Skin Cancer MNIST: HAM10000 disease Classification (Extended Project)

Author: Antreas Kasiotis Student Number: B8035526

Project overview

This project aimed in the development of a software solution that automated diagnosis for disease classification from a dataset of dermoscopic skin cancer images. The approach that was taken was to develop various machine learning and deep learning models that would tackle this task and comprehensively compare their performance.

Dermatology and Skin Cancer

Skin cancer is the most common type of cancer and it strikes one in five people by the age of 70 (SCF, 2022). With that being said, it is clear that this disease is a very big problem for our society. In the United Kingdom there is no national screening programme for non-melanoma skin cancer. Although non-melanoma is usually benign, it still has a 1% mortality rate (Cancer Research UK, 2022). According to the Skin Cancer Foundation, 99% of all skin cancer cases are curable if they are diagnosed and treated early enough (SCF, 2022). It is therefore very clear that the role of technology in this field in tremendously important. With the use of image classification machine learning and deep learning the world could benefit by being enabled to easily, quicly, and accurately detect skin cancer so that they can the act on it.

Work Plan

To carry out this project I will be implementing three image classifiers that will be able to differentiate a new, unlabeled image of skin cancer into one of the seven classes of skin cancer. To achieve the classification of the disease, the models used labelled data to learn the relationship between the input data (images) and the labels. The classifiers that I will implement are the cnn (convolutional neural network), the lstm (Long short-term memory RNN), and the svm (Support vector machine). For each type of classifier I will create two implementation, one of greyscale images and one for coloured ones. This will be done so that we can also observe the difference between the classifiers according to the data that they use.

EDA & Implementation

To give an overview of the work that was done I will be briefly explaining the process of developing and fitting the models, for more details please look at the "work-report" document.

To begin my investigation I started by carrying out some exploratory data analysis on the metadata of the images. From the metadata I was able to extract a plethora of information

about the patients and the disease. Some of my findings included frequencies of disease classes, age groups, frequencies of disease localizations on the body and gender frequencies.

The next stage of my work was to implement the classifiers and fit them to the two categories of data. I started working with the CNN classifier for which performed some pre-processing prior to building the model such as, sepating my data into predictor and response variable. I also oversample these variables to overcome the class imbalance I found in the EDA. Furthermore, I reshape and normalize my images and I also encode my labels to one-hot vectors so that they can be fitted to the model. For the LSTM I did the exact same process for preparing my data with some minor alterations. One of which was the reshaping of the image which had to be done differently so that it could be fitted into an LSTM. The SVM was slightly different. For this method still separated and oversampled my data and also reformatted and reshaped my data in a similar fashion. however, for the SVM I used a smaller sample from my oversampled data because my model was taking too long to train. Additionally, I also used LDA to reduce the dimentionality of my data and thus further increase training speed.

Results

All three classifiers were succesfully developed and fitted with both categories of image data therefore, in total there were six parts to this investigation. Each classification method was evaluated and metrics were derived about their accuracy, their loss of the data, their runtimes and their missclassification of particular classes. These results are presented as both reports and visualizations such as confusion matrices and more. The only issue in the evaluation of these methods came with the SVM model which had to be fitted with a different number of training data due to the amount of time it required to train the model. However, that was taken into consideration in the following comparison of the classifiers. The findings of the investigation suggest that the accuracy of a classifier is highly correlated with the image color spectrum. Additionally, it was interesting to see that these classifiers sometimes performed differently even while in the same environments. For example, when these models were trained with RGB images, they all had very different runtimes. Whereas in some other cases these classifiers had great similarities, such as they all had higher accuracy and lower loss over the dataset when trained with RGB images.

Future Work

This project primarily aimed to investigate two things, the differences of classifiers in similar environment and also the effect of the color spectrum on the performance of these classifiers. Therefore, the research that was conducted could be expanded in two way as well. Since we have decided to only invetigate the effect of the color spectrum on a classifier, the first way to further extent this work would be to also investigate whether the resolution of the image plays a role in the effectiveness of any classifier. Secondly, the task of image classification with the use of different classification methods was investigate. An extension of this work could be conducted in further expanding the range of classifiers used and compared.

Personal Reflection

This project was my first ever project in machine learning and deep learning and it has definately expanded my horizons on the field. In the begining i did not know much about image classification or even about handling image data programmatically but I have now aqcuired a plethora of knowledge that without a doubt be very usefull when I get to work with similar projects in the future. Python was also a programming language that I had never used prior to this project but seeing as I am of a computer science background I was able to quicly familiarize myself with its syntax and logic. However, the theorical concepts behind the machine learning classifiers were really hard to grasp in the beginning. I took me many tries and failures until I got things right. Endless hours of setting the models and training them just to then go back and make another minor change, again and again. Needless to say the process was educational but exhausting as it should. With the knowledge that I have now about the field I would change one thing about the process that I took in future cases where I will have to work on a similar projects, the goal setting for my investigation. The way that I would redo it is by setting more clear goals from the begining about what it is that I want to investigate. I believe this would greatly help me reduce the complexity of my work but also allow me to dedicate my focus on the key questions.