ABSTRACT

Breast cancer is the second most prevalent form of cancer worldwide. In this research paper we have proposed a computer aided model based on supervised multi-layered artificial neural network that can assist medical professionals in determination of breast cancer from results of fine needle aspirate (FNA) test on breast mass of the patient. The proposed approach was evaluated on a dataset made available by the University of Wisconsin and our model achieved significant accuracy on the same.

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 dataset used for this research was obtained from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29). 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 'B' or 'M' standing for ‘Benign’ and ‘Malignant’ respectively.

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

We found that there are 357 records with an outcome of ‘B’ (stands for ‘Benign’) whereas 212 records have ‘M’ (stands for ‘Malignant’) as outcome.



Figure : Distribution of output class

Thus, the distribution of the output data are slightly skewed in favour of ‘’Benign outcomes.

We subsequently analysed the distribution of the mean of each independent variable in the dataset and it was clear that there were a couple of features that has much higher range of values compared to the others.



Figure : Mean of individual features

This necessitated a statistical normalization to be conducted on the set of independent variables.

Equation 1 : Statistical Normalization

For which we used **min-max normalization** technique as our preferred method.

FEATURE SELECTION

**Principle Component Analysis** : After the normalization of the independent features was completed, we conducted a principle component analysis (PCA) on the feature set to understand the variance explained by each of them.

The test was conducted with increasing test set concentration (starting from 10% to 50% with a step of 5% in between) and in every single instance we found that around 99% of the variance in the data was explained by 15 of the 30 features in feature set.

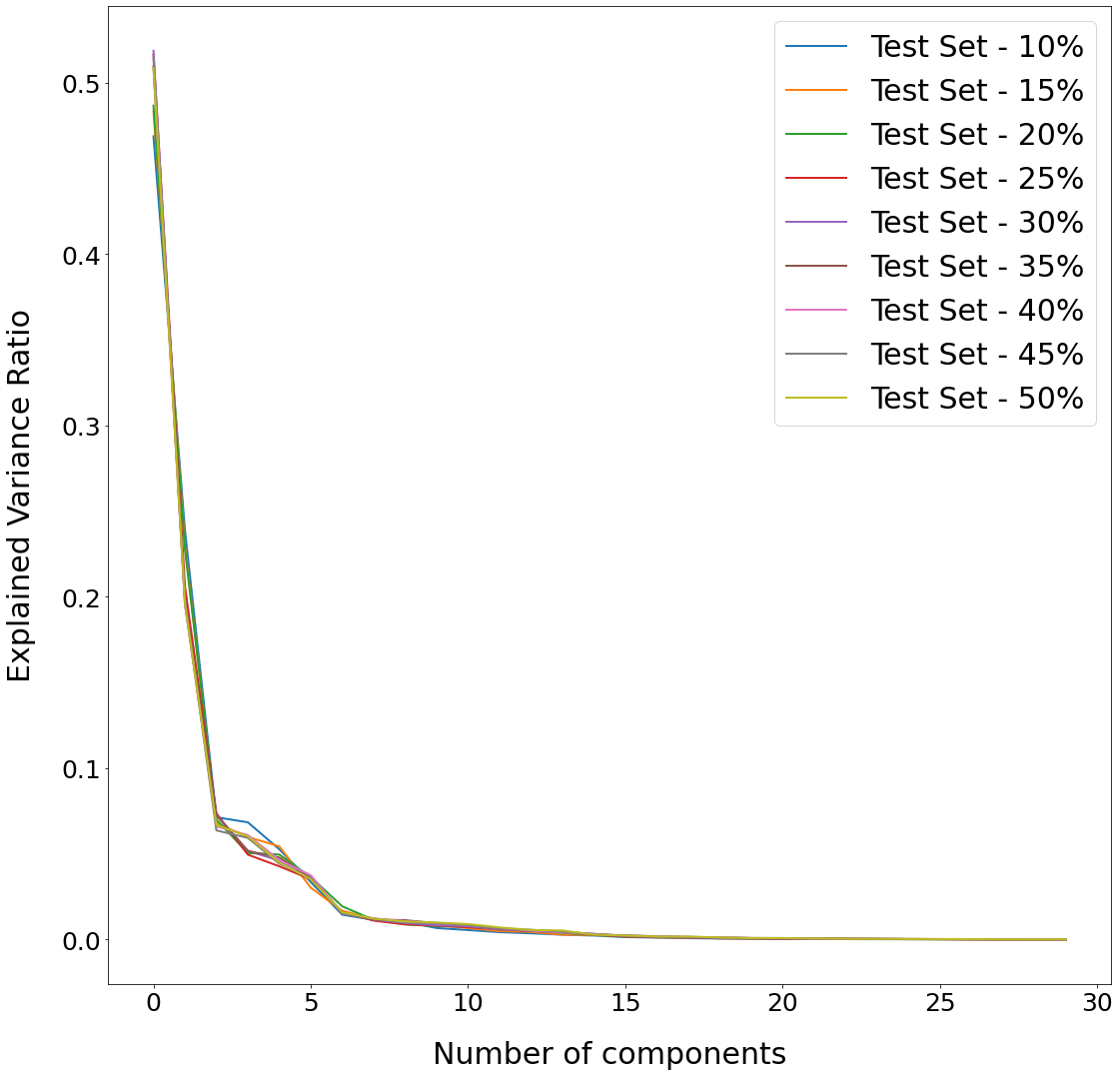


Figure : Result of Principle Component Analysis

Despite the observation we made in our PCA test, we decided to execute an analysis of variance (ANOVA) test on the feature set.

This decision was taken as the principle component analysis test only measures the extent of variance explained by the independent variables. It however does not consider the interaction between the independent variables and the output variable.

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

Figure : Result of ANOVA F-Statistics Test

We conducted this test with increasing test set concentration (starting from 10% to 50% with a step of 5% in between) as well.

We observed that the ANOVA F-Statistics test concluded around 25 of the 30 features to be useful. Though the extent of their usefulness was largely different as made clear by the range of scores obtained by the different variables.

As the main idea behind our feature selection process was to identify the features that had almost negligible impact on the outcome, the result we obtained from the ANOVA F-Statistics Test made it possible. Hence, we decided to eliminate the 5 independent variables from the dataset and use the remaining 25 for our models in the subsequent section.

MODELS :

In this section, we have first performed a comparative analysis of the prediction accuracy achieved on the Wisconsin Breast Cancer Diagnosis dataset by the various non deep learning classification models.

This served as a baseline for comparing the prediction accuracy achieved by our deep learning models.

**Model Training** : The model training in all scenarios were done by steadily increasing the test set concentration (starting from 10% to 50% with a step of 5% in between).

**Classification Models** : The classification models we used for this analysis are Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF). The prediction accuracy achieved by them are recorded in **percentage terms** in the table underneath.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Train | Test | NB | LR | KNN | SVM | RF |
| 90% | 10% | 89.47 | 98.24 | 96.49 | 100 | 94.74 |
| 85% | 15% | 91.86 | 97.67 | 97.67 | 100 | 96.51 |
| 80% | 20% | 91.23 | 96.49 | 97.37 | 97.37 | 96.49 |
| 75% | 25% | 91.61 | 96.50 | 95.80 | 96.50 | 96.50 |
| 70% | 30% | 90.64 | 95.32 | 96.49 | 97.66 | 97.66 |
| 65% | 35% | 91.00 | 95.00 | 96.50 | 97.50 | 95.50 |
| 60% | 40% | 91.23 | 94.29 | 96.05 | 96.93 | 94.30 |
| 55% | 45% | 92.22 | 94.16 | 95.33 | 96.50 | 94.16 |
| 50% | 50% | 93.33 | 95.08 | 95.79 | 97.19 | 95.09 |

Table 1 : Comparative Analysis of Classification Models

**Stacking :**

**Deep Learning Model** : Here, we propose a deep learning model based on a single-layered artificial neural network with the following hyperparameters as described in Table - 3. In Table – 4 we have recorded the median accuracy achieved by the model for varying training set concentrations.

For deep learning models, in order to eliminate skewness, the median accuracy obtained over 100 iterations was recorded for each training set concentration.

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

Table : 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 : 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

Efficient Approaches for Accuracy Improvement of Breast Cancer Classification Using Wisconsin Database

Deep Learning for Automatic Pneumonia Detection