Final Assignment - Projects in Data Science [BSPRDAS1KU] - GroupIT

Aniela Marta Ciecierska
anci@itu.dk
Francisco Gonçalves Medeiros Lemos Moreno
fmor@itu.dk
Jakub Piotr Gąsior
jaga@itu.dk
Jonas Drøivoldsmo Lesund
jles@itu.dl
Michaela Macejovska
mimac@itu.dk

Abstract

This paper showcases the route that the members of the group G took in order to explore the capabilities of a machine to identify melanoma, as well as their findings, through different programming algorithms and human evaluations. It contains a description of the process, auxiliary graphs and a conclusion of the results found.

1 Introduction

Artificial intelligence is becoming an undeniable force in technological advancements. Its true capacity is yet to be comprehended, and for this reason, people disagree on what it can or cannot do, as well as what it should or should not do. One of the fields where AI might become a usual tool is medicine. This report will focus on its possible presence in this field, more specifically for detection and diagnosis of melanoma. Melanoma is a form of skin cancer that develops from melanocytes, a type of skin cells. It occurs when these cells grow uncontrollably until it eventually becomes malign. This phenomenon can be identified based on different factors. In the report, the group will study the theory of if an algorithm can be trained to correctly detect and diagnose melanoma based on its asymmetry, color variability and the presence of a blue-white veil.

2 Data

2.1 Source

The PAD-UFES-20 dataset was created in collaboration with the Dermatological and Surgical Assistance Program (PAD) at the Federal University of Espírito Santo (UFES) in Brazil. This program offers free treatment for various skin lesions, focusing on those unable to afford private care. The dataset comprises 2,298 samples representing six distinct types of skin lesions. Each sample includes a clinical image and up to 22 clinical attributes, such as patient age, lesion location, Fitzpatrick skin type, and lesion diameter. Approximately 58% of the dataset's samples are biopsy-confirmed. The images in the dataset vary in size due to being captured by different smartphones and are stored in .png format. Metadata accompanying each lesion contains up to 26 features, available in a CSV file where each row represents a lesion and each column represents a metadata feature. Overall, the dataset comprises data from 1,373 patients, 1,641 skin lesions, and 2,298 images, with each image/sample linked to the corresponding patient and lesion in the metadata. Additionally, the group found that the images provided did not include enough features to conduct the tests mentioned later in the report. As such, more images (542 non-melanoma and 26 melanoma) were imported from Dermoscopedia, the National Institutes of Health (NIH) and ResearchGate, which included more of the features that came to be studied. These images were used for the training of the classifiers.

2.2 Data Cleaning

Upon analysis, the images were classified into two distinct categories: low quality and high quality. Upon further examination, it was evident that within the Low quality subset, some images still retained identifiable lesions. Consequently, the Low quality category was subdivided into usable and unusable segments. Notably, four pictures from the data set were deemed unusable (`PAT_246_377_159.png`; `PAT_153_233_45.png`; `PAT_356_4511_960.png`; `PAT_1618_2771_628.png`) due to the quality being too low to identify any feature relevant for this report.

3 Methodology

3.1 Image Reduction

The images from the data set include a large area that is not relevant for this research. For this reason, the group conducted a process of segmentation masks to crop the image down to a thin frame around the identifiable lesions. Not only did this process allow for a more adequate identification of the RGB channels of the lesion, but also promoted a swifter process and run time overall.

3.2 Feature Extraction

The group created different functions in order to extract specific features from the images. The features chosen for the tests were based on the medical suggestions that high color variability, low symmetry and presence of blue-white veils could indicate melanoma. Additionally, these factors are relatively more evident to the human eye, when compared to the other possibilities, which con-

tributed to the decision of the group for which features to work with.

Color: This code loads an image and its corresponding mask, calculates the variance of colors within the lesion area defined by the mask, based in the RGB values. Then, it computes a color score based on the sum of variances across all color channels. Finally, it categorizes the score into the ascending levels of 1, 2, 3 or 4. HSV test was also considered, however its methodology is based on the intensity of each color, which the human eye cannot evaluate as easily, thus making non-viable for a comparison between the computer test and the human test. Additionally, it also proved to not be as efficient for the data as the RGB test, as shown on Figure 1 and Figure 2.

Symmetry: The code used here computes the symmetry score of a lesion based on its shape and orientation. It starts by doubling the size of the input image with a black background, so that while rotating around the longest diameter it would not be outside of the image borders, then, it finds the longest diameter of the lesion in the mask. It aligns the longest diameter vertically or horizontally, depending on its angle, by rotating the mask. Finally, it crops it to the minimum bounding box. After this, it computes the number of pixels using binary masks (value 1 for a pixel outside the lesion). Then it calculates differences on both sides vertically and horizontally determining the fraction of similarity for both cases. As a result, it places each score into, again, four ascending levels (1, 2, 3 or 4).

Blue-White Veil: For the third test, the objective is to identify the occurrence of a blue-white veil, which can be indicative of melanoma. The code converts the image to the HSV color space, and defines ranges for detecting blue and white colors in it. It creates masks for the blue and white colors separately. Then, it combines them to identify areas with a blue-white veil, and it calculates the ratio of its area to the area of the whole lesion. Finally, it returns a binary score where 1 represents the presence of a blue-white veil and 0 the absence.

3.3 Group Test

For all three of the above mentioned tests, the members of the group conducted an individual rating of each image into the same categories as the code did. The average of these results was used to compare the interpretation of the images by the human eye and by a computer.

3.4 Comparison between Computer and Human

For each of the tests, the results of the computer and the mean of the group members were compiled into plots. This mean is then approximated to its closest rating (1, 2, 3 or 4). If this value matches the computer test, it creates a 'yes', otherwise, a 'no'.

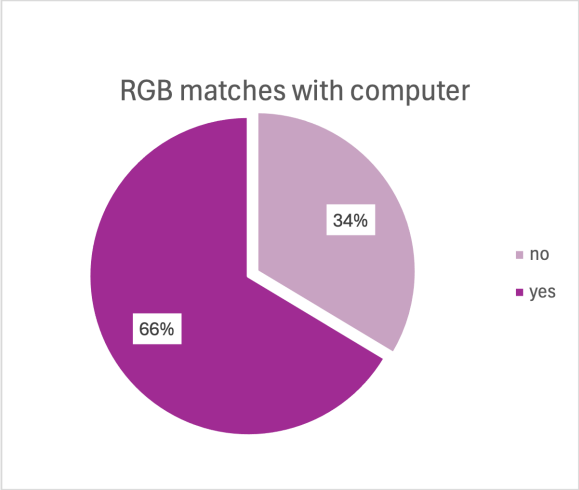

Figure 1: Matches between the computer and the human RGB tests

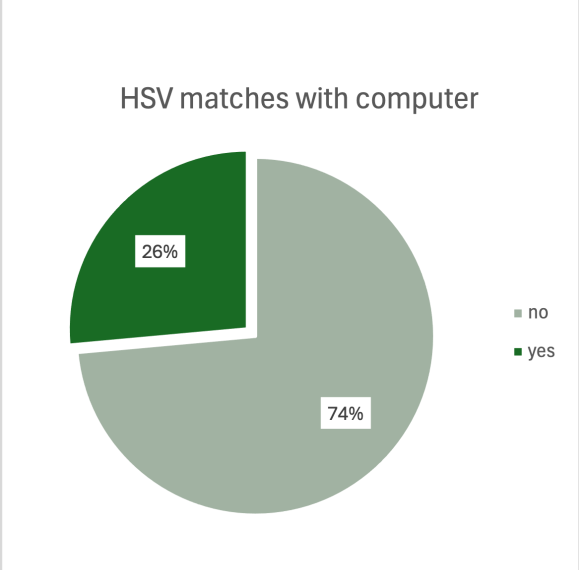

Figure 2: Matches between the computer and the human HSV tests

After analyzing both charts, the different inaccuracies of the RGB test (34%) and of the HSV

test (74%) was the reason why the group chose to use only the RGB test for the remaining of the study.

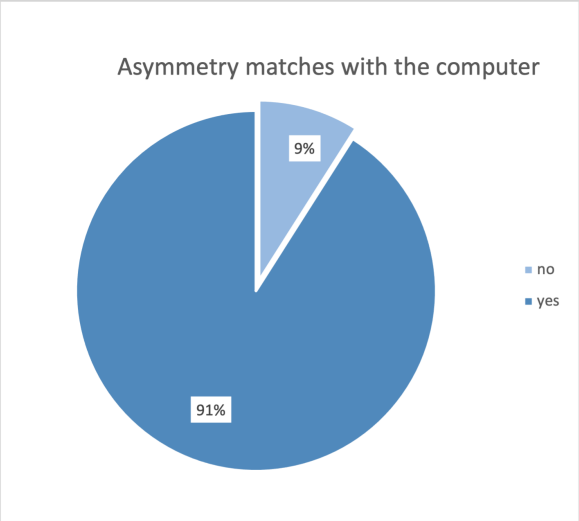

Figure 3: Matches between the computer and the human Asymmetry tests

This plot proves that the asymmetry of a lesion is visible to the human eye, as most results (91%) matched those of the computer test.

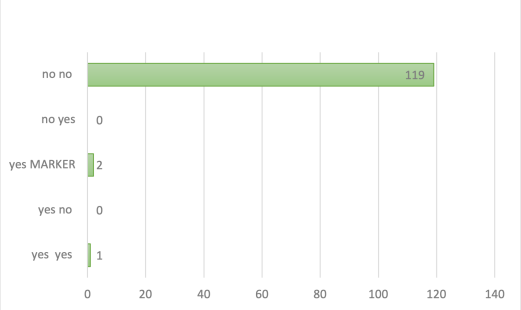

Figure 4: Matches between the computer and the human blue-white veil tests

In this plot, the first bar shows the instances where the computer and the human tests both detected no blue-white veil. For the 'no yes' and 'yes no', it is shown that there are no cases of the computer not detecting but the human doing so, or vice-versa. The 'yes MARKER' proves that there two occurrences of the computer detecting a blue-white veil, but upon human inspection it was found that those were actually misinterpretations of blue dots made by a marker or a pen. The last bar shows the only instance of a blue-white veil detected by both the computer and the human tests.

There is an additional graph in the appendix (Figure 13), which shows that although there are some instances where the mean does not match the computer, which generates a "no" as represented in Figure 1 , the deviation is usually only one point in the scale. This proves that the level of inaccuracy, although being 34%, is not as big as the graph would suggest.

3.5 Diagnostic

The data set used for the report includes the actual diagnosis of the images. This diagnosis classifies each image by the type of lesion it includes. The ones labeled with 'MEL' belong to the class melanoma, entirely biopsy-proven, and represent a '1' for in the binary classification used for the project, while any other label represents a '0' and belongs to the class non-melanoma.

4 Machine Learning: Classifiers

The group conducted a wide variety of tests to find the best classifier that can be used to detect melanoma. Three different options were considered: KNN, Decision Tree and Random Forest. For each classifier, a cross-validation setup was used to train them. Their performance was then evaluated with the help of learning curves.

4.1 KNN

K-Nearest Neighbors (KNN) is a machine learning algorithm used for classification and regression tasks. It works by finding the K closest data points to a new observation and using them to predict its label or value. KNN is based on the principle that similar things tend to be near each other. The group's algorithm begins by importing necessary libraries and setting a random seed for reproducibility. It then loads data from a CSV file and splits it into features and labels. After that, it prints the class distribution. Subsequently, the data is divided into training and test sets. Following this, a pipeline is constructed for data preprocessing (standardization) and classification using KNN with specified hyper-parameters. The code proceeds to train the classifier on the training data and generates learning curves to visualize the model's performance.

4.2 Decision Tree

The decision tree model partitions the data based on features, aiming to maximize the information

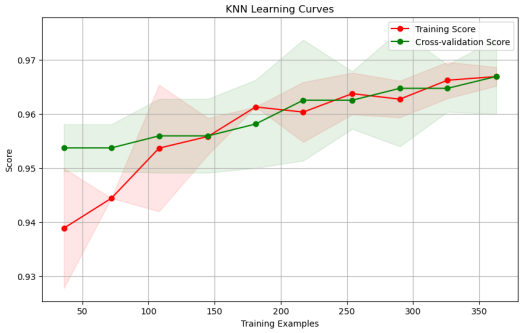

Figure 5: Learning curves for the KNN classifier

gain or minimize impurity at each step. As a result, it creates a hierarchical structure of decisions, with each node representing a decision based on a feature and each leaf node representing a class label. The group's code starts by importing necessary libraries and setting a random seed for reproducibility. It then loads data from a CSV file and splits it into features and labels. Next, it prints the class distribution. The data is subsequently divided into training and test sets. After that, a pipeline is constructed for data pre-processing (standardization) and classification using a Decision Tree Classifier with specified hyper-parameters. The model proceeds to train the classifier on the training data, and it generates learning curves to visualize the model's performance.

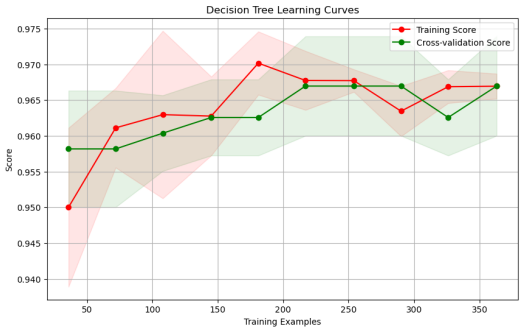

Figure 6: Learning curves for the Decision Tree classifier

4.3 Random Forest

Random Forest is an algorithm that utilizes an ensemble of decision trees. Each tree is trained on a random subset of the data and features, promoting diversity and reducing over-fitting. This code performs a machine learning classification task using a Random Forest Classifier. The group's model starts by importing necessary libraries and setting a random seed for reproducibility. Then, it loads data from a CSV file and divides it into features (X) and labels (y). It prints the class distribution of the labels. Next, the data is split into training and test sets. The code constructs a pipeline for data preprocessing (standardization) and classification using a Random Forest Classifier with specified hyperparameters. It then trains the classifier on the training data and generates learning curves to visualize the model's performance.

Figure 7: Learning curves for the Random Forest classifier

4.4 Decision and Algorithm

After analysing the three options above mentioned, the group decided to use the Random Forest classifier to train an algorithm with the images from both data sets mentioned in section 2.1. This decision came after analysing each learning curve. Although the learning curves for KNN show promising trends, the assigned probabilities, particularly concerning melanoma classification, exhibit a lack of granularity, with a restricted range of values (0.0, 0.2, 0.4, and 1.0). This limited variability in the probabilities suggests that the model may not adequately capture the intricacies or uncertainties present in the data compared to other classifiers. If alternative classifiers offer a broader spectrum of probabilities, they would show better levels of confidence or distinctions in their predictions. This attribute holds significant importance in medical contexts such as melanoma classification, where nuanced probabilities can enhance decision-making and risk assessment. In Random Forest models, the learning curves exhibit greater overlap compared to Decision Trees, indicating comparable performance between training and testing data. This similarity in performance motivated the selection of Random Forest

over Decision Trees. The outcome depends on the random seed selected by the group. It was determined that using seed 58 gets the most favorable outcomes. The efficacy is contingent upon the allocation of images for each stage of the cross-validation process and testing. Overall, all classifiers demonstrate the capability to accurately classify non-melanomas, with only a few instances of misclassification, typically involving 1-4 images. However, they encounter difficulties in labeling melanomas as such. This issue could potentially be mitigated with a larger dataset containing more melanoma images. Additionally, there is the possibility of generating new images using computer algorithms; however, this approach is unreliable and considerably less effective compared to utilizing actual melanoma images. Mean and standard deviation of training and validation scores were also calculated. Predictions were made on the test set, and further evaluation of the model's performance is conducted by generating a classification report and a confusion matrix.

```
Class Distribution:
0    542
1     26
Name: diagnostic, dtype: int64
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       109
           1       1.00      0.80      0.89         5

    accuracy                           0.99       114
   macro avg       1.00      0.90      0.94       114
weighted avg       0.99      0.99      0.99       114
```

Figure 8: Classification report for the classifier

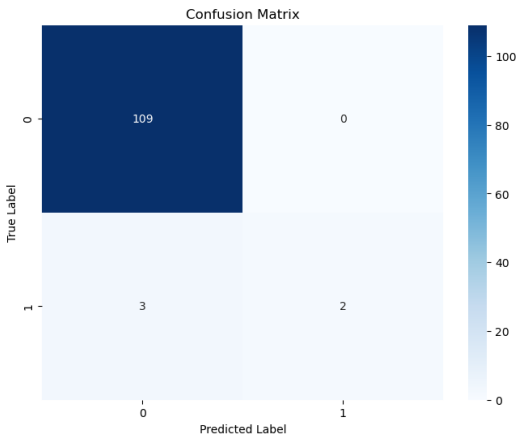

Figure 9: Confusion matrix for the classifier

This algorithm is meant to analyse images based on the factors Color variability, Symmetry and Presence of blue-white veil to conclude weather or not a lesion is melanoma, by comparing it with the results of a medical diagnose included on the data sets' descriptions. Its training was made by using the algorithm for the entire dataset (without cross-validation). First, the group trained the Random Forest classifier using the entire dataset and saved it. Then, applied it in the '3_in_1_code_with_trained_classifier' tool. This tool predicts the likelihood of an unseen lesion being a melanoma by analyzing original images and their masks. It assesses various features such as color, symmetry, and the presence of a blue-white veil. After scoring all images, the results are saved in a CSV file. Finally, using both these scores and the pre-trained Random Forest classifier, the tool predicts the probabilities. More details on how to use this tool can be found in the README.md file.

5 Limitations

This research raised some questions along its conduction, regarding its accuracy and usefulness. Firstly, it is important to denote that the amount of images used to test and train each code was not enough to deem them safe and trustworthy. Not only did the images not offer a diverse enough set of lesions or skin tones and textures, a proper machine learning model requires a much more extensive training. Most importantly, the scarcity of images with diagnosed melanoma imposed the greatest challenge to this study. Secondly, the way that the data was split for the machine learning algorithm could've caused a higher proportion of images with a blue-white veil on one of the pieces of data, which when dealing with such small quantities can cause large inaccuracies for the trained model. Lastly, the factors used for the project (color variability, symmetry and presence of blue-white veil) are not enough to scientifically classify a lesion as melanoma or not. Although high color variability, high asymmetry and the presence of a blue-white veil indicate a higher likelihood of melanoma, it is not clear how the proportions of each factor actually affect the incidence. For instance, an image rated 4 on the RGB test, but 1 on the symmetry and 0 on the blue-white veil tests does not necessarily mean it is more or less likely than an image rated 1, 4 and 0, respectively. Therefore, it is important that the lesions undergo a biopsy to detect cancerous cells to be able to diagnose it as melanoma or not.

6 Open Question: Differences Between Groups of Patients Based on Age

6.1 Research Question

How do the visual features of skin lesions (color, symmetry and blue-white veil) differ between various age groups?

6.2 Motivation

Understanding the differences in skin lesion characteristics across different age groups can provide insights into how melanoma and other skin conditions manifest in patients of different ages. This can aid in better diagnostic approaches tailored to age-specific presentations of skin lesions.

6.3 Data and Methodology

The group analysed both datasets mentioned in section 2.1 containing information about skin lesions, including visual features: color, symmetry and blue-white veil score. Then, divided the patients into five age groups: 0-20, 21-40, 41-60, 61-80, and 81-100. To determine if there are statistically significant differences in the visual features of skin lesions between these age groups, an ANOVA test was performed for each feature.

6.4 Results

1. Color Score

ANOVA Results: F-statistic: 4.3997 p-value: 0.0016

Interpretation: The color score shows significant differences across age groups. The mean color score tends to increase with age, indicating that older patients may have more colorful lesions.

2. Symmetry Score

ANOVA Results: F-statistic: 1.2273 p-value: 0.2981

Interpretation: No significant differences were found in symmetry scores across age groups. This suggests that age does not have a notable impact on the symmetry of lesions.

3. Blue-White Veil Score

ANOVA Results: F-statistic: 0.6439 p-value: 0.6314

Interpretation: No significant differences were observed in the presence of a blue-white veil across age groups. This implies that this feature is not strongly associated with patient age.

6.5 Visualizations

To further illustrate these findings, the group created bar plots showing the distribution of each score across the different age groups.

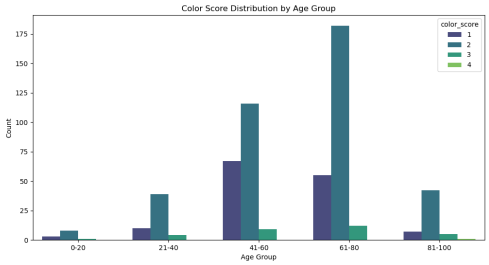

Figure 10: Color Score Distribution by Age Group

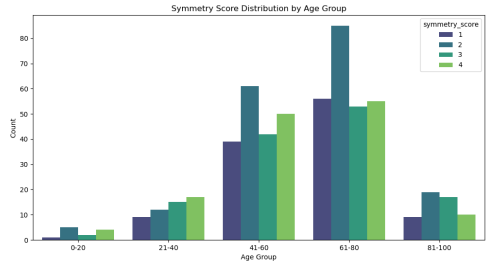

Figure 11: Symmetry Score Distribution by Age Group

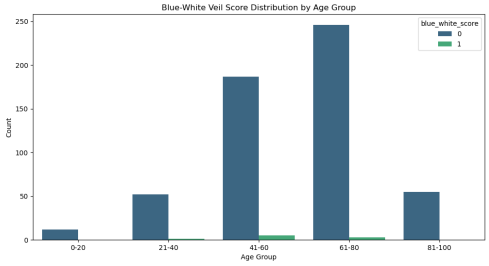

Figure 12: Blue-White Veil Score Distribution by Age Group

6.6 Conclusion

The findings suggest that while the colorfulness of skin lesions varies significantly with age, the symmetry and presence of a blue-white veil do not. The significant variation in color score might be due to biological changes in the skin as people age, leading to differences in lesion pigmentation. However, the lack of significant differences in symmetry and blue-white veil scores indicates

that these features are less influenced by age. This might be a relevant factor for dermatologists to consider during diagnosis. Further investigation through post-hoc tests can provide more granular insights into which specific age groups have significant differences in color variability scores. Additionally, other factors such as gender, smoking, and drinking habits can be analyzed to see if they contribute to differences in lesion characteristics.

7 Summary and Recommendations

This report details the group's exploration of machine learning classification methods for melanoma detection, focusing on color variability, symmetry, and the presence of a blue-white veil in skin lesions. The code employs various techniques, including color variance calculation, symmetry analysis, and blue-white veil detection, to assess lesion characteristics. Human and computer interpretations of these features are compared, highlighting discrepancies and informing algorithm selection. After evaluating KNN, Decision Trees, and Random Forest classifiers, the group opts for Random Forest due to its superior performance. It is important to admit that this study is not complete, however. As it was explored in the open question, different factors such as age can affect these results. Additionally, the limitations for the research are extensive, and prevent the trained model to be used in any trustworthy medical way. Finally, a functioning trained model should still not be interpreted as a tool for a diagnose, but as a recommendation to further concern, which to be to visit a doctor and go under the proper medical routes to get a diagnose.

References

Andre G. C. Pacheco and Gustavo R. Lima and Amanda S. Salomão and Breno Krohling and Igor P. Biral and Gabriel G. de Angelo and Fábio C. R. Alves Jr and José G. M. Esgario and Alana C. Simora and Pedro B. C. Castro and Felipe B. Rodrigues and Patricia H. L. Frasson and Renato A. Krohling and Helder Knidel and Maria C. S. Santos and Rachel B. Espírito Santo and Telma L. S. G. Macedo and Tania R. P. Canuto and Luíz F. S. de Barros 2020. *PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones*. Mendeley Data, V1. doi:10.17632/zr7vgbcyr2.1.

A. Madooei and M.S. Drew and M. Sadeghi and M.S. Atkins 2013. Automatic detection of blue-white veil by discrete colour matching in dermoscopy images. *Med Image Comput Comput Assist Interv*, 16(Pt 3):453-60. doi:10.1007/978-3-642-40760-4_57. PMID:24505793.

Cancer.Net. 2023. *Melanoma*. Retrieved from `https://www.cancer.net/cancer-types/melanoma/view-all` (Accessed Oct. 9, 2023).

National Comprehensive Cancer Network. 2023. *Melanoma: Cutaneous*. Retrieved from `https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1492` (Accessed Oct. 9, 2023).

Niederhuber JE, et al., eds. 2020. *Melanoma. In: Abeloff's Clinical Oncology.* 6th ed. Elsevier. Retrieved from `https://www.clinicalkey.com` (Accessed Oct. 9, 2023).

National Cancer Institute. 2023. *Melanoma treatment (PDQ) – Patient version*. Retrieved from `https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq` (Accessed Oct. 9, 2023).

National Cancer Institute. 2023. *Common moles, dysplastic nevi and risk of melanoma*. Retrieved from `https://www.cancer.gov/types/skin/moles-fact-sheet` (Accessed Oct. 10, 2023).

Rashid S, et al. 2023. *Melanoma classification and management in the era of molecular medicine*. *Dermatology Clinics*, doi:10.1016/j.det.2022.07.017.

National Cancer Institute. 2023. *Sunlight*. Retrieved from `https://www.cancer.gov/about-cancer/causes-prevention/risk/sunlight` (Accessed Oct. 10, 2023).

Allan C. Halpern, MD; Ashfaq A. Marghoob, MD; Ofer Reiter, MD. 2021. *Melanoma Warning Signs: What You Need to Know About Early Signs of Skin Cancer*.

Oriol Yélamos. 2023. *Blue White Structures*.

Ali Madooei, Mark Drew, Maryam Sadeghi, Margaret Atkins. 2013. *Automatic Detection of Blue-White Veil by Discrete Colour Matching in Dermoscopy Images*. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 16, pages 453-60. ISBN: 978-3-642-38708-1. DOI: 10.1007/978-3-642-40760-457.

8 Appendix

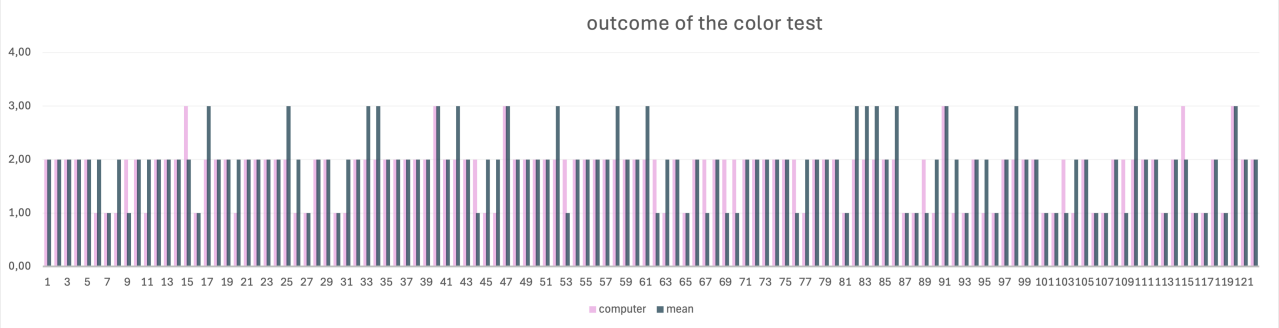

Figure 13: Outcome of the color test