 in HSV values, which contributed to outliers in our variance analysis. To address this, we adjusted the Hue values by 180 degrees to center around the prevalent red and pink hues, enhancing the consistency of our data. This adjustment ensures that our variance measurements remain robust across different

imaging conditions and skin tones.

### 3.2.2 Relative presence of important colors

During our research, we encountered two distinct methodologies for color extraction - one detailed in Kasmi and Mokrani article from 2016 [8], and the other presented by Ali, Li, and O'Shea in 2020 [9]. In both of them, the authors have identified 6 suspicious colors, indicative of a cancerous disease. These are light-brown, dark-brown, red, white, black and blue-gray. In the first article, using the RGB color space, the authors have chosen one single value for each of these colors. After that, they computed the Euclidean distance between every super pixel within the SLIC of the lesion and these designated color values. The smallest distance indicates that the pixel is closest to that color.

The second article followed a similar approach, but the authors have used the CIELab color space as it is considered to more closely represent the human eye perception of the colors. They have used the Minkowski distance, which is a generalization of the Euclidian and Manhattan distances, and they have used an interval to define each color rather than a single value.














Our approach combined elements from both articles while also incorporating adaptation based on the HSV color space. RGB is structured as a cube of size 255 and each of the three values, R, G and B, work as amount from each ingredient. CIELab color space allows for selecting an interval for a certain color, due to its structure. Conversely, HSV is structured as a cylinder of radius and height 100, with the three values working independently. The same change in each of them is comprehended differently by the human eye. For example, a change in saturation may make the color appear milder or sharper, while a change in hue could alter green to yellow.

To address these differences, we utilized the Manhattan distance, similar to the second article, treating each value individually. This approach is computationally efficient as it avoids calculating square roots, unlike the Euclidean distance.

In HSV, due to the range of H from 0 to 360 degrees, we encountered a challenge around the borders, as the color [359,x,y] and the color [2,x,y] are almost indistinguishable for the human eye, yet their Manhattan distance is 357, making them very different for the computer. To mitigate

this issue, we employed multiple different shades for each suspicious color, ensuring that the red (which lies around the H borders) has a shade on both ends of the scale.

Initially, we utilized the transformed HSV values for the six colors from the Kasmi and Mokrani (2016) article. Then knowing that HSV can also be represented as a cone with black located at the apex, we can cut the cone and add all pixels, whose V value is below 20, to the black color, as they are too dark and any differences in H and S values might be attributed to image quality. We then selected 13 different shades (3/3/3/2/2 for red/pink/brown/white/blue-gray).

Color shades					
1.		Red(Pantone)	8.		Pink Sherbet
2.		Red(RYB)	9.		Brown Sugar
3.		Ruby Red	10.		Brown
4.		Black Coral	11.		Bistre
5.		Blue Gray	12.		Yellow(Crayola)
6.		Persian Pink	13.		Snow
7.		Pink			

These shades were chosen after analyzing pixel values from lesions with high representation of each color and testing them on other lesions.

The light- and dark-brown have been combined, as distinguishing between them is primarily based on personal perception and also shadows on the lesions might make a light-brown section appear to be a dark-brown one. Additionally, we included pink, even though it is not a suspicious color, as it is present in many of the lesions and omitting it could lead to pink pixels being directed towards red or brown color resulting in an incorrect analysis.

### 3.3 General-purpose features

#### 3.3.1 Blue-white veil method

To extract the presence of blue-white veil, it was necessary to differentiate the blue-white portion of the lesion from other areas and subsequently assign a score indicating its presence. The primary approach for detection involved adjusting the HSV settings of the masked image to exclusively highlight the blue-white parts. Initially, a set of 60 images was selected—30 containing the feature and 30 lacking it—to optimize the HSV values and to serve as a dataset for testing the effectiveness of the detection method. Each image was analyzed alongside its filtered counterpart, allowing for the

fine-tuning of HSV settings until satisfactory extraction was achieved, at which point these settings were recorded as optimal.

However, initial tests showed that the detection accuracy was just over 50%, indicating a need for further adjustments. To enhance accuracy, the final score was set to either 1 or 0 based on a predefined threshold. Despite this improvement, some outliers remained. The threshold was dynamically adjusted using an automation script that tested various values to find the one that maximized accuracy.

Further refinement was achieved by enhancing the blue values in the image, which increased detection accuracy to approximately 95%. This significant improvement stemmed from the observation that most images featuring white structures did not actually depict lesions with a blue-white veil. The image quality generally was insufficient to accurately detect the presence of a white overlay.



Blue-white veil example

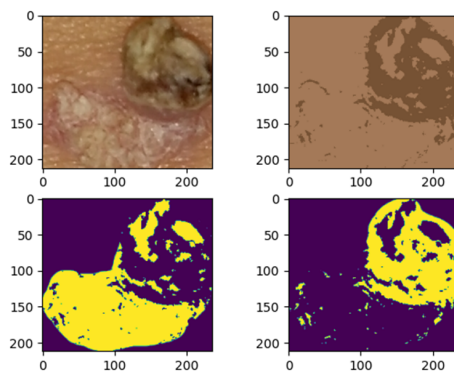
#### 3.3.2 Atypical pigmentation network method

To define what atypical pigmentation network (APN) means we looked at the example given in the Seven Point Checklist article [2] and an article that specializes in lesions in the face area that have this feature [10]. After that we developed our idea how to implement the extraction by looking at a method called Blind Color Deconvolution [11]. The only problem is that this method is applicable only if one can build filters that can divide the picture of the lesion into two canals. Since in our dataset we have a variety of pictures and each with distinct combination of colors we had to modify our approach. This is why we decided to switch the filtering with k-means clustering.

The scale of measuring of the atypical pigmentation network ranges from 0 (completely

atypical) to 1 (benign network).

Once the cropped version of the lesion and its mask are created, we perform k-means clustering on the cropped picture with the intend of getting only two segments inside it. In this way we separate the lesion into 2 major segments. One of them is going to be treated as origin of the lesion and the other - area of spreading. The two segments are converted into binary masks. As a result from our research about the nature of the deceases and the feature we arrived to the conclusion that most lesions that have this feature present start from a darker tumour or have darker pigmentation in their origin area. So we decided to compare the colour values for our two centroids using grey scale and select the segment with darker centroid and label it as the lesion's origin. After that we use the selected segment mask of the lesion origin and compare its area to the original mask of the cropped version. By calculating the proportion of overlaying we can measure the difference between the origin area and the whole area affected by the lesion. The grater the proportion the less atypical spread detected on the visible layer of the skin. Because the presence of this feature is connected to high colour variability the result should be interpreted with respect to the level of presence of different colours.



APN example

### 3.4 Automatic segmentation

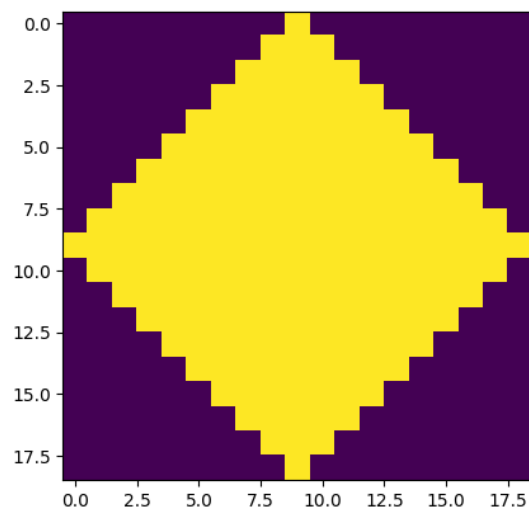
We briefly attempted doing the automatic segmentation. In order to do so we performed segmentation via classification followed by morphology processes, namely erosion and dilation.

Step by step process:

We trained KNN classifier for  $k=5$ . Algorithm decides for each of the We decided on using these features:

	Justification
Red, blue, green, gray color (4) intensities	Skin lesions most often have different color and brightness than patient's skin.
Gabor Filter	Gabor filter can capture texture differences. Some skin lesions have characteristics structs or elevations which this filter could capture. We used frequency=0.15, theta=0, and gaussian blur with sigma=3 in both dimensions.
(X,Y) coordinates	Most skin lesions are not in the edge area of the image, rather in the center.

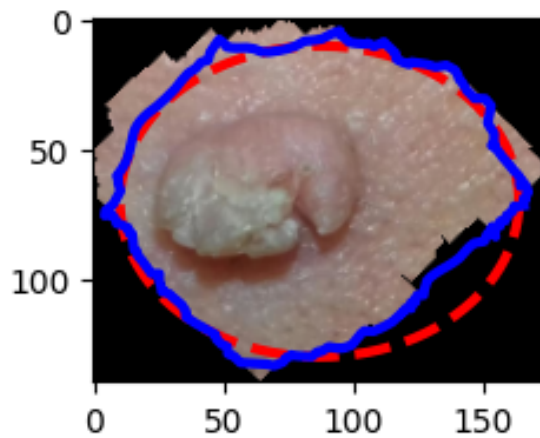
After that we applied erosion followed by dilation with a brush looking like:



Which is a diamond with the diagonals of 18 pixels in length.

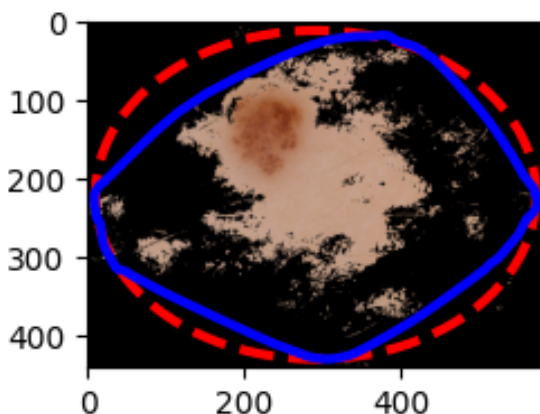
We decided on using segmentation via classification in a favour of classification via thresholding because we considered it to be more powerful (it can also include pixel intensity as a parameter). So as to obtain better segmentation, after KNN and morphology, we tried performing segmentation via shape (active contour) - the idea was to, apply an active contour on a piece of an

image selected by the mask obtained by KNN and morphology and cropped only to the bounding box. Even though the active contour was applied after trying it to on a few samples from validation dataset we saw no big improvement, so we decided not to proceed with it.



Auto seg. no ac

Active contour helps a little, but not much



Auto seg. ac

Here as well.

As our dataset we used 150 lesion images and corresponding of human annotated masks . 120 (75%) were used as train dataset and 30 (25%) as a validation dataset. As KNN is a lazy algorithm meaning that most computations are made during the inference, we decreased training sample size - from each image we sampled 100 pixels belonging to lesion and 100 not. So we got 30.000 learning data points. It solved the problem of imbalance as well.

	Dice Score	Hausdorff distance
Mean	0.56	185.60
St. Deviation	0.20	77.20

Because of the low results we decided on not carrying on the efforts, though from the visual inspection it is clear that most often predicted mask contains whole true mask which is a good sign.

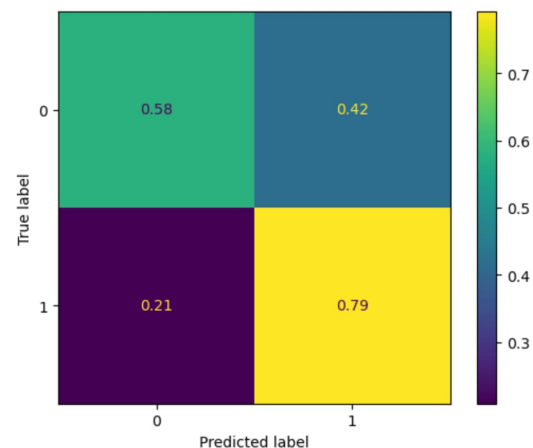
We suspect that perhaps tweaking morphology could better prepare intermediary mask for the active contour. There were many variable we have not checked, due to time constraints, like optimal number of neighbours, different classification algorithms.

## 4 Classification

Our classifier tries to solve the task of distinguishing cancerous from non-cancerous lesions. This is why we decided to combine the diagnoses MEL, BCC and SCC under the label of skin cancer (label 1) and ACK, SEK and NEV as non-cancerous conditions (label 0). We decided to test and implement the three different types of classifiers in the course: K-Nearest Neighbours, Logistic regression, and Decision tree. After a discussion how to evaluate the best performance, we decided to settle on the idea of keeping our model's train accuracy above 68% and after that maximizing recall score on testing. For all of our results we present confusion matrices.

### 4.1 Logistic regression - Maciej

Why, how we choose parameters.



Confusion matrix logistic regression model



## 4.2 KNN

K-Nearest Neighbours (KNN) is a classifier that follows the supervised learning approaches. The algorithm behind it is that when a new point has to be categorized, the model calculates the Minkowski distance between the new point and the closest k neighbours.

$$D(X, Y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

### 4.2.1 Parameter Search and Evaluation

If the sum of the distances is in favour of the cluster of cancerous lesions, the assigned label is 1, otherwise – 0. One disadvantage of the classifier is that the more features we give to the model, the more complicated this calculation gets. As a result, we get a high training time but low accuracy. In order to improve training, we need to reduce the numbers of features passed to the model but we are still concerned about keeping most of our information. To solve the problem, we choose to implement a Principal component analysis (PCA) which can select or combine some of the features that bring the most important information for classification. After implementation we examine that around 76% of the data can be explained just by 2 dimensions, 86% with 3 dimensions and 92% with 4. In order to evaluate which is the best number of dimensions we perform cross-validation with 5 folds made from our train dataset and we change the number of neighbours. According to [12] we can define the best number of neighbours by taking the square root of the number of our samples.

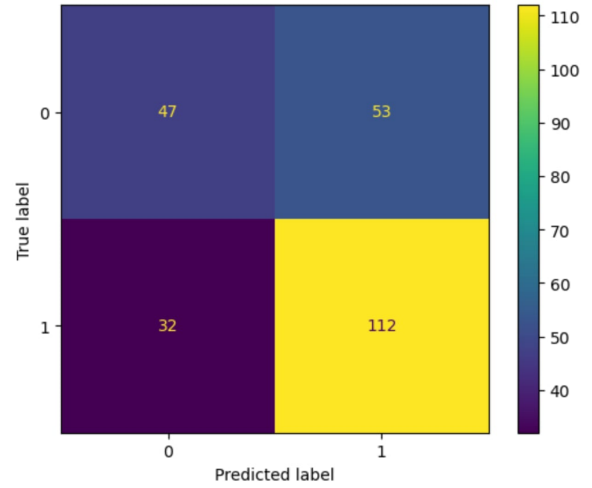
$$\text{max number of neighbours} = \sqrt{735} \approx 27$$

Now we cross-validate our example classifiers with range of neighbours from 1 to 27 and dimensions from 1 to 12. We record only the results for each dimension that best satisfies our pre-established condition. The results are represented in the following plot:

### 4.2.2 Final Model of KNN

The result from the training says that we find the best hyper parameters being 3 dimensions with 24 neighbours. Our results with test data give us:

Model: KNN  
Accuracy: 65.16% ± 3.03%  
Recall: 77.78% ± 3.49%  
Confusion Matrix:



Confusion matrix KNN model

We save both the PCA and the KNN model so that the test data can be reduced properly before the model predicts the diagnosis of the lesion in the test dataset.

## 4.3 Decision trees

In our approach for finding the lower and upper boundaries for the random forest classifier, we examined the research paper [13]. This paper focuses on optimal parameter settings for a random forest classifier, emphasizing the depth of the trees and the number of trees.

### 4.3.1 Parameter Search Space

The paper establishes a search space for the parameter `max_depth` defined by:

- **Minimum depth:** Calculated as  $\log_2(2) = 1$ . In our binary classification scenario (two classes), this results in a minimum depth of  $\log_2(2) = 1$ .
- **Maximum depth:** Determined by  $\frac{\text{Max Features}}{2}$ . With 12 features in our dataset, the maximum depth is  $\frac{12}{2} = 6$ .



Thus, the search space for `max_depth` ranges from 1 to 6.

For the `n_estimators` parameter, which defines the number of trees in the forest, the bounds are:

- **Lower bound:**  $7 \cdot \ln(\text{nob}_j) - 40$ , where  $\text{nob}_j$  is the number of samples in the training set. For our dataset with 734 images, this yields  $7 \cdot \ln(734) - 40 \approx 6$ .
- **Upper bound:**  $8 \cdot \ln(\text{nob}_j) + 45$ , resulting in  $8 \cdot \ln(734) + 45 \approx 98$ .

This establishes the number of trees' search space as between 6 to 98.

### 4.3.2 Model Optimization and Evaluation

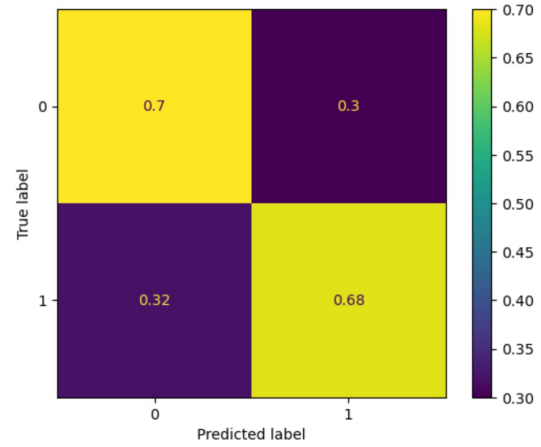
To optimize the classifier, we developed a script using `sklearn.metrics.roc_auc_score` to assess the Area Under the Receiver Operating Characteristics Curve (AUC-ROC). The AUC-ROC is a performance metric for binary classification problems, measuring the ability of the classifier to discriminate between classes. A higher AUC indicates better model performance, particularly in distinguishing between positive (cancer cases) and negative outcomes.

Additionally, we aimed to maximize recall to ensure the model effectively identifies as many true positive cases (actual cancer cases) as possible, thereby reducing the risk of false negatives.

Through iterative testing of various combinations of `max_depth` and `n_estimators` within the defined search spaces, we sought to find the optimal settings that achieve the best AUC and recall scores, thus ensuring a robust model for predicting cancer cases.

### 4.3.3 Final Model of Random Forest Classifier

We obtained the best model at 71 `n_estimators` and 6 depth with the following result:



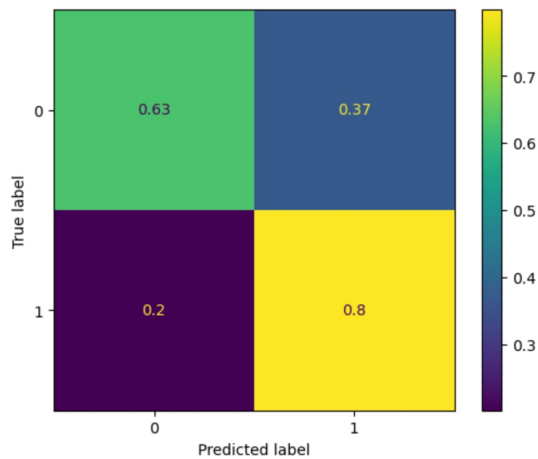
Confusion matrix random forest model

## 4.4 Results

After optimizing the parameters for our models to maximize AUC scores, we proceeded with testing. Each model's performance was assessed by first collecting accuracy and recall scores. We then conducted a sampling test, where we repeatedly sampled an equivalent number of images to our test dataset a thousand times without replacement. Each sample was run through each model, and the results (recall and accuracy) were recorded. From these results, we calculated the average accuracy and recall, along with their standard deviations.

The confusion matrices for the three optimized models indicate that the best standalone model is best model, achieving an accuracy of best model accuracy and a recall of best model recall. [SHOW CONFUSION MATRICES]

Rather than selecting the most reliable single model as our final classifier, we opted to combine all three tested models, given their promising results. The rationale for this improvement is that by integrating different models, the ensemble can mitigate individual errors, leverage diversity among the models, and generally achieve better generalization on unseen data. This concept closely parallels how the random forest classifier makes its final prediction by aggregating the predictions of its constituent trees, as discussed in "Random Forests" by Leo Breiman (2001).



Confusion matrix final model

## 5 Discussion and conclusion

### 5.1 Limitations

The study design introduces various types of limitations worth noting. Firstly, there are general limitations stemming from the diverse sources of the photos, captured by different individuals using mobile phone cameras. This causes an inconsistency in the quality due to different phones models, lighting conditions, technical expertise of the photographer, etc. Even if the photos were taken using a medical microscope, the model would have been trained for this exact device. This would result in worse accuracy when the test data comes from a different laboratory.

Secondly, the masks used in the study have been created by various annotators, lacking professional medical knowledge, which contributes to inconsistency in the masking. While the implementation of a reliable automatic segmentation tool would have been mitigated this issue, the inconsistent quality of the photos refrained us from that.

Another limitation arises from the time constraint; with more time, additional features from the 7-point checklist could have been implemented, resulting in a better overall performance of the model. Unfortunately, because of their high complexity, we could not implement any of them to a satisfactory level.

Furthermore, expanding the dataset would be beneficial so that it includes more examples of the less represented deceases such as MEL and SCC. This would facilitate the distinction between

cancerous deceases and open the possibility for building a multi classification models that can give specific diagnosis for unseen lesions.

### 5.2 Conclusion

Using the already established systems for evaluating skin deceases we successfully build a simple model that can detect cancerous lesions with high recall. Our best version uses three-way evaluation with logistic regression model, KNN model and random forest model, which increases the level of certainty for the final diagnosis. This means that our approach can offer highly reliable opinion but a misclassification in favor of false cancer is still possible.

## References

- [1] Wilhelm Stolz dermoscopia – Michael Kunz. Abcd rule — dermoscopia, 2023. [Online; accessed 20-December-2023].
- [2] Giulycalabrese Giuseppe Argenziano dermoscopia – Alina De Rosa, Teresa Russo. Seven point checklist — dermoscopia, 2023. [Online; accessed 20-December-2023].
- [3] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.
- [4] Ralph Braun dermoscopia – Scott Menzies. Menzies method — dermoscopia, 2023. [Online; accessed 20-December-2023].
- [5] William V. Stoecker, William Weiling Li, and Randy H. Moss. Automatic detection of asymmetry in skin tumors. *Computerized Medical Imaging and Graphics*, 16(3):191–197, 1992. Digital Imaging in Dermatology.
- [6] Lidia Talavera-Martínez, Pedro Bibiloni, Aniza Giacaman, Rosa Taberner, Luis Javier Del Pozo Hernando, and Manuel González-Hidalgo. A novel approach for skin lesion symmetry classification with a

deep learning model. *Computers in Biology and Medicine*, 145:105450, 2022.

- [7] Nikolay Metodiev Sirakov, Mutlu Mete, and Nara Surendra Chakrader. Automatic boundary detection and symmetry calculation in dermoscopy images of skin lesions. In *2011 18th IEEE International Conference on Image Processing*, pages 1605–1608, 2011.
- [8] Reda Kasmi and Karim Mokrani. Classification of malignant melanoma and benign skin lesions: implementation of automatic abcd rule. *IET Image Processing*, 10(6):448–455, 2016.
- [9] Abder-Rahman Ali, Jingpeng Li, and Sally Jane O’Shea. Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images. *PLOS ONE*, 15(6):1–21, 06 2020.
- [10] Linda Tognetti, Alessandra Cartocci, Elisa Cinotti, Martina D’Onghia, Magdalena Żychowska, Elvira Moscarella, Emi Dika, Francesca Farnetani, Stefania Guida, John Paoli, Aimilios Lallas, Danica Todorovic, Ignazio Stanganelli, Caterina Longo, Mariano Suppa, Iris Zalaudek, Giuseppe Argenziano, Jean Luc Perrot, Giovanni Rubegni, Gennaro Cataldo, and Pietro Rubegni. Dermoscopy of atypical pigmented lesions of the face: Variation according to facial areas. *Experimental Dermatology*, 32(12):2166–2172, 2023.
- [11] Neel Kanwal, Fernando Pérez-Bueno, Arne Schmidt, Kjersti Engan, and Rafael Molina. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access*, 10:58821–58844, 2022.
- [12] Max Bramer. *Principles of data mining*. Springer, 2007.
- [13] L A Demidova and M S Ivkina. An approach to determining the search range boundaries of optimal parameters values for the random forest algorithm. *Journal of Physics: Conference Series*, 1902(1):012112, may 2021.