Politecnico di Torino

Machine learning for pattern recognition

Gender identification 2023

Student Name	Student ID
1. Yufei Ma	S300570
2. Jiaqi Wu	S296187

Submission Date : 21/09/2023

1 Introduction

This project is using the 12-dimensional vectors as a feature embedding. We firstly analyze the distribution of data, then different classifier will be implemented to distinguish the gender.

The training set has 2400 samples, 720 of them are male(label = 0), the rest 1680 are female(label = 1). So the dataset is biased and female account for most of it (70%). The test dataset is made up of 6000 samples, it contains 4200 males and 1800 females.So the distribution is opposite of the training data.

According to the requirement, the primary working point is ($\pi = 0.5$ C_fn=1 C_fp=1). So in this report, we only consider this situation and calculate accordingly the minDCF of each model

We will implement K - Fold validation method to choose best hyperparameter of each classifier and best model. we set K equal to 5.

In order to evaluate the performance of our models, we use detection minimal function as metrics to compare between classifiers. Once the best model is sorted out, the actual detection cost function is calculated and recalibration will be applied on this model as well to get more accuracy score‘.

2 Feature

We applied PCA and draw a graph of the first two principle components :

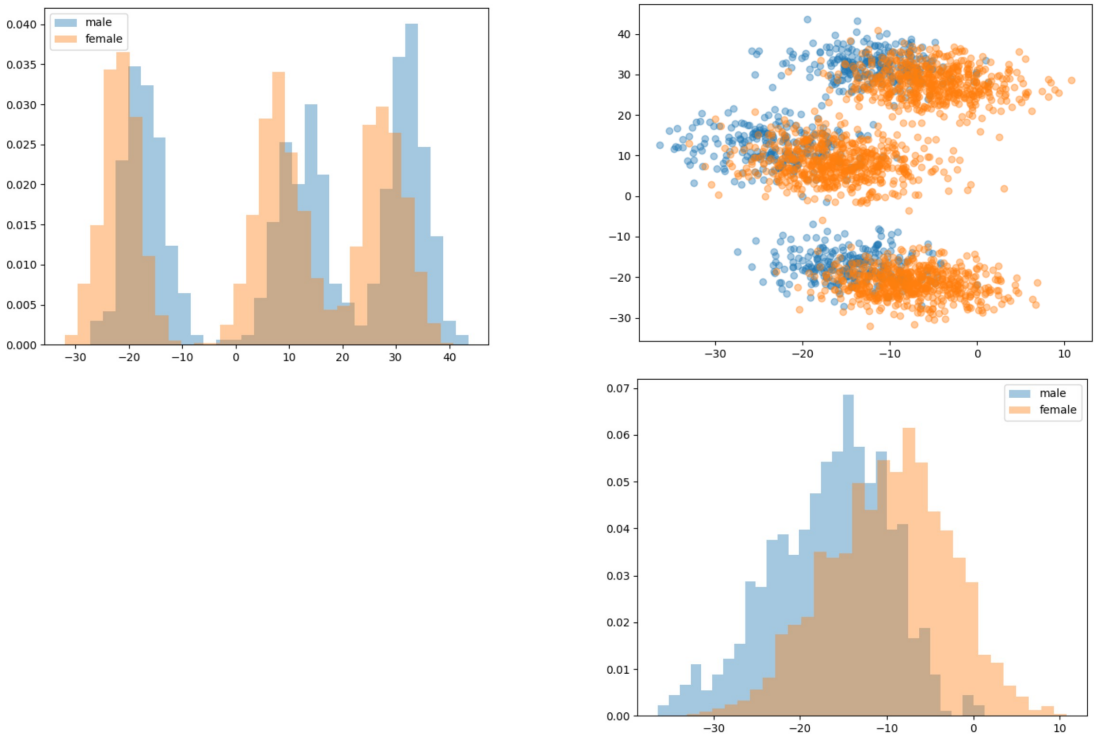


Figure 1: Histogram and 2D scatter plots of dataset features - principal components

From the top left graph which is the first principle component,it shows that GMM maybe shows better performance than a simple multivariate Gaussian model for dividing the gender.

Then we applied LDA to the data:

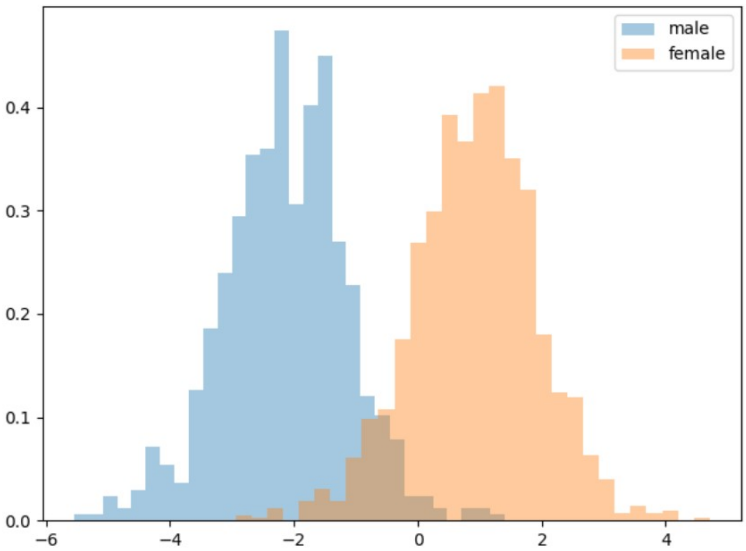


Figure 2: Histogram of dataset features - LDA direction

LDA shows that a linear classifier may have ability to discriminate the classes. LDA model has an ability to find maximum C-1 directions for our feature. In our project, we can find 1 dimension direction to classify the data since we have 2 type of label. However, regarding the features we observed in scatter plot, the GMM model with 3 components maybe will exhibit a better performance .

Now, we analyze the data from the aspects of the correlation between each attributes.

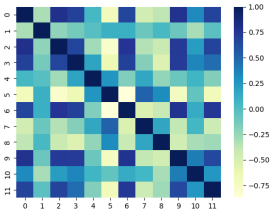


Figure 3: All training dataset

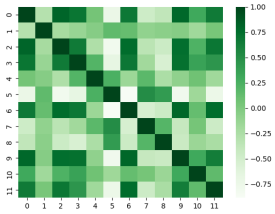


Figure 4: All Female training dataset

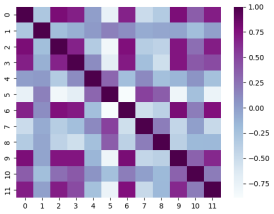


Figure 5: All Male training dataset

Dark color means high correlation between two features.

From the graph of all training dataset, it shows some features are significantly have large correlation with others no matter in which gender group. like feature 3 & 4, feature 7 & 4.

From the below analyze, it can be consider that the dimensional reduction with the assistance of PCA maybe useful to simple the calculation procedure . But it also contains the risk that losing the critical information during the reduction procedure.

Moreover, the variance graph is another useful way to evaluate how many dimension we want to retain.

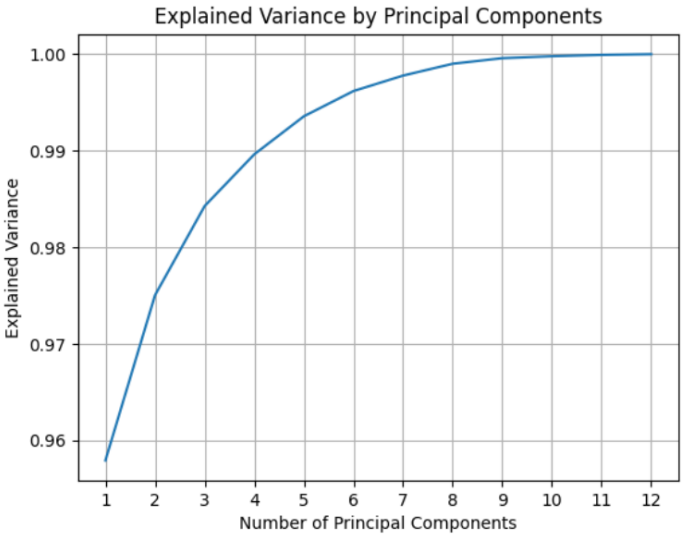


Figure 6: PCA − explained − variance

The graph shows when data keeps 10 dimension , 99% of the dataset variance is still remained. 96% will remains if we reduce the dimension into 9. 94% variance is left when the dimension jumps into 8 direction.

To start, we consider full dimension(12) till reduce to 8 dimension for PCA.

3 Classifiers for the task

We've employed a 5-fold protocol for K-fold cross-validation and evaluated performance using the metric of minimum costs (minDCF).Then, we will proceed to assess the actual DCF (actual C_prim)and perform score calibration.

3.1 Gaussian classifier

3.1.1 Full Covariance

We start with Gaussian classifiers. These classifiers encompass the full covariance (MVG), the diagonal covariance which assume covariance matrices Σ are diagonal, and tied assumption, which entails individual class means μ_c while computing the covariance matrix Σ over the entire dataset.

Firstly, we consider the full covariance, it means each class can have its own unique covariance matrix, which describes the relationships and variances among the features (variables) of the data. This approach allows for more flexibility in modeling complex data distributions but may require estimating a large number of parameters.

PCA	minDCF($\widetilde{\pi} = 0.5$)
-	0.114
11	0.124
10	0.166
9	0.190
8	0.193

Table 1: MVG classifier - minDCF

The results demonstrate that this model performs optimally when all features from the dataset are retained. As we progressively reduce the dimensionality, we observe a gradual increase in minDCF. Notably, when we reduce the dimension from 11 to 10, the minDCF sharply increase from 0.124 to 0.166. This observation strongly indicates that we may lost some important variables during the dimension reduction process.

3.1.2 Diagonal(Naive Bayes)

When we consider Naive Bayes , it assumes that all features (variables) are conditionally independent given the class label. When Naive Bayes is used with a diagonal covariance matrix, it means that the covariance matrix Σ is assumed to be diagonal, implying that there are no correlations between the features within each class. This simplification significantly reduces the number of parameters to estimate, making it computationally more efficient and less prone to overfitting when we have a small amount of data.

PCA	minDCF($\widetilde{\pi} = 0.5$)
-	0.463
12	0.119
11	0.123
10	0.168
9	0.195
8	0.198

Table 2: Diagonal MVG classifier - minDCF

We can see the results align with our observations of the dataset in previous chapter, which contains highly correlated attributes. Consequently, it's not surprising that the minDCF only reached 0.119, even without dimension reduction.

It's worth noting that this strong assumption, while reducing computational costs, the performance also get worse due to we neglect the correlation between attributes.

3.1.3 Tied

When we consider the MVG use Ties covariance materix, it assumes that all classes share a single common covariance matrix Σ computed over the entire dataset. In other words, instead of estimating a separate covariance matrix for

each class, we use one shared covariance matrix for all classes. This tied assumption can be useful in our case , because we only have limited data , so we don't want to have many parameters to be estimated , which may have the risk of overfitting. It essentially assumes that the covariance structure of the data is similar across all classes.

PCA	**minDCF($\widetilde{\pi} = 0.5$)**
-	0.114
11	0.118
10	0.162
9	0.186
8	0.189

Table 3: Tied MVG classifier - minDCF

We observe that even when significantly fewer parameters need to be estimated, the results keep as good as the Full covariance assumption, and this result reveal our male and female data have the similar covariance structure or pattern of relationships between their features (variables). We will use this as the optimal model, and we anticipate that this performance will extend to our consideration of GMM as well.

3.2 Logistic Regression

We now consider Logistic Regression models. It models the relationship between a dependent binary variable and one or more independent features by employing the logistic function. The logistic function transforms a linear combination of the input features into a probability score, bounded between 0 and 1, representing the likelihood of a data point belonging to a particular class. We use the raw data as input firstly

lambda	**minDCF($\widetilde{\pi} = 0.5$)**
1e-06	0.118
1e-05	0.118
1e-04	0.118
1e-03	0.118
0.01	0.118
0.1	0.118
1	0.133
10	0.219

Table 4: Logic Regression classifier - minDCF

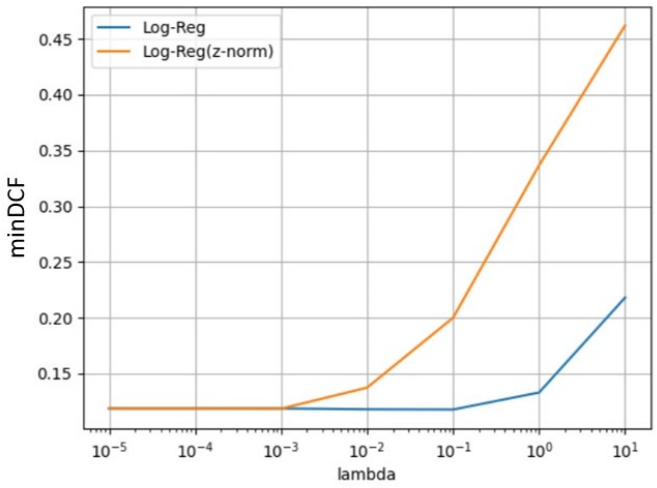


Figure 7: Logistic Regression with normed data

Surprisingly, although the input is raw data, the model's performance is not significantly worse . To further enhance our model's performance, we will explore the application of Principal Component Analysis (PCA). Since the z-normalization hasn't yielded favorable results, we apply PCA only to the original data.

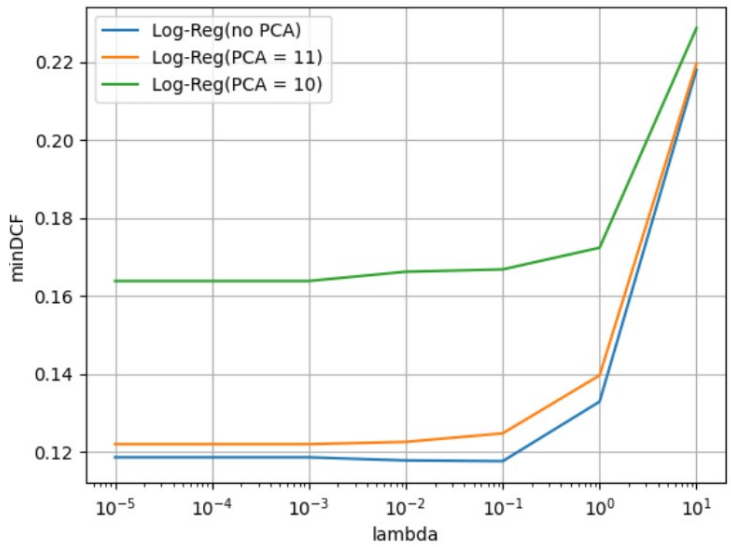


Figure 8: Logistic Regression with different PCA dimensions

It can be ssen that PCA has a limited impact on our model, and when the dimensionality increases to 10, there is a substantial cost increase from 0.12 to 0.16. Till now, the top-performing model is the one using the MVG with a Tied covariance matrix.

3.3 Support Vector Machine

SVM is always used consider discriminative model, its key objective is to find a hyperplane that optimally separates data into different classes while maximiz-

ing the minimum margin between them. In this report, we explore linear SVM's principles firstly and its hyperparameter tuning, then we consider nonlinearly SVM to separate data, this can be achieved through the "kernel trick" that maps data into higher-dimensional spaces. The linear SVM result shows below:

C	**minDCF($\widetilde{\pi} = 0.5$)**
1e-05	1
1e-04	0.137
0.001	0.120
0.01	0.115
0.1	0.115
1	0.115
10	0.0.115

Table 5: linear SVM classifier - minDCF

Next, we explore kernel SVM, beginning with polynomial kernels, and we exclusively employ the original dimensionality of the data without PCA.

For Poly kernel, the result shows below:
d =2 c = 0

C	**minDCF($\widetilde{\pi} = 0.5$)**
1e-05	1
1e-04	0.119
0.001	0.131
0.01	0.127
0.1	0.115
1	0.119

Table 6: Polymonimal SVM classifier - minDCF

We can found the result keeps the same as the linear one when C is equal to 0.1. Due to the hugh increased of the computationl, RBF also is considered k = 0.

logγ	**minDCF($\widetilde{\pi} = 0.5$)**
-3	0.2
-4	0.228
-5	0.306

Table 7: RBF SVM classifier - minDCF

It is surprised that in non linear SVM ,the utilization of the Polynomial (Poly) kernel yields results identical to the linear SVM model, and when the Radial Basis

Function (RBF) kernel is introduced, performance deteriorates. It suggests that for the data in this project, linear SVM is sufficient .

3.4 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) is a approach that represents data as a mixture of Gaussian distributions. This flexible approach allows GMMs to capture complex data patterns. In our project, we assume male and female are represented by several Gaussian distribution, the possible components candidate number would be [1,2,4,8,16]. we tried different combination. Because a dimension reduction doesn't bring a good result. we will only try original data

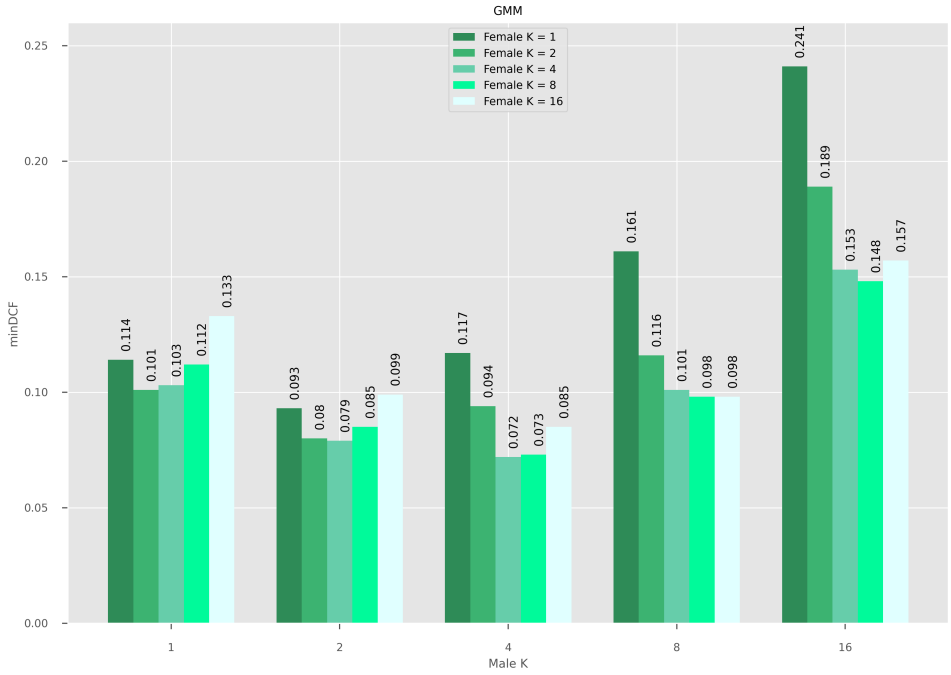


Figure 9: Gaussian Mixture Models with different components set

Since we found Tied Gaussian has a good performed in MVG models, so we will mainly consider Full Covariance(FC) and Ties(T) models. And from the above graph, we notice that the candidate components number would be 2 and 4 in male data. 4 and 8 in female data. so we tried different combination to find a better solution

Male K	**minDCF(Female K**	**minDCF($\widetilde{\pi} = 0.5$)**
2(T)	4(FC)	0.103
2(T)	8(T)	0.101
2(D-T)	4(FC)	0.221
2(D-T)	8(FC)	0.223
2(FC)	4(FC)	0.079
2(FC)	4(T)	0.083
2(FC)	8(FC)	0.085
2(FC)	8(T)	0.083
4(D-T)	4(FC)	0.126
4(D-T)	8(T)	0.127
4(T)	4(FC)	0.073
4(T)	4(T)	0.070
4(T)	8(FC)	0.078
4(T)	8(T)	0.073
4(FC)	4(FC)	0.072
3(T)	4(T)	0.071

Table 8: linear SVM classifier - minDCF

Male K	Female K	minDCF($\widetilde{\pi} = 0.5$)
2(T)	4(FC)	0.103
2(T)	8(T)	0.101
2(D-T)	4(FC)	0.221
2(D-T)	8(FC)	0.223
2(FC)	4(FC)	0.079
2(FC)	4(T)	0.083
2(FC)	8(FC)	0.085
2(FC)	8(T)	0.083
male K =4		
4(D-T)	4(FC)	0.126
4(D-T)	8(T)	0.127
4(T)	4(FC)	0.073
4(T)	4(T)	0.070
4(T)	8(FC)	0.078
4(T)	8(T)	0.073
4(FC)	4(FC)	0.072
3(T)	4(T)	0.071

Table 9: GMM classifier - minDCF

It can be seen that if male and female data are both divided into 4 components, with Tied model . the minDCF is lowest which is a rather better solution. Tied model has this outstanding solution means all the features in our dataset have a similar distribution. On the other side, if diagonal applied, it represents worse result . It means features may depend on each other which is also align with out analysis in heatmap. Finally, we implement on our best model: 4(T)

for male and female, we get the minDCF which shows a little improved(0.067)

Finally, we selected 4(T) -4(T) as our best model in GMM

Summary, our best model are

Model	$\mathbf{minDCF}(\widetilde{\pi} = \mathbf{0.5})$	$\mathbf{actDCF}(\widetilde{\pi} = \mathbf{0.5})$
GMM PCA = 12 Female: 4T Tied Male 4T	0.070	0.073
Linear SVM C = 0.01 , PCA = 12 raw data	0.115	0.123
Poly SVM d = 2 C = 0, PCA = 12 raw data	0.115	0.198
GVM Tied + Diag , PCA = 12 raw data	0.113	0.120
Log-Reg $\gamma = 0.001$, PCA = 12 Z-norm data	0.118	0.127

Table 10: RBF SVM classifier - minDCF

We don't consider RBF since this task already show good result on the linear classification in current dimension. It is no necessary to expand it to the higher dimension which may contain the risk of over fitting and high computation cost.

4 Calibration and fusion

4.1 Calibration

DET graph is able to show the trade-off between false acceptance rate and false rejection rate for a binary classification system.we apply this graph to visualize our system's performance across a range of decision thresholds or operating points.

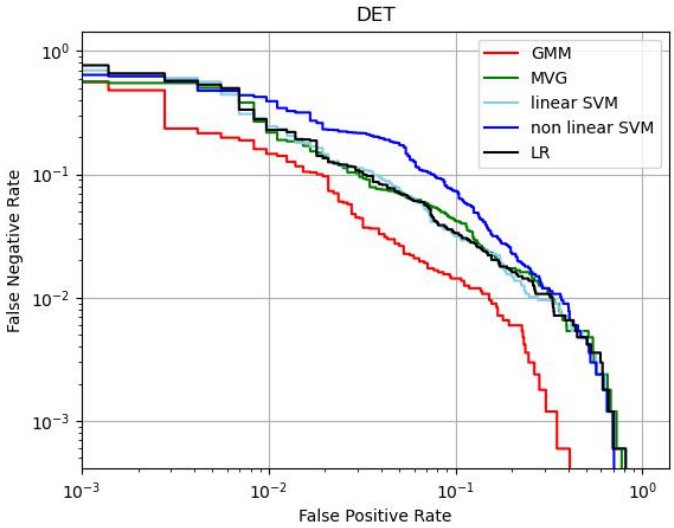


Figure 10: DET graph

It can be seen that GMM is still the best model since it has lowest FNR and FPR. No linear SVM shows modestly worse results, it means the linear model is already good enough to perform a classification job. The rest of the models have the similar behaviour.

We now consider calibration of there models and analyze if a score-level fusion could provide better improvement.

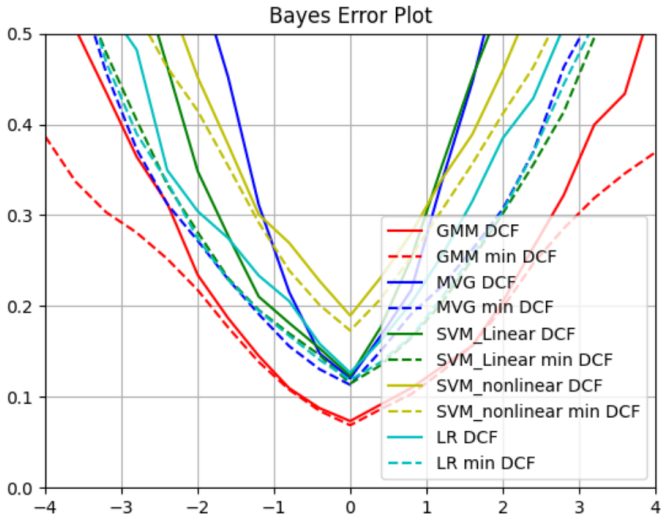
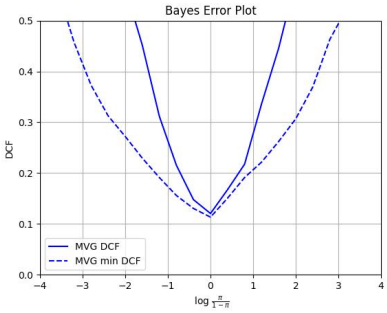
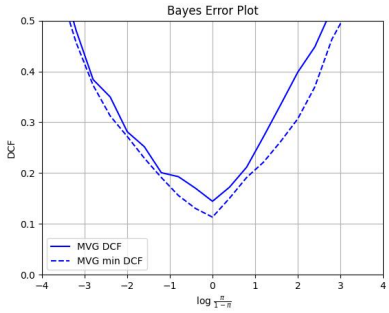


Figure 11: Bayes error plot for 5 best models - raw scores

From the graph, we can observe that the SVM linear model(light green line) and MVG(dark blue line) have a larger distance between actual and minDCF on each point. The recalibration will be applied in the next phase, we expect the distance will decrease which means models become more accurate due to a recalibration.



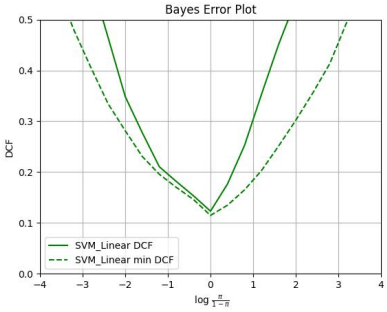
(a) MVG uncalibration



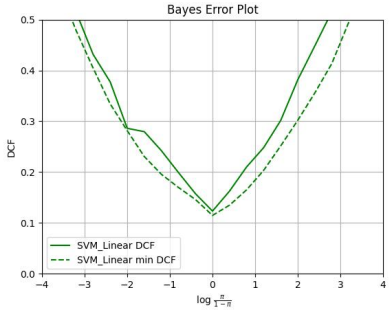
(b) MVG calibration

Figure 12: MVG calibration

The graph demonstrate that the distance between actual DCF and min DCF line become closer after recalibration. It means that the models benefit from the calibration process. Additionally, we also apply the procedure to SVM linear:



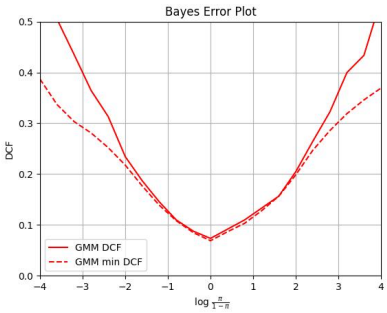
(a) SVM uncalibration



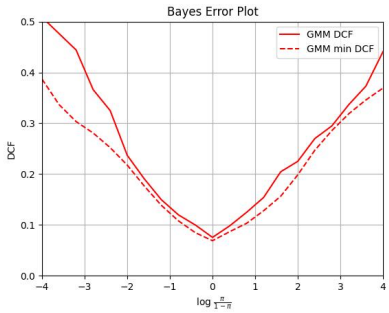
(b) SVM calibration

Figure 13: SVM calibration

The distance shows the similar behaviours as the MVG model. Finally,we applied calibration on the GMM model. From the result, it can be seen that the performance doesn't become more precise after the procedure. It is proved that the model is already accurate to perform the classficatio job.



(a) GMM uncalibration



(b) GMM calibration

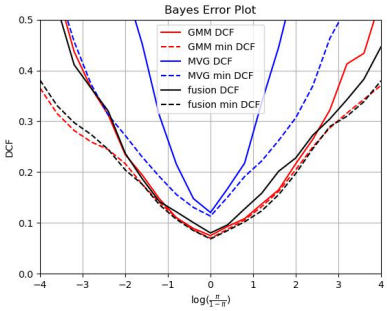
Figure 14: GMM calibration

From the analyze about the effect of calibration, we can notice that all three models show substantial improvements, with the MVG model particularly benefiting, despite a slightly larger gap between actDCF and minDCF compared to

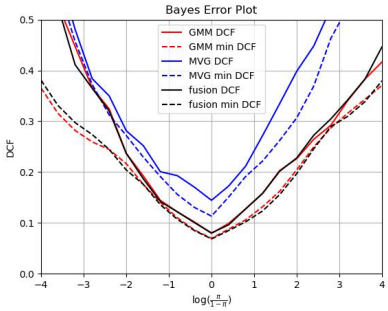
the other models. It is perhaps worth reflecting that calibration is strongly recommended when considering various operating points. However, when focusing on a specific range, such as our project where the operating point is set at 0.5, this procedure does not yield significant improvements

4.2 Fusion

Fusion is the procedure of combining multiple models to improve overall performance. It always be used because it harness the strengths of each individual model to produce a more accurate and robust classification. In this report, we try the following three combination : GMM + SVM; SVM + MVG and GMM + MVG and observe whether the whole performance improved.



(a) GMM+MVG uncalibration



(b) GMM+MVG calibration

Figure 15: GMM+MVG BayesErrorPlot

The black line is our Fusion result, we can observe that its minDCF and actDCF are lower than the single model, although the graph of GMM model is quite close to the fusion model. From the result in table, we notice fusion shows best minDCF among three models and GMM has the best/minmal actDCF value.

Model	MinDCF	actDCF
MVG	0.113	0.120
GMM	0.070	0.075
GMM+MVG	0.068	0.080

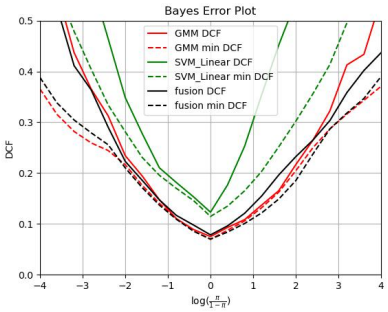
(a) UnCalibration

Model	MinDCF	actDCF
MVG	0.113	0.144
GMM	0.070	0.079
GMM+MVG	0.068	0.080

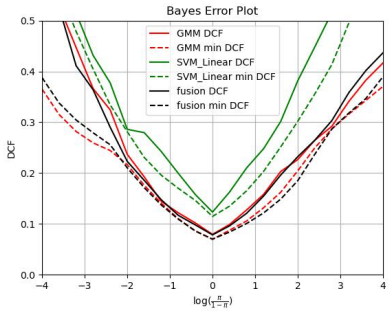
(b) Calibration

Table 11: minDCF and actDCF of single and Fusion model

When applimg GMM plus SVM model, although the fusion model doesn't achieve the best DCFs, we can still notice that in the region close to 0, the fusion model's DCF value is slightly lower than that of GMM.



(a) GMM+SVM uncalibration



(b) GMM+SVM calibration

Figure 16: GMM+SVM BayesErrorPlot

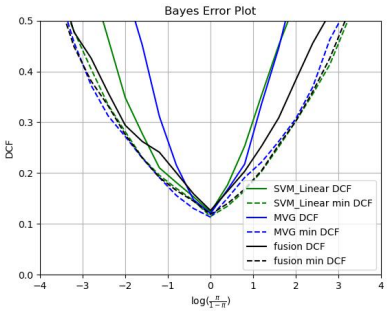
Model	MinDCF	actDCF
GMM	0.070	0.075
SVM	0.114	0.124
SVM+GMM	0.070	0.783

(a) UnCalibration

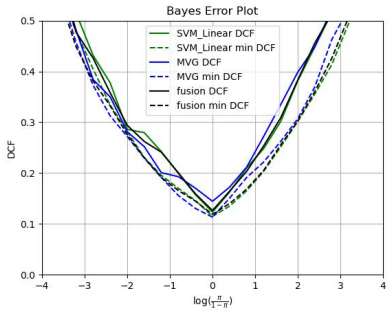
Model	MinDCF	actDCF
GMM	0.070	0.079
SVM	0.115	0.123
SVM+GMM	0.070	0.783

(b) Calibration

Table 12: minDCF and actDCF of single and Fusion model



(a) SVM+MVG uncalibration



(b) SVM+MVG calibration

Figure 17: SVM+MVG BayesErrorPlot

Model	MinDCF	actDCF
SVM	0.115	0.123
MVG	0.113	0.120
SVM+MVG	0.119	0.126

(a) UnCalibration

Model	MinDCF	actDCF
SVM	0.115	0.123
MVG	0.113	0.144
SVM+MVG	0.119	0.126

(b) Calibration

Table 13: minDCF and actDCF of single and Fusion model

The Linear SVM plus MVG exhibits slightly inferior performance when compared to other models. Additionally, the fusion model has only a little impact on our model's performance.

Based on the discussion below, we will restrict our evaluation to two fusion models: GMM+MVG and GMM+Linear SVM. We will also include the single model GMM in our assessment, as it demonstrates outstanding performance when

contrasted with the fusion models. Calibration and also uncalibration will be considered for the test data .

5 Experimental Evaluation

5.1 Model evaluation and its Bayes Error Plot

When we apply our model on evaluation dataset, what we need to do is to assess and measure the performance of the models rather than change any model's parameter. The minDCF and actDCF is still be used as a metric. And we will apply the evaluation set on the best models, they are GMM+MVGA fusion, GMM + Linear SVM and single GMM model. We will use uncalibration fristly.

We applied SVM + GMM firstly:

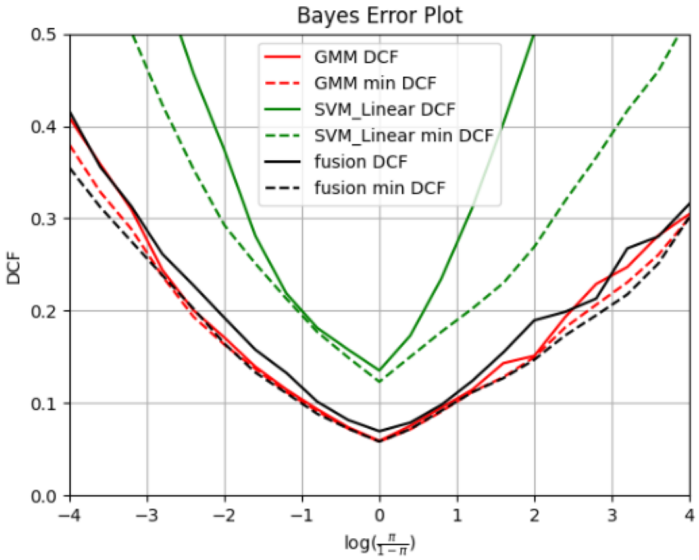


Figure 18: SVM + GMM Test DataSet without calibration

Model	minDCF	actDCF
GMM	0.058	0.059
SVM	0.123	0.135
GMM+SVM	0.058	0.070

Table 14: SVM GMM - DCF without calibration

Based on the results above, it can be observe that evaluation set yield a quite good result and the minDCF is even lower than the that of the training data.It is worth mention that GMM has better performance than Fusion model at the region which close to 0 area. It means if our working point is 0.5, GMM is good enough to tackle the classification gender job.

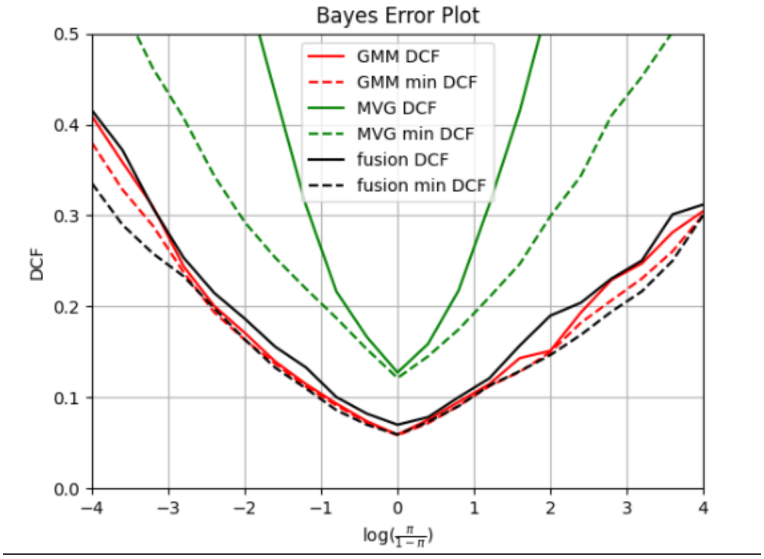


Figure 19: MVG + GMM Test DataSet without calibration

Model	minDCF	actDCF
GMM	0.058	0.059
MVG	0.121	0.127
GMM+MVG	0.059	0.070

Table 15: MVG GMM - DCF without calibration

When fusion is MVG with GMM. the perfromance is quite similar with the previous one.

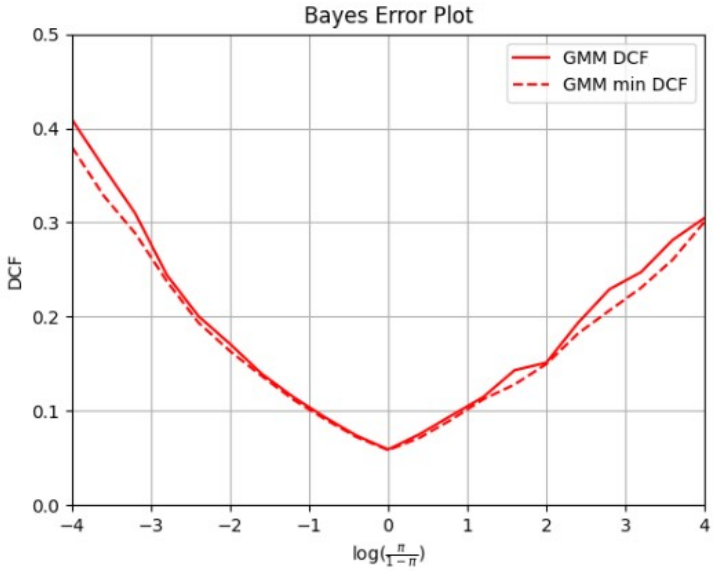


Figure 20: GMM Test DataSet without calibration

Model	minDCF	actDCF
GMM	0.058	0.059

Table 16: GMM - DCF without calibration

Now , we move to the calibration situation:

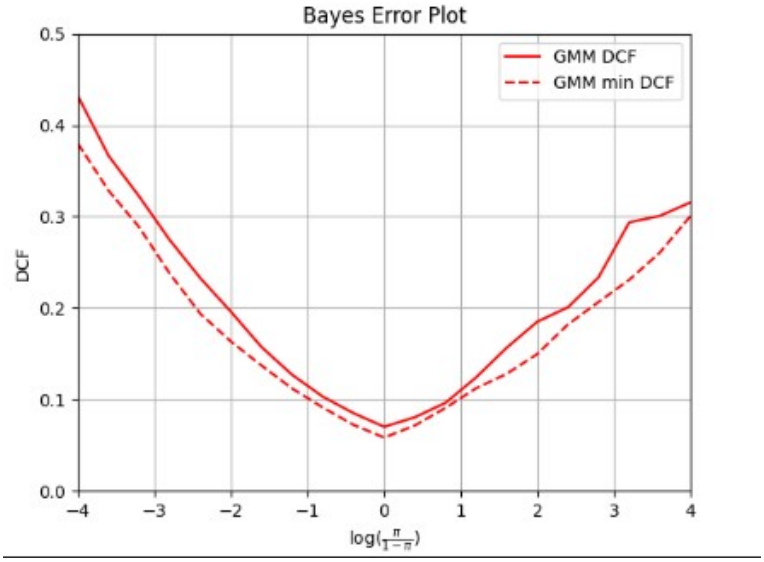


Figure 21: GMM Test DataSet with calibration

Model	minDCF	actDCF
GMM	0.058	0.070

Table 17: GMM - DCF

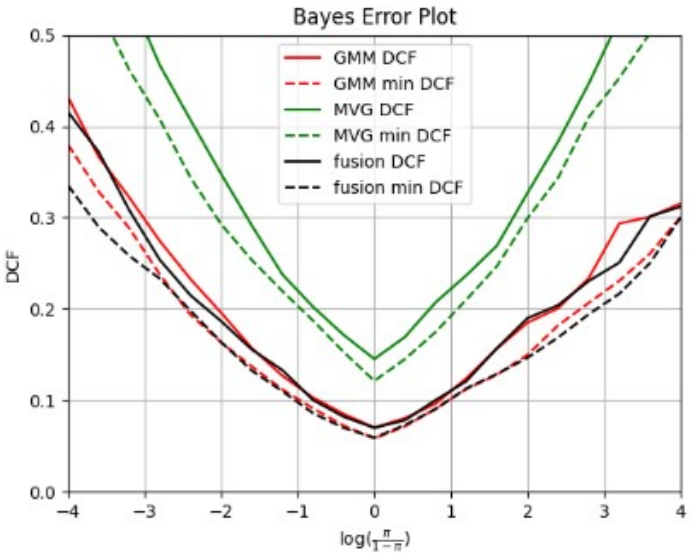


Figure 22: GMM MVG Test DataSet with calibration

Model	**minDCF**	**actDCF**
GMM	0.058	0.070
MVG	0.121	0.145
MVG+MVG	0.073	0.078

Table 18: GMM MVG - DCF with calibration

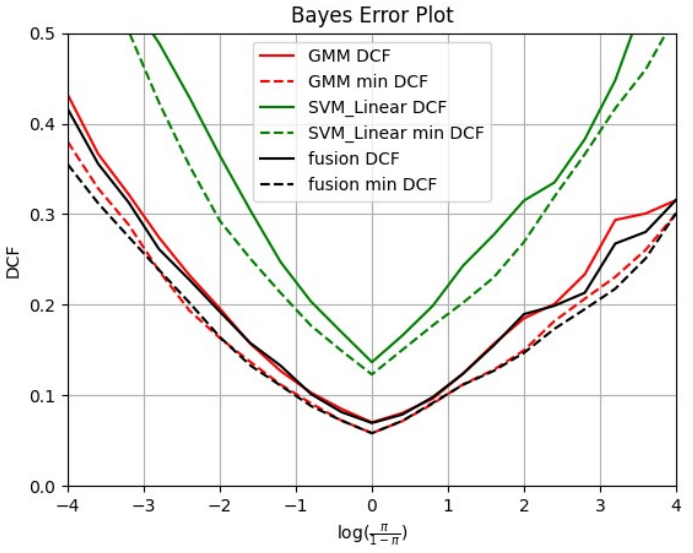


Figure 23: GMM SVM Test DataSet with calibration

Model	minDCF	actDCF
GMM	0.058	0.070
MVG	0.123	0.136
MVG+MVG	0.058	0.070

Table 19: GMM SVM - DCF with calibration

Through the comparison the minDCF between calibration and uncalibration model, we could found this method is much useful when we apply like linear SVM or MVG model. GMM show the similar result in both situation.

5.2 Gaussian Mixture Model on Test Data

Due to the good performance of GMM model, we emphasis is on evaluating GMM model behavior when applied to test data. As previously discussed, the input data will be 12 dimensions, aided by PCA. and no norm used since this procedure doesn't help a lot for the improvement of performance. And the X axis of following graph means the number of component for each class, y axis is the minDCF value.

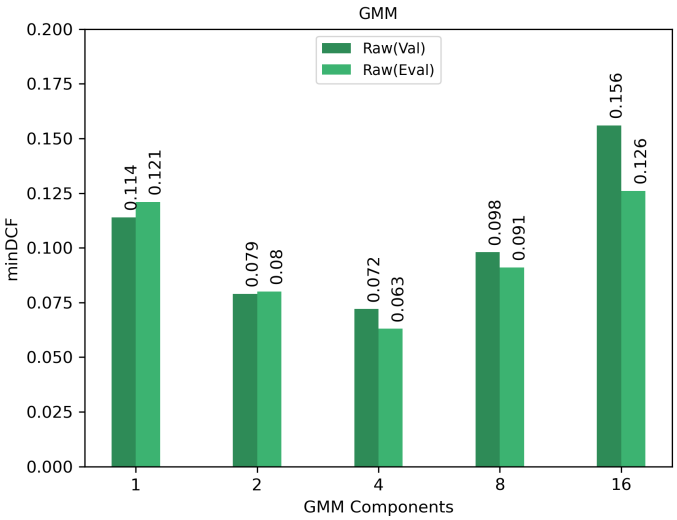


Figure 24: Full covariance assumption of GMM model on Test data

The graph reveals that, under the assumption of a full covariance matrix, the optimal performance is consistently achieved with four components for each class, resulting in minDCF scores of 0.063 for evaluation and 0.072 for validation data. However, when we increase the number of components (16 for each male and female), the cost sharply rises, leading to elevated minDCF values.

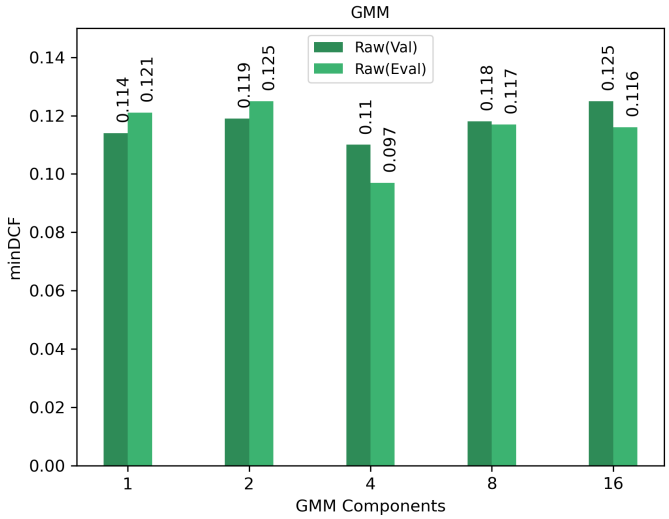


Figure 25: Diagonal assumption of GMM model on Test data

This result aligns with our earlier analysis, highlighting the relationships among certain attributes. When we make the assumption of no relationships between these attributes, we inadvertently neglect crucial information, leading to suboptimal outcomes. Notably, even in the case with just four components, considered the top candidate in the full covariance approach, the evaluation model yields a value of 0.097.

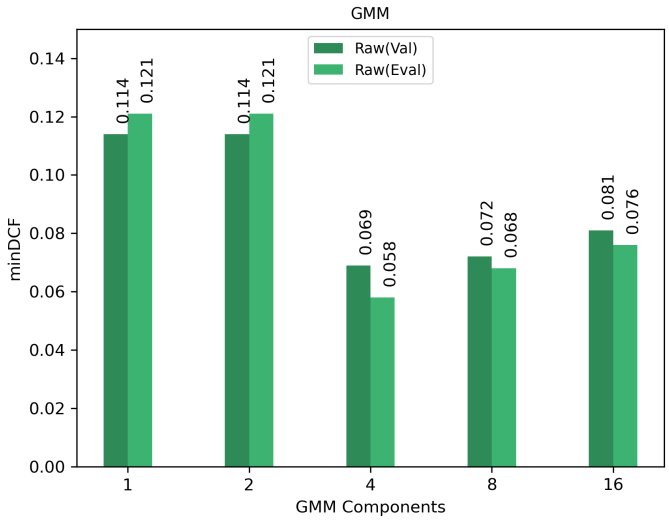


Figure 26: Tied assumption of GMM model on Test data

Assuming tied covariance results in high minDcf when use only 1 or 2 components. This suggests using less components may not adequately model the entire data(valuation and also the test data). However, dividing each class into 4 or more components yields favorable results, which align with our analysis for the data on the chapter 1, even though using 16 components is discouraged due to the risk of overfitting.

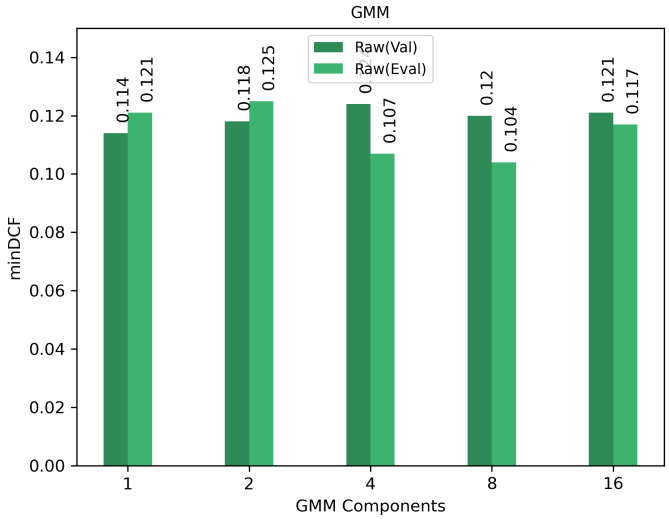


Figure 27: Tied and Diagonal assumption of GMM model on Test Data

Finally, we combine the Tied and Diagonal assumptions to observe the result. Since there exists a diagonal covariance matrix, the outcome is nearly identical to the pure diagonal case.

minDCF	Full	Diagonal	Tied	Tied + Diagonal
Valication	0.072	0.11	0.069	0.124
Evaluation	0.063	0.097	0.058	0.107

Table 20: GMM comparison in different covariance case(4 components)

In conclusion, when working with GMM, it's advisable to begin by observing the data distribution. This provides a valuable initial insight into the optimal number of components. Utilizing this information as a starting point can help the model-building process.

5.3 DET graph of models

DET graph shows following:

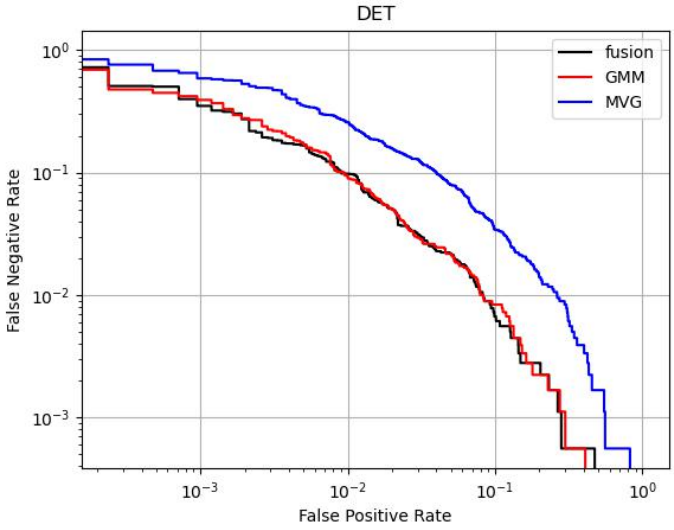


Figure 28: MVG + GMM Test DET Graph

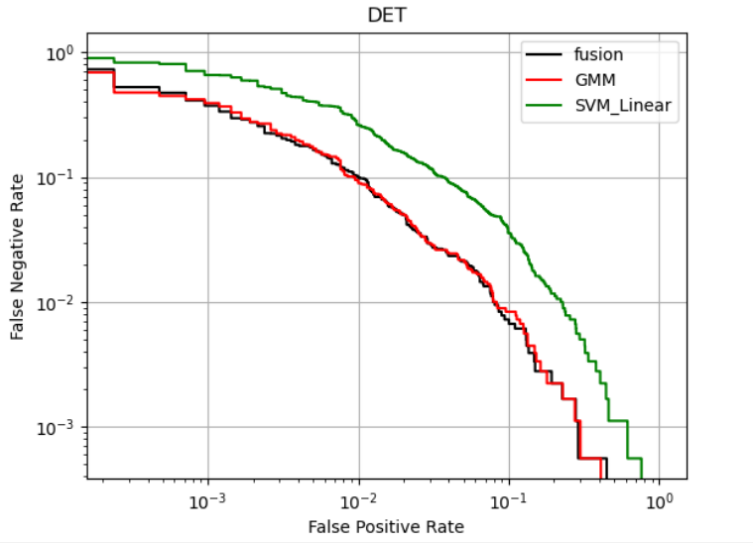


Figure 29: SVM + GMM Test DET Graph

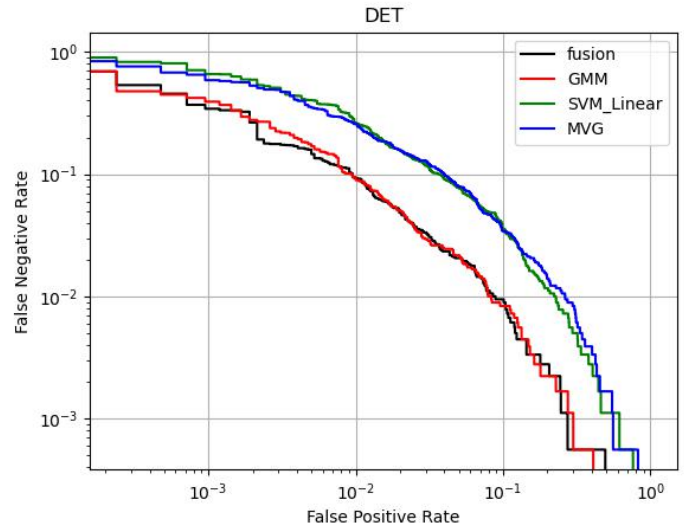


Figure 30: SVM + GMM + MVG Test DET Graph

GMM and the fusion which contains GMM has lower FNR and FPR, it indicates that the classification system is performing better in terms of both sensitivity (fewer false negatives) and specificity (fewer false positives). It means that the GMM model is achieving a better balance between correctly identifying positives and negatives, which is also a goal in our classification task.

6 Conclusion

Upon analyzing the evaluation set, we observed that our model effectively classifies the data. In terms of data input, we employed various methods to determine the optimal dimensionality for the model. Due to the limited dataset size, we discovered that using all 12 dimensions, without reduction, yielded the

best results. This highlights the importance of adapting dimension reduction techniques to specific scenarios; even highly correlated attributes should not be discarded when data is scarce. Preserving the entire dataset can lead to optimal solutions.

Furthermore, our investigation revealed that employing GMM with Tied covariance assumption, utilizing 4 components, particularly excels when our working point is around 0.5. For different working points, fusion models incorporating the GMM model also prove to be advantageous. Finally, it's worth noting that while a linear SVM yields favorable results, experimenting with non-linear SVMs not only substantially increases computational costs but also results in a higher minDCF.