ASSIGNMENT-3

SWAGATAM CHAKRABORTI(MT18146)

ASSUMPTIONS:

a. All the images have been converted to gray scale.
b. The covariance matrix elements have been rounded off to 2 decimal

1. Dimensionality Reduction

d. Projected data visualization after performing LDA

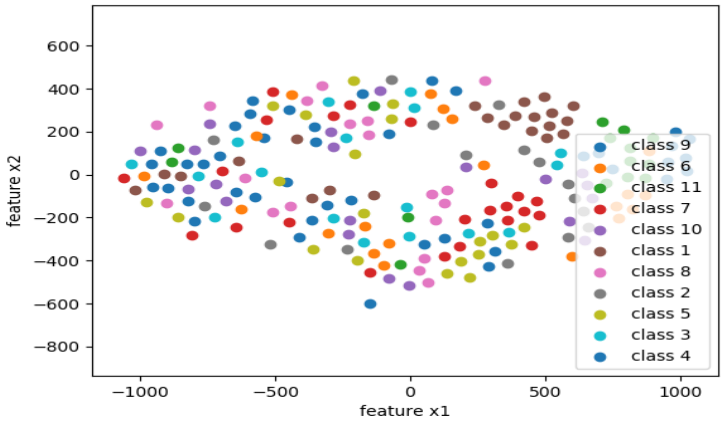

Figure 1: projected data visualization on face dataset

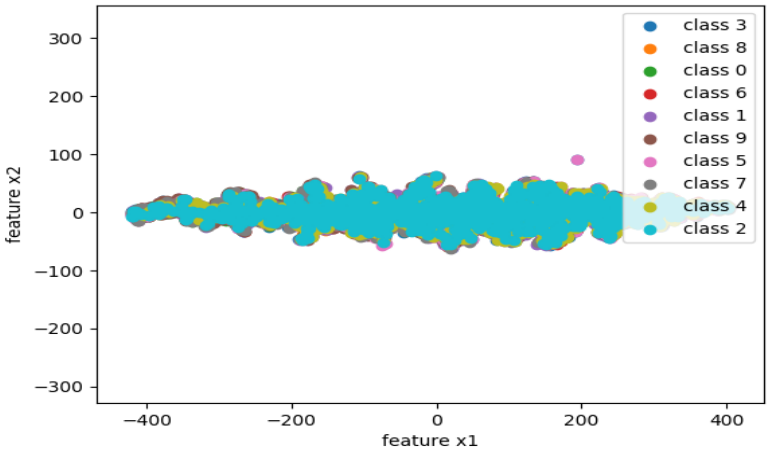

Figure 2: projected data visualization on cifar dataset

e. Classification of the LDA projected data using 5 fold cross validation

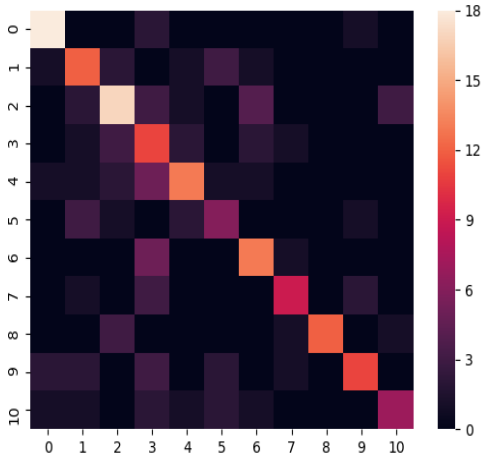

Figure 3: Face Data Confusion Matrix

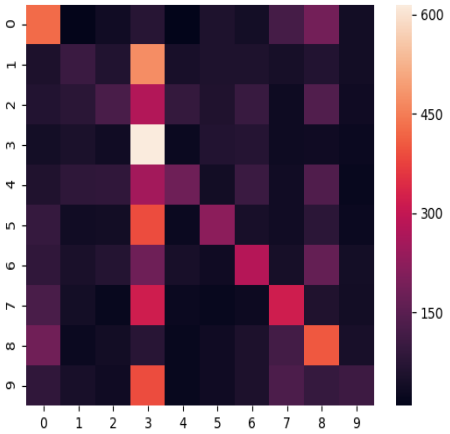

Figure 4: Cifar Data confusion Matrix

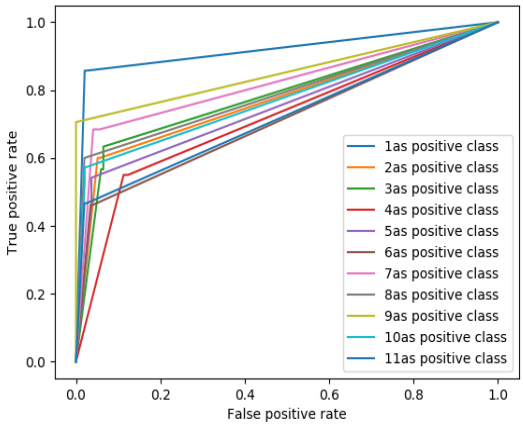

Figure 5: Face Data ROC curve

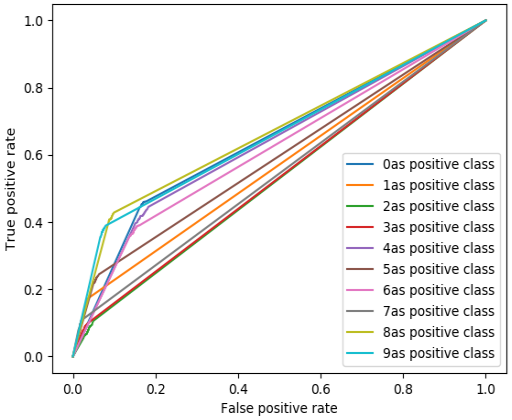

Figure 6: Cifar Data ROC curve

	Face Dataset	*Cifar Dataset*
Mean Accuracy	60.2	25.456
Standard deviation	7.782030583337486	0.4036632259693723
Final accuracy	60.0	26.619999999999997

Table 1: Result analysis across different dataset

f. Classification of the PCA projected data using 5 fold cross validation

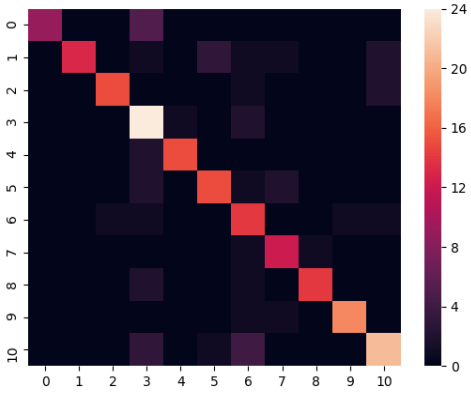

Figure 7: Face Data PCA Confusion Matrix

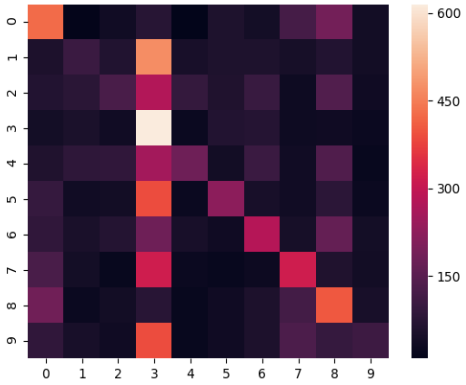

Figure 8: Cifar Data PCA Confusion Matrix

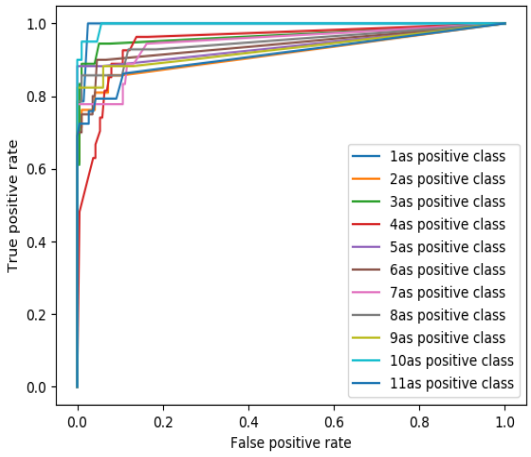

Figure 9: Face Data PCA ROC Curve

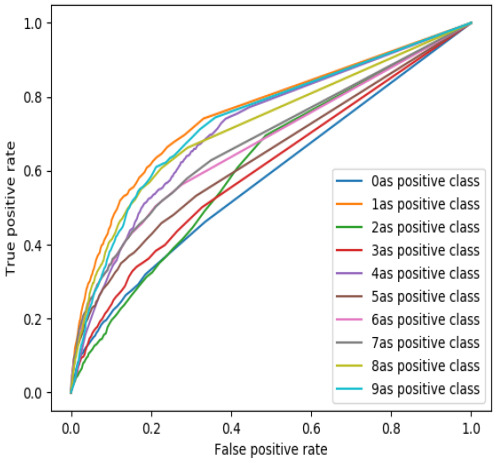

Figure 10: Cifar Data PCA ROC Curve

	Face Dataset	*Cifar Dataset*
Mean Accuracy	79.2	27.278
Standard deviation	4.445222154178574	0.6026408549044759
Final accuracy	79.06976744186046	27.83

Table 2: Result analysis across different dataset

g. Comparison between pca and lda across iterations

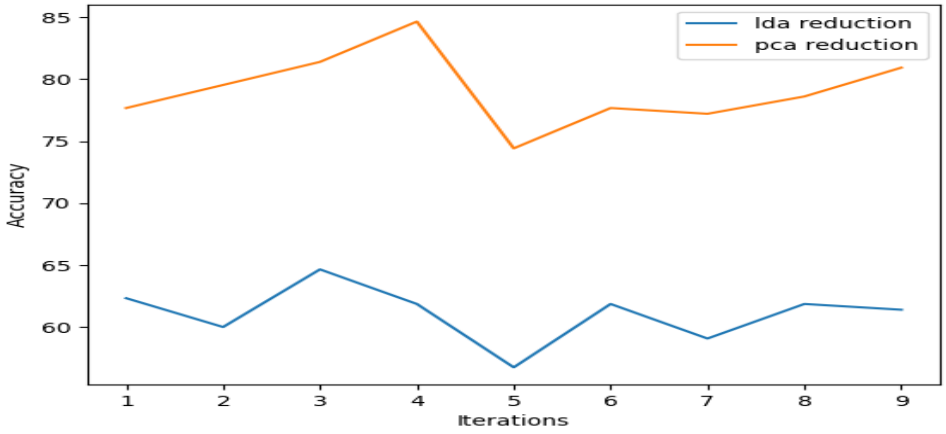

Figure 11: Comparison between pca and lda reduction techniques

INFERENCE: From the graph it is evident that pca showed better accuracy than lda projected datasets. We cannot comment that pca always shows better result than lda as both are dimensionality reduction techniques.

h. Visualization of the eigen vectors scorss datasets

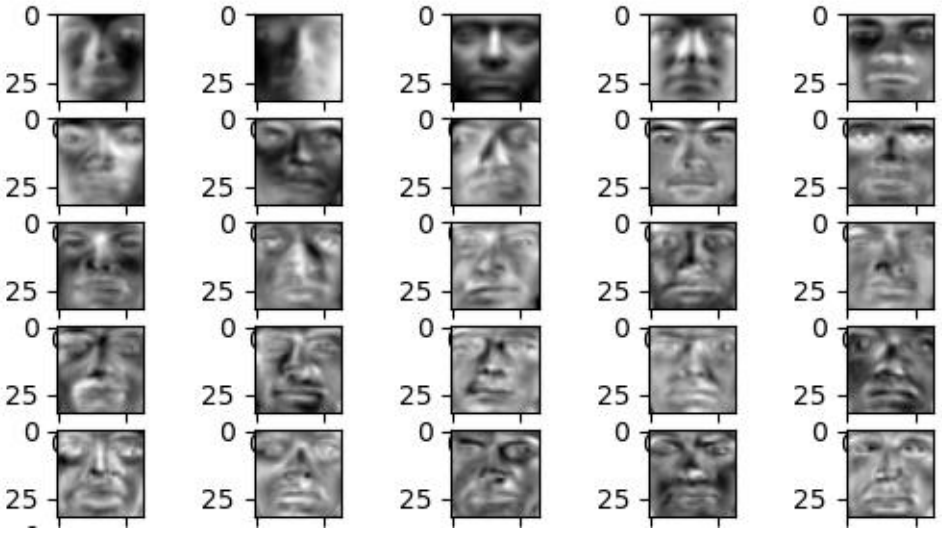

Figure 12: Eigen faces of top eigen vectors of dataset 1

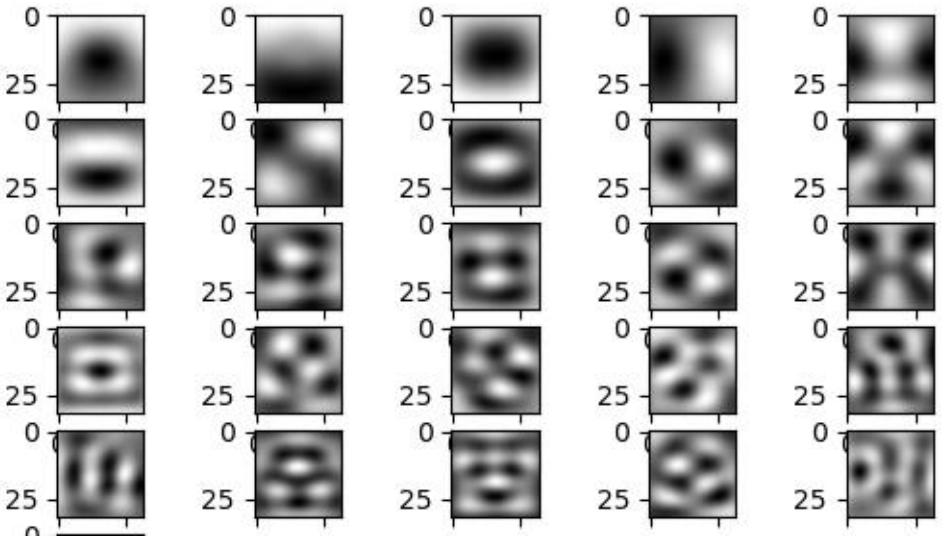

Figure 13: Images of top eigen vectors od dataset 2

Comparison over accuracy across different eigen energy over PCA

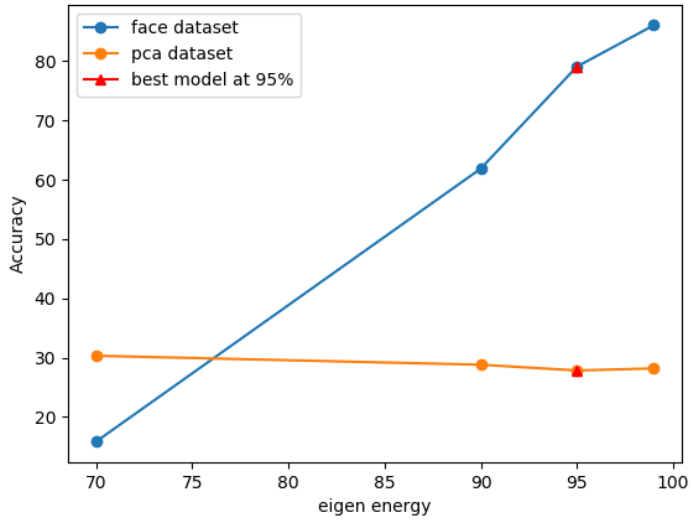

Figure 14: Comparison among the accuracy of different eigen energy across different datasets

INFERENCE: From the graph it is evident that with increase in eigen energy conservation the accuracy increase. But this is not the case for dataset 2.

i. Comparison between lda over pca and pca over lda across dataset

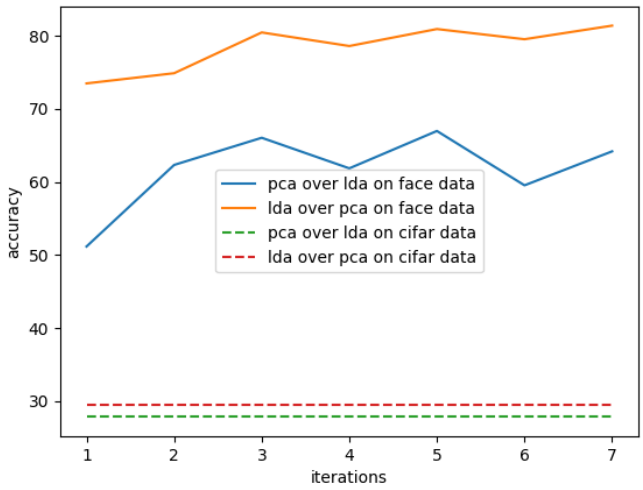

Figure 15: Comparison between LDA over PCA and PCA over LDA across datasets

INFERENCE: lda over pca is showing better result because on performing pca first we reduce the dimension by removing some features. Over that on performing lda it reduces the dimension in which the class segregation is better. On performing lda first doesn't give much good segregation of data, so pca over lda is bound to give less accuracy than former.

2. Ensemble Learning

a. Boosting

5 fold cross validation is performed is performed to obtain the tuning parameter. In boosting algorithm the tuning parameter is the number of iterations. Mean accuracy of the 5 models is obtained across all the tuning parameters and plotted in the below graph.

No. of Classifiers	*Mean Accuracy*	*Mean Error Rate*	*Mean Std Dev.*
20	27.27857142857143	72.72142857142856	3.648762720122412
50	35.81428571428571	64.18571428571428	1.0802966334884734
80	44.40714285714286	55.59285714285714	1.7961977982984603
100	44.15	55.85	1.5795892000188925
150	45.32142857142858	50.74999999999999	49.25000000000001
200	52.52142857142858	47.47857142857143	1.1716463212253168
250	54.35	45.64999999999999	1.5781350398751146

Table 3 Comparison of Different classifiers for tuning parameters

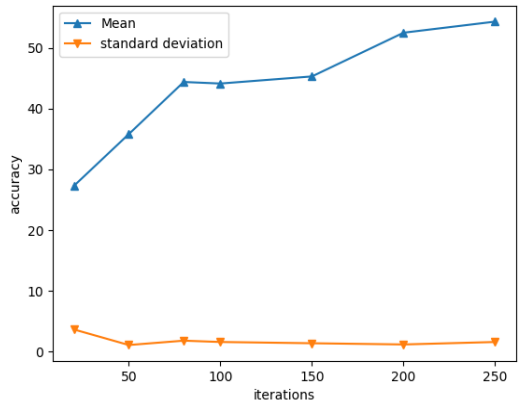

Figure 16: Accuracy across tuning parameters

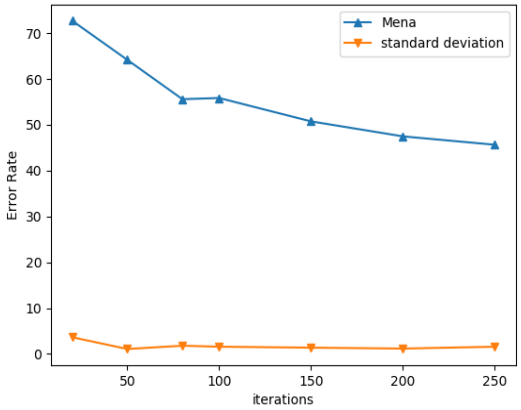

Figure 17: Error rate across tuning parameters

From the graph the best tuning parameter is selected on the basis of high mean accuracy and low error rate. So selected tuning parameter = number of iterations = 250.

On the best tuned parameter obtained from 5 fold cross validation following is the result obtained on the final accuracy.

Final Accuracy on test set = 51.*5333333* ***Final Accuracy on train data***= 53.114285714285714

b. Bagging

5 fold cross validation is performed is performed to obtain the tuning parameter. In boosting algorithm the tuning parameter is the number of iterations. Mean accuracy of the 5 models is obtained across all the tuning parameters and plotted in the below graph.

No. of Classifiers	*Mean Accuracy*	*Mean Error Rate*	*Mean Std Dev.*
20	17.457142857142856	82.5428572	1.8405023715257303
50	17.478571428571428	82.52142	1.6047760349808877
80	17.62142857142857	82.378572	1.5330708145578202
100	18.49285714285714	81.50715	1.5351992491835964
150	17.07857142857143	82.9214286	1.0452516541044856
200	18.164285714285715	81.8357143	2.0275651427668766
250	18.0	82.0000	2.0209363358278725

Table 4: Comparison of Different classifiers for tuning parameters

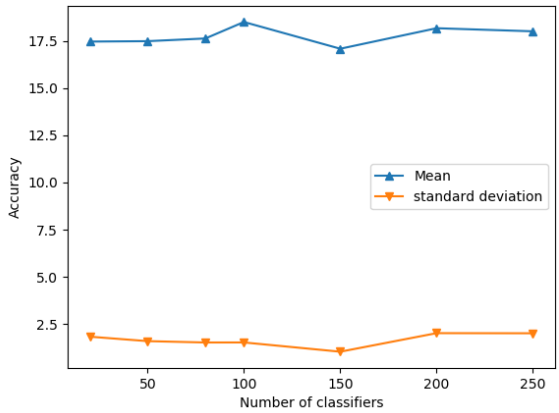

Figure 18: Accuracy across tuning parameters

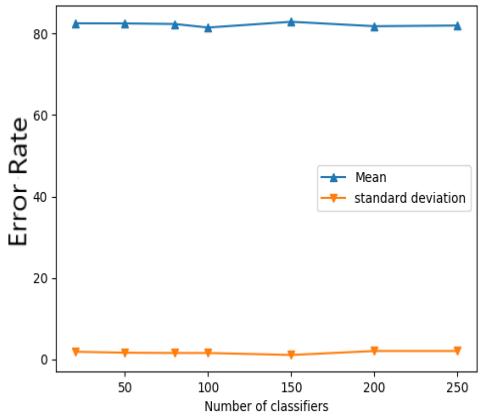

Figure 19: Error rate across tuning parameters

From the graph the best tuning parameter is selected on the basis of high mean accuracy and low error rate. So selected tuning parameter = number of iterations = 100.

On the best tuned parameter obtained from 5 fold cross validation following is the result obtained on the final accuracy.

	Train data	*Test data*
min_max	27.37142857142857	27.800000000000004
z score	28.34285714285714	28.65
tanh	28.34285714285714	28.65

Table 5: Accuracy among normalized techniques on training and testing dataset

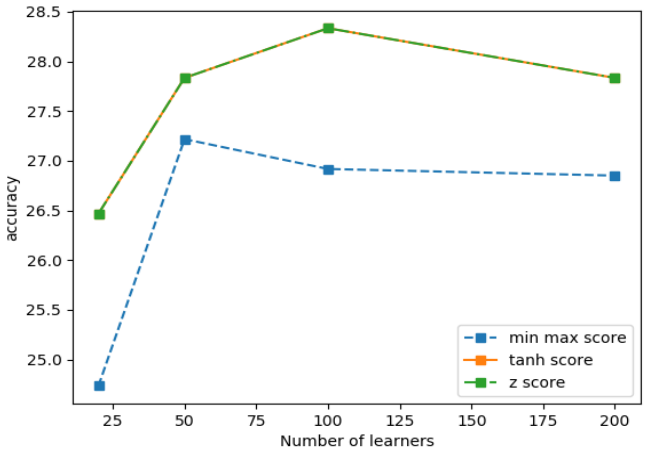

Figure 20: Comparison between different normalization techniques against accuracy

INFERENCE : From the graph it is evident that with min_max normalized technique shows less accuracy against the tanh and z score. And z score and tan h performance is almost similar as deduced from the graph.