Breast cancer

Shajahan Abdul Rehman Basith , Habeebulla Shaik , Chandrakanth seth

Introduction:

We are performing a classification task to predict if a patient has a benign or malign breast cancer, based on image features from a Fine Needle Aspiration (FNA).

An FNA is taken from the breast mass. This material is then mounted on a microscope slide and stained to highlight the cellular nuclei. A portion of the slide in which the cells are well-differentiated is then scanned using a digital camera and a frame-grabber board. The user then isolates the individual nuclei using an image processing software. When all of the nuclei have been isolated, values for each often characteristics of each nuclei are computed, measuring size, shape and texture.

Methodology:

We have used Matlab for classifying the data given and learning applications in classification learner. We have plotted scatter plots by selecting the features relation between them in terms of classifying the type of cancer.

Initially we import the data in to the work space and using the application by default we do five fold cross validation to the training data to get the scatter plot. It is a two dimensional data visualization that uses null points to represent the values for different variables. Here the data is seen by collection of points where the two dimensions determines the features.

We are calculating fisher ratio to determine the importance of the best features in classifying the data given. Fisher ratio is signal to noise ratio we can see fisher ratio in the figures below. This approach is usually done to remove some of the unwanted features to reduce the dimensionality.

Linear logistic regression uses a linear equation and it can classify only between two class since it is a binary classifier. We import the data into the work space, we do cross validation and we train using linear logistic regression application with best features. We repeat this process till we achieve high accuracy with optimal number of neurons.

Multi layer perceptron we try to prune it by adding and removing the features that resulted in least degradation of fisher index and till we get optimal number of MSE error.

Result:

1. Scatter plot for the relationships between different types of Breast cancer, here the dimensions represent the features.

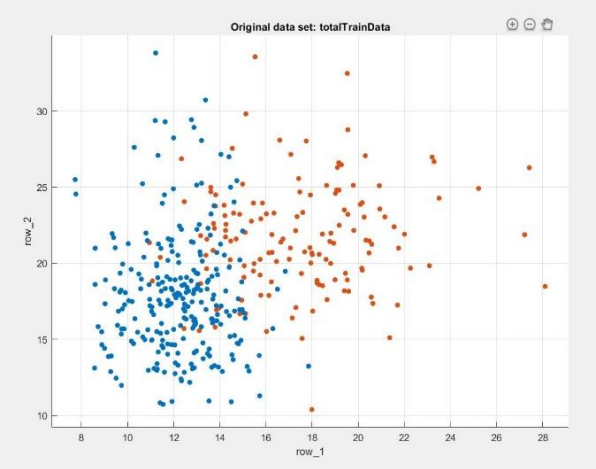

Figure:1

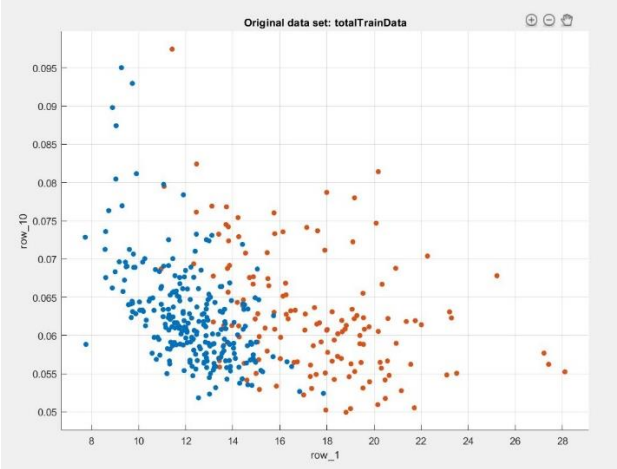

Figure:2

2. Fisher ratio for the given features

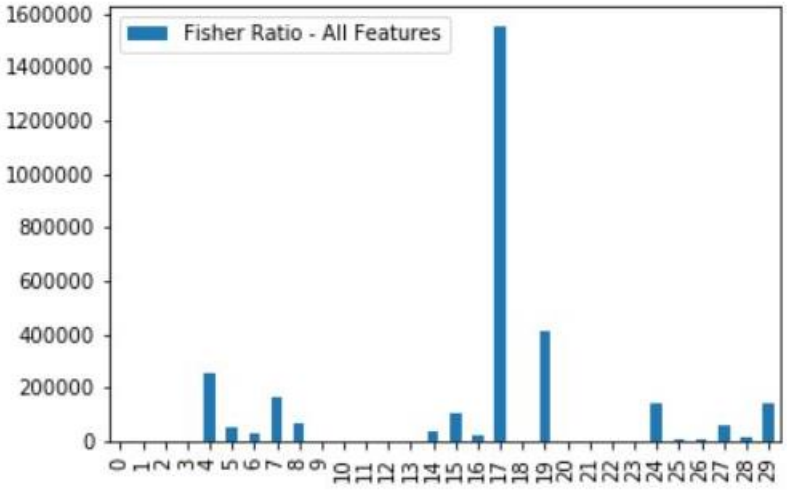

Figure:3

3. Linear logistic regression classifier which uses a linear equation and classifies only two classes.

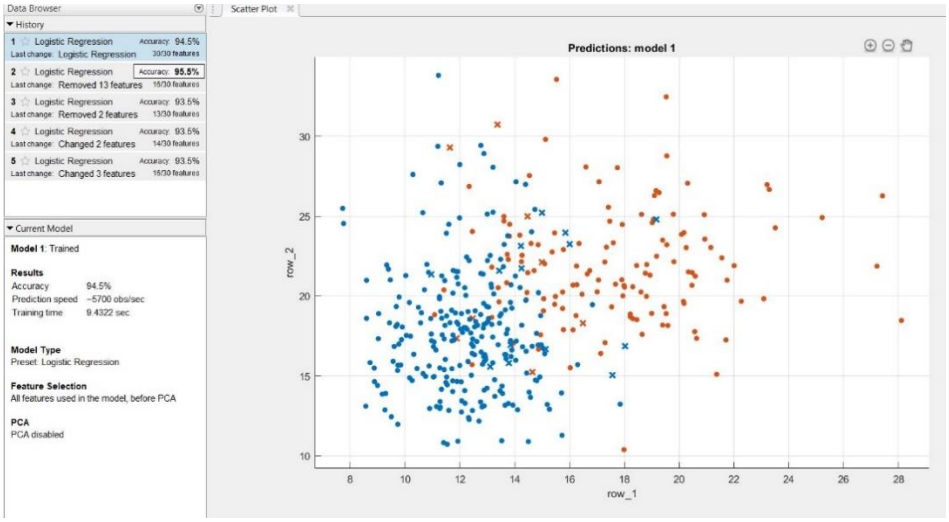

Figure:4

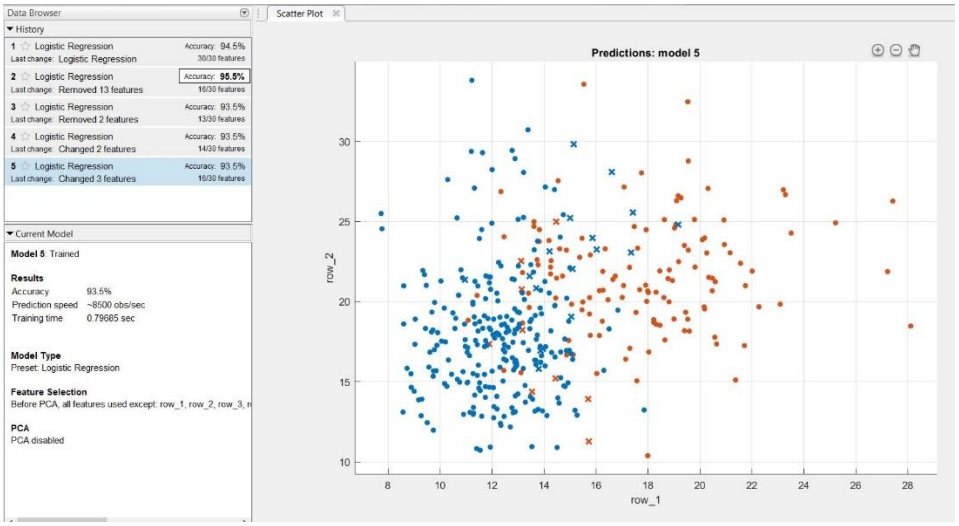

Figure:5

Best model with limited features

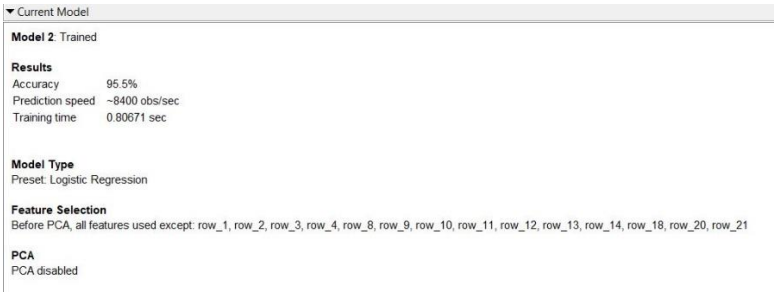

Figure:6

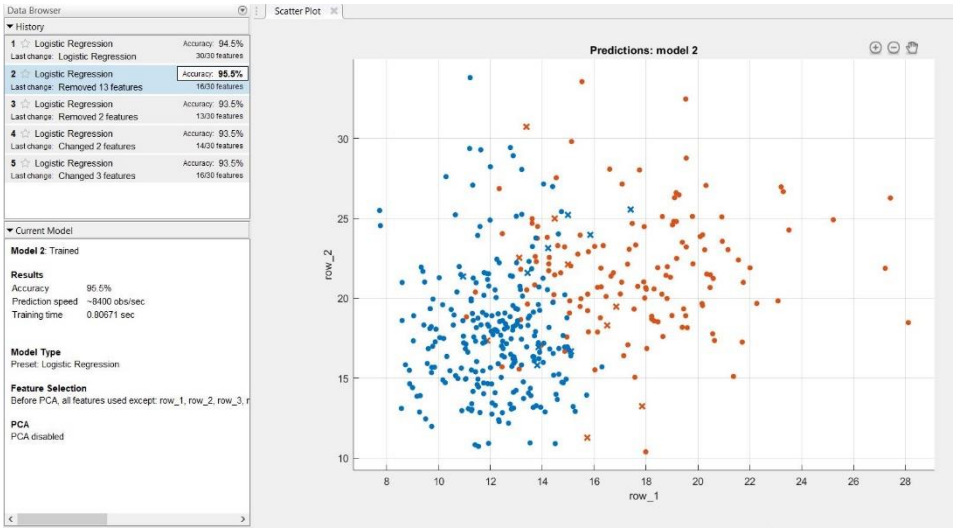

Figure:7

4,5. Multilayer perceptron (MLP) model using all the inputs and Generalization Error

- **All inputs for the Linear classifier**

Inputs: cancerTrainX
Targets: cancerTrainY

Training: 280 Samples(70%)
Validation: 60 Samples(15%)
Testing: 60 Samples(15%)

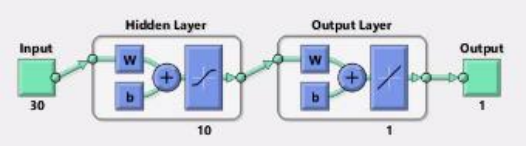

Figure:8

Trained Neural Network using algorithm: Bayesian Regularization (for more accuracy)

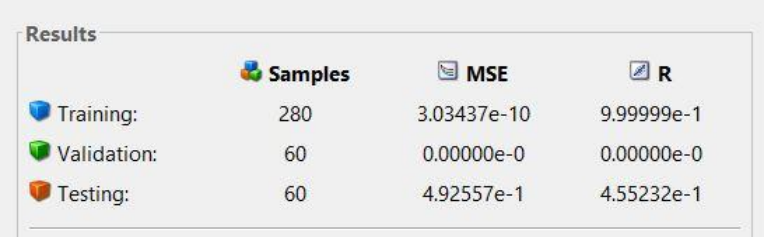

Figure:9

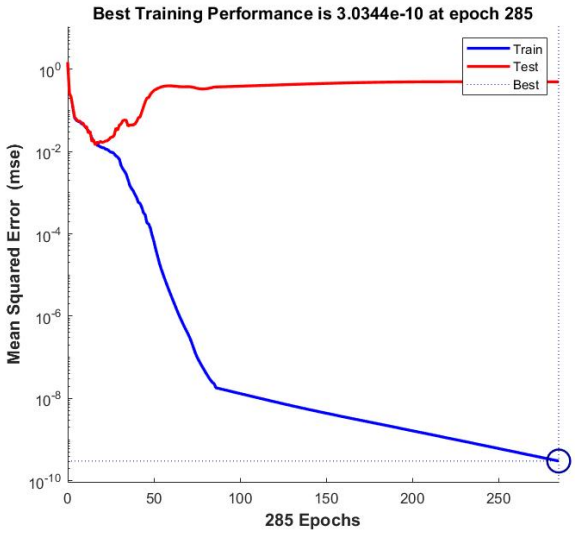

Figure:10

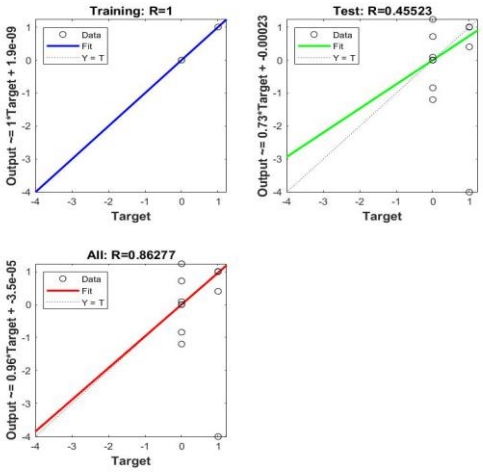

Figure:11

Best inputs for the Linear classifier: With 10 neurons

Inputs:	linear_bestfeatures
Targets:	cancerTrainY
Training:	280 Samples(70%)
Validation:	60 Samples(15%)
Testing:	60 Samples(15%)

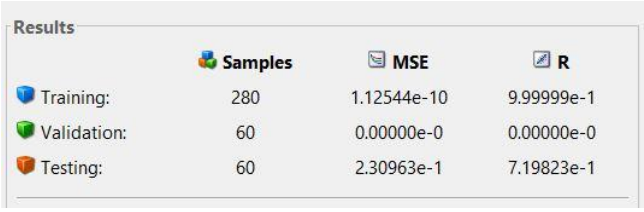

Figure:12

Figure:13

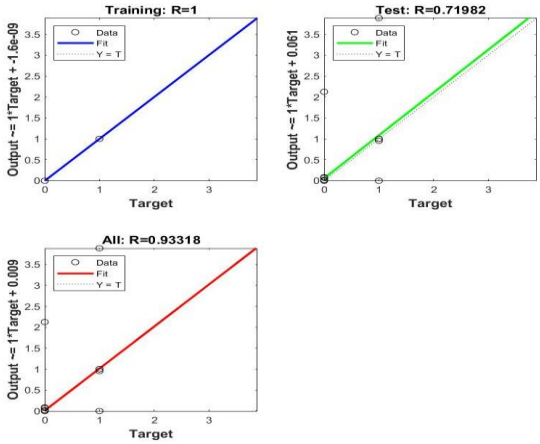

Figure:14

Best inputs for the Linear classifier: With 5 neurons

Inputs:	linear_bestfeatures
Targets:	cancerTrainY
Training:	280 Samples(70%)
Validation:	60 Samples(15%)
Testing:	60 Samples(15%)

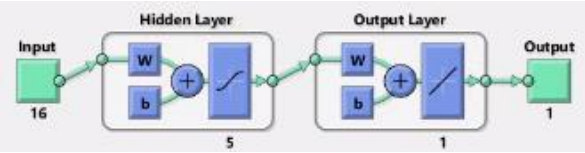

Figure:15

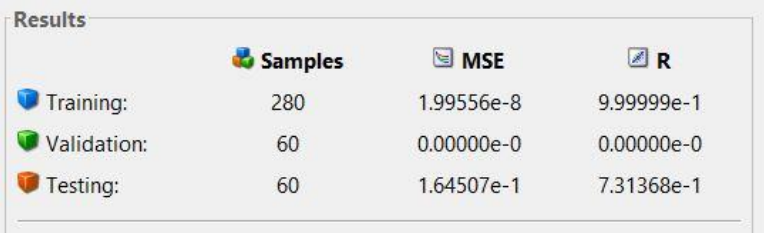

Figure: 16

Best inputs for the Linear classifier: With 50 neurons

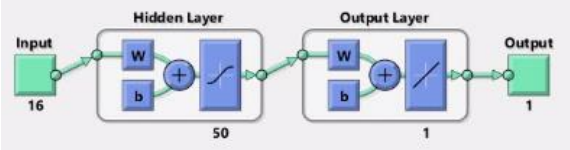

Figure:17

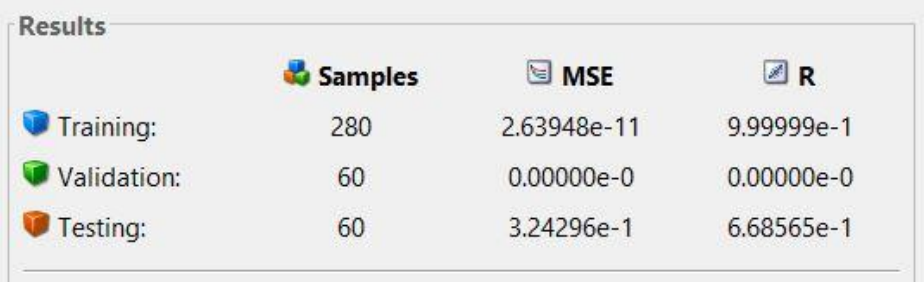

Figure: 18

4. **Outputs:** These files have been attached to the documents sent.


Linear regression Outputs.ods


MLP outputs.ods

Codes:


LinearRegression_All_code.m


LinearRegression_BestCode.m


LinearRegression_1_Code.m


MLP_All_Features_Code.m


MLP_Best_Features_Code.m