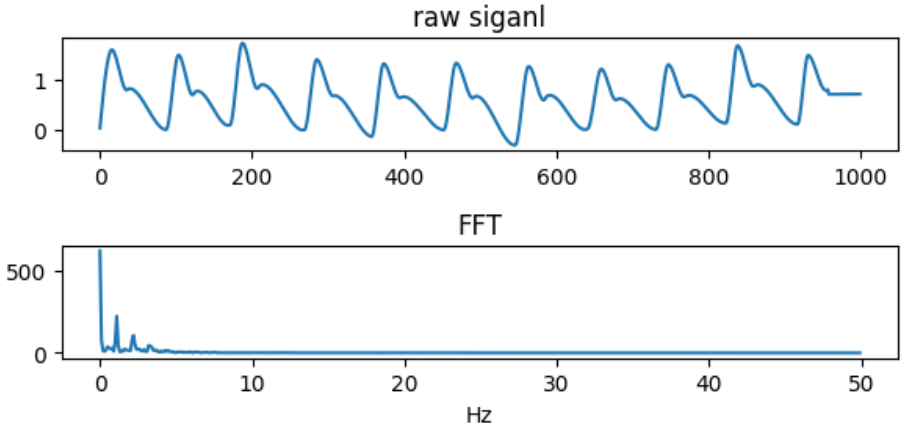
## Photoplethysmogram Signal Analysis

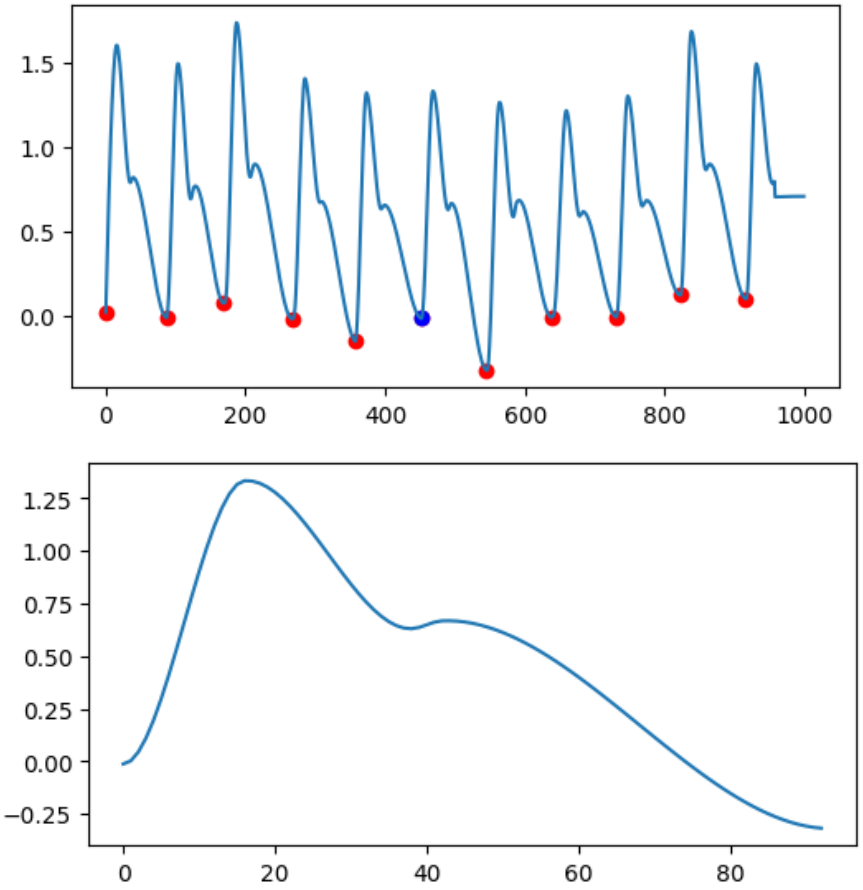
### 1. PPG Spectrum Analysis

Spectral analysis of the PGG signal from the neurokit2 package simulation observed that the ideal PPG signal spectrum should be concentrated in the low-frequency region.




### 2. Selection of the pulse in the middle as the R-peak-signal

Find the lowest point in the middle of the signal (blue) based on the lowest points of the signal (red), and select to the next lowest value point as the pulse for peak detection

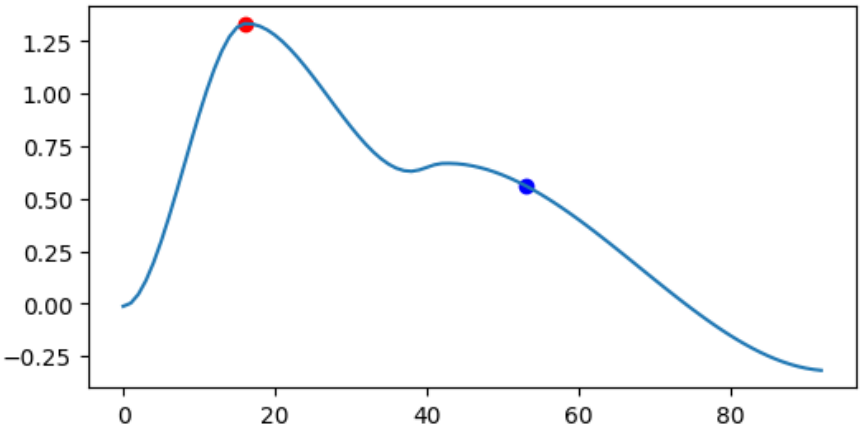



### 3. Peaks detection

When extracting heart beats from a signal, a marker is chosen that can reliably be detected at the same position on all heartbeat complexes in the signal. In the ECG the

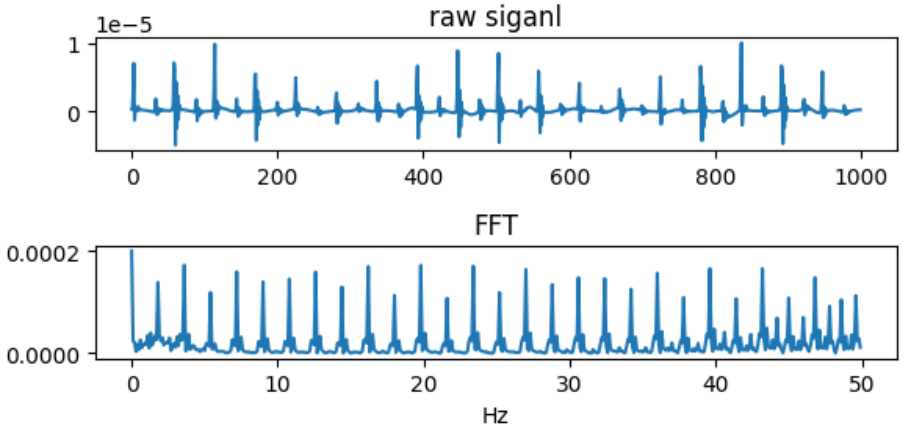
R-peak is often taken , in the PPG signal the maximum of the Systolic wave is usually marked (red).

We consider the diastolic peak to be the point at which the first derivative of the polynomial is equal to zero and the second derivative is negative. If there is no such point, then the point at which the second derivative is a local minimum is chosen as the diastolic peak (blue).

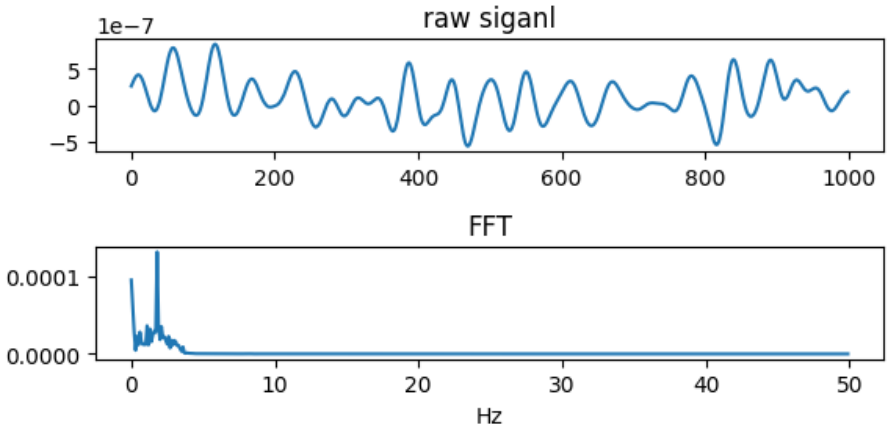



## Signal preprocessing for training data

### 1. Spectrum analysis



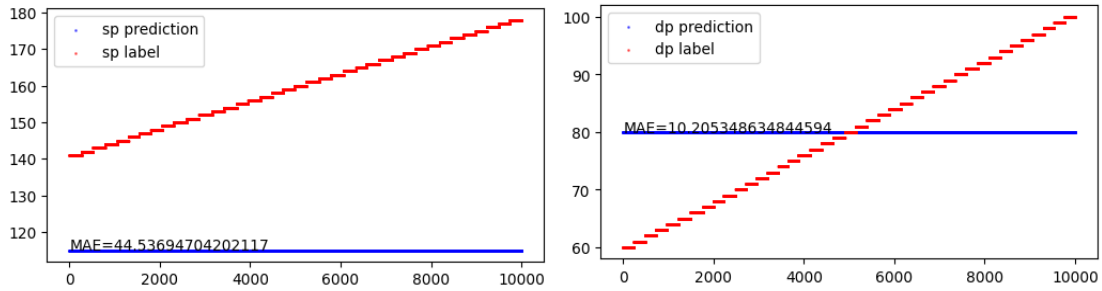

### 2. Low-pass filter




Several filtering methods were tried and none of them yielded a suitable PPG signal for peak detection, so the coordinates of S and D (X1, Y1, X2, Y2) were not successfully extracted as learned features.

## ML model for prediction

So, we tried to directly use S D as label for supervised training, training Ridge Regression models for systolic and diastolic pressures and tested them on the test data, and the MAE values obtained were **44.54, 10.21** respectively.




## Conclusion

I have documented the entire process in the *report.ipynb* file for your convenience. However, I must admit that the experimental results have not met our initial expectations.

One of the primary challenges I encountered was related to the preprocessing of the provided dataset. It appears that the choice of filtering methods and preprocessing techniques may not have been optimal, leading to difficulties in accurately extracting the coordinates of the R-peaks, systolic peaks, and diastolic peaks from the signal.

Furthermore, I attempted to employ a machine learning model directly to predict the S and D values from the data. Unfortunately, the model did not perform as anticipated, yielding Mean Absolute Error (MAE) values of 44.54 and 10.21, respectively. These values significantly exceed the specified target of MAE <= 3, which also implies that the extraction of signal features is very important for the prediction of the model.