IT UNIVERSITY OF COPENHAGEN

Medical Imagining

June 2nd, 2023

Renata Sapeta
resa@itu.dk

Raivis Lickrastins
rail@itu.dk

Florentina Fabregas
flfa@itu.dk

Julia Bijak
jbij@itu.dk

Michaela Rajcanova
mraj@itu.dk

This report is presented for the course
BSFIYEP1KU, First Year Project
2nd semester of BSc, Data Science
IT University of Copenhagen
Denmark

GitHub Repository
https://github.com/jbij/Group-of-5-people

Contents

1 Introduction

1.1 Introduction

Skin cancer is a disease that haunts people all over the world. As the skin is our biggest organ, having also more cells, the probability that cancer cells will develop there, is much bigger than in other parts of our bodies.

Given that the skin lesion is accurately classified in the early phase, correct disease diagnosis is much easier to achieve, therefore curing it before cancer spreads is more feasible. That is why early detection is so important. Considering the growing occurrence of skin cancer and the significance of prompt recognition, it became crucial in the medical world to develop an effective method to digitally classify skin cancer.[1]

This report aims to enhance our current knowledge of image classification tools in evaluating skin lesions, which have the potential to improve survival rates and assist in healthcare procedures. Especially, focusing on the three most significant features: asymmetry, border irregularity, and color variation. Throughout this paper, the discussion carried out in order to answer the following research questions, will be supported by the matter of limitations of the study and the depiction of the chosen approach. The main question that prompted the research is: can a predictive model accurately determine the malignancy of a given lesion based on specific diagnostic features and medical data? Additionally, we investigated more thoroughly if the fact that the patient or members of their family had previously been diagnosed with skin cancer is affecting the probability of getting skin cancer in the patient's case. Our motivation for exploring this topic is that we were curious whether skin cancer is primarily influenced by an individual's genetic makeup or if other factors play a more significant role in its occurrence.[2]

1.2 Literature review

Skin lesions are very common and often appear as a result of localized damage to the skin, i.e. sunburns or contact dermatitis. [3] Dermatologists use different scales to evaluate skin lesions, such as the ABCDE criteria for assessing pigmented lesions (asymmetry, border irregularity, color variegation, diameter larger than 6 mm, and whether the lesion is evolving over time).[5] Other factors include the shape, size, texture, and location of the lesion, as well as the patient's medical history and other symptoms.

These are the characteristics of skin damage that doctors look for when diagnosing and classifying melanomas. However, certain skin lesions need a biopsy. Such a procedure involves removing a small piece of skin from the lesion and subsequently testing this sample under a microscope, which seems to be invasive on the patient's body.[6]

While there are other types of skin lesions, our data set consisted of six different types:

Type	Full Name	Is it cancerous?
ACK	Actinic Keratosis	No
BCC	Basal Cell Carcinoma	Yes
MEL	Melanoma	Yes
NEV	Nevus	No
SCC	Squamous Cell Carcinoma	Yes
SEK	Seborrheic Keratosis	No

Table 1: Types of diagnostic in the data

The report on Bowen's disease [4] examines the association between family history of skin cancer and SCC risk. By providing results it assesses family history of skin cancer to be an important independent risk factor for SCCs. On the whole what we can understand from it, there exist an increased risk of developing SCC, if one of the parents has had it before. Moreover, people who have a family history of melanoma, have an increased risk of BCC.

2 Data

We were provided with PAD-UFES-20 dataset (collected at the Federal University of Espírito Santo), which consists of 2,298 samples of six different types of skin lesions mentioned above. The metadata associated with each skin lesion is composed of a clinical image and up to 22 features i.e. age, gender, skin cancer history and lesion diameter (see Appendix 2, Table 3). It is worth mentioning, that not all entries were filled, so there was some data missing. After analyzing the data, we observed that cases without any missing values were 64% of our whole dataset.

Investigating the data further, we have calculated the occurrence percentage of each diagnosis in the given data set. The results can be seen in the table below.

As we were not provided with the masks for the images, we had to do the lesion segmentation manually in a open source data labelling platform called Label Studio. In order to do that, we arbitrarily picked 150 pictures, regarding the percentage of each diagnosis, with no missing values in the metadata and created the masks which were crucial for training our classifier.

Moreover, we could observe that around 47% of people were diagnosed with a cancerous form of skin disease, which makes the data balanced in this regard.

Diagnosis	Amount	Percentage
ACK	730	31.8%
BCC	845	36.8%
MEL	52	2.2%
NEV	244	10.6%
SCC	192	8.4%
SEK	235	10.2%
Total	2298	

Table 2: Number of samples in provided data

Other interesting qualities of the data set were that 31% of the patients were females. It also had a relatively small percentage of people, who were labeled as smokers and drinkers, which discouraged us from investigating this part of the data any further. As for the age, most of the patients were in range between 50-80 years old. When it comes to skin color, it is not sufficiently spread, meaning in most of the cases it concerned people classified with second and third type on the Fitzpatrick scale.

3 Methods

3.1 Feature extraction

The ABCD rule, an initial dermoscopy algorithm designed to assist in distinguishing between benign and malignant melanocytic lesions [5], played a crucial role in our task. Given that the Asymmetry and Color have a significant role for this algorithm, the task included measuring those features manually ourselves prior to extracting them using code. (see Appendix 1). However, we did not choose to base a model on just two manually evaluated features.

Average color
To determine the average color of our skin lesion images, we initially applied a mask to the image and subsequently extracted the red, green, and blue color channels. The results obtained from our function consist of the mean values for the red, green, and blue channels, specifically for the lesion, disregarding any pixels outside the mask. Each of these values is treated as an individual feature.

Color Variability
The process involves applying a mask to the skin lesion image, followed by extracting the non-black pixels, which are those with color values other than black. Then, the region containing the lesion in the masked image is divided into 10 segments using the SLIC algorithm from the "skicit" python library. The color variability within each segment of the lesion is then determined by calculating the variance for each RGB channel. Finally, the mean value for all segments is

computed, resulting in a separate value for each RGB channel.

Asymmetry
To estimate the asymmetry of each lesion, we used our previously done masks. The masks were cropped and their borders were extended to provide the freedom of rotating the pictures. They were also centered, so that the "folding" picture in half was easily conducted. We rotated the pictures in 15-degree intervals from 0-90 degrees. Part of the picture that were not overlapping were representing the asymmetric areas and we summed up pixels of those areas. Lastly, the smallest area was picked as the asymmetry score of the lesion.

Border
To quantify the irregularity of the lesion's border, we developed a method to measure its compactness. It checks how much the lesion differs from a perfect circle in terms of its shape. We first calculate the area and perimeter of the lesion based on the provided mask image. The area is determined by summing the white pixels within the mask, i.e. the lesion, while the perimeter is obtained by identifying the boundary pixels. Then, we compute the compactness value which allows for a comparison of different lesions, providing varying levels of deviation from circularity.

3.2 Classification

Firstly, we decided to look at our data and visualize the distribution of the data with the pair plot of the different features.

Based on the Figure 1, we can see that the data is not well spread and does not have clear decision boundaries, nor visible linear relationships between variables. Due to these arguments, we carried out our classifier training on the non-parametric model, such as KNN. It seemed like a better fit for our data as it does not make assumptions about underlying data distribution and instead, it learns directly from the training data and can capture complex relationships and patterns.

Figure 1: Pair plot of different features

3.3 Evaluation

To evaluate our model, we decided to try the cross-validation method with different numbers of neighbors in the KNN model (1,5 and 8). We split the data into the training and test subsets, where testing data was about 30% of the pictures previously picked by us, and normalized them. Once we did that, we could perform the cross-validation on the training data, with the results given below.

We were focusing mostly on the F1 score, as it is more reliable to provide an informative evaluation. The accuracy score only refers to the overall correctness of predictions made by a classifier. It is a very straightforward method to measure the performance of the classifier. The F1 score takes into account both precision and recall, which makes it better to evaluate data sets, where it is im-

portant to keep in mind the existence of false positives and false negatives, which in our case are lesions classified as cancerous when they were not and vice versa.

	Knn 1 neighbor	Knn 5 neighbors	Knn 8 neighbors
F1	0.394167	0.538799	0.422876
Accuracy score	0.419048	0.523810	0.514286

Figure 2: Results of the cross-validation test

Based on these numbers, we decided to carry on with the KNN model of 5 neighbors.

4 Results

We have measured the performance of our classifier with accuracy and F1 score as we want as few false positive and false negative cases as possible. That means that we want to maximize the precision and recall of the predictions, so the higher the score, the better the performance.

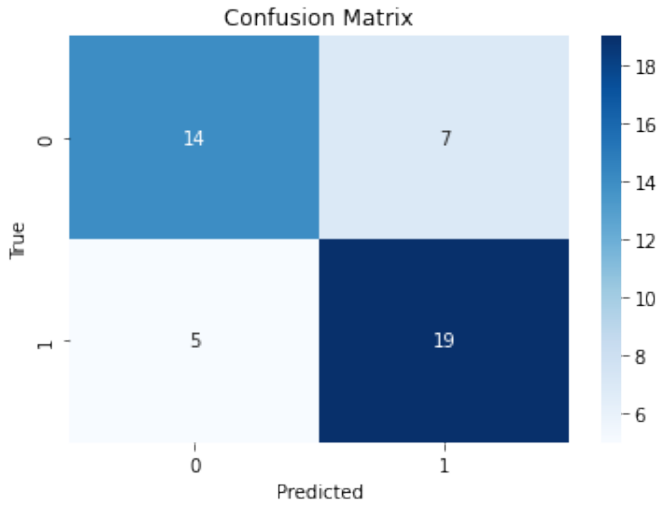

Figure 3: Confusion matrix for KNN 5

As we can see from the confusion matrix above, where "0" means that the lesion was diagnosed as cancerous and "1" - not cancerous, our test data generated 7 false negatives and 5 false positives. The chosen model performed with accuracy of 0.73 and F1 score of 0.76, which we recognize as a satisfying score.

Additionally, we decided to test whether the chosen non-parametric model did, in fact, perform better than a parametric one - logistic regression. After testing it on the test data, the following outcomes were obtained: 6 false negatives, 9 false positives, accuracy and F1 score both at 0.67. These results are notably worse than the ones from the KNN classifier, which suggests that we have chosen a better strategy.

As for the matter of skin cancer history, we conclude that a family history of skin cancer does not necessarily indicate that the patient is more prone to suffer from skin cancer.

In most cases non-melanoma skin cancer does not run in the DNA, however, the research has found that some families seem to have an abnormal number of such occurrences. These are very specific cases, where the correlation is more visible, but with provided data, it cannot be applied. By checking the data we were provided, we were able to see that the numbers of people who have been diagnosed with cancer were quite similar in both groups (with and without skin cancer history) at about 73%. Due to this high balance, we could not conclude that there exists a correlation between those two variables.

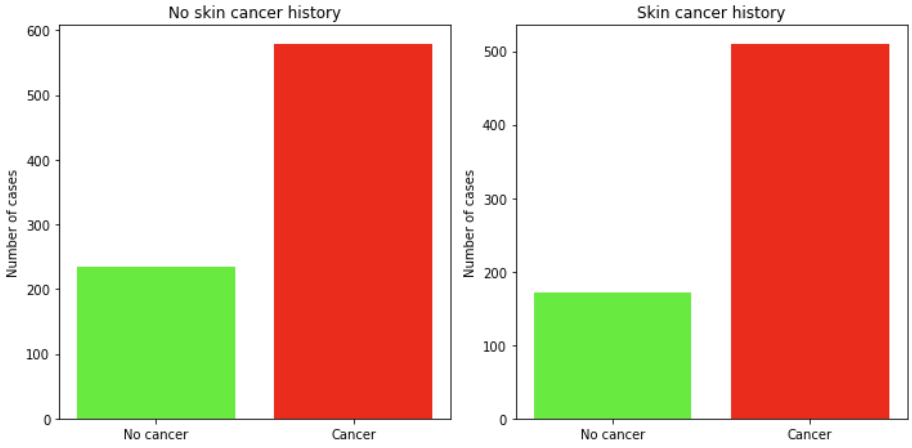

Figure 4: Distribution of diagnostic regarding skin cancer history

5 Conclusions

5.1 Limitations to our approach

Probably the most significant factor that influenced the development of our classifier, was the limited size of the data set we used for the entire project. As non-parametric models require more training data, expanding its volume would likely enhance the strength of our classifier.

There were also considerable numbers of lesions that were not visible in the images. That could be due to many reasons. During a group discussion, we took into consideration the following ones: poor image quality, interference from hair obstructing the skin or the similarity in color between the lesion and the skin.

Another difficulty that occurred as well was the lack of color variation within the data, since the used samples were limited mostly to individuals of White ethnicity. Therefore the accuracy of the model might be impaired.

We asked if there exists a correlation between a patient's history of skin cancer and their susceptibility to subsequent skin cancer diagnoses. The lack of relationship between those factors might be due to the relatively balanced data subset that we worked with.

5.2 Answers to research questions and general reflections

We tested ourselves that when the data is not clearly divided, the non-parametric classifier such as KNN model, can predict the diagnosis with relatively good accuracy and low number of false diagnoses. Taking into account the limitations of our approach, our model performed with satisfying scores. If we had more time and resources, for example pictures including more diverse skin colors, we could improve our classifier's performance even more. This shows that methods like digital segmentation of lesions and automatized diagnosis might be potentially used in the medical field.

Supplementary, based on the available evidence, it can be concluded that having a history of other types of cancer does not necessarily imply a higher likelihood of developing skin cancer. Mentioned research supports our observation in relation to non-melanoma skin cancers as most of them do not run in families [4].

6 References

[1] Yinhao Wu, Bin Chen, An Zeng, Dan Pan, Ruixuan Wang and Shen Zhao (2022) *Skin Cancer Classification With*

Deep Learning: A Systematic Review, frontiers in Oncology,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9327733/.

[2] Cara L. Benjamin and Honnavara N. Ananthaswamy (2006) *p53 and the Pathogenesis of Skin Cancer*,Toxicol Appl Pharmacol, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2080850.

[3] Kalyani Hari (2022) *Common Skin Lesions: Types, Causes, Symptoms Treatments*,vedix, https://vedix.com/blogs/articles/skin-lesions-types-causes-treatment.

[4] C.A. Morton, A.J. Birnie, D.J. Eedy (2013) *British Association of Dermatologists' guidelines for the management of squamous cell carcinoma in situ (Bowen's disease)*, British Journal of Dermatology, https://onlinelibrary.wiley.com/doi/full/10.1111/bjd.12766.

[5] Karsten Hoffmeyer (2023) *Diagnostic Strategies / Algorithms*, dermoscopedia, https://dermoscopedia.org/ABCD$_r$*ule*.

[6] Brian Krans (2017) *Skin Lesion Biopsy*, healthline, https://www.healthline.com/health/skin-lesion-biopsy.

7 Appendix

1. Annotation guide to measure features manually
Asymmetry:
0 - symmetric
1 - symmetric in one way
2 - very asymmetric

The lesions should be bisected in perpendicular axes. The distribution of colors and structures and the contour of the lesions are evaluated on either side of each axis. If asymmetry is absent with respect to both axes within the lesion, then the asymmetry score is 0. If there is asymmetry in only one axis then the asymmetry score is 1. If there is asymmetry in both axes, then the asymmetry score is 2. The asymmetry score (0-2) is multiplied with 1.3 in order to calculate the A contribution to the TDS.

Color:
The following six colors are considered important. The presence of each color (white,red,light brown, dark brown, blue-gray, black) counts a 1 point and the factor is 0.5.

Brown, black, blue-gray colors correspond with melanin distribution, white color corresponds with regression and red color reflects the degree of inflammation or neovascularization. The color white is considered to be present only if the area in question is lighter (whiter) in color than the adjacent color of normal skin.

2.

Table 3: Description of each attribute present in the metadata CSV file.

Attribute	Description
patient_id	a string representing the patient ID
lesion_id	a string representing the lesion ID
img_id	a string representing the image ID, which is a composition of the patient_ID, lesion ID, and a random number
smoke	a boolean to map if the patient smokes cigarettes
drink	a boolean to map if the patient consumes alcoholic beverages
background_father and background_mother	a string representing the country in which the patient´s father and mother descends
age	an integer representing the patient's age
pesticide	a boolean to map if patient uses pesticides
gender	a string representing patient's gender
skin_cancer_history	a boolean to map if the patient or their family has had skin cancer in the past
cancer_history	a boolean to map if patient or their family has had any type of cancer in the past
has_piped_water	a boolean to map if patient has access to piped water at home
has_sewage_system	a boolean to map if patient has access to sewage system at home
fitspatrick	an integer representing the Fitspatrick skin type
region	a string representing region on the body where lesion occured
diameter_1 and diameter_2	a float representing lesion's horizontal and vertical diameters
diagnostic	a string representing the skin lesion diagnostic
itch	a boolean to map if the skin lesion itches
grew	a boolean to map if the skin lesion has recently grown
hurt	a boolean to map if the skin lesion hurts
changed	a boolean to map if the skin lesion has recently changed
bleed	a boolean to map if the skin lesion has bled
elevation	a boolean to map if the skin lesion has an elevation
biopsed	a boolean to map if the diagnostic comes from biopsy