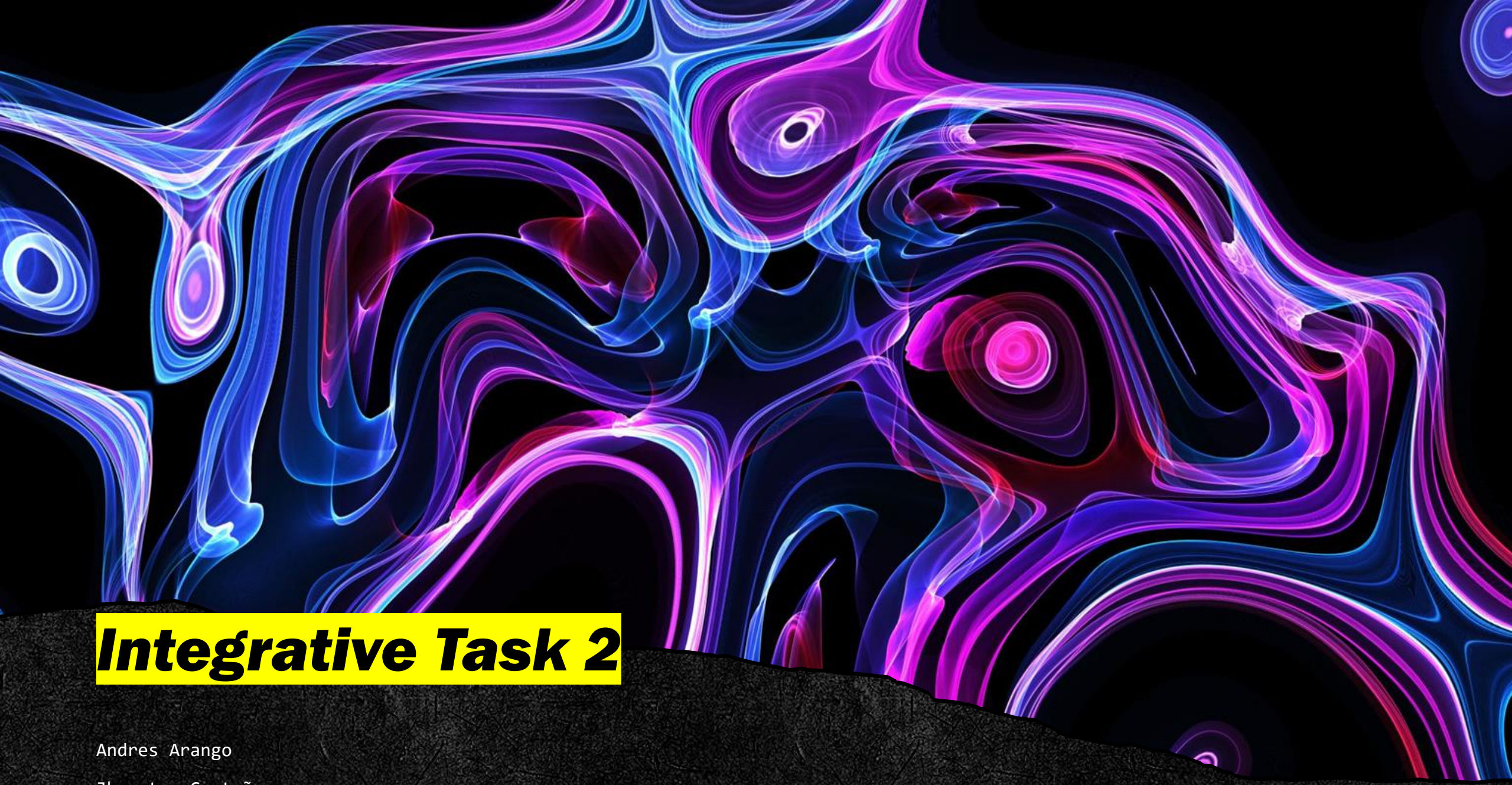


Integrative Task 2

Andres Arango

Jhonatan Castaño

Introduction

- Objective: Build a sentiment analysis model using supervised learning with vanilla Recurrent Neural Networks and LSTM
- Secondary Objectives:

1. Create a database with sentences and the type of sentiment of itself.
2. Tokenize the sentences to find a way to build a supervised learning model.
3. Implement a DummyClassifier for the model.
4. Implement a vanilla RNN sentiment analysis model.
5. Implement a LSTM sentiment analysis model.

Methodology

Define the problem

Gather Data

Explore and prepare data

Feature Engineerin

Split the Data

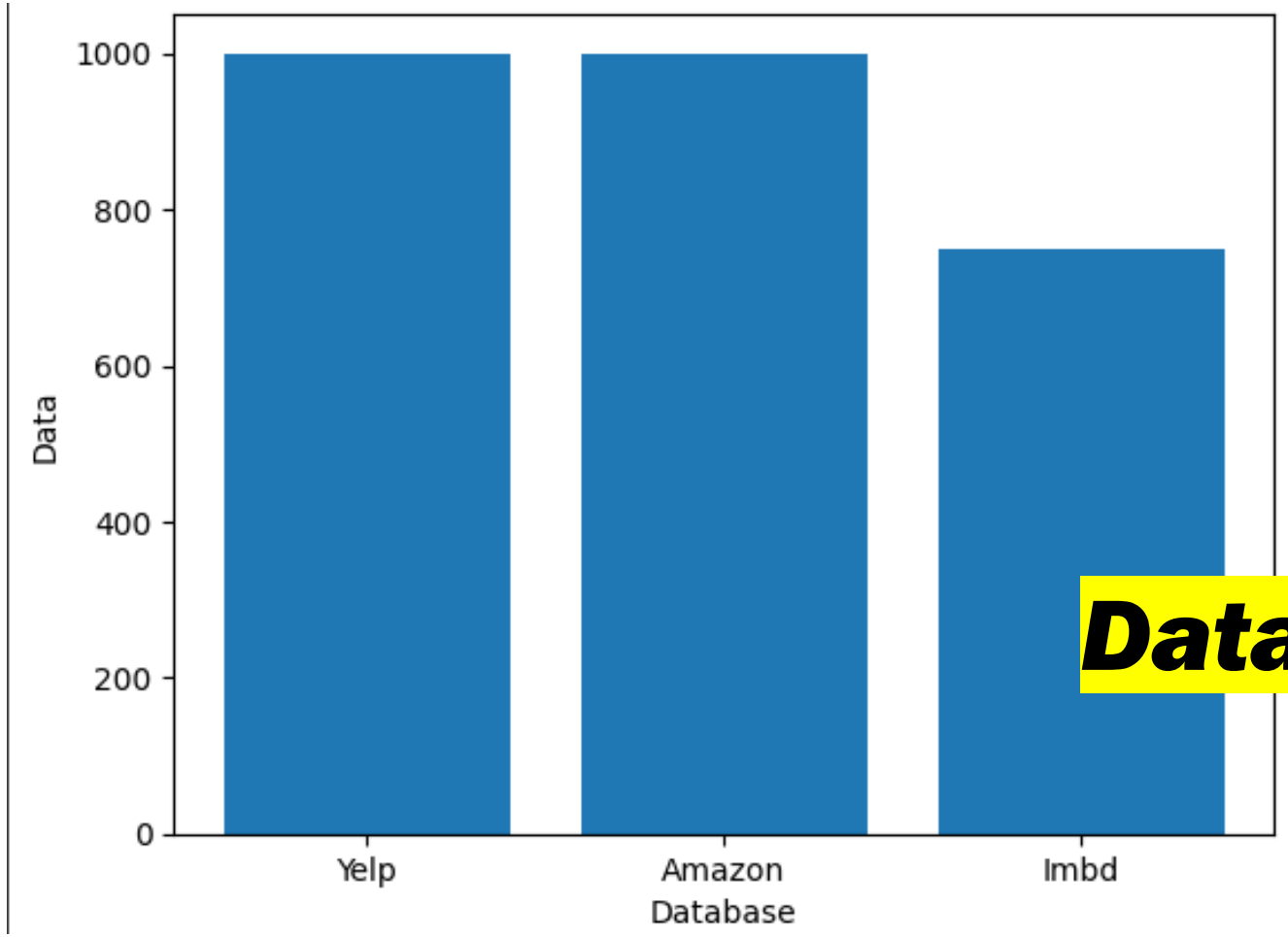
Choose a Model

Train the Model

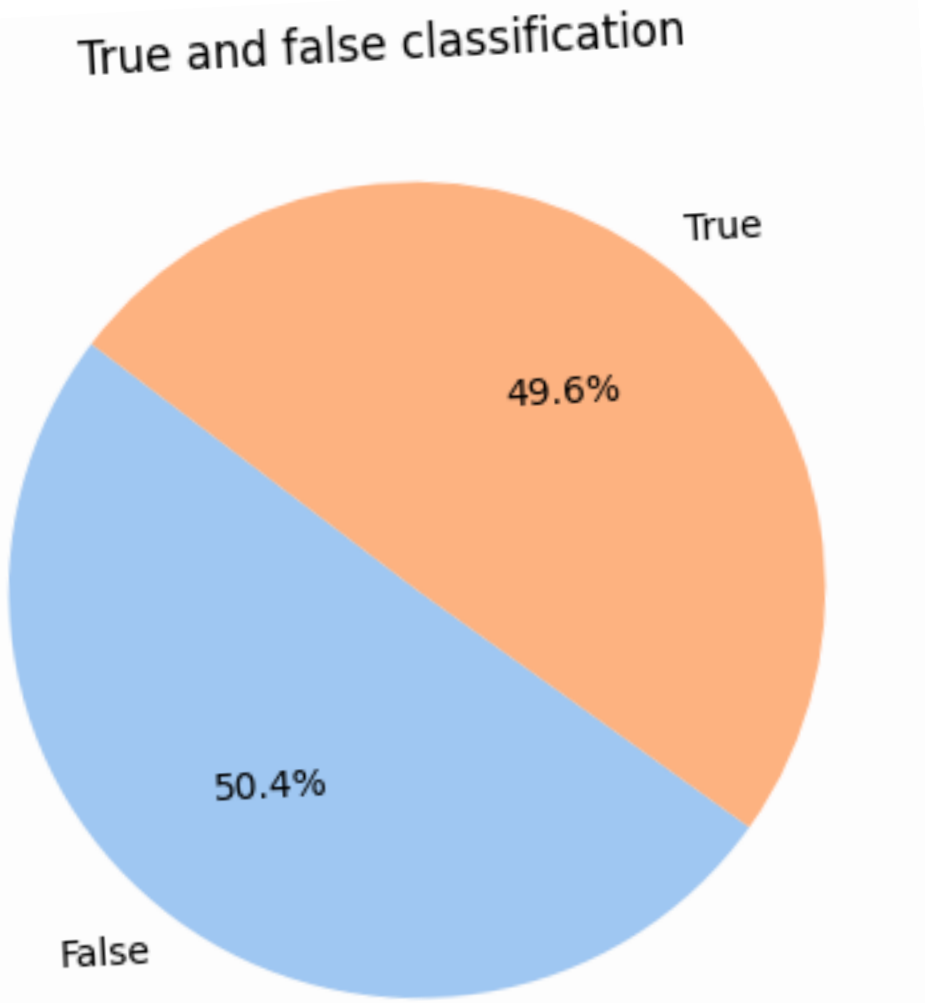
Evaluate Model

Hyperparameter Tuning

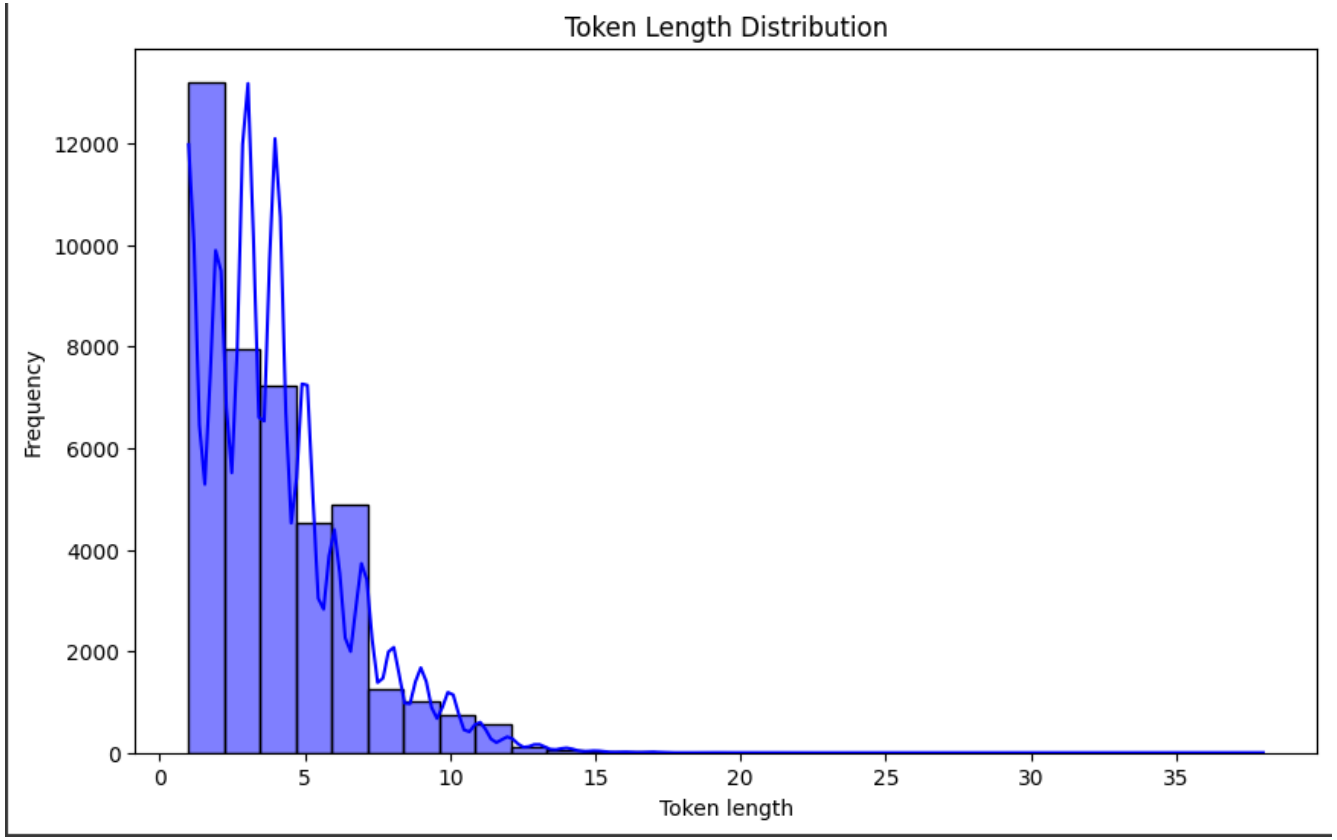
Deploy



Data From Each of the Databases

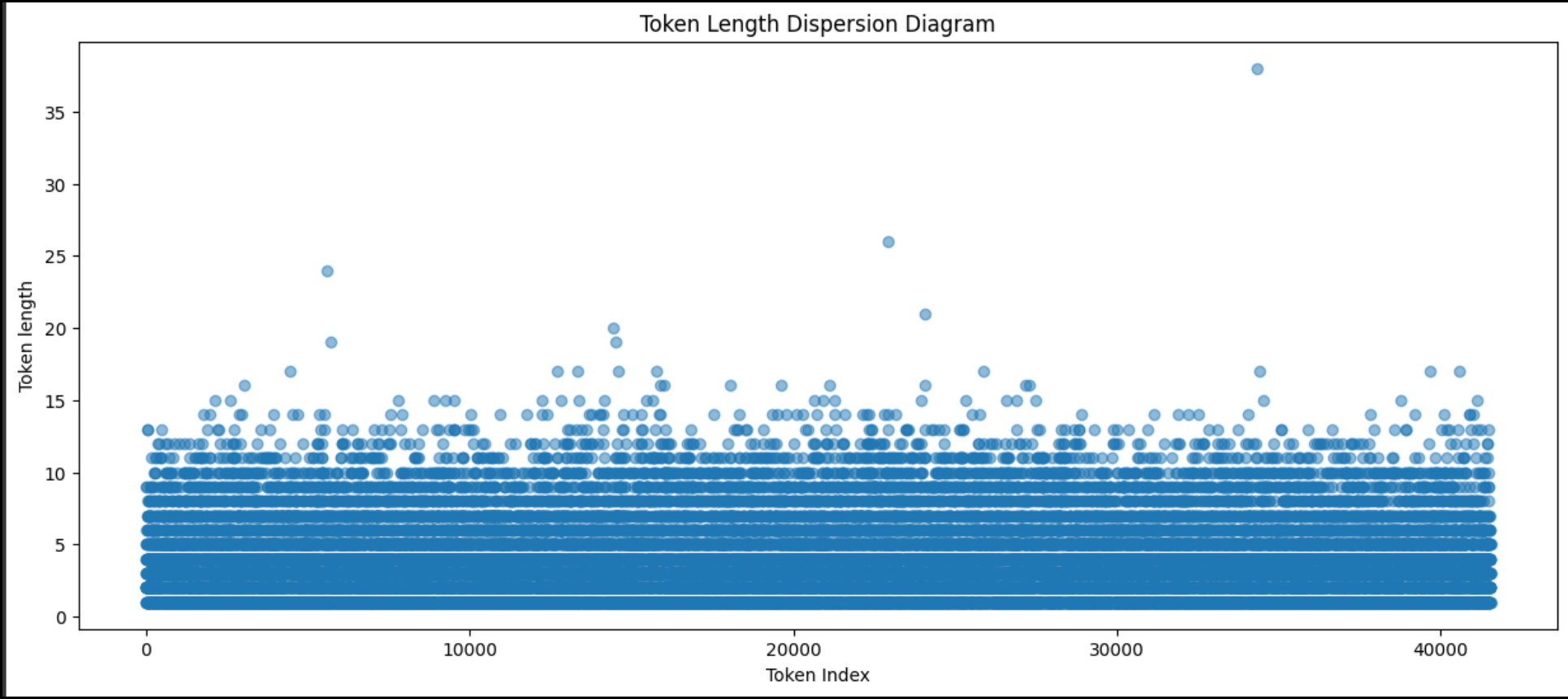


Clasiffication



Token length Distribution

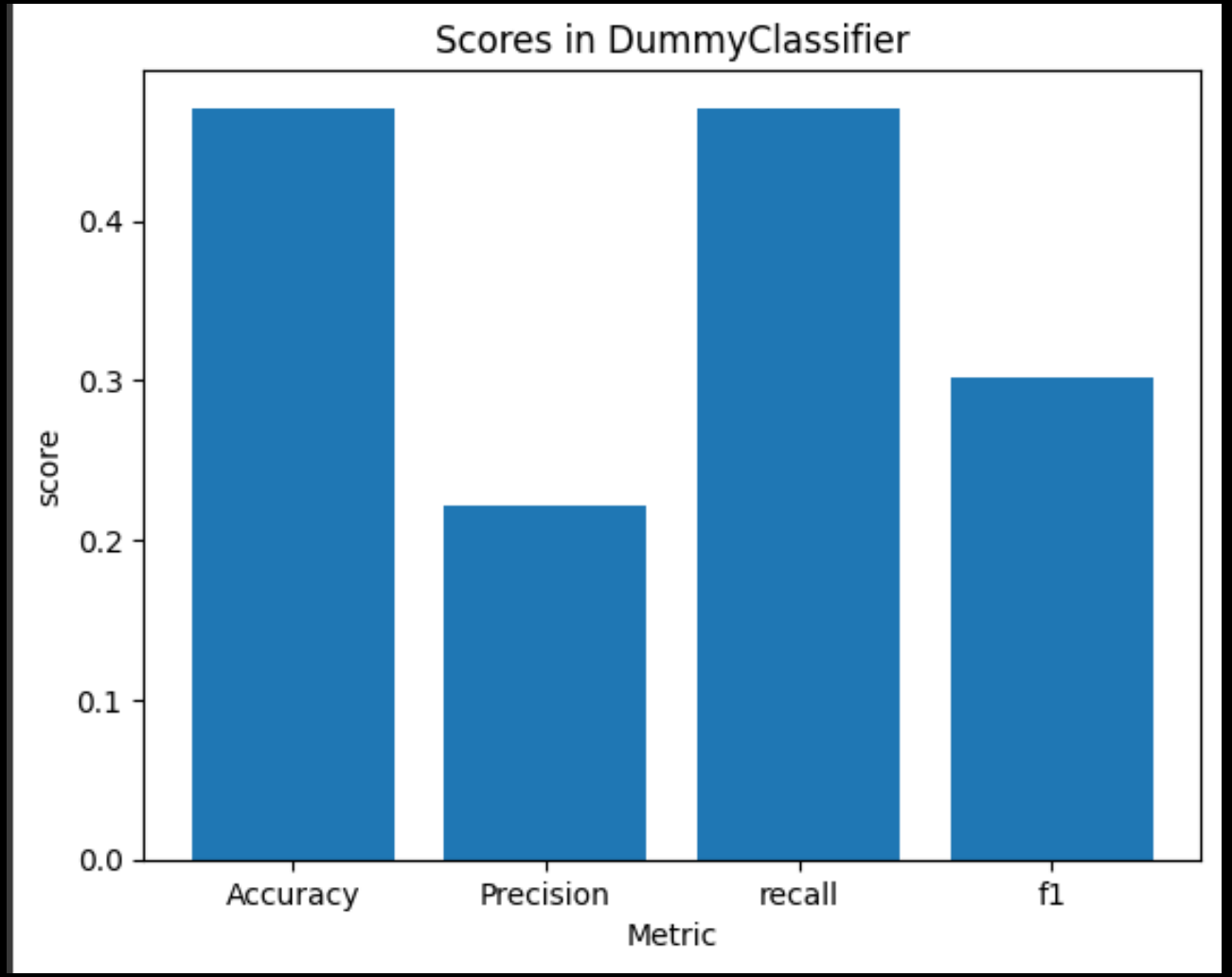
Token length Dispersion

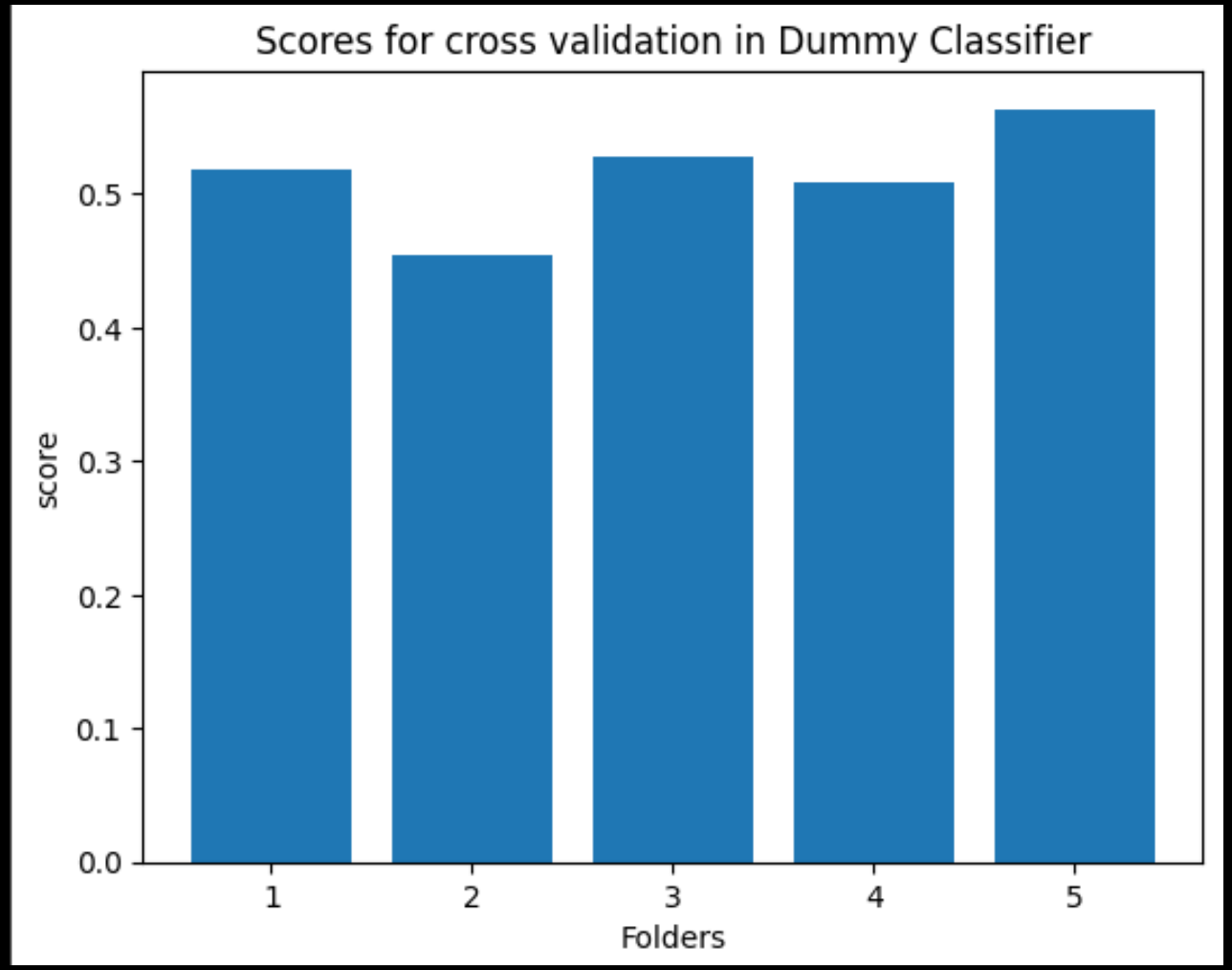


Token Word Cloud

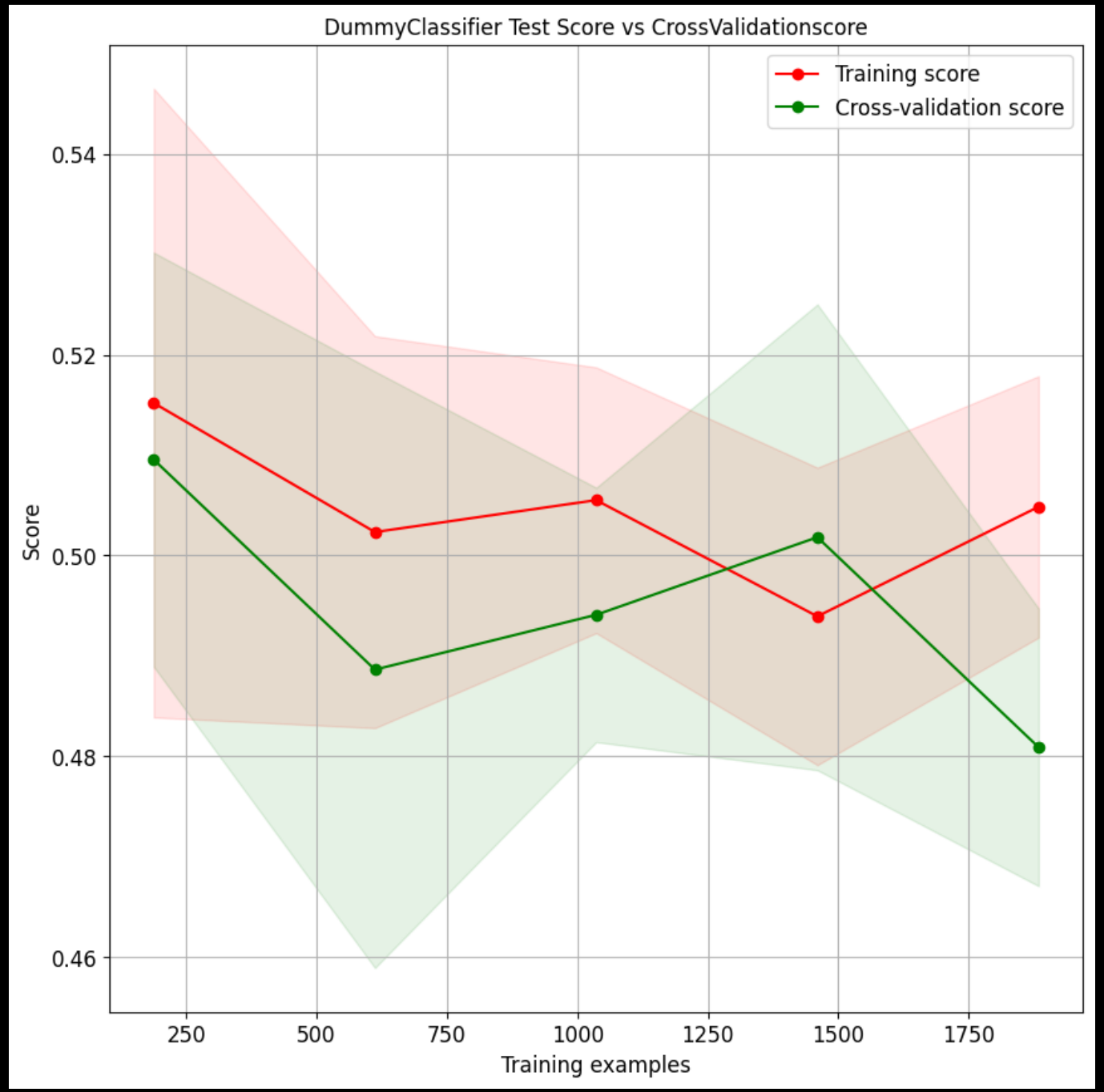
Dummy Classifier Implementation

Scores in DummyClassifier

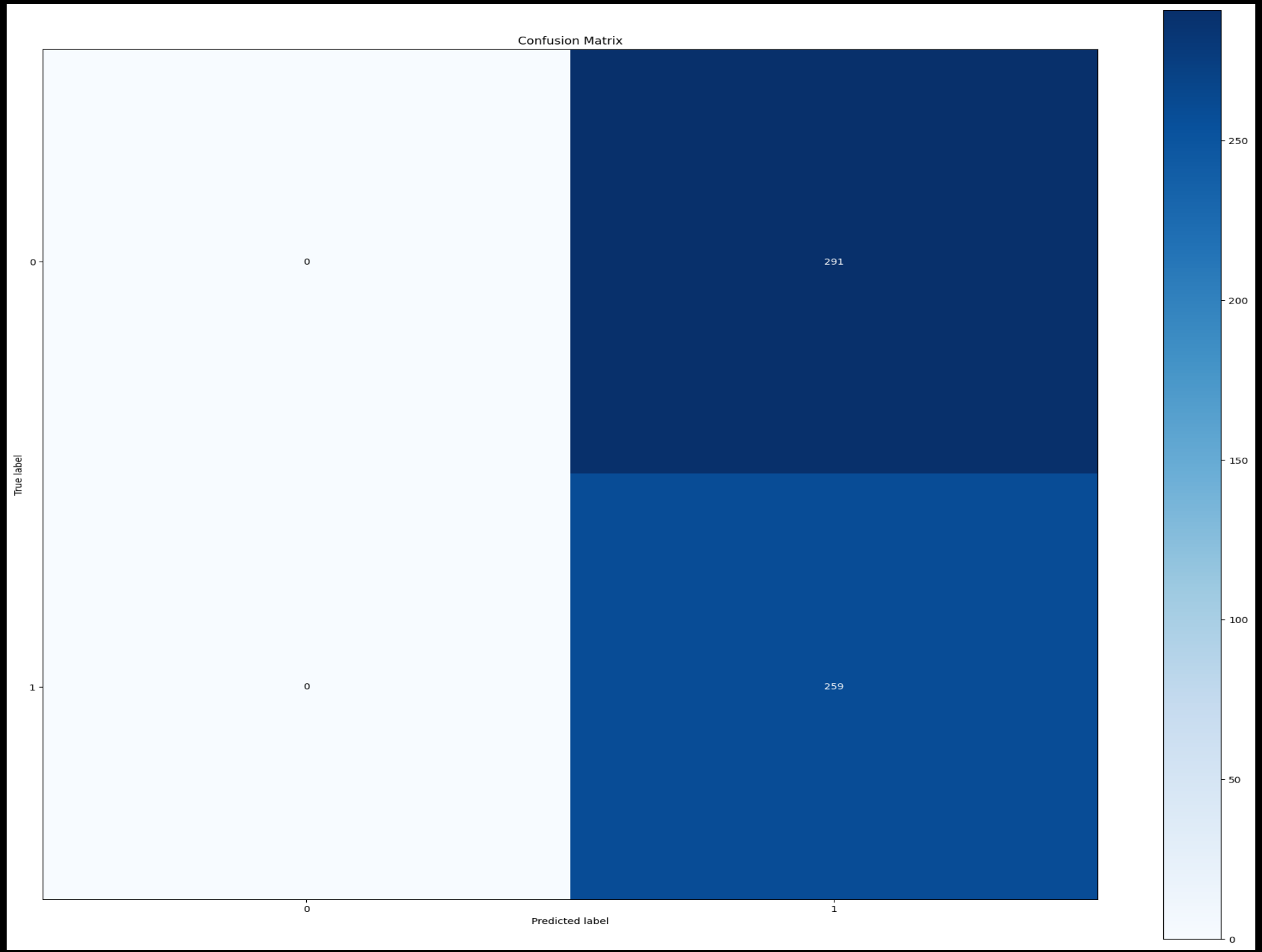




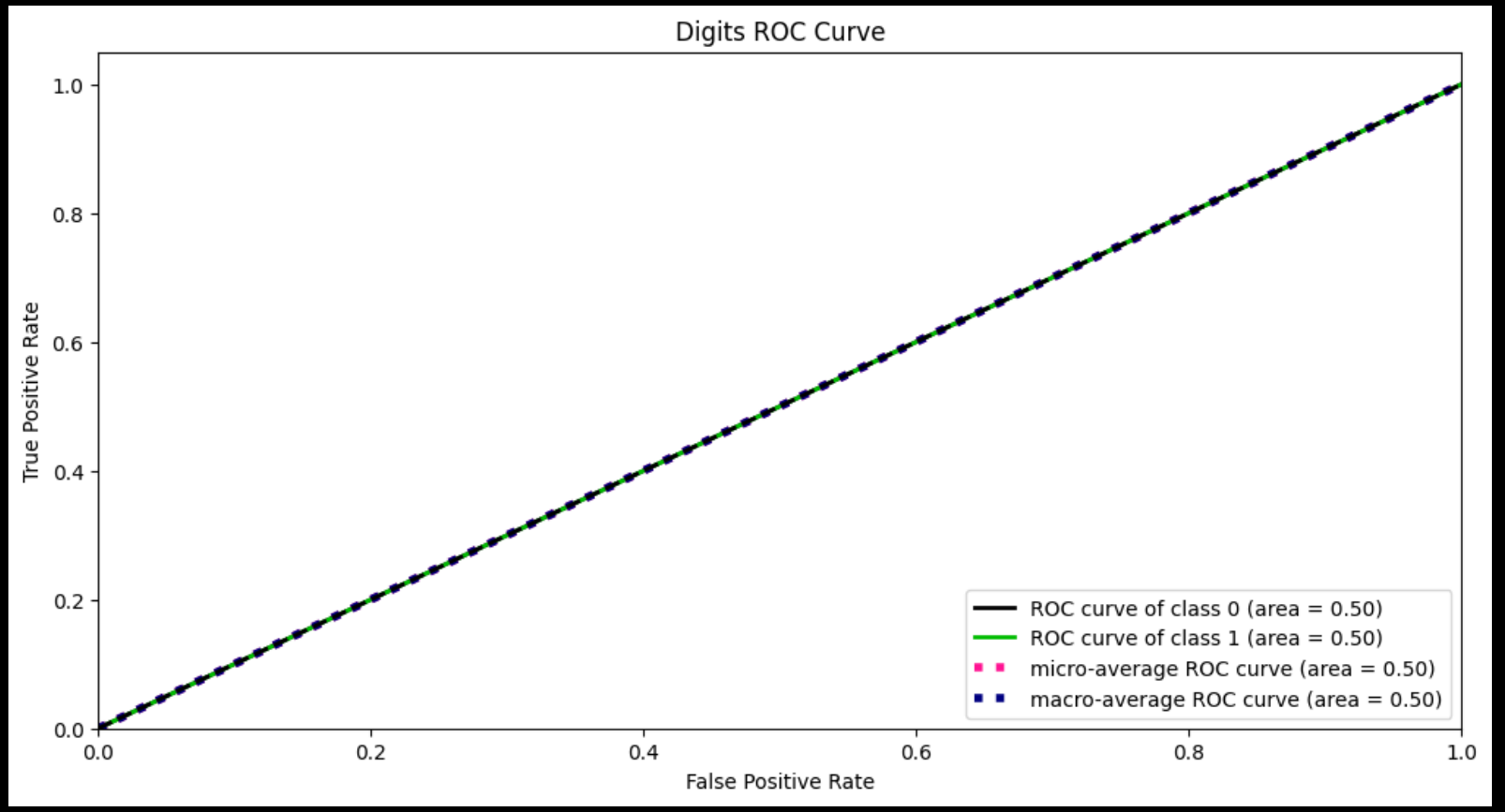
Scores for cross validation in DummyClassifier



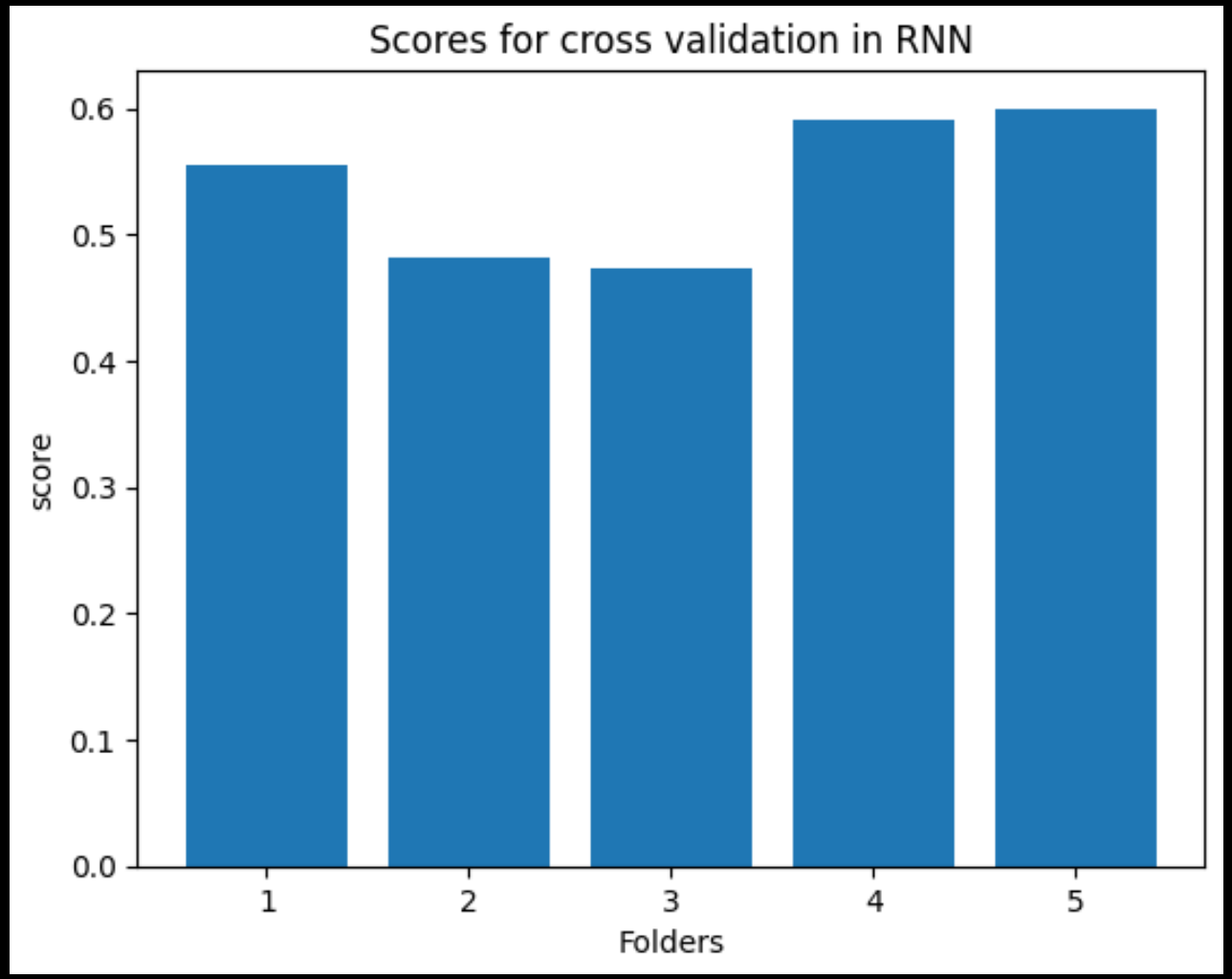
Cross validation vs Training results


Confusion Matrix
0
291
0
259
0
1
True label
Predicted label
250
200
150
100
50
0

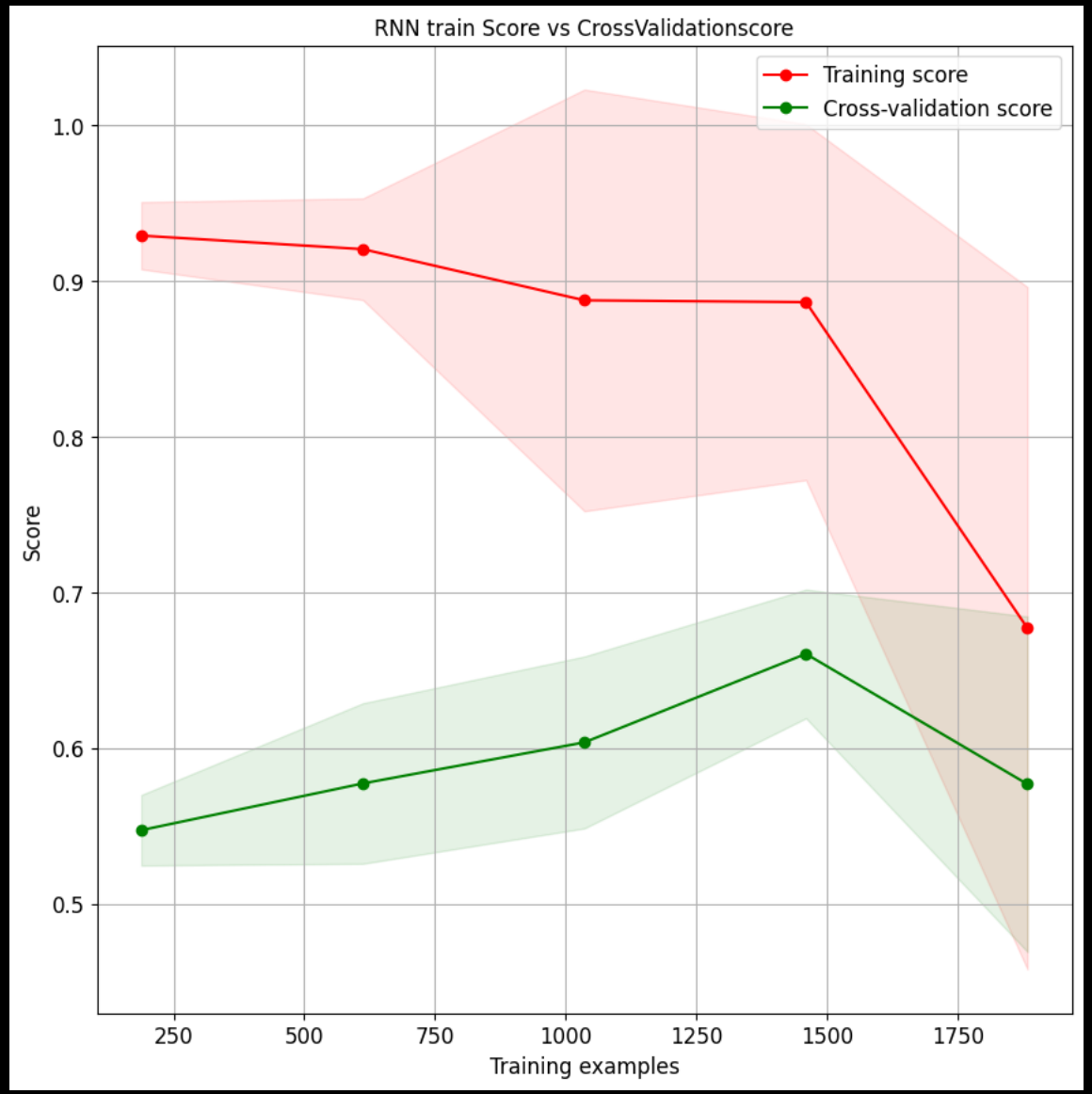
ROC curve


Digits ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.50)
ROC curve of class 1 (area = 0.50)
micro-average ROC curve (area = 0.50)
macro-average ROC curve (area = 0.50)

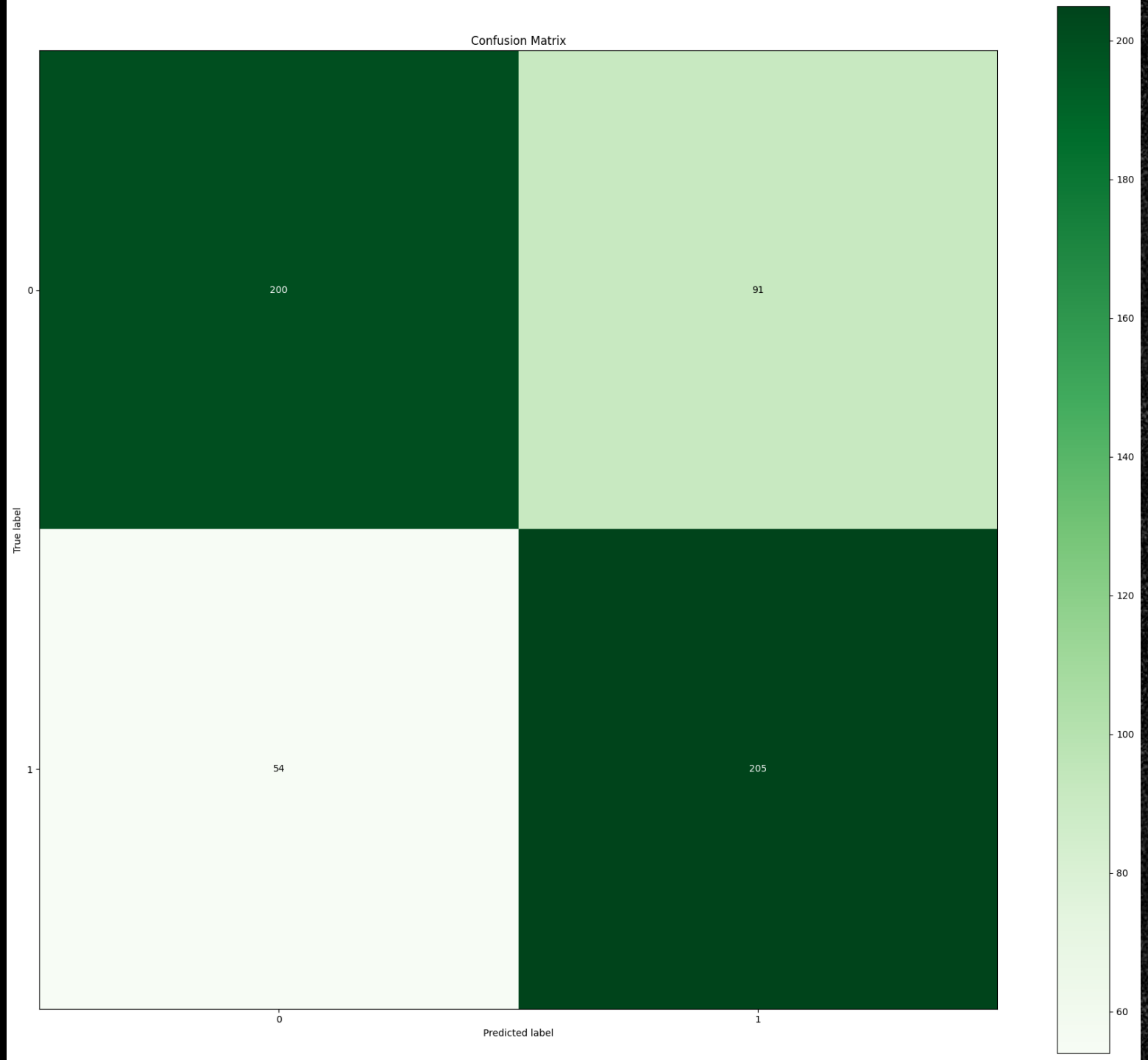
RNN Model



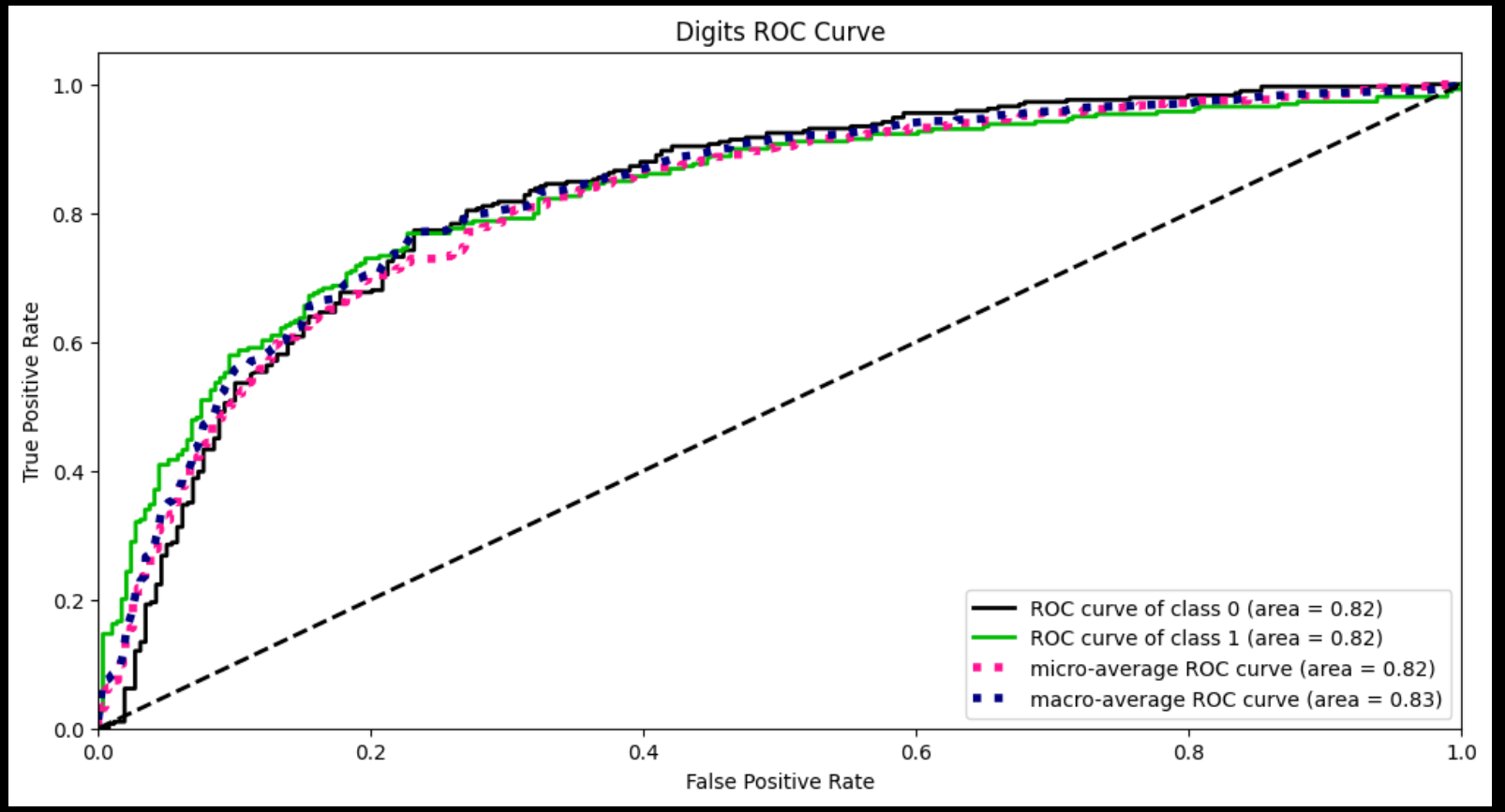
RNN Cross validation

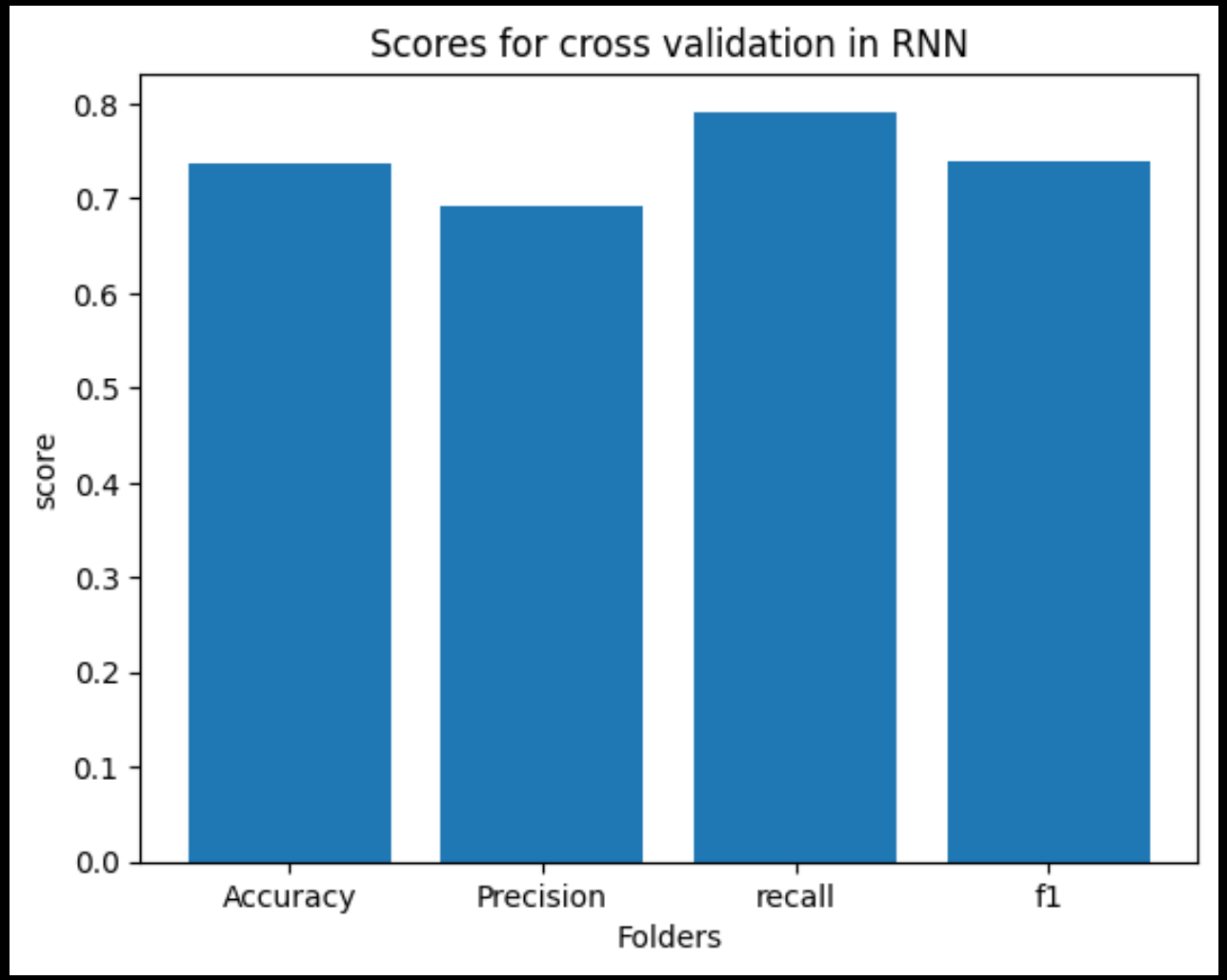


RNN Train Score vs Cross Validation


Confusion Matrix
200
91
54
205
0
1
True label
0
1
Predicted label
200
180
160
140
120
100
80
60

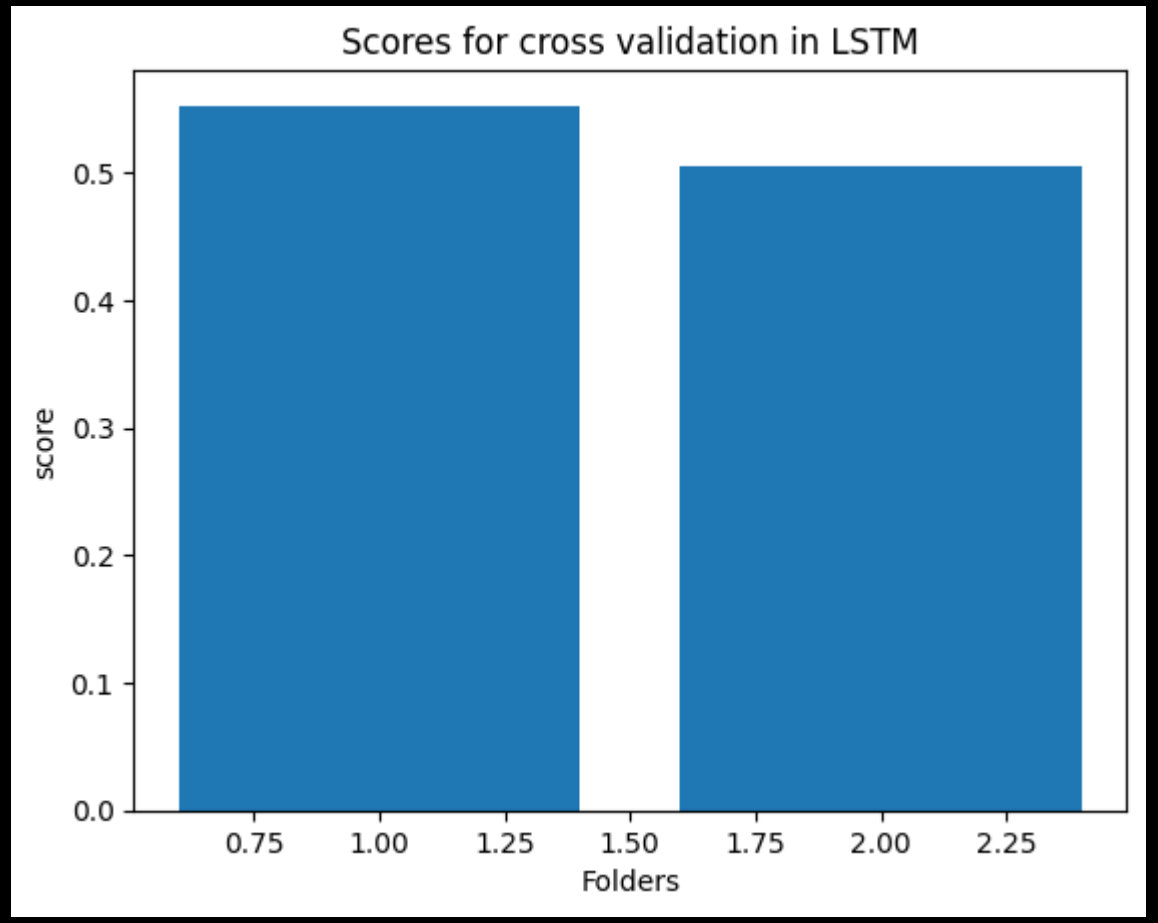
ROC curve


Digits ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.82)
ROC curve of class 1 (area = 0.82)
micro-average ROC curve (area = 0.82)
macro-average ROC curve (area = 0.83)

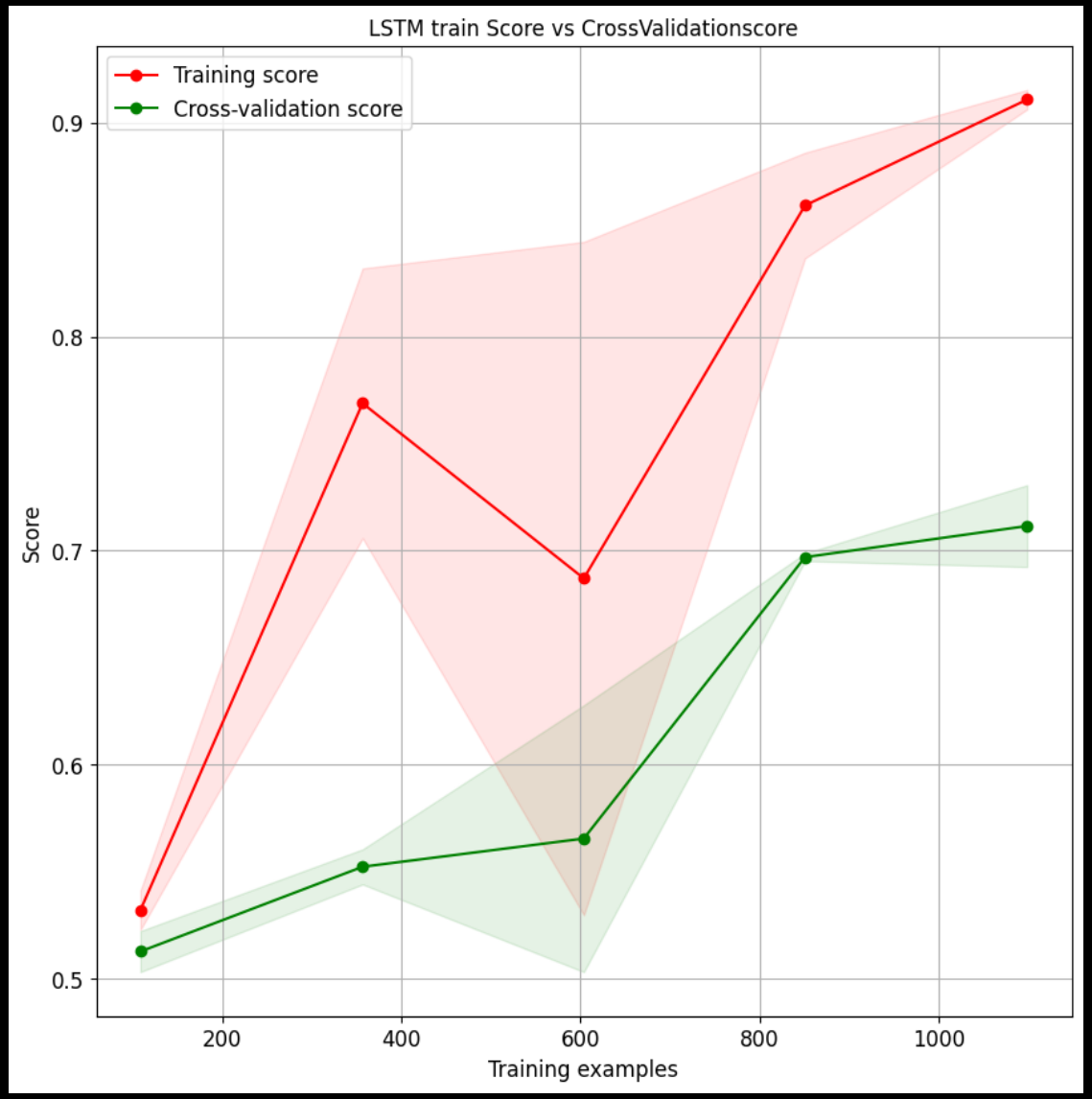


RNN Cross validation

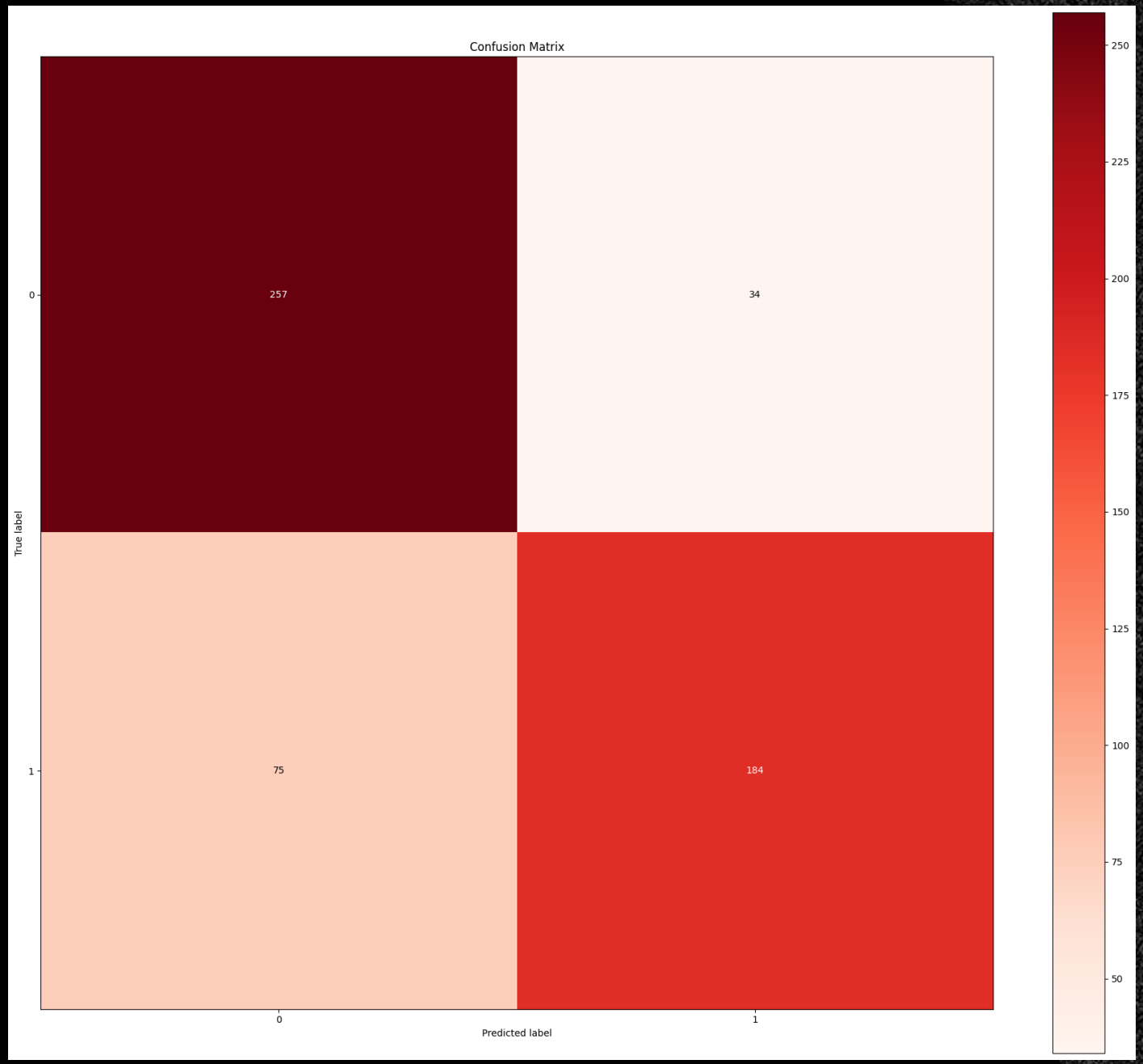
LSMT Model



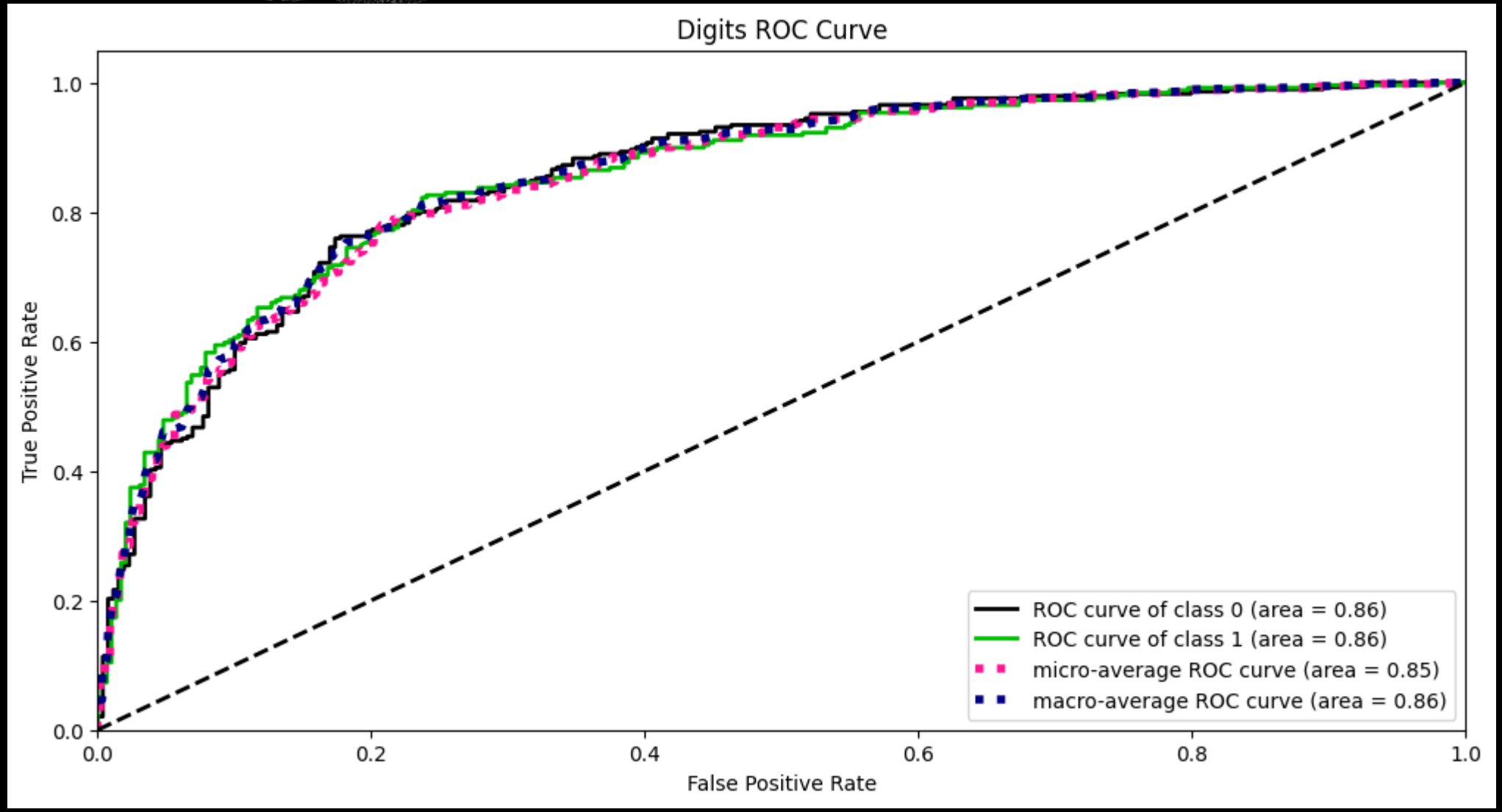
LMST Cross validation

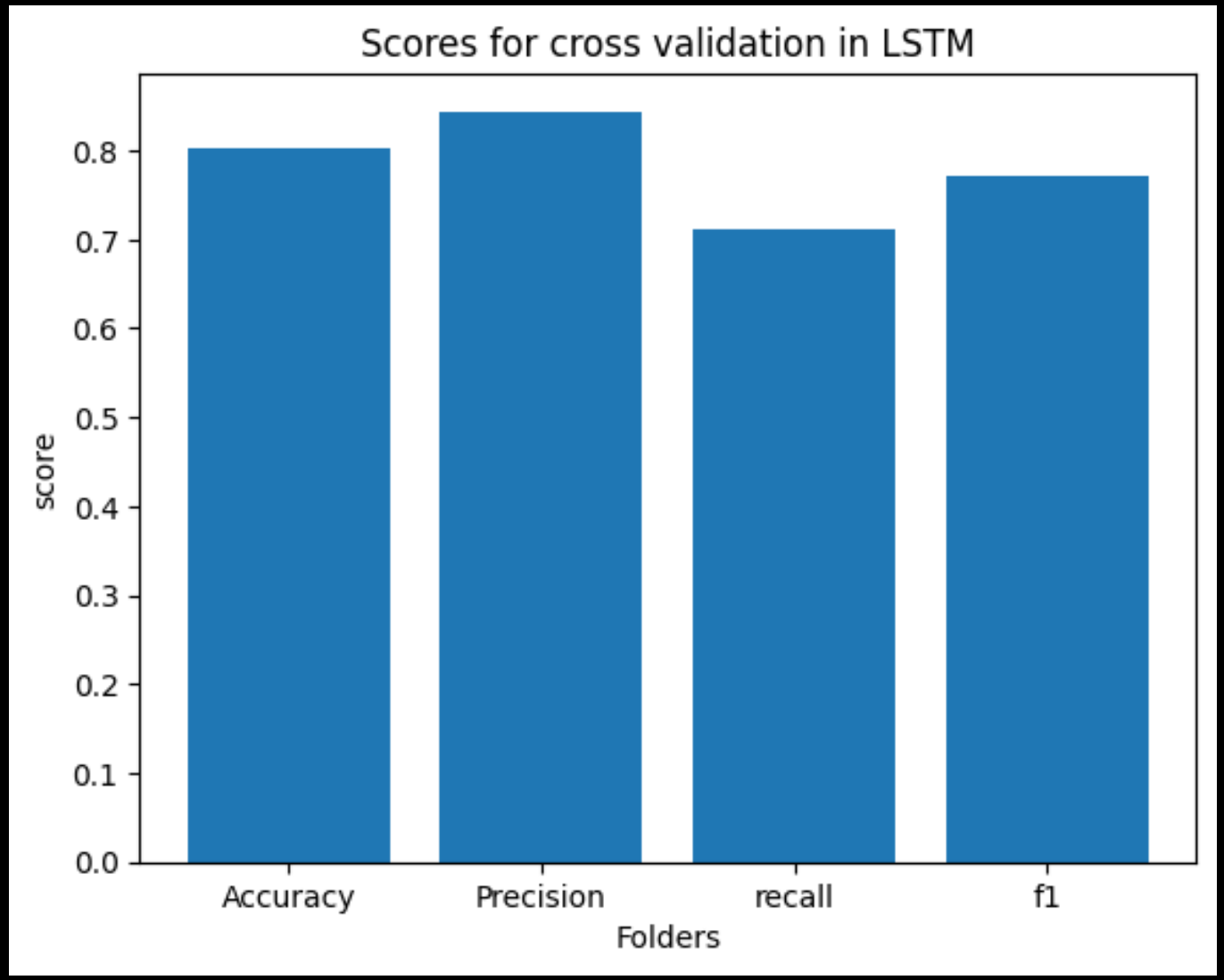


Train Score vs Cross Validation


Confusion Matrix
257
34
75
184
0
1
True label
0
1
Predicted label
250
225
200
175
150
125
100
75
50

ROC curve


Digits ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.86)
ROC curve of class 1 (area = 0.86)
micro-average ROC curve (area = 0.85)
macro-average ROC curve (area = 0.86)



RNN Cross validation