INTRODUCTION

Breast cancer is the second most prevalent form of cancer worldwide. In 2018 alone, there were more than 2 million new cases of breast cancer reported worldwide (as per data from World Cancer Research Fund). One of the most critical test in determination of breast cancer in a patient is a fine needle aspirate (FNA) of the patient's breast mass. A hollow needle attached to a syringe is used to withdraw the required amount of tissue from the area of suspicion. Features are subsequently computed from a digitized image of the sample. These features describes various characteristics of the cell nuclei present in the image in a 3-dimensional space. In this research paper we have focused on developing a computer-aided model based on multi-layered artificial neural network that is capable of conducting analysis of the prior mentioned features and predict malignancy with high degree of accuracy. This will assist medical professionals and minimize decisional errors that can be critical to human life.

DATASET

The labelled dataset used for this project is obtained from [Kaggle](https://www.kaggle.com/uciml/breast-cancer-wisconsin-data) (also available in [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)). The version of the dataset in used is V2. The dataset comprises of numerous parameters calculated from digitized image obtained after conducting fine needle aspirate (FNA) test on 569 patients. Each of the record in turn is labelled as either 'Benign' or 'Malignant'.

EVALUATION METRIC

The various model that has been developed as part of this project will be evaluated based on the area under their individual [Receiver Operating Characteristic (ROC)](https://en.wikipedia.org/wiki/Receiver_operating_characteristic) curve.

DATA ANALYSIS

First we found the distribution of the output class and turned out that there are 357 records with an outcome of ‘Benign’ whereas 212 records have a ‘Malignant’ outcome.



Figure : Distribution of output class

So the distibution of the output data are a little imbalanced in favour of Benign outcomes. In the next step we determined the mean of the individual features and it was immedeitly visible that a feature scaling would be required to eliminate tendencies of domination of the features with higher range values during traning of the models.



Figure : Mean of individual features

We used **statistical normalization** as our preferred method of feature scaling (Equation – 1).

Equation 1 : Statistical Normalization

Subsequently determined the importance of the individual features (Table-1) in the eventual determination of the target outcome.

|  |  |
| --- | --- |
| Features | Importance |
| area\_worst | .1390 |
| concave points\_worst | .1216 |
| perimeter\_worst | .1110 |
| radius\_worst | .1090 |
| concave points\_mean | .1087 |
| perimeter\_mean | .0639 |
| concavity\_mean | .0613 |
| area\_mean | .0442 |
| concavity\_worst | .0336 |
| radius\_mean | .0250 |
| perimeter\_se | .0218 |
| texture\_worst | .0207 |
| area\_se | .0199 |
| smoothness\_worst | .0166 |
| texture\_mean | .0156 |
| compactness\_worst | .0151 |
| radius\_se | .0094 |

Table 1 : Feature Importance (Descending Order)

It immedietly became apparent that 14 of the 30 features had less than 1% importance in determination of the target outcome. But before elimination of these features from the dataset we conducted a swarm plot analysis (Figure-3) on them as a double check.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

Figure : Swarm Plot Analysis

This analysis enabled us to ascertain if one (or more) of these 14 featurs can be used as a reasonably good classifier considering their distribution with respect to the target variable. But none of the features stood out and we decided to eliminate all 14 of those features from the dataset.

MODELS :

First we have made a comparative analysis of the prediction accuracy obtained from the traditional classification models namely Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF). The result from that analysis is recorded in **percentage terms** in the table underneath.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Train | Test | NB | LR | KNN | SVM | RF |
| 90% | 10% | 91.23 | 98.25 | 96.49 | 96.49 | 96.49 |
| 85% | 15% | 93.02 | 97.67 | 96.51 | 97.67 | 97.67 |
| 80% | 20% | 92.11 | 96.49 | 96.49 | 98.25 | 96.49 |
| 75% | 25% | 93.71 | 96.50 | 96.50 | 97.20 | 97.90 |
| 70% | 30% | 92.40 | 95.32 | 95.32 | 97.08 | 97.66 |
| 65% | 35% | 93.00 | 95.00 | 95.50 | 96.50 | 97.00 |
| 60% | 40% | 93.42 | 94.30 | 95.18 | 96.49 | 94.74 |
| 55% | 45% | 93.39 | 94.55 | 95.72 | 96.50 | 94.16 |
| 50% | 50% | 93.68 | 94.74 | 96.14 | 96.49 | 95.09 |

Table 2 : Comparative Analysis of Traditional Models

DEEP LEARNING MODEL

Here we used multi-layered artificial neural networks with the same dataset after our initial analysis with the more traditional models.

|  |  |
| --- | --- |
| Parameter | Description |
| Kernel Initializer | Uniform |
| Optimizer | SGD |
| Loss | Binary Crossentropy |

We initially started observed the prediction accuracy and convergence time with one hidden layer. Then subsequently increased the number of layers noted our observations in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Train | Test | Accuracy | Early Stop |
| 90% | 10% |  |  |
| 85% | 15% |  |  |
| 80% | 20% |  |  |
| 75% | 25% |  |  |
| 70% | 30% |  |  |
| 65% | 35% | 97.00% | 163 |
| 60% | 40% | 97.81% | 118 |
| 55% | 45% | 96.89% | 221 |
| 50% | 50% | 96.84% | 203 |

Table 3 : Result of ANN Models

The deep learning model we built here was able to attain a prediction accuracy of 98.00% when the split between testing and training was 50-50. That kind of accuracy was only obtainable with Logistic Regression during our research.

RESULTS :

DISCUSSION

CONCLUSION

REFERENCES

1. Kaggle Dataset : <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>
2. UCI : <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>