Part II

Challenge 1: ECG

5 Introduction

5.1 Structure of Part II

This section outlines the report's structure for the ECG part, briefly introducing the main content of each part.

First, section 5.2 gives a brief background about ECG signals and parts of ECG signals. Sections 5.3 and 5.4 describe the aims and main challenges associated with this project. Section 5.5 outlines the notations used in the study. Next, Section 6 describes the exploratory data analysis, dataset summary and feature engineering. Two different pre-processing approaches were pursued to clean and extract the relevant features from the signal. While the 1^{st} approach relies on signal processing concepts, the 2^{nd} approach uses more traditional statistical learning techniques.

5.2 Background

Electrocardiography (ECG or EKG) involves recording the heart's electrical activity by placing electrodes on the skin, generating a record of the heart's electrical activity across continuous cardiac cycles. These electrodes detect minor electrical changes arising from the depolarization and subsequent repolarization of the cardiac muscle during each cardiac cycle. Variations in the normal ECG pattern occur in numerous cardiac disease conditions (see page 1 of [9]). Each heartbeat cycle is expressed in the form of a PQRST complex, and the P wave represents atrial depolarisation, resulting in atrial contraction. The QRS wave group follows, indicating the initial stage of ventricular depolarisation and ventricular

contraction. The ST segment indicates the period when the ventricle is fully activated and before the T-wave. The T-wave indicates that the ventricle returns to a resting state (repolarisation) and completes a heartbeat cycle, as in the below image (see page 4 of [9]).

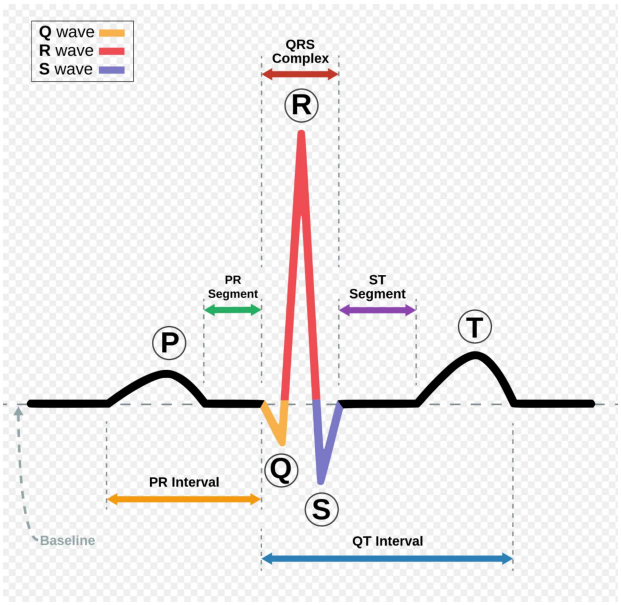

Figure 15: One heartbeat (reference [10])

Myocardial Infarction (MI), often referred to as a heart attack, is among the most severe forms of heart disease. It primarily occurs due to a significant reduction or complete blockage of blood flow to the myocardium, leading to the death of cardiac muscle cells. This condition can silently occur or manifest as a life-threatening event [11]. According to the British Heart Foundation (BHF), heart and circulatory diseases cause approximately 170,000 deaths each year in the UK, averaging about 480 deaths daily, equating to nearly one person every three minutes [12]. These statistics underscore the severity and urgency of addressing myocardial infarction within the UK's public health domain.

Cardiomyopathy involves a group of diseases that cause the heart muscle to become overly stretched, thickened, or stiff, thereby affecting the heart's pumping efficiency. Most cardiomyopathies are hereditary, though they may also result from other conditions. Cardiomyopathy can affect individuals of any age, with the main types being dilated car-

diomyopathy and hypertrophic cardiomyopathy, among others. The diagnosis typically relies on heart scans and tests, including electrocardiography (ECG), echocardiography, and cardiac MRI [13]. Although cardiomyopathies are generally incurable, symptoms and complications can be effectively managed through medications, lifestyle modifications, and medical procedures.

Therefore, utilizing ECG to promptly detect myocardial infarction and cardiomyopathy, effectively differentiating these diseases, is crucial for enhancing heart health in the UK and worldwide.

5.3 Aims

The main objective of this report is to develop a statistical machine-learning method for diagnosing heart conditions from Electrocardiograms (ECG). The dataset includes healthy individuals (coded as $y = 0$) and individuals with two types of pathology: myocardial infarction ($y = 1$) and cardiomyopathy ($y = 2$). The predictor variable x in each case is a 30000×1 vector containing the ECG time series for the specific individual.

The training set comprises ECGs (x) for 115 individuals with given health statuses (y). The test set contains the ECGs only for another 100 individuals. The task involves using these ECGs to predict the health statuses of individuals. The challenge aims to construct a model $\hat{y} = f(x)$ that maximizes the number of correctly classified individuals in the test sample, i.e., to maximize

$$\sum_{i=1}^{n_{\text{test}}} I(\hat{y}_i = y_i)$$

where I is the indicator function, which equals 1 if the predicted health status $\hat{y}_i$ equals the actual health status y_i, otherwise 0. This necessitates the development of a machine learning model capable of accurately distinguishing between healthy individuals, myocardial infarction patients, and cardiomyopathy patients.

5.4 Challenges

The ECG challenge had a very high dimension of features, therefore the training data had to be processed extensively. Further, the ECG signals are nonlinear, non-stationary, and highly random weak physiological signals. The causes for these are noise introduced by instrumentation, human activity, the operator, and environmental factors. The main sources of noise in ECG signals include baseline drift, muscle artifacts, and power-line interference. Baseline drift, often caused by respiration and body movements, manifests as slow-changing noise with a frequency below 0.5 Hz. Among these, baseline drift has the most significant impact on this classification task because the ST segment, a crucial parameter for diagnosing myocardial infarction, has a frequency close to that of baseline noise. Therefore, the filtering process must avoid affecting the ST segment as much as possible, as incorrect filter choices may alter the ST segment, impacting the diagnosis of myocardial infarction. Thus, preprocessing, especially noise reduction through filtering, becomes particularly important [14].

Moreover, analysis and statistical data indicate that the ECG signals of cardiomyopathy patients exhibit different patterns, sometimes resembling those of healthy individuals, and sometimes mimicking myocardial infarction. Hence, the sensitivity of ECG signal features in diagnosing cardiomyopathy might be low [15]. Additionally, the extreme imbalance in the number of categories within this ECG dataset might cause the classifier to underrepresent rare samples, thereby complicating their classification and significantly affecting model performance.

Finally, the feature engineering for ECGs is crucial to the success of this classification task. How to extract features from the PQRST complex will largely influence the classifier's outcome. Overall, preprocessing, especially the feature extraction part, will require extensive knowledge of ECG terminology, and signal processing, and learning and applying this knowledge proficiently in a short period is another challenge of this project.

5.5 Fundamental Notation

To ensure clarity and rigor in our study, the notation for datasets, features, and model parameters used in this ECG classification task are defined below.

5.5.1 Datasets

- $X = [x_1, x_2, ..., x_{115}]$: Represents ECG data of all patients in the dataset
- $x_i(n)$: The value of the ECG signal for patient i at discrete time index n. i.e., $n = 1, 2, \ldots, 30000$.
- m_j: Represents the number of heartbeats for patient j
- m_{ij}: Represents the i^{th} heartbeat for j^{th} patient
- p_{ij}: Represents the i^{th} heartbeat of the j^{th} patient after transforming the signal to a fixed sequence of length $p = 200$
- s_{ij}: Represents the spline regression fitted values for the i^{th} heartbeat of the j^{th} patient
- fs: Represents the estimated sampling rate of the ECG for this study
- Y: Represents the class label of the patients. In this study, Y can take the following values:

5.5.2 Features

- $f_i(X)$: Represents the i^{th} feature extracted from the ECG signal $X(n)$, where $i = 1, 2, \ldots, M$, and M is the total number of features. The purpose of feature extraction is to transform the raw ECG signal into a set of quantitative indicators that aid in classification.

5.5.3 Model Parameters

- θ: Represents the set of parameters for the machine learning model used for ECG signal classification.
- $F(X;\theta)$: Represents the predictive output of the model with parameters θ for the input signal $X(n)$, typically the probability distribution predicted for each category.

5.5.4 Model Performance Metrics

- P: Represents the performance metrics of the model on the test set, such as accuracy, recall, F1 score, etc.

6 EDA and Feature Engineering

6.1 Exploratory Data Analysis (EDA)

6.1.1 Data Overview

Dataset Size and Structure The train set comprises a total of 115 ECG signals of patients, each consisting of 30,000 data points, reflecting the continuous recording of cardiac activity.

Category Labels The ECG signals in the train set are divided into three categories(see figure 16):

- Healthy individuals (labeled as 0): comprising 26 samples.
- Myocardial Infarction (MI, labeled as 1): comprising 80 samples.
- Cardiomyopathy (CMP, labeled as 2): comprising 9 samples.

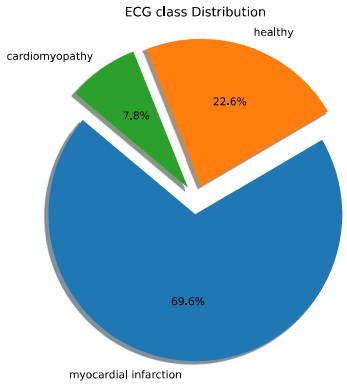

Figure 16: Distribution of ECG Classes

6.1.2 Statistical Summary and Visualization

Time-Domain Plots Healthy individuals' ECG signals exhibit typical rhythms and waveforms without significant abnormalities. Below is a representation of the first 10,000 data points from a sample see figures 17 .

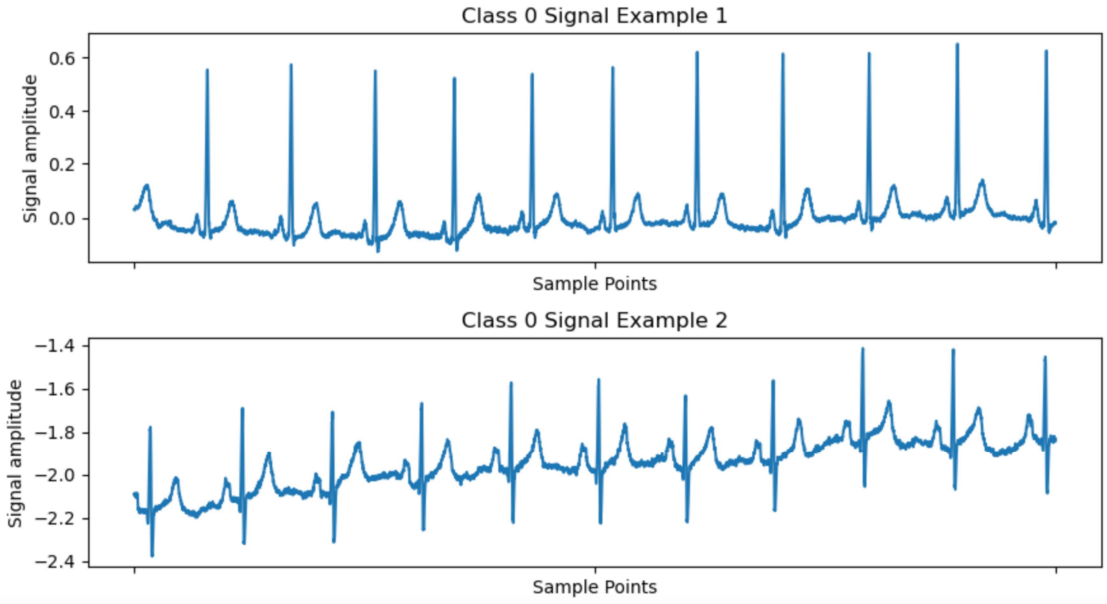

Figure 17: ECG of a Healthy Individual (First 10,000 Data Points)

ECG signals of individuals with myocardial infarction may show elevation or depression of the ST segment and abnormal Q waves. Below is the ECG of a myocardial infarction patient sample showcasing the first 10,000 data points(see figures 18).

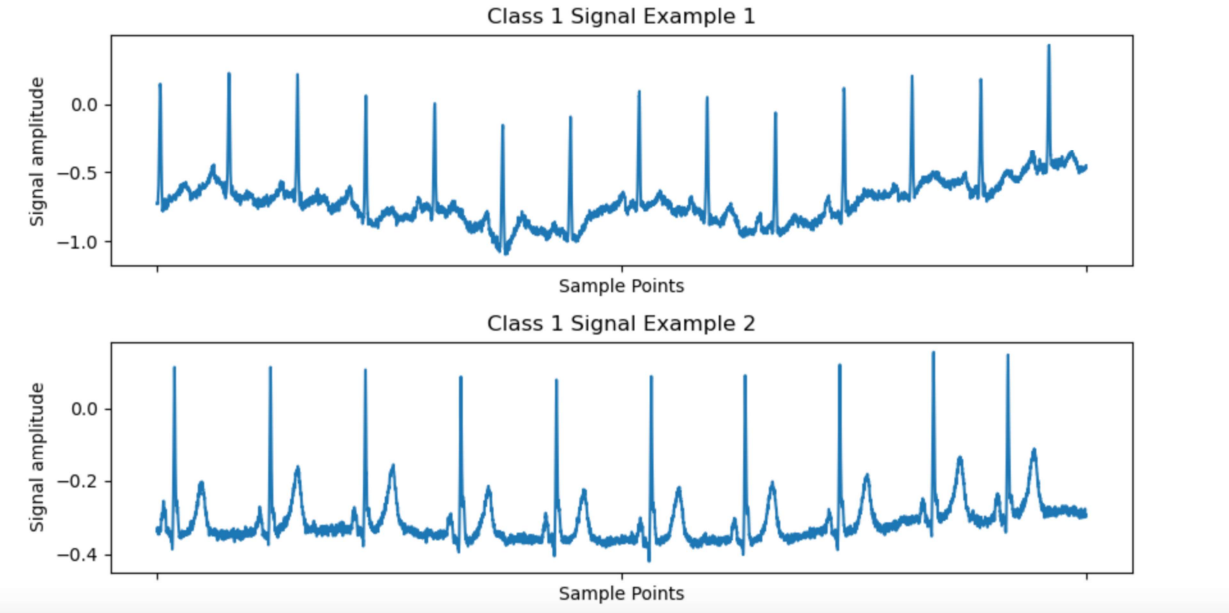

Figure 18: ECG of a Myocardial Infarction Patient (First 10,000 Data Points)

ECG signals of cardiomyopathy patients may exhibit features such arrhythmia or irregular heartbeat rhythm. It may also show that abnormal QRS complex and ST segment or T-wave inversion. Below is the ECG of a cardiomyopathy patient sample (see figures 19)

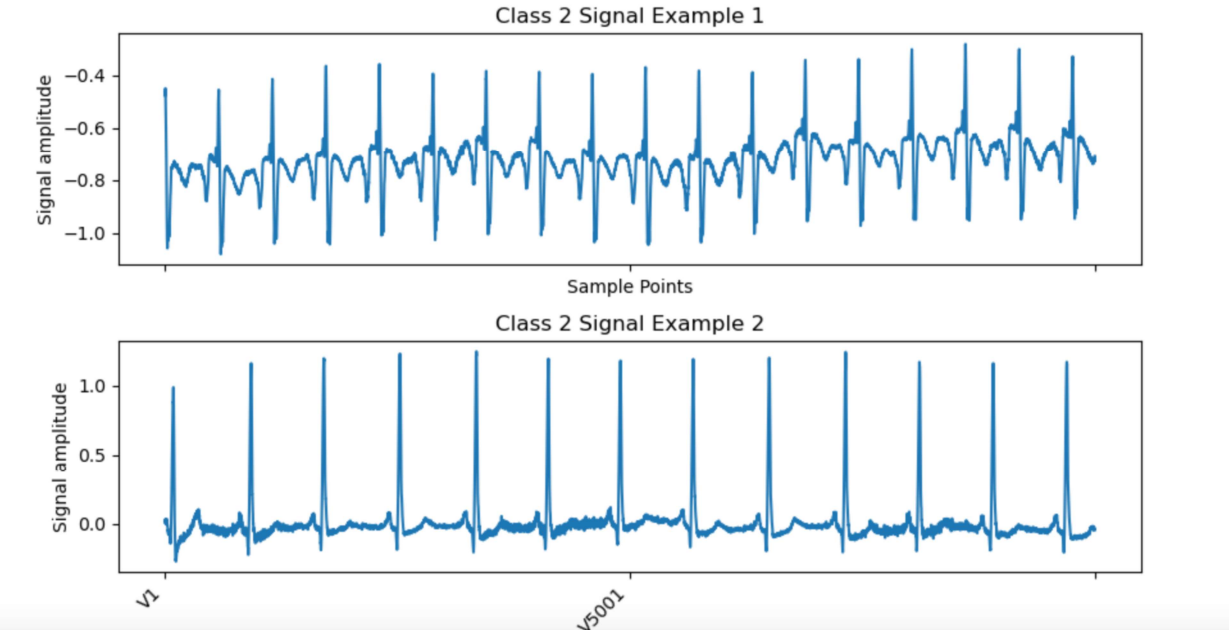

Figure 19: ECG of a Cardiomyopathy Patient (First 10,000 Data Points)

Frequency-Domain Plots After performing a Fast Fourier Transform (FFT) on the ECG signals, the frequency-domain plots reveal the energy distribution across different frequency components. The larger amplitude values close to zero frequency 0-1hz are typically due to respiration, slow variations, or instrument drift, corresponding to baseline drift in noise [14]. This baseline drift is particularly evident for the sample figure 20. It shows that the signal next to the 0Hz is up to 3.

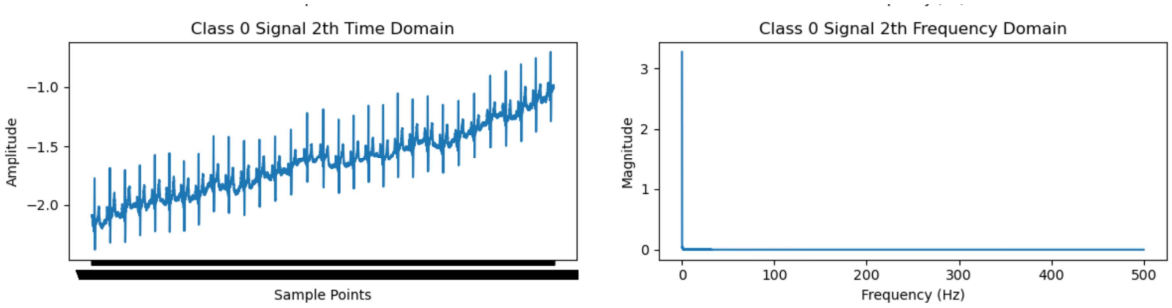

Figure 20: Sample No.2, Time-Domain and Frequency-Domain Comparison

Meanwhile, frequency-domain analysis reveals that frequency components above 40Hz are relatively minor, indicating that the effective information in ECG signals primarily lies within the 0-40hz frequency range (see figures 21, and 22). These observations are crucial for subsequent data processing steps and set a benchmark for the setup of filters.

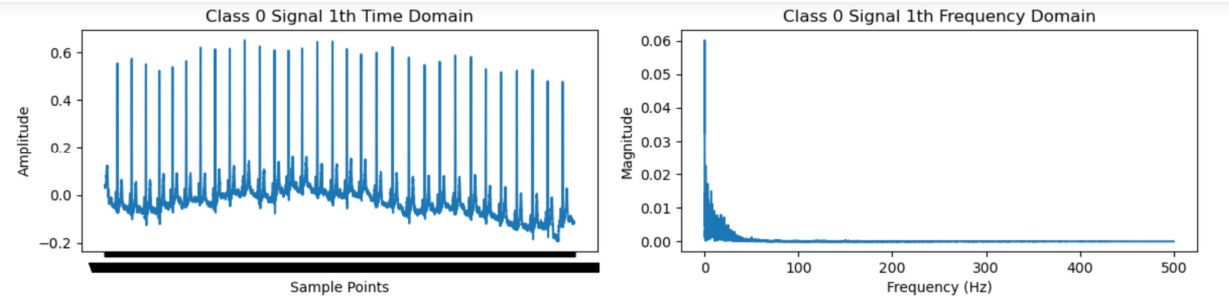

Figure 21: Sample No.1, Time-Domain and Frequency-Domain Comparison

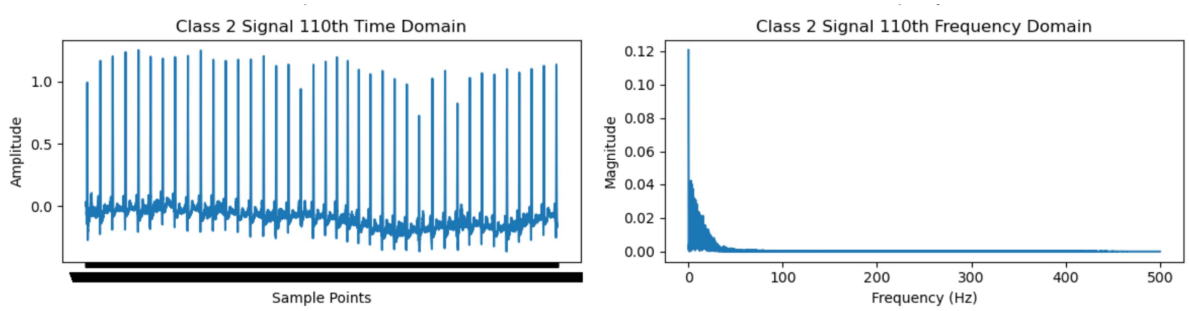

Figure 22: Sample No.110, Time-Domain and Frequency-Domain Comparison

6.1.3 EDA Summary

After conducting preliminary Exploratory Data Analysis (EDA), it was found that time-domain features are crucial for distinguishing between healthy individuals, myocardial infarction (MI), and cardiomyopathy (CMP) patients. Particularly, the width of the QRS complex, the amplitude of P waves and T waves, and their intervals provide the most crucial information to classify an individual.

6.2 Preprocessing - 1st Approach

This section explains the 1st pre-processing approach which uses concepts from the field of signal processing.

6.2.1 Designing a Butterworth Filter

Filtering is a key preprocessing step in ECG signal processing aimed at removing noise and the baseline drift from the signal. Here is the brief process for designing a Butterworth band-pass filter [16]. The basic idea of Butterworth filter design is to maintain frequencies within a certain range while attenuating frequencies outside this range (see figure 23). It is characterized by a flat frequency response within the passband. Ensure that the filter has minimal impact on the signal amplitude within the desired frequency range.

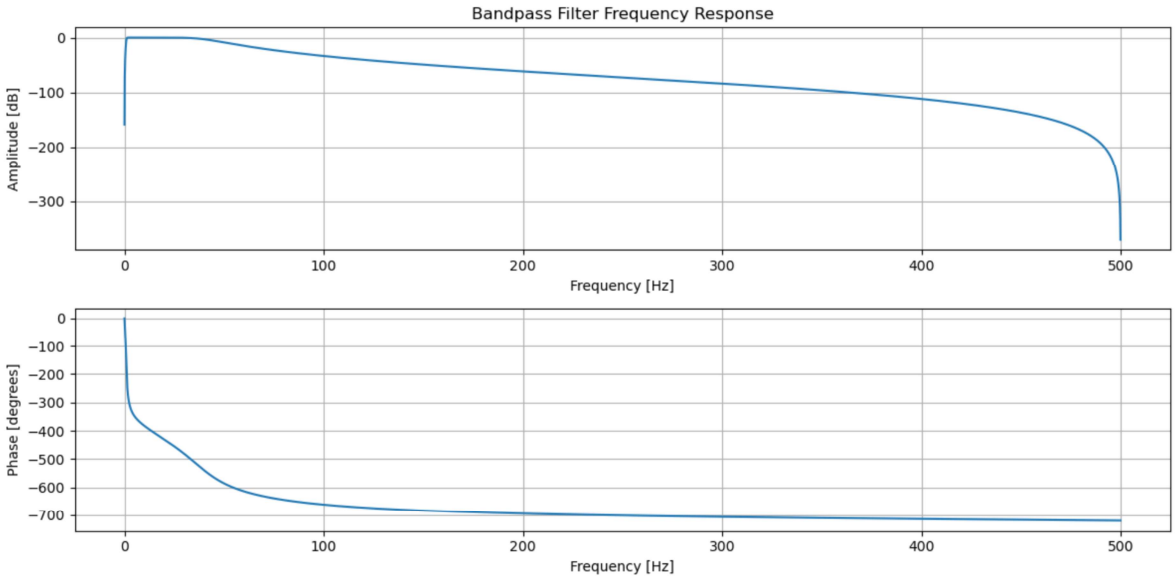

Figure 23: ECG filter design

Based on the exploratory data analysis (EDA) findings, particularly the frequency domain diagrams, it is observed that the predominant frequencies of the ECG signals lie within the 1-40Hz range[16]. Consequently, the 4th-order Butterworth filter was config-

ured to have a passband frequency of 1-40Hz, denoted as $f_{p_1} = 1$ and $f_{p_2} = 40$. Given the estimated heart rate, the ECG's sampling frequency is determined to be 1000Hz. Therefore, the conversion from the analog domain to the digital domain is achieved using the following equations:

$$w_1 = 2\pi \cdot \frac{f_{p_1}}{f_s} = 2\pi \cdot \frac{1}{1000} \tag{6.2.1}$$

$$w_2 = 2\pi \cdot \frac{f_{p_2}}{f_s} = 2\pi \cdot \frac{40}{1000} \tag{6.2.2}$$

The final output imagery demonstrates significant removal of baseline drift and minor noise, preserving the original signal details, as depicted in Figure 24 and 25.

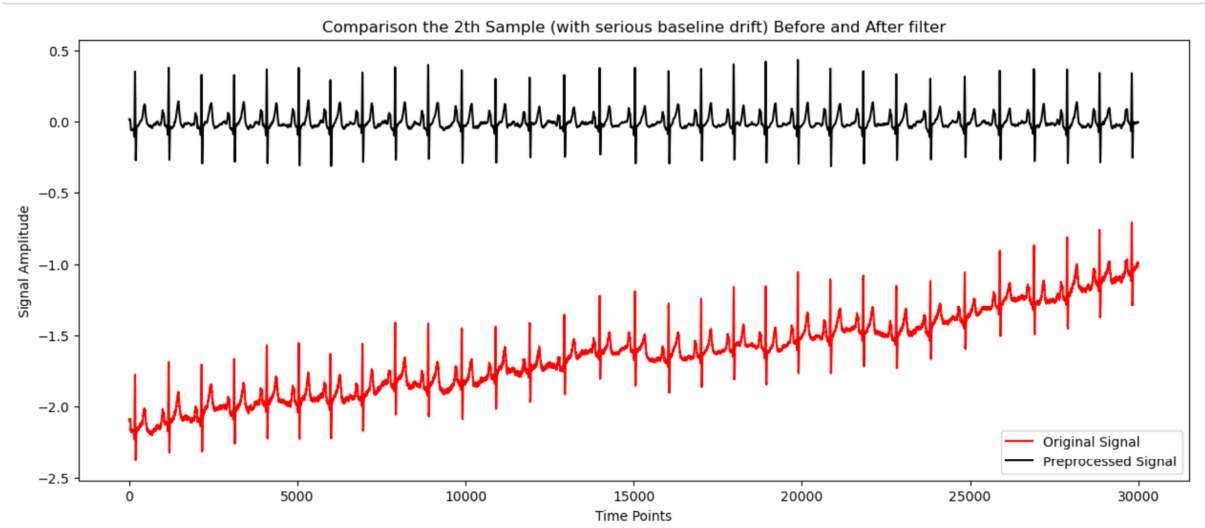

Figure 24: ECG filter work on baseline drift

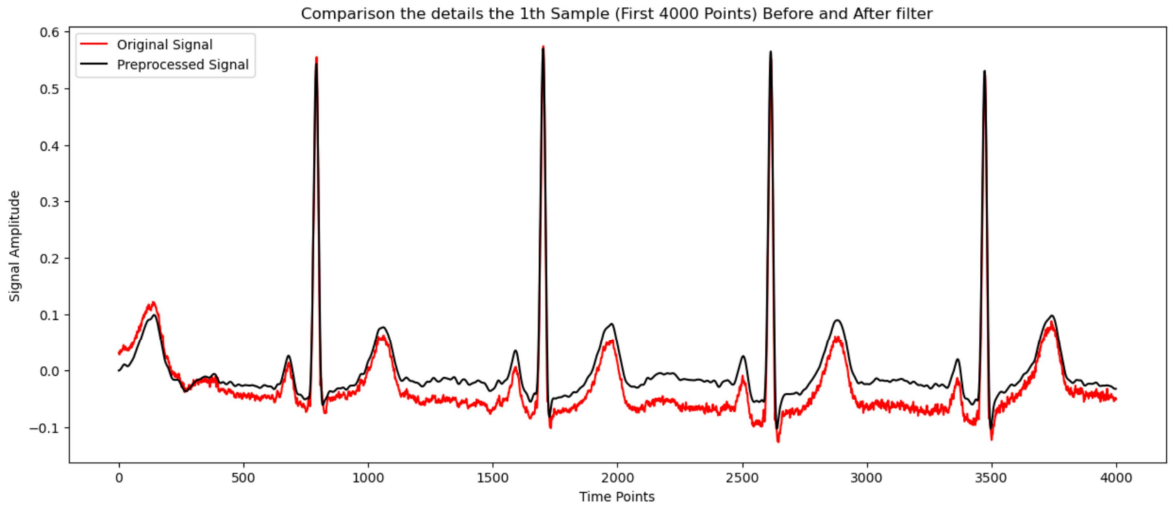

Figure 25: ECG filter work on signal detail

6.2.2 Limitations of the first preprocessing approach

Inability to effectively process certain peculiar signals, leaving segments that appear unclean, as shown in Figure 26. An approach to address this involves analyzing only the clean segment of the signal, such as extracting the signal from index 7000 to 20000 for this example.

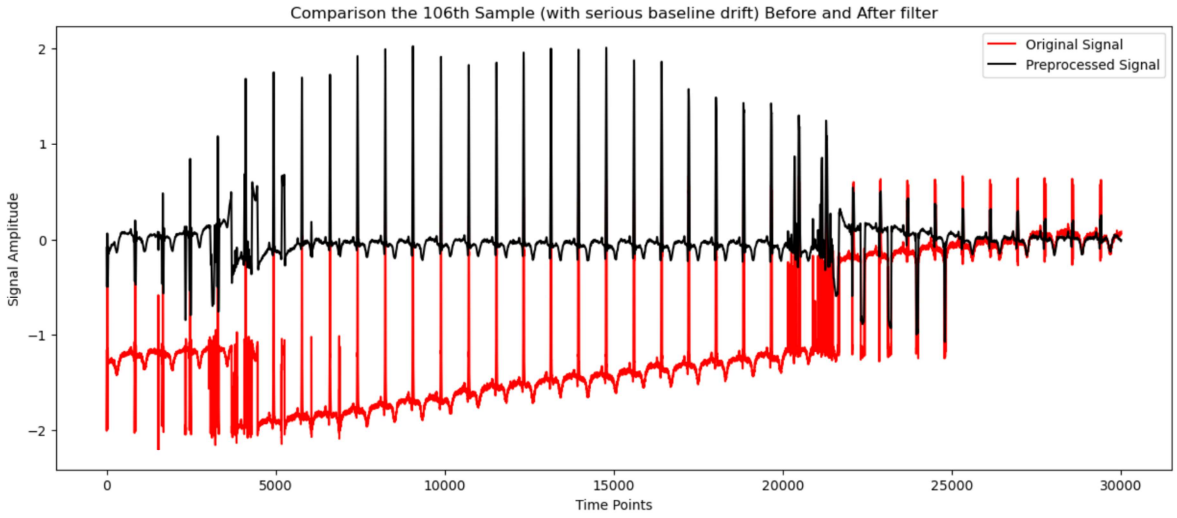

Figure 26: A peculiar ECG signal example

Potential loss of information in the ST segments due to filtering frequencies between 0-1Hz, which may include ST-segment data [14]. Such coarse processing could impact subsequent feature extraction and slightly reduce sample sensitivity, affecting classification outcomes.

This section provides a concise overview of preprocessing steps, illustrating the transition from theoretical principles to practical application, specifically in handling baseline drifts and noise through the Butterworth filter. Despite achieving visually satisfactory results, under certain conditions alternative preprocessing strategies are necessary.

6.3 Preprocessing - 2ndApproach

This section describes the 2nd preprocessing approach to process the ECG signal into useful information for modeling later. This approach improves upon the starter code shared with the challenge.

6.3.1 Removing Linear and Non-Linear Drift

The ECG signal drift was seen in the ECG plots of multiple patients. The "detrend" function in R is used, which fits a linear regression line on the ECG signal. The values of the linear regression line are subtracted from the ECG signal to obtain a detrended signal.

It worked well for patients with only a linear drift, however many patients still had a non-linear trend. To remove this, a higher-order polynomial was used but it could not capture the trend completely. This may be because the non-linear trends were sometimes sudden, erratic, and not equally spaced, and increasing the polynomial order beyond a point was not computationally optimal. Finally, the Penalized B-spline regression is used with a basis of 31 and a penalty of 10^{-4}. The values of these parameters were calculated through trial and error and looking at the visual fit of the B-spline line on the most challenging non-linear ECG cases. A custom R function called "p_splinefn" was defined which enabled

it to iterate and visualize the fit on different parameters and patients. Figure 27 shows an example of penalised spline capturing the non-linear trend in ECG signal.

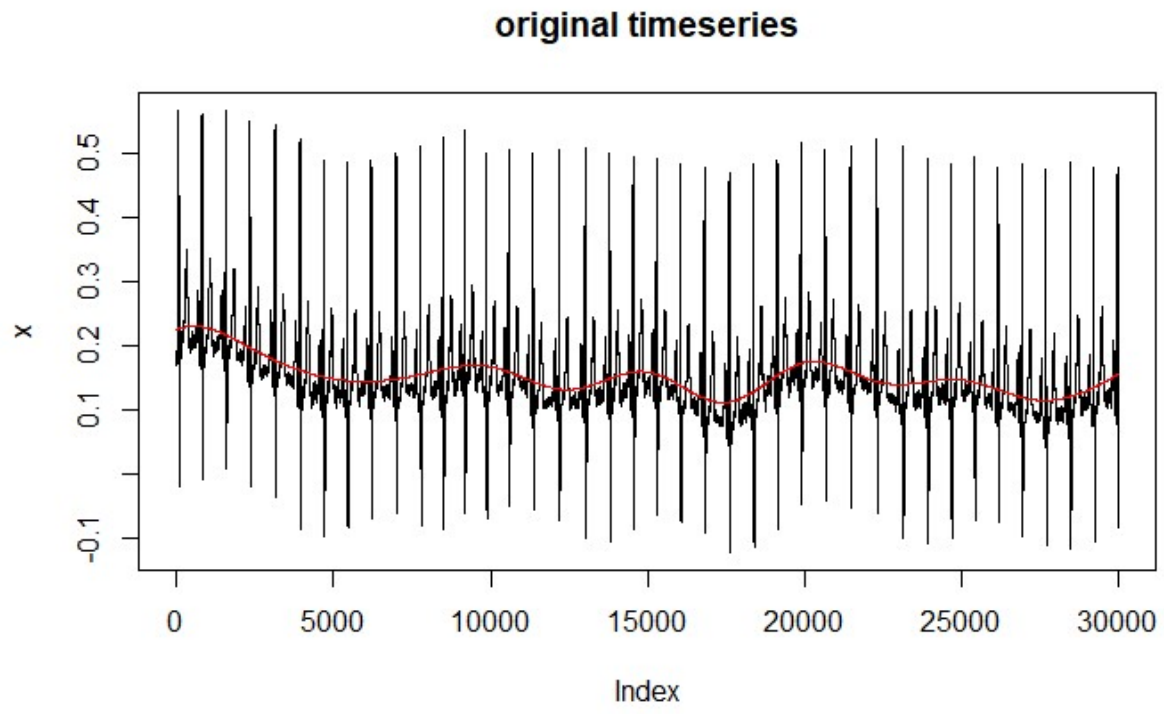

Figure 27: Penalized Spline line in Red on ECG signal

The first 2000 and last 2000 points are discarded since these sections are typically more noisy. Since there are m_j heartbeats for patient j, discarding these 4000 points doesn't lead to much loss of information.

6.3.2 Outlier Treatment

In week 3, while analyzing misclassified patients, it was noticed that few ECG signals showed a huge spike. These spikes were disturbing the peak detection algorithm which is discussed in the next section. A modified version of Inter Quartile Range (IQR) method was implemented to detect and treat the outliers since sometimes the IQR method erroneously

labeled R-wave peaks as outliers.

$$\begin{aligned} NR &= 99P - 1P \\ LB &= 1P - 0.5(NR) \\ UB &= 99P + 0.5(NR) \end{aligned} \tag{6.3.1}$$

First, the 99 percentile (99P) and 1 percentile (1P) are identified, and the Normal Range (NR), Lower Bound (LB), and Upper Bound (UB) are calculated as shown in equation 6.3.1. The values above the UB and below the LB are labeled as outliers and replaced by the median of the ECG signal. Figure 28 shows the visualization of UB and LB calculated by this method. The R code for the function called "rm_outlier" is attached in the Appendix.

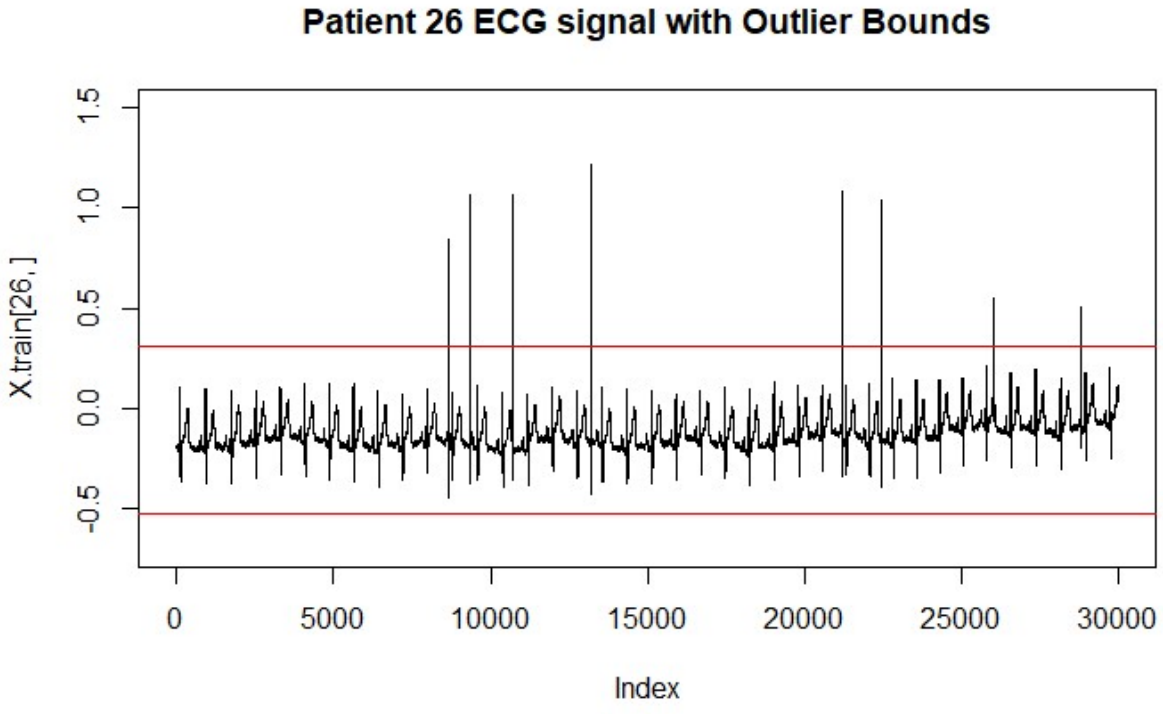

Figure 28: Upper and Lower bounds (in red) to detect outliers

6.3.3 Re-scaling and Finding R-wave Peaks

Since the y-axis scales are different for all patients, to make it comparable all the ECG signals are re-scaled between 0 to 1. Next, to find the R-wave peaks, the R function called "findpeaks" is used from the "gsignal" package, which is a signal processing library. The working of the findpeaks function is not explained in detail since it may divert from the focus of the study. The two main parameters supplied to the function are $MinPeakHeight$ (Minimum height of the peak) and $MinPeakDistance$ (minimum distance between consecutive peaks). After multiple iterations, the parameters were set to $MinPeakHeight = 0.65$ and $MinPeakDistance = 450$ respectively. If the parameters are set lower then along with the R-wave peak the T-wave peak and noisy peaks might be detected.

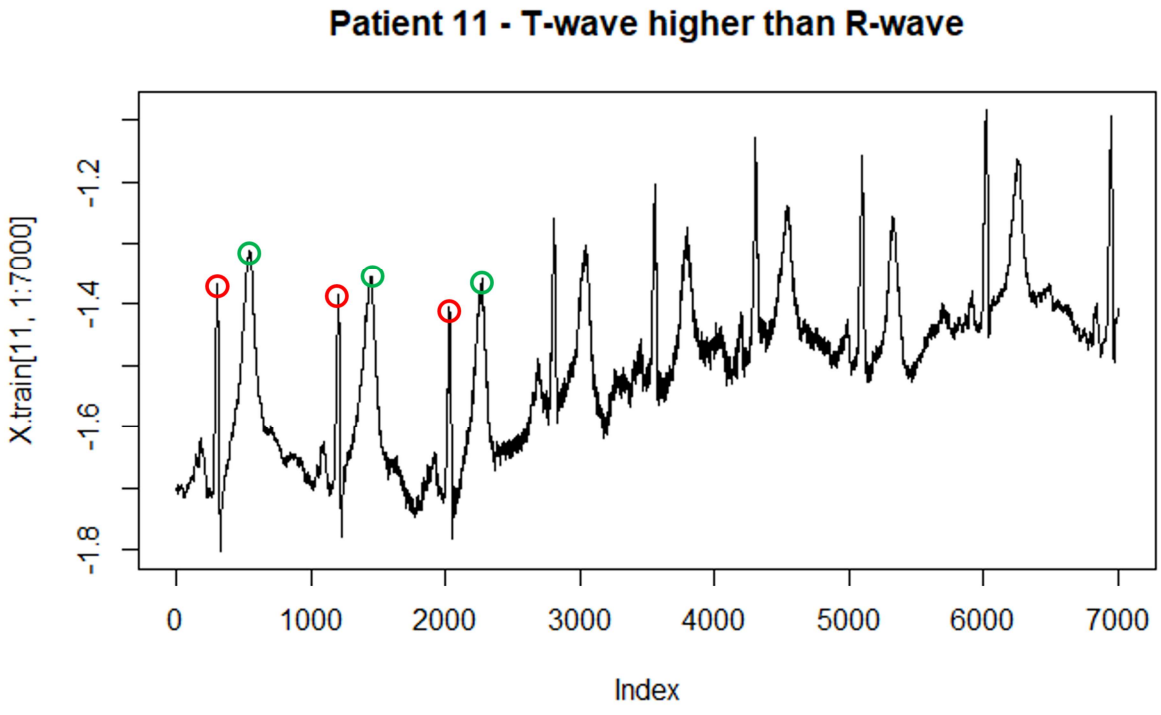

Figure 29: Penalized Spline line in Red on ECG signal

This method is not perfect, since there are situations where the T-wave peak is higher

than the R-wave peak. In figure 29 the R-wave in red is bigger than T-wave in green. This may be due to noise or the patient's condition. In such cases, wrong R-wave peaks are detected and some heartbeat intervals are incorrect. This problem is tackled by discarding such erroneous heartbeat intervals discussed in the next section.

6.3.4 Slicing and selecting the normal heartbeats

Once the R-wave peaks are detected the ECG signal is chopped into m_j heartbeats. Using the "boxplot.stats" function on the heartbeat length, the outlier heartbeats are detected and discarded. The outlier is caused due to wrong detection of R-peaks.

$$Heartbeat\_variance = Var(len(m_{1j}), len(m_{2j}), ...len(m_{ij})) \tag{6.3.2}$$

Next, variance in heartbeat is calculated using 6.3.2. The heartbeat variance is a feature used later in the modelling step. The "seq" inbuilt R function, as used in the starter code helps to convert the heartbeat into a lower dimension vector of dimension p. Here $p = 200$, which closely resembles the input heartbeat.

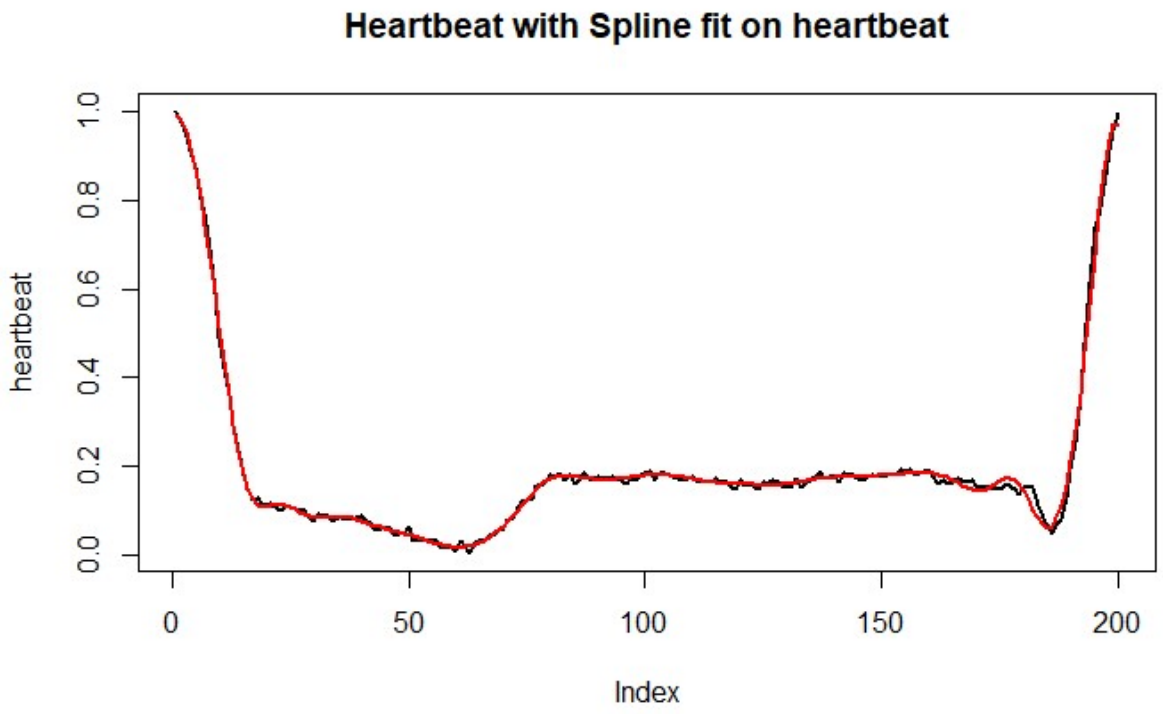

Figure 30: Penalized Spline line in Red on a heartbeat

6.3.5 Smoothing the ECG signal

The raw ECG signal contains a lot of noise which makes future tasks challenging. Moving average is used to make the ECG signal smoother but it has the drawback of losing a few points at the start and the end of the signal. Instead, the B-spline regression is again utilized to fit a non-linear curve which showed very good results as shown in figure 30. The number of basis selected was 31 after testing on various cases.

6.3.6 Locating P-Q-R-S-T indexes in ECG

The R points are already detected earlier during peak detection. The B-spline fit easily provides the 1^{st} and the 2^{nd} derivative which is useful to locate the P, Q, S, and T points. The algorithm for this is explained in this section.

For the S and Q points, a simple heuristic is used since they are the local minima just after and before the R-wave respectively. The minimum value between index 1-27 should be the S point and the minima from index 176-200 should be the Q point. While this method is not error-proof, it performs reasonably good as seen in figure 31.

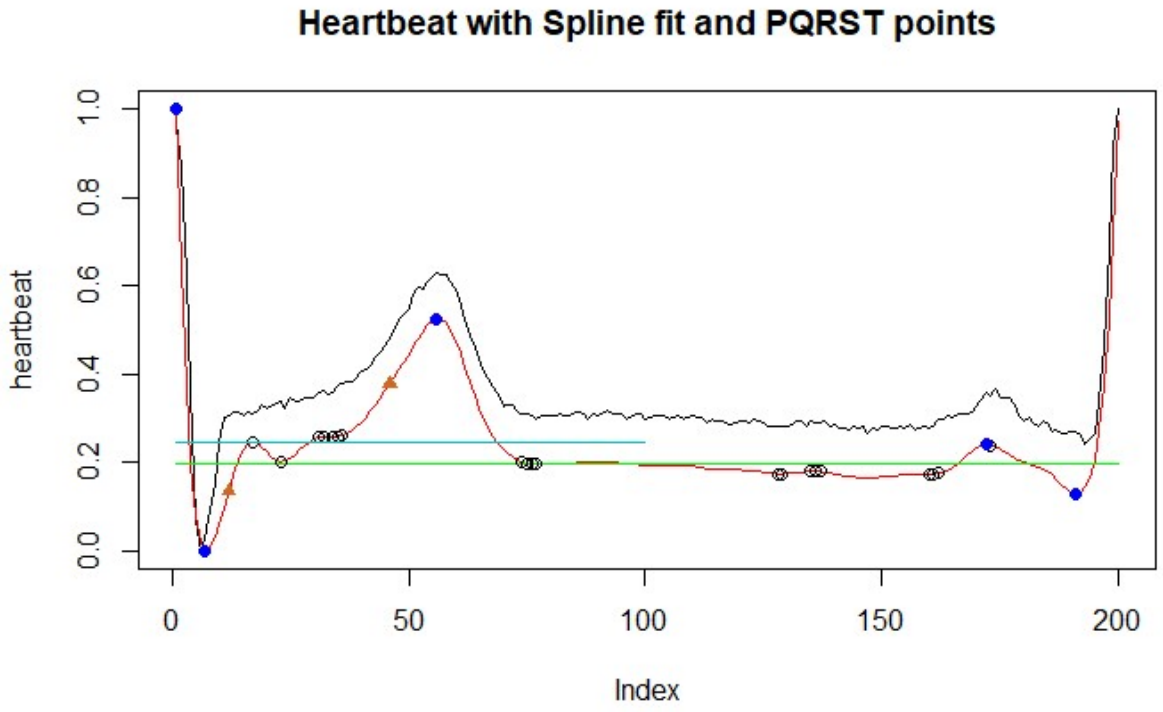

Figure 31: PQRST, baseline and ST elevation

The heuristic to find points P and T are very similar with a slight difference. Both are located away from the edges of a heartbeat. While the P-wave is generally upright, that is a local maxima, the T-wave can be both upright or inverted, that is local maxima or minima. Particularly, T-wave inversion can indicate the presence of cardiac diseases [17]. The first and second derivative tests are used to find these local extrema. The test says the first derivative should be close to zero and the second derivative should be non-zero. The second derivative should be positive for an inverted wave and negative for an upright wave.

The following steps describe the heuristic for locating the T-wave position index,

1. **Calculate heartbeat baseline:** $Baseline = Median(m_{ij})$

$$\begin{aligned} Abs(f'(s_{ij})) &<= 0.003 \\ Abs(f''(s_{ij})) &>= 0.001 \end{aligned} \tag{6.3.3}$$

2. **Find candidate T-wave indexes:** Find the index at which the first derivative of the ECG signal is zero and the second derivative is non-zero. Practically, as shown in equation 6.3.3 the values are not exactly set to 0 since the ECG signal is discrete and not continuous.

3. **Apply Location Constraint:** T-wave should lie between index 45-125. This is based on observation. Shortlist candidates from the previous step that satisfy this location criteria.

4. **Select the T-wave indexes:** If multiple candidates are left, select the candidate index which has maximum deviation from the heartbeat baseline. If a single candidate remains, that is the T-wave index

5. **Check if T-wave is inverted:** If $f''(s_{ij}) > 0$ then T-wave is inverted, else it is upright. Store this information as a feature for modeling later.

The heuristic for P-wave detection is similar and the complete code is attached in the appendix for reference. It is noted for some instances, due to noise or limitation of the heuristic, it may not locate indexes for P and T waves. Such heartbeats were discarded. The first and second derivative tests could have been used for locating S and Q waves, but it was seen in heartbeats with T-inversion, sometimes the transition from S index to T-index was a near continuous downward slope and the $Abs(f'(s_{ij})$ was not close to 0. It was a trade-off between tuning the parameter for this rare instance and generalizing over all the patients.

6.3.7 Feature Engineering on ECG signal

The location of critical points in the ECG - P, Q, R, S and T are used to calculate 24 features. Most of these features are self-explanatory and simple to calculate, hence, the formulas are mentioned only where necessary in this section. For all the details refer to the custom R function "get_ecg_feature()" in the appendix.

- **Wave Amplitudes Features:** It is the height of the wave relative to the heartbeat baseline. Literature shows high voltage values for R-wave and T-wave could indicate a cardiac disease [18]. The amplitude is calculated for all 5 points, namely P,Q,R,S, and T. Example, $R-wave\_amplitude(R\_amp) = m_{ij}[R\_wave\_idx] - median(m_{ij})$

- **Duration Features:** Literature shows various durations between key points can help identify the condition. Calculate the R-duration, QT interval, and PQ interval.

- **Height Difference Features:** The ratio of R to S wave height (R/S ratio) can be used as a diagnostic feature [19]. The height difference between the following pairs of points - Q & S, S & T, R & T is also calculated.

- **ST Elevation:** One of the most popular features to identify myocardial infarction is the presence of elevated ST segment. In the literature this is a common subclass of heart attacks called **STEMI (ST-elevation myocardial infarction)** CITE. The ST-segment begins at the offset point of the S-wave and onset point of the T-wave. In figure 31, the horizontal light blue line marks the ST segment level. The offset and onset points can be quite challenging to find and due to limited time some assumptions on their location were made. The ST segment elevation is the deviation of the ST segment from the heartbeat baseline .

$$\text{S\_wave\_offset} = \text{S\_wave\_idx} + 5$$

$$\text{T\_wave\_onset} = max((\text{T\_wave\_idx-25}), (\text{S\_wave\_idx+10}))$$

$$\text{ST\_Elevation} = median(m_{ij}[\text{S\_wave\_offset} : \text{T\_wave\_onset}]) - \text{heartbeat\_baseln}$$

- **Shape Features:** The shape of the waves can lead to identifying the cardiac condition. For example, a "dagger-like" Q-wave is a feature of Hypertrophic Cardiomyopathy CITE. The second derivative at the peak points P, Q, S, and T is stored as a feature. If this value is large it indicates a sharper peak and if it is small then it's a flatter peak. The mean of the slopes at the 3^{rd}, 4^{th} and 5^{th} index before and after the peak of T-wave and P-wave are also stored as features. If the slopes before and after are similar it means the wave is symmetric, otherwise the waves must be skewed.

- **Other Features:** The T-wave inversion and heartbeat variance form the final 2 features. In earlier sections, it was highlighted how these features were derived.

6.3.8 Making the final dataframe

For each heartbeat, a feature vector of length 24 is obtained. Each patient's ECG has multiple heartbeats, but some of the heartbeats are discarded due to reasons mentioned earlier. An average of these heartbeats is taken to arrive at a single representative feature vector for each patient. For the feature T-inversion, since it is a binary variable, the Mode is calculated instead of the average. Finally, iterating over n patients the final data frame if formed of the shape nxp. Where n is the number of patients and $p = 24$ is the number of features.

6.4 Classification Model Development

This challenge is a classification problem with 3 possible classes. This section explains the modelling methodology and reasoning behind modelling decisions for the study. First, the baseline model trained during the first week is discussed. Next, the intermediate modeling steps are discussed. Finally, the best model used in the final week to get the highest score is presented in detail.

6.4.1 Baseline Model

In week 1, the time was limited to explore the dataset completely and become aware of all the data issues. Hence, using the starter code only implemented 2 pre-processing steps - removing the linear trend and re-scaling the data were implemented. The average heartbeat of each patient was calculated as a vector of length 200. The training dataset had 200 features compared to only 115 samples.

Since the dataset had a high feature dimension and fewer samples, the Naive Bayes which works well in such situations was tried. It is reasonable to assume that patients with similar conditions should have similar-shaped ECG. Therefore another model called K-Nearest Neighbours was tried since it would identify the closest K ECG vectors of patients in 200 dimenesion space and predict the patient condition. Different values of K were checked, from $K = 2$ to $K = 5$.

Model	**Validation Accuracy**
Naive Bayes	63%
KNN (K=3)	60%
KNN (K=4)	62%
KNN (K=5)	68%

Table 9: Baseline Model Accuracies

The training data was split into train and validation sets with a 70-30 split, no cross-validation was done in the initial week. KNN with $K = 5$ gave the best accuracy of 68% on the validation dataset. The table 9 shows the comparison of baseline models. The submitted predictions gave an accuracy score of 64% on the test set.

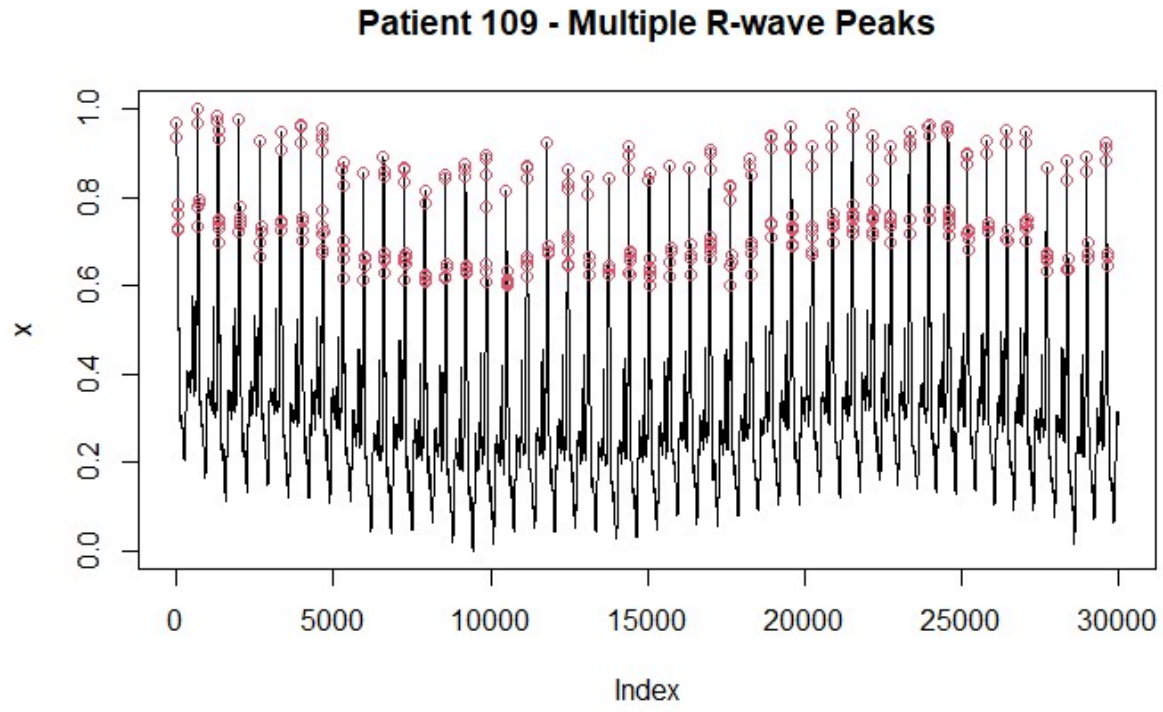

Figure 32: Incorrect R-wave peaks detected

This model had big limitations because it did not address the non-linear trend in the data as seen in figure 27 and the R-wave peaks detection was failing for some patients as seen in figure 32. The next couple of weeks of the study were focused on tackling these two problems through two different pre-processing approaches. Later weeks focused on feature engineering and model selection.

6.4.2 Intermediate Model

The key issues from the baseline model were identified and addressed by 2 different pre-processing approaches. The pre-processing approach 2 was partially developed by removing the non-linear trend using a B-Spline fit and using a better peak-detection algorithm. No change was made in the modeling step and the KNN model with $K = 5$ gave the best accuracy. It saw a 2% jump in the test set accuracy, from 64% in week 1 to 66% in week 2.

Simultaneously the pre-processing approach 1 was using the Butterworth filters and de-noising algorithms were tried. For the feature extraction, the statistical features such as mean, standard deviation, maximum, minimum, kurtosis and skewness were extracted. Additionally, some papers show that the wavelet feature is also effective in such classification tasks. The wavelet transform was used to decompose the signal into simpler elements and then extract the feature (see section 3 of [20]). A cross-validation method was implemented to get a robust accuracy score for the models. New models like SVM-PSO (Particle Swarm Optimization was used to find the best parameter of the SVM model) and dimension reduction techniques (PCA) were tested. This eventually led to a further jump of test set accuracy to 67% in weeks 3 and 4.

Some new challenges and limitations were identified during this time. First, there were huge spikes and outliers in a few patient ECGs which were causing wrong features to be extracted as seen in figure 28. Second, since the classes were severely imbalanced the majority of predictions were for myocardial infarction ($y = 1$). When classes are imbalanced, accuracy is not the preferred metric since it can be misleading. Precision and Recall for each class are more reliable to measure the model performance across all the classes. It was debated whether the models should be tuned by accuracy or precision. However, since the metric for the challenge was accuracy, accuracy was finally used. If the distribution of classes in the test set had many cardiomyopathy patients (class 2), then the model trained based on the accuracy metric is likely to perform poorly.

6.4.3 Final Model

In the final weeks, the focus moved to creating tailor-made features based on information from literature. The pre-processing approach 2 was fully implemented. One critical decision was made to discard the heartbeat cycles which were incorrectly sliced or if any of the PQRST points could not be located. This led to improvement in the quality and reliability of features.

Model	**Accuracy**	**Accuracy (After scaling)**
Naive Bayes	67%	69%
KNN (K=3)	61%	73%
SVM	60%	62%
LDA	65%	68%
Random Forest	73%	73%

Table 10: Cross-validated Model Accuracy on Validation Data

To calculate robust accuracy values that would generalize well, a 4-fold cross-validation with 4 repetitions was implemented. In table 10 the results of cross-validated accuracy on the validation set are shown. The table shows the Random Forest as the best-performing model with 73% accuracy, followed by Naive Bayes and LDA. The predictions using this Random Forest model were submitted and the accuracy on the test jumped by 5% to get the best score of 72%. The interpretation of results from Random Forest and LDA offers some interesting insights.

Feature Name	Importance
T_amp	2.96
RT_ht_diff	2.78
T_wave_slope_bf	2.43
ST_ht_diff	2.42
T_wave_slope_af	2.27
T_wave_2nd_derv	2.11
R_duration	1.58
ST_elev	1.25

Table 11: Top 8 Features found by Random Forest

The table 11 shows the feature importance as calculated by the Random Forest model. A method called **permutation feature importance** is used to estimate these values. It works by measuring the drop in accuracy of the model when the values of one of the features are randomly shuffled. If the feature was not important, the shuffling of its values would result in a very small performance drop. The most important feature is the T amplitude which aligns with the literature. If the amplitude is negative then the T-wave is inverted and the patient most probably has some disease condition. Also, a big positive amplitude of the T-wave is abnormal and points to cardiac disease. 7 out of the 8 top features are based around the T-wave. The ST-elevation turns out to be an important feature that is a known predictor for myocardial infarction. Although the LDA had the 3rd best performance, it is useful to visualize the separation of all the patient classes. In figure 33 it can be seen, that the patients are separated into 3 different classes with some overlap when plotted on an axis of the 2 linear discriminant components.

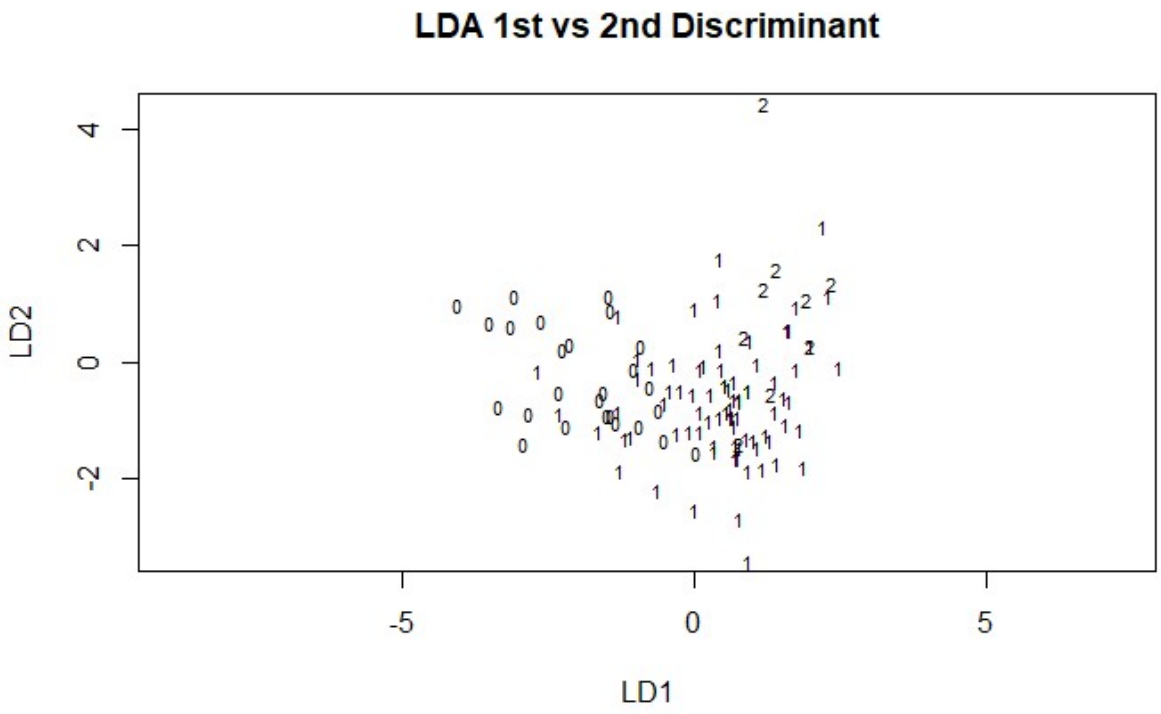

Figure 33: Plot of 1st vs 2nd Discriminant for Training Data

At the threshold of $LD1 = 0$, it separates class 0 from the rest, that is, the normal patients from cardiac disease patients. Similarly, the threshold of $LD2 = 0$ separates class 2 from the rest of the patients.

Finally, it was realized after submitting the final predictions that the features were not re-scaled to comparable scales. Therefore, distance-based models like KNN and SVM performed very poorly. Whereas, Random forest, LDA, and Naive Bayes which are invariant to feature scaling were the best models initially. After re-scaling it is seen in table 10 that KNN performs as well as the Random Forest with cross-validated accuracy of 73% on the validation set.