CS5691 Assignment 3 Report - Team 14

Ajith Reddy - CS19B014, Aswin Ramesh - CS19B007

Kmeans and GMM:

1. Image Dataset:

   Current Approach : Each 6*6 pixel(each pixel - 23 dimensional feature vector) image belonging to a class is taken as 36 data points with 23 dimensional feature vector belonging to the class. The Kmeans and GMMs are run on the data points of each of 5 classes. Common covariance and mean across all the pixel blocks in this model

   Initial approach : The Kmeans and GMMs are computed individually for all the 36 blocks in each image(1st block seperate GMMs, 2nd block ... , 36th block). But with this approach there were many nan values (det of covariance matrix is becoming 0). Different covariance and mean across all the 36 pixel blocks in this model

   The accuracy of the model is around 60% for values of K from 5-15. The model's accuracy is almost the same for k = 2-15 (slightly decreasing with increasing the value of k). The confusion matrix for the model for k = 15 is shown in fig-1. The class numbering is (1 - coast,2 - forest,3 - highway,4 - mountain,5 - opencountry).

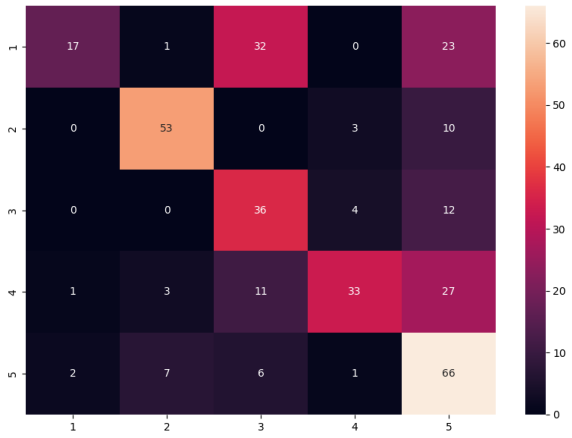

Fig 1.1 - Confusion matrix for k = 10 (Indexing on the y-axis is the actual class, On the x-axis is predicted by the model)

   The probability of the model predicting a coast correctly is very less whereas the prediction of forest and open country is a bit higher. The ROC and DET curves are shown in Fig 1.2 and Fig-1.3 respectively.

   The model almost the same across various values of K(slightly better for higher values of k(15) and lower values of k(2)).

2. Synthetic Dataset:

   For k-means clustering, the number of iterations of the EM algorithm was set to 10, and as expected, the accuracy was very high. The accuracy of k-means was different every iteration due to the randomness caused by the selection of the initial k centroids randomly.

   The scatter of the dataset is shown below in fig-1.4. The data points for each class are in the shape of 3.

   Below figure 1.5 shows the decision boundaries for few values of k(5,10,15).

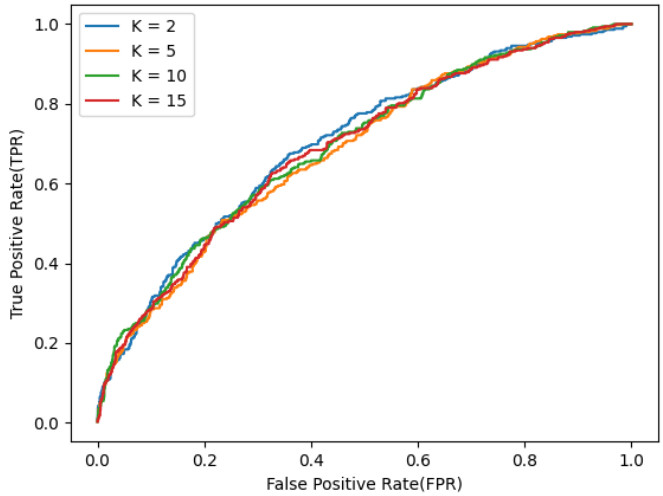

Fig 1.2 - ROC curve for various values of K.

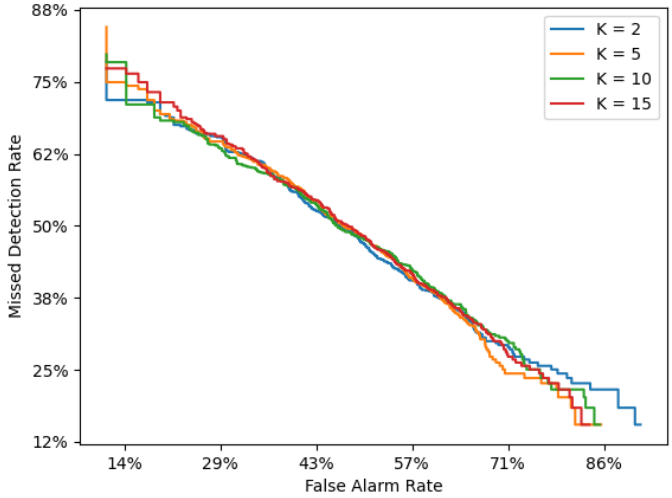

Fig 1.3 - DET curve for various values of K.

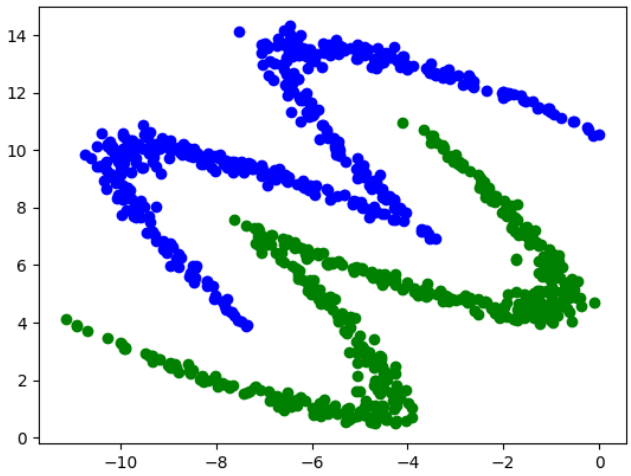

Fig 1.4 - Scatter of the datapoints blue(class 1), green(class 2).

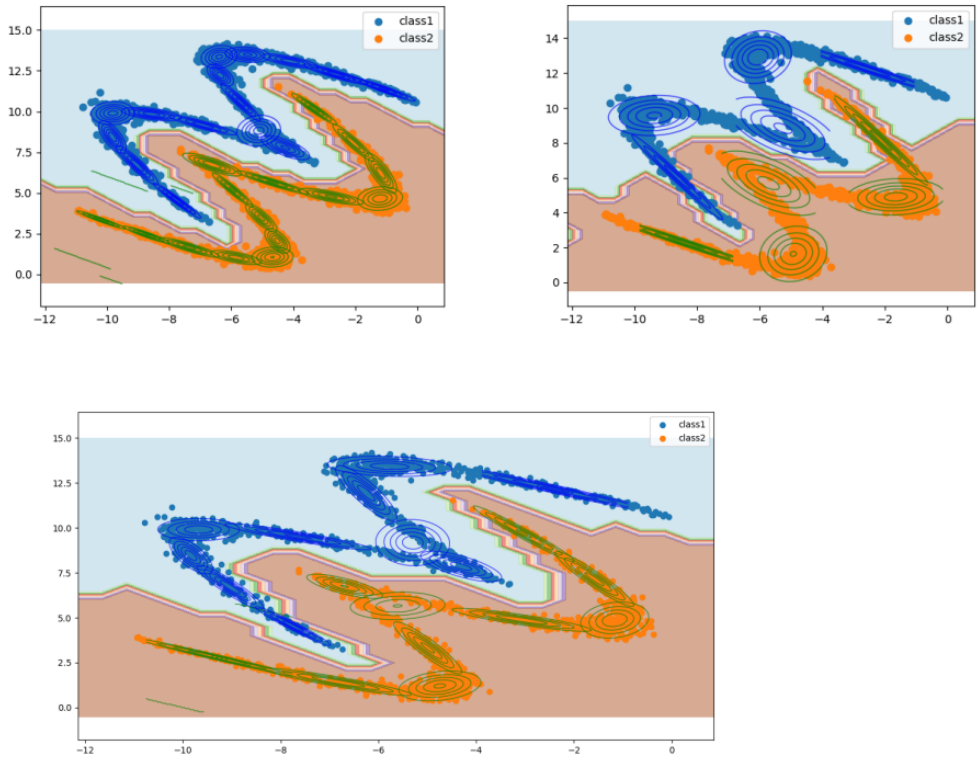

Fig 1.5 - Decision boundaries for k = 5(top right), k = 10(bottom), k = 15(top left).

The accuracy of Kmeans is >99% from k ≥ 5(It is almost >99.5% for many of the iterations. Since the shape of the datapoints is 3, it makes sense that from 5 clusters(5 edges in digit 3) the model is almost accurate.

The ROC curves for k = 2,3,5 are given in fig-1.6.

DTW and HMM:

1. DTW:

   - **Isolated Spoken-Digit dataset**:

     The accuracy of the model is high, it classified 59 correct and 1 wrong with k = 5(k being the consideration of minimum k distances for each class to classify). The given dataset it (1,3,4,8,o). The only wrong classification is of 4(the model predicted it as o), this is because the audio of the 4 has close resemblance to that of sound of o(f sound of 4 is very low in the audio) and hence the probable reason for that one wrong prediction.

     The model got better with increasing k(k=1 - 2 wrong predictions, k=5 - 1 wrong prediction). All the above details are after normalising the data.

     The confusion matrix, ROC and DET curves are shown in fig-2.1 and fig-2.2 respectively.

   - **Online Handwritten-Character dataset :**

     The accuracy of the model is 100%(After normalising the data). Before normalising it was slightly lesser(around 90%). There were not many things to experiment in this as the model was predicting perfectly well on the development dataset. Normalising the dataset gave better results.

     The confusion matrix(fig-2.3), ROC curve and DET curve(fig-2.4) are shown below.
     These DTW models are very accurate and can be used in real life also.

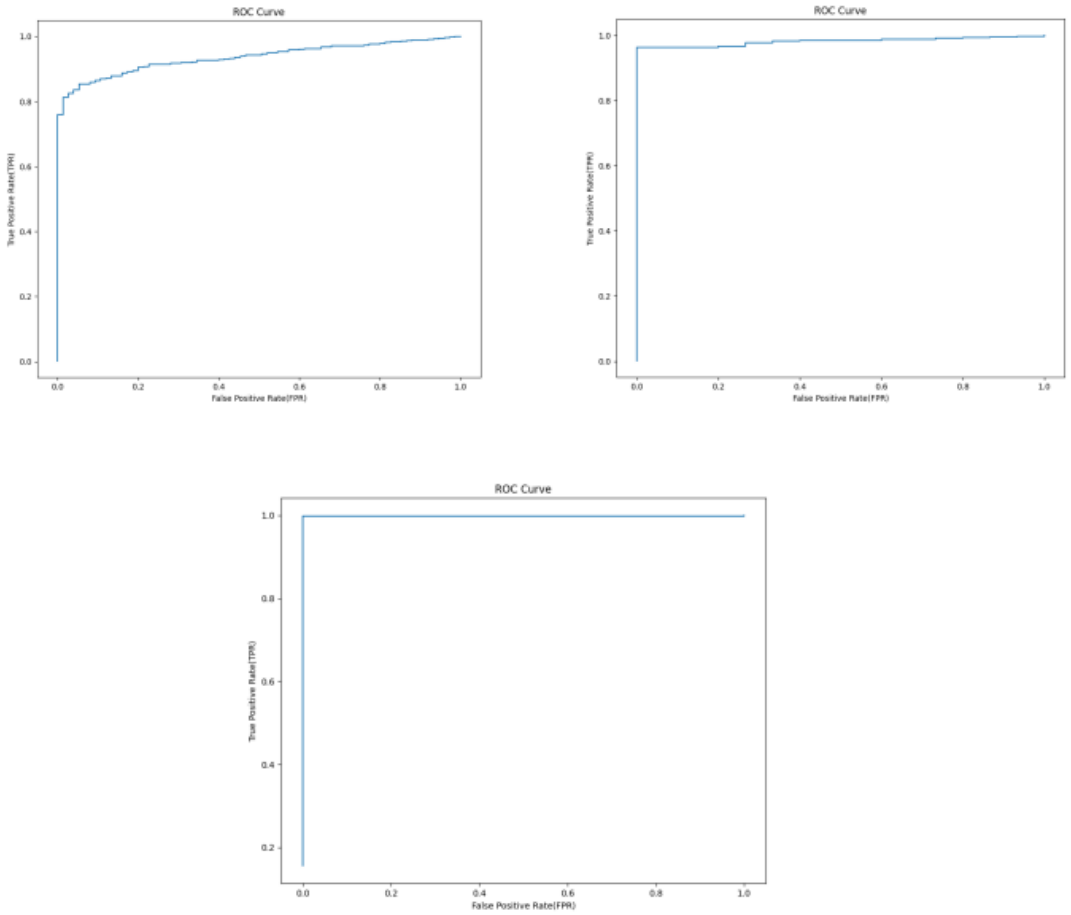

Fig 1.5 - ROC curves for k = 2(top left), k = 3(top right) and k = 5(bottom).

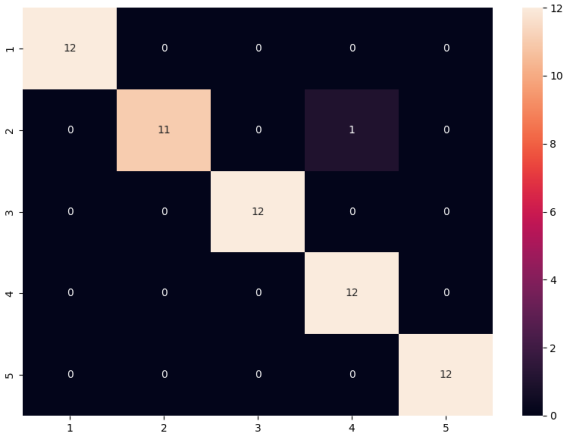

Fig 2.1 - Confusion matrix of digit dataset.

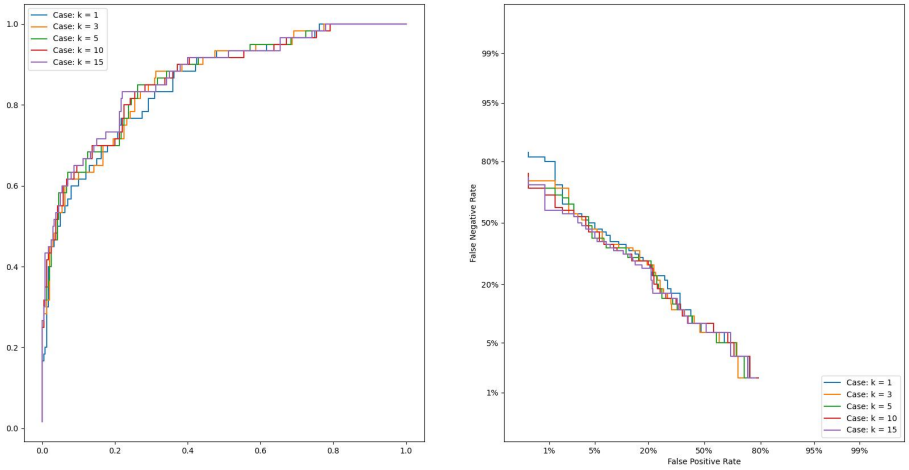

Fig 2.2 - ROC(left) and DET(right) curves of digit dataset(For different k).

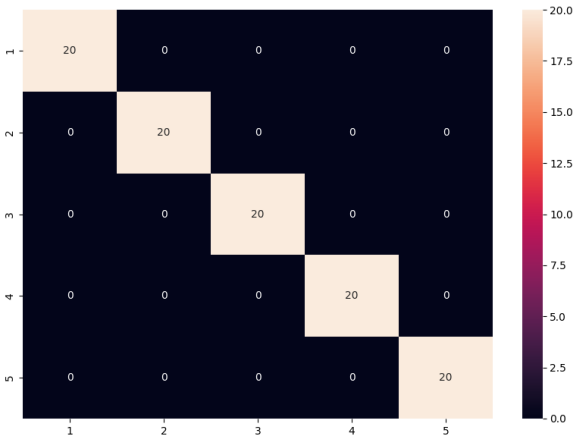

Fig 2.3 - Confusion matrix of telugu letter dataset.

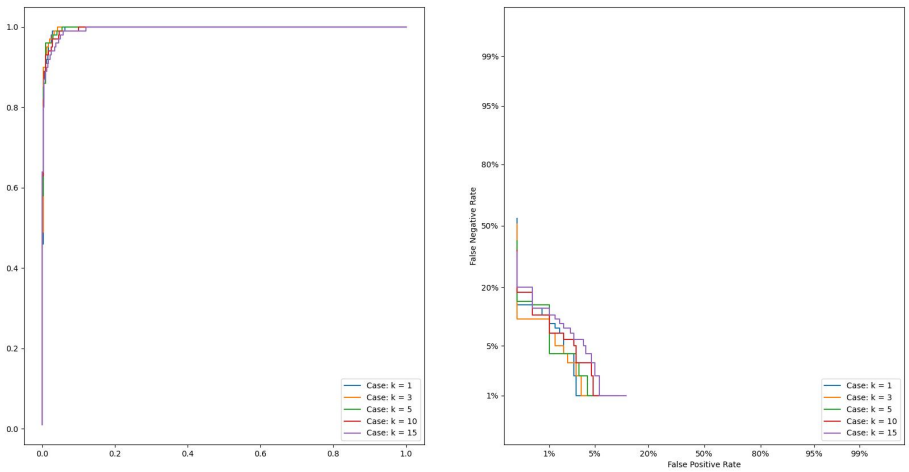

Fig 2.4 - ROC(left) and DET(right) curves of telugu letter dataset(For different k).

2. HMM:

- **Isolated Spoken-Digit dataset** :

  The highest accuracy achieved using various values of k(the number of clusters used for clustering the training data (*i.e*) the number of sequeces used to train the model) and num_states(the number of hidden states used in the markov models) was 100%. The values of k and num_states were 15 and 5 respectively, and sometimes even 12 and 5 respectively.

  The ROC(Fig 2.6), DET(Fig 2.7) and confusion matrix(Fig 2.5) graphs are as follows:

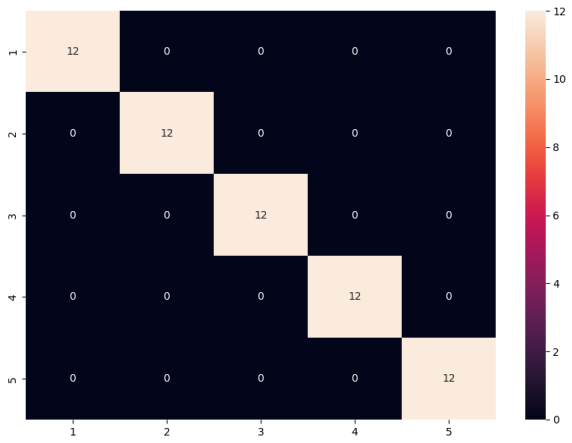

Fig 2.5 - Confusion matrix of Digits dataset for the highest accuracy

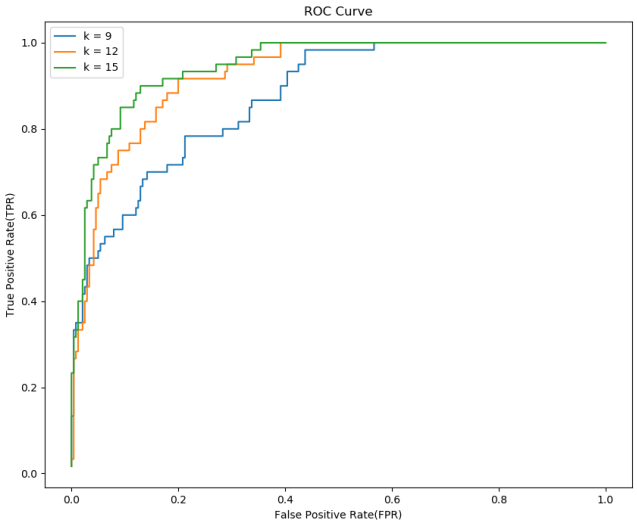

Fig 2.6 - ROC curve of Digits dataset for various values of k

The accuracy of the HMM models increased with increase in the value of k(checked till 15) and also with the increase in the value of num_states(checked till 5)

- **Online Handwritten-Character dataset** :

  The highest accuracy achieved using various values of k and num_states was 87%. Keeping the num_states fixed at 5, with variation in k from 9 to 15, there weren't any huge jumps in accuracy, but often the highest accuracy was found at k=12.

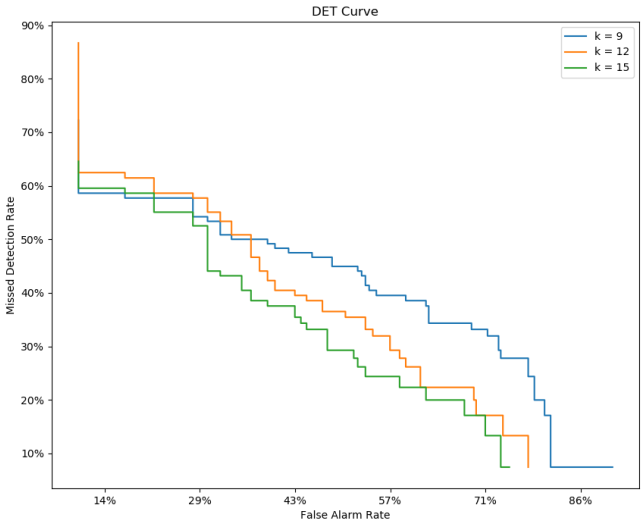

Fig 2.7 - DET curve of Digits dataset for various values of k

The ROC(Fig 2.9), DET(Fig 2.10) and confusion matrix(Fig 2.8) graphs are as follows:

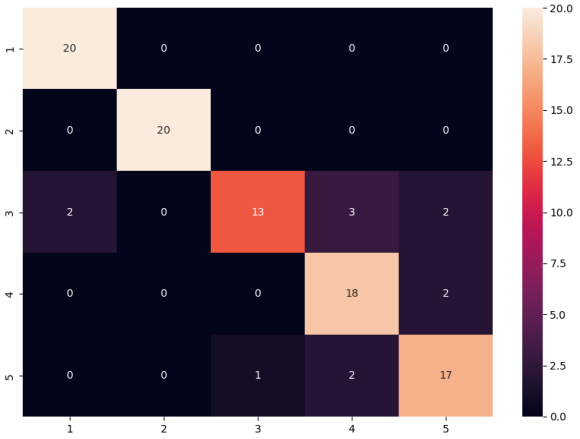

Fig 2.8 - Confusion matrix of Handwriting dataset for the highest accuracy

The accuracy of the HMM models increased with increase in the value of k, till around 12, and then decreased with increase in k(checked till 16). Similar to the models used for the digits dataset, the accuracy increased with the increase in the value of num_states(checked till 5)

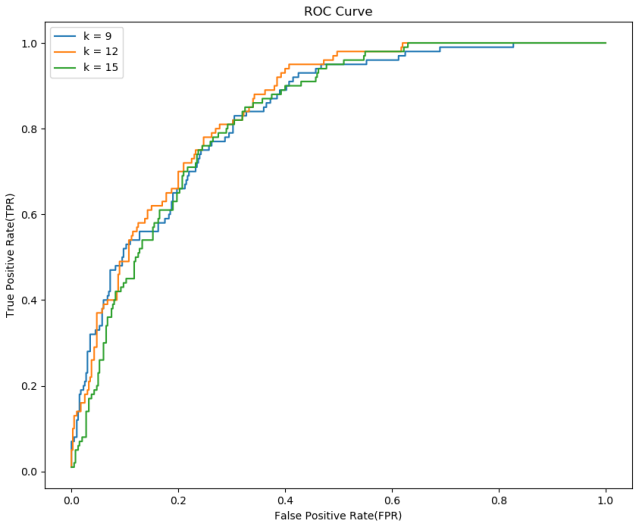

Fig 2.9 - ROC curve of Handwriting dataset for various values of k

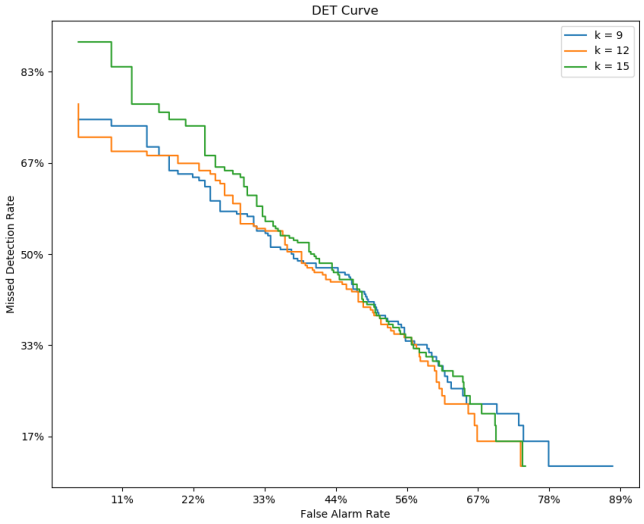

Fig 2.10 - DET curve of Handwriting dataset for various values of k