BREAST CANCER PREDICTION USING MACHINE LEARNING

BHUVANA KORRAPATI
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
bkorrapati@gmail.com

***Abstract*—Breast cancer is the leading cause of mortality from cancer for women between the ages of 20 and 59 and the second leading cause for women over the age of 59. To stop the disease's progression and lower its morbidity rates, it is crucial to diagnose and treat this pathology as soon as possible. Several factors are considered for the accurate prediction of the cancer such as radius, texture, compactness, concavity. So this paper presents the detail description of the features and prediction of the stroke using two different algorithms that are Support Vector Machine(SVM) and Convolution neural networks(CNN).**

I. Introduction

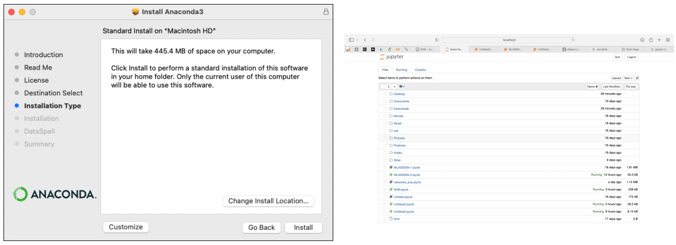
The most prevalent illness in women is diagnosed as breast cancer. Every year more than 2.23 million women are diagnosed with the breast cancer. In this project we analyse the sample data that is collected from UCI macine learning dataset repository[1]. The machine learning algorithms that is Support Vector Machine(SVM) and Convolution neural network(CNN) are used to analyse and predict the dataset. The predicted values should be highly accurate to diagnose the disease, since the comparison between the algorithms are made. CNN provides the higher accuracy comparitively for the given dataset.

II. Setting up Environment

Mostly python or R language is used for the machine learning programming. Anaconda environment was downloaded from the official Anaconda website[2] and for proper installation followed the steps mentioned in the website[3]. Python provides access to large number of libraries and frameworks. Anaconda is a free open-source environment which consists of over 250 automatically installed packages. Executing tensorflow in the mac M1 chip has some difficulty, so i used online compiler that is google collab. For the programming, jupyter notebook is installed with all the necessary packages and used for the code description, prepossessing and algorithm application to the given dataset. The libraries to run scikit-learn and TensorFlow is installed in the jupyter notebook.

III. Dataset Description

The dataset that is considered for the prediction of the cancer consists of the features such as diagnosis, radius, texture, perimeter, area, smoothness, compactness, concavity. These are the dimensional measure of the tumor. All the mean values, worst values are considered. Here the compactness represents the volume of the tumor, concavity represents the concave positions. The labels that is 'M' represents the malignant tumor and 'B' represents the benign tumor. Malignant is the process in which the cancer cells are multiplied and are harmful, finally leads to the cancer. Benign is a standard tumor where this is not harmful and cannot spread the cancer cells. We can have the count of the features using countplot.



Standard Install on "Macintosh HD"

This will take 445.4 MB of space on your computer.

Click Install to perform a standard installation of this software in your home folder. Only the current user of this computer will be able to use this software.

Select items to perform actions on them.

Name		Last Modified	File size
Desktop		29 minutes ago	
Documents		19 days ago	
Downloads		29 minutes ago	
Movies		18 days ago	
Music		18 days ago	
opt		18 days ago	
Pictures		19 days ago	
Postman		16 days ago	
Public		19 days ago	
Sites		9 days ago	
MLASSIGN-1.ipynb		18 days ago	1.81 MB
MLASSIGN-2.ipynb	Running	12 hours ago	50.3 kB
networks_proj.ipynb		a day ago	1.14 MB
SVM.ipynb	Running	3 hours ago	239 kB
Untitled.ipynb		18 days ago	172 kB
Untitled2.ipynb	Running	5 hours ago	26.5 kB
Untitled3.ipynb	Running	8 hours ago	8.14 kB
html		17 days ago	0 B

Fig. 1: Left: anaconda for mac. Right:jupyter notebook.

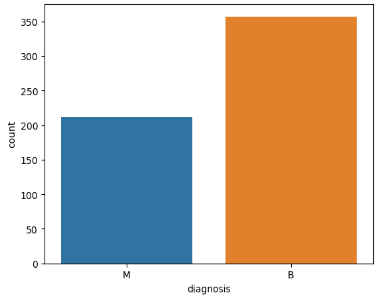


Fig. 2: diagnosis count

IV. Data Preprocessing

The data can be considered as three parts that is number of observations, features and the labels. The number of observations are considered to be the total instances. The "features" that make up a data set used to describe or communicate the data set itself. This may depend on factors including size, location, age, time, color, etc. Features are sometimes referred to as attributes, variables, fields, and characteristics. They appear as columns in datasets. The label is to differentiate the classes and is only observed in supervised machine learning models. The data can be imbalanced, inaccurate or incorrect, by some simple statistical approaches we can clean the noise data. the processing of the data includes identifying and dropping unnecessary columns or features, detect null values and fill it with the mean or median values as this can cause bias for data processing.we can even describe the data and check the values for each feature and find the correlation as well.

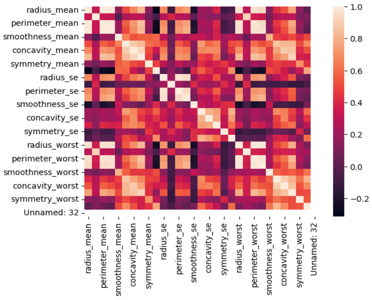


Fig. 3: correlation heatmap

V. Proposed Methodology

Machine Learning algorithms are used for the classification and prediction of the data, The dataset is divided by 80:20 rule that is 80 percent of training data and 20 percent of testing data. It is a proven rule to split the dataset for better accuracy as well and to reduce the problem of non seen data. The training and testing datasets are saved as a csv files. Here the training data is quantitative measure and validation and testing comes under the qualitative measures.

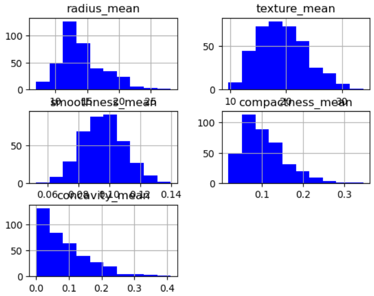


Fig. 4: Training data graphs

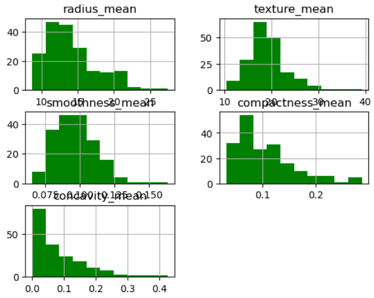


Fig. 5: Testing data graphs

A. SVM CLASSIFICATION

The support vector machine is a supervised machine learning algorithm. This is the mathematical model where the input data is not scalable. The data cannot get added after the data is trained. We use C-Support Vector Classification to classify this dataset. The steps includes importing the library and calling the function to perform the svm model.Then the data is sepertaed to training and testing data, and apply the svm model. We can check the model with the qualitative measures from the confusion matrix that is generated from the data.

Table.1.

Our Accuracy Score	0.925531914893617
Our Precision Score	0.9090909090909091
Our Sensitivity Score	0.9836065573770493
Our Specificity Score	0.8181818181818181

TABLE I: SVM Measures

B. Convolution Neural Networks

The CNN is a deep learning algorithm. This is the hierarchical model where the input data is scalable. The data cannot get added after the data is trained. We use C-Support Vector Classification to classify this dataset. The steps includes importing the library and calling the function to perform the CNN model.An epoch means training the neural network with all the training data for one cycle, here we considered 50 observations to train at once.An epoch is made up of one or more batches, where we use a part of the dataset to train the neural network.
we used a sequential tensorflow keras model, and we normalized using batch normalization, it applies a transformation that maintains the mean output close to 0 and the output standard deviation close to 1. The Dropout layer acts as a mask, erasing some neurons' contributions to the subsequent layer while leaving all others unaltered. For training the model in CNN we we give the set of data of each epoch and find the accuracy and loss. this determine the proper training and testing the data. Adam optimiser is used for the model optimisation. An algorithm for gradient descent optimization is called adaptive moment estimation. When dealing with complex problems involving a lot of data or factors, the strategy is incredibly effective. It is effective and uses little memory.

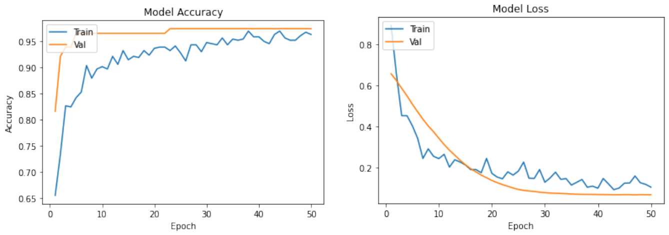


Fig. 6: Left: accuracy graph. Right:loss graph.

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, It is a summary of prediction results on a classification problem. We can get all the four values to calculate the qualitative measures from the algorithms, that are true positive, false positive, true negative, false negative. All the measures that are accuracy, precision, sensitivity and specificity can be calculated from the confusion matrix.

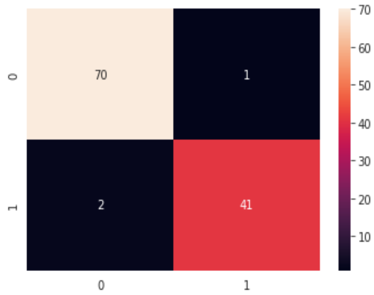


Fig. 7: confusion matrix

Table.2.

Our Accuracy Score	0.9736842105263158
Our Precision Score	0.9761904761904762
Our Sensitivity Score	0.9534883720930232
Our Specificity Score	0.9859154929577465

TABLE II: CNN Measures

VI. Conclusion

According to the performance evaluation of both the algorithms, both CNN and SVM classifiers achieve comparable results. CNN is even scalable with different epochs and batch range. We can add the new data when ever available as it is one observation training. Comparatively CNN has high accuracy and precision values and is recommended in this case.

VII.

References

[1] http://www.springer.com/us/book/9781489976406

[2] http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

[3] https://docs.anaconda.com/anaconda/

[4] https://www.anaconda.com/products/individual

[5] https://scikit-learn.org/stable/install.html

[6] https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html?highlight=svm

[7] https://en.wikipedia.org/wiki/Convolutionalneuralnetwork