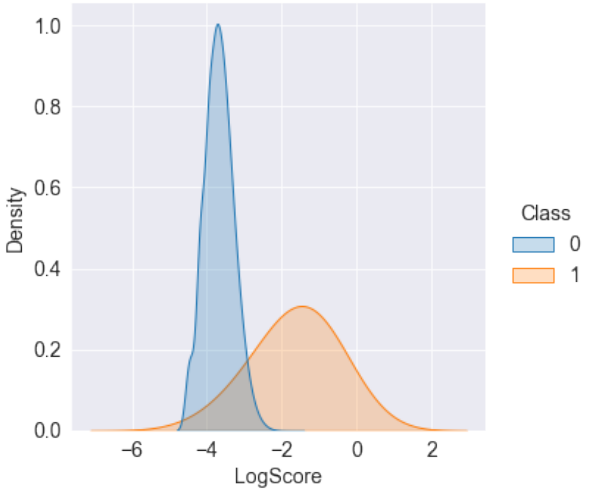
Baré Alexandre (r0912072)

Assignment 1 : Evaluating Performance of Biometric Systems

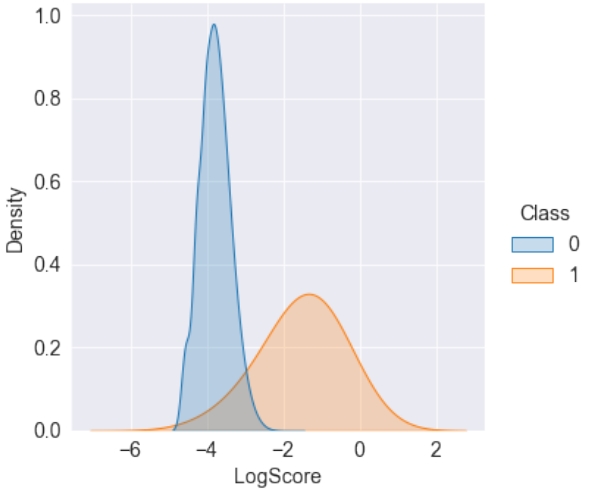
Biometrics System Concepts

Katholiek Universiteit Leuven
Faculty of Engineering Science
Academic Year 2021-2022

1 Q1 : Score distributions



((a)) li



((b)) ri

FIGURE 1 – Genuine and Impostor Score Distributions

We plot the distributions using *displot* of the library *seaborn* by Gaussian kernel density estimation (see the code below). We need to normalize independently the genuine and the impostor score distribution to plot relevant results (*common_norm* = *False*). Indeed, the classifier seems to perform very well : the impostor distribution is so narrow/condensed on a very small range of score values that the genuine distribution would not be visible alongside the impostor distribution without proper normalization. We also plotted the distribution on a logarithmic scale of the score for the same reasons.

```
log_scores = np.log(scores)
df_scores = pd.DataFrame({'Class': genuine_id.astype('int'), 'Score': scores, '
    LogScore': log_scores})
sns.displot(df_scores, x='LogScore', hue='Class', kind='kde', common_norm=False,
    fill=True, bw_adjust=4.5)
```

For the analysis, we limit the score range as suggested, say to approximately the range $[-4.5; 0]$ of log scores. We see that the impostor class is **limited to a very narrow region** from $[-4.5, -2]$ (resp. $[e^{-4.5}; e^{-2}] = [0.011; 0.135]$ in standard scale) and has its peak around -4 in logarithmic scale (resp. $e^{-4} = 0.183$ in standard scale). As a consequence, the FAR will most probably be very low for most decision thresholds. On the contrary, the genuine distribution **spread over a broader range of values** from -4.5 to 3 (resp. $[e^{-4.5}, e^{0}] = [0.011; 1]$) and has its peak around -1 (resp. $e^{-1} = 0.368$ in standard scale).[1] As a result, the FRR will probably be much more dependent of the decision threshold we go for.

Looking at these distributions that do not overlap much (especially given the fact that the x-axis is logarithmic and the distributions are normalized), we should expect that with an optimal decision threshold, the FAR and FRR will be very low. Indeed, the 2 classes are for the most part of the score range very distinctive.

At this stage, we hardly see the difference between the biometric system of the left index and the right index.

1. The numbers here are very approximate and are just used for the purpose of illustrating a qualitative analysis.

2 Q2 : ROC Curves

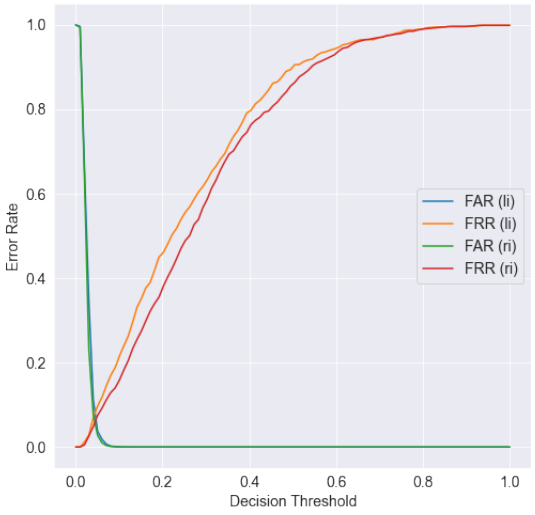
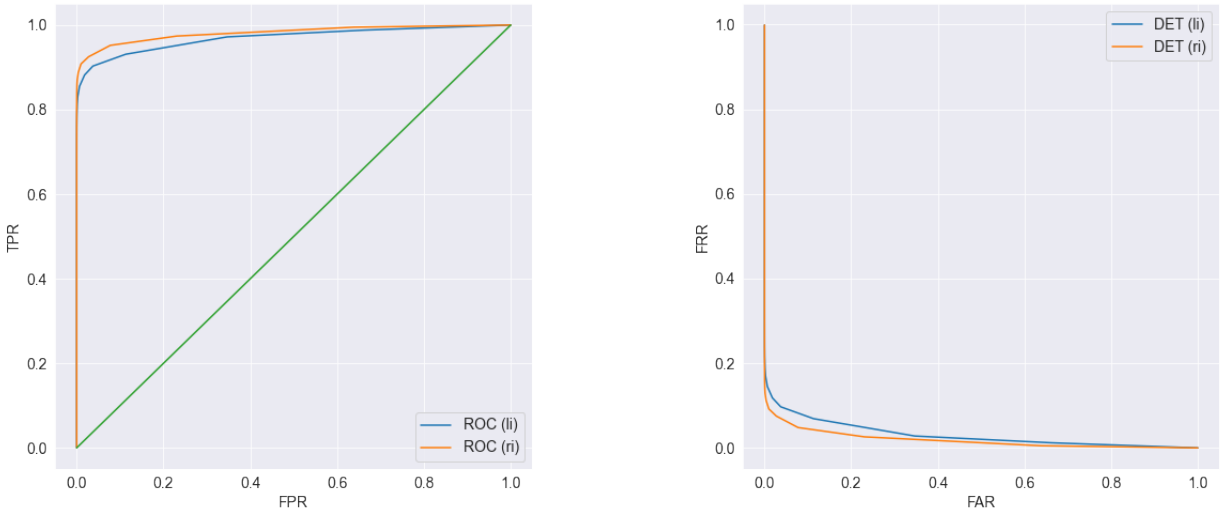


FIGURE 2 – FAR - FRR

In both systems, we notice that the FAR curves are very steep and that it is less the case for the FRR curves. It confirms our observations from question 1. We however better see the difference between the right index and left index systems here on the FRR curves. It seems that the right index system reject falsely less samples that are from the genuine class whatever the decision threshold. It already gives us a good indication that the right index system is a bit better at authentication. The difference between both systems is however negligible in the extreme values of the decision threshold.



((a)) ROC

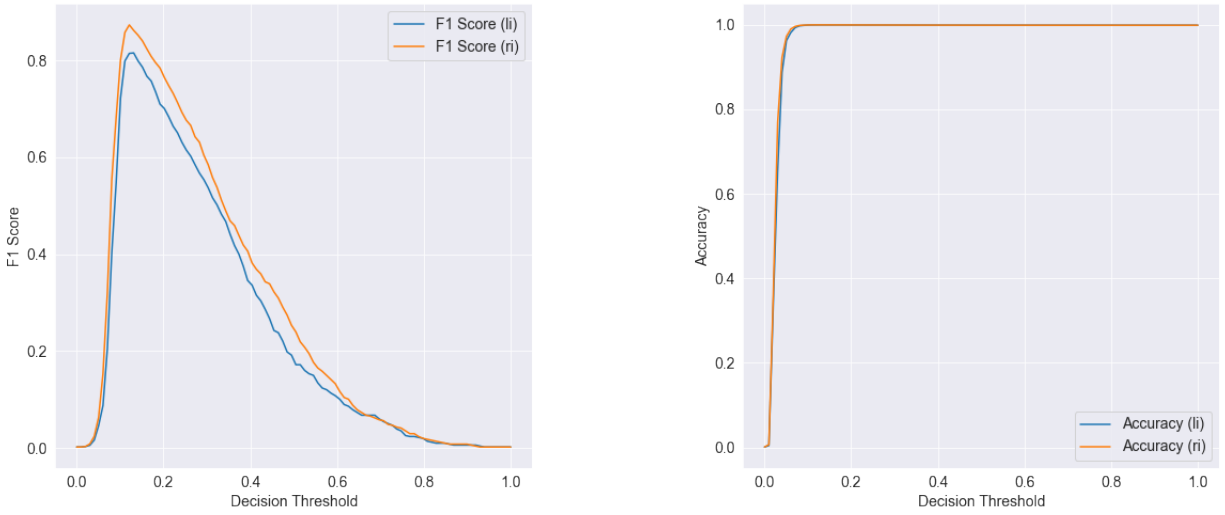
((b)) DET

FIGURE 3 – ROC & DET Curves

We observe through the ROC curves that the classifiers can be really good class discriminators. The flat slope of the ROC curves at most points indicates the extent to which the FPR can be lowered without damaging the TPR. In other words, given a good decision threshold, the biometric systems can have very low FPR and very high TPR at the same time.

For the DET curves, we observe again a very flat slope on a broad range of the x-axis. It means that the FAR (resp. security) can be reduced (resp. increased) without increasing (resp. decreasing) much the FRR (resp. biometric system convenience of use). This observation makes sense with what we have discussed for the ROC curves as FRR = 1 - TPR and FPR = FAR.

3 Q3 : Classification Metrics



((a)) F1 Score

((b)) Accuracy

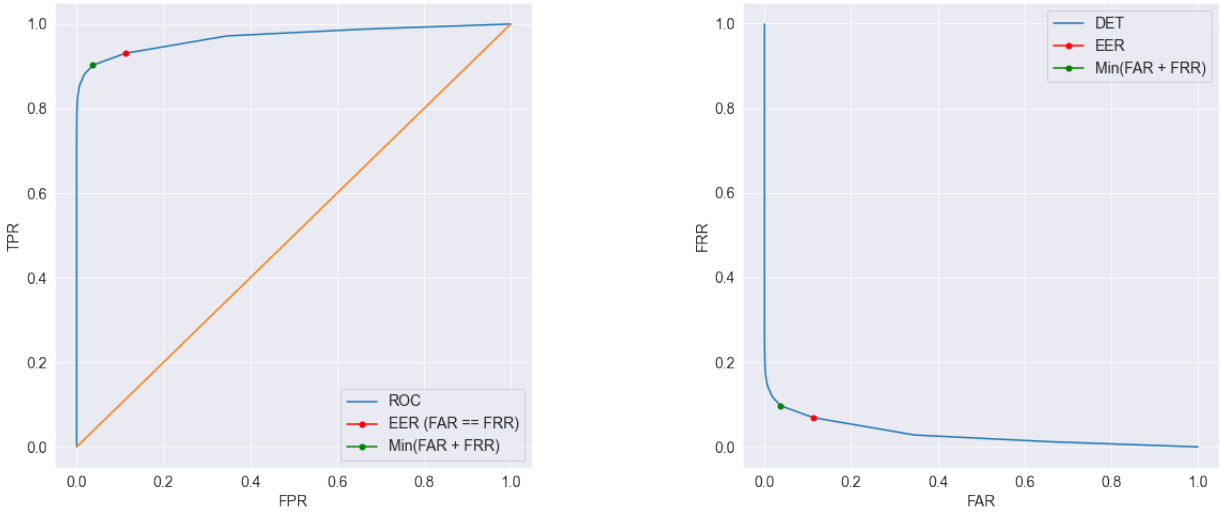
FIGURE 4 – F1 Score & Accuracy

The F1-score is (almost) overall higher for the right index system. It means that extreme values of precision and recall are more penalized, ensuring a better balance between these 2 metrics with the right index system. The peak 0.87 for the ri system (resp. 0.815 for the li system) lies at a decision threshold of 0.121 (resp. 0.131). The point at which F1 is maximal makes a decent compromise between precision (the fraction of genuine samples correctly detected) and recall (the probability of correctly detecting a genuine sample). This metric focuses on the classification of the genuine class. Nevertheless, depending on the application, a balance focusing on both classes could be sought instead.

Evaluating both biometric systems on the accuracy does not help in this case to differentiate them much. They have both a very high accuracy (= 1 - error rate) of 0.9998 from a threshold of 0.121 for ri (resp. 0.131 for li) onwards. But we know from the FAR and FRR curves, that this is mainly due to a very low FAR = 1 - TNR (i.e. a very high specificity). We know that FRR = 1 - TPR increases (i.e. sensitivity decreases) with a decision threshold increase. The accuracy plot does not seem to be taking that into account. Accuracy is thus not an appropriate metric to find a good balance between sensitivity (i.e. convenience) and specificity (i.e. security). It only promotes a good specificity in our case because there are far more negative than positive pairs of samples in a verification system, i.e. there are far more possible imposters than samples of genuine identity.

Both F1-score and accuracy happen to advise on the same decision threshold in our case although in general F1-score is a more trustworthy metric that is not influenced by class imbalance. It however focuses on the correct classification of positive samples while accuracy focuses on the correct classification of negative samples due to high class imbalance.

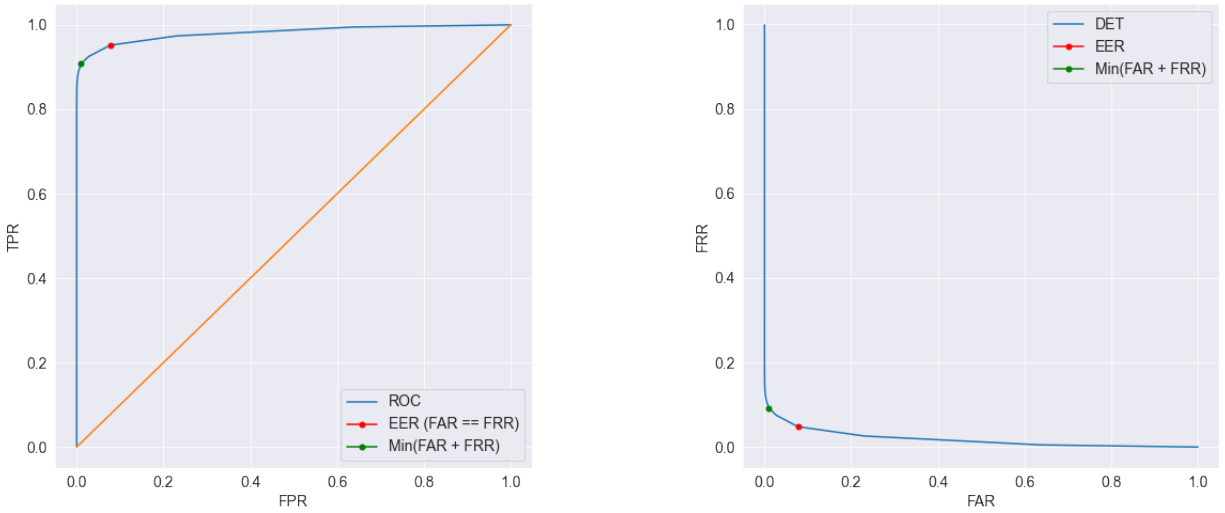
4 Q4 : AUC, EER and alternatives



((a)) ROC for li

((b)) DET for li

FIGURE 5 – ROC & DET Curves for li



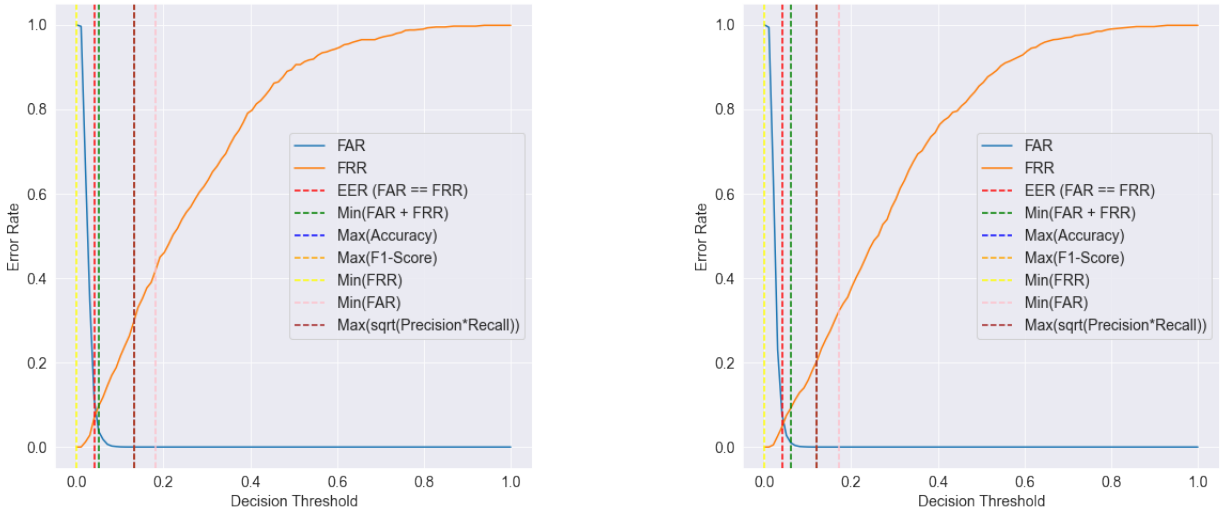
((a)) ROC for li

((b)) DET for li

FIGURE 6 – ROC & DET Curves for ri

As expected, the area under the ROC curve is quite big : 0.983 for ri (resp. 0.971 for li). The AUC provides an aggregate measure of performance across all possible classification thresholds. It can be interpreted as the probability that the model scores a random genuine sample more highly than a random impostor. The closer it is to 1 (100% of correct classification), the better it will classify irrespective of the classification threshold. However, in real-world applications, the classification threshold remains an important factor that greatly differs from a case to another.

Nonetheless, looking at the ROC curves from figure 3, we can not spot a region where the li system performs better than the ri system although at the extreme FPR values, both systems converge (which system is chosen then becomes irrelevant). We know from figure 2 that high decision thresholds result in a low FPR but among these thresholds, the system can differ a great deal in FRR. Comparatively, very low decision thresholds result in a high FPR and low FAR. Regions of convergence of the ROCs at extreme values of the decision threshold are thus not relevant in most applications. Given the fact that there is no region where the li system performs better than the ri system, the AUC could be thus a relevant metric in our case to compare both systems even though it does not discriminate on the choice of the decision threshold.



((a)) li ((b)) ri

FIGURE 7 – FAR - FRR

The EER point is the fairest choice (in the sense that FRR = FAR) between a low FAR (i.e. high security) and a low FRR (i.e. high convenience). As already mentioned before, FAR and FRR's importance are however application specific. We may want to lower one more than the other.

If we sum the FAR and FRR and minimize the resulting quantity, we do not prioritise one over the other but one may be reduced more to balance an increase of the other. The sum is in general not equal to the total classification error. Indeed, $FAR = \frac{FP}{FP+TN}$ and $FRR = \frac{FN}{FN+TP}$.

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ &= TPR * \frac{P}{P+N} + TNR * \frac{N}{P+N} \\ &= (1-FRR) * \frac{P}{P+N} + (1-FAR) * \frac{N}{P+N} \end{aligned}$$

Accuracy (= 1 - error rate) can not be simplified without further assumptions. From the formula above, we can however note that maximising accuracy is equivalent to minimizing FAR + FRR in the case of a balanced dataset (P = N) but not in our case. Minimising FAR + FRR is in fact equivalent to maximizing a weighted accuracy adjusted for imbalanced dataset (*balanced_accuracy_score* in *sklearn*) which makes this sum advantageous as a metric in such a context.

Other strategies exist to find an "optimal threshold" such as minimising only the FAR, i.e. maximising security, at the expense of the FRR, i.e. convenience of use (or vice-versa). We can minimize the geometric mean of the precision and recall although in practice it is similar to the case of the F1-score but the extremes are less penalized. In our case, it leads to the same optimal point than with the F1-score and the accuracy.We could also imagine weighting the precision and recall differently in the F1-score or the geometric mean to give more importance to one or the other.

5 Q5 : Precision-Recall curves and related summary measures

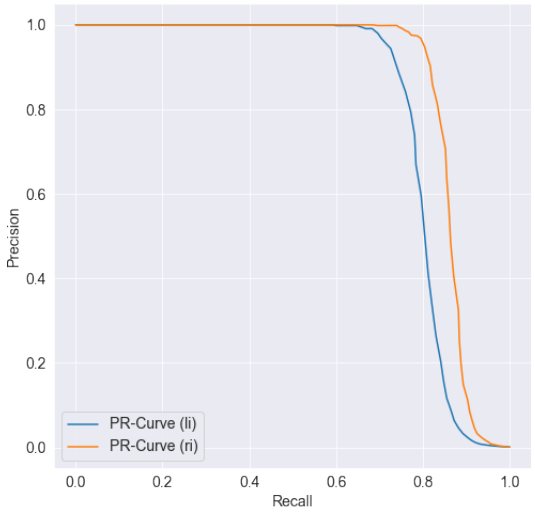


FIGURE 8 – Precision-Recall Curves

The precision-recall curves are especially interesting in place of the ROC curves in a verification scenario (such as in our case) where there is typically a high class imbalance. It focuses mainly on the genuine class (via PPV and TPR) in minority and cares less about the frequent impostor class. We see that the curves are quite flat under low recall values. It therefore implies that the recall can be increased effortlessly without affecting the precision. However, from a recall of 0.8 onwards for the ri system (resp. 0.7 for the li system), we observe a steep drop in precision. It becomes very costly in terms of precision to improve the sensitivity (= recall or 1-FRR, i.e. convenience of use). We notice that for decent recall values, the precision of the ri system is better than that of the li system.

We observe also that the baseline of the PR-curves is very low (close to 0). It is determined by the ratio of genuine samples among the total population which is unsurprisingly very low given the important class imbalance.

The area under the PR-curve is big : 0.863 for ri (resp. 0.803 for li). We compute it as follows

```
pr_auc = sklearn.metrics.auc(recall, precision)
```

It is an aggregate measure of the precision across all possible recall values. The higher it is, the better the system is at correctly detecting genuine samples. It is, as for the ROC-AUC, classification threshold independent. The classification threshold is however a crucial parameter to adjust differently for each application. We will prefer the PR-AUC over the ROC-AUC in a verification system because there are far more impostor pairs than genuine pairs.

The PR-AUC can actually be seen as the average precision scores.

```
average_precision_scores = sklearn.metrics.average_precision_score(true_classes,
    matching_scores)
```

It yields close but still different results. It is explained in sklearn documentation that "[*average_precision_score*] implementation is not interpolated and is different from computing the area under the precision-recall curve with the trapezoidal rule, which uses linear interpolation and can be

too optimistic."[2] We will thus preferably look at the second implementation in place of the first one. The average precision score for ri is 0.860 and for li is 0.799. Once again, the ri system performs better than the li system.

6 Q6 : CMC curves

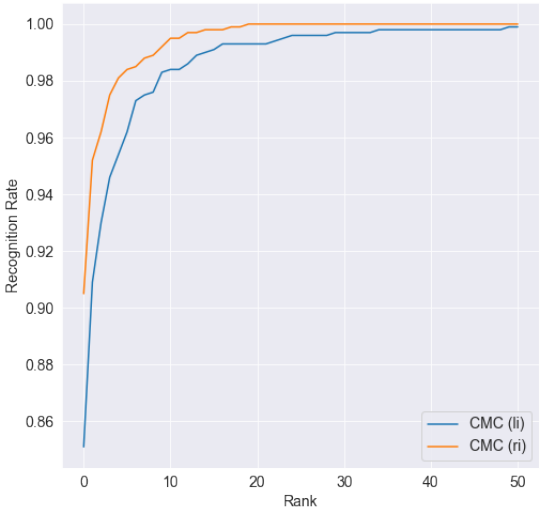


FIGURE 9 – CMC

A CMC curve is used in an identification system to indicate how high the true positive identification rate (TPIR) is for different ranks t. From the similarity matrix, we are able to rank the biometric sample of the users enrolled in the system that best matches with a given sample. The TPIR is then the expected proportion of identification by users enrolled in the system where, among the t identities that match the most, the correct identity is present. If, in a specific application, security requirements are lower than in a verification system, we can see that increasing the rank considerably helps to increase the TPIR.

The rank-1 recognition rate is equal to 0.905 for ri (resp. 0.851 for li) which corresponds to the average of the recall for each user in a scenario where we only want to detect the best match. The ri system outperforms again the li system.

7 Q7 : Evaluate different biometric systems

See the answers to the questions above.

2. `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html#sklearn.metrics.average_precision_score`