
# UNIVERSITÀ DEGLI STUDI DI TRIESTE

---

**Dipartimento di Matematica, Informatica e Geoscienze**

**CORSO DI DOTTORATO DI RICERCA IN**

## APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

**CICLO XXXIX**

# TIME SERIES ANALYSIS FOR SEISMIC WAVE DETECTION AND PREDICTION

Tanaka Hernandez, Ken

June, 2026

# Contents

# List of Figures

# Introduction

Earth's litosphere is a very complex and dynamic system as it is constantly subject to stresses and deformations in response to the forces and rheological properties of the materials acting on it. The lithosphere is composed of the crust and the uppermost part of the mantle and it is divided into tectonic plates that float on the semi-fluid asthenosphere. The relative motion of these plates is responsible for the formation of mountain ranges, ocean basins, earthquakes and volcanic activity. The study of the lithosphere and its dynamics is of great importance as it helps us to understand the processes that shape the Earth's surface and to predict natural hazards such as earthquakes and volcanic eruptions.

After the earthquake in L'Aquila, Italy, in 2009, which caused 309 deaths and left 65,000 people homeless, the Italian government decided to establish the Istituto Nazionale di Geofisica e Vulcanologia (INGV) to monitor and study seismic activity in the country. The INGV has deployed a network of over 400 seismometers throughout Italy to detect and locate earthquakes and to study the propagation of seismic waves through the Earth's crust. These vast amounts of data collected by the seismometers are processed and analyzed by seismologists to estimate the seismic hazard of the region. Traditional methods, relying on physics-based models and empirical relationships, have significantly advanced our understanding of earthquake processes. However, these methods have some limitations, such as the need for simplifying assumptions and the inability to capture the complexity of the lithosphere. Furthermore, the traditional methods of detecting and locating earthquakes rely on the manual inspection of seismic data by seismologists, which is a time-consuming and subjective process. The inherent complexity, nonlinearity, and variability of seismic phenomena have motivated the exploration of new methodologies that can handle large amounts of data and uncover patterns that are difficult to detect using conventional techniques.

In recent years, there has been an increasing interest in the use of machine learning (ML) techniques to analyze seismic data and to improve the detection and location of earthquakes. Machine learning algorithms have been shown to be effective in detecting seismic events, in classifying seismic signals and in estimating the magnitude and location of earthquakes. The ability of ML models to recognize patterns, make inferences, and generalize from data makes them well-suited for the complex and data-rich field of seismology. Seismological datasets are often vast and multidimensional, offering a fertile ground for the application of ML techniques.

In this thesis, we explore the use of ML techniques for the detection and prediction of seismic waves. We focus on the analysis of seismic waveforms recorded by seismometers in the North Eastern Italy region known as AdriaArray. Our goal is to develop ML models that can automatically detect earthquakes from seismic waveforms and predict the arrival times of seismic waves in real-time. This research, as implied by the Piano Nazionale di Ripresa e Resilienza (PNRR) objectives, has the potential to improve the efficiency and accuracy of the seismic monitoring system in Italy and to enhance the understanding of seismic phenomena through digital innovation. Furthermore, we develop a cloud-based platform publicly available, infrastructure provided by the Istituto Nazionale di Oceanografia e di Geofisica Sperimetale (OGS), for the scientific community to access and analyze seismic data, to train ML models, and to share the results of the analysis under the Terabit Network for Research and Academic Big Data in Italy (TeRABIT) project.

# Chapter 1

# Data Description

In this chapter, we describe the dataset of continuous waveforms that are used for the detection and prediction of seismic waves. The dataset consists of continuous waveforms recorded by seismic stations around the North-Eastern Italy region known as AdriaArray. Access to the data is provided by the OGS and the INGV as part of the PNRR and the TeRABIT project.

## 1.1 Datasets

### 1.1.1 Original

The original dataset contains continuous waveforms recorded by seismic stations around the North-Eastern Italy region. The waveforms are labeled with the arrival times of the P and S waves, which are the first and second arrivals of a seismic event, respectively. The waveforms are divided into three classes: P waves, S waves, and void. The P and S waves are the seismic phases that are of interest for the detection and prediction of seismic events, while the void class represents the absence of any seismic phase. The dataset contains a total of 100,000 waveforms, with 33,333 waveforms in each class.

### 1.1.2 INSTANCE

The Italian Seismic Dataset for Machine Learning (INSTANCE) dataset was compiled by the Istituto Nazionale di Geofisica e Vulcanologia (INGV) and contains

- 54,008 earthquakes for a total of 1,159,249 3-channel waveforms

- 132,330 3-channel noise waveforms
- 115 metadata for each waveform providing information on *station*, *trace*, *source*, *path*, and *quality*
- 19 networks
- 620 stations
- Magnitude range of 0 to 6.5

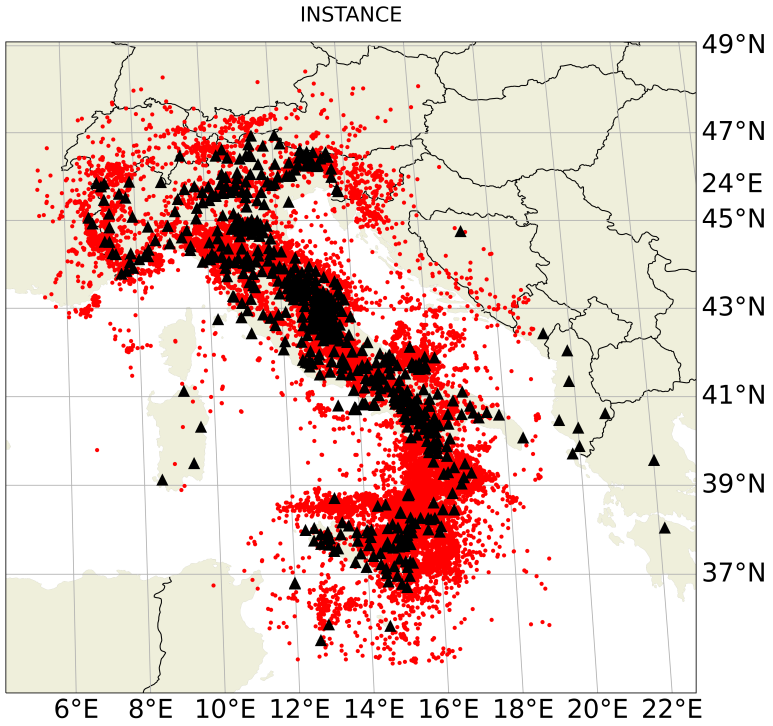



Figure 1.1: Seismic stations used for waveforms extraction in INSTANCE dataset

### 1.1.3 SCEDC

The Southern California Earthquake Data Center (SCEDC) dataset was compiled by the SCEDC and contains all publicly available recordings of seismic events in Southern California Seismic Network. The dataset contains  8.1 million manually picked waveforms from 2000 to 2020.

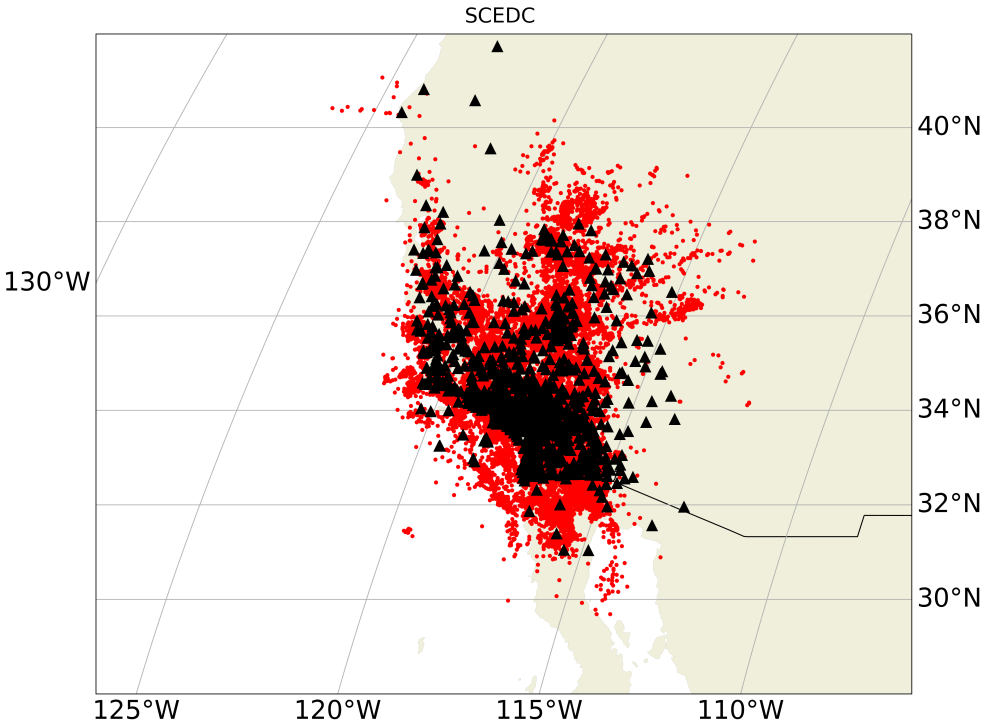



Figure 1.2: Seismic stations used for waveforms extraction in SCEDC dataset

### 1.1.4 STEAD

The STanford EArthquake Dataset (STEAD) dataset was compiled by the STEAD project and contains

- 1.2 million time series
- 100,000 noise examples
- 450,000 earthquake examples

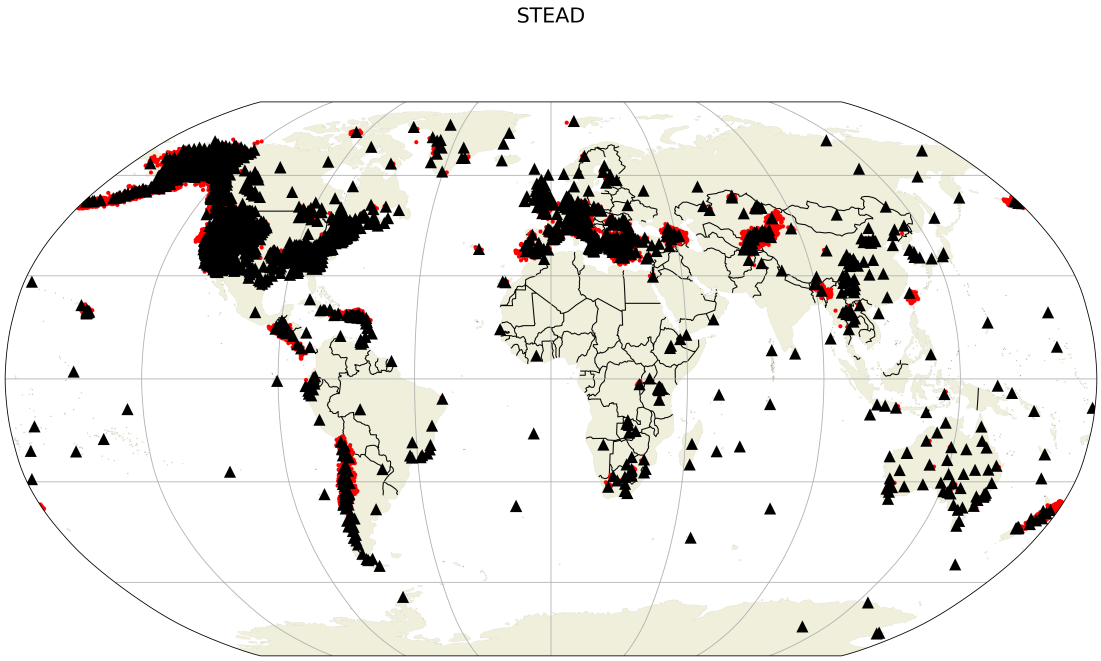



Figure 1.3: Seismic stations used for waveforms extraction in STEAD dataset

## 1.2 AdriaArray Region

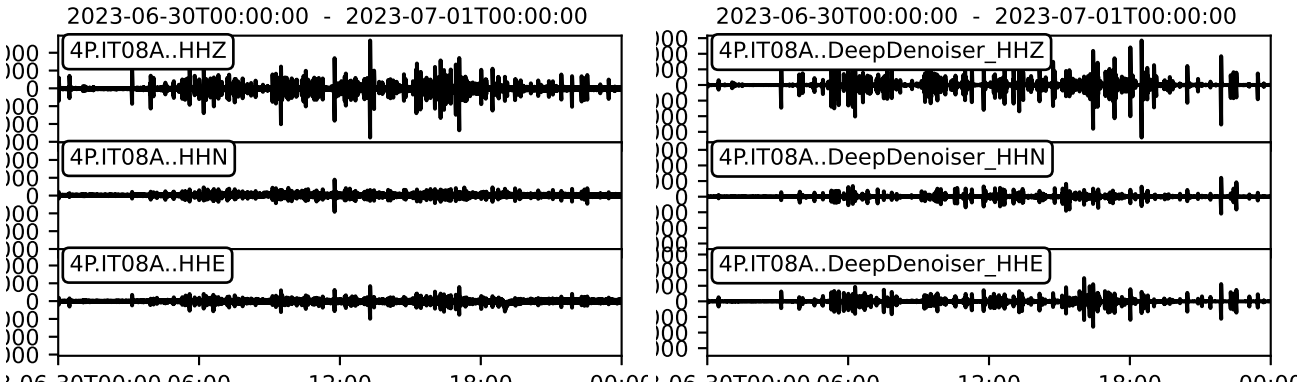
The North-Eastern Italy region known as AdriaArray is a seismically active region with a high density of seismic stations. The region is located in the

## 1.3 Waveforms

## 1.4 Labels

## 1.5 Data Preprocessing

The continuous waveforms may vary in length and amplitude given the different seismic instruments used to record the data. Some continuous waveforms contain accelerometer data which are sampled at a frequency of 200 Hz and have a lower amplitude, while others contain velocity data, sampled at a frequency of 100 Hz with higher amplitude which may exceed the dynamic range of the instrument. Accelerometer can be identified by the **Accelerometer data is converted to velocity**




**data by integrating the accelerometer data to obtain the velocity data**. The waveforms are then resampled to a fixed frequency of 100 Hz to ensure that the models are able to learn the patterns in the data.

As the ML models used in this study were trained on waveforms with a fixed frequency of 100 Hz, the data is resampled to match this frequency. The waveforms are also standardized to a fixed window of one day (*i.e.* 8640000 samples). The seismic instruments record data in three components: Z, N, and E. The waveforms are preprocessed by selecting the Z component, which is the vertical component of the seismic wave, and removing the N and E components.

# Chapter 2

# Literature Review

ADD traditional methods for seismic wave detection and prediction ADD a table of the hyperparameters of the different models

## 2.1 Methods for Seismic Pick Detection

### 2.1.1 Traditional Methods

Traditional methods for seismic pick detection rely on the manual identification of seismic phases in continuous waveforms. These methods are based on the analysis of the waveform data to identify the arrival times of the P and S waves, which are the first and second arrivals of a seismic event, respectively. The arrival times of the seismic phases are used to locate the source of the seismic event and to estimate the magnitude of the event. Traditional methods for seismic pick detection are time-consuming and labor-intensive and are prone to errors and inconsistencies. These methods are also limited in their ability to detect and predict the arrival times of the seismic phases in real-time.

### 2.1.2 Deep Learning Methods

Deep learning methods for seismic pick detection are based on the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to learn the complex patterns in the data that are indicative of seismic phases. These methods are able to detect and predict the arrival times of the seismic phases in real-time and are able to generalize well to unseen data. Deep learning

methods for seismic pick detection have been shown to outperform traditional methods and are able to detect and predict the arrival times of the seismic phases with high accuracy and low latency. These methods are also able to automatically extract features from the waveform data and are able to learn the relationships between the features and the seismic phases.

#### PhaseNet

PhaseNet is a deep learning model for the detection of seismic phases in continuous waveforms. The model is based on a convolutional neural network (CNN) architecture that takes as input a raw waveform and outputs the probability of the presence of a seismic phase at each time step. The model is trained on a large dataset of seismic waveforms with labeled phase picks and is able to learn the complex patterns in the data that are indicative of seismic phases. PhaseNet has been shown to outperform traditional methods for phase detection and is able to detect seismic phases with high accuracy and low latency.

#### EQTransformer

EQTransformer is a deep learning model for the detection and classification of seismic signals in continuous waveforms. The model is based on a convolutional neural network (CNN) architecture that takes as input a raw waveform and outputs the probability of the presence of a seismic signal at each time step. The model is trained on a large dataset of seismic waveforms with labeled signal picks and is able to learn the complex patterns in the data that are indicative of seismic signals. EQTransformer has been shown to outperform traditional methods for signal detection and classification and is able to detect seismic signals with high accuracy and low latency.

## 2.2 Methods for Seismic Phase association

There are several methods that have been developed for the association of seismic phases in continuous waveforms. In this section, we will compare the performance of three methods, namely the Gaussian Mixture Model Associator (GaMMA), the k-nearest neighbors (KNN) algorithm, and the support vector machine (SVM) algorithm, for seismic phase association. The methods were trained on a large dataset of seismic waveforms with labeled phase picks and are able to learn the

complex patterns in the data that are indicative of seismic phases. The performance of the methods was evaluated using the following metrics: the cumulative number of detections, the multi-class classification results, the true positive, false negative, and recall, and the time displacement.

### 2.2.1 Traditional Methods

Traditional methods for seismic phase association rely on the manual identification of seismic phases in continuous waveforms. These methods are based on the analysis of the waveform data to identify the arrival times of the P and S waves, which are the first and second arrivals of a seismic event, respectively. The arrival times of the seismic phases are used to locate the source of the seismic event and to estimate the magnitude of the event. Traditional methods for seismic phase association are time-consuming and labor-intensive and are prone to errors and inconsistencies. These methods are also limited in their ability to detect and predict the arrival times of the seismic phases in real-time.

### 2.2.2 Deep Learning Methods

Deep learning methods for seismic phase association are based on the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to learn the complex patterns in the data that are indicative of seismic phases. These methods are able to detect and predict the arrival times of the seismic phases in real-time and are able to generalize well to unseen data. Deep learning methods for seismic phase association have been shown to outperform traditional methods and are able to detect and predict the arrival times of the seismic phases with high accuracy and low latency. These methods are also able to automatically extract features from the waveform data and are able to learn the relationships between the features and the seismic phases.

#### GaMMA

GaMMA is a probabilistic model for the detection and classification of seismic signals in continuous waveforms. The model is based on a Gaussian mixture model (GMM) architecture that takes as input a set of features extracted from the waveform and outputs the probability of the presence of a seismic signal. The model is trained on a large dataset of seismic waveforms with labeled signal picks and is able to learn the complex patterns in the data that are indicative of seismic signals.

GaMMA has been shown to outperform traditional methods for signal detection and classification and is able to detect seismic signals with high accuracy and low latency.

## 2.3 Models for Seismic Wave Detection

There are several models that have been developed for the detection of seismic waves in continuous waveforms. In this section, we will compare the performance of three models, namely PhaseNet, EQTransformer, and GaMMA, for seismic wave detection. The models were trained on a large dataset of seismic waveforms with labeled phase picks and are able to learn the complex patterns in the data that are indicative of seismic phases. The performance of the models was evaluated using the following metrics: the cumulative number of detections, the multi-class classification results, the true positive, false negative, and recall, and the time displacement.

### 2.3.1 PhaseNet

PhaseNet is a deep learning model for the detection of seismic phases in continuous waveforms. The model is based on a convolutional neural network (CNN) architecture that takes as input a raw waveform and outputs the probability of the presence of a seismic phase at each time step. The model is trained on a large dataset of seismic waveforms with labeled phase picks and is able to learn the complex patterns in the data that are indicative of seismic phases. The model is designed to be robust to noise and can generalize well to unseen data. PhaseNet has been shown to outperform traditional methods for phase detection and is able to detect seismic phases with high accuracy and low latency.

### 2.3.2 EQTransformer

EQTransformer is a deep learning model for the detection and classification of seismic signals in continuous waveforms. The model is based on a convolutional neural network (CNN) architecture that takes as input a raw waveform and outputs the probability of the presence of a seismic signal at each time step. The model is trained on a large dataset of seismic waveforms with labeled signal picks and is able to learn the complex patterns in the data that are indicative of seismic signals. The model is designed to be robust to noise and can generalize well to unseen data. EQTransformer

has been shown to outperform traditional methods for signal detection and classification and is able to detect seismic signals with high accuracy and low latency.

## 2.4 GaMMA

GaMMA is a probabilistic model for the detection and classification of seismic signals in continuous waveforms. The model is based on a Gaussian mixture model (GMM) architecture that takes as input a set of features extracted from the waveform and outputs the probability of the presence of a seismic signal. The model is trained on a large dataset of seismic waveforms with labeled signal picks and is able to learn the complex patterns in the data that are indicative of seismic signals. The model is designed to be robust to noise and can generalize well to unseen data. GaMMA has been shown to outperform traditional methods for signal detection and classification and is able to detect seismic signals with high accuracy and low latency.

# Chapter 3

# Time Series Analysis

## 3.1 RNN

RNN is a type of neural network that is designed to handle sequential data. The network is composed of a series of interconnected nodes, each of which represents a time step in the sequence. The nodes are connected by weighted edges, which determine the strength of the connection between the nodes. The network is trained on a large dataset of sequential data and is able to learn the complex patterns in the data that are indicative of the underlying structure. RNN has been shown to outperform traditional methods for sequential data analysis and is able to model the temporal dependencies in the data.

## 3.2 LSTM

LSTM is a type of RNN that is designed to handle long sequences of data. The network is composed of a series of interconnected nodes, each of which represents a time step in the sequence. The nodes are connected by weighted edges, which determine the strength of the connection between the nodes. The network is trained on a large dataset of sequential data and is able to learn the complex patterns in the data that are indicative of the underlying structure. LSTM has been shown to outperform traditional methods for sequential data analysis and is able to model the temporal dependencies in the data.

## 3.3 Transformer

The Transformer is a type of neural network that is designed to handle sequential data. The network is composed of a series of interconnected nodes, each of which represents a time step in the sequence. The nodes are connected by weighted edges, which determine the strength of the connection between the nodes. The network is trained on a large dataset of sequential data and is able to learn the complex patterns in the data that are indicative of the underlying structure. The Transformer has been shown to outperform traditional methods for sequential data analysis and is able to model the temporal dependencies in the data.

### 3.3.1 EQTransformer

Earthquake Transformer (EQTransformer) is a deep learning model for the detection and classification of seismic signals in continuous waveforms. The model is based on a Transformer architecture that takes as input a raw waveform and outputs the probability of the presence of a seismic signal at each time step. The model is trained on a large dataset of seismic waveforms with labeled signal picks and is able to learn the complex patterns in the data that are indicative of seismic signals. The model is designed to be robust to noise and can generalize well to unseen data. EQTransformer has been shown to outperform traditional methods for signal detection and classification and is able to detect seismic signals with high accuracy and low latency.

# Chapter 4

# Discussion & Results

## 4.1 Pick detections

The performance of the models was evaluated using the following metrics: the cumulative number of detections, the multi-class classification results, the true positive, false negative, and recall, and the time displacement. The metrics are used to assess the performance of the models on the test set and compare the performance of the models with each other.

### 4.1.1 Classification results

The classification results of the models were evaluated using the confusion matrix. The confusion matrix shows the true positive, false negative, false positive, and true negative values of the models. The confusion matrix is used to evaluate the performance of the models on the test set and compare the performance of the models with each other.

#### Cumulative number of detections

The cumulative number of detections is the total number of seismic phases detected by the models over time. The cumulative number of detections is used to evaluate the performance of the models on the test set and compare the performance of the models with each other.

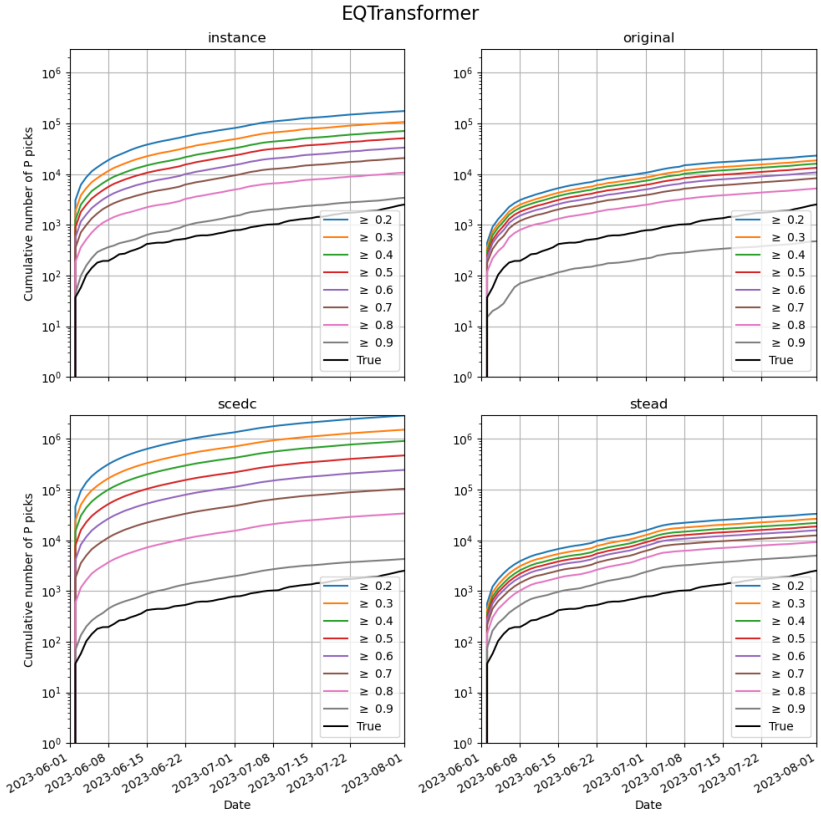



Figure 4.1: Cumulative number of detections in time for the EQTransformer architecture comparing different trained datasets.

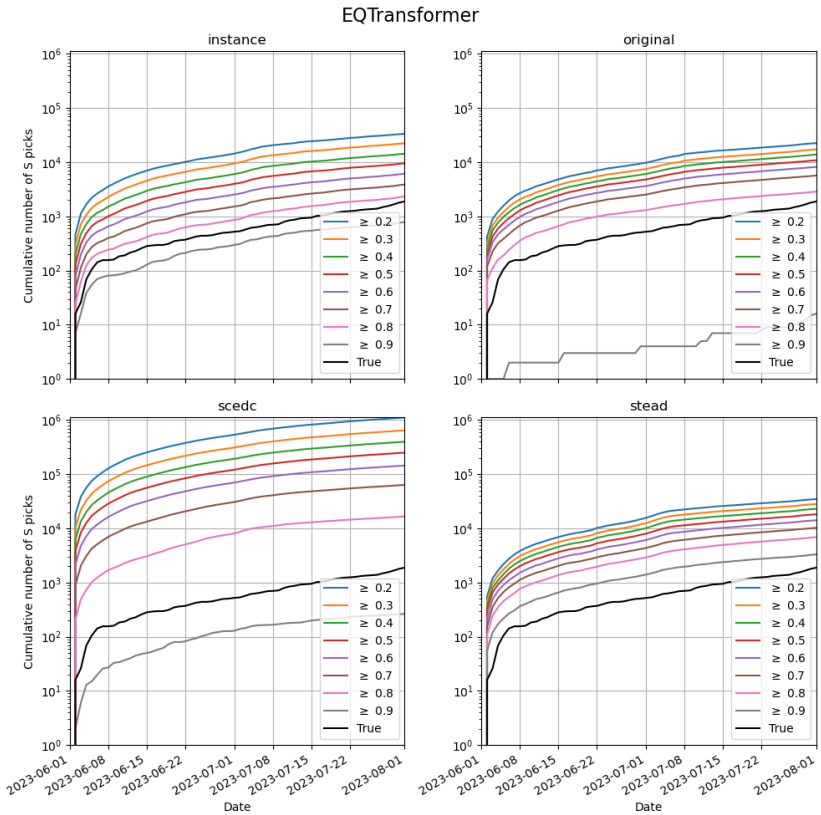



Figure 4.2: Cumulative number of detections in time for the EQTransformer architecture comparing different trained datasets.

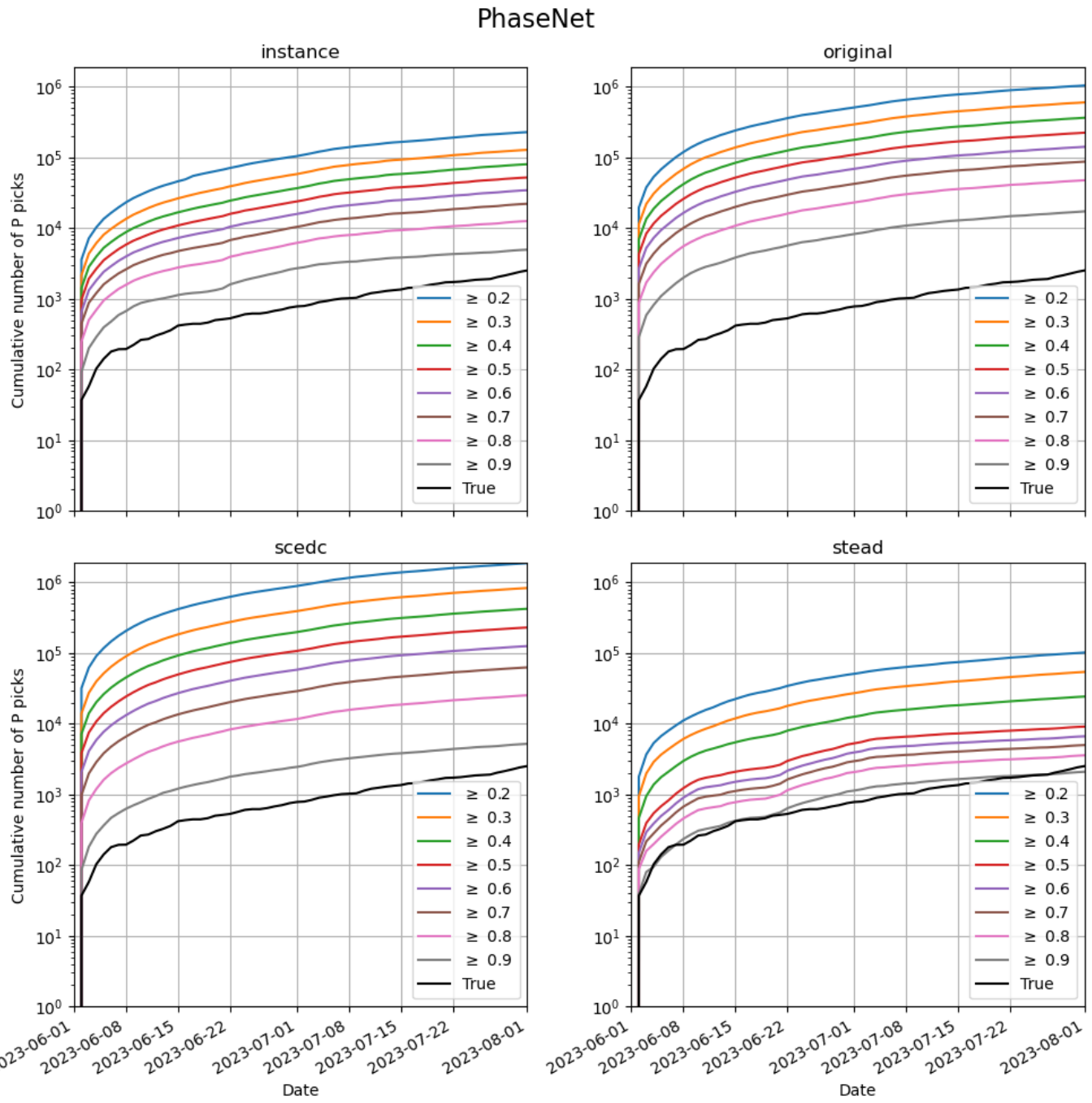



Figure 4.3: Cumulative number of detections in time for the PhaseNet architecture comparing different trained datasets.

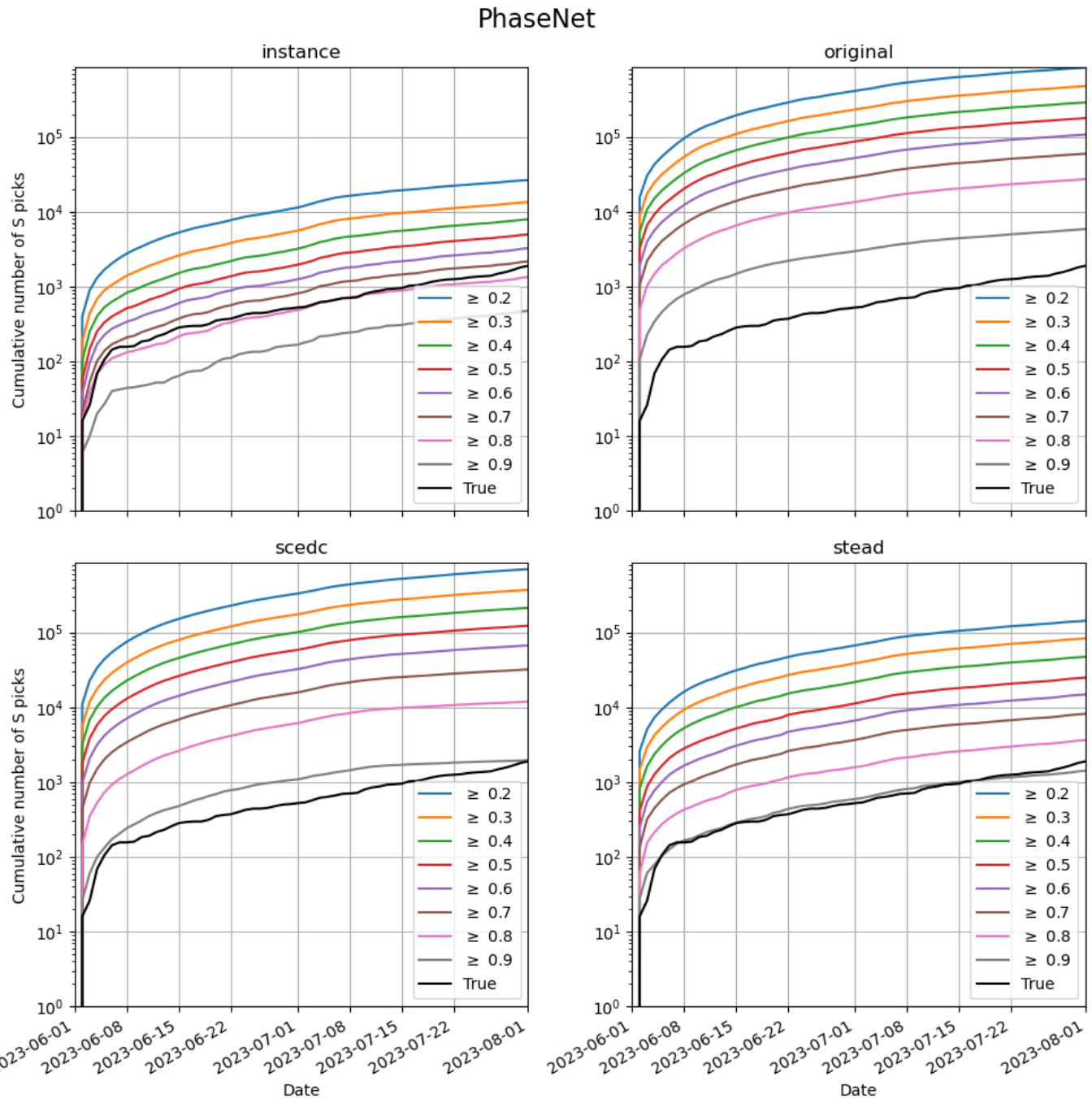



Figure 4.4: Cumulative number of detections in time for the PhaseNet architecture comparing different trained datasets.

**Confusion matrix**

The confusion matrix is a table that shows the true positive, false negative, false positive, and true negative values of the models. The confusion matrix is used to evaluate the performance of the models on the test set and compare the performance of the models with each other.

| | Pred. P | Pred. void |
|---|---|---|
| **True P** | TP | FN |
| **True void** | FP | TN |

| | Pred. S | Pred. void |
|---|---|---|
| **True S** | TP | FN |
| **True void** | FP | TN |

Table 4.1: Confusion matrix for the binary classification of P and S waves, left and right, respectively.

| | Pred. P | Pred. S | Pred. void |
|---|---|---|---|
| **True P** | cTP | pTP | FN |
| **True S** | pTP | cTP | FN |
| **True void** | FP | FP | TN |

Table 4.2: Confusion matrix for the multilabel classification of seismic waveforms.

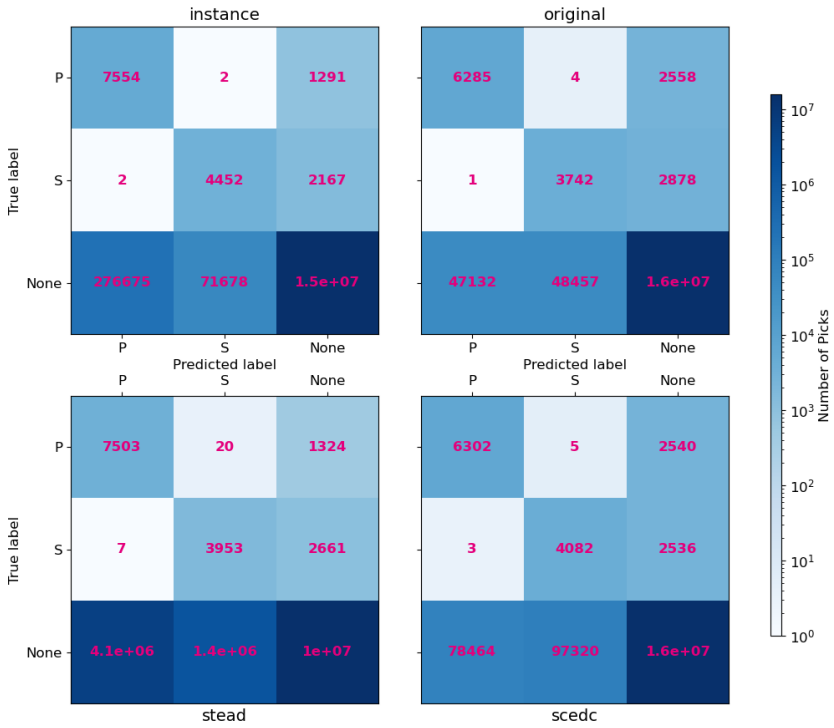



Figure 4.5: Multi-class classification confusion matrix of the EQTransformer architecture with 0.3 threshold.

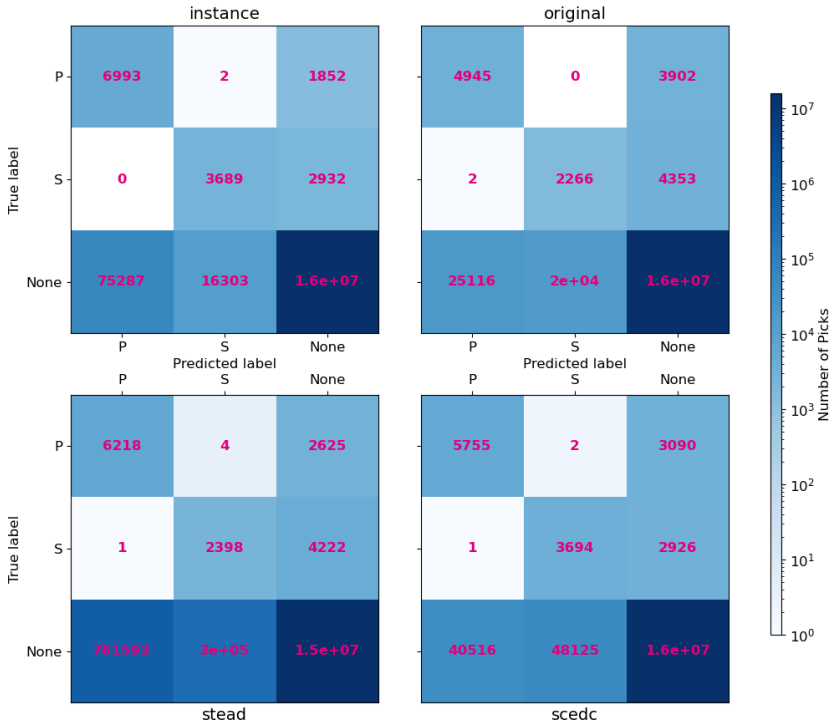



Figure 4.6: Multi-class classification confusion matrix of the EQTransformer architecture with 0.6 threshold.

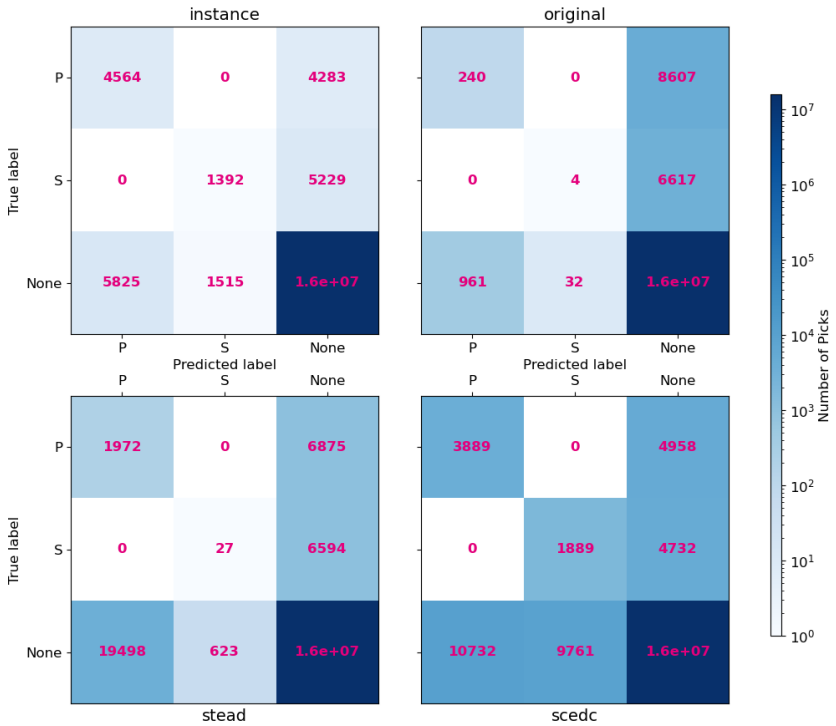



Figure 4.7: Multi-class classification confusion matrix of the EQTransformer architecture with 0.9 threshold.

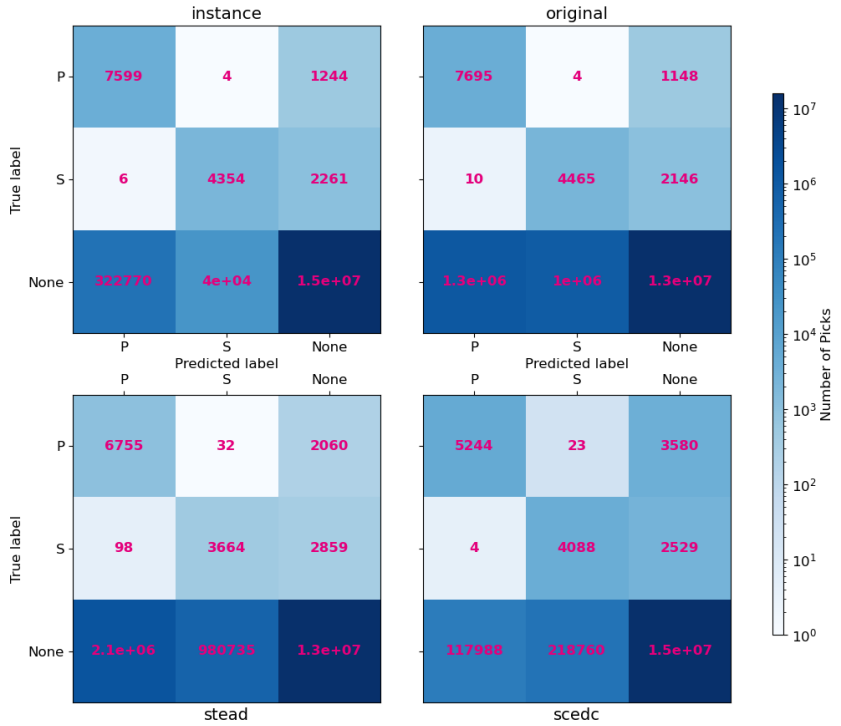



Figure 4.8: Multi-class classification confusion matrix of the PhaseNet architecture with 0.3 threshold.

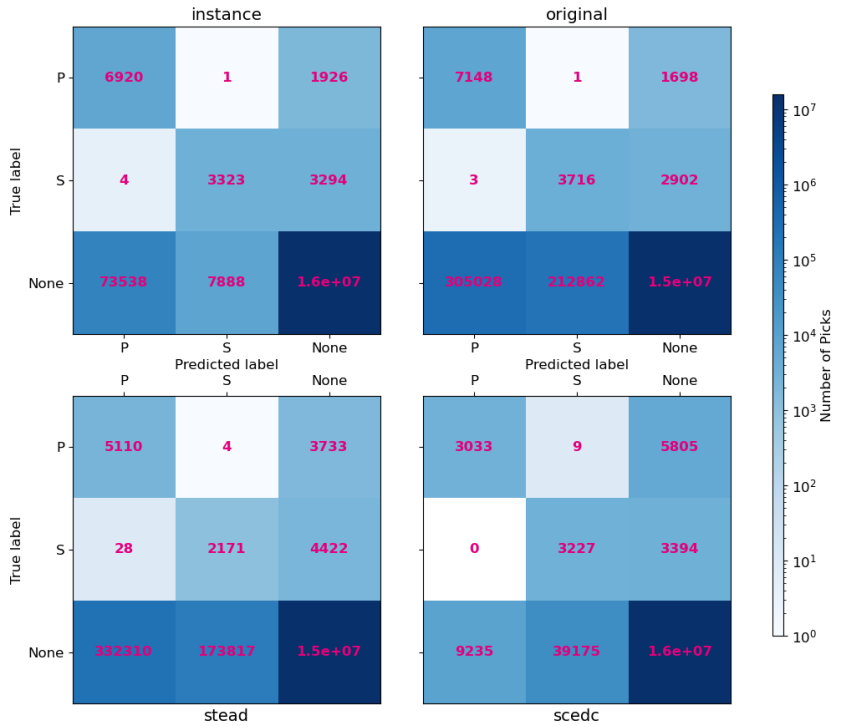



Figure 4.9: Multi-class classification confusion matrix of the PhaseNet architecture with 0.6 threshold.

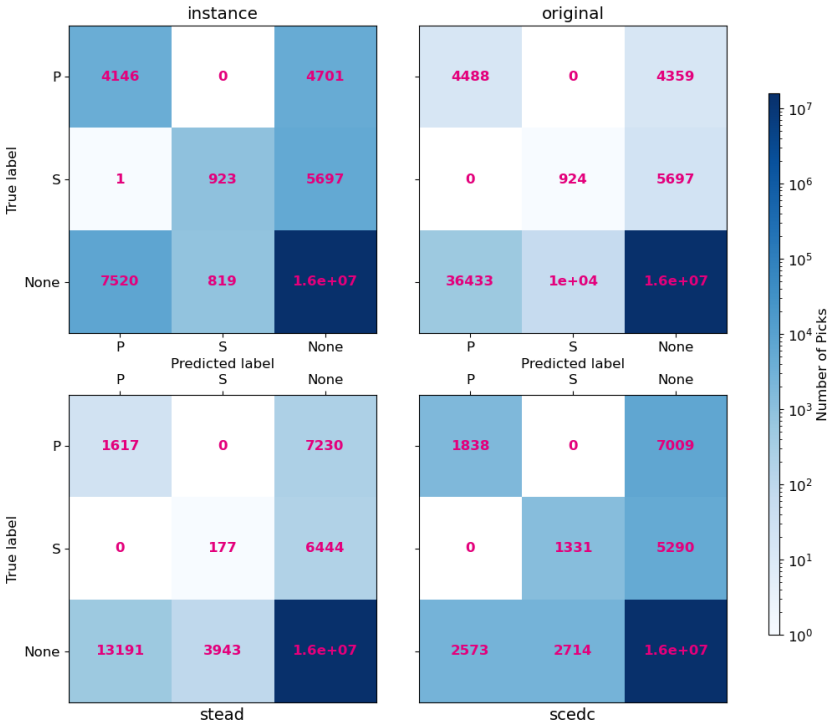



Figure 4.10: Multi-class classification confusion matrix of the PhaseNet architecture with 0.9 threshold.

### Metrics

The performance of the models was evaluated using the following metrics: the true positive, false negative, false positive, and true negative values of the models. The metrics are used to assess the performance of the models on the test set and compare the performance of the models with each other.

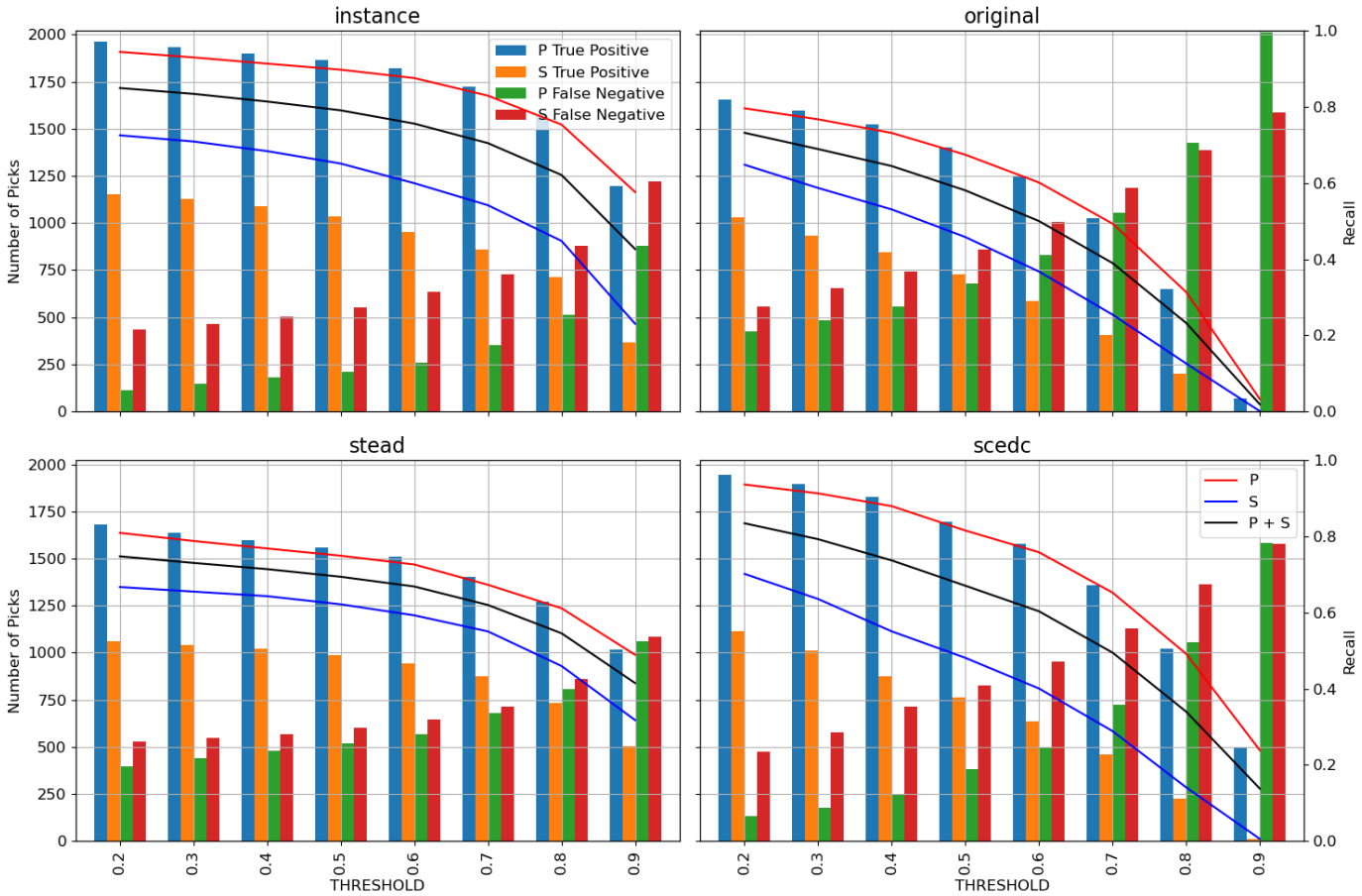



Figure 4.11: True Positive, False Negative detections and the recall per threshold for the EQTransformer architecture trained on the instance dataset.

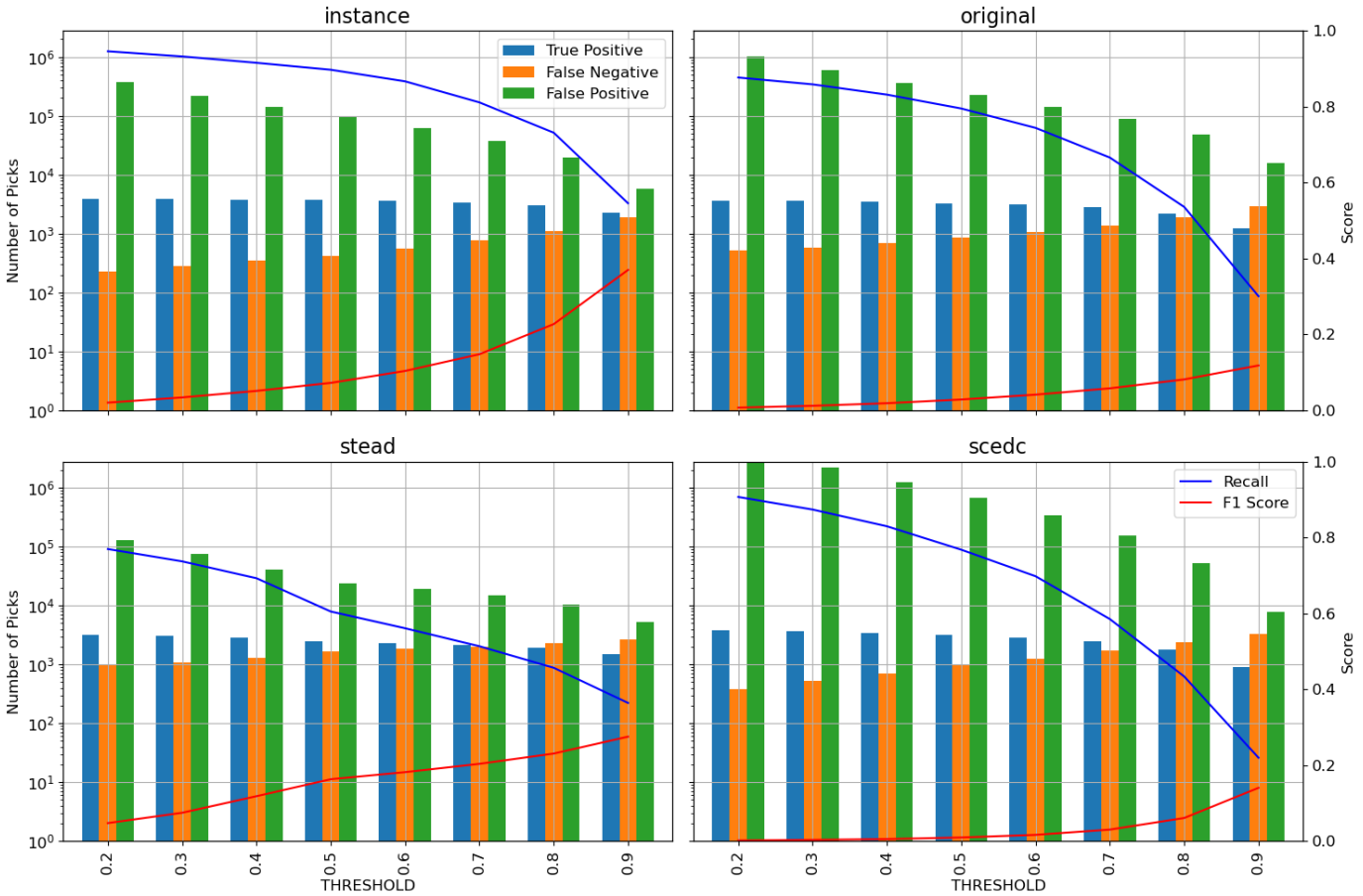



Figure 4.12: True Positive, False Negative detections and the recall per threshold for the EQTransformer architecture trained on the instance dataset.

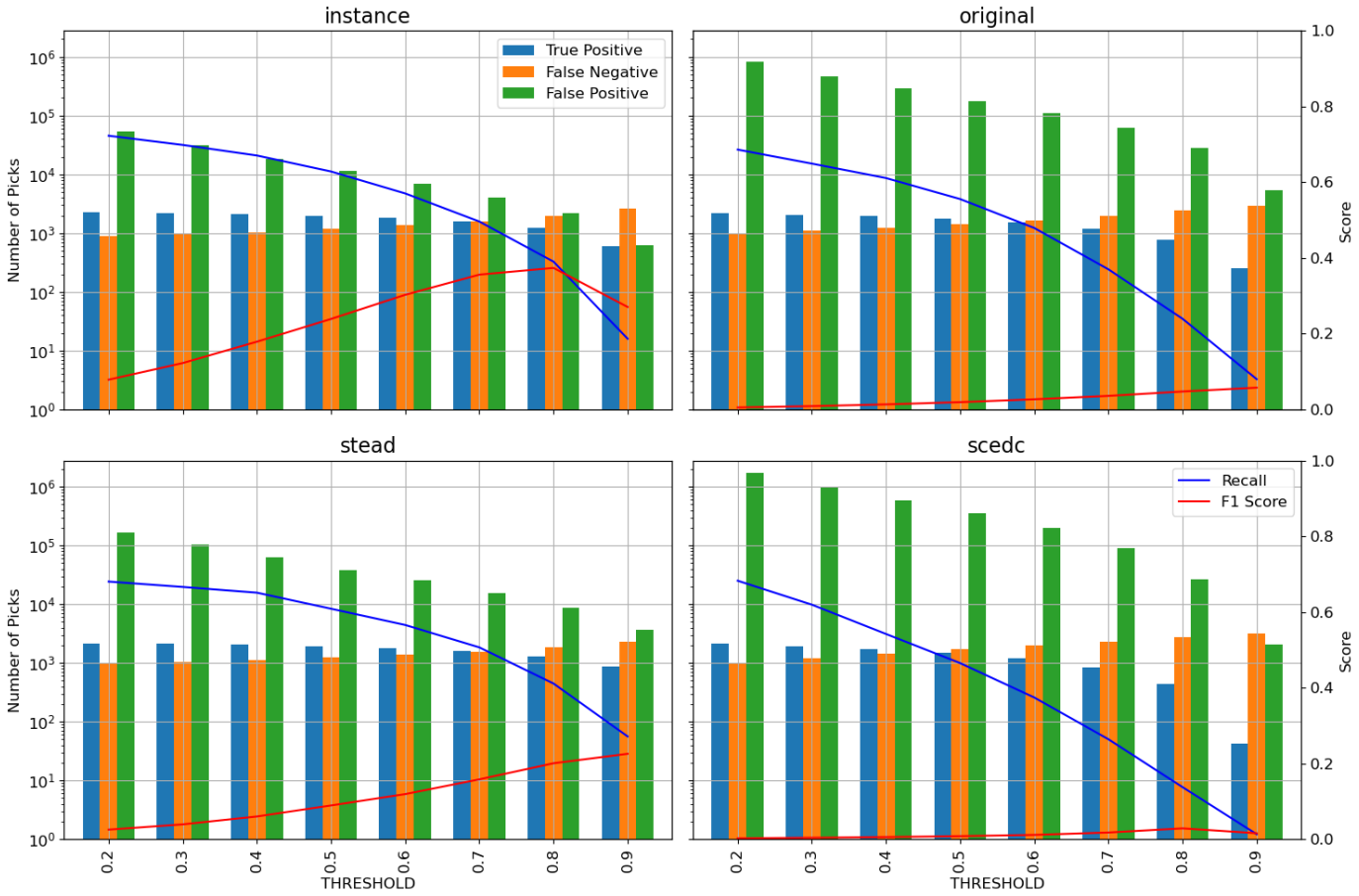



Figure 4.13: True Positive, False Negative detections and the recall per threshold for the EQTransformer architecture trained on the instance dataset.

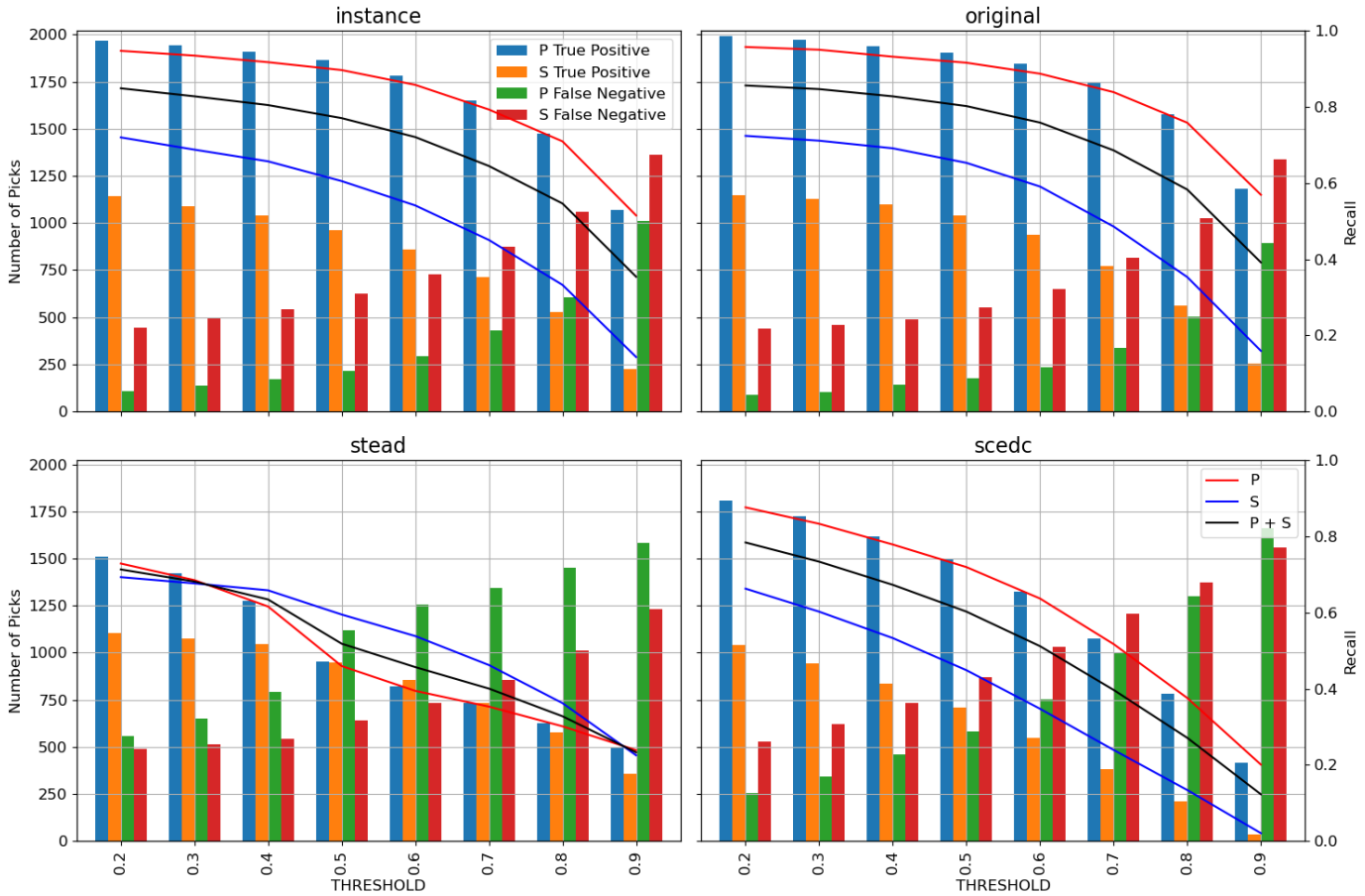



Figure 4.14: True Positive, False Negative detections and the recall per threshold for the PhaseNet architecture trained on the stead dataset.

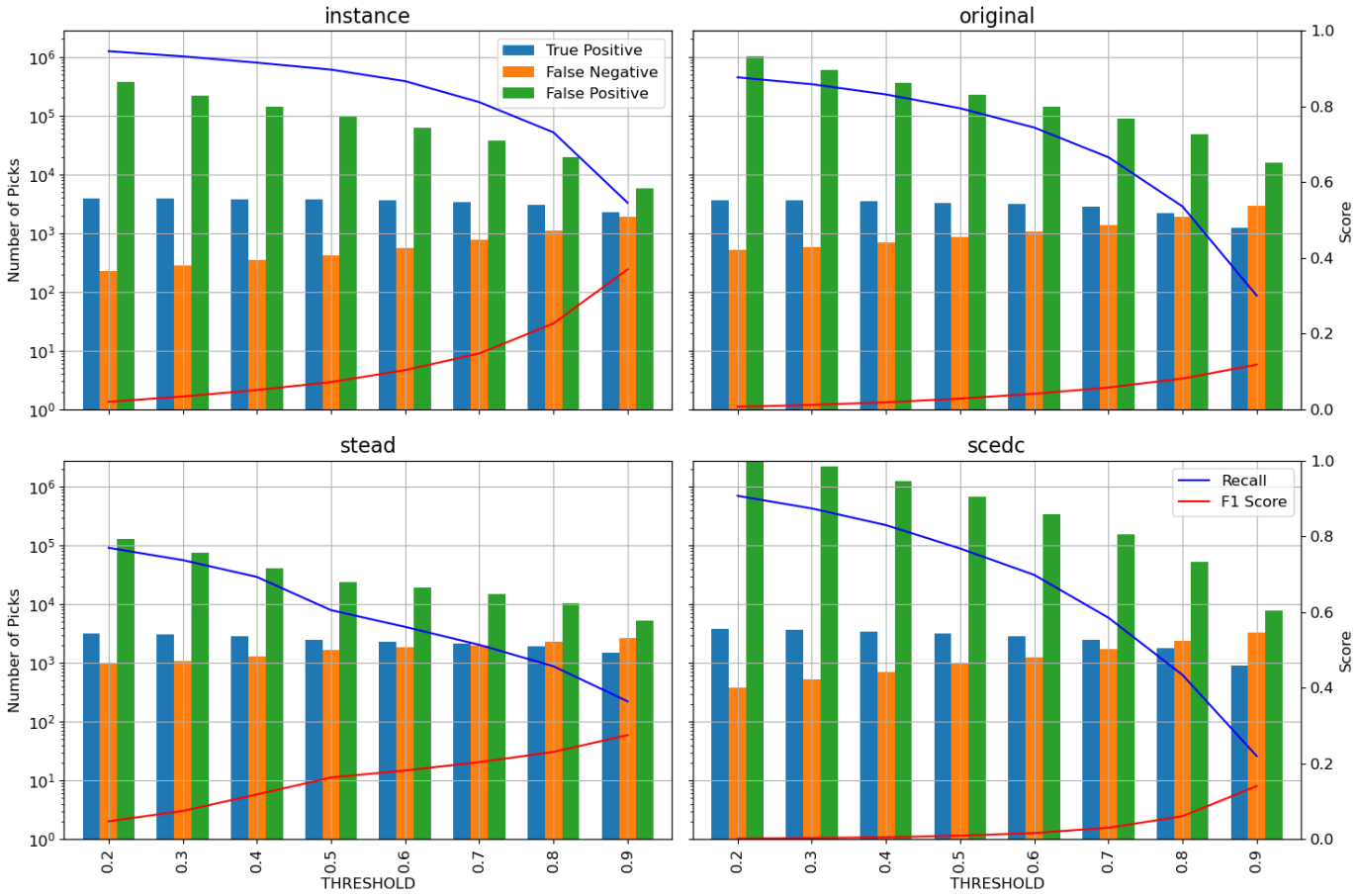



Figure 4.15: True Positive, False Negative detections and the recall per threshold for the PhaseNet architecture trained on the stead dataset.

| Model | True Positive | False Negative | Recall | F1 |
|---|---|---|---|---|
| EQTransformer instance | 0.0 | 0.0 | 0.0 | 0.0 |
| EQTransformer original | 0.0 | 0.0 | 0.0 | 0.0 |
| EQTransformer scedc | 0.0 | 0.0 | 0.0 | 0.0 |
| EQTransformer stead | 0.0 | 0.0 | 0.0 | 0.0 |
| PhaseNet original | 0.0 | 0.0 | 0.0 | 0.0 |
| PhaseNet instance | 0.0 | 0.0 | 0.0 | 0.0 |
| PhaseNet scedc | 0.0 | 0.0 | 0.0 | 0.0 |
| PhaseNet stead | 0.0 | 0.0 | 0.0 | 0.0 |

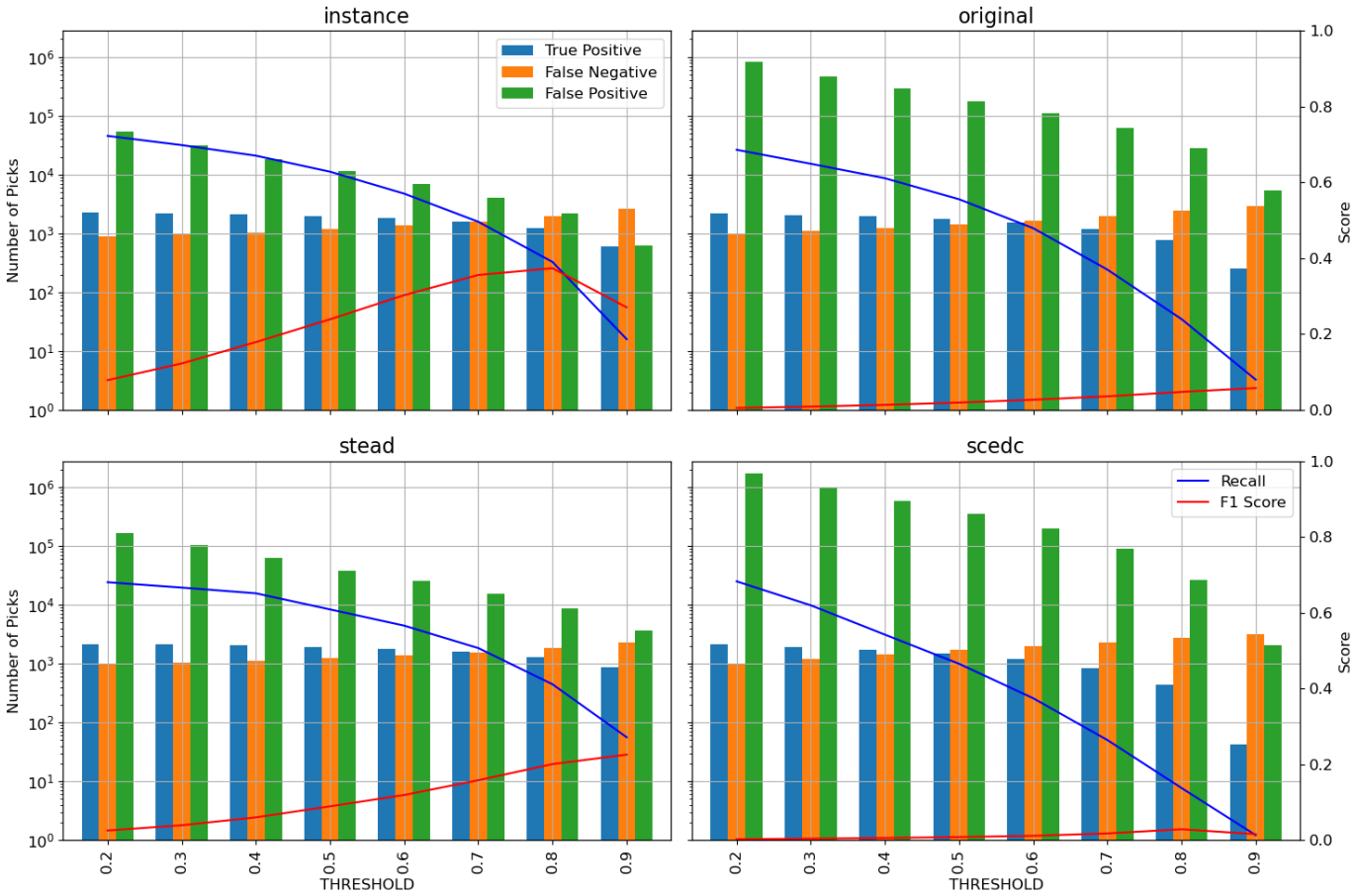



Figure 4.16: True Positive, False Negative detections and the recall per threshold for the PhaseNet architecture trained on the stead dataset.

### 4.1.2 Time displacement

The time displacement is the difference between the predicted arrival times of the seismic phases and the true arrival times of the seismic phases. The time displacement is used to evaluate the performance of the models on the test set and compare the performance of the models with each other.

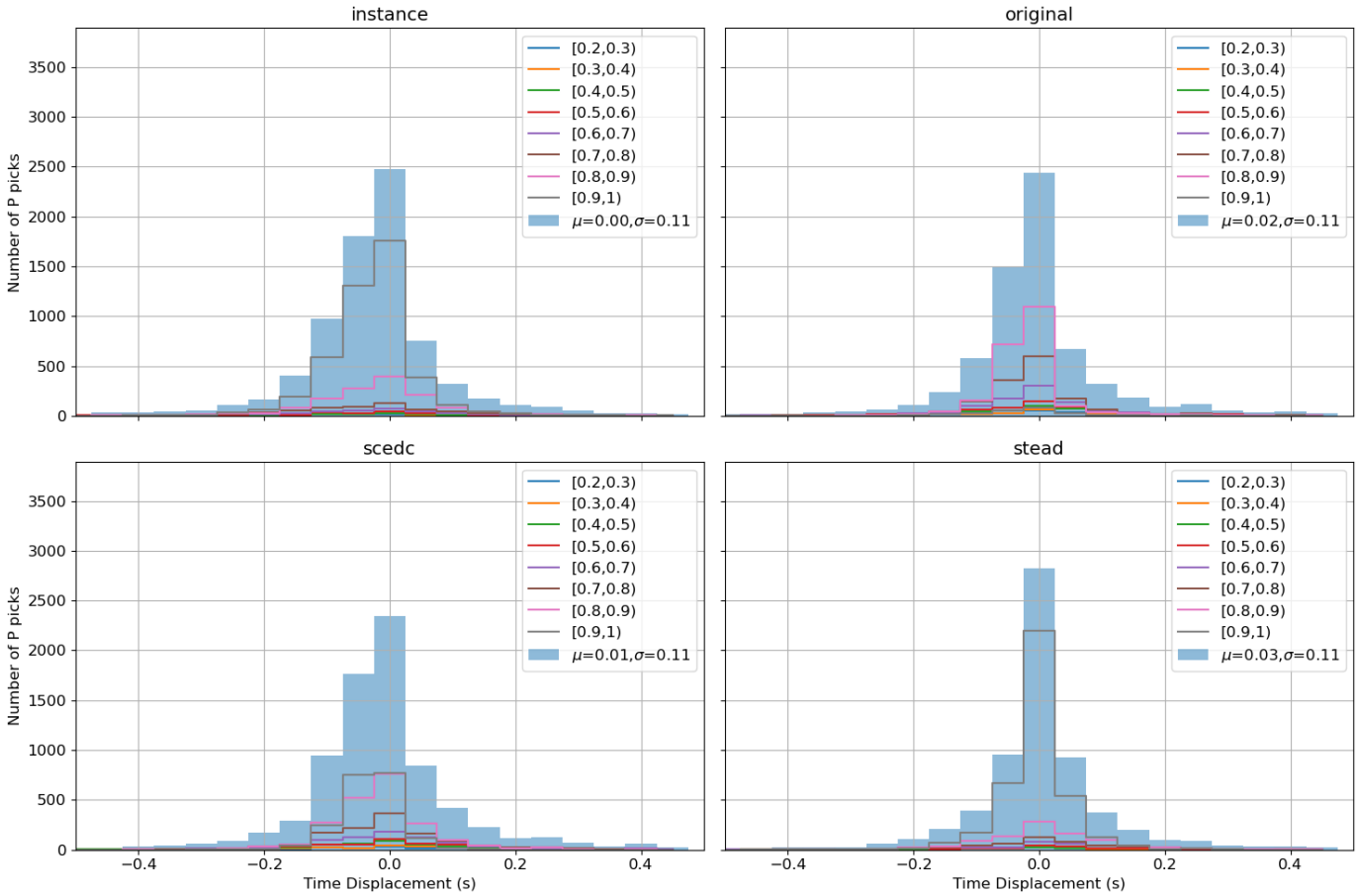



Figure 4.17: Time difference for the True Positive detections for the EQTransformer architecture trained on the stead dataset.

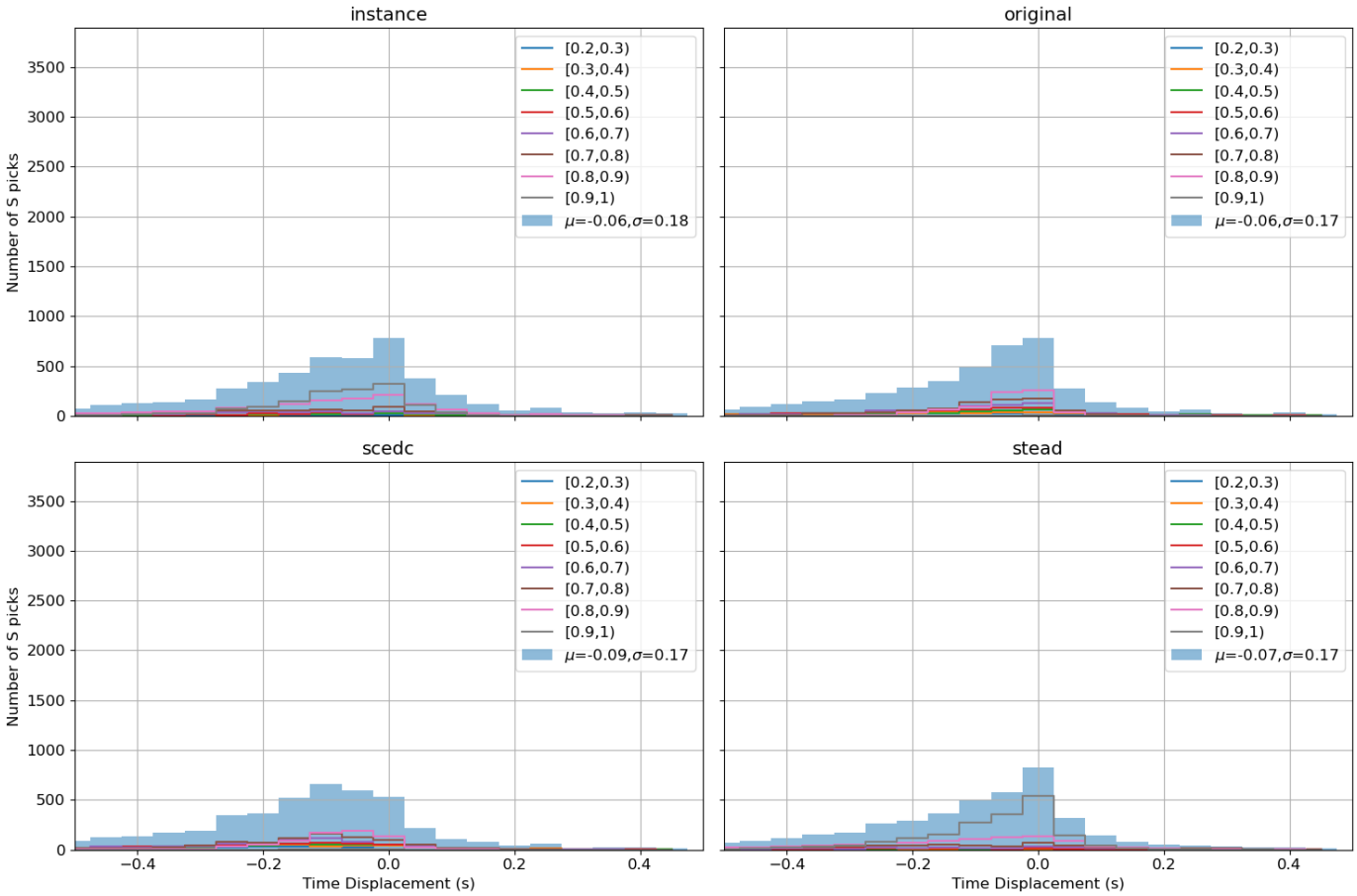



Figure 4.18: Time difference for the True Positive detections for the EQTransformer architecture trained on the stead dataset.

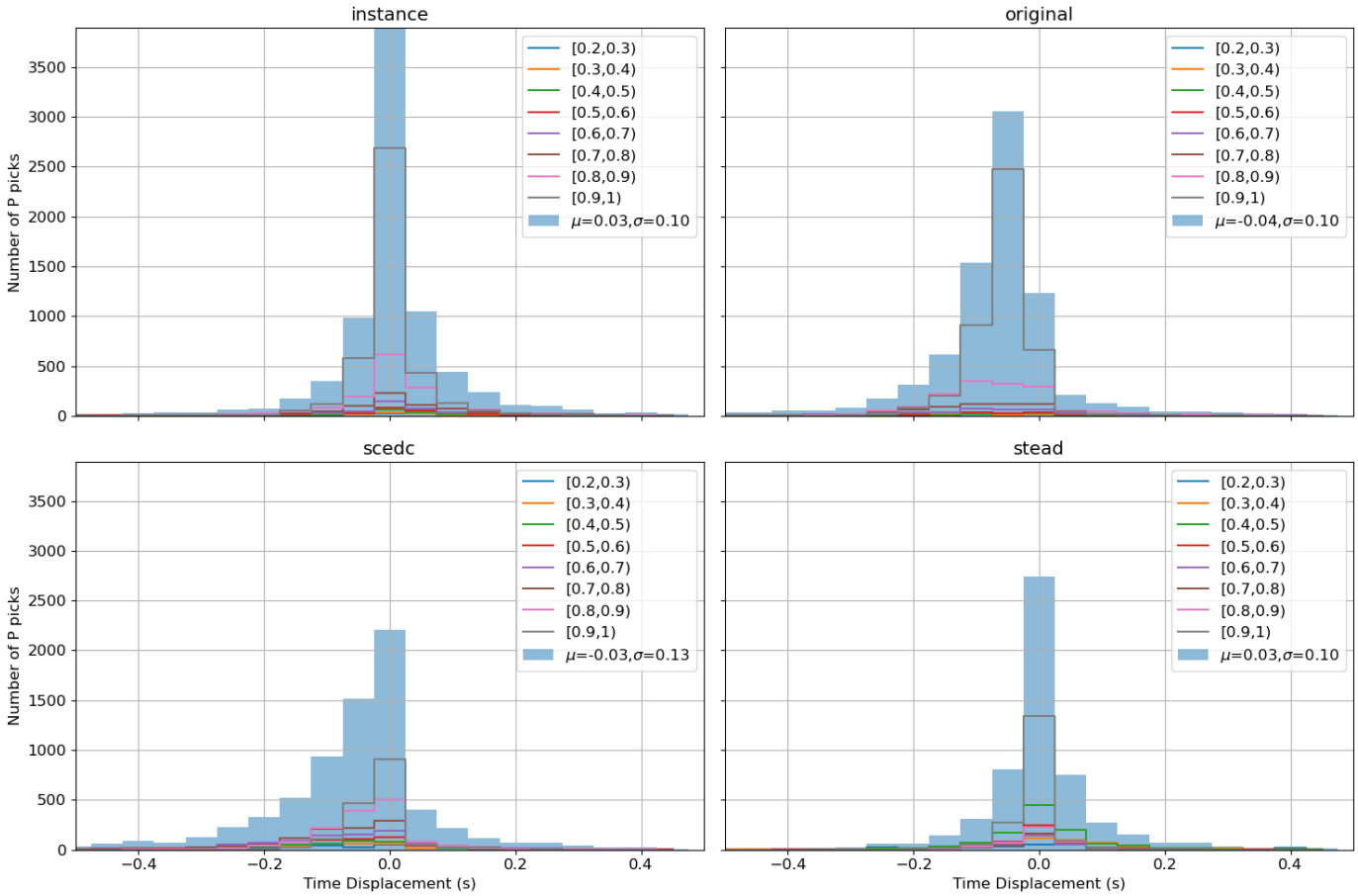



Figure 4.19: Time difference for the True Positive detections for the PhaseNet architecture trained on the stead dataset.

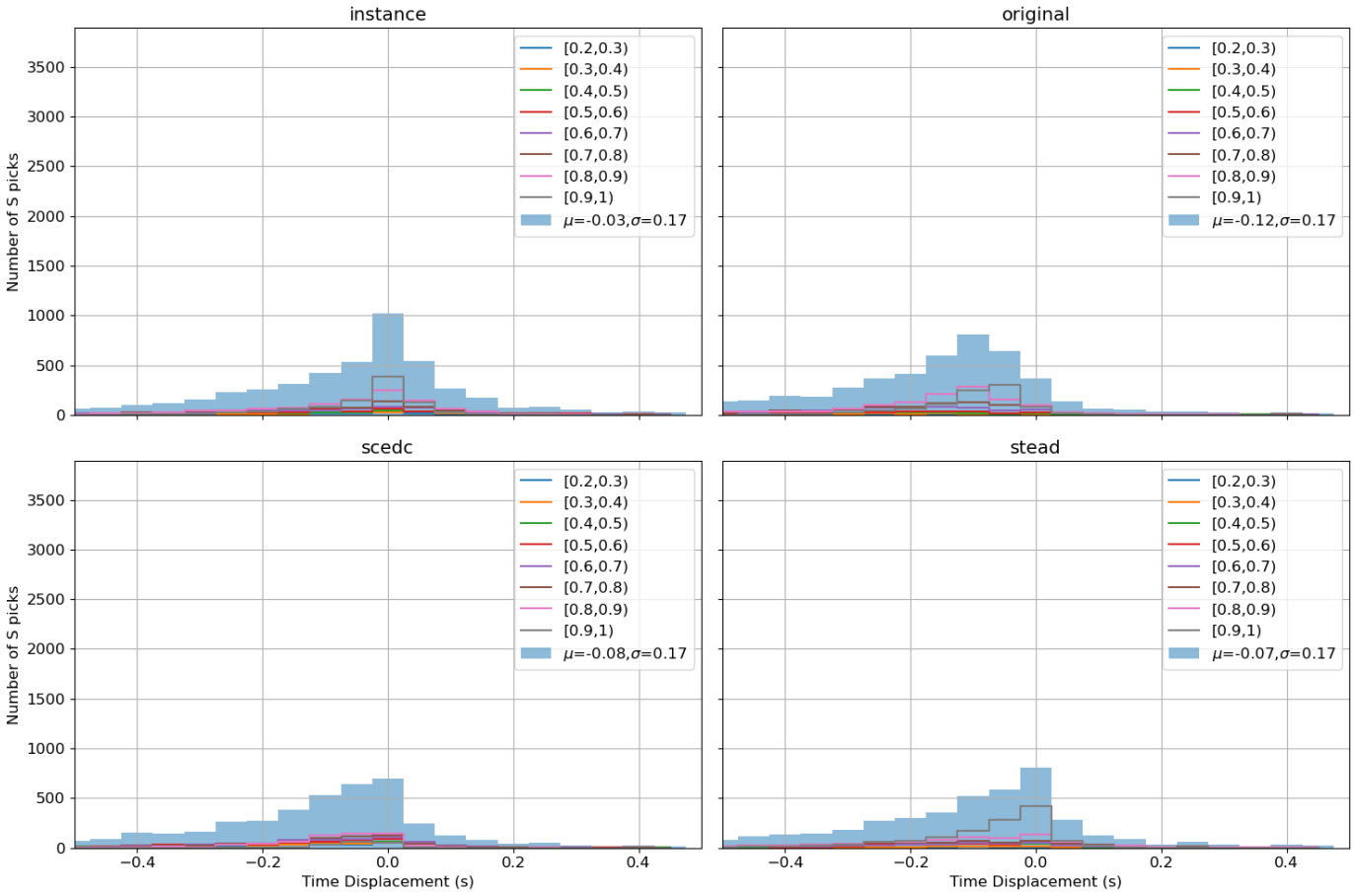



Figure 4.20: Time difference for the True Positive detections for the PhaseNet architecture trained on the stead dataset.

## 4.2 Phase association

The phase association is the process of associating the detected seismic phases with the true seismic phases. The phase association is used to evaluate the performance of the models on the test set and compare the performance of the models with each other.

# Chapter 5

# Computation Performance

# Chapter 6

# Cloud Infrastructure

The cloud infrastructure is provided by CINECA on Ada Cloud under the research project IscrC_AI4Seism_C. The cloud infrastructure is provided on a DUal-Socket Dell PowerEdge platform under the management of OpenStack version Zed. The infrastructure consists of a cluster of 71 nodes, each equipped with 2 Intel CascadeLake 8260 processors, 24 cores each, operating at 2.4 GHz, with Hyperthreading enabled for a total of 48 cores per node. Each node is equipped with 768 GB of DDR4 RAM. The nodes are interconnected by a 100 Gbps Ethernet network.

# Chapter 7

# Conclusions