IT UNIVERSITY OF COPENHAGEN

FIRST YEAR PROJECT 2

Classification of Melanoma in Skin Lesions by Image Analysis

Authors
Viggo U. GASCOU
Frida N. LAUSTSEN
Magnus S. JAHN

Authors
Marie H. PETERSEN
Gustav KRISTENSEN

Supervisor
Veronika CHEPLYGINA

April 7 th, 2022

1 Introduction

Melanoma is the deadliest type of skin cancer. It kills thousands of people worldwide every year (Global Coalition & Euro Melanoma, 2020). It is estimated that the number of cases of melanoma from 2020 to 2040 will have increased by 13.2 % for European countries. For the same period the number of deaths will have increased by 25.5% (Ferlay et al., 2020). Early diagnosis of melanoma increases the patient's survival rate to 97.6 % if diagnosed in stage I-II. (U. S. National Institutes of Health, n.d.) Although melanoma can only be confirmed with a biopsy, it is usually identified by a physical examination. The doctor looks for specific features, such as the "ABCDE's", that can indicate a malignant skin lesion.

The objective of this report is to help a local dermatologist develop image analysis algorithms to extract features from images of skin lesions. To help them save time in the clinic, we were asked to develop a model from the extracted features with the purpose of classifying skin lesions. To achieve this, we have determined which features to extract and which image analysis algorithms to use for the task, as well as the reliability of those features. We aim to shed light on the following research question:

> *To what degree can we optimize a dermatologist's process of determining melanoma in skin lesions using a classifier trained on features extracted from images of skin lesions?*

In addition to this primary research question, we were asked to investigate another research question using some of our data and findings from the investigation of the first question. We decided to look at some of the additional metadata available for the images from ISIC. We used some of this available metadata to train an additional classifier. We will look at how this classifier performs differently from the one that only uses images and segmentations. By doing this, we aim to shed light on the question

> *"How does the inclusion of additional image metadata in the model impact the performance of the classifier?"*

2 Data

We were provided with data sets for 150 images, including image files, segmentations, additional metadata containing information about age and sex of the subjects as well as a ground truth file. In addition to this, we downloaded from the same ISIC 2017 challenge a training data set containing 2000 images, corresponding segmentations, age/sex metadata, ground truth, additional json metadata for each image and a superpixel mask for each image. 17.35% of the skin lesions in the data set are melanoma, the rest are benign. The majority of the images are of high quality, with few exceptions. We decided to keep the low quality images, since filtering these would either have to be done manually with some self-developed guidelines or by developing a preprocessing classifier of which images to discard. We downloaded an additional data set of 600 test images from the ISIC 2017 archive for our secondary research question.

To decide whether a skin lesion is cancerous, there are different features to take into account. These features are commonly referred to as 'ABCDE'. Malignant skin lesions are very often asymmetrically shaped, whereas benign lesions are typically uniform and symmetrical in shape. The border of the skin lesion is also important to take into account. When the border of the lesion is not well defined, irregular or scalloped, the lesion is often cancerous. A benign skin lesion has a well defined and smooth border. Melanoma often has varying colors in different sections of the lesion, which makes it easier to distinguish from benign lesions. A benign skin lesion will usually have one shade of brown while malignant skin lesions will typically have several shades of brown, tan, black and in some cases even red, white or blue. (Halpern et al., 2021)

It is essential to take into consideration that malignant skin lesions vary greatly in appearance. The malignant lesions may show all of the above mentioned features, while other lesions may only show one or two. For the manual annotations, we were asked to rate whether melanoma was present in an image of a skin lesion based on the indicators mentioned above. We also rated the asymmetry, border and color on a scale from 0 to 10. Even though we all used the same scale when rating the aforementioned features, the scale might be different on an individual level. Each person might have a different interpretation of where a given feature of a lesion lies on the scale. We manually classified 60 of the 150 provided skin lesions. The data generated from this was saved in a csv file, to later be compared with the measurements from our classifier.

3 Feature Extraction

We will now look at the different features we have measured with functions to use in the classifier.

The first couple of features fit in the category asymmetry. Most melanoma skin lesions are asymmetrical. That means if you draw a line through the middle of the lesion, the two halves don't match. To implement this, we first need to evaluate whether the lesion segmentation mask is worth measuring. We measured whether 25% of the lesion was within 1% of the image border. If it was, we gave it the maximum asymmetry score of 4.

If the image was not given the maximum score initially, we continued with measuring the asymmetry. We rotated the segmentation, so the widest point in the mole became aligned with the line going vertically through the center of the image. Then we cropped the image to exactly fit the rotated lesion. To evaluate the asymmetry, we flipped the segmentation array and added it to the non-flipped array. Then we got the amount of non-overlapping pixels and divided by the total area of the lesion to get a score between 0 and 2, where 2 is most asymmetric. This was done both vertically and horizontally. In addition, we smoothed the rotated segmentation using a Gaussian blur. We repeated the flipping and addition to find asymmetry scores for the smoothed segmentation as well. This was done as many of the segmentation's had very jagged edges, and we wanted to see if smoothing them would yield a more informative asymmetry feature. The vertical and horizontal asymmetry scores were added together, giving two features ranging from 0 to 4.

We also had third feature as a measure of asymmetry. We constructed a circle at the center of the rotated and cropped segmentation. The circle had a diameter equal to the width of the cropped image. We would then measure the amount of pixels in the lesion that were not overlapping with the circle. This was divided by the area of the circle, to return the percentage overlap with the circle. While this does measure asymmetry, it assumes the lesion is shaped like a circle, while a symmetric lesion could be oval shaped.

Another important lesion feature to measure is the border. One way to measure the irregularity of the lesion's border is to compare its area to its perimeter. To measure the area of the lesion, we simply summed the binary segmentation array. The actual area of the lesion was not possible to measure, as the images are not all provided with scales. The perimeter of the lesion was measured by convolving a kernel on the segmentation array. The resulting array would have negative values where the edges were, and positive values elsewhere. The amount of negative pixels were summed as the perimeter measure. An additional feature compactness can be calculated as $c = (l^2 4\pi)/A$. This score indicates the compactness of the shape compared to a circle, where a circle has compactness 1 and all other scores fall between 0 and 1.

Color is the last important lesion feature that we were able to measure. We computed three different features in this category. For the first one, we applied the segmentation masks to the images and divided the lesion into multiple sections using Simple Linear Iterative Clustering (SLIC). For the first color feature, we computed the average color of each section. Then we computed the euclidean distance from each color to all other colors and took the mean. For the second feature we computed the standard deviation of all the average colors of the sections. We did this for all combinations of using 5 and 10 segments and sigma values 5 and 10 for the SLIC function. Lastly, we replicated a method previously used by other researchers to detect how many different colors were present in each lesion (Majumder & Ullah, 2019). Each picture was processed and given a color score between 1 and 6. 1 meaning one color is present and 6 meaning all six predetermined colors present. The colors and their respective thresholds are taken from table 1 of (Majumder & Ullah, 2019).

We used multiple methods for evaluating which features to chose to train our classifier. We looked at KDE plots for each feature to see the overlap between the feature scores for benign and malignant lesions. The less the overlap, the better the feature is on its own. No features were completely separated, but some were clearly better than others. For example, the area and perimeter have less overlap than color distance with 5 segments and sigma 10 (figure 1) While kernel density plots gives a good intuition about how two classes (e.g. melanoma and not melanoma) differs, it does not provide us with something quantifiable. Therefore, to support our intuition about which features to pick for our classifier, we computed the mutual information for each feature. This provided us with a tangible metric to guide our final feature selection (figure 2.)

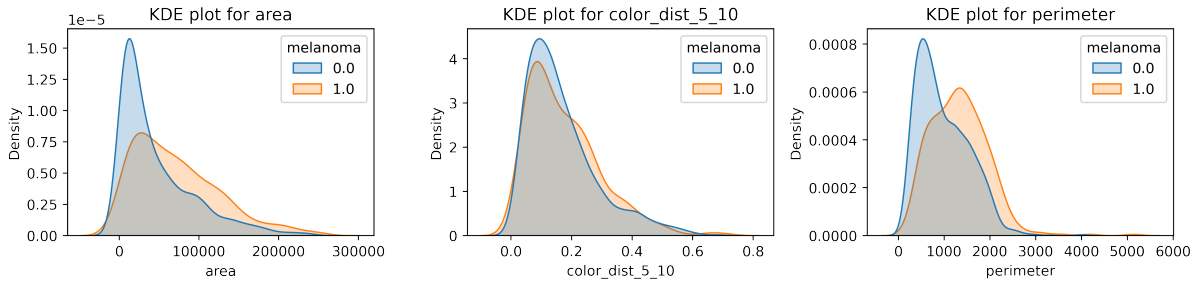

Figure 1: KDE plots for features

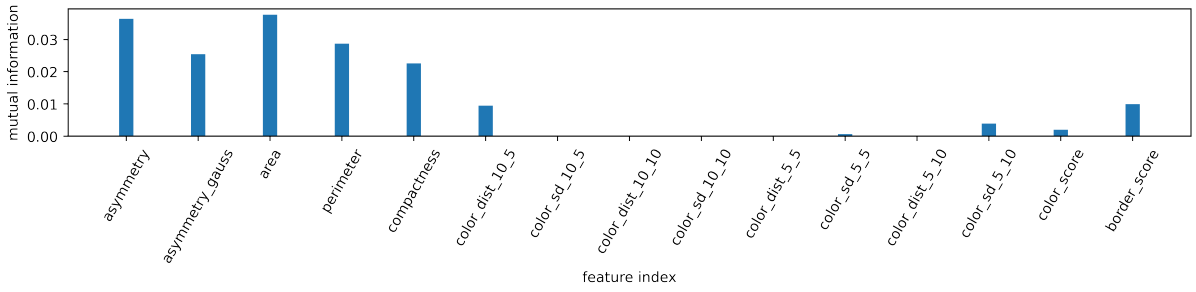

Figure 2: Mutual information for features

Additionally, we tried selecting features using sequential feature selection, both forwards and backwards. We selected 6 features in both directions, and saw color_dist_10_10, perimeter, compactness and color score in both. From all the different feature selection methods, we selected 5 features that performed well in one or more of the selections. Some features were discarded, as overlapping features are not ideal. For example, we chose area and perimeter, but not compactness, as compactness is calculated with the other two.

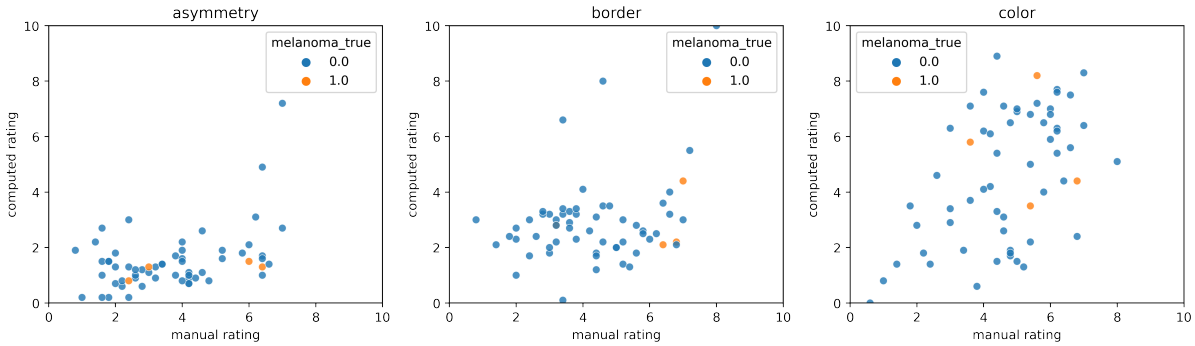

Figure 3: Comparison of manual and computed feature scores

We also also plotted the color standard deviation against the average color distance from each picture, and found a 97% correlation with a significant p-value, so we scrapped the color standard deviation. The color distance measure with 10 sections and sigma 10 performed best and was chosen. Different feature combinations were tested and evaluated based on the score of a KNN classifier trained on them. The combination of features that were ultimately chosen for the classifier were area, asymmetry, color distance, color score and perimeter.

We compared the measured features from the feature extraction to our manual classification. To do this, we transformed the features into the same A, B and C that we manually measured, on a scale from 0 to 10. The features were normalized to values between 0 and 1, then split into the categories as described above and averaged. Then they were multiplied by 10 to end up between 0 and 10. For our manual classifications, we averaged our measurements for each feature, and for the melanoma label we took the consensus majority as the label. We plotted the measurements from feature extraction against our manual measurements (figure 3). On this particular subset of pictures, our manual classification bested our model on AUC-ROC score ($0.80 > 0.35$), but the accuracy of the model was slightly better ($0.52 < 0.62$).

4 Classification

After selecting the features to use for the classifier, we found the best parameters to use for our model. We used a K-Nearest Neighbors classifier for the model and the training data set. A KNN-model is trained on labeled data, and when given a data point to classify, it finds the k nearest points and gives a label based on the majority label. The KNN algorithm assumes that similar things exist in close proximity (Harrison, 2019). We started by scaling the features to values between 0 and 1, so they would all have equal influence in the model. To test the performance of the classifier, we split our data set into 5 using `sklearn.model_selection.StratifiedShuffleSplit`. For each set of training/validation data, we trained the classifier on the training set and predicted labels and probabilities on the validation set. The resulting accuracy and AUC-ROC scores were saved and averaged at the end. We tried using different values of k to find the optimal one. The accuracy of the model increased with k, and started flattening out at around $k = 9$. With a k-value at 9, the results of the AUC score was 0.644 ± 0.01.

For a clear visualisation of the performance of our model, we created a confusion matrix. The confusion matrix, also known as an error matrix, represents the true label on the vertical axis and the predicted label on the horizontal axis. The predicted label was found by applying a threshold to the predicted probabilities of melanoma from the classifier.

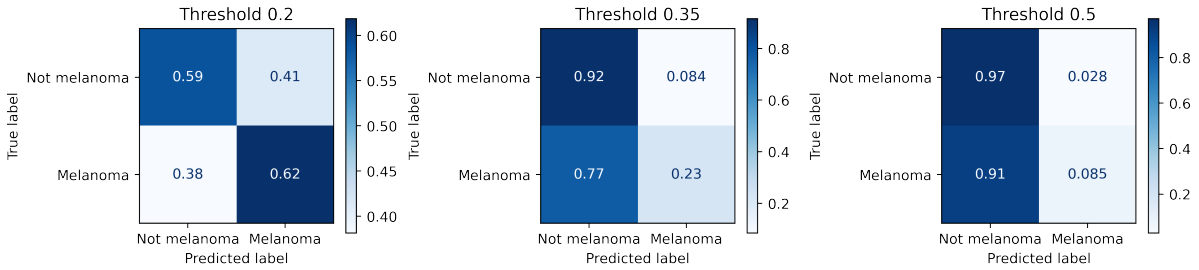

Figure 4: Normalized confusion matrices for different thresholds

In order to motivate our choice of threshold, we made confusion matrices with different thresholds. Decreasing the threshold corresponded to an increase in the amount of lesions being classified as melanoma (see figure 4).

We chose a threshold of 0.2 to balance the amount of false positives and false negatives. Ideally we would like to minimize the amount of false negatives. A false negative means a patient with a cancerous lesion has their lesion evaluated as benign. This could have deadly consequences, so we want to minimize that number. On the other hand, classifying too many false positives will almost all patients must have their lesion evaluated properly by a dermatologist, defeating the purpose of the classifier.

We decided to use additional data from ISIC to look at another research question: *"How does the inclusion of additional image metadata in the model impact the performance?"* To answer this question, we used the superpixel masks and JSON metadata available for the images on ISIC's website. The JSON metadata contains binary arrays for four features, where a 1 at index i of the array means the feature is present at the superpixel with index i. ISIC provides pseudocode for a function `decode_sp_index` that reads the RGB value of each index in the superpixel mask and converts it to the corresponding superpixel index. The four features in the JSON metadata are pigment networks, negative networks, milia like cysts and streaks. Pigment networks are networks of darker colored areas forming a grid in the skin lesion. Negative networks are the negatives of pigment networks, lighter areas comprising a grid around darker areas in the skin lesion. Negative networks are highly specific to melanoma (Dermoscopedia, 2019a). Milia like cysts are small cysts on the lesion that mostly occur in seborrheic keratosis as opposed to melanoma (Stricklin et al., 2011). Streaks are pigment projections at the border of a lesion. They occur during the growth of a lesion, which could likely indicate melanoma (Dermoscopedia, 2019b).

To utilize the JSON data for feature measurement for an image, we converted the image's superpixel mask to a 2D-array where each element has the value of the corresponding superpixel index. Then for each of the four JSON features, a new binary array was created, where each element had value 1 if the feature was present in the corresponding superpixel index. Then we labeled the connected components in each array, and the sizes of the components were saved. This was done for all 2000 training images, and then we computed 4 equipopulated bins for each feature. With the bins computed, the measuring function for the JSON features made a histogram of the component sizes with the calculated bins. This gives a total of 16 features, 1 for each bin, for each feature. As an additional 4 features, we computed the percentage of superpixels with a feature present out of the total possible amount, for each feature. This gives a total of 20 features extracted from the superpixel masks and JSON metadata.

We trained a the same KNN-classifier with $k = 9$, but with the JSON data included. The classifier performed poorly with all the features included in the fitting, so we selected the 5 best JSON features using a mutual information selector. The classifier was then trained with both the ABC features and the chosen JSON features.

Classifier	Accuracy score	AUC-ROC score
ABC only classifier	0.507	0.596
Classifier with JSON features	0.563	0.61

Table 1: Accuracy scores for the two classifiers

To test the difference between the two classifiers, we tested them both on the same 600 image test set. The resulting accuracy and AUC-ROC scores are shown in table 1. Based on our results, it would not make sense to spend time labeling the features represented in the JSON metadata. The JSON classifier barely outperforms the ABC classifier. At the same time, it is a requirement for the JSON classifier that images to be classified must come with corresponding superpixel masks and JSON metadata. Labeling the metadata is a task that is not worth doing, when the improvement in the model is so marginal. However, it is possible that more could be interpreted from the metadata. Through a better understanding of the features and their correlation to melanoma, and more advanced algorithms for measuring higher level features from the metadata.

5 Discussion and Conclusion

During the research we discovered some limitations with the available data sets. First of all the quality of the provided images varied a lot. Some of the images were taken close up in order to get more detail of the skin lesion. This resulted in some of the lesions being outside the frame of the image, making it difficult to measure its asymmetry. A portion of the images also had hair covering the skin lesion, which could affect our models' measurements of color. An additional limitation related to the data set is the inability to measure diameter and evolution, the D and E of the ABCDE features. With no scale on most images and only one image per lesion, there is no way to measure the diameter of the lesion or its evolution over time. Another major limitation was related to work space constraints. We had very limited time to conduct the research. With more time we could have gathered more knowledge on both skin lesions and image analysis. This would have allowed us to develop better algorithms to measure features in the lesions, which could improve the accuracy of the classifier.

One of the goals of this research was to create a model that could be used by dermatologists in conjunction with physical inspection of lesions to alleviate their workload. It should be stated that our model is not nearly accurate enough to be used in a clinical setting. Even with our attempt to minimize false positives, the model still identifies only about half of melanoma cases, and falsely labels about half of all benign cases. It could be interesting for future research to look at data sets containing the necessary information to measure diameter and evolution of lesions. The evolution of melanoma lesions is nuanced, as they can evolve in both size, shape, color and elevation. Very often large lesions (diameter greater than 6mm) are melanoma, so the diameter could be an informative feature. It would also be interesting to do further research into image processing and improvement of our feature extraction algorithms or development of new ones.

In conclusion, we have developed a classifier that can identify melanoma with a certainty better than a random guess. But it is not accurate enough to use as a tool alongside physical inspection of lesions as envisioned. Further research must be done into the research area to improve the feature extraction algorithms and the performance of the classifier. Including additional features from provided metadata can very slightly increase the performance of the classifier. However, the marginal performance gain from the inclusion of this metadata does not justify the workload that would go into creating the metadata in a clinical setting.

References

Stricklin, S., Stoecker, W., Oliviero, M., Rabinovitz, H., & Mahajan, S. (2011). Cloudy and starry milia-like cysts: How well do they distinguish seborrheic keratoses from malignant melanomas? *Journal of the European Academy of Dermatology and Venereology : JEADV*, *25*(10), 1222–1224. https://doi.org/10.1111/j.1468-3083.2010.03920.x

Dermoscopedia. (2019a). *Negative pigment network - dermoscopedia* [Dermoscopedia.org]. Retrieved April 7, 2022, from https://dermoscopedia.org/Negative_Pigment_Network

Dermoscopedia. (2019b). *Streaks - dermoscopedia* [Dermoscopedia.org]. Retrieved April 7, 2022, from https://dermoscopedia.org/Streaks

Harrison, O. (2019, July 14). *Machine learning basics with the k-nearest neighbors algorithm* [Medium]. Retrieved April 7, 2022, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Majumder, S., & Ullah, M. A. (2019). Feature extraction from dermoscopy images for melanoma diagnosis. *SN Applied Sciences*, *1*(7), 753. https://doi.org/10.1007/s42452-019-0786-8

Ferlay, J., Laversanne, M., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2020). *Global cancer observatory: Cancer tomorrow. lyon, france: International agency for research on cancer.* Retrieved April 7, 2022, from https://gco.iarc.fr/tomorrow

Global Coalition, & Euro Melanoma. (2020). *2020 melanoma skin cancer report*. Retrieved April 7, 2022, from https://melanomapatients.org.au/wp-content/uploads/2021/09/2020-campaign-report-GC-version-MPA_1.pdf

Halpern, A. C., Marghoob, A. A., & Reiter, O. (2021). *Melanoma Warning Signs and Images*. Retrieved March 31, 2022, from https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/

U. S. National Institutes of Health. (n.d.). *Seer training modules, five-year survival rates. u. s. national institutes of health, national cancer institute.* Retrieved April 7, 2022, from https://training.seer.cancer.gov/melanoma/intro/survival.html