BREAST CANCER DETECTION PROBLEM FORMULATION

Abdelrahman Karawia
Department of Computer Science
Mansoura University, FCIS
Mansoura

Research Report

March 25, 2022

Abstract: This project aims to detect or to be more specific to classify the given data (Radiology or Numerical Medical Results for a breast tumor) which will be given as input to the program is benign or malignant tumor. The problem of tumor classification for its malignancy level has been studied deeply and has a dozen of applications here and there for everyone who's trying to improve a model which has a very reasonable level of predictability. There are enormous number of algorithms to be used in machine learning for such a classification problem and this would a great challenge to choose between, not only this we have also to figure out which parameters for every algorithm we chose (Tuning Process). We will move through two process or steps, the first one is to use different classical machine learning algorithms and observe the differences between them and try to choose the relatively best one among them, then we will turn into neural network model and try to figure out the best number of hidden layers and activation functions to use in this problem. All of this would be in a program with a quite good GUI and the user will have the capability to input either Radiology or Numerical Medical Results and get the results immediately.

Keywords: Machine Learning, Algorithms, Image Classification, Binary Classification , Tuning, Performance, Neural Network, ANN, TensorFlow, KNN, GUI, Python.

1. RESEARCH PROBLEM

1.1 Problem Area

Binary Classification is the main method or field in the project for a medical existing problem which is just a choice between two options for the pathologists to make, but of course it is not that easy for them and there is no 100 percent certainty in their choice. Breast Cancer is considered as the most common type of cancer between females and the first cause of cancer death among them In Egypt, it constitutes 33% of female cancer cases and more than 22,000 new cases diagnosed each year. This is expected to rise exponentially over the next years given the enlarging population [1]. Early Detection and Diagnosis, it's not just a combination of two words it's the key for treatment and the only way to be far away from the danger zone [2].

Our goal is to develop a binary classifier using Machine Learning and Deep Learning technologies and make a quick response or initial prediction with only the radiology image for the patient or the Numerical Medical Results. This is not just a classifier which will be cliché if there no other feature in the classifier, therefore, accuracy is going to be another challenge for us as we must be sensitive in these fields particularly.

1.2 Practical Problems

There's a quite good number of researchers which talk about binary classification and breast cancer classification particularly, and they separated into two different groups one of them decided to use just classical ML algorithms and compare each other by looking at accuracy as a primary comparator, the others headed to use genuinely the Deep Learning technology with that huge number of data available from different sources.

We genuinely decided to make a great combination here by using the classical algorithms and NNs as well and try to grab the highest accuracy from both, hoping to crack the accuracy has been made from other researchers which up to 90 percent. Over-fitting is a problem faced most of the researchers tried to make this classifier especially in deep learning method, so we are going to focus as well in a way like **regularization** to overcome this problem to make a high-test accuracy.

Most of work has been done is just a bunch of codes have to be run to get the results which might not be understandable for normal users which a problem, and here we are trying to make a quite good **Graphical User Interface** for the project with interactive buttons and widgets to ease the process for the user and show the results in an understandable way and showing the limitations of dependencies.

1.3 Aims and Objectives

Framing the problem by addressing the aims and objective beneath the project, this research or project aims to investigate large-scale dataset of a breast cancer tumor on Kaggle or any other sources and it will be separated into two different projects one for just inputting a radiology and the other one for whose have numerical results for the tumor and make high-test accuracy as much as possible regarding the limitations which may face the learning process from images lacking or positions, … etc..

Data analysis will be done as an initial step for the whole process to investigate the data for better understanding and grabbing some conclusions with different statistical methods available. No surveying for collecting the data are going to happen just grabbing dataset from the internet as it's available with numerous amounts.

Implementing Different machine learning algorithms on the same dataset mentioned above like:

- Logistic Regression
- K- Nearest Neighborhood (KNN)
- Random Forest Model
- Support Vector Machine (SVM)

Which they are the almost the most important algorithms used in classification problems but here we must check the four algorithms using different evaluations methods known and choose the best one for our program to be the one.

GUI will be the end step in the project by creating a pretty application for users to deal with using an outstanding library to do so, which is **Tkinter** Tk/Tcl has long been an integral part of Python. It provides a robust and platform independent windowing toolkit, that is available to Python programmers using the Tkinter package [3].

Works Cited

[1] A. H. Abdelaziz, "Breast Cancer Awareness among Egyptian Women," *Clinical Oncology Department, Faculty of Medicine, Ain Shams University,* vol. 17, no. 1-8, 2021.

[2] J. M. A. G. S. L. A. N. a. A. J. C. E. DeSantis, "Breast cancer statistics, 2017, racial disparity in mortality by state," *Cancer Journal for Clinicians,* vol. 67, no. 6, 2017.

[3] Python, "Python Docs," [Online]. Available: https://docs.python.org/3/library/tk.html.

[4] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, Sebastopol: O'Reilly Media, Inc, 2019.