PHOTOPLETHYSMOGRAPHY SIGNALS ANALYSIS WITH ML/DL-BASED SOLUTION

DR. GIRALDI ANGELA

PROJECT OVERVIEW

1

Research

Initial phase of research in literature about photoplethysmography (PPG) signals & existing ML approaches utilising PPG data

2

Data Analysis

Perform exploratory data analysis to understand the characteristics of the input data and target variable

3

Preprocessing

Explore the need to apply necessary preprocessing steps (e.g. filtering, segmentation, normalisation...) to prepare the data for ML models

4

Models Development

Set up various ML/DL architectures and approaches to predict the continuous labels in this time series regression challenge

5

Model Selection

Evaluation and selection of the best performing model based on appropriate metrics

6

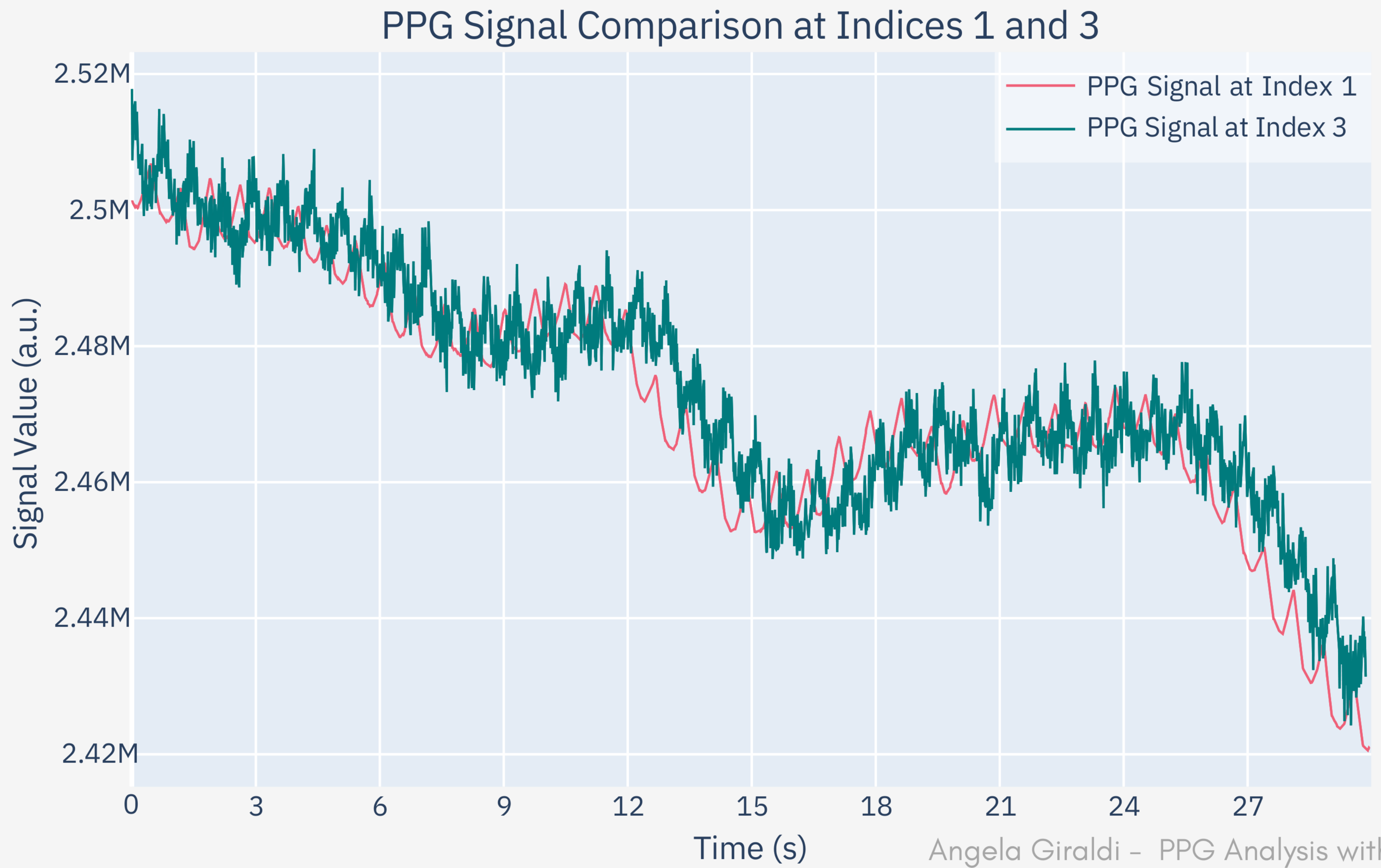
Summary& Outlook

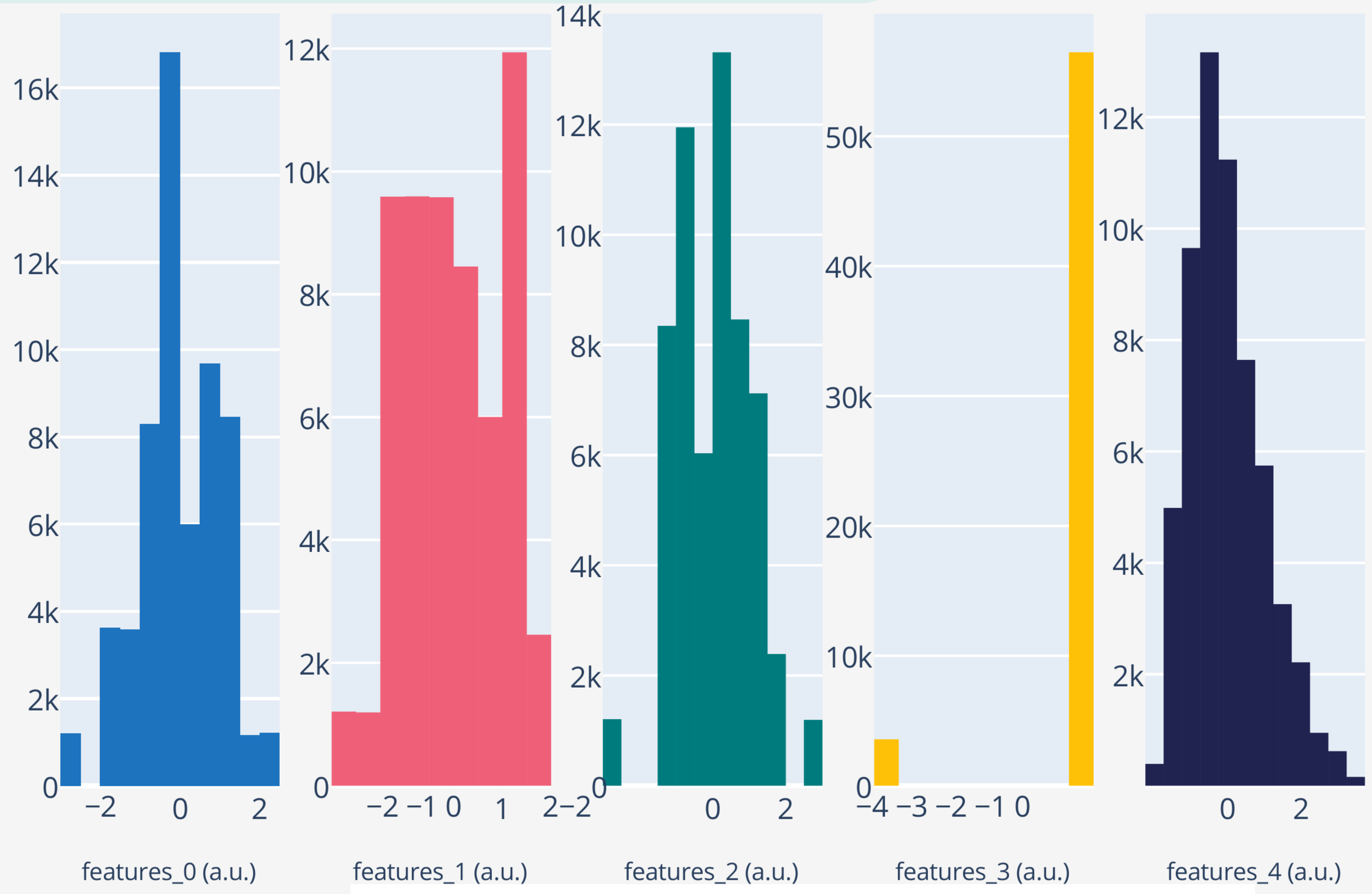
Summary of findings, potential improvements and future directions

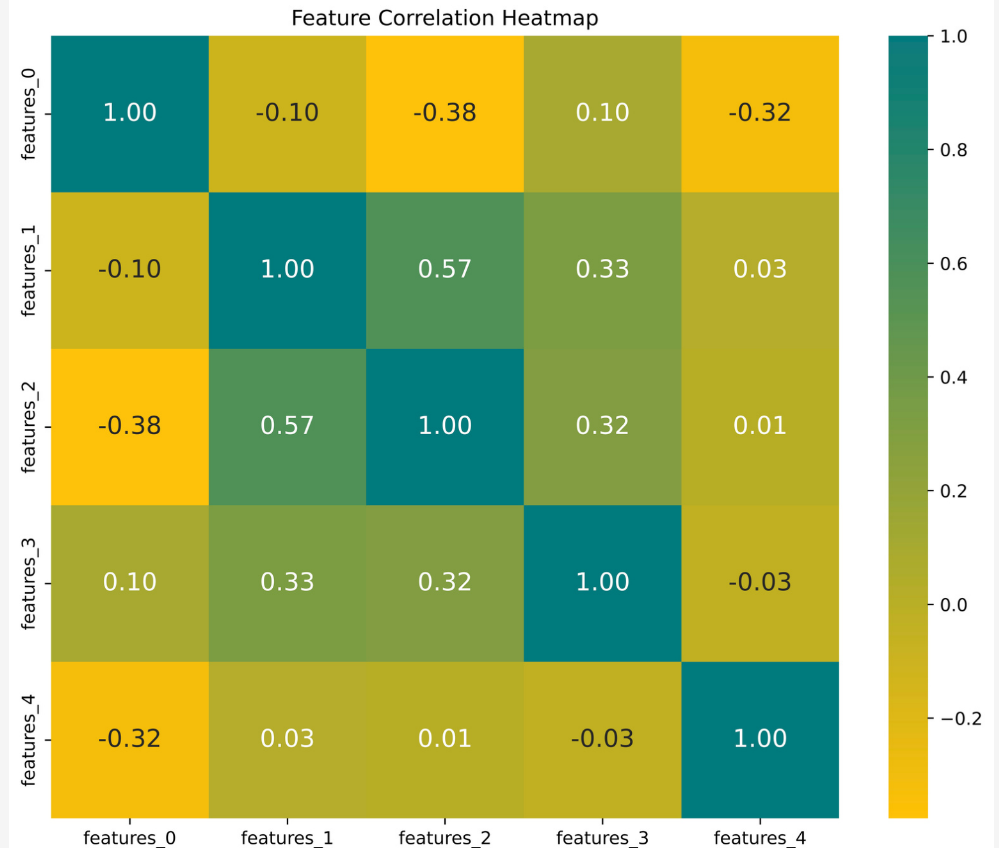
INPUTS & PREPROCESSING

01 - INPUT DATA

- **60,000 PPG signals**, each a 30-second time series sampled at 100 Hz
  - Illustration of signal variability
- Additional **5 engineered features** and their **correlations**





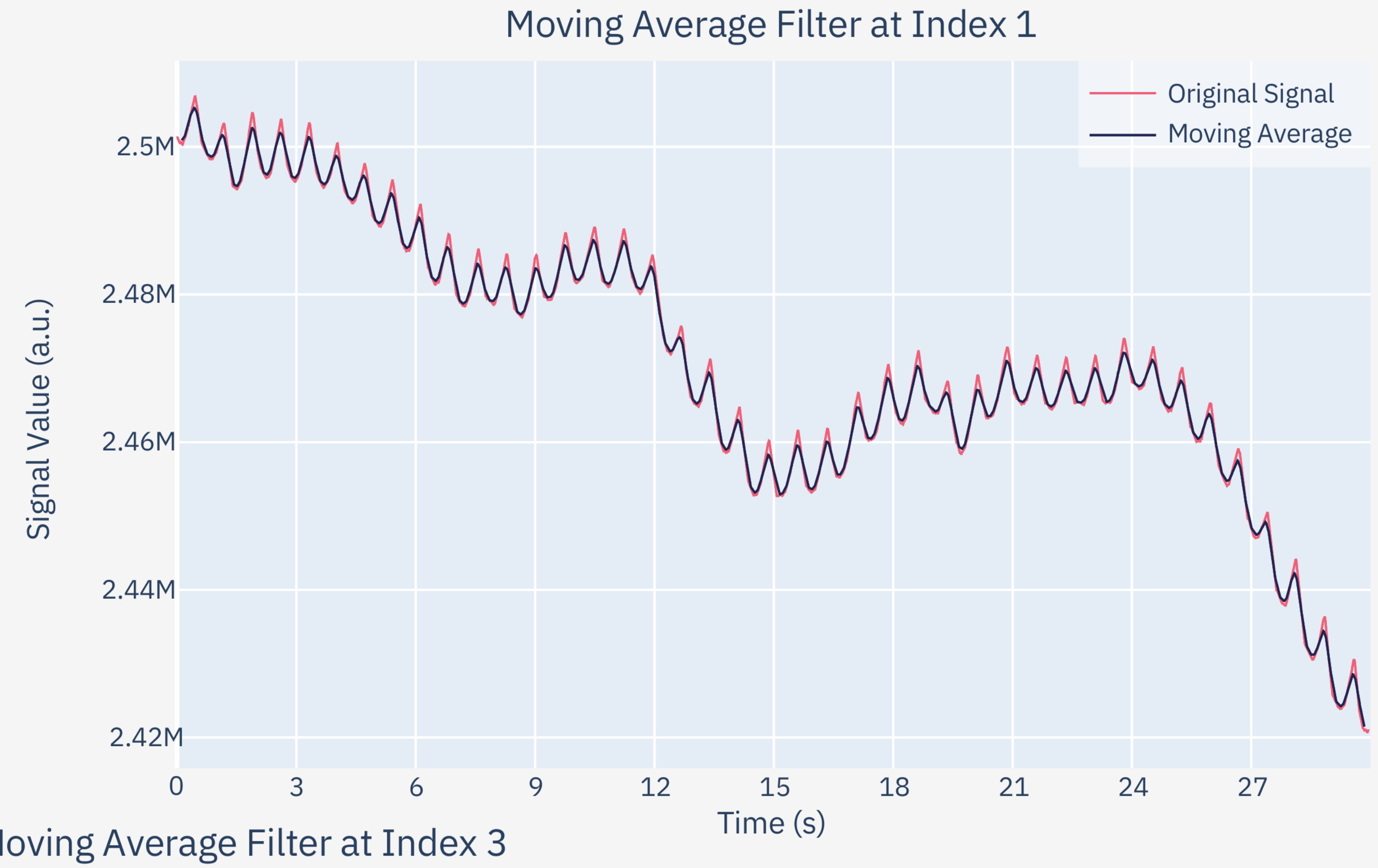


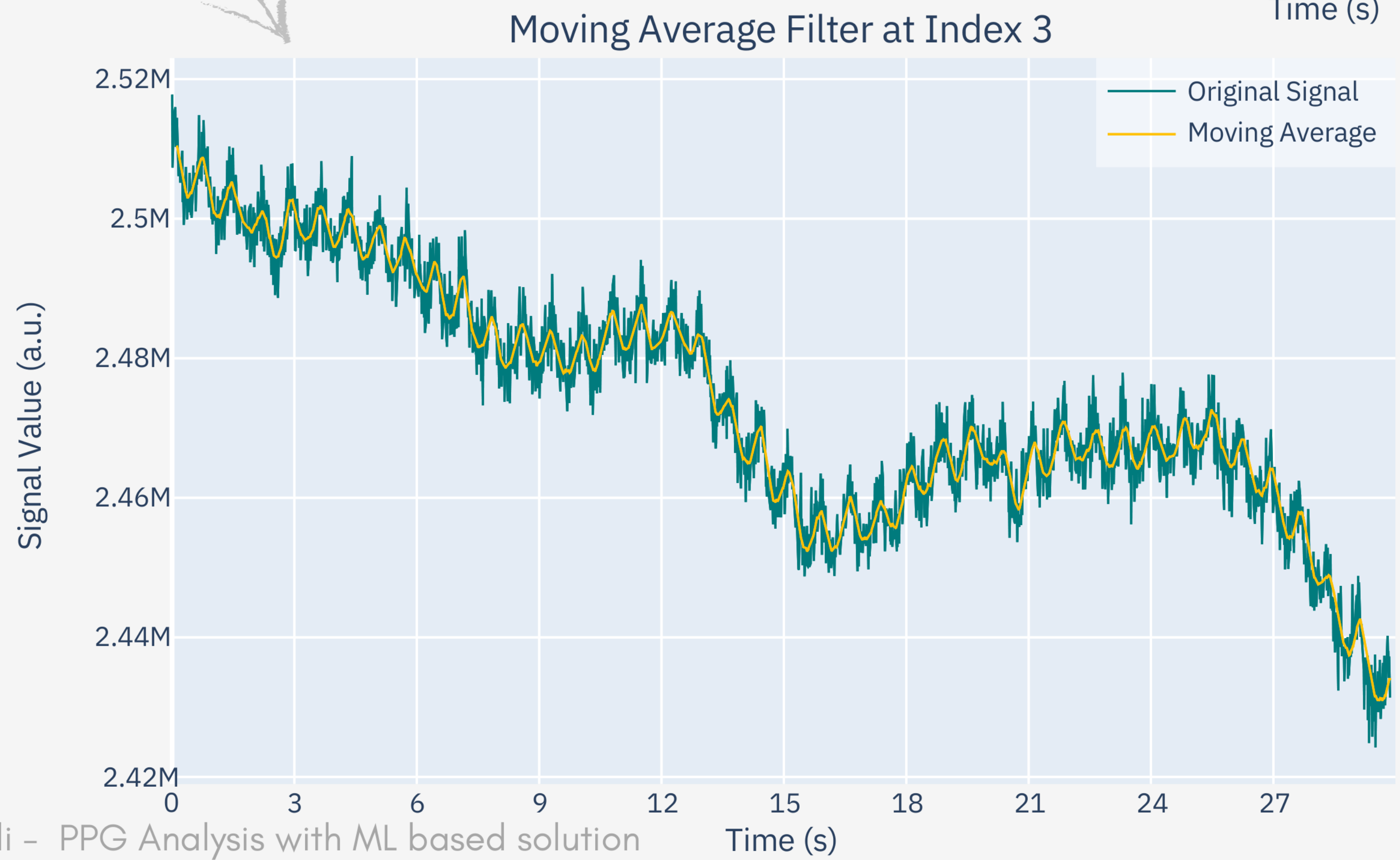
INPUTS & PREPROCESSING

01 - INPUT DATA

02 - PREPROCESSING

- Applied alternatively:
  - **Moving Average Data Filter**: Applied to smooth the data and reduce noise
  - **Low-pass Filter**: Used to eliminate high-frequency noise from the signals
- **Quality Assumption**: All signals are assumed to be of good quality
- **Normalization**: Essential to ensure consistency in signal amplitude
- Applied same steps also to test data!
- **Data Storage**: Preprocessed data saved in .h5 format for efficient handling





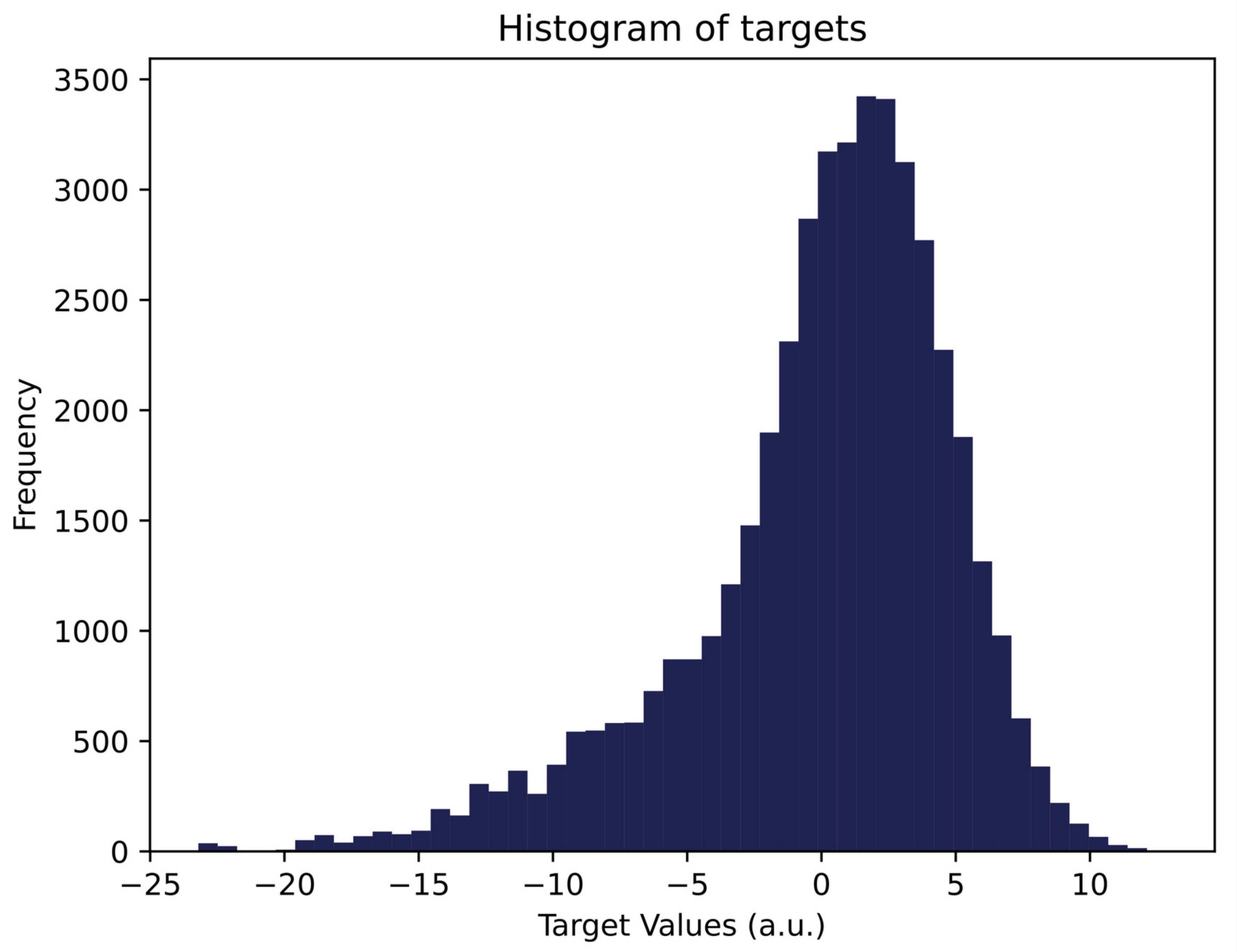
INPUTS & PREPROCESSING

01 - INPUT DATA

02 - PREPROCESSING

03 - TARGET

- **Continous labels**
  - Range [−23.2, 12.8]
  - Mean 0.006
  - Median 0.89
  - Std 4.98



Given the target distribution, it might help to weight the inputs so the ML models can learn better in regions with less statistics

MACHINE LEARNING MODELS

RF

RANDOM FOREST

It is versatile and effective in handling a large number of features and capturing non-linear relationships

RMSE*: 1.79

LSTM

LONG SHORT TERM MEMORY

It is suitable for time series data like PPG signals and able to capture long-term dependencies

RMSE: 5.18

RNN

RECURRENT NEURAL NETWORK

It is designed for sequential data, effective in capturing short-term patterns and learn contextual patterns

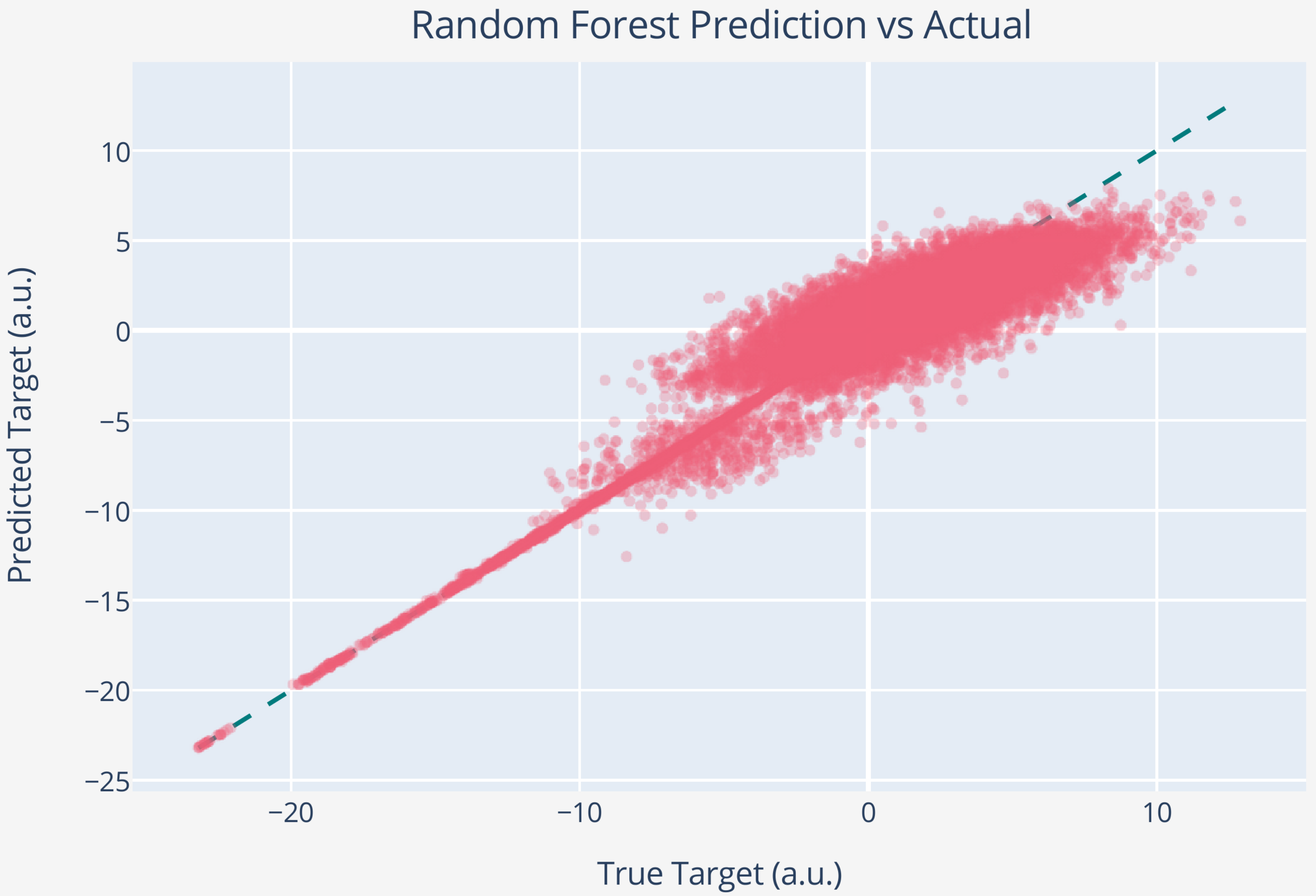
RMSE: 4.28

CNN

CONVOLUTIONAL NEURAL NETWORK

It is adaptable for time series data and efficiencient in detecting local patterns

RMSE: 4.71



Model Training General Details:

- Data Split: Training & Validation (25%)
- Batch Normalization: to improve stability and performance
- L2 Regularization: to prevent overfitting
- Optimizer: Adam, with a learning rate of 0.001-0.01

Hyperparameters optimization could be extremely helpful in finding the best architecture for the models

*root mean squared error (RMSE)

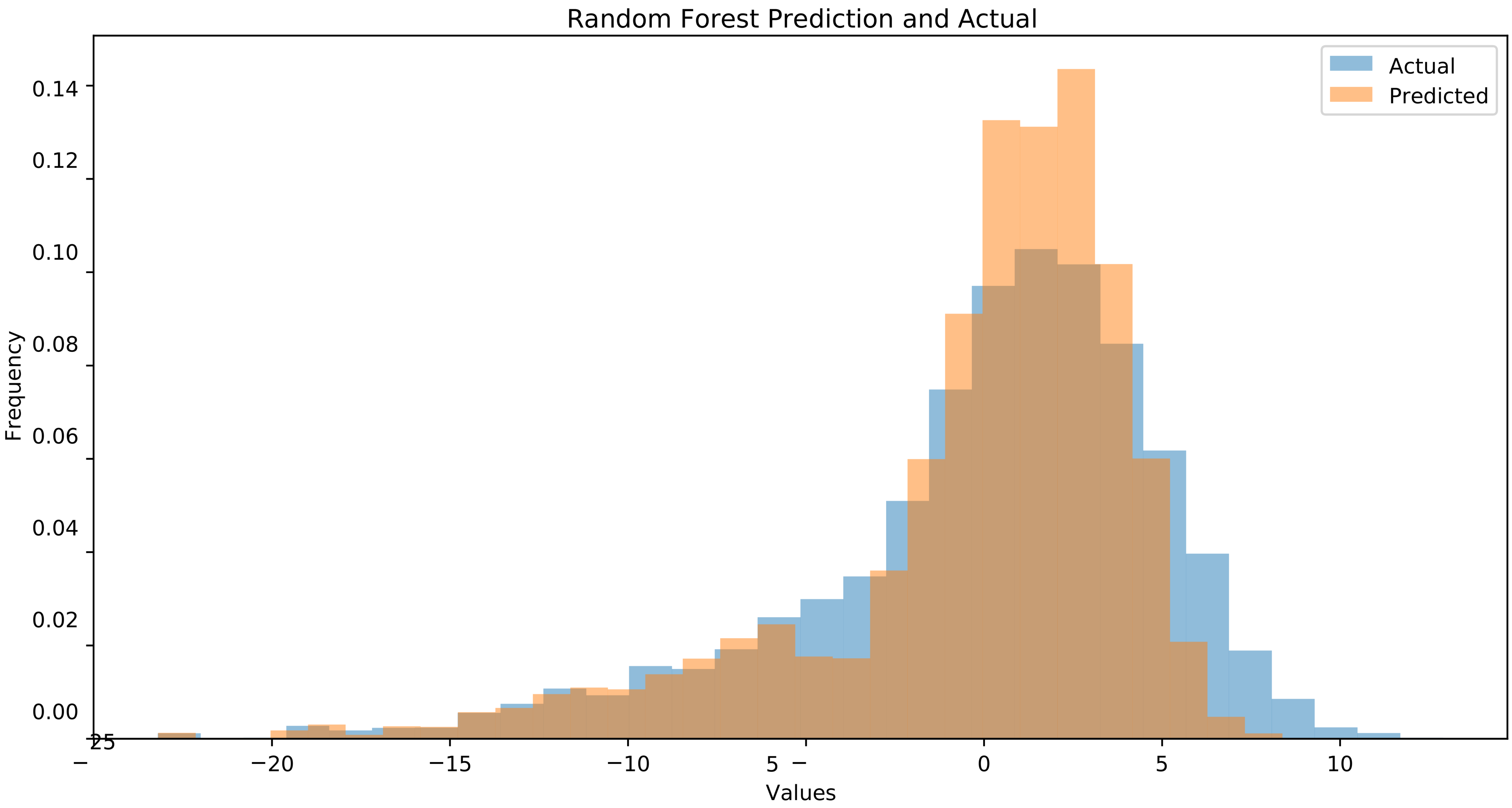
SUMMARY & OUTLOOK

- **Accomplishments**:
  - Analyzed, preprocessed, and utilized PPG signals for a time series regression task to predict a continuous target
  - After preliminary training of various ML models, the Random Forest was chosen for its better performance and robustness

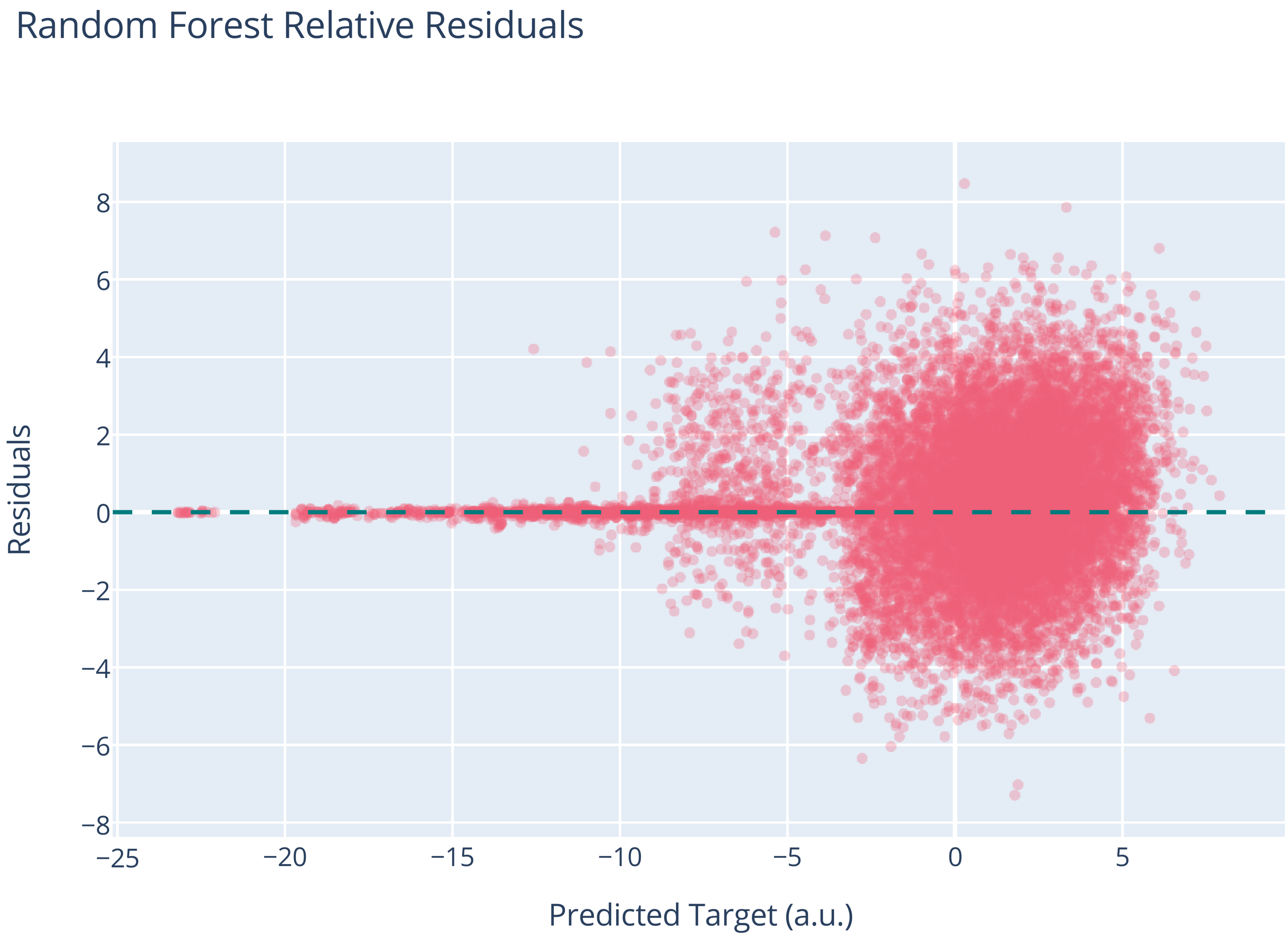
- **Potential Improvements**:
  - Explore different filtering methods in preprocessing
  - Implement weighting in training to address imbalances
  - Investigate combining models, such as using CNN for smart feature extraction followed by LSTM for time series analysis
  - Selecting most important input features to identify the ones with more predictive power
  - Consider adding NEW additional engineered features to enhance performance
  - Conduct hyperparameter optimization to fine-tune the models

THANK YOU

ADDITIONAL MATERIAL

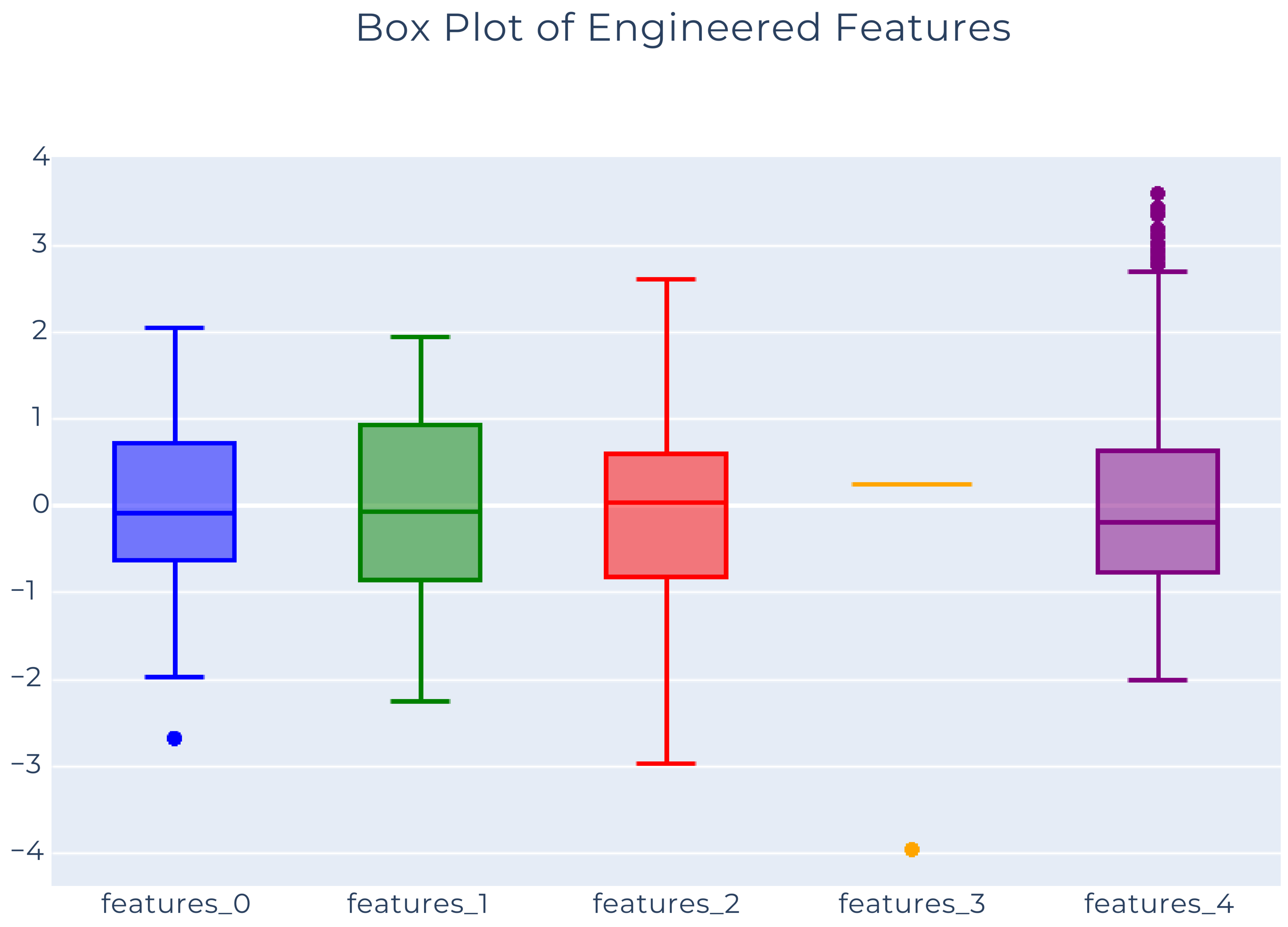


Random Forest Prediction and Actual
Actual
Predicted
Frequency
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−25
−20
−15
−10
−5
0
5
10
Values

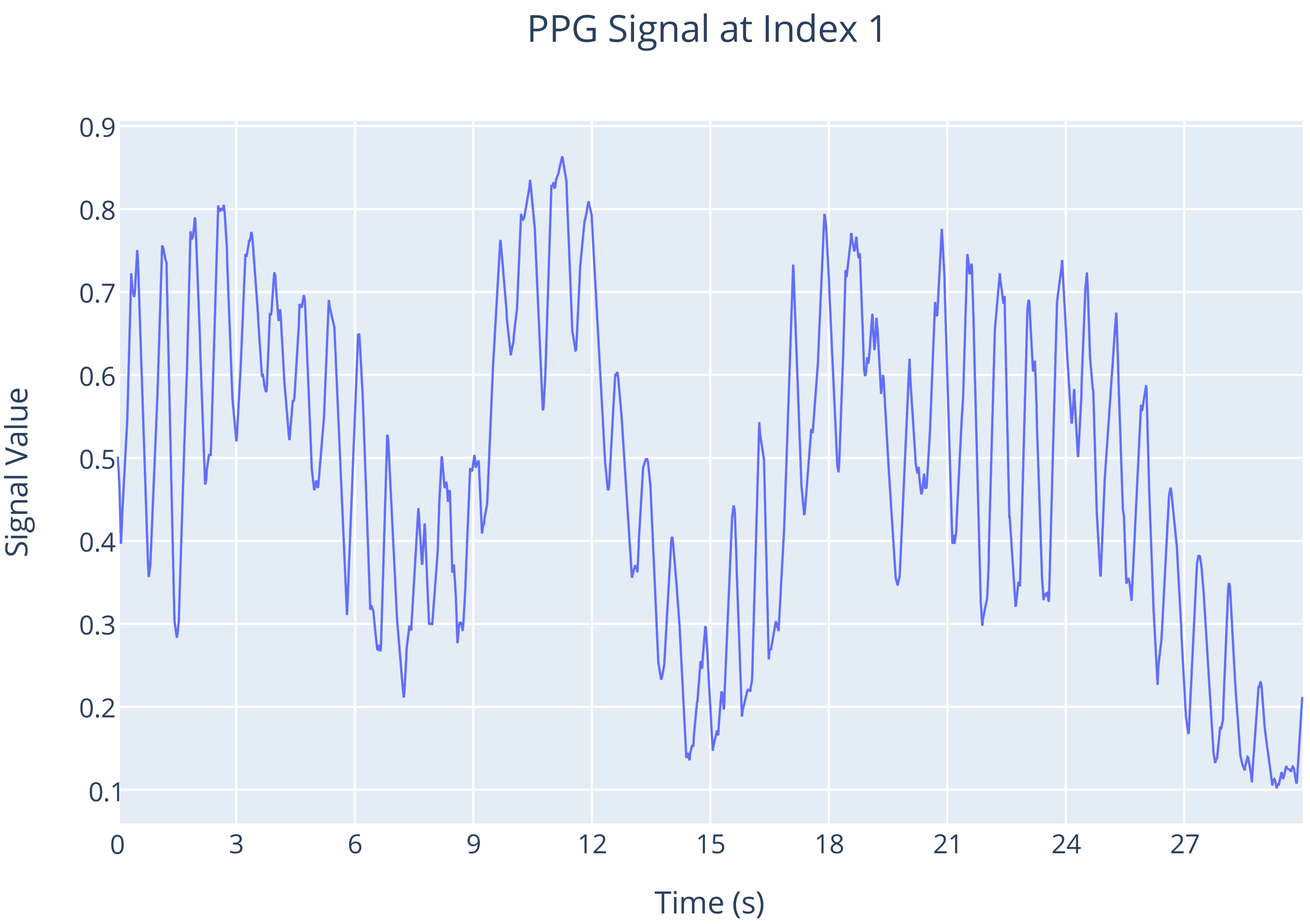


Random Forest Relative Residuals
Residuals
Predicted Target (a.u.)
8
6
4
2
0
−2
−4
−6
−8
−25
−20
−15
−10
−5
0
5

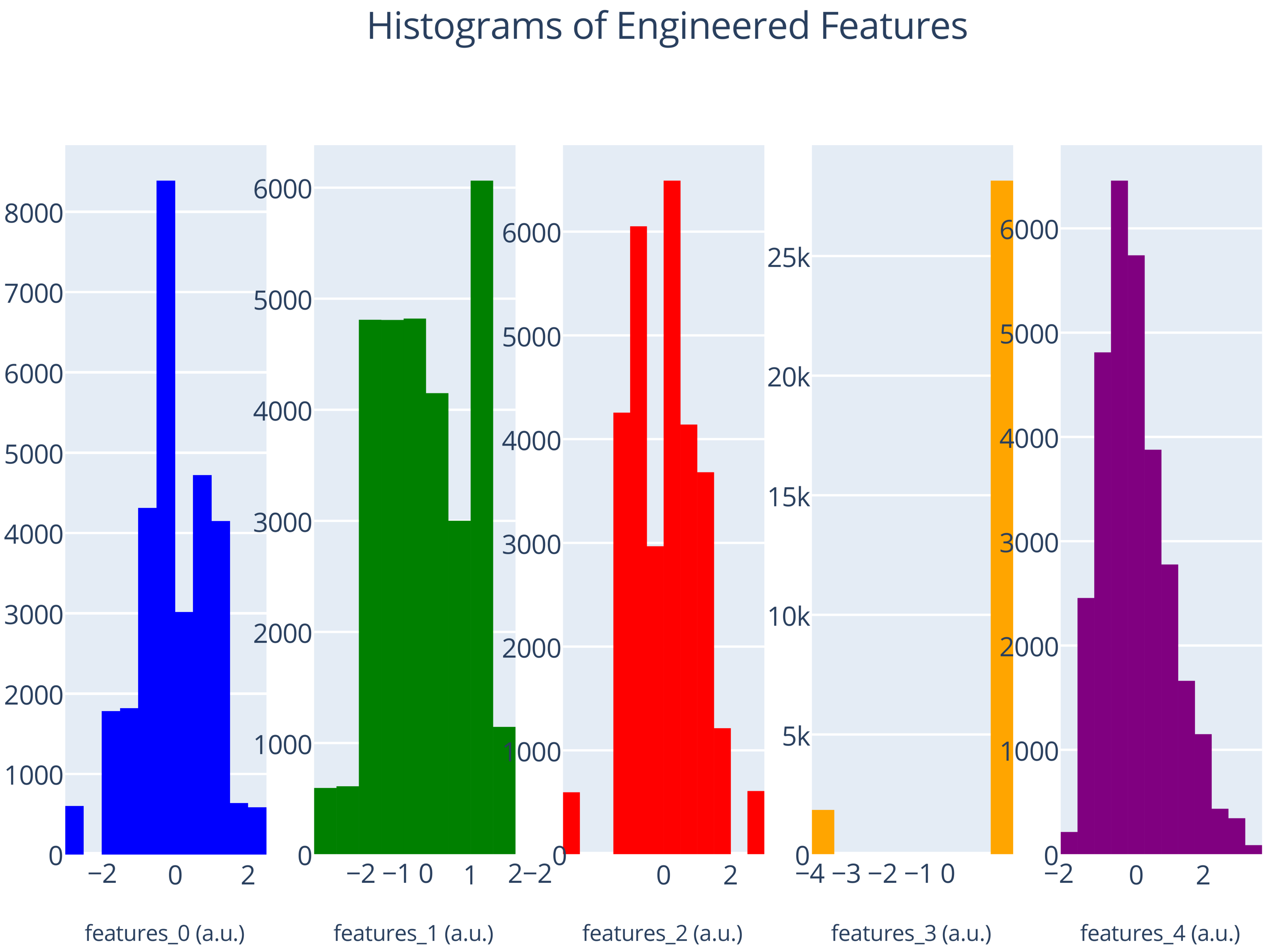
Training set



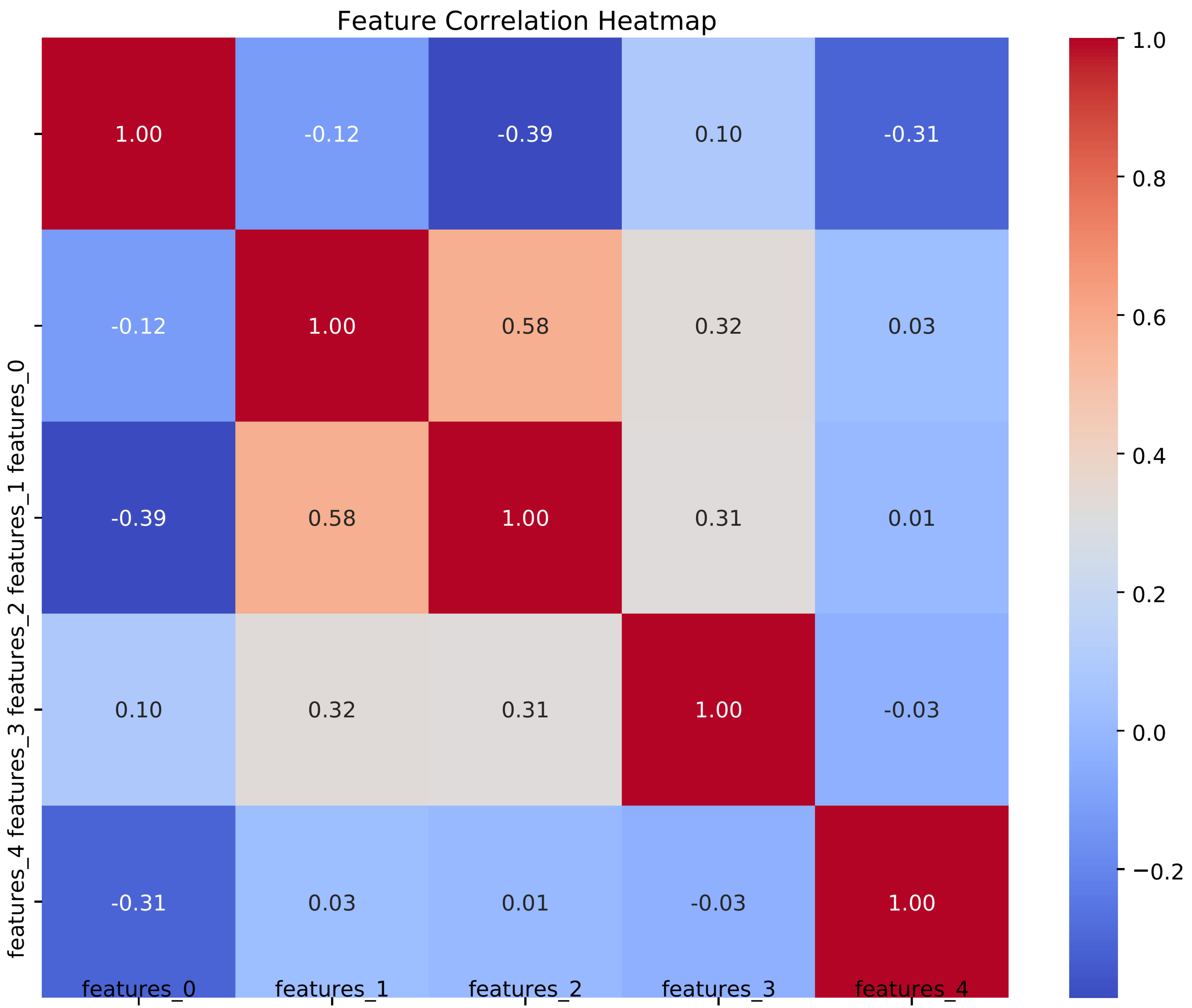
Training set



Test set


Histograms of Engineered Features
features_0 (a.u.)
features_1 (a.u.)
features_2 (a.u.)
features_3 (a.u.)
features_4 (a.u.)
0
1000
2000
3000
4000
5000
6000
7000
8000
5k
10k
15k
20k
25k
−4 −3 −2 −1 0
−2
0
1
2

Test set



Feature Correlation Heatmap
1.00
-0.12
-0.39
0.10
-0.31
-0.12
1.00
0.58
0.32
0.03
-0.39
0.58
1.00
0.31
0.01
0.10
0.32
0.31
1.00
-0.03
-0.31
0.03
0.01
-0.03
1.00
features_0
features_1
features_2
features_3
features_4
1.0
0.8
0.6
0.4
0.2
0.0
−0.2