Programming Assignment-2

SNEHA G.P.S (EP19B032), M.K. Vishal Rishi (CH18b013)

April 7, 2022

1 Experimental results:

1.1 Synthetic-Data :

1.1.1 K-Means :

We apply K-Means to the data , to find optimal number of clusters we use Elbow method . We plot distortion vs k (No. of Clusters) graph and pick the elbow of the curve as Optimal value

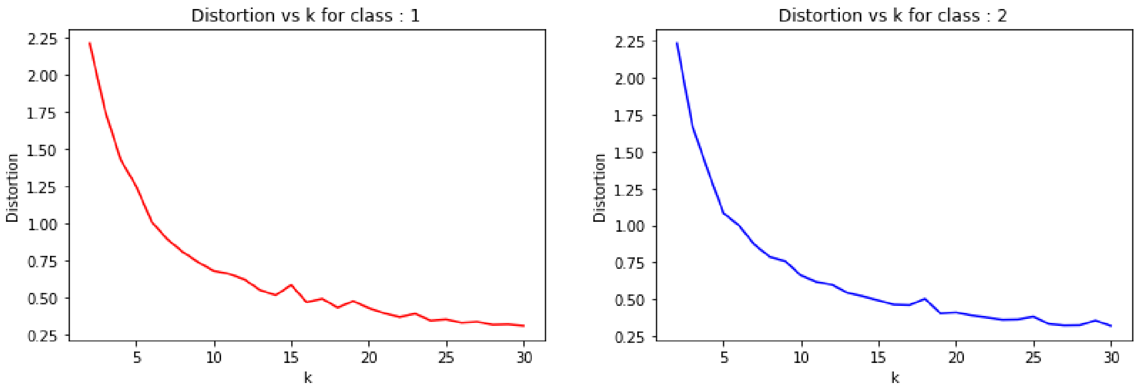


Figure 1: Distortion vs Number of Clusters

To find Optimal value of iteration we plot Distortion vs Iteration for k = 16

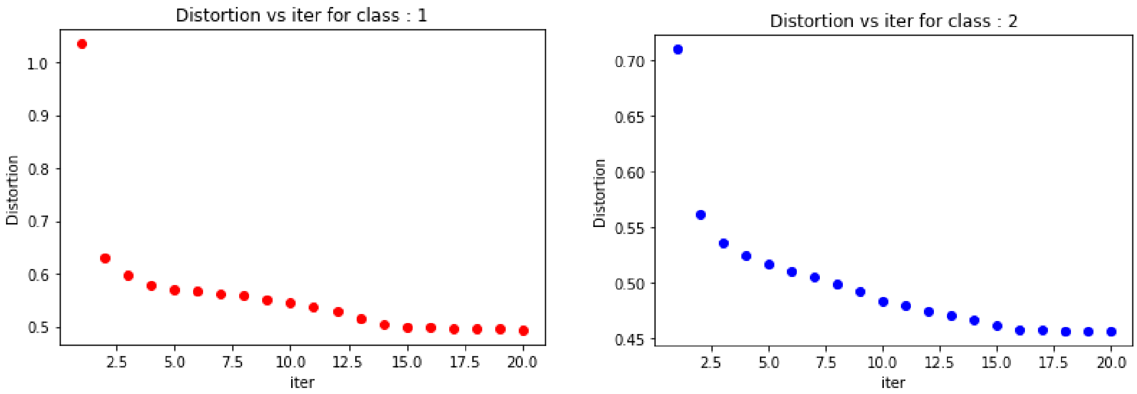
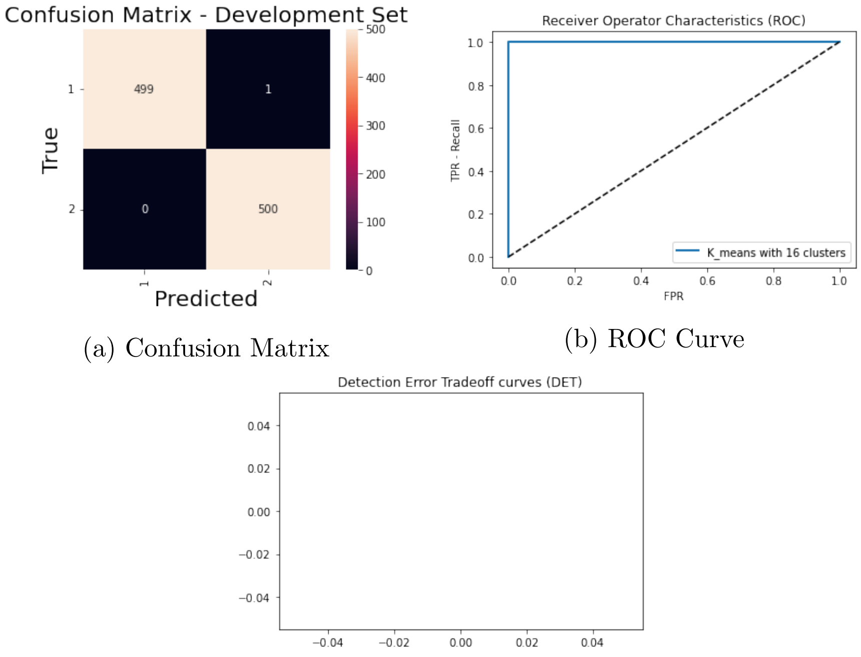


Figure 2: Distortion vs Number of Iteration

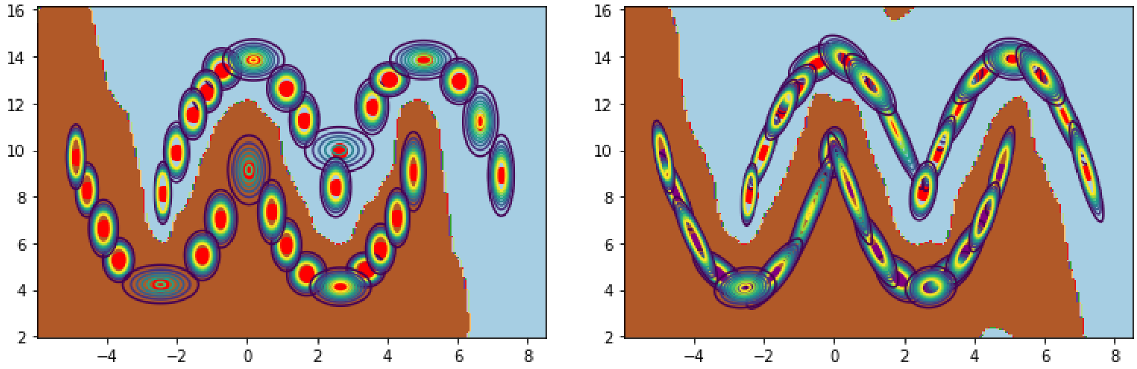


(a) Confusion Matrix

(b) ROC Curve

(c) DET Curve

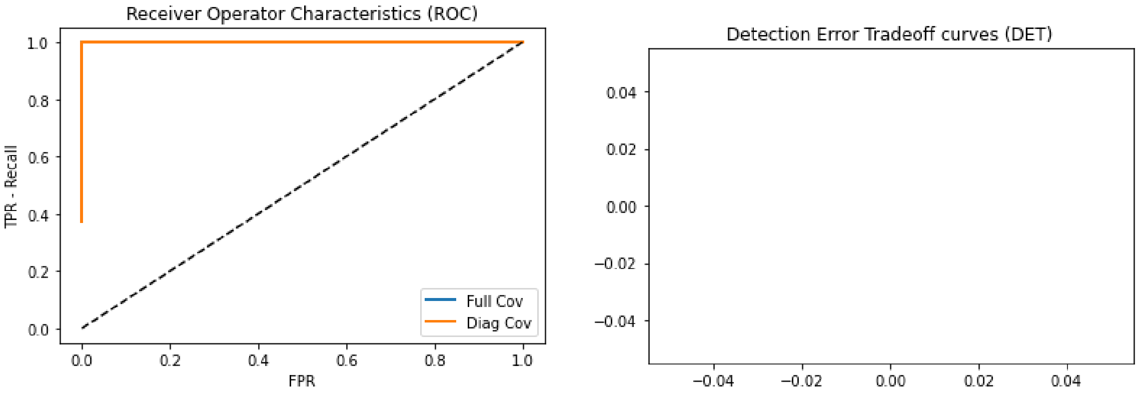
1.1.2 Gaussian Mixture Model :



(a) Diagonal Covariance

(b) Non Diagonal Covariance

Figure 4: Constant Density Curves



(a) ROC Curve

(b) DET Curve

Figure 5: ROC and DET Curves for Diagonal and Non-diagoanl Covariance

Here we are getting good classification results for both diagonal and non diagonal covariance . As we expect the constant density curve for diagonal covariance are parallel to axes and non diagonal are not parallel to axis .From k-means we get optimal number of iteration to be 16 (using elbow method) . We initialize GMM using centroids obtained from k-means so we run 10 iterations in GMM (can be reduced to 5 also).ROC and DET curves shows that both diagonal and non-diagonal covariance perform well in classifying the data .

1.2 Image

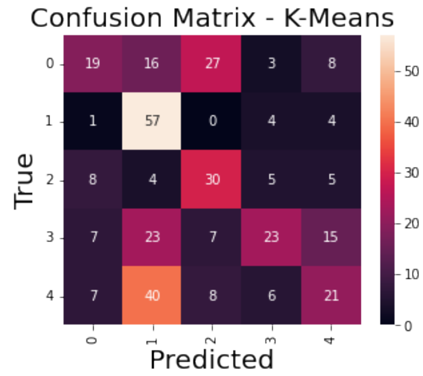
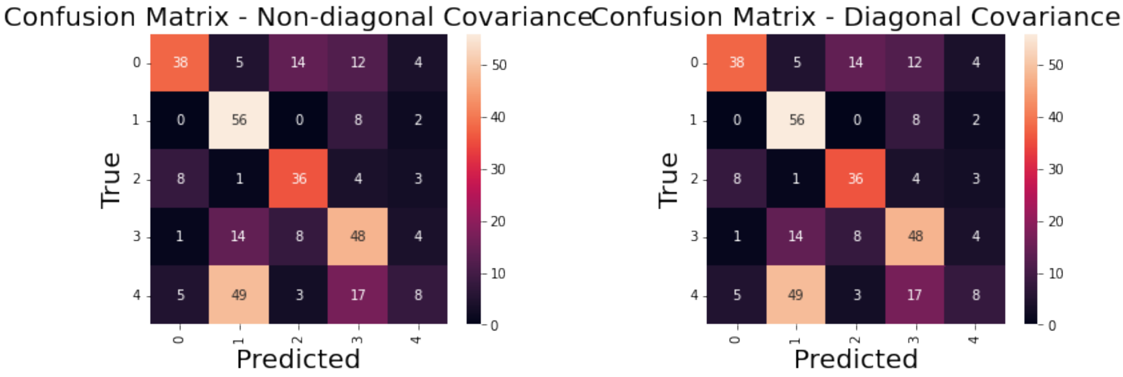


Figure 6: Confusion Matrix k-Means



(a) Non Diagonal

(b) Diagoanl

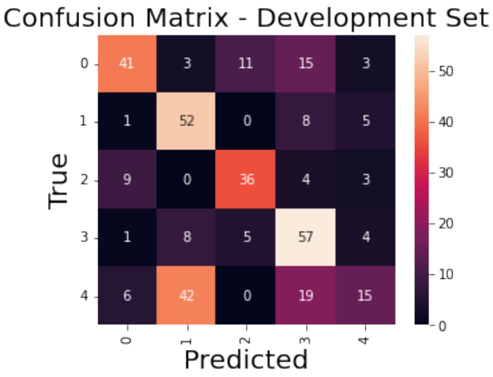


Figure 8: Confusion Matrix k=16

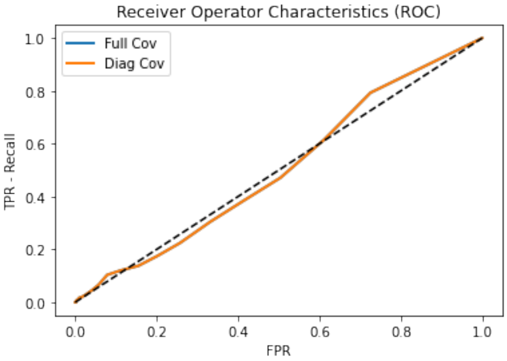


Figure 9: ROC Curve for Image

We see that confusion matrix for k = 16 gives good results for coast class (0) and for class 4 (opencountry) we are not getting good results because this image data set has a variety of images which vary from each other so increasing cluster for this class to higher value will help improve this classification .

DYNAMIC TIME WARPING:

Vector Quantization:

Vector Quantization is performed using K-Means set to **20 clusters**. All data points from all classes are combined and clustered using K-Means. In the given datasets, each data point is a two-dimensional matrix, with dimensions ***Nf x Nc*** (***Nf*** varies for each data point)***.*** With ***N*** being the number of data points from all classes, the data matrix being fed to K-Means is of dimensions ***(N x Nf) x Nc***. After clustering, each row of a data point is replaced with the cluster that is closest (in terms of euclidean distance) to that particular row. Hence, a quantized data point has the dimension ***1 x Nf***.

Dynamic Time Warping:

The L1-norm is chosen as the cost function for performing DTW. For performing classification, the test data point undergoes vector quantization. The average DTW score is computed between each class and the quantized data point. The data point is assigned to the class that has the lowest average DTW score. We present the confusion matrices, ROC curves, and the DET curves (on the development data) after performing DTW classification on the *Isolated Spoken Digit dataset* and the *Online Handwritten-Character dataset*.

Results:

Isolated Spoken Digit dataset:

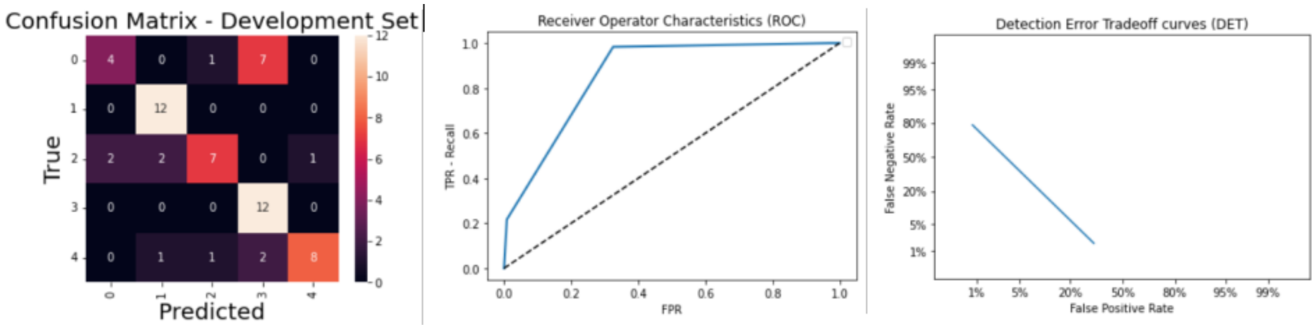


Figure 1: Confusion matrix (left), ROC curve (middle), DET curve (right) - Isolated Spoken Digit Dataset

Online Handwritten-Character dataset:

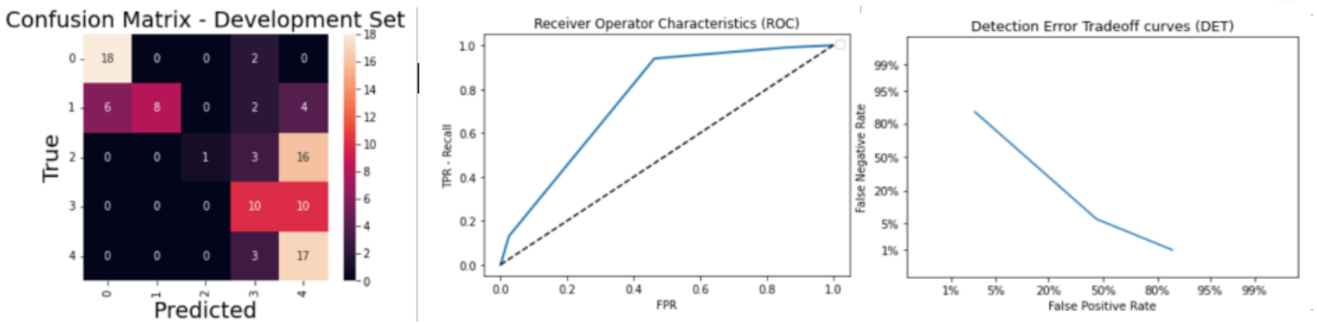


Figure 2: Confusion matrix (left), ROC curve (middle), DET curve (right) - Online Handwritten-Character Dataset

HIDDEN MARKOV MODEL:

Similar to DTW, we perform vector quantization for both the datasets. After performing vector quantization, we train a hidden markov model for each class, using the given code. During the testing phase, we compute the probability of the test data point from the trained hidden markov models. The data point is assigned to the class whose corresponding model gives the highest probability.

The parameters that define the hidden markov model are
a) The number of hidden states
b) The number of discrete symbols that are being observed. This is also equal to the number of clusters while performing vector quantization
We keep these parameters the same for all the classes. We provide results for various values of these parameters on the development datasets.

Results:
Isolated Spoken Digit dataset:
Number of states = 15
Number of symbols = 20

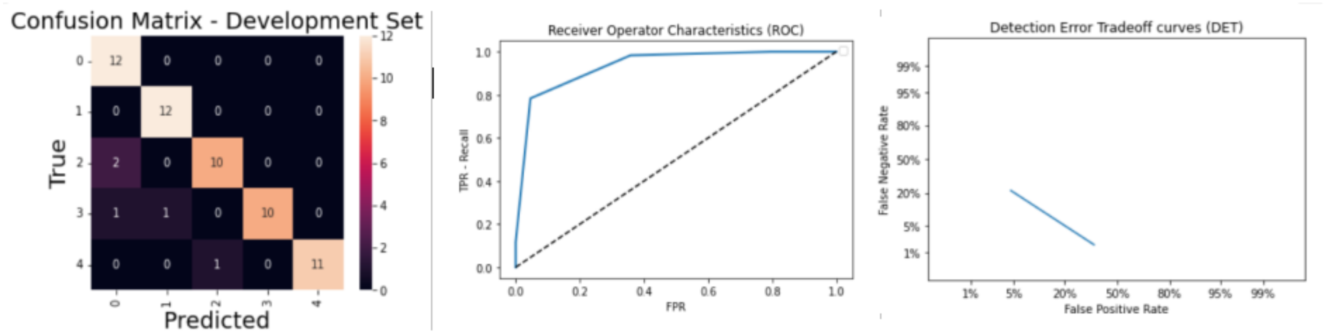


Figure 3: Confusion matrix (left), ROC curve (middle), DET curve (right) - Isolated Spoken Digit Dataset

Number of states = 5
Number of symbols = 20

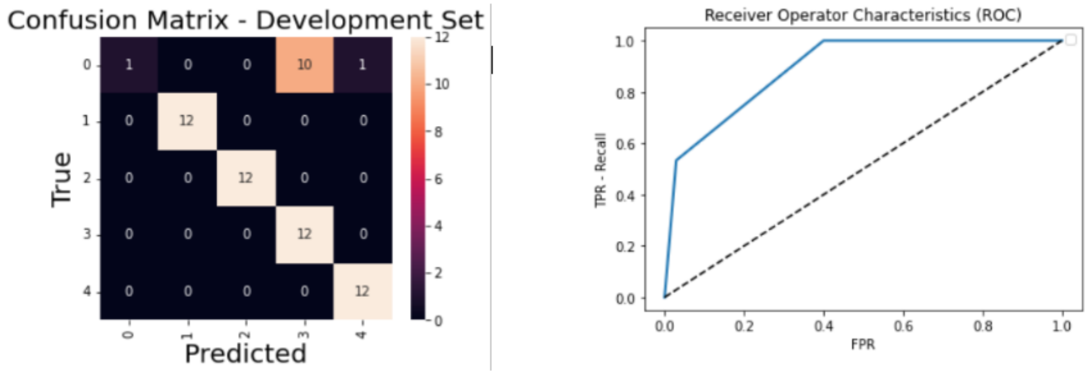


Figure 4: Confusion matrix (left), ROC curve (right) - Isolated Spoken Digit Dataset

Number of states = 25
Number of symbols = 20

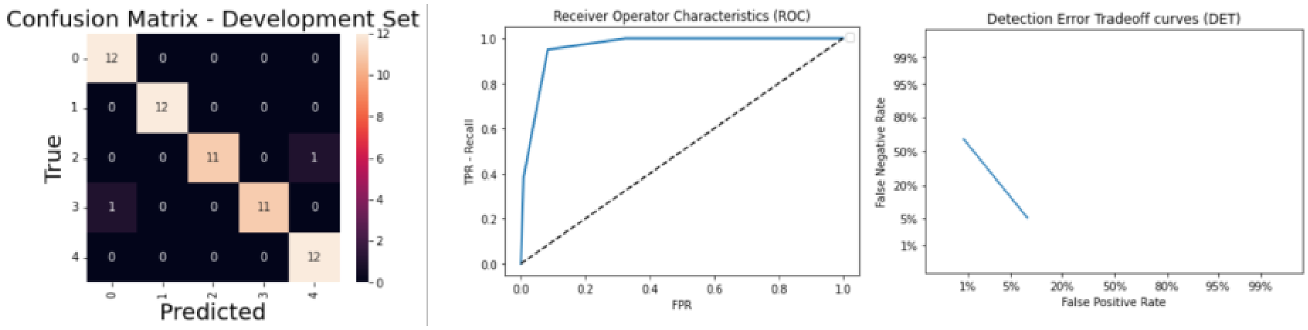


Figure 5: Confusion matrix (left), ROC curve (middle), DET curve (right) - Isolated Spoken Digit Dataset

Number of states = 15
Number of symbols = 30

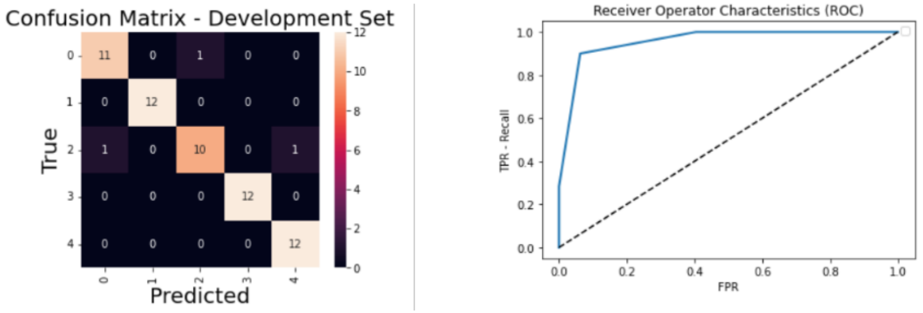


Figure 6: Confusion matrix (left), ROC curve (right) - Isolated Spoken Digit Dataset

Number of states = 5
Number of symbols = 30

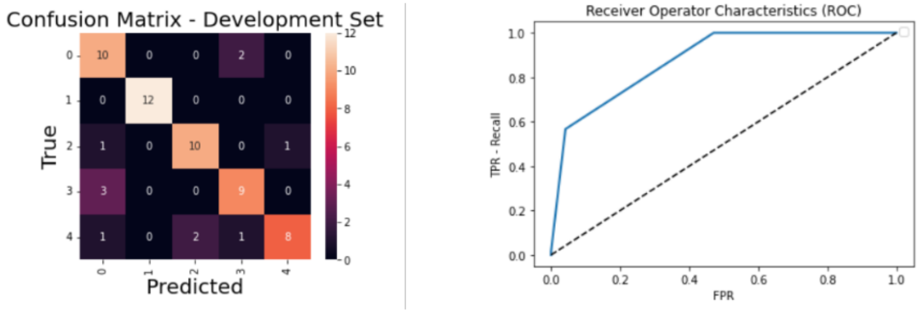


Figure 7: Confusion matrix (left), ROC curve (right) - Isolated Spoken Digit Dataset

Number of states = 25
Number of symbols = 30

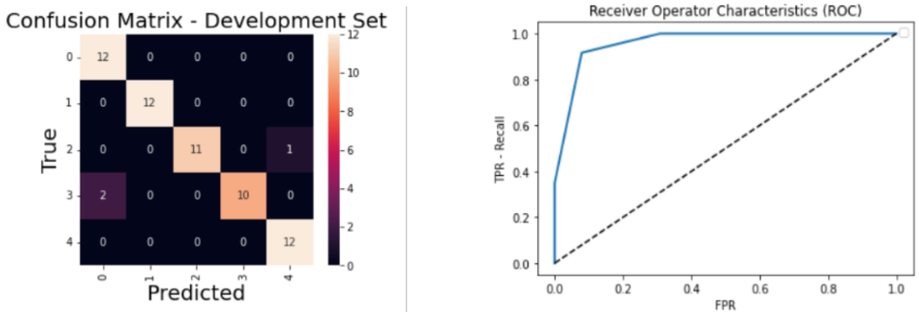


Figure 8: Confusion matrix (left), ROC curve (right) - Isolated Spoken Digit Dataset

Online Handwritten-Character dataset:
Number of states = 20
Number of symbols = 10

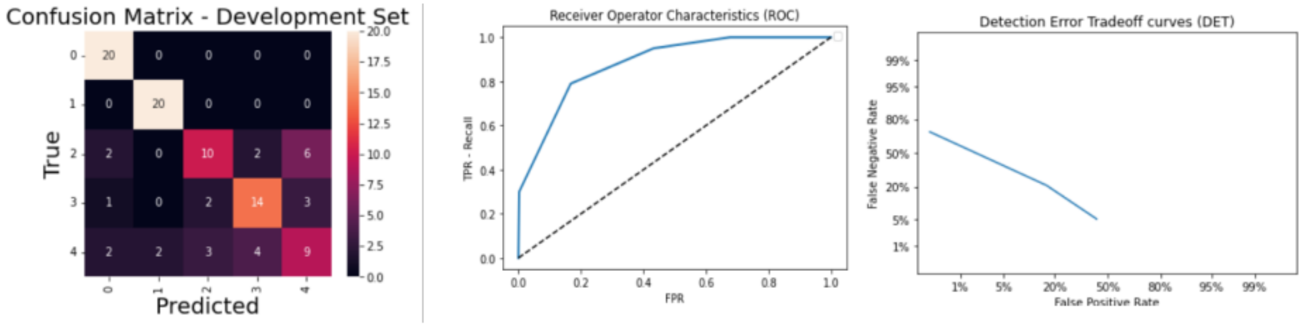


Figure 9: Confusion matrix (left), ROC curve (middle), DET curve (right) - Online Handwritten-Character Dataset

Number of states = 5
Number of symbols = 10

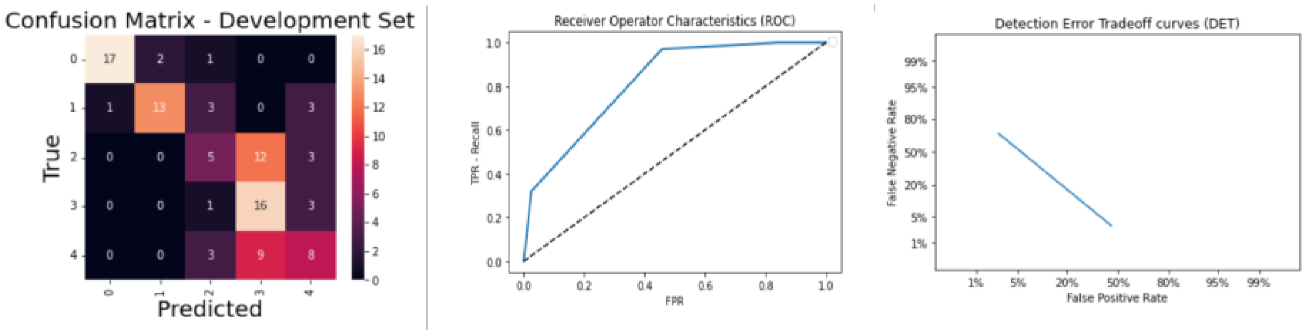


Figure 10: Confusion matrix (left), ROC curve (middle), DET curve (right) - Online Handwritten-Character Dataset

Number of states = 5
Number of symbols = 5

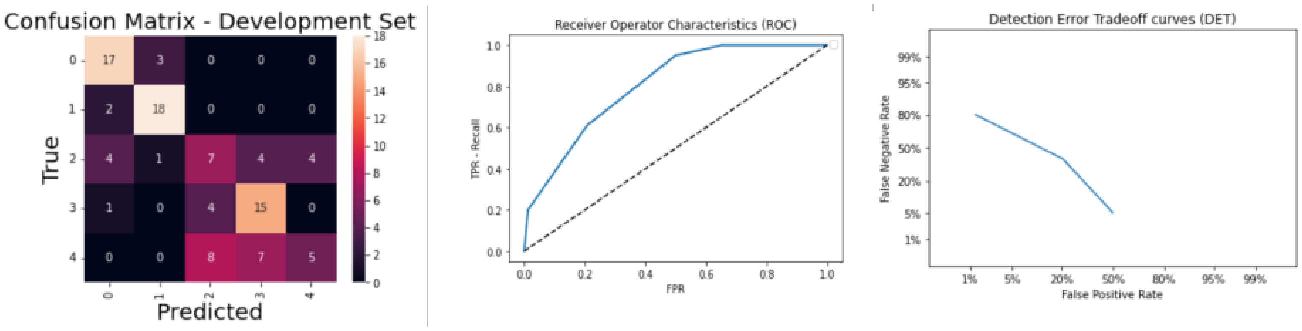


Figure 11: Confusion matrix (left), ROC curve (middle), DET curve (right) - Online Handwritten-Character Dataset

Number of states = 20
Number of symbols = 20

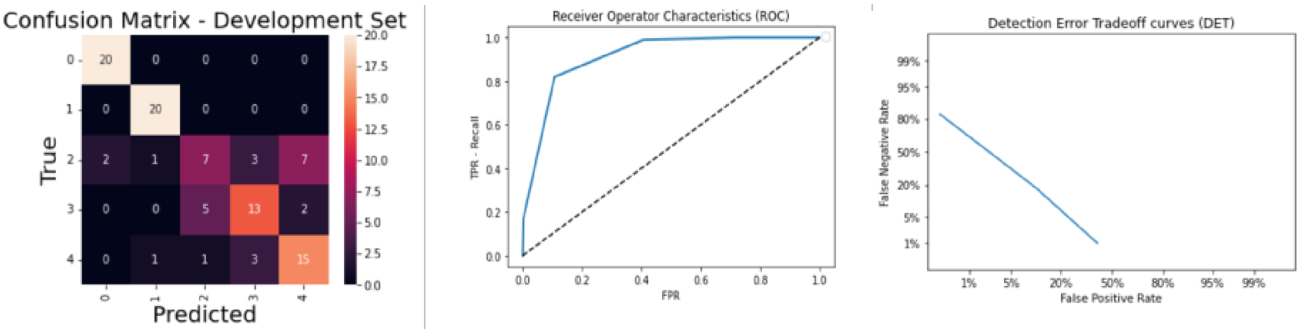


Figure 12: Confusion matrix (left), ROC curve (middle), DET curve (right) - Online Handwritten-Character Dataset

Number of states = 20
Number of symbols = 30

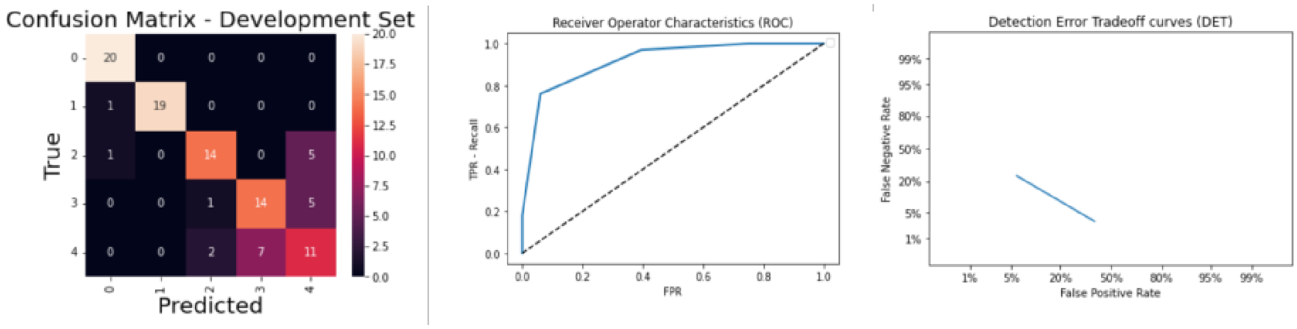


Figure 13: Confusion matrix (left), ROC curve (middle), DET curve (right) - Online Handwritten-Character Dataset

INFERENCE:

1) The HMM architecture that gives the best results in

Isolated Spoken Digit Dataset:

Number of states = 25
Number of symbols = 20

Online Handwritten-Character Dataset:

Number of states = 20
Number of symbols = 30

2) HMM is able to perform better in both the datasets compared to DTW

3) In the *Online Handwritten-Character Dataset,* we can see that the classification performance enhances when the capacity of the HMM is increased (when we increase the number of states and symbols)

4) In the *Isolated Spoken Digit Dataset,* the classification performance remains the same for all values of HMM parameters