Projects in data science

Project 1: Summary

By Elias Fischer Hegelund, Mikkel Emil Jensen and Kristoffer Schmidt (Nighthawks)

(https://commons.wikimedia.org/wiki/File:Common_Nighthawk_%2814605341423%29.jpg)

Table of contents

Introduction: Briefly introduce the dataset and its purpose.

For this project we are working with the dataset PAD-UFES-20 to analyze and create masks for given photos containing skin lesions. The dataset contains thousands of different pictures of skin lesions, with the with and other data on the lesion being available in the metadata. In this project we will work with a specific small portion of the dataset called “n”. In the dataset there are seven different types of lession, with 3 of them being cancers. The point of this part of the project (project 1) is to mark the lesions with the labelstudio-software so that we at a later point can train a model to recognize them, and use the ABCDE-rules to make a qualified estimation on whether or not a lesion is cancerous. By focusing on the ABCDE-rules, using the metadata information will not be strictly necessary, though it could most definitely prove beneficial in the future, if we were to make a more precise model that could differentiate more in its diagnostics.

Dataset Overview: Describe the structure of the dataset, including the types of data available (images, metadata) and any missing or low-quality data.

The dataset that this project revolves around (PAD-UFES-20), contains 2,298 samples which are of six types of skin lesions wherein 3 are types of cancer and the rest are other skin diseases. Each of these samples contains a clinical image and contains up to 22 clinical features, which include: Age, location of skin lesion, diameter of lesion , etc. Still, it is worth noting that many of the patients have a lot of missing data with some data being more important than other data such as fitzpatrick or diameter, since skin lesions with a diameter of above 6mm require more awareness because it can be a potential factor for skin cancer (link). There are also other important factors to look for when studying skin lesions, for example the ABCDE, where D is the diameter. Though for our dataset we do not have information from the metadata on either A: Asymmetry, B: Border, C: Color (multiple colors) or E: Evolving (link).

ABCDE-Rules

The ABCDE-Rules are the most common way of identifying whether or not a lesion is cancerous or dangerous. The ABCDE-Rules are not a definitive ruling, but if the rules point at a lesion being bad, it should be of utmost importance to make a doctor's appointment as soon as possible, so that a proper diagnosis can be made and a treatment can be given. Since we do not train or data with the metadata diagnosis, the ABCDE-rules will be our only or at least our primary way of making the automatic evaluation of the lesions. The rules are as follows:
A: Asymmetry,
B: Border
C: Color (multiple colors)
D: Diameter (link).
E: Evolving (link).
But our intention is that we can analyze at least A and B and possibly C.

Quality of photos:

The photos are of different quality and of different zooms, since the photos aren't unanimously taken by the same type of photographic mechanism, but by different smartphones. This results in significantly differing quality of the photos provided, with some being of low quality, which might make the process of creating the masks of the photos vary in precision.

What types of diagnoses are there, and how do they relate to each other?

For this dataset, we are looking at seven types of skin lesions, The different types of skin lesions in the dataset can be seen in the table below:

Name	**Abbreviation**	**Is cancer**
Basal Cell Carcinoma	(BCC)	True
Squamous Cell Carcinoma	(SCC)	True
Melanoma	(MEL)	True
Actinic Keratosis	(ACK)	False
Seborrheic Keratosis	(SEK)	False
Bowen’s disease	(BOD)	False
Nevus	(NEV)	False

As the Bowen’s disease is considered SCC in situ, we clustered them together, which results in six skin lesions in the dataset, three skin cancers (BCC, MEL, and SCC) and three skin disease (ACK, NEV, and SEK)

Conclusion:

In summary, analyzing the PAD-UFES-20 dataset has revealed a wide range of skin diseases, from minor to potentially serious types of skin cancers. When looking at our data, hereby the photos, we see a lot of variance in the quality of these photos, being image quality and/or clear definition of the skin lesion area. Understanding these images and diagnoses better is essential for future work and documentation of this project. By

acknowledging the dataset's strengths and limitations, especially regarding photo quality, we're better equipped to contribute to dermatological research and diagnosis.