Solutions for CCE : AIMIA Homework 2 Solutions

November 29, 2022

1 Q1

Perform principal component analysis (PCA) and project the high dimensional CT embeddings to 100 and 50 dimensions. Also, report the reconstruction error as a plot of number of principal components.
Solution:-

Principal component analysis (PCA) is a statistical technique that is used reduce the dimensionality. The PCA represent higer dimensional data with reduce number of principal componenets without losing much of the information contained in the data.

The steps involved in PCA are as follows

- Step 1: Data Standardization- make the feature values have mean zero and standard deviation as one.
- Step 2: Determine the covariance matrix for dataset features- remove the highly correlated features suggesting redundant information
- Step 3: Compute eigenvalues and eigenvectors for the covariance matrix.
- Step 4: Sort eigenvalues and their corresponding eigenvectors- sort the eigenvectors based on their eigenvalues in descending order to figure out the important principal components first.
- Step 5: Choose k eigenvalues and form a matrix of eigenvectors- selected principal components that are sufficient to represent the data without the loss of much of the information
- Step 6: Projection to principal components-Transofrm the original matrix by reorienting the data from the original axes to the ones described by the principal components. Thus the reduction in dimensions.

The Covariance matrices can be rally big with large number of features and calculation of eigenvectors could be very slow. In this kind of senarios, the singular value decomposition(SVD) could be employed to find the first top k eigen vectors. Given any mxn matrix A, algorithm to find matrices U, V, and W such that

$$A = UWV^T$$

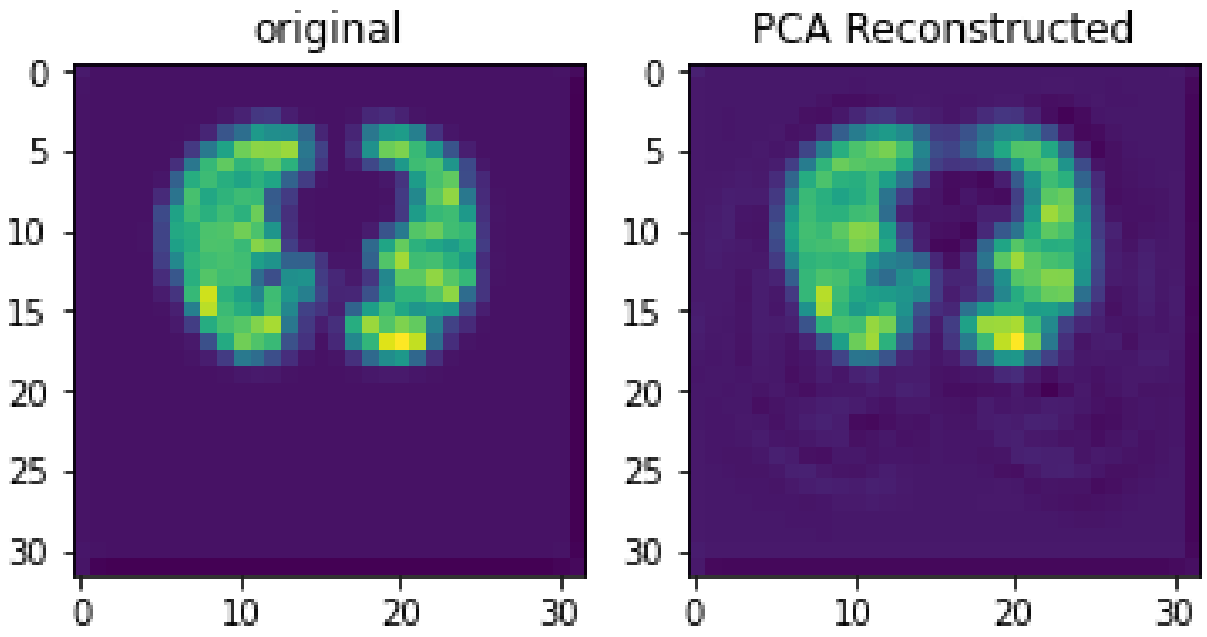

Figure 1: The original and PCA reconstruted CT slice

where U is mxn and orthonormal, W is nxn and diagonal, V is nxn and orthonormal

PCA using SVD algorithm

- Step 1: Start from m by n data matrix X
- Step 2: Recenter: subtract mean from each row of

$$X–Xc < –X–Xbar$$

- Step 3: Call SVD algorithm on Xc – ask for k singularvectors
- Step 4: Principal components: k singular vectors with highest singular values

$$(rowsofV^T)$$

- Step 5: Coefficients-project each point onto the new vectors

The PCA has been employed to reduce the 1024 features to 100 and 50 features . The PCA analysis results on the given dataset of embeddings are as follows

The PCA reconstructed CT slice has been depicted in Figure.1

The Reconstruction MSE versus Number of Principal Components is represented in Figure.2 and Figure.3 for 50 and 100 components respectively.

The scree plots(Figure.4 and Figure.5) highligting the principal components and the variance explained.

:

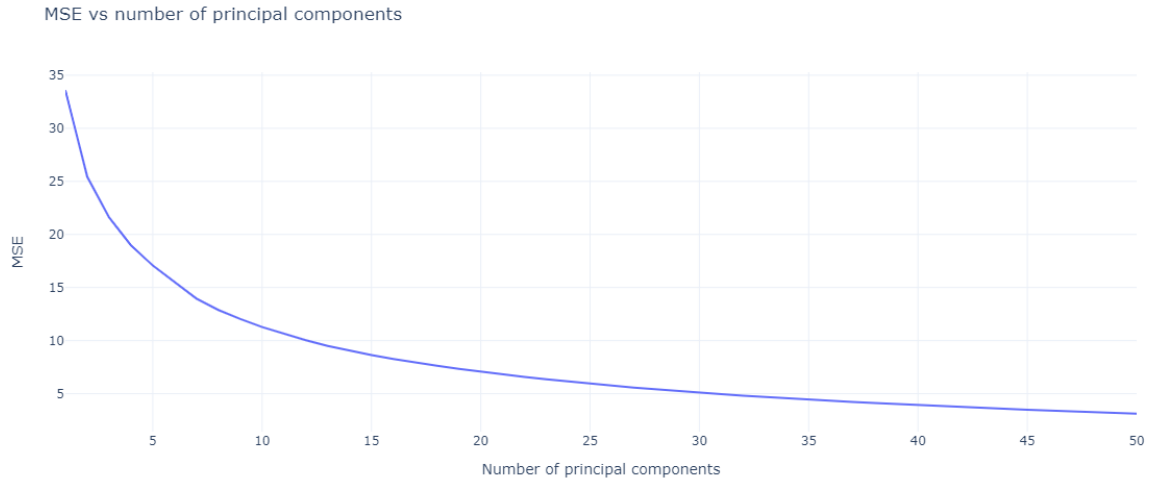

Figure 2: The MSE vs. No. of Principal components

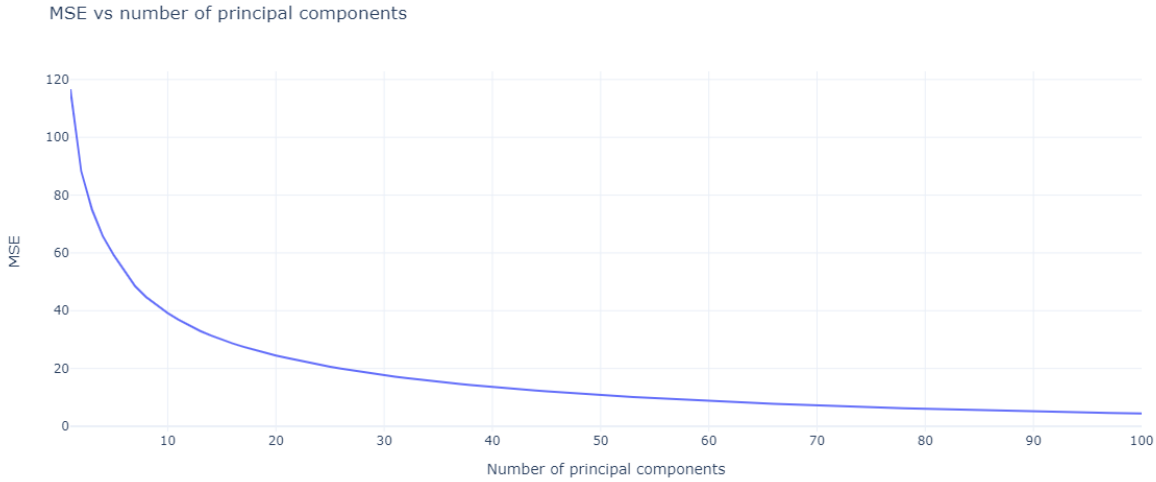

Figure 3: The MSE vs. No. of Principal components

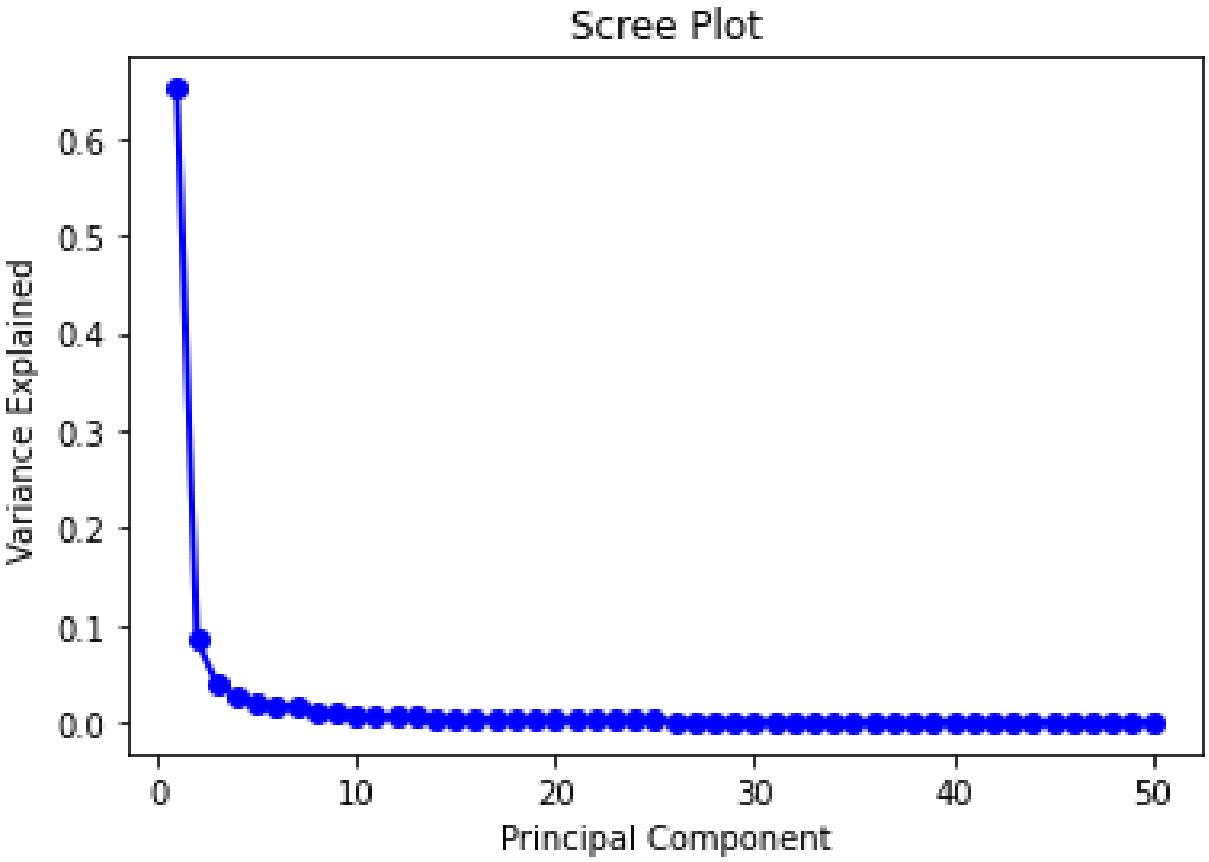

Figure 4: The Scree plot of PC vs. Varience Explained

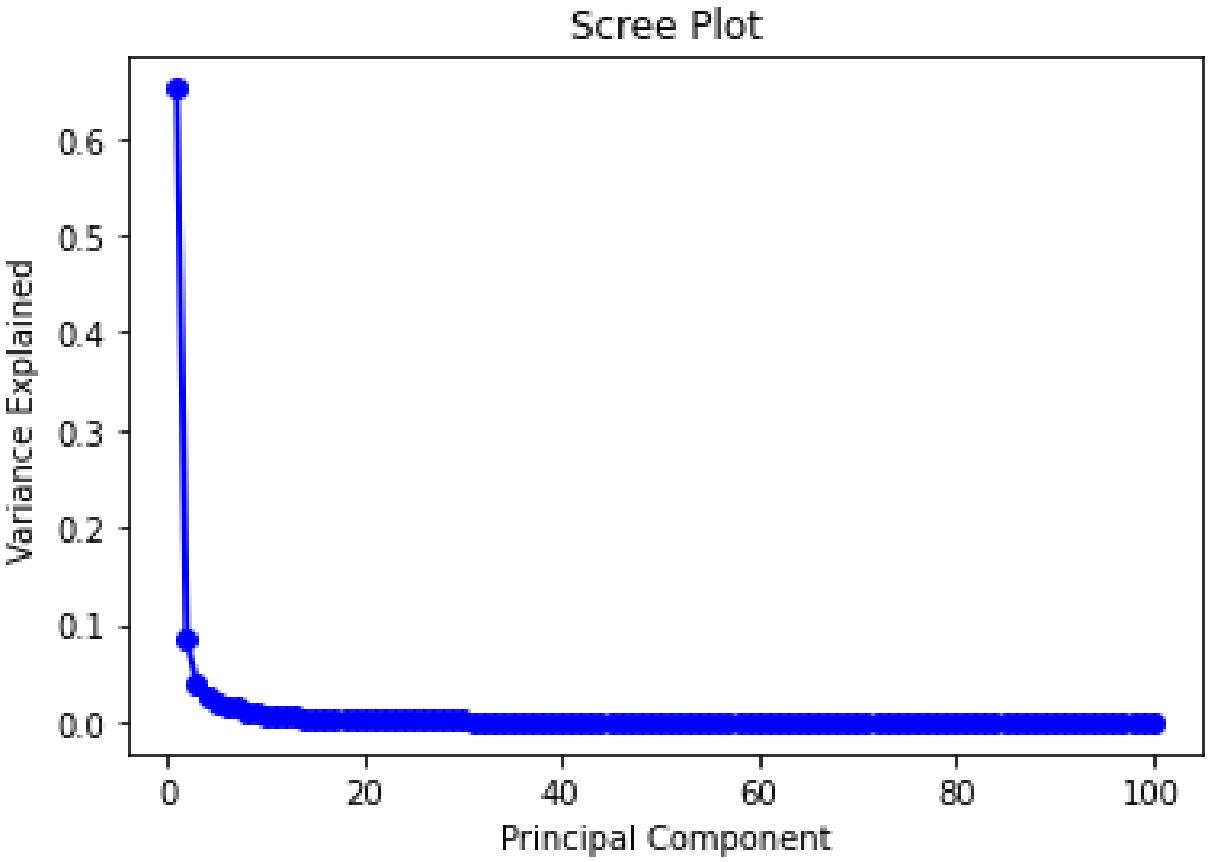

Figure 5: The Scree plot of PC vs. Varience Explained

2 Q2

Split the PCA projected embeddings into 70% training, 10% validation and 20% testing. Apply SVM with Linear kerenl and RBF kernel to classify these embeddings into Normal, Mild and Severe categories. Report the Accuracy and F1-score on training and testing set for all the classes.

Solution:-

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for solving both classification and regression.SVM can solve linear and non-linear problems. The algorithm creates a line or a hyperplane which separates the data points into different categories. There are two different types of SVMs, Simple SVM(used for linear regression and classification problems) and Kernel SVM (which has flexibility for non-linear data and can add more features to fit a hyperplane instead of a 2D space.) The SVM is effective in cases where number of features is greater than the number of data points.SVM makes use of a subset of training points in the decision called support vectors which makes it memory efficient.

Different kernel functions can be specified for the decision function.The most widely used Kernel functions are Linear and Radial Basis Function (RBF) The linear kernel works really well when there are a lot of features. Linear kernel functions are faster than most of the others and will have fewer parameters to optimize. The function which defines the linear kernel is as follows:

$$f(X) = w^T * X + b$$

Radial Basis Function (RBF) one of the most powerful and commonly used kernels in SVMs and it is the opt choice for non-linear data. The equation for an RBF kernel:

$$f(X1, X2) = exp(-gamma * ||X1 - X2||^2)$$

The PCA Projected Embeddings have been split in to train alidation and test sets as follows

- Train ratio = 0.70
- Validation ratio = 0.10
- Test ratio = 0.20

The SVM is applied with Linear and RBF kernels. The performance metrics are as follows.

SVM with Linear Kernel

- The Validation accuracy is: 92.1%
- The Model accuracy is: 93.4%

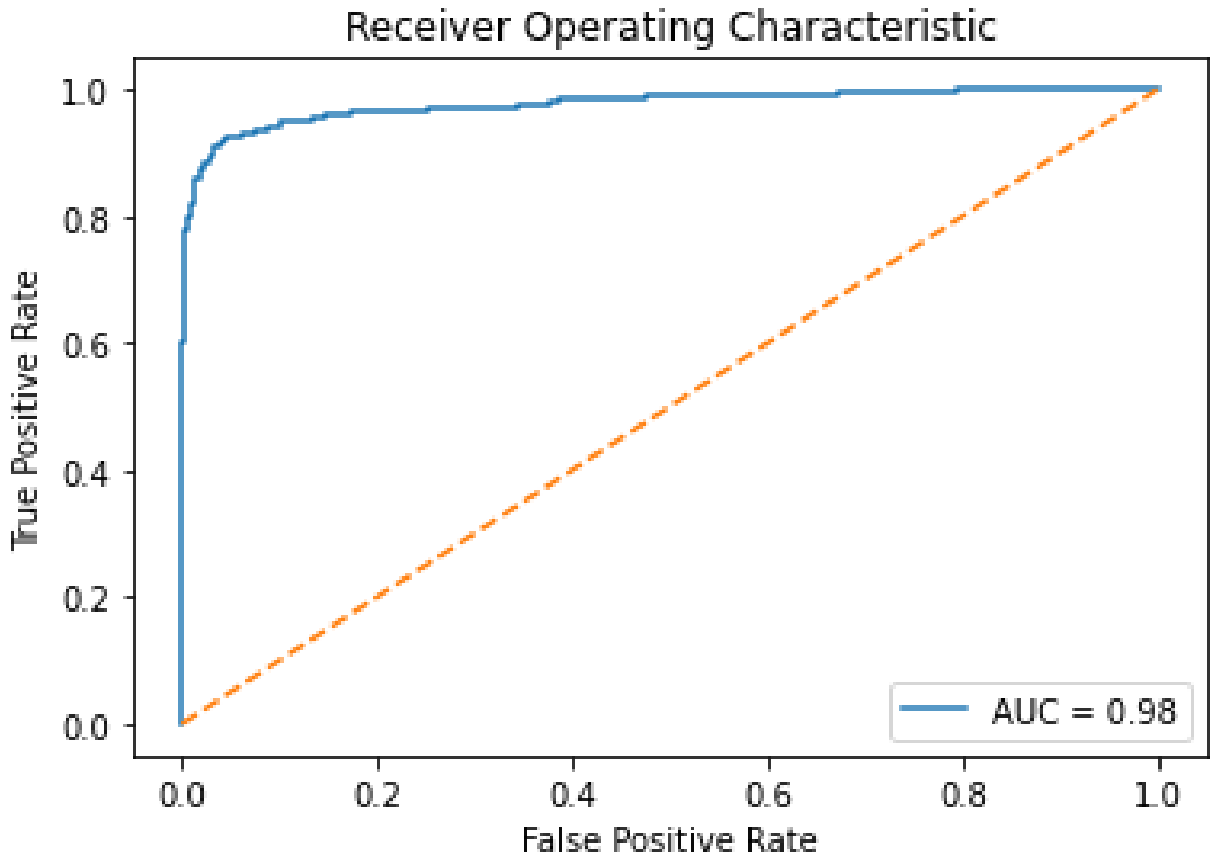

Figure 6: The RoC-SVM-Linear-Kernel

- The F1 Score is : 0.93

  The RoC curvefor SVM with Linear kernel is shown in figure 6.

  SVM with RBF Kernel

- The Validation accuracy is: 89.6%
- The Model accuracy is: 92.3%
- The F1 Score is : 0.92

The RoC curvefor SVM with RBF kernel is shown in figure 7.

The Confusion Matrix Depicting the Prediction versus actual values is shown in figure 8.

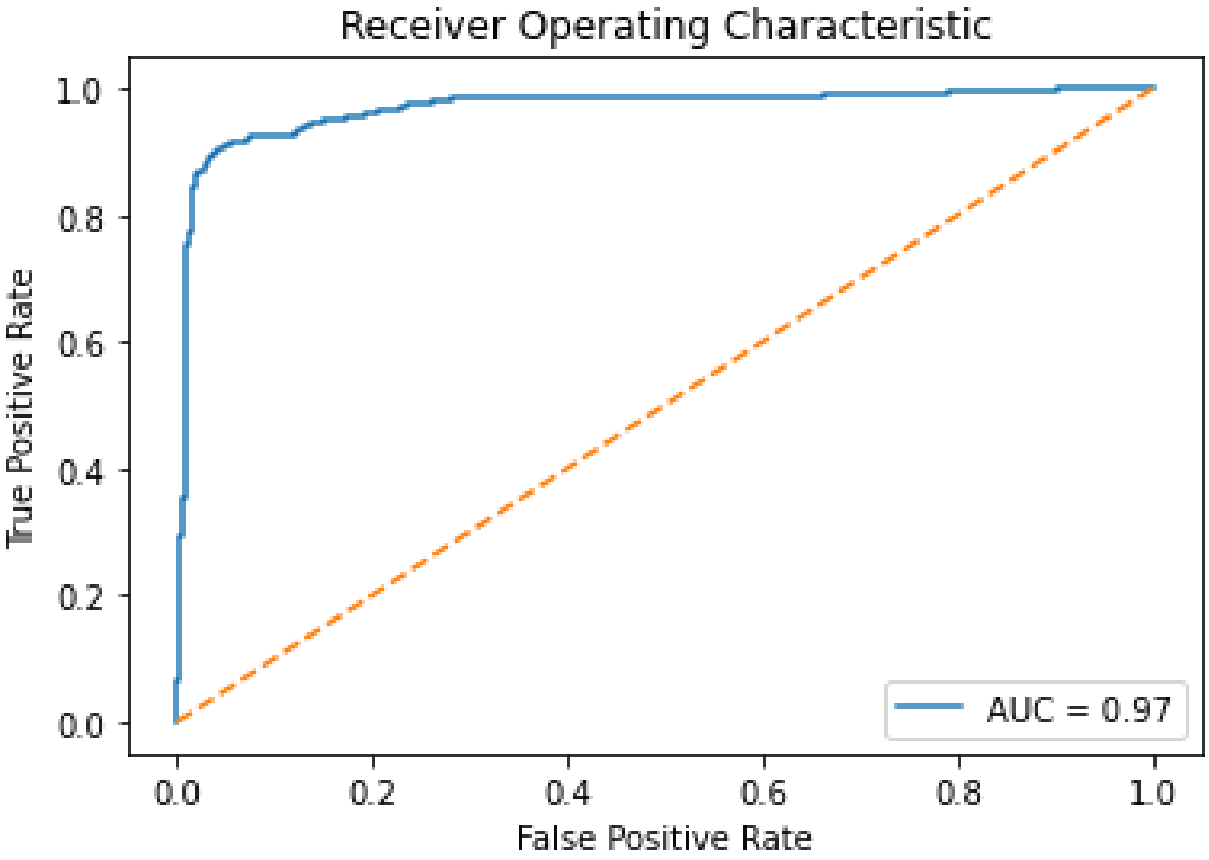

Figure 7: The RoC-SVM-RBF-Kernel

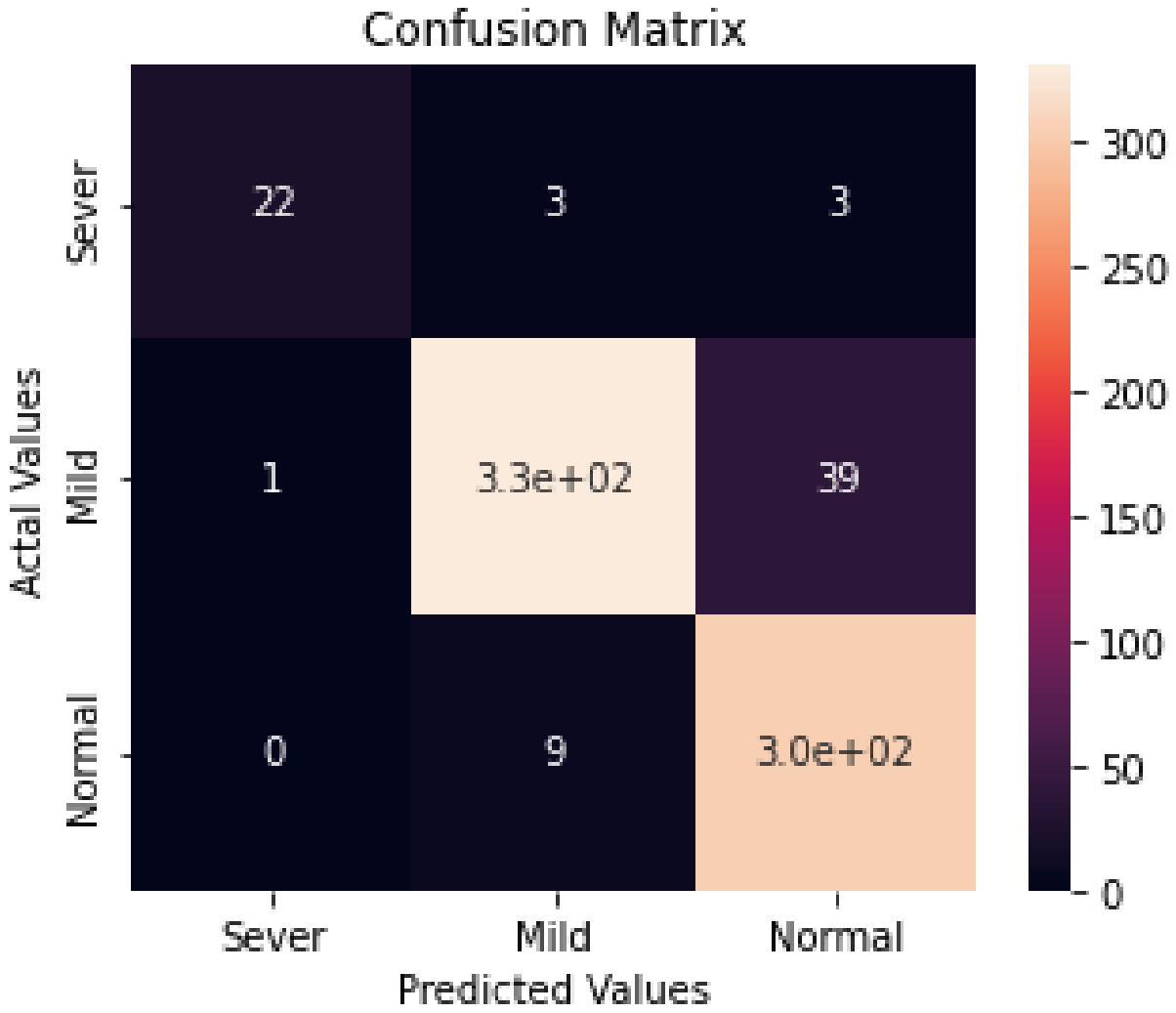

Figure 8: Confusion Matrix RBF

3 References

1. Link https://scikit-learn.org/stable/modules/svm.html
2. Link https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
3. Link https://medium.com/@pranjallk1995/pca-for-image-reconstruction-from-scratch-cf4a787c
4. Link https://rpubs.com/Sharon_1684/454441
5. Link https://scikit-learn.org/stable/modules/model_evaluation.html
6. Link https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-
7. Linl https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/
8. Link https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

All the links accessed on 29-11-2022