BREAST CANCER DETECTION USING DIFFERENT CLASSIFIER MODELS

Botchway, Kodjo Opoku

A20338464

INTRODUCTION

Breast cancer is a disease that stems from the uncontrollable increase in the cells that are situated in the breast region. This cancer occurs mostly in women and very rarely in men account for a current large percentage of cancers that affect people in the world today. It has also been recorded as the most common form of cancer in women. The early detection of breast cancer is a really important step forward in the treatment process to enable doctors and physicians save time, reduce severity and in drastic cases prevent fatal endings. The plan for approaching the problem is to categorize and correctly predict the diagnosis of breast cancer data, they, being malignant or benign. This accuracy of our model will help with detection of the cancer in patients with the available recorded data. This is the main problem being investigated.

PROBLEM STATEMENT

The main idea of this project is to use the data from the digitized images of a breast mass to detect breast cancer. The dataset that will be employed for the project is obtained from http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. To be able to effectively accomplish this, we are going to use different machine learning models and weigh and compare the accuracy of the resulting models. Using the features provided in the dataset, the plan is to use 3 traditional ML classifier models, as well as a multi-layer feedforward neural network model and a simple CNN model. The reason for testing these different models is for obtaining the model with the highest classification accuracy. We plan to use different hyperparameters to optimize the models and ensure that the complexity and the efficiency of the model are not compromised.

There has been some work done on the detection based on the actual images from patients. This is going to be the baseline to understand whether the results are in the positive direction. This is not particularly necessary for this project, but we wanted some domain knowledge to be able to weigh our results against the real-life instances. The results are going to be evaluated both qualitatively and quantitatively for context. Plots will also be made between the predicted and actual values among others. We will use different binary evaluation measure for calculating our model performance, including R-square value, misclassification score, precision and the F1 score.

TIMELINE

Week 1 – Understanding the problem internally, putting together the dataset, cleaning and pre-processing the data.

Week 2 – Preparing framework of the models and the discussion and initialization and the starting stages of building.

Week 3 – Model building and optimization factoring in hyper-parameter tuning and optimization. Model finalization and results

Week 4 – Finalization of project and completing final report.