IT UNIVERSITY OF COPENHAGEN

Project 2: Medical Imaging

First Year Project

BSFIYEP1KU

Group 2:

Anna Gnat (agna@itu.dk)
Petya Petrova (pety@itu.dk)
Alexis Serruys (serr@itu.dk)
Bogdan Mihaila (bomi@itu.dk)

Github: https://github.com/m3bogdan/Project-2-Medical-Imaging

BSc in Data Science
IT University of Copenhagen
June, 2023

Contents

1 Introduction

1.1 Background

Skin cancer is the most common cancer type[Foundation, 2020]. In fact, it is so common that in the U.S. more people are diagnosed with skin cancer than all the other cancer types combined. Early detection of skin cancer is crucial when it comes to the survival rate of the patient. Visual examination and imaging of the skin lesion are the most frequent methods for diagnosing skin cancer. If there is a suspicion that the lesion is cancerous, a biopsy is then performed. Therefore the research on automated detection of cancerous skin lesions is important as it can increase the chances of detecting skin cancer early on.

1.2 Problem statement

The goal of our report is to conduct an investigation into the feasibility of reliably measuring specific features of skin lesions through the utilization of image analysis algorithms. We also set out to investigate the importance of the consistent quality of images for developing prediction models.

1.3 Research objectives

In order to evaluate the performance of the trained machine learning models, we will predict the diagnosis of unseen skin lesion images and assess the models' performance using appropriate evaluation metrics on different quality datasets. The expected outcomes of this investigation include the development of image analysis algorithms capable of extracting relevant features from skin lesion images, the creation of machine learning models trained to accurately classify skin lesions as benign or malignant, and the evaluation of their performance compared to human dermatologists' diagnoses. Through this research, we anticipate gaining valuable insights into the potential benefits and limitations of using automated systems for skin cancer detection, especially taking into account the consistency of image quality, paving the way for improved diagnostic capabilities and enhanced patient care.

1.4 Literature review

Automated skin lesion classification techniques are the subject of extensive research, as early skin cancer identification is important for effective treatment. As the basis for our study, we used several different articles and research papers. The most significant aspects are summarized in this section.

1.4.1 Automation methods for skin lesion diagnosis

The automation for the process of skin lesion detection and classification has rapidly increased over the last few years. The possible different computer-aided systems for diagnosing cancerous lesions vary in complexity, computational power and accuracy of the predictions [Kassem et al., 2021]. The methods span from traditional machine learning methods to complicated deep learning models. Different methods have different advantages and disadvantages, so choosing an appropriate method for the project can be crucial.

1.4.2 Algorithms for skin lesion classification

In dermatology, there are several different classification methods when it comes to recognizing cancerous skin lesions and there is no standard across the field. The most widely used ones are [Carrera et al., 2016]:

- ABCD rule,
- the Menzies method,
- the 7-point checklist,
- chaos and clues,
- and CLASH.

All of the mentioned classification methods yield similar results when used by trained professionals. According to [Argenziano et al., 1998], using the 7-point checklist can lead to better results when used by less experienced observers. This was the reason why we decided to proceed with the 7-point checklist as our method of choice.

1.4.3 7-point checklist method

The seven-point checklist was updated in 2011 and the weighting of the features was modified [Carrera et al., 2016]. It went from a split of three major and four minor features to giving equal weight to each of the seven features. We used the current version of the classification system, as shown in Table 1 [Argenziano et al., 2011], to design the feature extraction algorithms.

ELM criterion	Definition	Score
Atypical pigment network	Black, brown, or grey network with irregular meshes and thick lines	1
Blue-whitish veil	Confluent, grey-blue to whitish-blue diffuse pigmentation associated with pigment network alterations, dots/globules and/or streaks	1
Atypical vascular pattern	Linear-irregular or dotted vessels not clearly combined with regression structures and associated with pigment network alterations, dots/globules and/or streaks	1
Irregular streaks	Irregular, more or less confluent, linear structures not clearly combined with pigment network lines	1
Irregular pigmentation	Black, brown, and/or grey pigmented areas with irregular shape and/or distribution	1
Irregular dots/globules	Black, brown, and/or grey round to oval, variously sized structures irregularly distributed within the lesion	1
Regression structures	White areas (white scar-like areas) and blue areas (grey-blue areas, peppering, multiple blue-grey dots) may be associated, thus featuring so-called blue-whitish areas virtually indistinguishable from blue-whitish veil	1

Table 1: Criteria and scores according to the 7-point checklist method

2 Data Exploration and Preprocessing

2.1 Data Description and size

We were provided with 3.4GB dataset in which we can find 1373 patients from different countries, 1641 skin lesions, and 2298 images. The metadata contains up to 22 clinical features for every patient, including age, skin lesion location, Fitzpatrick skin type, and skin lesion diameter. The skin lesions can be split in two main categories: skin cancer (Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Melanoma (MEL)) and skin disease (Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), and Nevus (NEV)). Additionally, in order for us to test the impact of image quality on the quality of automated diagnosis, we gathered another dataset made of 2.7GB containing 1227 high-quality images. The metadata provided with this dataset only contains the patient ID and the type of skin lesions the patient has.

2.2 Data cleaning process

After exploring the csv file that contained various metadata connected to skin lesion images, we divided them into six categories based on the skin lesion diagnosis type. To prevent the imbalance in the training data, we took 20 pictures from each category. That sums to 120 pictures in total, half cancerous and half non-cancerous that we have used to train and test our model. In order to investigate our research question, we had to find another dataset with better quality pictures which contains the following categories: MEL,ACK,NEV,DF (dermatofibromas),BCC. We chose 25 pictures out of each category, except for pictures with DF (choosing only 7). In the data-set with better-quality pictures, we chose the same amount of cancerous and non-cancerous pictures.

3 Methods

3.1 Data selection and splitting

We focus on achieving data balance by employing techniques such as oversampling or under-sampling to address the potential bias arising from imbalanced classes. We determine a split ratio of 80 percent for training and 20 percent for testing, ensuring a substantial training set while reserving a separate portion for an unbiased evaluation. We randomize the data-set prior to splitting to eliminate ordering bias and ensure that there was not more than one image from the same patient present in either the training or testing set, but not both, to prevent information leakage. These measures aim to create a balanced and representative data-set, enabling our classifier to generalize effectively and accurately identify cancerous and non-cancerous lesions.

3.2 Manual feature extraction and classification

To establish a common understanding of the 7-point checklist method, we created a few training sets of images to manually extract the features and classify the lesion as either cancerous or non-cancerous. First, we took 10 random images, discussed each potential feature together and assigned scores, accordingly to the chosen method. We saved the results in a CSV file for future reference. Afterwards, we created additional two sets of ten images each and annotated those individually (see appendix B for an example of individual scoring). Each set of images was annotated by two group members so that we were able to compare the results of our extraction and further discuss the features if there were any discrepancies.

3.3 Algorithm Design and Feature Extraction

In order to automatically extract our features we used some of OpenCV's and SciKit-Image's powerful functions. The code for all the features can be found on the GitHub page for this project (link on the title page).

3.3.1 Feature 1: Atypical pigment network

Using OpenCV's image processing capabilities, the function extracts and enhances relevant features, generating a binary mask that separates the pigment network from the background. The coverage percentage, obtained by comparing pigment pixels to the total number of pixels.

3.3.2 Feature 2: Blue whitish veil

To detect a blue veil, we had to loop through each pixel of an image and try to satisfy this condition:

$$B > 60, R - 46 < G < R + 15$$

[Madooei et al., 2019] in the RGB space. This threshold was set by Ogorzalek, who used wavelet [Piatek and Mroczek, 2022] decomposition to set these boundaries. Wavelet decomposition is a way of breaking down or separating a signal of an image into different parts to analyze it in more detail.

3.3.3 Feature 3: Atypical vascular pattern

To measure atypical vascular patterns we decided to measure the amount of "red" pigmentation in the images - following our strategy from manual assessment. First, we enhance the red channel of the image and convert the image to the HSV colour scale. Based on that, we determine the numerical range for the colour red and create a binary mask. The code then returns the number of pixels that were within that range, therefore classified as red.

3.3.4 Feature 4: Irregular dots/globules

We measure the presence of globules in the lesion by implementing a blob detection technique using one of the algorithms from the skimage collection [Scikit-Image,]. First, we convert the skin lesion image to grayscale and invert it. Then we employ the blob-detection function. The result is the amount of detected globules in the image.

3.3.5 Feature 5: Irregular streaks

In order to detect irregular streaks we color the image gray and apply an adaptive threshold so we can detect the contours. After detecting the contours we compute the lesion area and the lesion border. We use the formula stated in [Golston et al., 1992] to detect the irregularity.

3.3.6 Feature 6: Irregular pigmentation

In order to quantify the irregular pigmentation coverage percentage in medical images. To begin, we transformed the image into grayscale and applied Otsu thresholding, creating a binary image that emphasized areas of irregular pigmentation. By labeling the regions in this binary image and calculating their circularity, we were able to identify regions with circularity below 0.6 as indicative of irregular pigmentation. With this information, we determined the coverage percentage by comparing the number of irregular pixels to the total pixel count

3.3.7 Feature 7: Regression structures

Firstly, we transform the image color format to Hue, Saturation, and Value (HSV). Next, we select a lower bound ([0, 0, 150]) and an upper bound ([180, 30, 255]). After that, we create a mask using these bounds. Finally, we count the number of pixels that fall within these specified bounds.

3.4 Feature extraction techniques

Considering our model's seven dimensions, we employed Principal Component Analysis (PCA) for feature extraction. PCA, a widely recognized technique in the field [Addison et al.,], is known for its exceptional performance in a variety of cases.

3.5 Classifier training and evaluation methods

In the process, we've implemented several classifiers: logistic regression (LR), K-nearest neighbours (KNN) and decision tree classifier (DTC). We work with both full models with all the features included and with simplified models using PCA feature extraction technique. All models are trained on extracted features from our dataset. We also apply a 5-fold cross-validation method to ensure our results are robust across different sample selections. Then we evaluate the models with different performance metrics. The model with the highest F1 score, a harmonic mean of precision and recall, is selected as the best model for this task. This model is then ready to make predictions on new data, demonstrating its utility in ”real-world” applications.

4 Result and Analysis

4.1 Presentation of classification results

The results from training all the models can be seen in the tables below. The statistics we decided to show are:

1. Precision: $\frac{TP}{TP+FP}$ - the percentage of true cancerous lesions among all the predicted cancerous lesions,

2. Recall: $\frac{TP}{TP+FN}$ - the percentage of predicted cancerous lesions compared to all the true cancerous lesions,

3. F1 - score: $2 * \frac{Precision*Recall}{Precision+Recall}$ - a harmonic mean of precision and recall.

Classifier	Precision	Recall	F1-score
LR (with PCA)	0.667	0.444	0.533
LR (without PCA)	0.623	0.611	0.617
KNN (with PCA)	0.686	0.648	0.667
KNN (without PCA)	0.686	0.648	0.667
DTC (with PCA)	0.500	0.463	0.481
DTC (without PCA)	0.490	0.444	0.466

Table 2: Relevant scores for different versions of trained classifiers trained on low-quality images.

Classifier	Precision	Recall	F1-score
LR (with PCA)	0.782	0.907	0.840
LR (without PCA)	0.782	0.933	0.897
KNN (with PCA)	0.783	0.867	0.823
KNN (without PCA)	0.783	0.867	0.823
DTC (with PCA)	0.689	0.680	0.685
DTC (without PCA)	0.786	0.733	0.759

Table 3: Relevant scores for different versions of trained classifiers trained on high-quality images.

Classifier	Precision	Recall	F1-score
LR (with PCA)	0.636	0.581	0.607
LR (without PCA)	0.610	0.814	0.698
KNN (with PCA)	0.629	0.682	0.654
KNN (without PCA)	0.631	0.690	0.659
DTC (with PCA)	0.634	0.659	0.646
DTC (without PCA)	0.624	0.605	0.614

Table 4: Relevant scores for different versions of trained classifiers trained on images with high and low quality

To further investigate the impact of the quality of images on the effectiveness of our classifiers, we took the best-performing classifier (based on the scores mentioned above) from each category of image quality and tested it on different image samples. These samples again reflected the three categories of image quality. The results are shown in the tables below.

Image quality	Precision	Recall	F1-score
low	0.666	0.666	0.666
high	0.125	0.500	0.200
mixed	0.367	1.000	0.537

Table 5: Test scores for KNN model trained on low-quality images.

Image quality	**Precision**	**Recall**	**F1-score**
low	0.000	0.000	0.000
high	0.000	0.000	0.000
mixed	0.367	1.000	0.537

Table 6: Test scores for LR model trained on high-quality images.

Image quality	**Precision**	**Recall**	**F1-score**
low	0.000	0.000	0.000
high	0.000	0.000	0.000
mixed	0.333	0.545	0.414

Table 7: Test scores for LR model trained on mixed-quality images.

4.2 Evaluation metrics and interpretation

The K-Neighbors Classifier performs great in both cases, with or without PCA, with low-quality pictures, as shown in Table 2. It has the highest precision and recall scores compared to other classifiers. So, for the dataset that we fed the model with, the K-Neighbors Classifier outperforms the others.

As shown in Table 3, the Logistic Regression classifier consistently achieves high precision and recall scores in both cases, with and without PCA, with high-quality pictures. The F1-score is slightly better in the model without PCA, meaning that the PCA method does not improve the performance. Furthermore, The K-Neighbors classifier also performs well. On the other hand, the Decision Tree Classifier performs poorly compared to the other classifiers. Overall, based on the metrics, the Logistic Regression classifier shows the best performance compared to the other classifiers in the case of training on good-quality images.

All classifiers, that have been fed with the mixed-quality images with PCA, show very similar performance. The PCA method is improving only the model using the decision tree classifier, meaning that it did not improve the rest of the models that are using LR and KNN classifiers. The model using Logistic Regression without PCA outperforms the other models by having a 0.698 F1 score and the lowest number of false negatives.

Therefore we decided to save the following models: kNN without PCA trained on low-quality images, LR without PCA trained on high-quality images, and LR without PCA trained on mixed-quality images and test them in order to investigate which one performs best and choose one.

Table 5 gives the results of KNN classifier, trained on low-quality images but tested on all possible data sets: low-quality, high-quality, and mixed-quality images. When comparing all the F1-scores of these models, we would expect our models to perform similarly, which is not the case. Therefore it suggests that our models might be over-fitted. The F1-score for the low-quality images was reduced from 0,66 to 0,2 for KNN tested with high-quality. KNN tested on mixed-quality images versus tested on low-quality images give very close

F1 scores, having the one tested on low-quality having slightly higher scores. Overall, KNN tested on low-quality data is a better model compared to KNN tested on mixed-quality data, because of its precision score of 0.66 compared to 0.36.

Table 6 gives the results, of the LR classifier trained on high-quality images tested on all the possible datasets. The F1-score for LR tested on low-quality images and high-quality images is zero compared to LR tested on mixed-quality data, which has an F1 score of 0.536. This indicates the over-fitting of the models.

Table 7 gives the results of the LR classifier trained on the mixed-quality dataset and tested on all possible datasets. The results are similar to LR trained on high-quality images. The F1-scores are zeroes for LR tested on low- and high-quality data, whereas LR tested on mixed-quality data gives the F1-score of 0.413, which again indicates over-fitting.

Overall, based on the testing we concluded that all of our models were over-fitted, meaning that they did well on known data, however, they could not output good predictions when tested, except for KNN classifier trained on the low-quality images, which scored 0.666 F1 score on predictions and precision of 0.666 when tested with low-quality images. Since KNN did well when tested on both low-quality images and mixed data compared to LR, which performed well only on mixed data, we can conclude that KNN tested on low-quality images performs best.

5 Discussion

5.1 Summary of Findings

In this study, we successfully utilized machine learning methods to train and evaluate classifiers for the task of medical image analysis. Here are the key findings and insights we gained:

1. **Effective Feature Extraction:** Through our custom feature extraction methods, we were able to derive meaningful attributes from the images, such as pigment network coverage, number of globules, and presence of vascular structures, among others. These features captured important characteristics of the images that significantly aided in the classification process.

2. **Importance of Dimensionality Reduction**: Given the difference in the performance of our models on the training data versus testing data, we can see the high impact of over-fitting the data on the performance of the models.

3. **Balanced Performance is Crucial**: By using balanced classes to train the model and the F1-score as our primary evaluation metric, we ensured that our chosen model performed well across all classes, not just

the majority ones. This is vital in medical applications where incorrect classifications can have serious implications.

4. **Scope for Model Improvement**: We need to acknowledge that the model has massive shortcomings when it comes to generalization. The potential improvements should include using a larger dataset and counteracting the over-fitting by deploying more complex feature extraction/selection techniques.

In summary, our study demonstrates the potential of machine learning methods in medical image analysis. The insights gained provide a basis for further refining our models and techniques, ultimately contributing to advancements in this crucial field.

5.2 Limitations and reflections

We faced a lot of limitations in this project. Mainly our lack of skills in the machine learning domain, the small dataset containing bad-quality images, which also led to inaccurate automatic image segmentation, and limited time and resources. Considering the ethics, for our dataset it is mentioned on the website that ”the program is managed by the Department of Specialized Medicine and was approved by the university ethics committee.” (see Appendix A).

6 Conclusion

6.1 Reflection on Research Objectives

The goal of this project was to predict a diagnosis of skin cancer in an automated process and examine the impact of using higher-quality images. Overall, the main objectives of this project were met. In terms of using better-quality images, we found that training the models with higher-quality images might not necessarily improve the performance. However, further research on this subject is necessary for any definitive conclusions.

6.2 Implications and Future Research

The statistics from training data potentially suggest that image quality has an impact on the accuracy of the automated detection of cancerous skin lesions. Because of the limitations of this study, our conclusion cannot be definite. Therefore, potential future research should focus on creating models that can be generalized and investigate if there is a discrepancy when testing the model with different quality images.

References

[Addison et al.,] Addison, D., Wermter, S., and Arevian, G. A comparison of feature extraction and selection techniques.

[Argenziano et al., 2011] Argenziano, G., Catricalà, C., Ardigo, M., Buccini, P., De Simone, P., Eibenschutz, L., Ferrari, A., Mariani, G., Silipo, V., Sperduti, I., and Zalaudek, I. (2011). Seven-point checklist of dermoscopy revisited. *British Journal of Dermatology*, 164:785–790.

[Argenziano et al., 1998] Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., and Delfino, M. (1998). Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of Dermatology*, 134:1563–1570.

[Carrera et al., 2016] Carrera, C., Marchetti, M. A., Dusza, S. W., Argenziano, G., Braun, R. P., Halpern, A. C., Jaimes, N., Kittler, H. J., Malvehy, J., Menzies, S. W., Pellacani, G., Puig, S., Rabinovitz, H. S., Scope, A., Soyer, H. P., Stolz, W., Hofmann-Wellenhof, R., Zalaudek, I., and Marghoob, A. A. (2016). Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma. *JAMA Dermatology*, 152:798.

[Foundation, 2020] Foundation, S. C. (2020). Skin cancer facts statistics - the skin cancer foundation.

[Golston et al., 1992] Golston, J. E., Stoecker, W. V., Moss, R. H., and Dhillon, S. (1992). Automatic detection of irregular borders in melanoma and other skin tumors. *Computerized Medical Imaging and Graphics*, 16:199–203.

[Kassem et al., 2021] Kassem, M. A., Hosny, K. M., Damaševičius, R., and Eltoukhy, M. M. (2021). Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics*, 11:1390.

[Madooei et al., 2019] Madooei, A., Drew, M. S., and Hajimirsadeghi, H. (2019). Learning to detect blue–white structures in dermoscopy images with weak supervision. *IEEE Journal of Biomedical and Health Informatics*, 23:779–786.

[Piatek and Mroczek, 2022] Piatek, and Mroczek, T. (2022). Analysis and classification of melanocytic skin lesion images. 207:1911–1918.

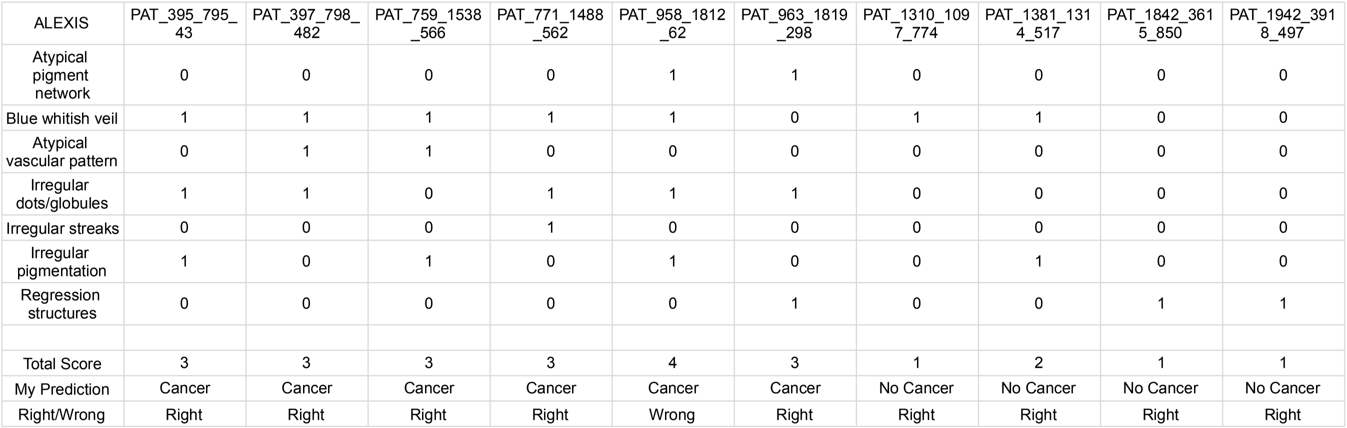
[Scikit-Image,] Scikit-Image. Blob detection — skimage v0.20.0 docs.

7 Appendix

A Appendix 1

"Ethics statement: The dataset was collected along with the Dermatological and Surgical Assistance Program (PAD) of the Federal University of Espírito Santo. The program is managed by the Department of Specialized Medicine and was approved by the university ethics committee (nº 500002/478) and the Brazilian government through Plataforma Brasil (nº 4.007.097), the Brazilian agency responsible for research involving human beings. In addition, all data is collected under patient consent and the patient's privacy is completely preserved." - https://data.mendeley.com/datasets/zr7vgbcyr2/1

B Example: CSV file with manual extraction



ALEXIS	PAT_395_795_43	PAT_397_798_482	PAT_759_1538_566	PAT_771_1488_562	PAT_958_1812_62	PAT_963_1819_298	PAT_1310_1097_774	PAT_1381_1314_517	PAT_1842_3615_850	PAT_1942_3918_497
Atypical pigment network	0	0	0	0	1	1	0	0	0	0
Blue whitish veil	1	1	1	1	1	0	1	1	0	0
Atypical vascular pattern	0	1	1	0	0	0	0	0	0	0
Irregular dots/globules	1	1	0	1	1	1	0	0	0	0
Irregular streaks	0	0	0	1	0	0	0	0	0	0
Irregular pigmentation	1	0	1	0	1	0	0	1	0	0
Regression structures	0	0	0	0	0	1	0	0	1	1
Total Score	3	3	3	3	4	3	1	2	1	1
My Prediction	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	No Cancer	No Cancer	No Cancer	No Cancer
Right/Wrong	Right	Right	Right	Right	Wrong	Right	Right	Right	Right	Right