Assignment 1

Rojina Kashefi (r1018183)

March 2025

Section 1: Score distributions

1.1

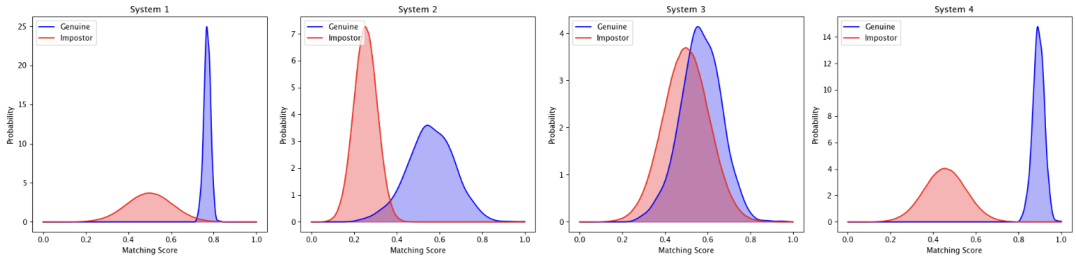

Figure 1: Genuine and impostor score distributions

1.2

In the sim2scores function, when reading the genuine and impostor scores from the similarity matrix, we normalize them to the range [0, 1]. Since the scores are already normalized, no further normalization is required when calculating and comparing the distributions.

1.3

A good biometric system should produce high genuine scores (strong matches for the same user) and low impostor scores (weak matches for different users). In addition, the amount of overlap between the genuine and impostor score distributions determines the system's misclassification. Less overlap indicates better performance, as overlap suggests incorrectly matching different users as the same or the same user as a different user. Based on the plot, System 4 shows no overlap and genuine scores are very close to 1, which means the best performance. System 1 has a small amount of overlap, which will be second best. System 2 has moderate overlap, and System 3 shows significant overlap with the worst performance. Therefore, the order from best to worst is: **System 4** > **System 1** > **System 2** > **System 3**.

1.4

- **System 1:** The genuine distribution is narrow, showing small intra-user variations. There is a small overlap between the genuine and impostor distributions, suggesting small inter-user similarity.
- **System 2:** The genuine distribution is more spread out, showing large intra-user variations. Despite of the overlap, there is a large inter-user distance, as the distributions are distinguishable.
- **System 3:** The genuine distribution is wide, showing large intra-user variation. The genuine and impostor distributions overlap significantly, showing large inter-user similarity. This system has the worst performance.
- **System 4:** The genuine distribution is narrow, showing small intra-user variations. The overlap between the genuine and impostor distributions is minimal, showing large inter-user distance. This system has the best performance.

Section 2: ROC Curves

2.1

Table 1 shows that system 4 has the best performance. It accepts all real users (TPR: 100%) and blocks all impostors (FPR: 0%). On the other hand, System 3 has the weakest performance, (low TPR: 0.636% and high FPR: 0.364%). Also, A higher EER threshold (like 0.809 in System 4) means the system is more strict and only accepts highly confident matches, while a lower threshold (like 0.537 in System 3) means the system is more forgiving and accept less confident matches. In general, systems with higher thresholds and better TPR/FPR values are more reliable and accurate. Therefore, the order from best to worst is: **System 4** > **System 1** > **System 2** > **System 3**.

System	EER Threshold	TPR	FPR
System 1	0.733	0.985	0.015
System 2	0.351	0.962	0.038
System 3	0.537	0.636	0.364
System 4	0.809	1.000	0.000

Table 1: TPR and FPR at EER Threshold

2.2

In Figure 2, we can see that as the threshold increases, the false acceptance rate (FAR) decreases because the system becomes more strict in accepting impostor inputs. On the other hand, the false rejection rate (FRR) increases, as the stricter threshold may also reject some genuine inputs.

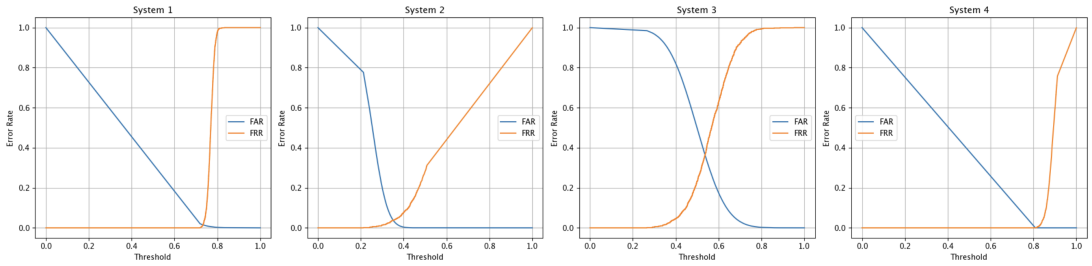

Figure 2: FAR and FRR as a function of the decision threshold.

2.3

In the ROC curve, the closer a line is to the top-left corner, the better the performance. This is because the True Positive Rate (TPR) is high and the False Positive Rate (FPR) is low in that region. We can see that Systems 1, 2, and 4 move toward the top-left corner from the very beginning, which makes it necessary to analyze them on a logarithmic scale. However, from the ROC curve, we can see that System 3 performs the worst. It has a much lower TPR for the same FPR compared to the others. Its performance is close to the diagonal line, which means it behaves similarly to random guessing. In the log-scale ROC curve, we observe that System 2 performs better than System 4 at very low FPR values. However, as the threshold changes, System 4 eventually shows almost perfect classification. Although System 1 starts off performing worse than random, its TPR increases rapidly at low FPR values and eventually outperforms System 2. This shows that System 1 is sensitive to the threshold. System 2 also performs well, maintaining a high TPR across all thresholds.

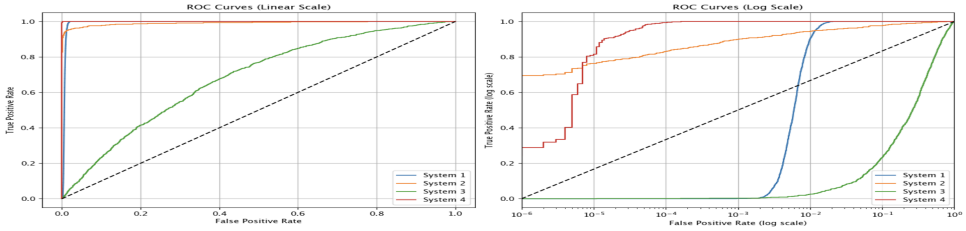

Figure 3: ROC curve

2.4

The Detection Error Trade-off (DET) curve is commonly used to evaluate detection systems in biometrics. Unlike the ROC curve, which plots the True Positive Rate, the DET curve displays the False Negative Rate (FNR) against the False Positive Rate (FPR). The ideal point on the DET curve is at the bottom-left corner, where both FNR and FPR are zero. By analyzing this curve, we can gain a better understanding of how different systems perform in terms of errors and compare their behavior in failure scenarios. This curve aligns with the ROC results. System 4 performs the best overall, except at the very beginning. As we move along, System 1 also outperforms System 2. System 3 struggles the most, with both high false positive and false negative rates. In other words, it ends up letting in many impostors while also rejecting genuine users.

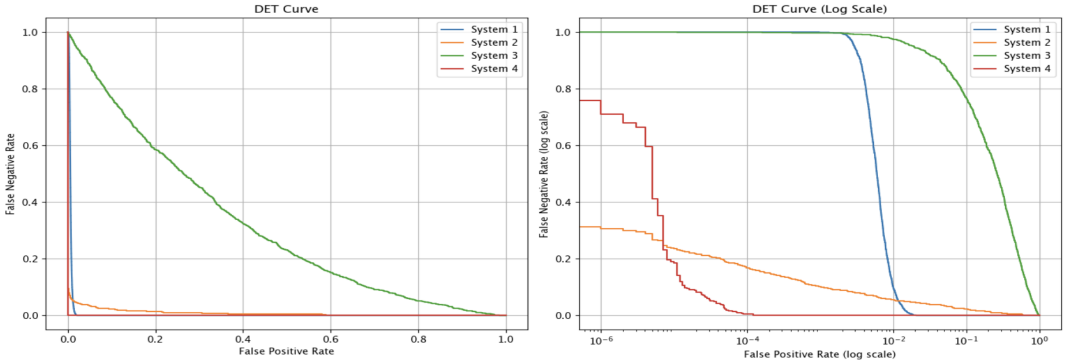

Figure 4: DET curve

Section 3: Classification Metrics

3.1

In Figure 5, we observe that as the threshold increases, the system's accuracy also increases. This is because the dataset is imbalanced, containing 999,000 negatives (impostors) and only 1,000 positives (genuine). As a result, correctly predicting more negatives (true negatives) dominates the accuracy formula and causes it to rise even when the model misses all the positives. On the other hand, the F1 score, which balances the actual positive rate and the predicted positive rate, peaks at a threshold where both precision and recall are reasonably high. After this point, recall tends to drop faster than precision improves, causing the F1 score to decrease.

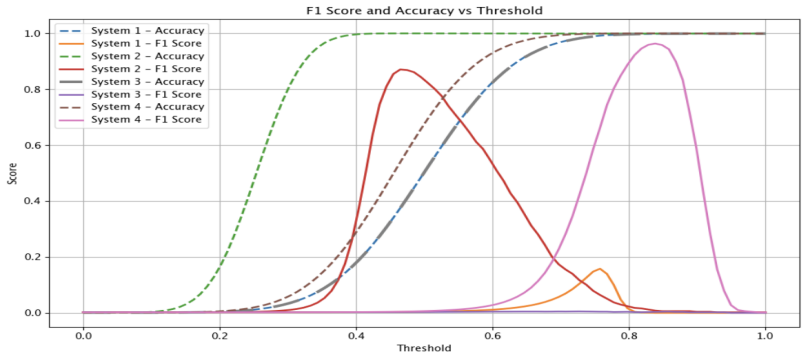

Figure 5: F1 score and accuracy vs. threshold

3.2

In Table 2, we can see that all systems reach high accuracy, close to 1, at the threshold where the F1 score is maximized. This happens because the F1 score is highest when both precision and recall are balanced. This means the model is making correct predictions across both classes, resulting in high overall accuracy.

System	Max F1 Score	Threshold at Max F1	Accuracy at Max F1
System 1	0.1579	0.7576	0.9917
System 2	0.8706	0.4646	0.9998
System 3	0.0052	0.7273	0.9820
System 4	0.9634	0.8384	0.9999

Table 2: Performance metrics for each system at the threshold that maximizes F1 score.

3.3

In Table 3, we notice that Systems 1 and 3 reach around 99% accuracy, but their F1 scores is zero. This tells us that they're likely predicting everything as negative (impostor) (999000/1000000), completely missing the positive cases. It shows that max accuracy does not always align with max F1, especially in imbalanced datasets. A better strategy is often to look for the F1 score, which gives a more balanced view of performance across both classes. On the other hand, System 4 hits both its highest accuracy and F1 score at the same threshold, meaning it's doing a good job at recognizing both positive and negative samples. This makes it a more reliable and practical choice for biometric cases. System 2 behaves similarly, with its best F1 and accuracy happening at close thresholds.

System	Max Accuracy	Threshold at Max Acc	F1 at Max Accuracy
System 1	0.9990	1.0000	0.0000
System 2	0.9998	0.4747	0.8685
System 3	0.9990	1.0000	0.0000
System 4	0.9999	0.8384	0.9634

Table 3: Performance metrics for each system at the threshold that maximizes Accuracy score.

3.4

No, accuracy isn't a good metric here. F1 score is better because it balances precision and recall, which helps us detect the positive class. Accuracy can be misleading when most samples are negative. For example, if 96 out of 100 samples are negative, a model that always predicts negative gets 96% accuracy but misses all the positive cases, which matter most in biometrics. F1 score shows how well the model handles those rare but important cases.

Section 4: AUC, EER and alternatives

4.1

The ROC AUC is a common metric used to evaluate how well binary classifiers perform, especially in verification systems. A score of 1.0 indicates perfect separation between genuine and impostor samples, while a score of 0.5 means the classifier is guessing randomly. ROC AUC is useful because it measures performance independently of any decision threshold, which is especially important when working with imbalanced datasets. Table 4 shows that Systems 4, 1, and 2 have high AUC scores, showing strong separation between genuine and impostor samples. However, System 3 has an AUC close to 0.68, meaning it often confuses impostors with genuine users.

System	ROC AUC
System 1	0.9935
System 2	0.9916
System 3	0.6833
System 4	1.0000

Table 4: ROC AUC scores for each system.

4.2

EER is a widely used operating point, as it represents the threshold where the false acceptance rate (FAR) equals the false rejection rate (FRR), indicating a balanced trade-off between the two types of

errors. However, if the application prioritize one error type over the other, EER may not be the most appropriate metric. As shown in Figure 6, Systems 1, 2, and 4 have low EER values. This shows fewer false detections. In contrast, System 3 has a high EER of 36%, suggesting poor performance.

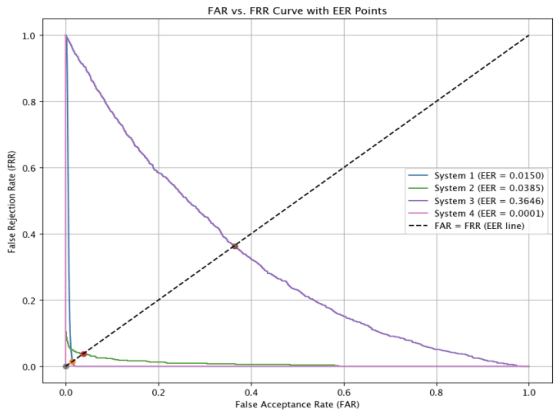

Figure 6: FAR vs. FRR curves with EER points and EER line.

4.3

If the dataset is balanced, the threshold that minimizes the sum of FAR and FRR also minimizes the total classification error. In such cases, this threshold often aligns with the one that maximizes overall accuracy. However, when the dataset is unbalanced, the total misclassification error (1 - accuracy) can give misleading results, while the sum of FAR and FRR still provides a clearer and more reliable measure of the system's performance. In Table 5, we can see that System 4 has the lowest total error, followed by System 1 and System 2, while System 3 has the highest error.

System	Min Total Error	Threshold
System 1	0.0196	0.7235
System 2	0.0627	0.3778
System 3	0.7172	0.5166
System 4	0.0002	0.8086

Table 5: Minimum total error and corresponding threshold.

4.4

In a very secure system, minimizing the False Acceptance Rate (FAR) is most important because you want to prevent unauthorized users from being accepted, even if that means occasionally rejecting authorized users (higher FRR). On the other hand, in a very convenient system, minimizing the False Rejection Rate (FRR) is more important to ensure that genuine users are not wrongly denied access, even if a few impostors might get accepted (higher FAR). The balance between FAR and FRR depends on whether the system prioritizes security or user convenience.

Section 5: Precision-Recall curves and related measures

5.1

The Precision-Recall curve shows how well a system finds genuine users while avoiding impostors, which is useful when working with imbalanced data. Precision means of all the users the system accepted, how many were actually genuine. Recall means of all the genuine users, how many did the system correctly accept. The closer the curve is to the top-right corner, the better the system's performance. In Figure 7, System 4 performs the best, maintaining high precision and recall across all thresholds. System 2 also performs well, though its precision drops slightly as recall increases. This may be because increasing recall means the system is trying to catch more genuine users, which usually involves lowering the decision threshold. While this helps find more true positives, it also allows more impostors through.

System 1 starts off too strict, rejecting almost everyone, and only slightly improves later. System 3 performs the worst, with low precision across all recall levels. Overall, Systems 4 and 2 are the best, while Systems 1 and 3 struggle to distinguish between genuine users and impostors.

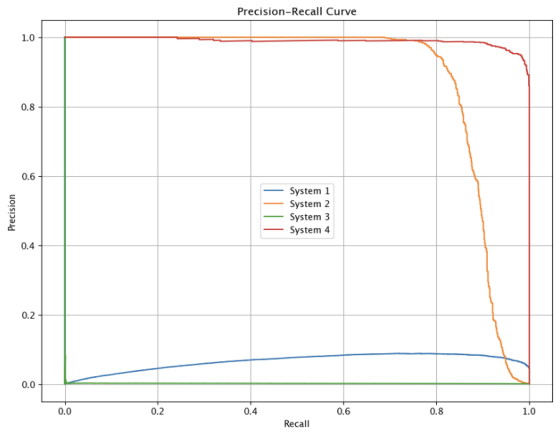

Figure 7: Precision-Recall curve

5.2

The Precision-Recall Curve (PRC) is more appropriate for evaluating these four systems because we are dealing with highly imbalanced data (1,000 genuine users and 999,000 impostors). PRC focuses on how well the system identifies the positive class (genuine users), which is especially important in biometric applications where false positives can be critical. In contrast, the ROC curve can give an overly optimistic view in imbalanced settings, as it includes true negatives in its calculation. ROC curves are more suitable when the number of positive and negative samples is roughly balanced.

5.3

The Area Under the Precision-Recall Curve (PR AUC) measures how well a system balances precision and recall across all thresholds. A higher PR AUC means the system is better at correctly identifying genuine users while avoiding impostors. Based on calculations we see System 4 > System 2 > System 1 > System 3, which is similar to the conclusion we got from the PR curve analysis.

System	PR AUC
System 1	0.0653
System 2	0.8877
System 3	0.0019
System 4	0.9893

Table 6: Precision-Recall AUC values for each system

5.4

Average Precision is the area under the Precision-Recall curve, but instead of just calculating the area like PR AUC, it takes the average of the precision values at different levels of recall. It gives more importance to higher recall, meaning the system is rewarded more when it catches more genuine users without letting in too many impostors. Higher AP values shows stronger performance. When the Precision-Recall curve is jagged, the PR AUC and Average Precision scores may differ. However, when the curve is smooth, the two scores are often very close or even identical. In this case, they are similar because the systems have smooth PR curves.

System	Average Precision
System 1	0.0654
System 2	0.8877
System 3	0.0020
System 4	0.9893

Table 7: Average Precision values for each system

Section 6: CMC curves

6.1

The CMC curve shows how likely it is that the correct identity appears within the top-k matches. A steeper curve means better performance, it shows the system can recognize users correctly at lower ranks. In Figure 8, we can see that System 4 reaches the top first, followed by System 1, then System 2, and finally System 3.

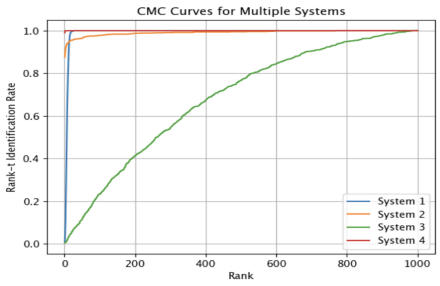

Figure 8: CMC curve

6.2

System 4 performs the best, achieving nearly perfect Rank-1 accuracy (99.10%) and reaching 100% identification right away. System 1, while having a very low Rank-1 rate (only 1.00%), reaches to 100% in just a few ranks. This means that even though it often misses at the first guess, it usually includes the correct identity within the top 30 results. That can still be useful in some scenarios, though not in biometrics, where early accuracy is critical and even small mistakes matter. System 2 starts strong with a high Rank-1 rate of 87.70%, but its CMC curve rises slowly, only reaching 100% around rank 700. So if System 2 makes a mistake at Rank-1, it takes a long time to recover, making it less reliable than System 1 in terms of consistency, despite having a better Rank-1 score. Finally, System 3 performs the worst across all metrics. It has the lowest Rank-1 accuracy (0.40%) and a very slow CMC curve, taking more than 800 ranks to reach full identification.

System	Rank-1 Recognition Rate (%)
System 1	1.0
System 2	87.7
System 3	0.4
System 4	99.1

Table 8: Rank-1 Recognition Rates

Section 7: Evaluate different biometric systems

7.1

The d-prime value (d'), also known as Cohen's d, is a statistical measure used to assess how well a verification system can tell apart genuine users from impostors. It captures the separation between their score distributions, measured in units of standard deviation.

$$d' = \frac{\sqrt{2} \cdot |\mu_1 - \mu_0|}{\sqrt{\sigma_1^2 + \sigma_0^2}}$$

Here, μ_1 and σ_1 are the mean and standard deviation of the genuine score distribution, while μ_0 and σ_0 refer to the mean and standard deviation of the impostor score distribution. A larger d' value means the two distributions are more separated, so the system can more easily and accurately distinguish genuine users from impostors. In Table 9, we can see that System 4 has the highest d-prime value, followed by System 2, then System 1, and finally System 3.

System	d-prime (d')
System 1	3.507
System 2	3.513
System 3	0.666
System 4	6.081

Table 9: d-prime values for different systems

7.2

In different applications, not all errors carry the same importance. For example, in a security-focused system, a false positive (allowing an unauthorized user) can be more serious than a false negative (blocking an authorized user). Conversely, in user-friendly applications, false negatives may be more important, as they directly affect the user's experience. To account for these differences, a metric called the **Weighted Error Rate (WER)** is used. This metric assigns different importance (weights) to each type of error:

$$\text{WER} = w_{\text{FNMR}} \times \text{FNMR} + w_{\text{FMR}} \times \text{FMR}$$

Here, w_{FNMR} is the weight assigned to the False Non-Match Rate (FNMR), and w_{FMR} is the weight assigned to the False Match Rate (FMR). For instance, in high-security applications, w_{FMR} might be higher to minimize the risk of unauthorized access. In contrast, in user applications, w_{FNMR} might be higher to avoid frustrating users.
In Figure 9, we illustrate two scenarios: one security focused and the other user convenience focused. In the security scenario, we set the weight of FMR to 10. In the user-convenience scenario, we assign a weight of 10 to FNMR. In the security-focused case, higher thresholds make the system stricter, reducing false accepts and lowering the WER. On the other hand, the user-convenience scenario prioritizes being easy to access, so increasing the threshold results in more false rejects, causing WER to rise. From the security plot, we can see that **System 2** performs the best among the four. It reaches nearly zero WER around $\tau = 0.41$, much earlier than the other systems. This suggests it achieves strong security (very few false accepts) without requiring a high threshold. The next best performers are System 4, and then System 1. The worst is System 3, which only reaches its lowest WER at $\tau = 0.94$. In the user-convenience scenario, lowering the threshold makes the system more user-friendly. At lower thresholds, **System 2** again performs the best, followed by System 4, then System 1, and finally System 3. However, at higher thresholds (when the system becomes more strict), system 1 and 4 still maintain low error rates. This means they remain reliable.

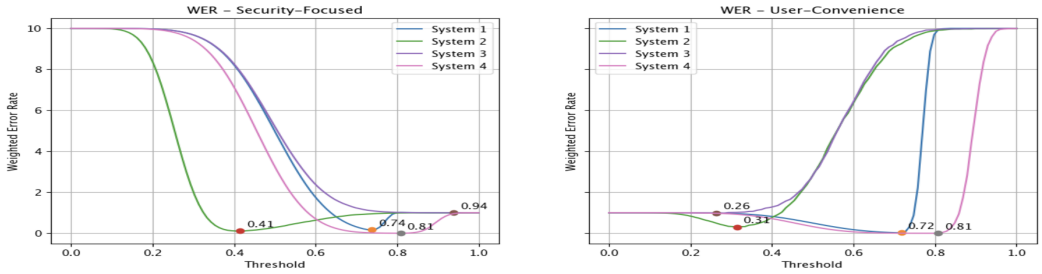

Figure 9: WER across different scenarios