European Championship in Trustworthy AI
Prof. Dr. Adam Dziedzic and Prof. Dr. Franziska Boenisch
TASK #6: Dataset Inference for Vision APIs

Task: Execute a Dataset Inference Attack and achieve the highest attack success (TPR@FPR=0.01) and area under curve (AUC).

The goal of a Dataset Inference Attack is to determine whether a given dataset was part of the dataset the model in question was trained on. In this task, your goal is to launch a successful Dataset Inference Attack on an Image Classifier, only being given access to the final logits.

As an attacker, you get access to:

- Query access to a resnet18 Image Classifier that returns logits for an input image
- 1000 sets of 100 Images per set (available here)

Additional information that you should know

The data subsets are the datasets on which you should conduct dataset inference. All images in a given subset are either all part of the training data, or none of them are. 50% of the subsets were used for training.

What you don't have access to:

- The training dataset used to train Resnet-18 model (obviously)
- The architecture and parameters of the Restnet-18 model

What's your task:

- Your goal is to determine which subsets of data were part of the training data, and assigning these subsets the value 1, and all others the value 0.

Evaluation Metrics

- TPR@FPR=0.01
- AUC

How to submit your results?

Submit a CSV-File containing only two columns: One with the ordered IDs of the subsets (0-999), and one with the membership indicator.