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 1 : 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 2 : 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 3 : 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 :

In this section, first I have performed a comparative analysis of the prediction accuracy obtained from five well known supervised machine learning algorithms on the Wisconsin Breast Cancer dataset. Then recorded the prediction accuracy obtained from a deep learning model based on a single-layered neural network

**Model Training** : The model training in both cases was done by steadily decreasing the training set concentration (and equally increasing concentration of the test set). Then once the model was trained, its prediction accuracy was measured on the test set. In order to eliminate skewness from the prediction accuracy, median of 100 iterations were measured in each case, instead of one.

**Supervised Models** : The models involved in the analysis are 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

I noted that Random Forest takes significantly higher time to converge compared to the other traditional models.

**Deep Learning Model** : I tested the same dataset with a single-layered artificial neural network. The following hyperparameters were used in for all training experiments.

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

Table 3 : List of Hyperparameters

In this case as well, the neural network was trained with steadily decreasing training set concentration and subsequently its predictive power was tested on simultaneously increasing test set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Train | Test | Median | Minimum | Maximum |
| 90% | 10% | 100.00% | 96.49% | 100.00% |
| 85% | 15% | 100.00% | 94.18% | 100.00% |
| 80% | 20% | 97.36% | 58.77% | 98.24% |
| 75% | 25% | 97.90% | 93.70% | 98.60% |
| 70% | 30% | 97.66% | 92.98% | 98.24% |
| 65% | 35% | 97.50% | 95.99% | 98.00% |
| 60% | 40% | 97.80% | 95.17% | 98.24% |
| 55% | 45% | 97.27% | 95.71% | 97.66% |
| 50% | 50% | 96.49% | 95.78% | 97.19% |

Table 4 : Result of ANN Models over 100 iterations

RESULTS :

The predictive power achieved by the single-layer artificial neural network used in my analysis either bettered or at least equaled the best accuracy garnered by using any of the traditional classifier in every variation of test set concentration (with one exception). Eventually obtaining an accuracy of 96.49% for the highest test set (lowest training set) concentration.

CONCLUSION

I observed that the time taken by the deep learning model to converge was consistently greater in comparison to those taken by the traditional classifiers for the same training set concentration. In addition, the convergence time of the deep learning model increased steadily with decrease in training set concentration. Thus the use of a deep learning model instead of traditional classifiers came with a trade-off between accuracy and time. This can be problem in several real life applications. Hence, there is visible scope for future improvements in this area with development of deep learning models with better learning rate and faster convergence time.

REFERENCES

1. Kaggle Dataset : <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>
2. Efficient Approaches for Accuracy Improvement of Breast Cancer Classification Using Wisconsin Database
3. Deep Learning for Automatic Pneumonia Detection