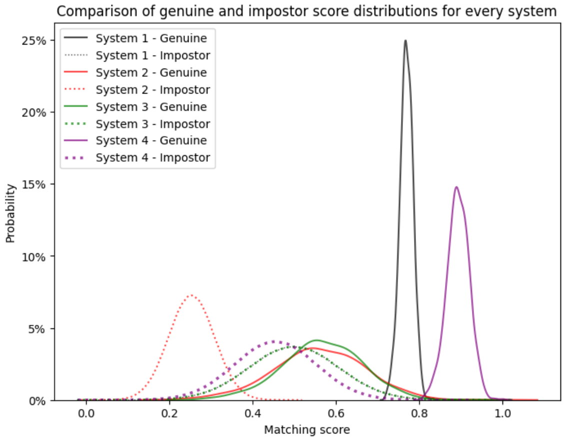
Q1:

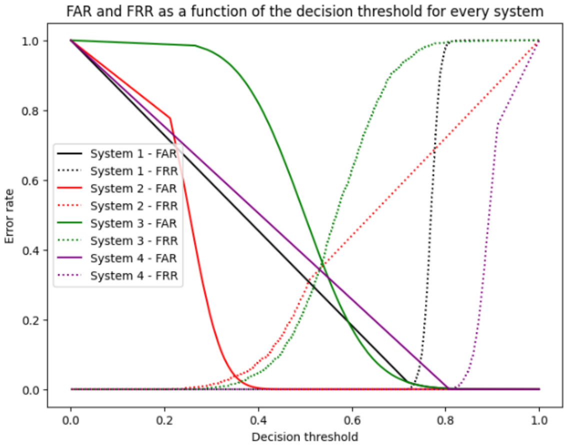


The score distributions need to be normalised by dividing the occurrence of each score by the total number of observations within its respective class, resulting in a probability distribution where the area under the curve is equal to 1. This ensures that the score distributions of two classes - genuine and impostor - are comparable between each other. In our case of imbalanced data, where 99.9% of observations belong to the impostor class and only 0.1% to the genuine class, an unnormalised score distribution would make the underrepresented genuine class almost invisible. Normalization addresses this imbalance, ensuring a meaningful comparison between the two distributions.

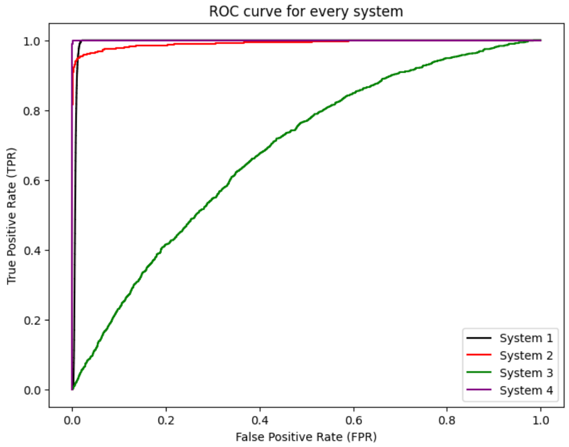
Based on the combined plot, System 4 performs the best, as it has the smallest uncertainty interval of overlap between the genuine and impostor score distributions. This allows the decision threshold for System 4 to be defined in such a way that a low sum of False Rejection Rate (FRR) and False Acceptance Rate (FAR) is achieved. In contrast, System 3 performs the worst, as the uncertainty interval is the largest. The performance of System 1 and System 2 is in the middle, because the overlap between the genuine and impostor score distributions is moderate for them.

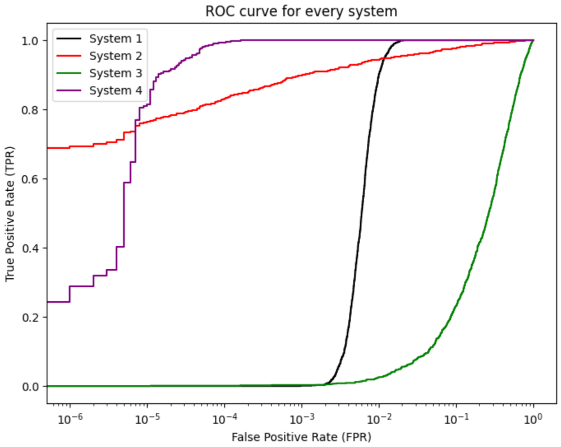
System 4 corresponds to B - small inter-user similarity, i.e. it is unique, as the averages of genuine and impostor scores are far apart, and small intra-user variations, i.e. it is permanent, as the standard deviation of genuine scores is low. System 3 corresponds to C - large inter-user similiarity, as the averages of genuine and impostor scores are close to each other, and large intra-user variations, as the standard deviation of genuine scores is high. System 2 corresponds to D – large inter-user user distance, i.e. small inter-user similarity, as the averages of genuine and impostor scores are far apart, and large intra-user variations, as the standard deviation of genuine scores is high. System 1 corresponds to B or A with medium inter-user similarity, as the averages of genuine and impostor scores are quite far apart, although closer than in System 4, and small intra-user variations, as the standard deviation of genuine scores is low.

Q2:

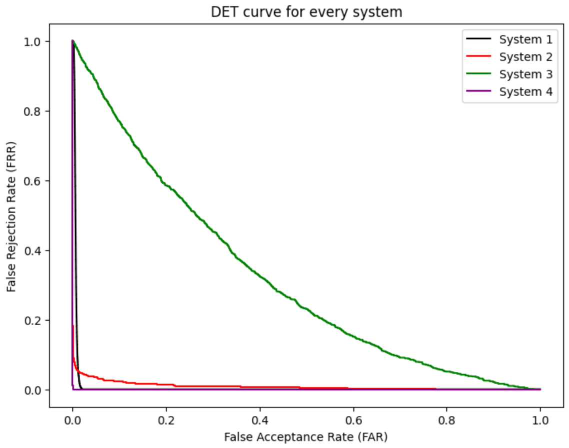


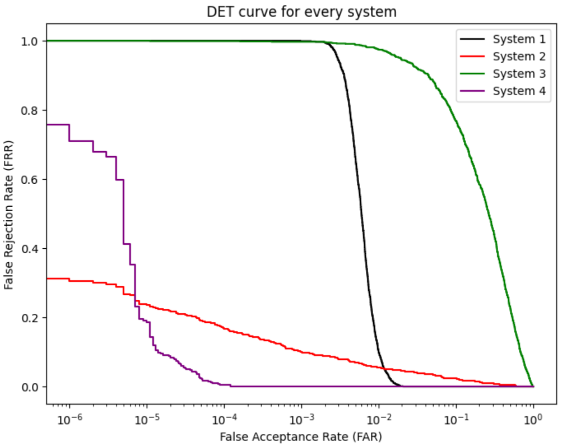
As the decision threshold increases (becomes more conservative), the False Acceptance Rate (FAR), or False Positive Rate, decreases. For lower decision thresholds (more liberal), the False Rejection Rate (FRR), or 1-True Positive Rate (TPR), is lower. Further, we analyse the significance and behaviour of these two metrics for every system. It is important to note that the decision thresholds should be evaluated relatively, considering the threshold for each system separately, rather than comparing between absolute values, as their interpretation depends entirely on the underlying score distributions of the systems, e.g. the averages of the genuine and impostor scores of System 2 are below the averages of the genuine and impostor scores of Systems 1 and 4, respectively, as given in Q1, and so on.





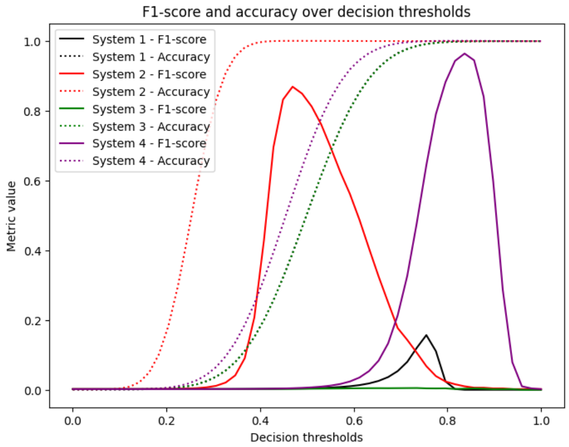
It can be observed that System 4 achieves the highest True Positive Rate (TPR) for most of the False Positive Rate (FPR) values given. This means that for most decision thresholds, System 4 performs best of all the systems. However, at the very low FPR values, System 2 outperforms System 4 and other systems by achieving higher TPR values. This indicates that at low FPR values, System 2 accepts more genuine cases than System 4, making it a better choice for highly secure biometric systems that require low False Positive Rate with relatively more conservative decision thresholds. However, as the FPR increases and the decision threshold becomes more liberal, System 4 outperforms System 2 and the other systems, making it the better choice for more convenient systems. Furthermore, System 3 is the worst system for each decision threshold. For very convenient systems, System 1 is better than System 2.





The Detection Error Trade-off (DET) curve has the same horizontal line as the ROC curve, with the False Acceptance Rate (FAR) on the x-axis, which corresponds to the False Positive Rate (FPR) in the ROC curve. However, the y-axis of the DET curve represents the False Rejection Rate (FRR), which is 1-True Positive Rate (TPR), whereas the y-axis of the ROC curve represents the TPR. Therefore, the conclusions drawn from the DET curve are fundamentally the same as those drawn from the ROC curve, as both curves provide the same information. The only difference is that the DET curve uses the TPR information as FRR=1-TPR on the y-axis. Thus, for the DET curve, the closer the curve is to the point (0,0), the better the system, as it indicates low FRR and low FAR. In contrast, for the ROC curve, the closer the curve is to the point (0,1) the better, where the TPR is high and the FPR is low.

Q3:



The F1 score can be interpreted as a harmonic mean of precision and recall, where the relative contribution of precision and recall to the F1-score is equal. The precision describes how good a model is at predicting the positive class and the recall is the ability of the classifier to find all the positive samples. It ranges from 0 (worst) to 1 (best) and is calculated as follows:

$$\text{F1}=2*\frac{\text{Precision*Recall}}{\text{Precision+Recall}}=\frac{2*\text{TP}}{2*\text{TP+FP+FN}}, \text{ where Precision}=\frac{\text{TP}}{\text{TP+FP}} \text{ and Recall}=\frac{\text{TP}}{\text{TP+FN}}$$

The results of maximising the F1 score are as follows:

System	Max F1-Score	Decision Threshold	Accuracy
System 1	0.1567	0.7551	0.9912
System 2	0.8684	0.4964	0.9998
System 3	0.0051	0.7347	0.9847
System 4	0.9364	0.8367	0.9999

It is an interesting operating point because the F1-score metric is a suitable metric for the imbalanced data of our case, though with some caveats. Since F1-score does not depend on True Negative (TN), it does not reward correct prediction of the majority class of negative users (impostor users). This prevents from being dominated by the rejection of impostors, which could otherwise inflate the metric value. However, it is important to note that precision, a component of the F1-score, is class prior dependent because it combines results from both positive and negative samples - including False Positive (FP), and is thus influenced by the number of majority class of impostors. The Harmonic Mean of precision and recall penalizes the extreme values and make the relative contribution of precision and recall to the F1-score equal, thus a high metric value requires both precision and recall to be high. However, a higher F1-score does not always mean a better model, as precision or recall might be more critical depending on the use case.

Although the F1 score is an effective metric for evaluating models when both high precision and recall are required and their contributions should be equal. System 4 performs best, achieving the highest F1-Score of 0.9364, which means that it can provide high precision and recall at the decision threshold of 0.8367. The worst system is System 3, which never achieves a high F1-score, with its best value being 0.0051 at a threshold of 0.7347. System 2 also performs significantly better than System 1, being able to achieve a much higher maximum F1-score.

Accuracy measures the proportion of correctly classified observations (both genuine and impostor) over the total number of observations, given as follows:

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}=\frac{\text{TP}}{\text{TP+FN}}*\frac{\text{P}}{\text{P+N}}+\frac{\text{TN}}{\text{TN+FP}}*\frac{\text{N}}{\text{P+N}}=\text{sensitivity}*\frac{\text{P}}{\text{P+N}}+\text{specificty}*\frac{\text{N}}{\text{P+N}}$$

The results of maximising accuracy are as follows:

System	Max Accuracy	Decision Threshold	F1-Score
System 1	0.9990	1	0
System 2	0.9998	0.4694	0.8684
System 3	0.9990	1	0
System 4	0.9999	0.8367	0.9634

It is not an interesting operating point as accuracy is not a good performance metric in our case of imbalanced data. As can be seen in the table, simply setting the decision threshold to 1, which assigns every case to impostor, results in an accuracy of 99.9%, as given in system 1 and system 3, due to the fact that we have 99.9% of impostor cases in our data. Therefore, setting the decision threshold high would result in high accuracy by favouring the true prediction of the majority class - impostor class in our case, namely favouring high true negative (TN). It may falsely suggests high performance, while in fact the model fails to identify genuine cases (F1-score is 0).

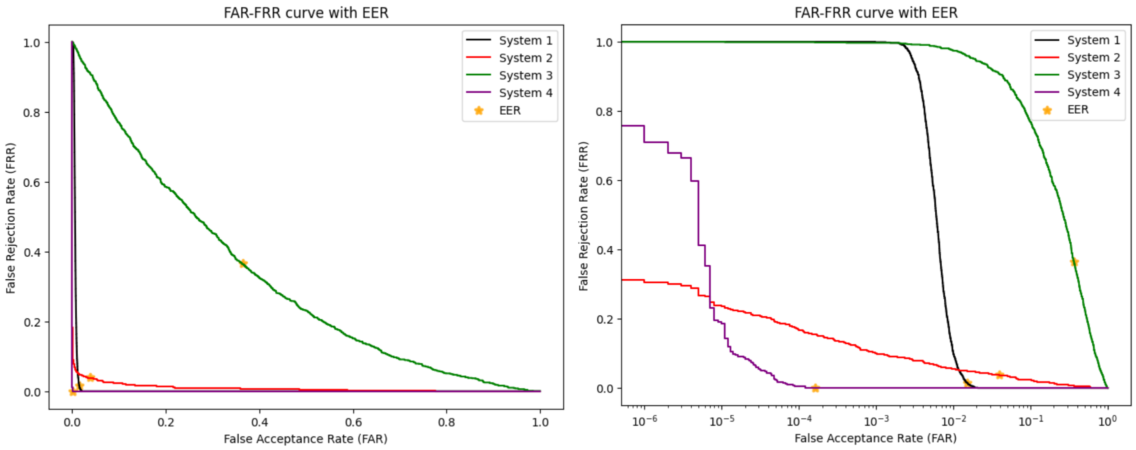
The results differ from the previous one because maximising accuracy and F1-score is a task of maximising metrics that are defined differently, potentially resulting in different thresholds when maximising them separately. It may also be the case that the same threshold maximises both, but this depends on the balance between precision and recall and the balance between sensitivity and specificity as well as data class balance. Accuracy is biased towards the majority class, while F1-score is not. Thus, in our case, F1-score is a better metric to use to evaluate models.

Q4:

System	ROC AUC
System 1	99.3515%
System 2	99.1636%
System 3	68.3340%
System 4	99.9992%

The Area Under the Receiver Operating Characteristic Curve (ROC AUC) is the area under the curve of the True Positive Rate (TPR) versus the False Positive Rate (FPR). The AUC score ranges from 0 to 1, with 0.5 representing a random baseline model. The higher the ROC AUC score, the better the overall performance of the model. System 4 performs best overall, achieving the highest AUC of 99.9992%, indicating a very good model overall. System 3 performs worst overall, achieving an AUC of 68.3340%, indicating a poor model as it is not much higher than 50%. System 1 and System 2 are good models overall achieving above 99%. System 1 has a slightly higher AUC than System 2, indicating slightly better overall performance. However, it is possible for a classifier with a lower AUC to outperform a classifier with a higher AUC in a particular region, as explained in Q2 when discussing the ROC curve results. System 1 and System 2 have similar AUC values with two very different ROC curves. Furthermore, ROC curves can be misleading on imbalanced datasets, giving an optimistic picture of the model on datasets with a class imbalance such as our case, as the ROC curve uses true negatives (TN) in the FPR.

System	EER	Decision Threshold
System 1	0.0151	0.7327
System 2	0.0391	0.3503
System 3	0.3652	0.5365
System 4	0.0002	0.8086



The Equal Error Rate (EER) is the point at which the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). It represents the rate at which the system incorrectly accepts impostors and incorrectly rejects genuine users equally. At this point, a given proportion of genuine users are rejected at the cost of accepting a given proportion of impostors. A lower EER indicates a better system, as it means that both types of error are less frequent when equal. Thus, System 4 performs best with an EER of 0.0002 at a decision threshold of 0.8086 and System 3 performs worst with an EER of 0.3652 at a threshold of 0.5365. System 1 and System 2 perform well, with System 1 having a slightly lower EER value being slightly better. EER is a good operating point to evaluate and compare models if our idea is to have equally good systems in terms of incorrectly accepting an unauthorised user and incorrectly rejecting an authorised user. However, if we prefer security or convenience, it is important to analyse FAR-FRR trade-off using DET curve or ROC curve, as done in Q2.

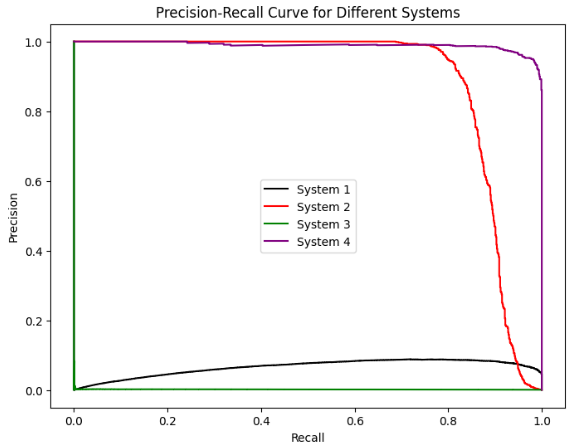
System	Min sum of FAR and FRR	FAR	FRR	Decision Threshold
System 1	0.0196	0.0186	0.0010	0.7235
System 2	0.0627	0.0117	0.0510	0.3778
System 3	0.7172	0.4362	0.2810	0.5166
System 4	0.0002	0.0002	0	0.8086

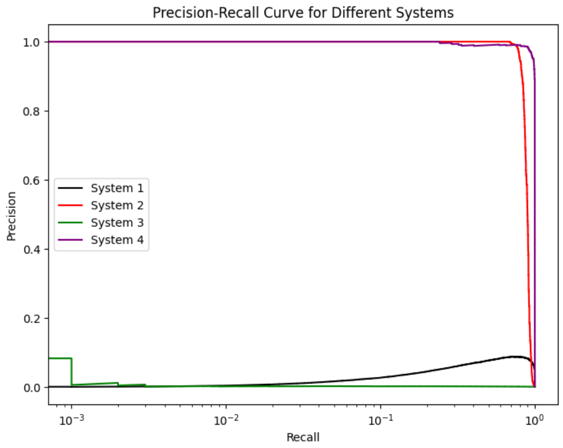
The decision threshold where the sum of FAR and FRR is minimized is a good operating point when there is no specific preference between the two errors, and the goal is to keep their sum as low as possible. System 4 performs best with a minimum sum of 0.0002 at a threshold of 0.8086. System 3 performs the worst with the lowest sum of FAR and FRR of 0.7172. System 1 and System 2 perform well as they are also able to achieve a low sum of FAR and FRR of 0.0196 and 0.0627 respectively.

The decision thresholds obtained by minimising the sum of FAR and FRR may not be similar to those minimising the total classification error. Minimising classification error is equivalent to maximising accuracy, as when minimising False Positives (FP) and False Negatives (FN), True Positives (TP) and True Negatives (TN) are maximised (and the number of observations remains the same). Minimising FP and FN is defined differently from minimising the sum of FAR=$\frac{FP}{FP+TN}$ and FRR=$\frac{FN}{FN+TP}$, because of the denominators. Minimising the raw FP and FN counts does not take into account the class imbalance, whereas minimising FAR and FRR does. Therefore, the decision threshold for minimal classification error may be different from that obtained by minimising the sum of FAR and FRR. The decision thresholds obtained by minimising the sum of FAR and FRR are also not the same as those obtained by calculating the EER, as FAR and FRR may be different.

The importance of FRR and FAR depends on whether we want a very secure or a very convenient system. For a very secure system, a lower FAR is preferred, even if the FRR is higher, as we do not want to accept impostors. For a very convenient system, a lower FRR is preferred, even if the FAR is higher, as we do not want to reject genuine users. Therefore, it is important to analyse the whole curve of the FRR-FRR curve or ROC curve, as done in Q2.

Q5:





The Precision-Recall curve shows the trade-off between precision and recall for different decision thresholds. The precision describes how good a model is at predicting the positive class and the recall is the ability of the classifier to find all the positive samples. The PR curve shows how precision changes as recall increases. The baseline of the curve is determined by the ratio of positives (P) to negatives (N), $\frac{P}{P+N}$, which in our case is 0.001. System 2 achieves the highest precision at lower recall levels. It is therefore possible to set a relatively stricter decision threshold of System 2 to have a model with high precision and lower recall, which may accept fewer genuine users, but most of its predicted labels are correct when compared to the ground truth, as minimizing False Positive ensuring higher precision. Thus, as suggested earlier, it is possible that System 2 may be suitable as a very secure system (e.g. banking application authentication). System 1 achieves slightly higher precision than System 2 at high recall values, making it more suitable for a more convenient system where we want to ensure that genuine users are not rejected, as minimizing False Negative ensuring higher recall. System 1 is however outperformed by System 4 at high recall values, so System 4 would still be the best choice for a convenient system. System 1 performs poorly overall, failing to achieve high precision at all thresholds, as struggles with minimizing FP - precision is class prior dependent and is thus influenced by the number of majority class of impostors, so good to comapre with a baseline of 0.001 in PR curve. System 4 performs best overall, being able to achieve both high precision and high recall, making it ideal for systems that require both convenience and security. System 3 performs worst, achieving both low precision and low recall at every threshold, and is unreliable for both security and convenience.

In our case of an imbalanced dataset, the PR curve is preferred to the ROC curve. ROC curves give an optimistic picture of the model on imbalanced datasets, because true negatives are used in the False Positive Rate in the ROC curve. Consequently, the use of an ROC curve with an imbalanced data can be misleading and lead to incorrect interpretations of the model's performance. As Precision-Recall curves do not take true negatives into account, PR curve is better to use for highly imbalanced datasets where we are optimising for the positive class of our case, giving a more honest representation of the result than the ROC curve. In addition, the PR curve also takes into account the imbalance of our data when determining the baseline random model of 0.001, exposing the imbalance in the data.

System	PR-curve AUC	Average Precision
System 1	0.0653	0.0654
System 2	0.8877	0.8877
System 3	0.0019	0.0020
System 4	0.9893	0.9893

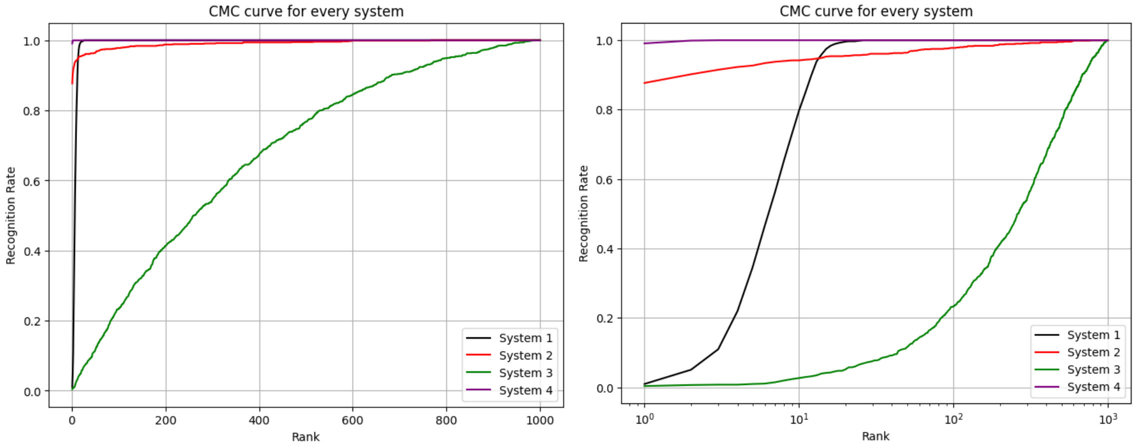
The PR-curve AUC measures the area under the Precision-Recall curve. The higher the AUC, the better the overall performance of the system in terms of precision and recall combined. For a perfect classifier, AUC = 1. The best overall performance is given by System 4, which has a very high AUC value of 98.93%, meaning that it achieves both high precision and recall across the thresholds. The worst overall performance is given by System 3, which achieves a very poor AUC of only 0.19%, which is only slightly above the baseline (0.1%), indicating that it fails to meaningfully distinguish between genuine and impostor users. System 2 shows strong overall performance with a AUC of 88.77%, meaning that it performs well across most thresholds. As discussed earlier, it can even outperform System 4 in certain applications of very secure systems within regions of a stricter threshold, but overall System 4 remains better. System 1 has a relatively low AUC of 6.53%, indicating that while it may be better than System 2 at high recall levels, its overall performance is weak compared to System 4 and System 2.

Average Precision (AP) is an alternative to AUC for summarising a precision-recall curve. It is calculated as the weighted average of the precision achieved at each threshold, with the increase in recall from the previous threshold as the weight:

$$AP=\sum_{n}(R_n-R_{n-1})P_n,$$

where P_n and R_n are the precision and recall at the nth threshold. AP therefore only uses observed recall-precision pairs and is not interpolated, which differs from calculating the area under the precision-recall curve with the trapezoidal rule, which uses linear interpolation and can be too optimistic. However, in our case of rather smooth precision-recall curves of systems with gradual transitions between values, the Average Precision and PR curve AUC values are almost identical for each system.

Q6:



CMC curve for every system

CMC curve for every system

System	Rank-1 Recognition Rate
System 1	0.01
System 2	0.877
System 3	0.004
System 4	0.991

Our dataset consists of 1000 test samples (probes), each with 1000 entries (galleries), with only one genuine user per probe. The Cumulative Match Characteristic (CMC) curve plots the probability that a correct identification is returned within the top-x ranked matching scores of a sample. Given the data, the x-axis ranges from 1 to 1000, representing the top-k ranks, and the y-axis represents the Recognition Rate, which is the average fraction of probes where the genuine match is found within the top-k ranks. The Rank-1 Recognition Rate represents the value on the CMC curve where rank equals 1. This metric indicates that System 4 performs the best, being able to detect 99.1% of genuine users within the highest score. System 2 achieves 87.7%, which is also a good result but slightly worse. Systems 1 and 3 perform extremely poorly, achieving 1% and 0.4% respectively. As rank increases, all systems clearly accept more and more users, but System 1 in particular shows the sharpest increase, outperforming System 2 beyond about rank 15. This suggests that it is eventually accept genuine users, but is struggling to recognise them within the highest top ranks. System 4 converges quickly to 100%. System 3 performs the worst with its recognition rate growing slowly as rank increases.

Q7:

The previous questions introduced the F1-score, where the relative contribution of precision and recall to the F1-score is equal. If we want to focus more on recall or precision, the F-beta score is better instead, which is the weighted harmonic mean of precision and recall:

$$F_{\beta}=\frac{(1+\beta^2)*\text{Precision}*\text{Recall}}{\beta^2*\text{Precision}+\text{Recall}}=\frac{(1+\beta^2)*TP}{(1+\beta^2)*TP+FP+\beta^2 FN}$$

The β parameter represents the ratio of recall importance to precision importance: β>1 gives more weight to recall and should be used if we are focusing on the convenience of the system, while β<1 favours precision and should be used if we are focusing on the security of the system.

System	Max F beta=0.1 (threshold)	Max F beta=20 (threshold)
System 1	0.0874 (0.7551)	0.9472 (0.7347)
System 2	0.9954 (0.5102)	0.9228 (0.3878)
System 3	0.0326 (0.9388)	0.3428 (0.5102)
System 4	0.9870 (0.8776)	0.9993 (0.7959)

Using β values of 0.1 and 20 in the F-beta score, we prioritise precision (security) and recall (convenience) respectively to extreme degrees. It confirms that System 2 can be a good very secure model of the highest F-beta 0.1 of 0.9954 at threshold of 0.5102. System 1 can be a good classifier for convenient system as it achieves high F-beta 20 value of 0.9472 at threshold of 0.7347. System 4 is the best overall, being very good at both metrics. System 3 is the worst as it is variably bad on both metrics. The decision threshold also depends on β, as a higher threshold (more conservative) is for a lower β value, where we prioritise security. We should consider what β value we want to set.

Furthermore, the Geometric Mean (GM) of sensitivity and specificity can be introduced because, unlike accuracy, it is not biased by imbalanced classes, since sensitivity itself depends only on the positive class and specificity depends only on the negative class:

$$GM=\sqrt{\text{sensitivity}*\text{specificty}}=\sqrt{\frac{TP}{TP+FN}*\frac{TN}{TN+FP}}$$

It is important to note that this metric treats sensitivity and specificity as equally important. In case we need a more convenient (sensitivity-focused) or a more secure (specificity-focused) system, it is important to consider the weighted geometric mean. However, if there is a need to evaluate a model and set a decision threshold to find a system that achieves both high sensitivity and specificity with equal contribution, this is an effective summary metric:

System	Max GM (threshold)
System 1	0.9884 (0.7143)
System 2	0.9669 (0.3673)
System 3	0.6371 (0.5306)
System 4	0.9999 (0.7959)

GM gives an optimistic picture of the model on imbalanced datasets of our case, because true negatives are used in specificity, and the negative class of impostors is the large majority of our data. However, since GM is the geometric mean, both sensitivity and specificity must be high for the metric to be high. It can be concluded that system 4 performs best overall, system 3 performs worst overall, and system 1 and system 2 both achieve a high GM, with system 1 slightly higher.

However, as GM and F-beta scores are two summary metrics, it is always important to consider analysing the curve as well, such as the ROC or DET curve for GM and the PR curve for F-beta scores, as this gives a better overview of performance over decision thresholds, as was done in the previous question discussing usability and performance for each system. Nevertheless, the metrics discussed give a good idea of performance and threshold determination.