# Uncertainty in Machine Learning \ Assignment 3

Massimiliano Falzari (s3459101)

July 1, 2022

## Contents



## 1 Out of Distribution Detection

I have used an ensemble (10 NN) for epistemic uncertainty quantification and the machine learning model is a standard Convolutional NN

The result can be seen in the table 1. The model that has the best performance is the ensemble model. However, the difference is not too evident.

From the results, we can observe that the AUC is higher in the ensemble than in the baseline model. The accuracies are almost identical both for ID and OD. In 1 and 4 we can see that the Baseline model has a higher peak around 0 for in distribution data. This is because this model does not use any Uncertainty quantification method and therefore is more confident resulting in lower entropy. For out-of-distribution data, there is no clear difference in entropy between the two models. The same is true for the max probability graphs (2,5), there is no clear difference between the two models. The ROC curves (3,6) are also quite similar, however, from the AUC value in 1 we can see that the ensemble method is slightly better than the baseline model.

Table 1: Results for Out of Distribution Detection

| model | ID | OD | ID accuracy | OD accuracy | AUC |
|---|---|---|---|---|---|
| ensemble | mnist | $\text{fashion}_{\text{mnist}}$ | 0.9044 | 0.1145 | 0.853628295 |
| baseline | mnist | $\text{fashion}_{\text{mnist}}$ | 0.9025 | 0.1027 | 0.846553755 |

## 2 Reverse OOD Detection

No, we do not obtain the same performance according to AUC. This is because of different reasons. First, the training process is stochastic so we should not expect the same results. Second, the features in the two datasets are different and therefore, in general, it is not the case that reversing the dataset in an OOD framework results in similar results.

In 7 we can see a quite different pattern both for in and out distribution data with respect to 1. This is probably because Fashion mnist has more features/ more information and therefore it is harder to learn compared to the standard mnist dataset. Indeed, we can see that the entropy is considerably higher both for in and out distribution data, which means less confidence. Similar pattern can be seen in 8. For in-distribution data the model is not really confident. For out-of-distribution data, the distribution has a really high peak of 0.2 which is completely different in 2. This again shows that Fashion mnist is harder to learn or it needs different architecture/hyperparameters.

Table 2: Results for Reverse Out of Distribution Detection

| model | ID | OD | ID accuracy | OD accuracy | AUC |
|---|---|---|---|---|---|
| ensemble | $\text{fashion}_{\text{mnist}}$ | mnist | 0.7565 | 0.0715 | 0.827956175 |
| baseline | $\text{fashion}_{\text{mnist}}$ | mnist | 0.7471 | 0.0575 | 0.821769015 |

## 3 Calibration

The calibration errors are all really close to each other. The one with the lowest calibration error is the ensemble trained on the $\text{Fashion}_{\text{mnist}}$.

If we compare the 3 reliability plot 13 24 and 35, we can see that in general they are all underconfident. However, the 13 is slightly overconfident when the accuracy is low. Something similar can be seen in 24. The number of bins used was 20.

Finally, for the reliability plot per class, they can be read as a standard reliability plot taking into account only one class at the time. Therefore, if the red line is above the black one, then the model is underconfident; if the red line is below, then the model is overconfident. To construct these plots the idea was to use as prediction always the class for which we are doing the reliability plot. The true classes were the same as a normal plot and the confidence was the confidence for that particular class predicted by the model.

Table 3: Calibration errors

| model | ID | Calibration Error |
|---|---|---|
| ensemble | mnist | 0.11720183312464813 |
| baseline | mnist | 0.11773799104029882 |
| ensemble | $\text{fashion}_{\text{mnist}}$ | 0.10139520830329869 |

## 4 PLOTS

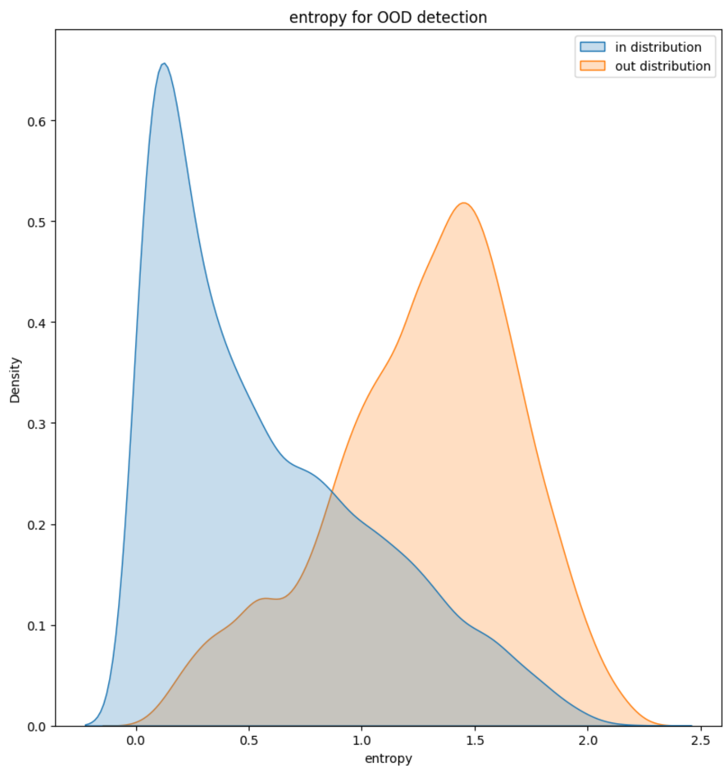


Figure 1: Ensemble (ID mnist) Entropy

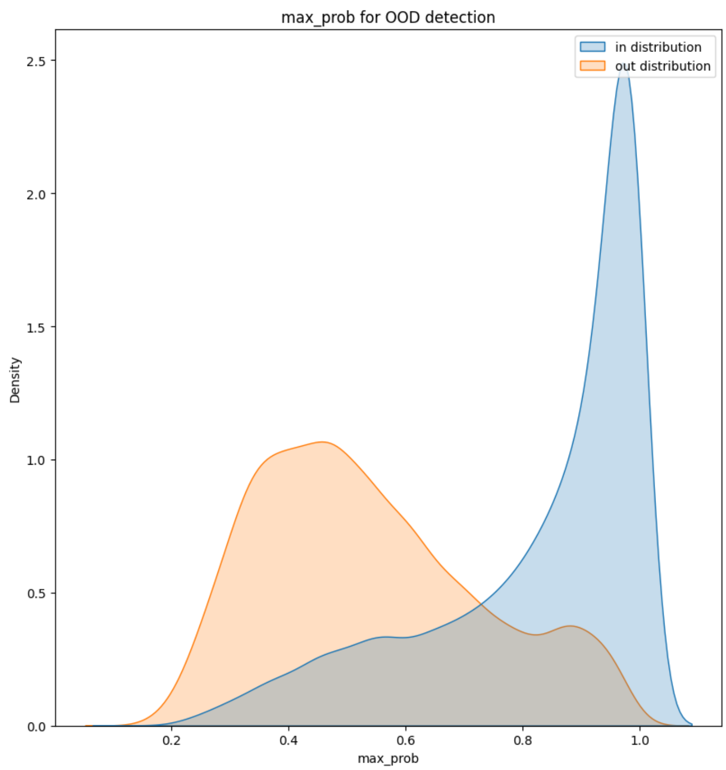


Figure 2: Ensemble (ID mnist) Max Probabilities

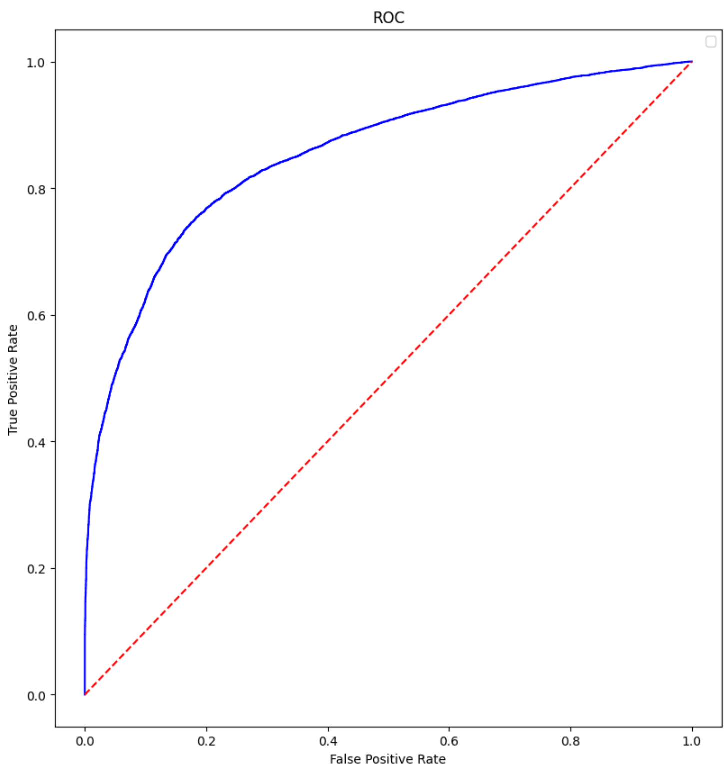


Figure 3: Ensemble (ID mnist) ROC curve

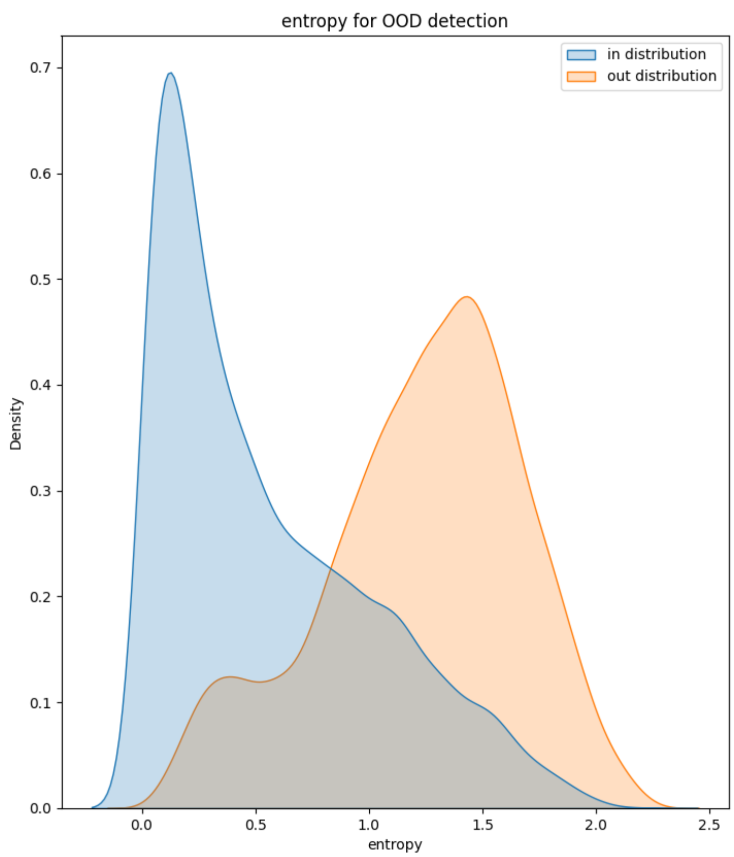


Figure 4: Baseline (ID mnist) Entropy

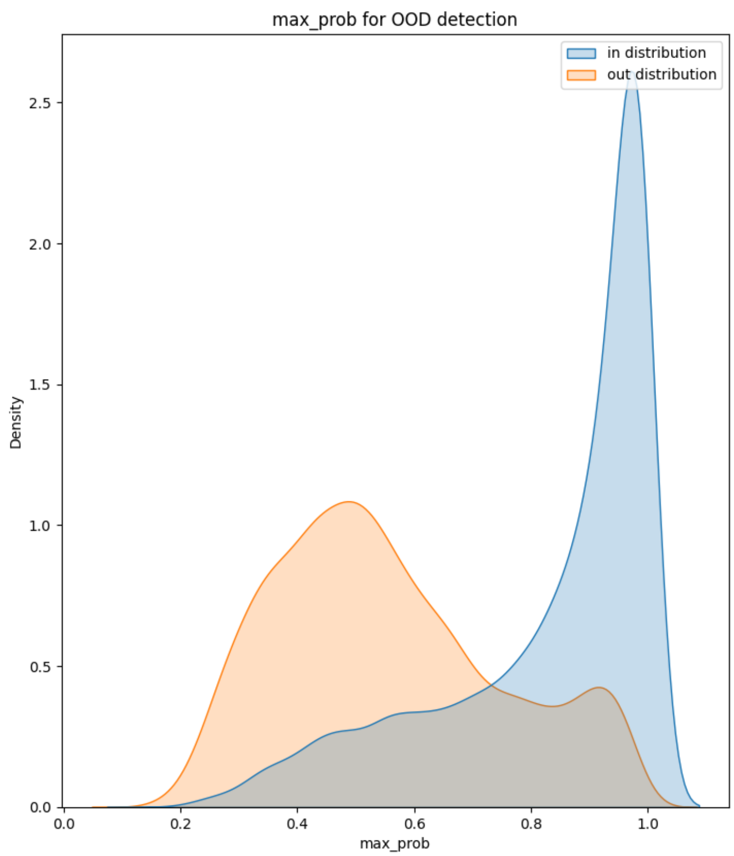


Figure 5: Baseline (ID mnist) Max Probabilities

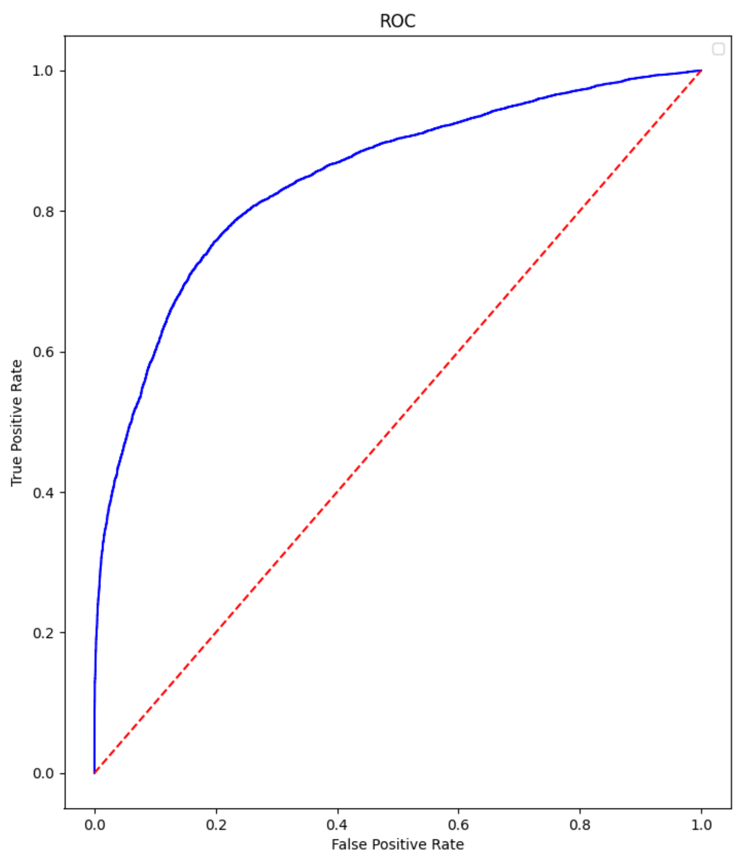


Figure 6: Baseline (ID mnist) ROC curve

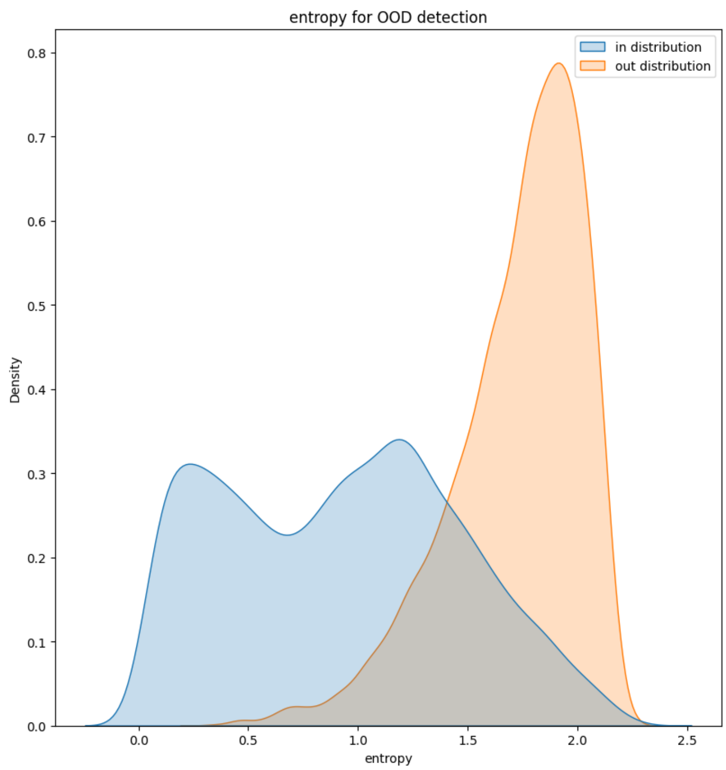


Figure 7: Ensemble (ID $\text{fashion}_{\text{mnist}}$) Entropy

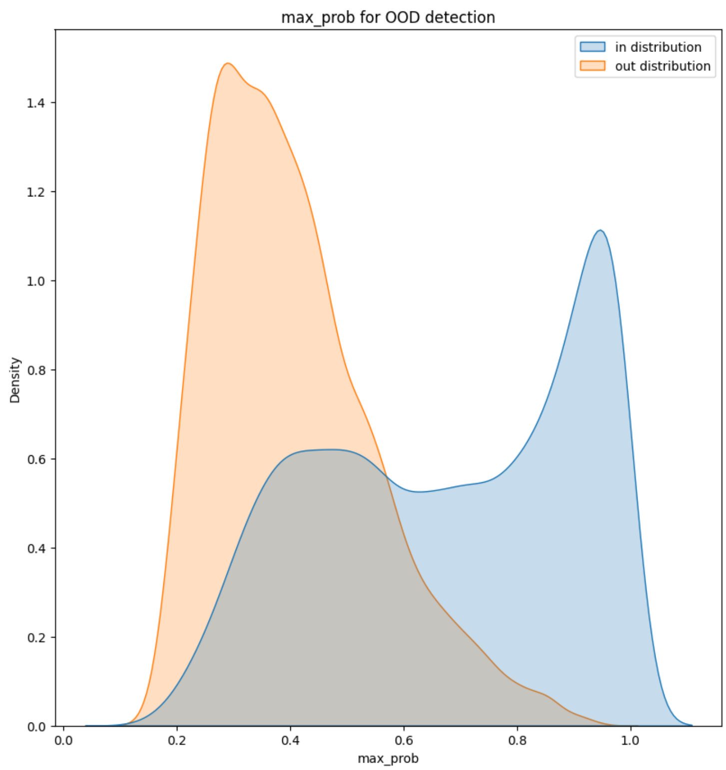


Figure 8: Ensemble (ID $\text{fashion}_{\text{mnist}}$) Max Probabilities

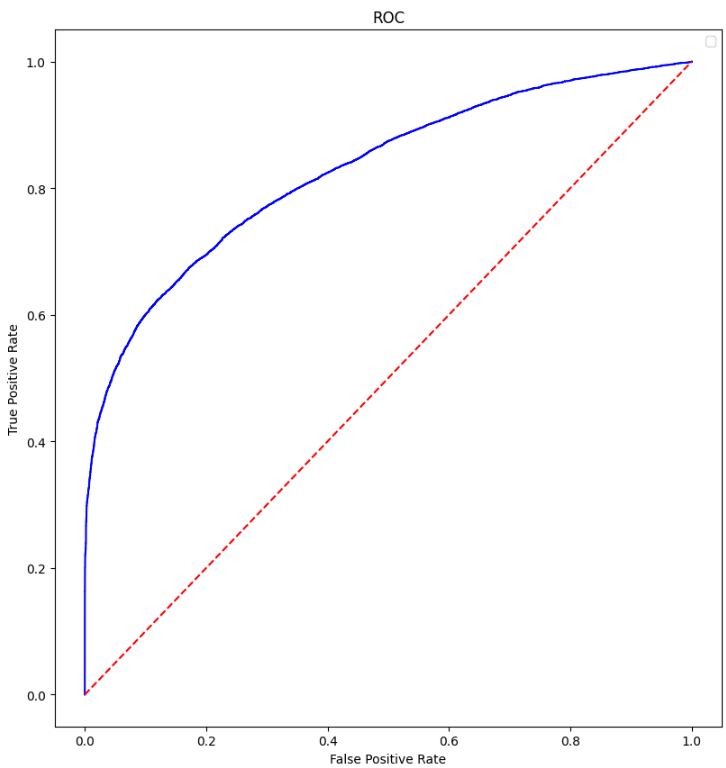


Figure 9: Ensemble (ID $\text{fashion}_{\text{mnist}}$) ROC curve

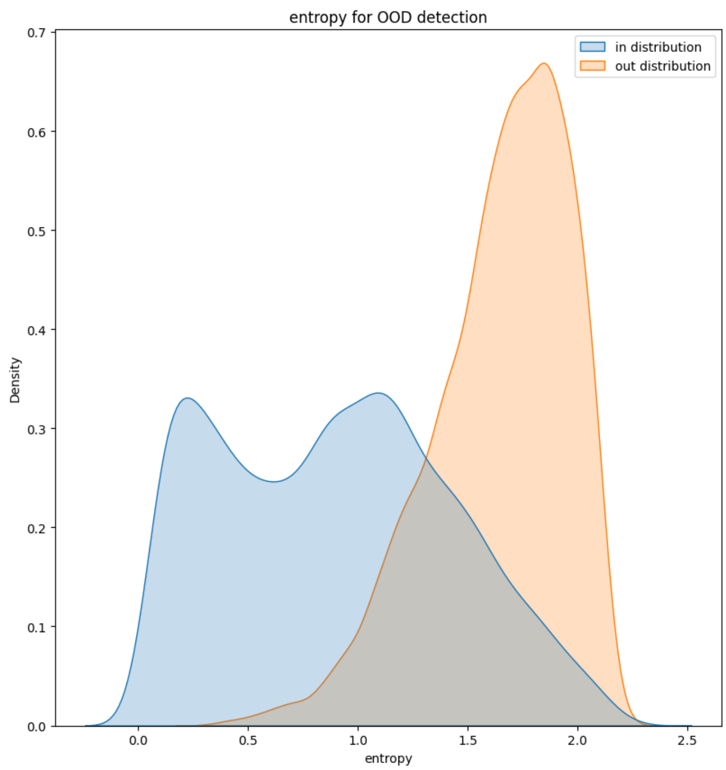


Figure 10: Baseline (ID $\text{fashion}_{\text{mnist}}$) Entropy

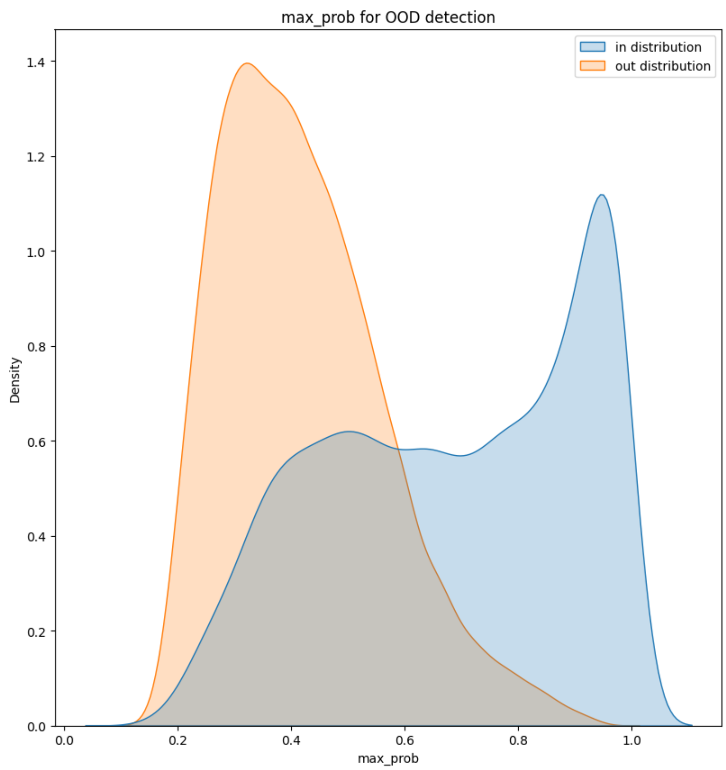


Figure 11: Baseline (ID $\text{fashion}_{\text{mnist}}$) Max Probabilities

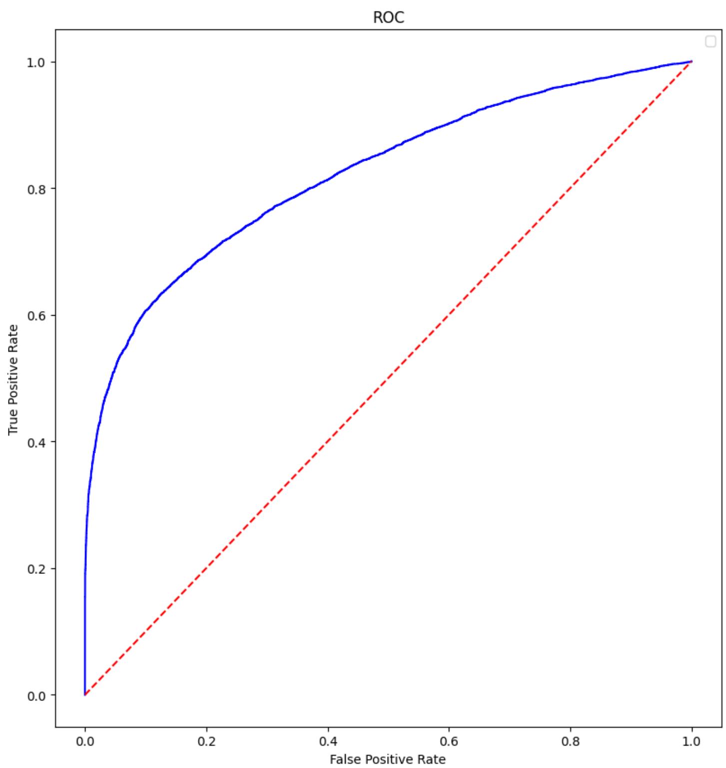


Figure 12: Baseline (ID $\text{fashion}_{\text{mnist}}$) ROC curve

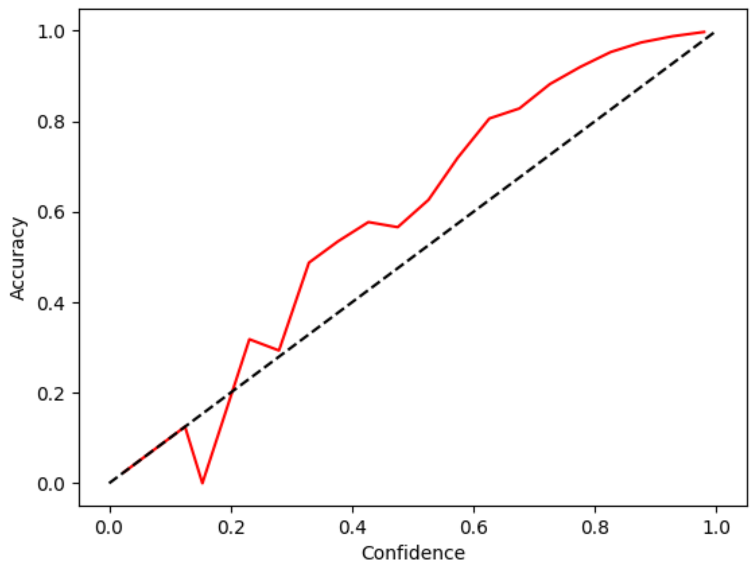


Figure 13: Ensemble Reliability Plot (ID mnist)

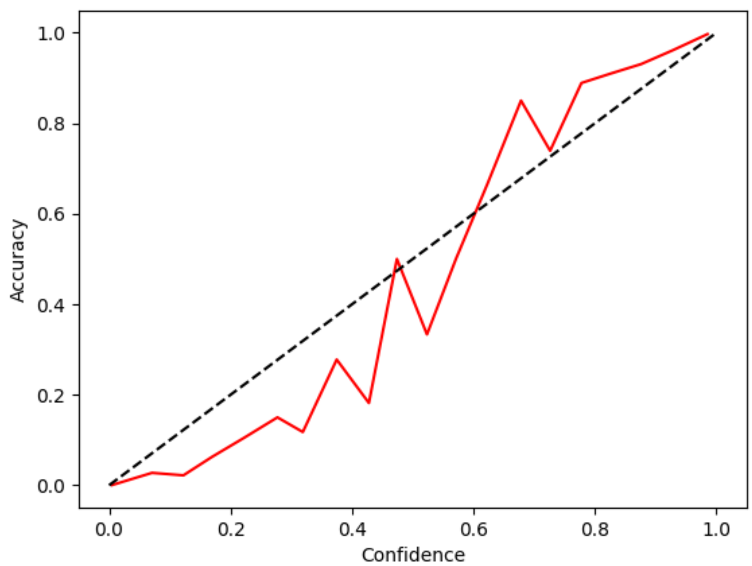


Figure 14: Ensemble Reliability Plot for class 0 (ID mnist)

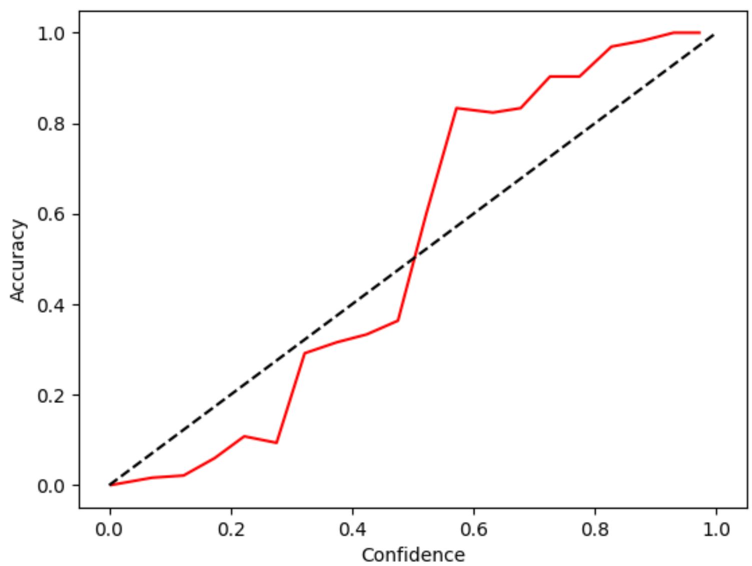


Figure 15: Ensemble Reliability Plot for class 1 (ID mnist)

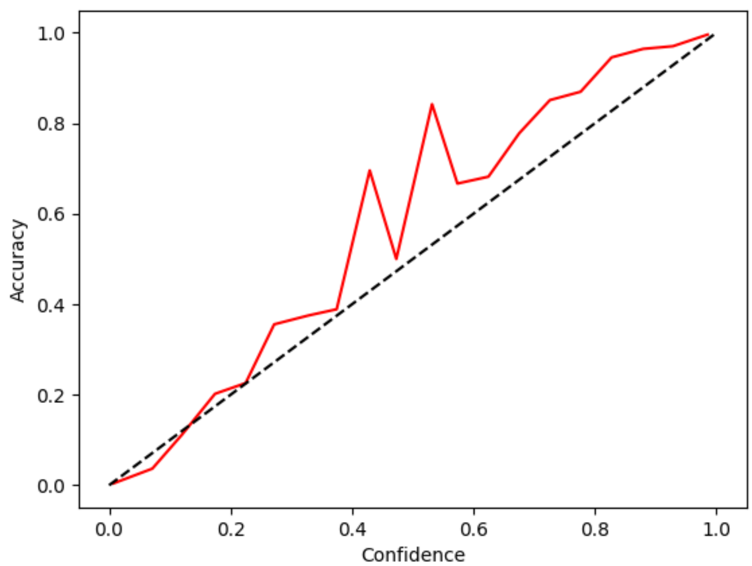


Figure 16: Ensemble Reliability Plot for class 2 (ID mnist)

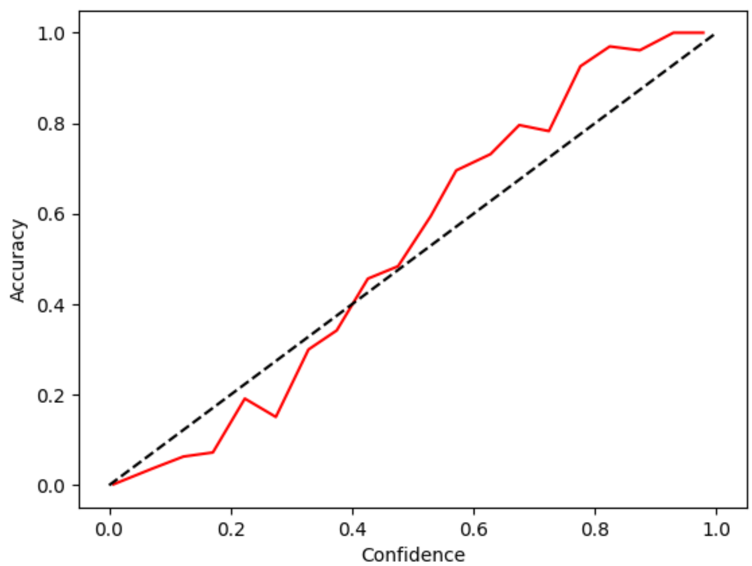


Figure 17: Ensemble Reliability Plot for class 3 (ID mnist)

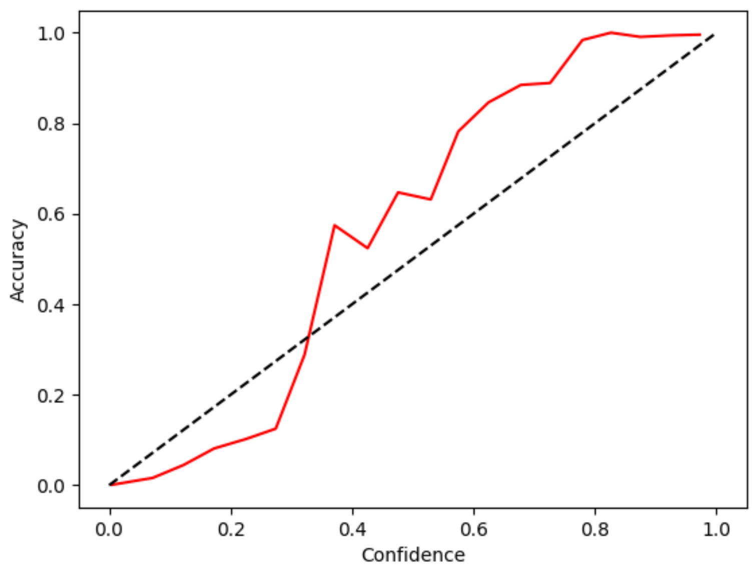


Figure 18: Ensemble Reliability Plot for class 4 (ID mnist)

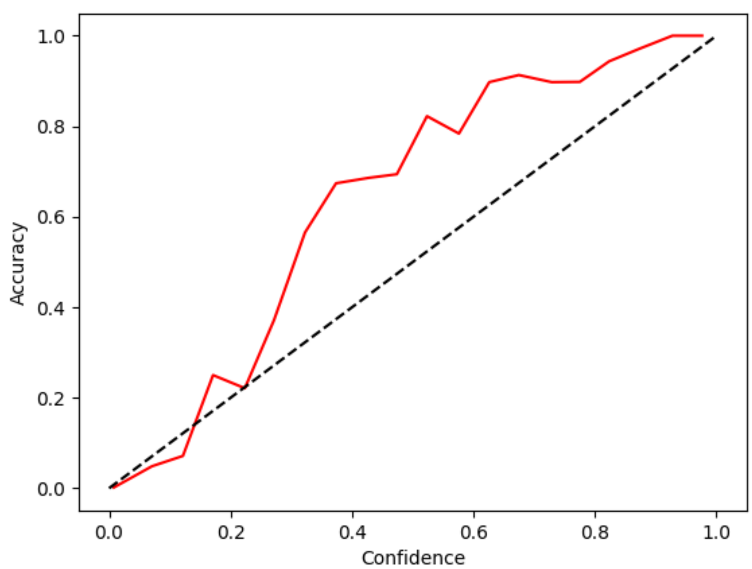


Figure 19: Ensemble Reliability Plot for class 5 (ID mnist)

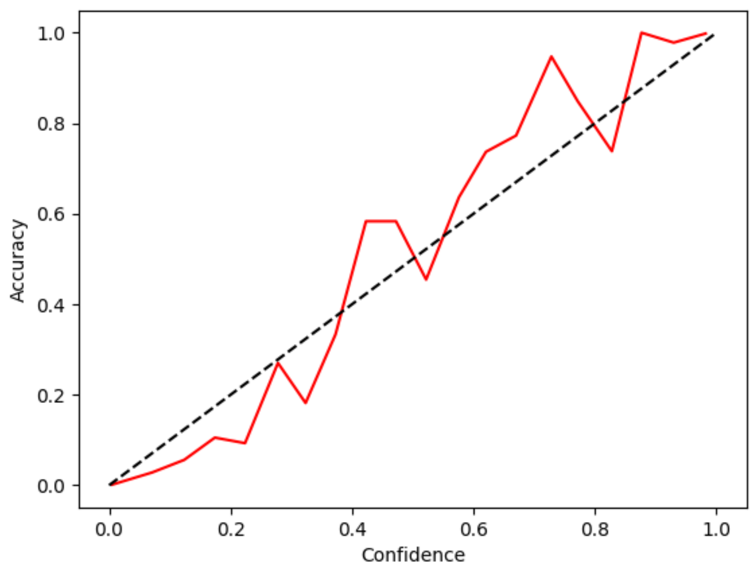


Figure 20: Ensemble Reliability Plot for class 6 (ID mnist)

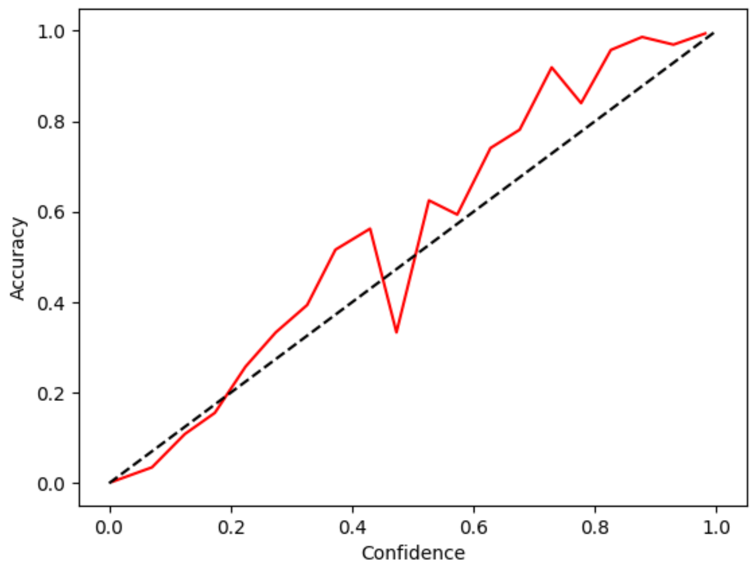


Figure 21: Ensemble Reliability Plot for class 7 (ID mnist)

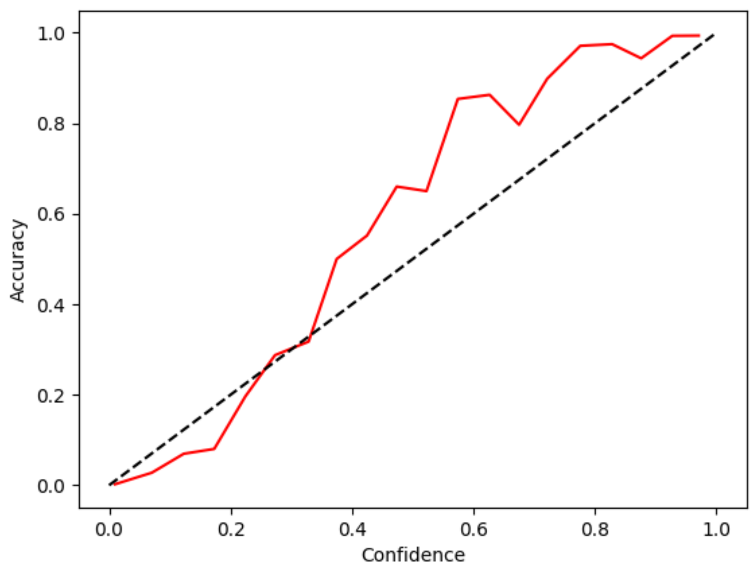


Figure 22: Ensemble Reliability Plot for class 8 (ID mnist)

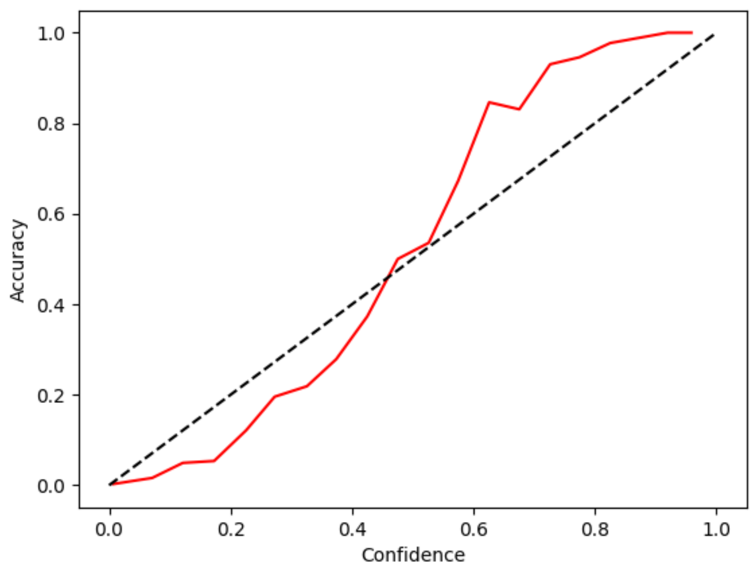


Figure 23: Ensemble Reliability Plot for class 9 (ID mnist)

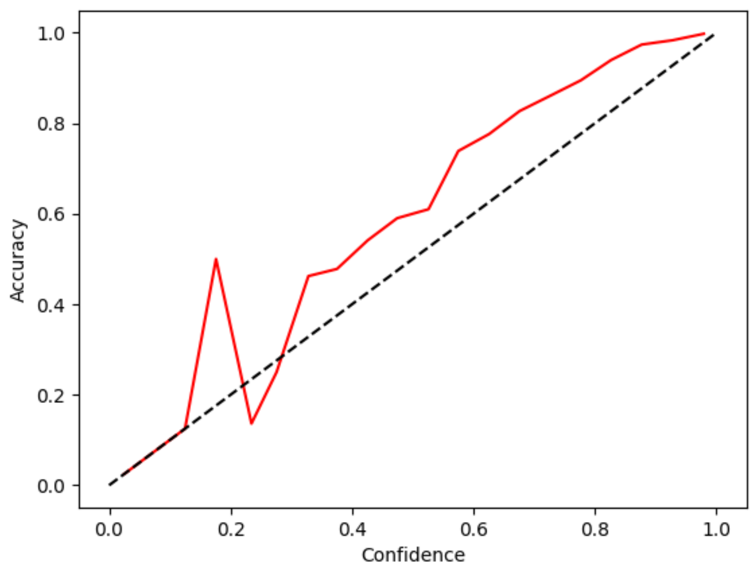


Figure 24: Baseline Reliability Plot (ID mnist)

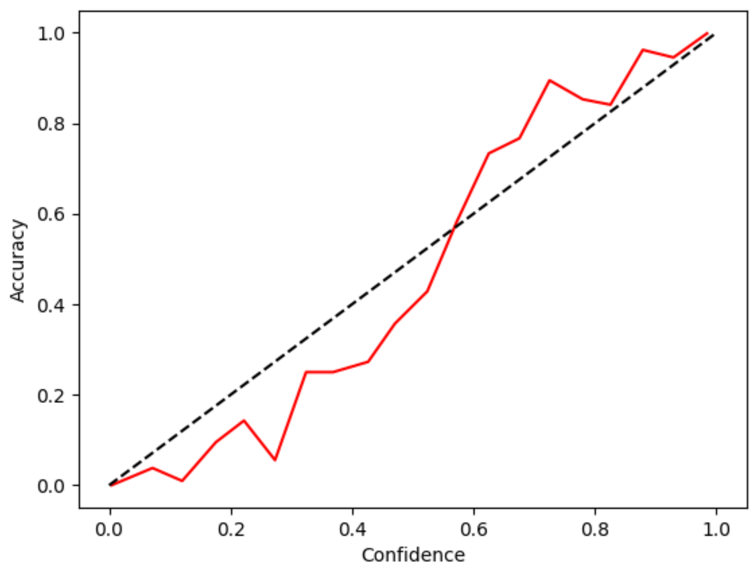


Figure 25: Baseline Reliability Plot for class 0 (ID mnist)

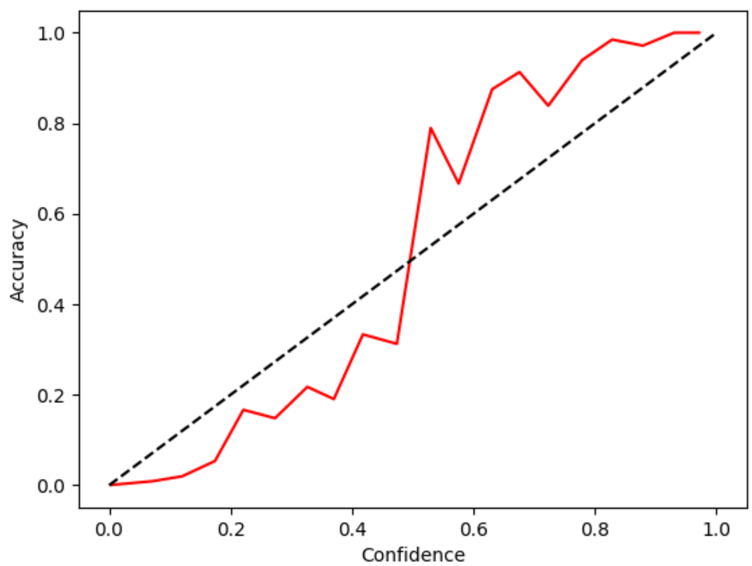


Figure 26: Baseline Reliability Plot for class 1 (ID mnist)

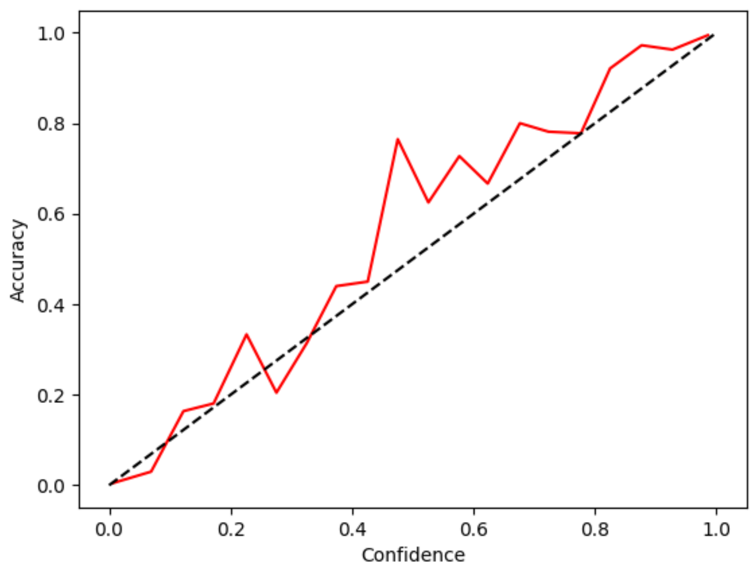


Figure 27: Baseline Reliability Plot for class 2 (ID mnist)

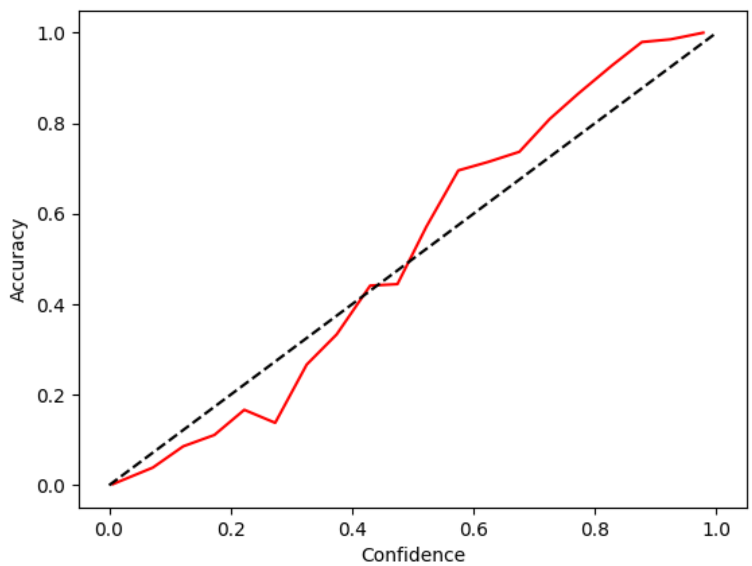


Figure 28: Baseline Reliability Plot for class 3 (ID mnist)

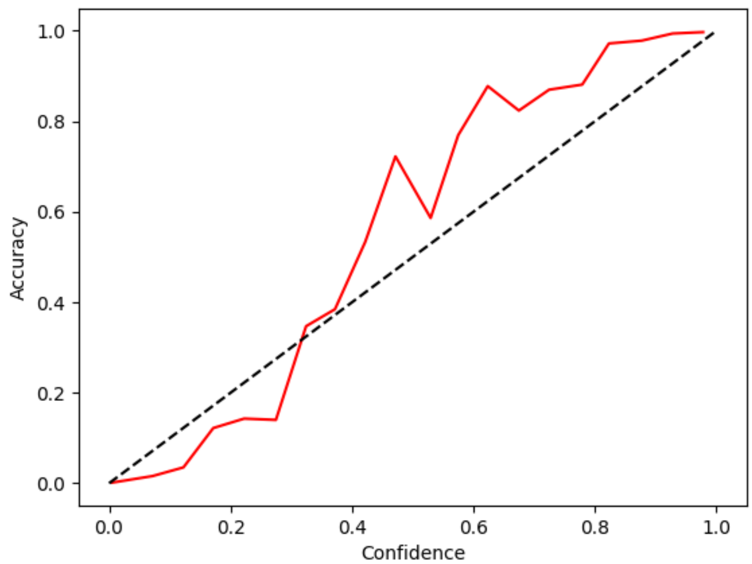


Figure 29: Baseline Reliability Plot for class 4 (ID mnist)

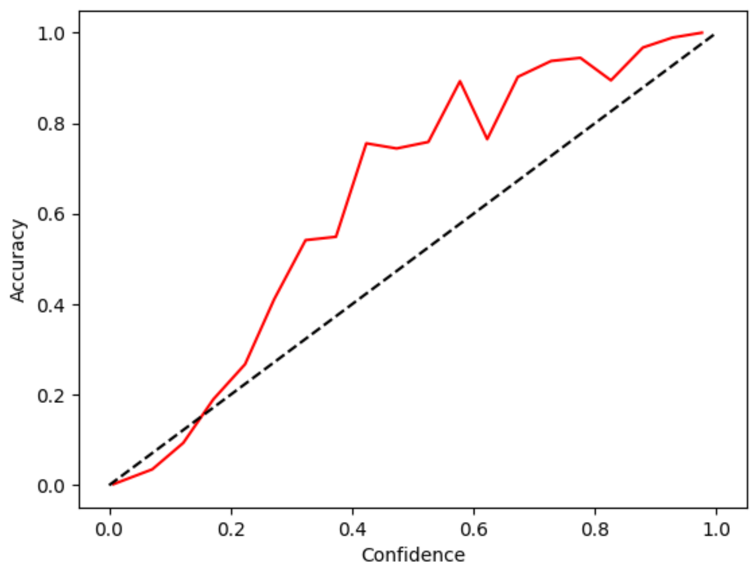


Figure 30: Baseline Reliability Plot for class 5 (ID mnist)

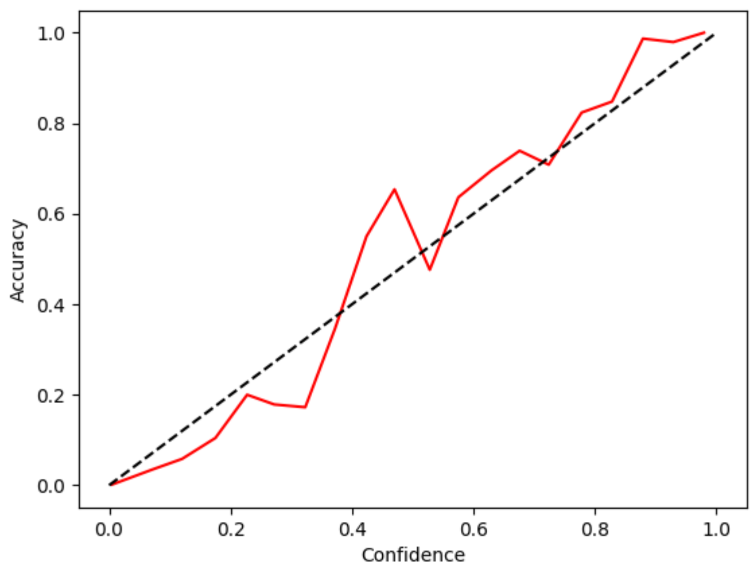


Figure 31: Baseline Reliability Plot for class 6 (ID mnist)

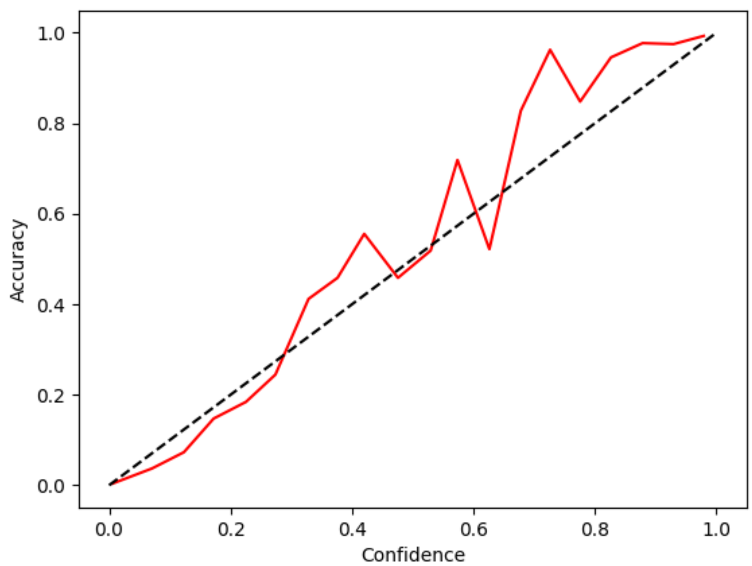


Figure 32: Baseline Reliability Plot for class 7 (ID mnist)

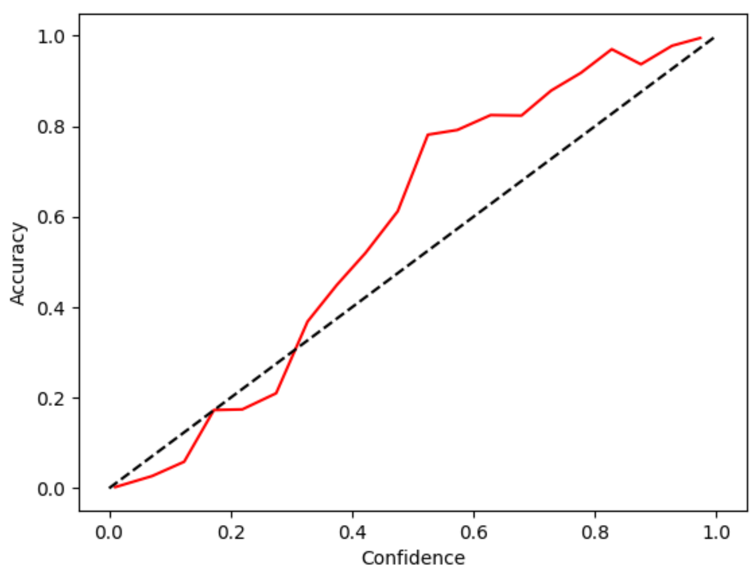


Figure 33: Baseline Reliability Plot for class 8 (ID mnist)

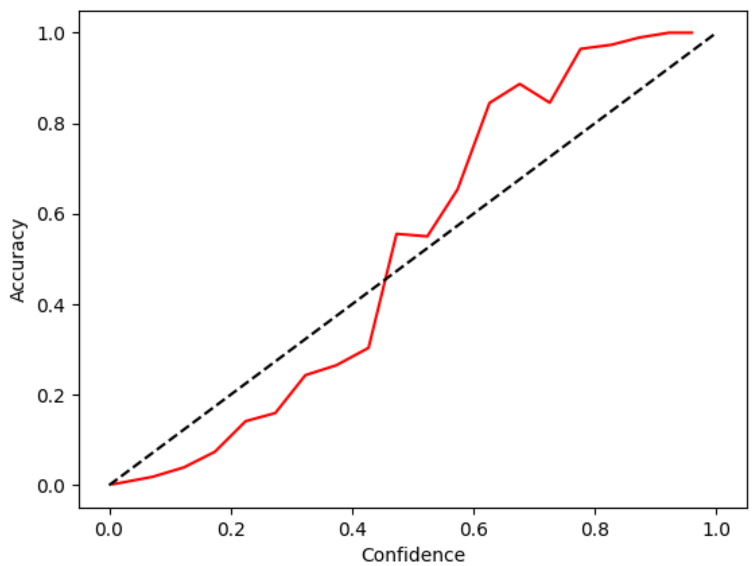


Figure 34: Baseline Reliability Plot for class 9 (ID mnist)

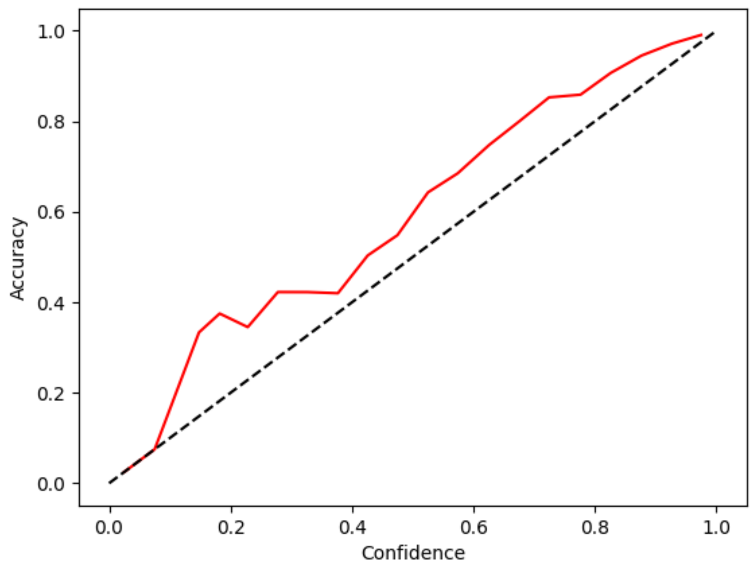


Figure 35: Ensemble Reliability Plot (ID fash)

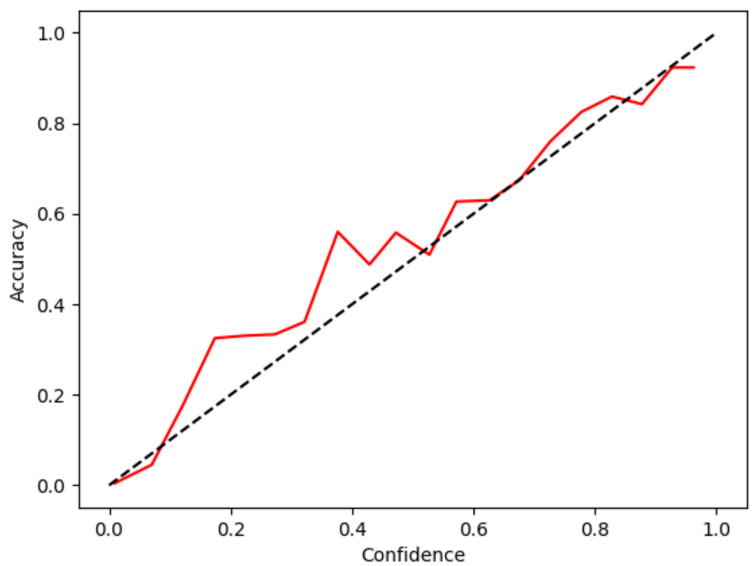


Figure 36: Ensemble Reliability Plot for class 0 (ID fash)

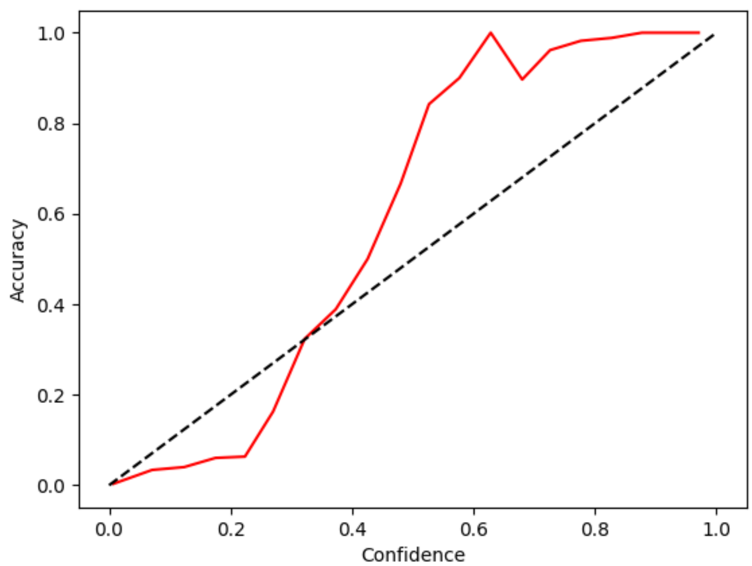


Figure 37: Ensemble Reliability Plot for class 1 (ID fash)

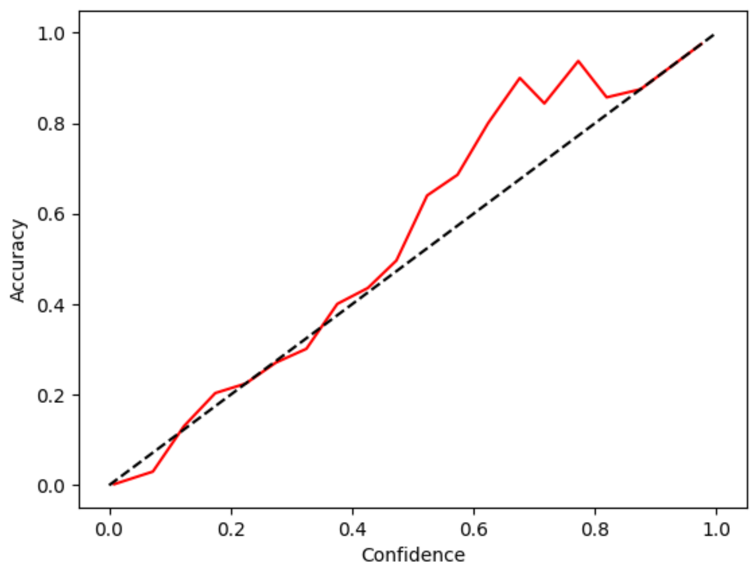


Figure 38: Ensemble Reliability Plot for class 2 (ID fash)

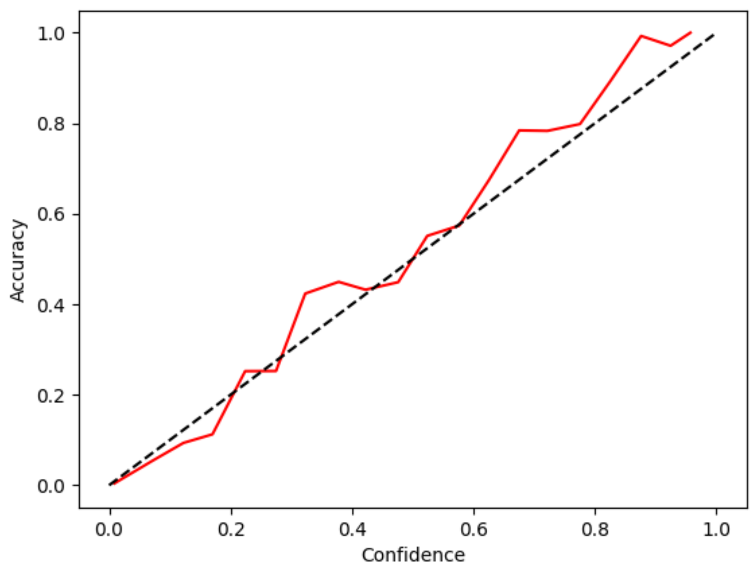


Figure 39: Ensemble Reliability Plot for class 3 (ID fash)

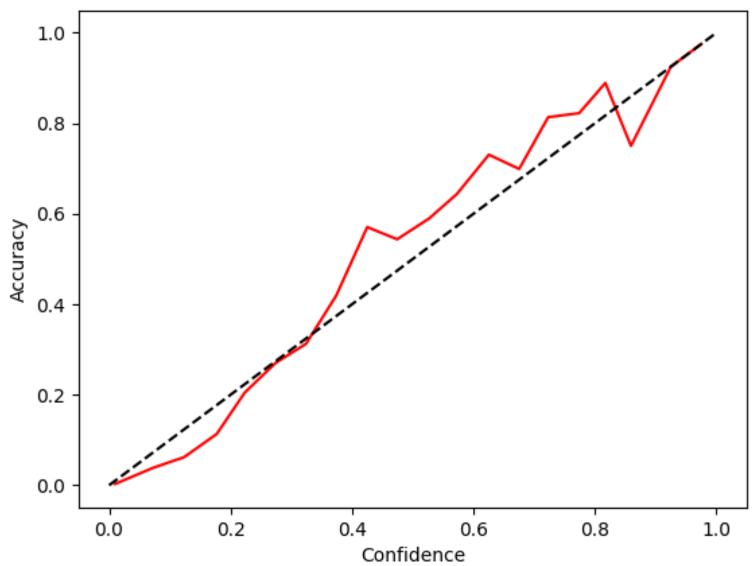


Figure 40: Ensemble Reliability Plot for class 4 (ID fash)

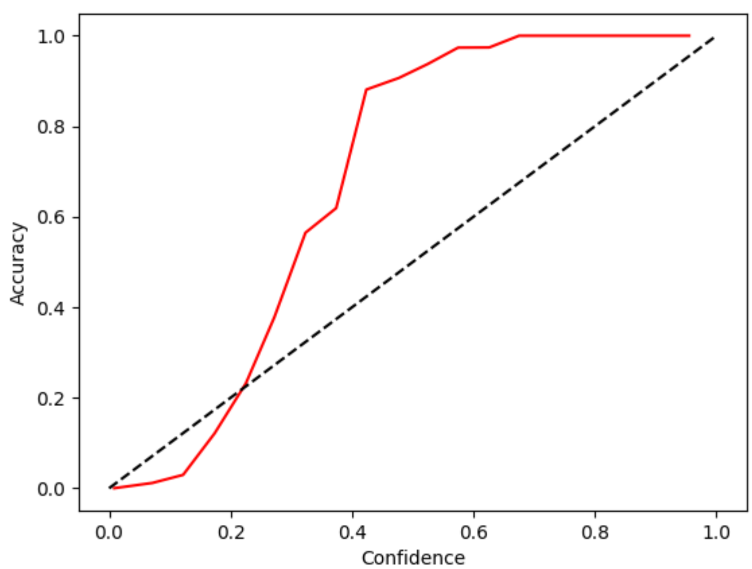


Figure 41: Ensemble Reliability Plot for class 5 (ID fash)

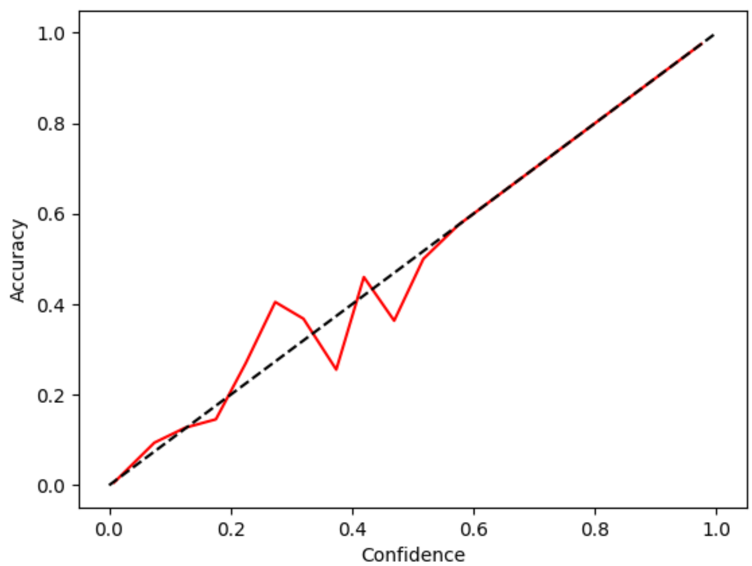


Figure 42: Ensemble Reliability Plot for class 6 (ID fash)

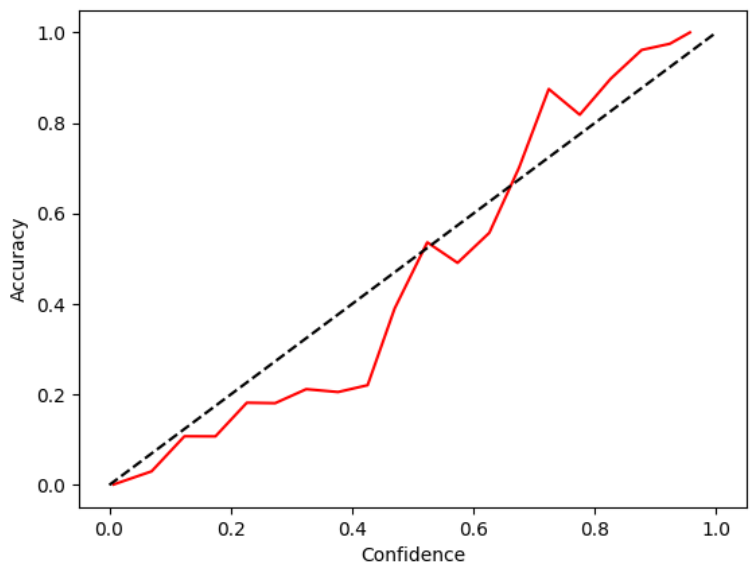


Figure 43: Ensemble Reliability Plot for class 7 (ID fash)

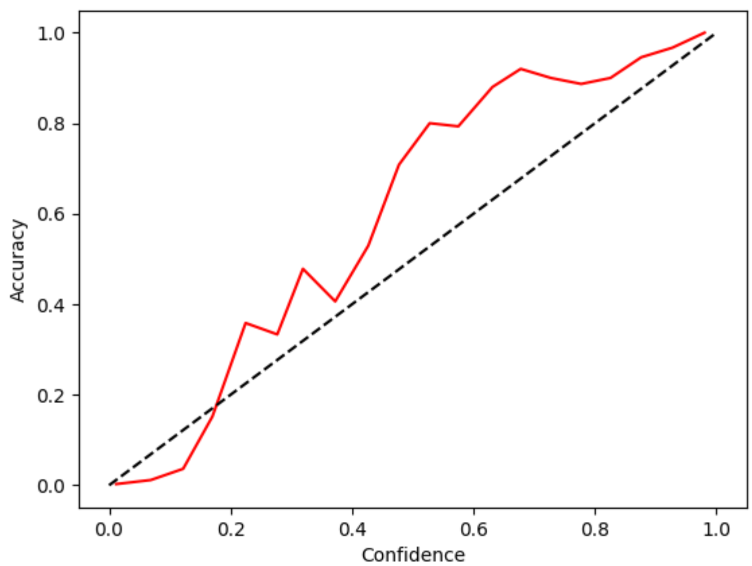


Figure 44: Ensemble Reliability Plot for class 8 (ID fash)

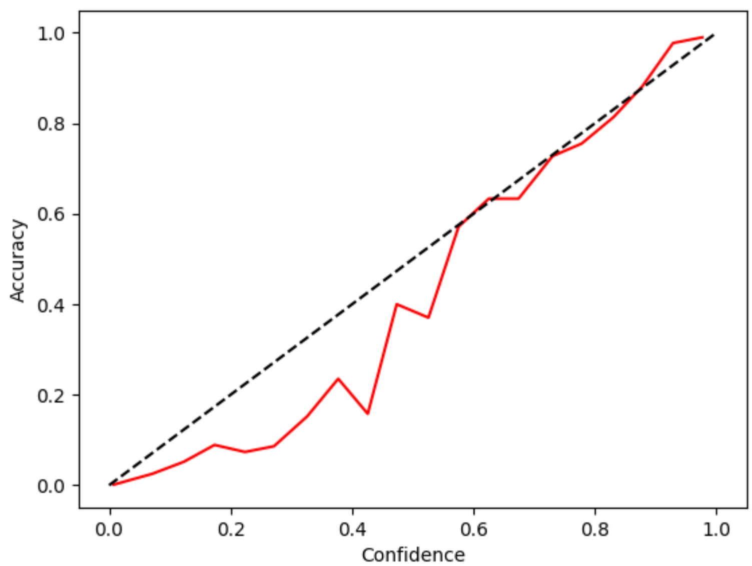


Figure 45: Ensemble Reliability Plot for class 9 (ID fash)