Detection of Breast Cancer using various Machine Learning methods

Declaration:

I(Rekha Kondeti) hereby declare that I am the sole author of this whole paper and have used the textbook as reference.

Abstract:

The dataset (Title: Wisconsin Diagnostic Breast Cancer (WDBC)) used in this project differentiates malignant and benign tumors. This contains 569 instances(samples) and 32 attributes(features) including ID number and Diagnosis (M = malignant, B = benign). This gives 212 malignant, 357 benign tumors.

This paper investigates different classification models like LDA, DLDA, QDA, DQDA, SVM(RBF), SVM(Linear), SVM(polynomial with degree 2) and SVM(polynomial with degree 3) to classify samples into malignant and benign tumors. Finally used ensembling using sum,min, max product rules and also stack generalization to get a better performance model. The accuracies of models LDA, DLDA, QDA, DQDA, SVM(RBF), SVM(Linear), SVM(polynomial with degree 2) and SVM(polynomial with degree 3) are 95.9%, 94.7%, 95.9%, 94.7%, 98.2%, 95.3, 79.5%, 89.4% respectively. Also the the accuracies of sum,min,max, product rules are 92.1% ,88.6%, 96.2%,94.7% and stacking using logistic regression is 97.6%. The main goal of using different methods is to compare all the classifiers and to get the best classifier with better accuracy.Among all these, the Stacking ensemble method gives better accuracy of 97.6%. The investigation of this paper is conducted on python which has special libraries for machine learning.

Introduction:

Machine learning is a subclass of Artificial Intelligence which is used to find the patterns in the data. Though the idea of machine learning has existed since decades, with the growth of data and computational resources it gained its significance.

These days breast cancer is a common disease in women which causes death. Hence,diagnosis of this disease should be easy and accurate to differentiate between malignant and benign tumors.Hence the aim of the paper is to study and compare different machine learning methods and also get best classifier which perfectly differentiates between malignant and benign tumors

Methods (*How*) :

Normalization: It is part of data preparation.It is a technique of scaling values into range[0,1].It changes data into common scale without getting change in range of values.When features have different ranges it is required [3] page 333-335.

Linear Discriminant Analysis(LDA): It is generally known as Fisher's Linear Discriminant.LDA is a supervised algorithm which computes the direction and maximizes the multiple class separations. It gives linear separation of data.Unlike PCA, LDA finds the direction of the vector where data are projected and also classes are well separated [1], page 140-145.

Scatter matrices should be nonsingular to perform LDA where in real world it may be in singular and hence this fails.To avoid this first we need to perform PCA and then perform LDA on the same [1], page 140-145 .

Quadratic Discriminant Analysis(QDA): It gives non linear separation of data.Unlike from LDA, QDA has different covariance matrix separately for each class.It is more flexible and is better than LDA but it has more number of parameters when compared with LDA [2], page 96-105.

Diagonal Linear Discriminant Analysis(DLDA): It is a special case of LDA where off diagonal elements of the covariance matrix are assumed to be "zero". It belongs to the family of Naive Bayes classifiers.The advantage of this class is it has less parameters to estimate when we have a large number of features[2], page 96-105.

Diagonal Quadratic Discriminant Analysis(DQDA): It is a special case of QDA where off diagonal elements of the covariance matrix are assumed to be "zero". It belongs to the family of Guassian naive Bayes [2], page 96-105.

Support Vector Machine(SVM): It is a supervised algorithm used in both classification and regression problems.It finds hyperplanes which separates different classes with maximum margin [4,]page 349-350.

Kernel Trick: It internally performs complex data transaformations.It is a technique which converts non separable problem to separable problem.It transforms low dimensional data to high dimensional data [5], page 359-361.

We have different types of Kernels like RBF, Linear and polynomial.Linear is used for linear hyperplanes. RBF and Polynomial are used for non linear hyperplanes

Ensemble Method: Ensembling is a combination of decisions from different models to get improved performance. It gives more accurate results than a single model would give. There are multiple ways to combine models [6], page 488-494..

Few of the combination rules we use are sum(Sum of scores),min(min of scores),max(max of scores) and product(product of scores).Also we used the Stack generalization(Logistic Regression) concept to get improved performance [6], page 488-494.

Classification Matrix: it shows the performance of the classifier in binay way [7], page 560-565.

Roc: It is a plot showing classifier model performance at all thresholds.It is plot of GAR vs. FAR [7], page 560-565.

AUC: It gives the aggregate measure of the classifier across all thresholds [7], page 560-565**.**

Results and Discussions (*Analysis and Justification of Results*)

The goal of this paper is to check the classifiers performance and to compare them.Further more to get the highest accuracy classifier out of these.

Comparison of results of LDA,QDA,DLDA and DQDA:

Below table shows the accuracies of testing data for the models LDA,QDA,DLDA and DQDA:

	LDA	**QDA**	**DLDA**	**DQDA**
AUC	1	0.99	0.99	0.98
Specificity	94%	99%	94%	97%
Sensitivity	98%	90%	94%	90%
Overall performance	95%	95%	94%	94%

As we know, the parameters of DLDA and DQDA are reduced, it makes computation easy but it reduces the performance of the model. From the above,

performance of LDA and QDA are better when compared to DLDA and DQDA. Bias is introduced when the covariance matrix is reduced. If the covariance matrix is not reduced, variance is introduced.This very clear from the above table [8], page 106.

LDA and QDA accuracies are the same but QDA has more parameters as it has different covariance matrices.Hence it is good to consider a simple model which can fit better on testing data because it is more robust(Occam's Razor) [9], page 468.

Comparison of results of SVM(RBF),SVM(Linear) and SVM(polynomial with degree 2) and SVM((polynomial with degree 3):

Below table shows the accuracies of testing data for the models LDA,QDA,DLDA and DQDA:

	SVM(RBF)	**SVM(Linear)**	**SVM(polynomial with degree 2)**	**SVM((polynomial with degree 3)**
AUC	1	1	0.82	1
Sensitivity	100%	98.1%	78.9%	86.1%
Specificity	95.4%	90.4%	81.5%	100%
Overall performance	98.2%	95.3%	79.5%	89.47%

From the above table,it is clear that overall performance of SVM(RBF) is higher than all other models. SVM(Linear) has also better accuracy which is comparatively less than SVM(RBF). Though accuracy of SVM(RBF) is high ,it is a complex model and as per Occam's Razor we should select a simple model with better accuracy. Hence consider SVM(Linear) as the model with better accuracy [9], page 468.

Comparison of results of top 3 ensembles using SUM,min,max and product rule along with stack generalization (Logistic regression):

Accuracy of top 3 ensembles with combination rules sum,min,max and product rules are 94%,92% and 96%.

Accuracy of Stack generalization using logistic Regression is 97%.

From the above we can see that stack generalization gives better accuracy than all the others.As we give different model outputs as inputs to the metalearnerie Stack generalization,it gives better accuracy. [10], page 509-511

Conclusion:

The methods used in this paper are few and there are many more methods(classifiers) in machine learning.It all depends on the ratio of how we split training and testing data. Here we divided in the ratio of 70%(training) and 30%(testing) and worked on the above classifiers.Among all the classifiers used, Stack generalization(logistic regression) gives the highest accuracy of 97%. Hence it is considered as the better model with better accuracy.

Reference

[1] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 6, page 140-145.

[2] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 5, page 96-105.

[3] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 12, page 333-335.

[4] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 13, page 349-350.

[5] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 13, page 359-361.

[6] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 17, page 488-494.

[7] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 19, page 560-565.

[8] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 5, page 106.

[9] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 16, page 468.

[10] Ethem Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014, section 17, page 509-511.

Appendices:

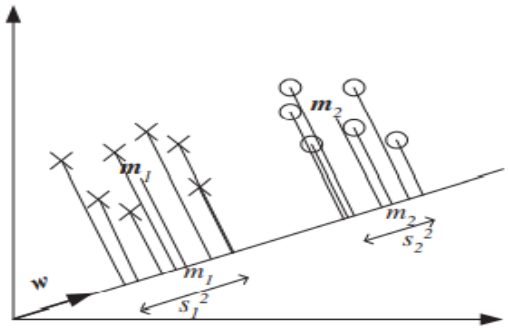

Figure 1: Two-dimensional, two-class data projected on w [1], page 140-145.

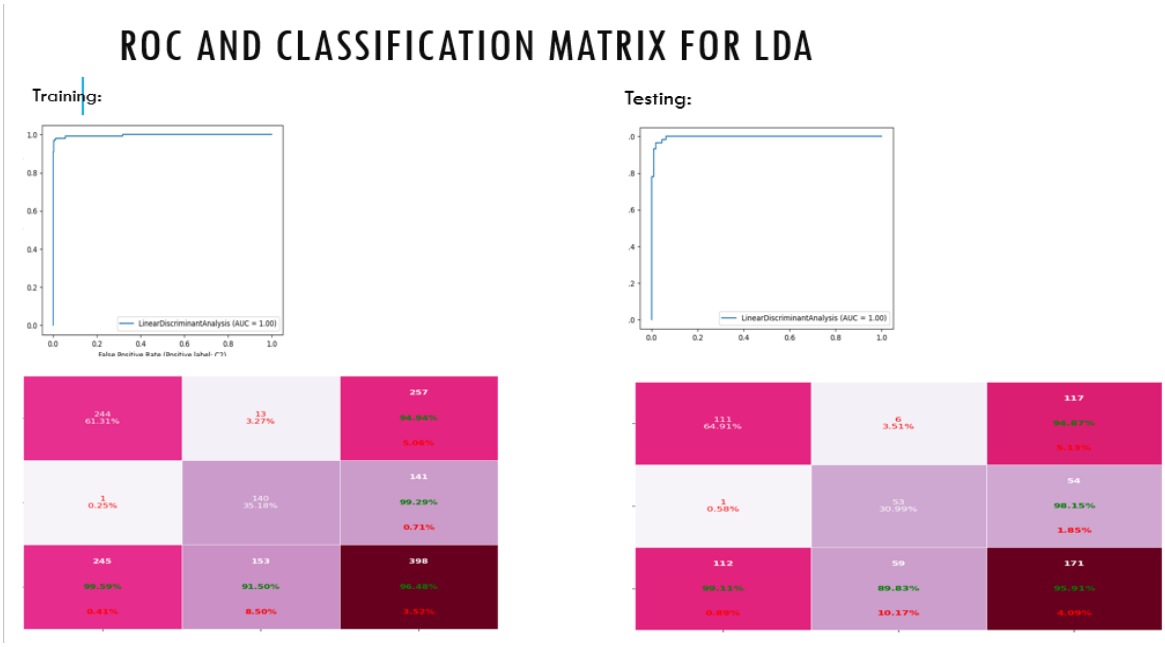

Fig2: Roc and Classification matrix for training and Testing forLDA

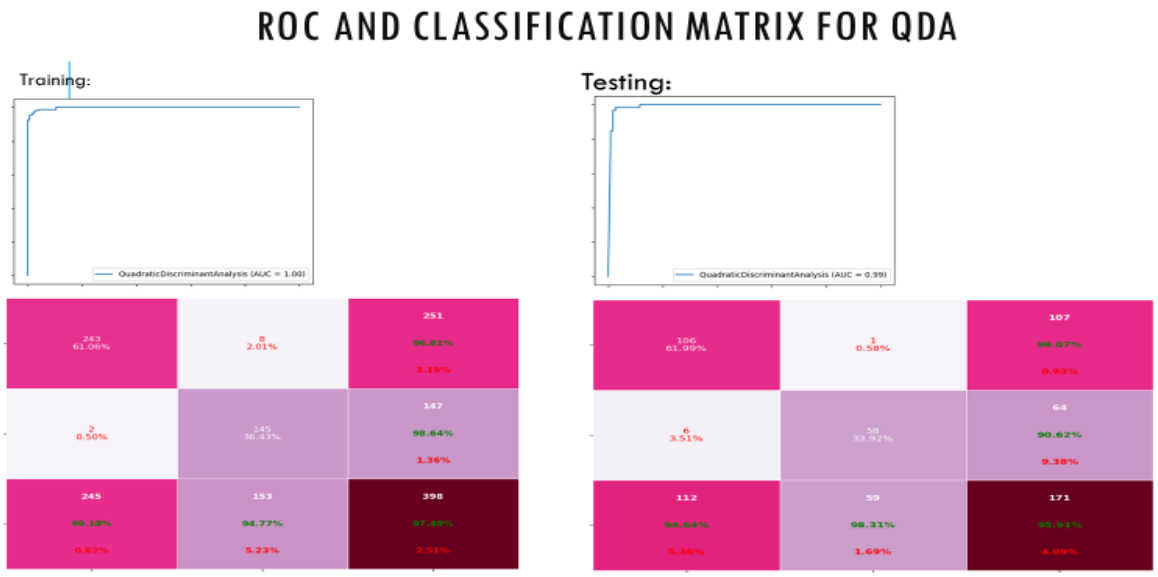

Fig3: Roc and Classification matrix for training and Testing for QDA

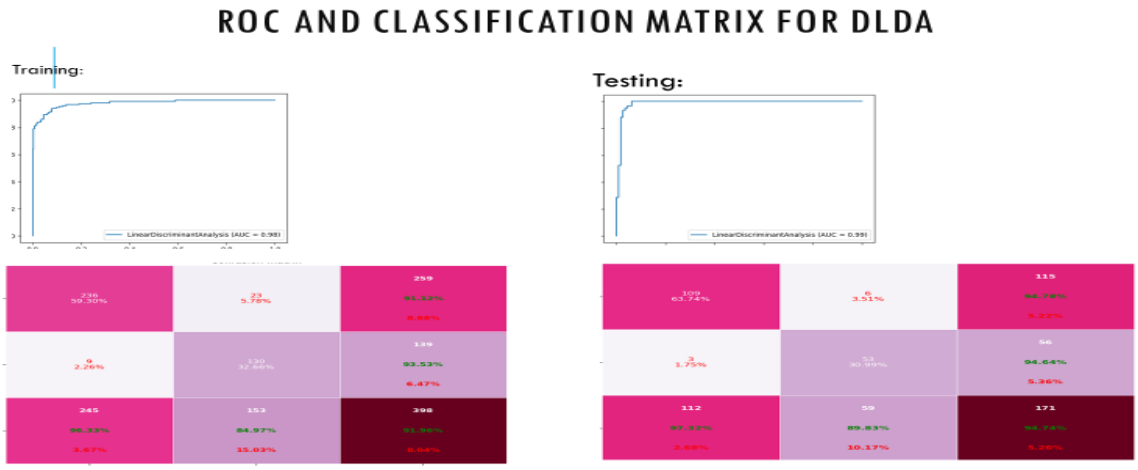

Fig4: Roc and Classification matrix for training and Testing for DLDA

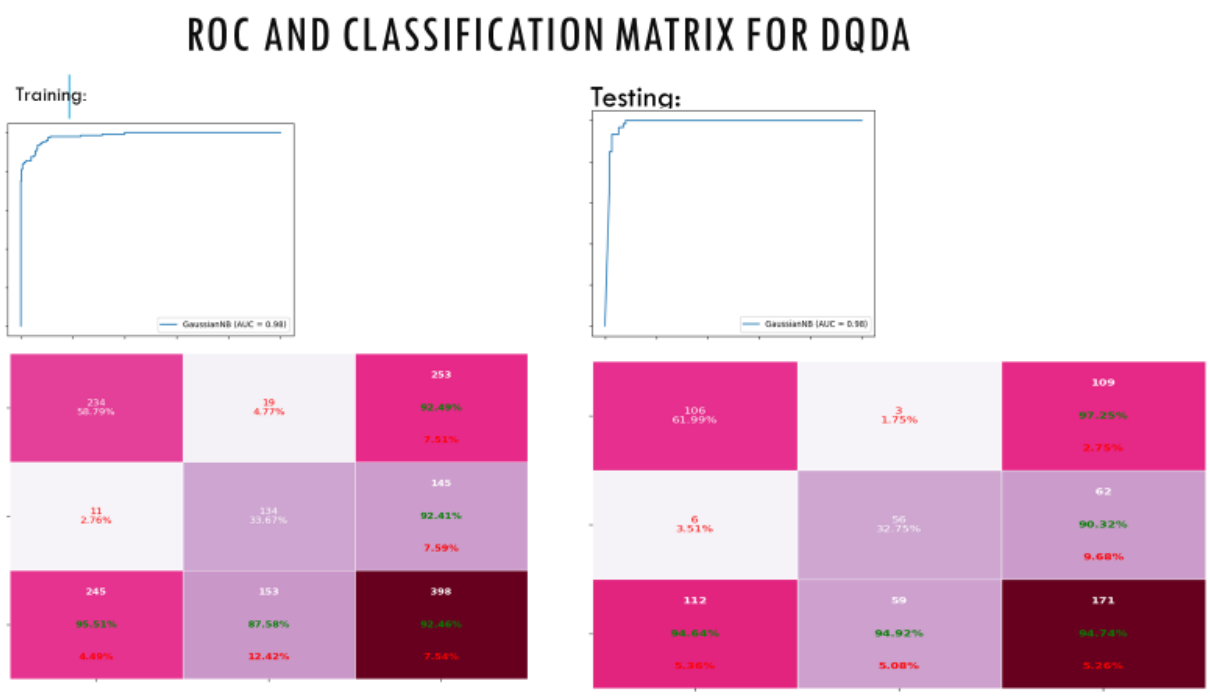

Fig5: Roc and Classification matrix for training and Testing for DQDA

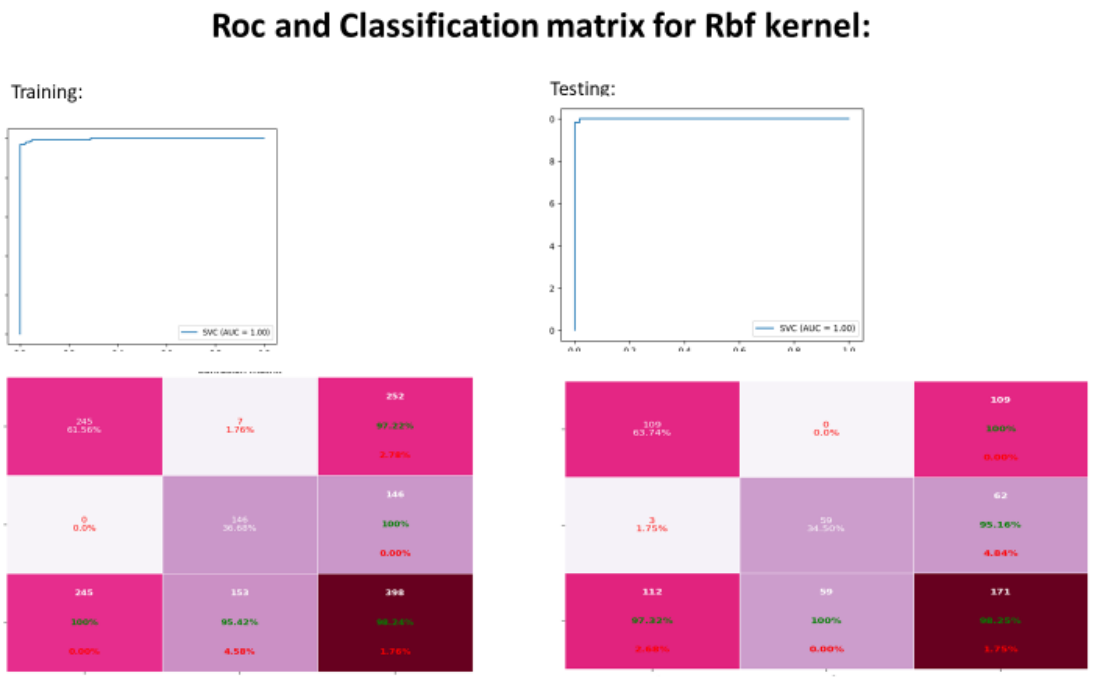

Fig6: Roc and Classification matrix for training and Testing for SVM(RBF)

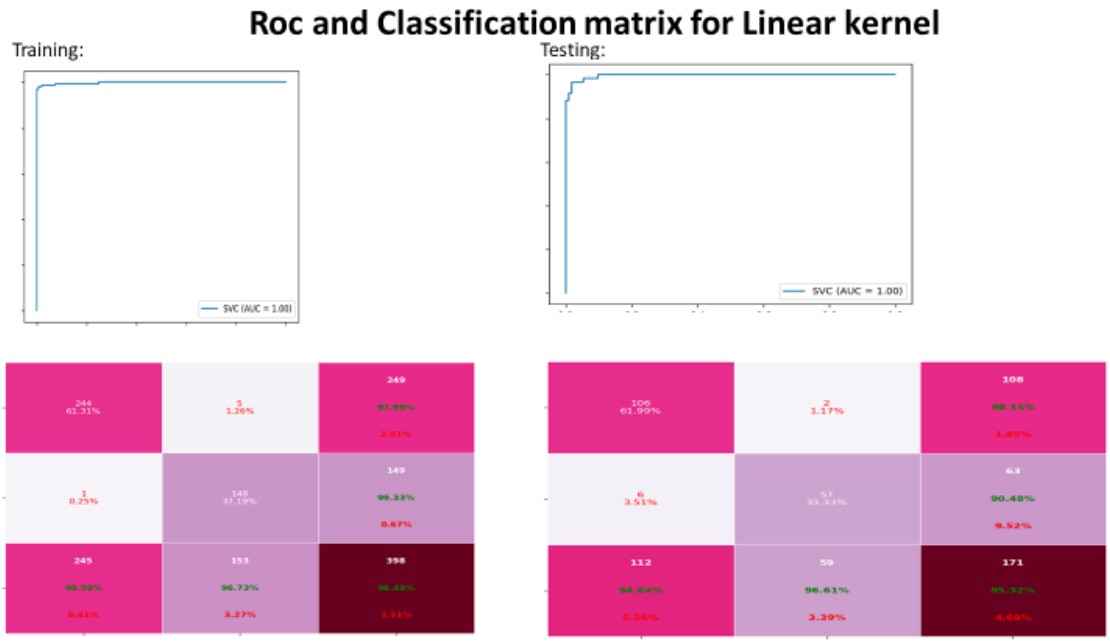

Fig7: Roc and Classification matrix for training and Testing for SVM(Linear)

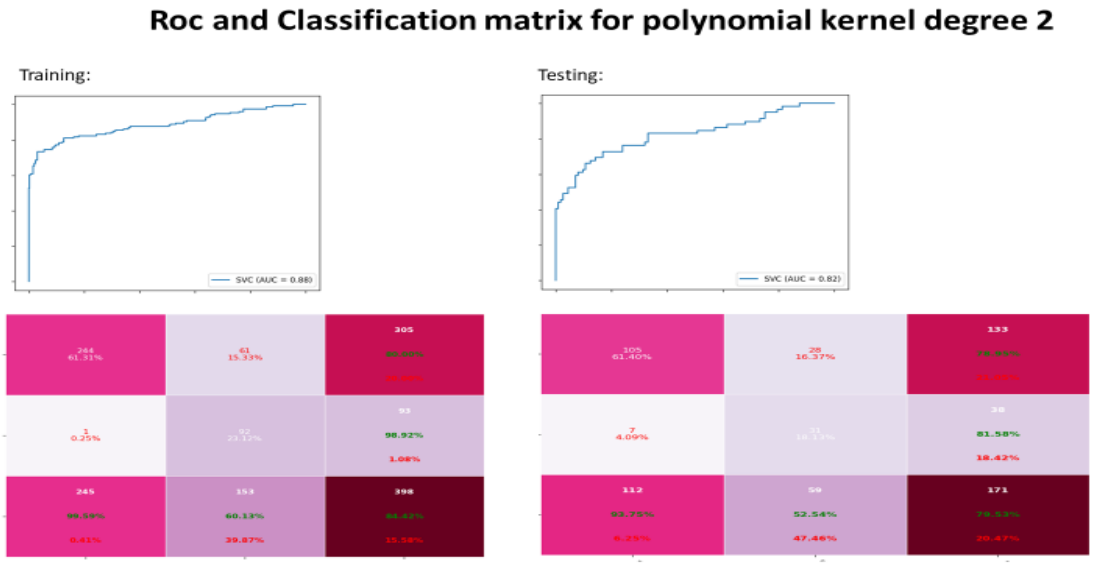

Fig8: Roc and Classification matrix for training and Testing for SVM(Polynomial with degree 2)

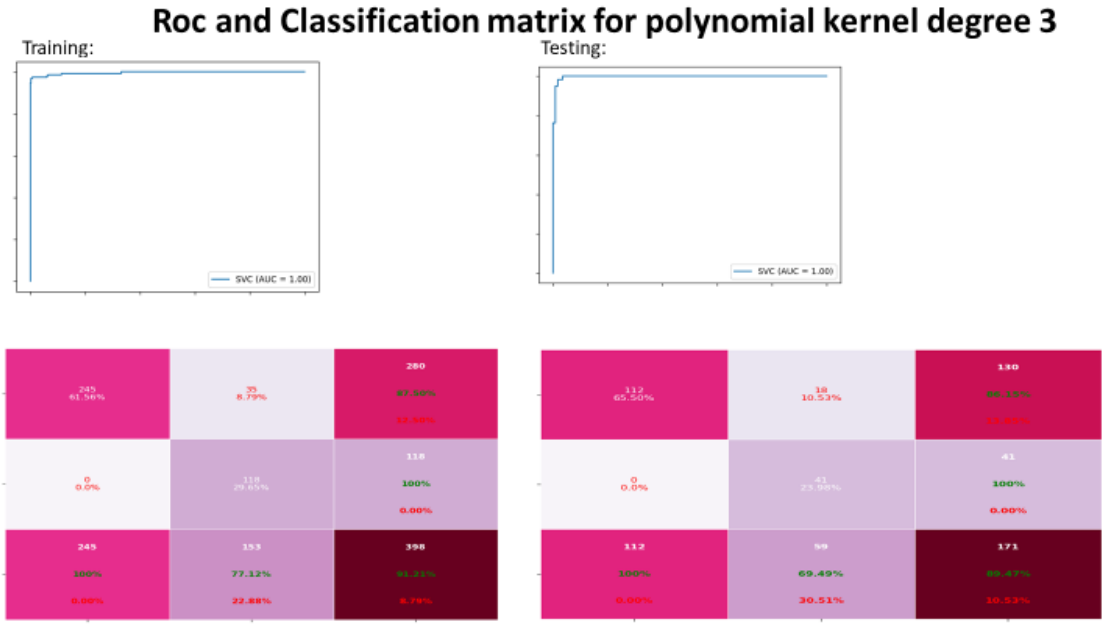

Fig9: Roc and Classification matrix for training and Testing for SVM(Polynomial with degree 3)

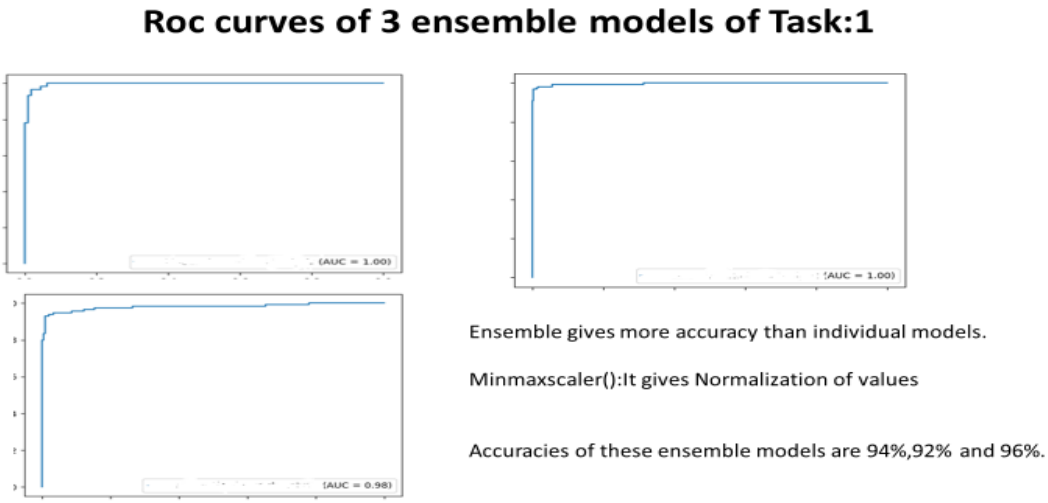

Fig10: Classification matrix for Testing using sum,min,max and product

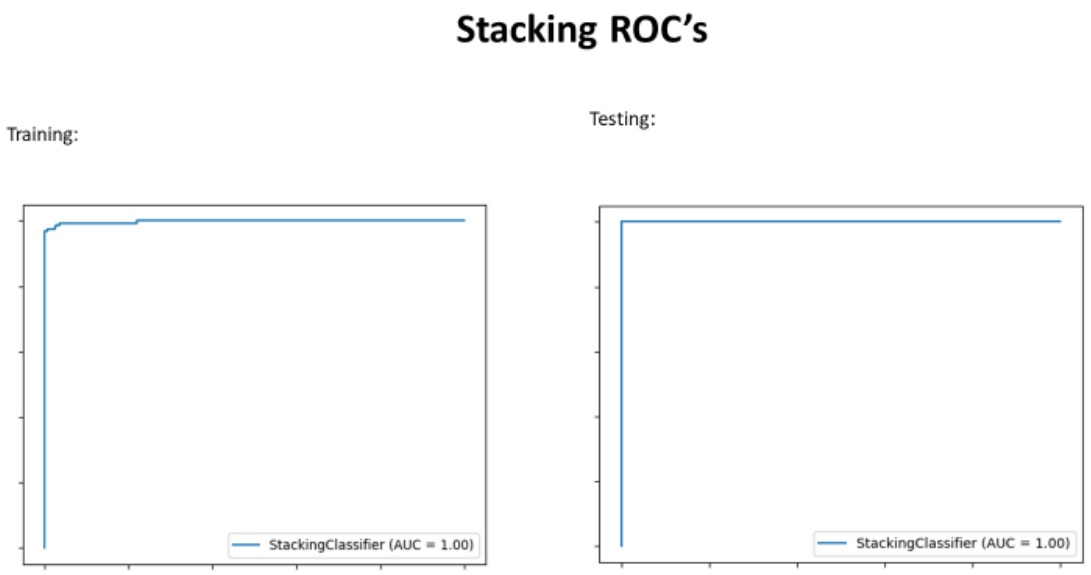

Fig11: Classification matrix for Training and Testing using Stack Generalization(Logistic Regression)

Table1: Confusion Matrix

		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive **+**	**False Negative (FN) Type II Error**	True Positive **(TP)**

FAR (False Accept Ratio) = FP/(TN+FP) known as type II error

FRR (False Reject Ratio) = FN/(FN+TP) known as type I error

GAR (Genuine (true) Accept Ratio) = TP/(TP+FN) known as ‘Sensitivity’

GRR (Genuine Reject Ratio) = TN/(FP+TN) know as ‘Specificity’