IT UNIVERSITY OF COPENHAGEN

Medical Imaging

Course: First Year Project
Course Code: BSFIYEP1KU

Course Manager: Louise Meier Carlsen
Teachers: Veronika Cheplygina, Amelia Jiménez Sánchez

Group 1 - Bozhidara Pesheva (bope@itu.dk), Lili Raleva (lilr@itu.dk),
Gabriela Zhelyazkova (gazh@itu.dk),
Cristina Avram (cria@itu.dk), Cristian-Andrei Pricope (crpr@itu.dk)
https://github.com/ImBrun/FYP.git

May 2023

Contents

1 Introduction

Skin lesions are a common sighting in our everyday life. On one side are the occasional benign moles and birthmarks, but on the other – the life-threatening cancerous skin diseases. These dangerous lesions have specific characteristics that make them detectable by dermatologists and, when in an already advanced stage – recognizable by the patients themselves. The ABCD rule [1] (Asymmetry, Border irregularity, Color variegation, and Diameter) provides a checklist of alarming features which have the potential to be signs of malignant formations. The rule is not for exclusive use by dermatologists and physicians only but comprehensible for every individual who is in doubt. However, if suspicion arises, one should immediately seek medical assistance.

As skin cancer diseases constitute the majority of identified types of cancer globally [2], many mobile applications that diagnose skin lesions have been developed in an attempt to aid the early detection of malignant formations. The investigation our group conducted regards this topic. After being presented with an extensive data set consisting of images of different malignant and benign skin lesions, we had to decide on a combination of characteristics that would be extracted from the pictures. These features would later be used to train a machine learning model, which, in turn, would be tasked with the diagnosis prediction of skin formations. The model is meant to give an evaluation of the lesion – labeling it as harmless or potentially malignant.

In this report, our group documents the process of crafting a skin cancer classifier - from familiarizing ourselves with the data set at our disposal, deciding on and extracting the distinctive features of the lesions, to testing different classification methods and tracking their performances. By utilizing the abovementioned ABCD rule, machine learning techniques, and the programming language Python, we yielded a skin cancer classifier that may be of help for the early detection of dangerous skin formations and will, prospectively, motivate more individuals to get a medical checkup.

2 Exploring the Data

The PAD-UFES-20 data set was collected along with the Dermatological and Surgical Assistance Program (in Portuguese: Programa de Assistência Dermatológica e Cirurgica - PAD) at the Federal University of Espírito Santo (UFES-Brazil). The data set consists of 2,298 samples of six different types of skin lesions. Each sample consists of a clinical image and up to 22 clinical features (smoke, drink, the background of the father, the background of the mother, age, pesticide, gender, skin cancer history, cancer history, if the patient has piped water, if the patient has sewage system, Fitzpatrick region, diameter, diagnostic, itch, grew, hurt, changed, bleed, elevation, image id, biopsied).

The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen's disease (BOD), Melanoma (MEL), Nevus (NEV).

All BCC, SCC, and MEL are biopsy-proven. The remaining ones may have clinical diagnoses according to a consensus of a group of dermatologists. In total, approximately 58% of the samples in this data set are biopsy-proven. This information is described in the metadata. The images in the data set are different sizes because they are collected using different smartphone devices. In total, 1,373 patients, 1,641 skin lesions, and 2,298 images are present in the data set. Each image/sample references the patient and the skin lesion in the metadata.[3]

[1] American Cancer Society Signs and Symptoms of Melanoma Skin Cancer.
https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/signs-and-symptoms.html

[2] World Health Organization Skin cancer. https://www.iarc.who.int/cancer-type/skin-cancer/

[3] https://data.mendeley.com/datasets/zr7vgbcyr2/1

As the data set contains 2298 images in total, we decided to extract and analyze 10% of them because of our time limitations. Since the different diseases were not distributed equally, we decided to use each illness' distribution in the original data set to construct a new one that would help us create and modify a model for further training and testing. The distribution is as follows:

Disease	Distribution in %
BCC	36.8%
ACK	31.8%
NEV	10.6%
SEK	10.2%
SCC	8.4%
MEL	2.3%

From the metadata, we only considered the feature informing about whether the lesion is cancerous or not, and all of the other features that were going to be further used were extracted directly from the images.

3 Segmentation

The images in the data set were characterized by numerous parameters with high variability(such as shape, size, color, etc.). Therefore, the better approach for the aim of the investigation was for images to be segmented in a way such that only the lesion is considered (without the background). We first decided on an automatized method (algorithm) for segmentation. The general idea was to create a function that takes an image as input and returns a mask of this image (where the region of pixels in the lesion has a value of 1, and the region that represents the background has a value of 0).

The first method we tried was the snake method, also known as active contour or active contour model. It starts with an initial contour, which can be a simple shape defined by a set of points. The contour is then adjusted to fit the object boundaries in the image. Unfortunately, the results were not promising, and it took a long time to set the proper parameters for the function, so we decided to try a second approach.

The second method tried was the Otsu II algorithm. Otsu's method is a simple technique for automatic image thresholding. It aims to find an optimal threshold value to separate foreground and background regions in an image. In Otsu's method, the algorithm calculates the variance within two classes of pixels (foreground and background) for all possible threshold values. However, the results were also not satisfactory enough to work with them.

The last method discussed, and also the one we used, is manual labeling. In order to achieve this, we used the software Label Studio. We segmented the images manually, one by one. Then the masks and the original image were stored with the same name for easier further work. During the segmentation process, some images were dropped due to being blurry or not being properly visible. The time limitations allowed us to work only on 10% of the data, which has been considered a possible risk for the lower accuracy of the model.

By deciding on manual labeling (segmenting/masking), where it is assumed the correctness of the work done, there is left space for further analysis by extracting features directly and only from the masks or/and images, which eliminates the possibility of differences in the new features passed or/and incompleteness.

4 Feature Extraction

Upon completing the segmented masks, we have decided to create features to follow the ABCD (Asymmetry, Border Irregularity, Color, Diameter) rule, as it is the main way dermatologists detect skin cancer and has been tested for decades. We decided we could not extract Diameter as a feature because the images in our dataset are in different scales and from different sources (cameras, smartphones, etc.) and are therefore not comparable.

To explore to different ways of representing the features and to test the accuracy of our functions, we observed 5 images and decided on a value for their features on a scale from 1-5 for Asymmetry and Border Irregularity, where 1 is not very asymmetric/irregular and 5 is very asymmetric/irregular. For the color features our annotation turned out to not be useful, because we ended up extracting different features that cannot be decided on visually.

4.1 Asymmetry

Following the ABCD rule, firstly, we took the asymmetry into consideration. After examining a couple of previous scientific works, we opted for the approach of building a vector of three measurements, as per Abder-Rahman Ali et al.[4]. They "split the extracted lesions vertically and horizontally across the center into four equal halves". Doing so allowed us to compare the four halves and determine the pixel difference percentage.

We have written a function that returns a score, rounded to 3 decimals, displaying how asymmetric the image is in percentages. We have divided the masks into four sections – top, bottom, left, and right. Subsequently, the right and the top sides were flipped so that we could compare them to the left and the bottom sides, respectively. As a result, we had four cropped masks with pixel values of either 0 or 1. Next, we evaluated the asymmetry by summing up two opposite sides (left - right and top - bottom) and obtaining a final mask consisting of three pixel values – 0, 1, and 2. Since 0 and 2 indicate overlap (0+0 = 0 and 1+1 = 2), we only take the pixels that have a value of 1 - the mismatches. We did this horizontally and vertically, consequently attaining the sum of all the pixels that do not overlap.

After comparing the two halves, the output sum represented the pixel difference of the whole mask. To compute the asymmetry, we divided the overall pixel discrepancy by the total amount of pixels on the mask's inside. Thereby, the result depicted the lesion's dis-symmetry in %, and, for simplicity, we only took the first three decimals. This percentage will be used to train and test the classifier.

4.2 Border Irregularity

The shape of the lesion proves itself to be an essential feature for the assessment of skin cancer. To measure it, we used the perimeter and area of the potentially cancerous skin formations. The values were taken from the manually crafted masks and combined into a single number, providing us with the compactness of the lesions. Compactness is used to tell us how smooth the border of the skin formation is.

As mentioned, we utilized the binary masks to extract the area and perimeter of the lesions. After resizing the images, the values in their representative arrays were rounded so that we could obtain an array of 0s and 1s, where the 1s mark the pixels inside the skin formation. By summing up the array, we acquired the area.

[4] Ali, Abder-Rahman, et al. "Towards the Automatic Detection of Skin Lesion Shape Asymmetry, Color Variegation and Diameter in Dermoscopic Images." PloS One, 16 June 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7297317/.

To compute the perimeter, we used a structuring element. It was utilized to "eat away" the pixels around the border of the lesion. Subsequently, by subtracting the eroded mask from the original input mask, we were presented with the outline of our lesion. By finding the sum of 1s in its representative array, we attained the perimeter of said lesion.

Finally, combining the two features, we yielded a single value representing the shape of the skin formation – the compactness. To measure it, we used the formula[5]

$$compactness = \frac{(perimeter)^2}{4\pi.area}$$

Lower compactness values imply smoother shapes, while higher values indicate more irregular borders.

4.3 Color

Initially, we used the Python library ColorThief to identify the predominant color shades and extract their RGB values as features. However, upon completing this task, we have found several issues with it. Only taking into account the most predominant shades would result in a significant loss of information, as there could be relevant colors linked to skin cancer that would not be extracted because they are not considered predominant. It would also detriment the assessment of color variegation - one of the main indicators of skin cancer. [6] This method would also create many dimensions, as each channel value for each predominant color must be saved into a separate feature. In order to minimize the loss of information, we decided to extract the percentages of common colors instead. We segmented the image using the SLIC algorithm and then calculated the mean color of each segment, which we map to a base color by finding the RGB color values' minimal Manhattan distance to colors in a color dataset[7].

$$d = |R - R_i| + |G - G_i| + |B - B_i|$$

The color with the minimum Manhattan distance calculated using the formula shown above will be the one estimated to be in that segment. By segmenting the image using SLIC and considering the segments, we ensure that the code will be much faster than if we had to check each pixel, or each nth pixel, while still extracting information with as little data loss as possible. We acknowledge the main issue with this approach, mainly the human error introduced by using a data set with color values of certain shades, the choice of which is subjective and based on human perception. Because of this, we also researched introducing completely objective statistical features as an alternative approach. Another issue with this approach is that, because we are using the average color of each SLIC segment, any color average being close enough to shades of black or white to be detected as such from the minimal Manhattan distance will be very rare. Any black or white in the segment will lighten or darken the average color without bringing it close enough to true black or true white unless most of that segment is originally black or white.

As a statistical feature of the colors, we looked into calculating the standard deviation of the three color channels, R, G, and B. Upon more research on the topic, we found that RGB is not an ideal color representation to use because it is heavily influenced by lighting, and our data set, as well

[5]Michael A. Wirth, Ph.D. Shape Analysis & Measurement. Computing and Information Science Image Processing Group. 2004

[6]Ali Abder-Rahman, et al. "Towards the Automatic Detection of Skin Lesion Shape Asymmetry, Color Variegation and Diameter in Dermoscopic Images." PloS One, 16 June 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7297317/.

[7]https://github.com/codebrainz/color-names/blob/master/output/colors.csv

as other real-world data, may vary greatly in the brightness of the picture. Due to the channels being influenced by lighting conditions and correlated to each other, the idea of RGB color representation relies on the relationship between the R, G, and B components, making the individual channels as features a less useful representation. Instead, we found the alternative of using the HSV color format, which is a simple mathematical transformation of the RGB model. The advantage of HSV is that it maps the colors into dimensions easier for humans to comprehend, namely hue, saturation, and value.

In order to use more objective, statistical features to describe colors, we decided to calculate the standard deviation of the HSV channels, which is a common statistical feature used to evaluate color variegation in the context of automatic skin cancer detection[8] After extracting the H, S, and V channels separately and computing their standard deviation values, we added one more statistical feature related to color, representing the variety of colors on each lesion. After observing each channel (H, S, V), the following statements were concluded:

- Hue - the change of the hue value changes the main color of the pixel
- Saturation - the change of the saturation value affects the gray scale amount of the color (the bigger the S, the more vivid the color of the pixel)
- Value - the value channel changes the lightness of the pixel (the bigger the V, the lighter the pixel)

From the above statements, it can be concluded that the variety of colors depends on the hue channel. To better explore and display the range of hue values, we used the Interquartile Range (IQR) . The statistical formula for it is:

$$IQR = Q_3 - Q_1$$

where Q_1 is the median of the first half of the sorted list of values, and Q_3 is the median of the second half. By subtracting them, the range of colors' hue values is given. It can be assumed that the bigger the IQR, the greater the color dispersion. Therefore, the IQR value can be used to show the variability of colors in a lesion, which can be used as a factor for cancer.

4.4 Feature selection

After feature extraction, we had a total of 18 features. To choose the best combination of features, we initially decided to use the SelectKBest function from the *scikit-learn* library. However, the ranking of the features it presented placed Border Irregularity and Asymmetry as being some of the least important. This went against the widely known system dermatologists use for diagnosing cancer, which places great importance on those two characteristics of lesions. That is why we decided to remove correlated features and colors that our color extraction function did not extract very accurately, namely black and white. Ultimately, the chosen features were Asymmetry, Border Irregularity, IQR of the hue channel, standard deviation of the saturation and value channels, and the suspicious colors red, gray-blue, and dark-brown. We also left pink, as it showed in its density plot that it is good at separating the non-cancerous images from the cancerous ones.

[8]Ebrahim Mohammed Senan, Mukti E Jadhav, "Analysis of Dermoscopy Images by Using ABCD Rule for Early Detection of Skin Cancer." Global Transitions Proceedings, 23 Jan. 2021, www.sciencedirect.com/science/article/pii/S2666285X21000017.

4.5 PCA

After doing feature selection and some analysis, we decided on ten features, eight features that dealt with color in addition to Border Irregularity and Asymmetry. We decided to reduce dimensionality by principal component analysis (PCA). This is a method of replacing redundant features with a few new features that adequately summarize the information contained in the original feature space [9]. PCA uses linear projection to transform data into the new feature space. Each of the new components explains a different percentage of the variance. In general, we want to choose the components that explain the biggest part of the variance. After some tests, we decided to transform the entire feature space and bring it down to 5 principal components, which contained 72% of the variance and gave us the best results in classification later on.

5 Classification

For the data classification, we chose binary classification, where the class predicted will be 0 when the image is deemed non-cancerous and 1 for cancerous. Because of the small sample of data points we will be working with, we concluded this approach would be more accurate than trying to do multi-class predictions for each disease. We tested three types of classifiers introduced in the course - K-Nearest Neighbors, Logistic regression, and Decision tree. They are all types of supervised classifiers, meaning that they use the true labels of the data points in the training process. Our method of evaluating the classifiers used machine-learning metrics such as F1 score, accuracy, and recall. They have different formulas that take the results from the confusion matrices (numbers of False negatives, True positives, False positives, and True Negatives).

5.1 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies a data point based on the labels of the closest nearest neighbors in the feature space. When we applied it, we decided to use grid-search to help find the optimal number of neighbors and type of metric/distance. In the end, we decided not to manipulate the distance metric and kept the default Minkowski distance, as we were afraid of overfitting.

Then, we predicted the probability of the labels for each data point. It was up to us to choose the threshold over which the binary class label is set to 0 or 1. Our main focus was on minimizing the errors in classifying cancer, or the False negatives (FN), so we decided to increase the threshold for non-cancer, so more cancerous images will be accurately predicted. This was important to us because skin cancer is a life-threatening disease, and it is crucial for people to get tested even if the probability of them having skin cancer is 40%.

Unfortunately, that increased the number of False positives in the testing set (non-cancerous lesions predicted as cancer). Nevertheless, we decided that this is a necessary trade-off we can make.

Following the "rule of thumb" for the number of neighbors, we tried numbers of neighbors between 5 and 13. The best performance we got for k=7 (seven neighbors). The average results from the cross-validation sets can be seen below. The classifier was called 100 times in a row, and the average value of Accuracy, Recall, and F1 score was computed and reported as follows:

[9] Zheng, Alice, and Amanda Casari. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly, 2018.

Number of neighbors	F1 score	Accuracy	Recall
k=5	0.67	0.62	0.85
k=7	0.68	0.64	0.81
k=9	0.68	0.66	0.79
k=11	0.69	0.67	0.79
k=13	0.69	0.67	0.78

5.2 Logistic regression

Logistic regression is a parametric classifier used in cases where the outcome variable, the predicted class, is categorical in nature. It takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. It also has weights for each predictor variable/feature, which show how the predictors and outcome variable relate to one another. The weights are updated using gradient descent, which minimizes the cost function or the error between predicted and actual values of the classes. The classifier was called 100 times in a row, and the average value of Accuracy, Recall, and F1 score was computed and reported as follows:

F1 Score	Accuracy	Recall
0.69	0.68	0.78

5.3 Decision tree

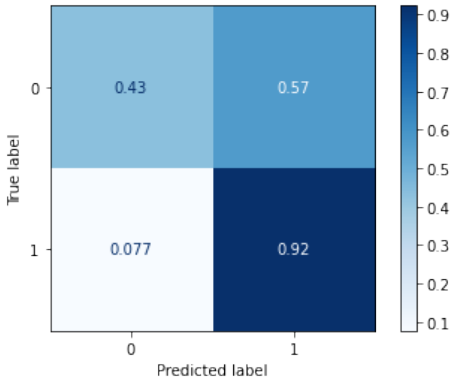
"Decision tree methodology is a commonly used data mining method for establishing classification systems or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure".[10]. The classifier was called 100 times in a row, and the average value of Accuracy, Recall, and F1 score was computed and reported as follows:

F1 Score	Accuracy	Recall
0.56	0.59	0.57

5.4 Results

In summary, based on the overall performance seen from the confusion matrices and the average scores, our classifier was determined to be K-Nearest Neighbors with a number of neighbors set to 5. This was settled based on the desire to have the highest recall while maintaining high enough accuracy and F1 scores.

[10]Y;, Song YY;Lu. "Decision Tree Methods: Applications for Classification and Prediction." Shanghai Archives of Psychiatry, pubmed.ncbi.nlm.nih.gov/26120265/. Accessed 1 June 2023.



(a) Confusion matrix of final model

F1 Score	Accuracy	Recall
0.75	0.71	0.92

(b) Table of metrics from final model

6 Discussion

6.1 Additional study

For the open question part of the project, it was decided that we wanted to investigate how some of the features in the metadata of the given dataset will improve one of the unused classifiers. After observing the metadata, we found out that all of the features which do not have too many missing values are boolean. That gave us the idea to use them and try to improve our Decision tree classifier.

We made that choice because of the structure of the classifier - its most simple form is the Binary decision tree, which makes decisions based on binary features (1 or 0, True or False). The boolean features we decided to use ("itch", "hurt", "bleed", "elevation") aimed to simplify the classification of the model, and as they are part of the diagnosis of many skins diseases, we were also hoping that they will improve the performance of the model. The results from the test are shown below. Again, the classifier was called 100 times in a row, and the average value of Accuracy, Recall, and F1 score was computed and reported.

	THRESHOLD = 0.5	THRESHOLD = 0.6
Accuracy	0.71	0.72
Recall	0.69	0.70
F1 Score	0.69	0.69

It is visible that the Decision tree classifier improves a lot and stabilizes even though it was tried with two different thresholds. This empirical examination of the hypothesis that the boolean features improve a Decision tree classifier showed that involving them in the model can benefit its performance. Furthermore, the data used for the model is provided and is considered true. That is also a factor to be considered, as all other features were extracted directly from the images and are objects to limitations.

Upon completing different combinations of features and implementing several classifiers, the results vary. First, the aim was to reach the highest accuracy possible (without over-fitting the model), but during the work process, it was concluded that the highest overall accuracy might cause the risk of a higher percentage of false negatives. Regarding the topic investigation, it was decided that we can focus on lowering the percentage of false negatives (fewer people not be informed about

their possible cancerous lesion). While this may make our model not good in the sense of accuracy, we look at the task as creating a tool to shorten the period of identifying cancer in earlier stages.

Our expectations, in general, were for better performance due to the large number of features supported by scientific literature. That turned out not to be true. Also, the features themselves differed from the ones we manually annotated. We could not compare the exact values because of the different scales used. While annotators mostly agreed, and the values of asymmetry and border irregularity corresponded in a ranking with the ones perceived by us, the biggest difference was in the number of suspicious colors. There the real feature value corresponded less than 50% with our annotations. This is a very good example of human bias in visual perception and the difference between human vision and algorithms.

6.2 Limitations

Although our model has achieved satisfactory results, there are several limitations we would like to discuss and acknowledge, as well as propose future steps which could be taken to improve the classifier further.

Our classifier has been trained on only 280 images, although the data set we were provided had several thousand images available. This is due to the manual approach we have decided to take regarding image segmentation, as well as a lack of medical knowledge and time limitations. We believe that segmenting more images, using pictures from additional data sets to get a wide-ranging training set, and developing a segmentation algorithm to reduce human error and automate the process would improve the performance of our model.

Another noteworthy limitation of our model is our color extraction features, especially how we decided to extract color percentages. The extraction of color percentages depends on a data set consisting of RGB values mapped to color shades based on human perception and thus introduces an actual concern of human error and bias in the process. We attempted to solve this problem by using statistical measures such as the standard deviation of the predominant colors' channels in HSV format, together with the Interquartile Range of the average Hues of lesion segments delimited using the SLIC algorithm. Nonetheless, these features did not provide enough information to replace the color percentages, and our model performed very poorly when trying to remove them, so we decided to compromise by using both the color percentages and the statistical color measures.

In the future, we would like to develop color features with as little human error as possible and find a mathematical way to describe color, providing enough information for our model.

We also believe that our model could be further optimized to reduce the time complexity of the feature extraction process, especially the color features. We have tried to mitigate this issue by approximating the mean colors of lesion segments delimited using the SLIC algorithm instead of having to check every pixel or every nth pixel, which would be much slower.

Our mean color approach has introduced another issue - the inaccurate detection of white and black. As we are taking the mean of each segment, black and white pixels will only lighten or darken the mean color. Yet, they will only bring it close to true black or true white if most of the segment is made out of black or white pixels, which may not be the case because of the way they are distributed, even if the lesion has an overall high percentage of black or white pixels.

7 Conclusion

Our project aimed to find and extract features from images of skin lesions and then use them for classifying skin cancer. We followed the ABCD rule, which dermatologists use to diagnose

skin cancer, and extracted features regarding asymmetry, border irregularity, and color. We then trained the K-Nearest Neighbors classifier to predict the diagnosis of cancer. While we did not get satisfactory results for predicting non-cancerous lesions, our model performs well on cancerous ones, which was our main focus, given the severity of the disease. Even though we tried to extract and use good features based on the existing literature, our model's general performance is influenced by the small sample of the data set available and the color-extracting method that relies heavily on a set of information created by the human eye's perception of colors.

8 References

1 American Cancer Society Signs and Symptoms of Melanoma Skin Cancer. https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/signs-and-symptoms.html . . . 3

2 World Health Organization Skin cancer. https://www.iarc.who.int/cancer-type/skin-cancer/ 3

3 https://data.mendeley.com/datasets/zr7vgbcyr2/1 . . . 3

4 Ali, Abder-Rahman, et al. "Towards the Automatic Detection of Skin Lesion Shape Asymmetry, Color Variegation and Diameter in Dermoscopic Images." PloS One, 16 June 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7297317/. . . . 5

5 Michael A. Wirth, Ph.D. Shape Analysis & Measurement. Computing and Information Science Image Processing Group. 2004 . . . 6

6 Ali Abder-Rahman, et al. "Towards the Automatic Detection of Skin Lesion Shape Asymmetry, Color Variegation and Diameter in Dermoscopic Images." PloS One, 16 June 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7297317/. . . . 6

7 https://github.com/codebrainz/color-names/blob/master/output/colors.csv . . . 6

8 Ebrahim Mohammed Senan, Mukti E Jadhav, "Analysis of Dermoscopy Images by Using ABCD Rule for Early Detection of Skin Cancer." Global Transitions Proceedings, 23 Jan. 2021, www.sciencedirect.com/science/article/pii/S2666285X21000017. . . . 7

9 Zheng, Alice, and Amanda Casari. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly, 2018. . . . 8

10 Y;, Song YY;Lu. "Decision Tree Methods: Applications for Classification and Prediction." Shanghai Archives of Psychiatry, pubmed.ncbi.nlm.nih.gov/26120265/. Accessed 1 June 2023. . . . 9