IT-UNIVERSITETET I KBH

3rd Project Report - Group 13

Josephine Holst
jtih@itu.dk

João Vieira da Silva
joap@itu.dk

Ludek Cizinsky
luci@itu.dk

Disclosure

The group was reduced from 4 to 3 people due to one member dropping out (mnib@itu.dk). Besides this, all group members contributed equally to the project. Due to the real-time collaborative platform we used, the attached git log will not reflect this correctly.

Introduction

Skin cancer is one of the most common types of cancer worldwide. Even though the number of deaths related to Melanoma are more than 60,000 per year[1], early detection of the disease can lead to a 99% 5-year survival rate[2]. Since detecting Melanoma at an early stage is crucial, computer aided diagnostic systems are constantly improved to aid dermatologists assessing and detecting malignant lesions. In recent years, various mobile apps have been developed to provide suggested on-the-go diagnoses based on images of skin lesions. As these technologies have the potential to improve survival rates and assist healthcare procedures, this report seeks to complement the current understanding of how reliably skin lesions can be evaluated using image classification tools. This approach takes three common characteristics into account for describing features of early Melanoma, namely asymmetry, border irregularity and color variation. By designing and measuring these "handcrafted" features in images of both malignant and benign skin lesions, three simple classification models were trained on 4 data-sets with different distribution of given cancer type. These results were then evaluated, answering the two-folded question: How well can measures of asymmetry, border irregularity and color variation reliably detect Melanoma and Keratosis? In the following pages the results of this exploratory approach will be presented and discussed along with the main limitations of this study. This includes a description of methods for image data collection, filtering and feature extraction.

Data

Data Collection and Transformation

The results of this report builds on data from the ISIC 2017 challenge. The complete data set consists of 2750 unique benign and malignant records with no missing values. Each record is associated with a corresponding image and segmentation mask, either labeled as Melanoma, Keratosis or neither (healthy). In addition, three types of metadata were included using the ISIC API. This includes internal API id of the record and image width and height. The record id was used to query image data from the API, making the reproducibility of results more convenient. As images varied in resolution, sizes exceeding 2100 x 3200 pixels were filtered out to improve performance, due to high computational complexity for larger images. Finally, images were checked to only include lesions located more than 50 pixels from the border of the image, this was done to minimize the risk of errors during rotation when measuring asymmetry as well as to exclude invalid images.

Data for Prediction: Variations on Balance

Given the domain of skin cancer detection, it is reasonable to assume that the ratio between malignant and benign images will be imbalanced in real life applications. This also becomes apparent in the ISIC data set. Therefore, four subsets of data were randomly sampled (without replacement) from the ISIC data set, each of size 400. These reflect two different balances, namely an imbalanced data set (20%

[1] https://melanomapatients.org.au/wp-content/uploads/2020/04/2020-campaign-report-GC-version-MPA$_1$.*pdf*

[2] https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html

malignant vs. 80% benign) and a more balanced data set (40% malignant vs. 60% benign) for both Melanoma and Keratosis (see Table 1). This was done to explore how imbalanced data might affect the models' predictive performance, specifically for the minority class (malignant). Features were then extracted for all four subsets and saved to a .csv file for analysis. It is important to note that this preprocessing pipeline is not extensive. This will be further discussed in limitations.

Name	Final size	Melanoma (%)	Keratosis (%)	Healthy (%)
model_input_mel.csv	399	40	0	60
model_input_mel_imb.csv	399	20	0	80
model_input_ker.csv	398	0	40	60
model_input_ker_imb.csv	399	0	20	80

Table 1: Balance of unique data sets with extracted features (filtered for bordering lesions)

Results and Discussion

Measuring Features

Asymmetry

The binary segmentation masks were used to estimate the level of asymmetry of each lesion. By 'folding' the images horizontally, non-overlapping parts were assumed to represent asymmetric areas. These areas were then quantified by their amount of pixels and standardized relative to the area of the lesion (in pixels). To ensure the best fitting rotation for each 'fold', masks were rotated in 30 degree intervals from 30 - 180 degrees. Finally, the smallest area was selected as the metric (see Figure 1).

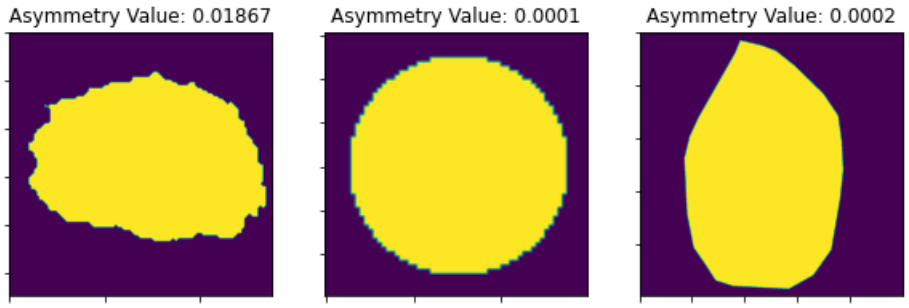

Figure 1: Example of Asymmetry measurements

Border Irregularity

Using morphology, dilation was used to extract the boundary pixels for each lesion by subtracting the dilated pixels from the original mask. To minimize the effect of the boundary's curved shape the border was divided into 50 equally sized segments. For each segment, the euclidean distance was calculated from each border pixel to the center point. These distances were then transformed into z-scores. Finally, the percentage of points that were above or below the mean by 1.9 standard deviations was calculated. The goal of this approach was to focus only on segments that showed a significant variation and take varying lesion sizes into account. Finally, the standard deviation of the

percentages for each segment was returned (see Figure 2). A second approach to quantify border irregularity was implemented by measuring the compactness of each lesion. This was assumed to be suitable for recognizing how much the border would deviate from a perfect circle.

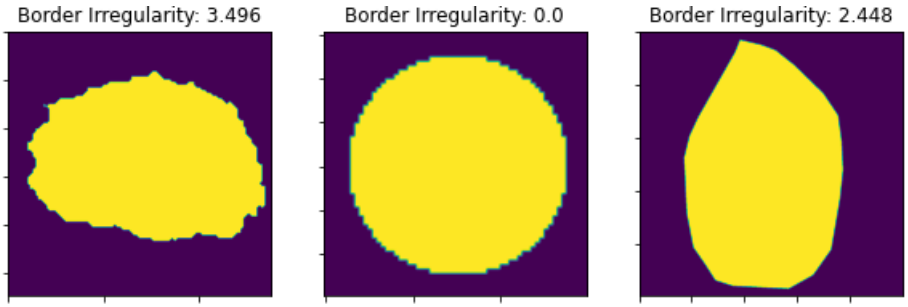

Figure 2: Example of border irregularity measurement

Color Variation

To estimate the range of colors within each lesion, color images were cropped using the segmentation masks and divided into approximately 250 regions (super-pixels). This was done using a SLIC segmentation tool which uses k-means clustering to segment the image into similar areas. All black regions were excluded as these would represent areas outside the lesion. Variation 'levels' were then measured by converting the average regional colors from RGB to HSV and accessing hue, saturation and value (brightness) channels for each region. This approach aimed to minimize the effect of potential color noise as well as to make the variation of the average regional RGB colors more comparable. The level of color variation was then measured by taking the standard deviation of each channel across all regions. Finally, the IQR of brightness was added to test for a different measure of spread, resulting in a total of four different color features.

Normalization, Data Splitting and Feature Selection

A total of seven features were extracted for classification. To ensure that each feature would carry equal weight, all values were normalised before using them as an input for classification. Each of the four data sets was split into training (80%) and test sets (20%). Given the smaller size of each data set, dimension reduction was implemented to improve the models' performance. This was done using a filter approach, automatically selecting the two best scoring features for each data set. As data consisted of numeric features and binary classes, scores were accessed from a selector object using mutual information to perform a uni-variate evaluation.

Classification Models

Before choosing a classifier a visual inspection of the class distributions was made by looking at grouping patterns when plotting features against each other. These varied depending on the diagnosis and the data balance, but clusters seemed to suggest a general tendency of class overlaps. In certain scenarios (e.g. Keratosis classification) it seemed more suitable to use a linear model such as Support Vector Classification (SVC) as classes seemed to form a linear relationship. Other combinations showed non-

linear relationships (e.g. Melanoma classification), which suggested the use of a K-Nearest Neighbors (KNN) or Decision Tree classifier. To fine tune the hyper-parameters of the selected classifiers, a cross validation method was implemented with 5 folds. Recall was chosen over accuracy and precision as the main assessment metric. This was due to the real life context where it is usually more preferable to have healthy lesions labeled as cancerous than leaving a cancerous lesion undetected. The final test results can be viewed in Table 2 below.

Data set	Classifier	Accuracy (%)	Precision (%)	Recall (%)
model_input_mel.csv	KNN-5	42.5	29.4	31.25
model_input_mel_imb.csv	KNN-5	71.25	0	0
model_input_ker.csv	KNN-5	66.25	56.7	65.6
model_input_ker_imb.csv	KNN-5	85	75	37.5

Table 2: Test results for best performing classifiers

Results for Melanoma Classification

Looking at the primary metric, recall, results show that the trained classifier would miss around 70 % of Melanoma images for a balanced data set. From an end user perspective, this would be unacceptable. One of the possible reasons for such a poor performance could be the implementation of border irregularity measurement which according to previous studies should be one of the key identifiers. However, in our case it performs poorly (not in top 2 features). By reducing the number of Melanoma samples, the classifier then achieves the worst possible performance based on the recall metric (0%). There are two possible explanations. Firstly, due to the small number of Melanoma samples within training set (64 samples), the classifier was not able to learn properly what distinguishes Melanoma apart from the healthy lesion. Secondly, the implemented general features do not describe well the difference between healthy and Melanoma lesions.

Results for Keratosis Classification

In contrast to Melanoma, the trained classifier for Keratosis would miss only about 35% of Keratosis images when built on a balanced data set. On the other hand, out of all images that were labeled by the classifier as Keratosis, only 56.7% of them were actually Keratosis. Finally, given the best performance being of the KNN-5 classifier on training data (and, in contrast, a considerably worse performance on test data) it means that the model was overfitted. For the imbalanced data set, the classifier's ability to be precise increased compared to the balanced data set, despite the small number of samples. Yet, the recall remained considerably low. Similarly to the results from Melanoma, this could be explained by the limited number of positive samples as well as inadequate feature implementations.

Limitations

Data: Four unique data sets were created for analysis, resulting in smaller sets of input records for each model. Model prediction could thereby be strengthened by adding more data or using re-sampling techniques.

Feature Extraction: As the results of report rely heavily on how the features were extracted limitations in these procedures are important to take into account. Common to all measurements are the use of provided segmentation masks. As smoothness, precision and border distinctness vary across masks, this includes an underlying uncertainty in the exactness of the features values. Furthermore, since features were selected based on prior knowledge on malignant lesions this could lead to relevant information getting lost as representation removes information.

Image filtering: From a visual inspection, it became clear that lesions were photographed under varying conditions and standards influencing both lighting and resolution. Furthermore, several images had noise in the form of hairs and/or air bubbles. Since this analysis does not extended to filtering nor neutralizing these potential variations, it cannot be excluded that these may influence the measurement of color variations.

Concluding Remarks and Future Work

This study had one main objective - to answer how well measures of asymmetry, border irregularity and color variation can reliably detect Melanoma and Keratosis. In order to answer these questions, feature extraction methods were first implemented based on the available literature. To expand the analysis, two data sets for each cancer type were created. One with a more balanced representation of healthy and malignant samples in order to reduce complexity of training models and look at the problem from a theoretical level. The second data set was designed to be unbalanced in order to simulate real life circumstances. Test results showed that the trained classifier was persistently inadequate for detecting both Melanoma and Keratosis based on the top two scoring features - even for the more balanced data sets. This demonstrates that it is very challenging to implement general feature(s) which would adequately distinguish between malignant and healthy skin lesion without being computationally too heavy. On the other hand, it is important to consider the many limitations of this report: the size of data sets, non-standardized images and a limited image processing pipeline. To conclude, despite the poor classification results, this research is believed to provide value by at least demonstrating the many challenges that arise when working with image classification problems of a medical nature.
Future research could evidently focus on improving the feature measurements as well as developing more extensive image filtering procedures while maintaining computational efficiency.