ECG Heartbeat Categorization Dataset Exploration report

This report is intended to explore the ECG heartbeat classification dataset, the MIT-BIH Arrhythmia Dataset, and to provide a foundational understanding of the data for subsequent modeling.

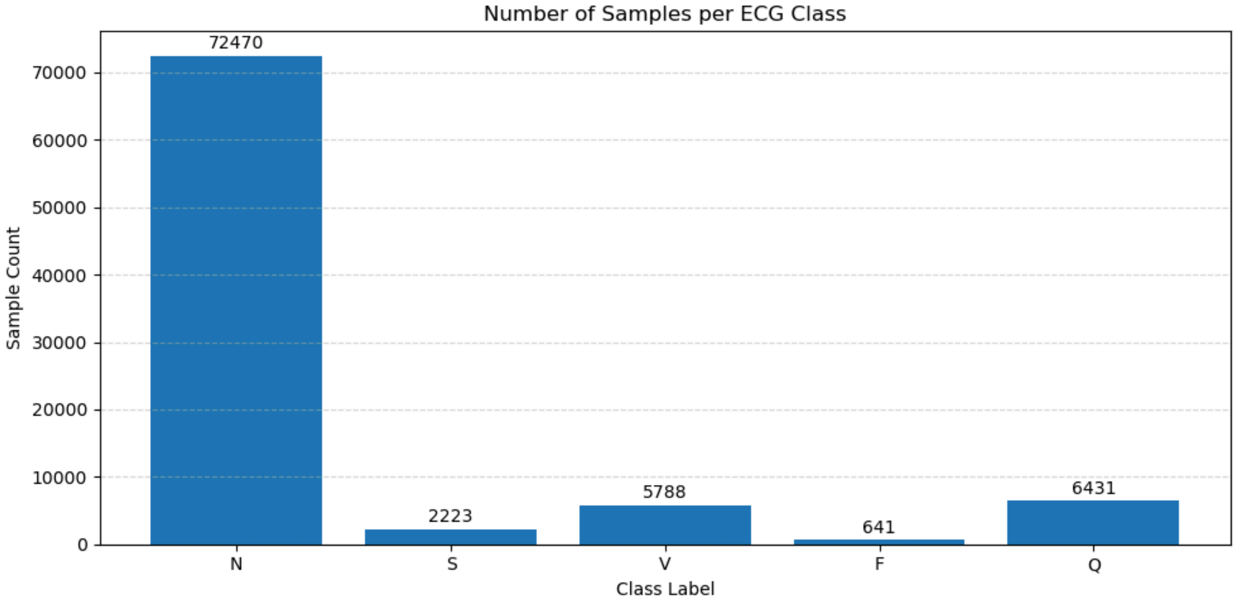
1. Dataset description

Number of Samples: 109446
Number of Categories: 5
Sampling Frequency: 125Hz
Data Source: Physionet's MIT-BIH Arrhythmia Dataset
Classes: ['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4]

Category	Code	Meaning	Indication
N	0	Normal beat	Healthy
S	1	Supraventricular ectopic beat	Pathological
V	2	Ventricular ectopic beat	Pathological
F	3	Fusion beat	Pathological
Q	4	Unknown beat	Potentially pathological

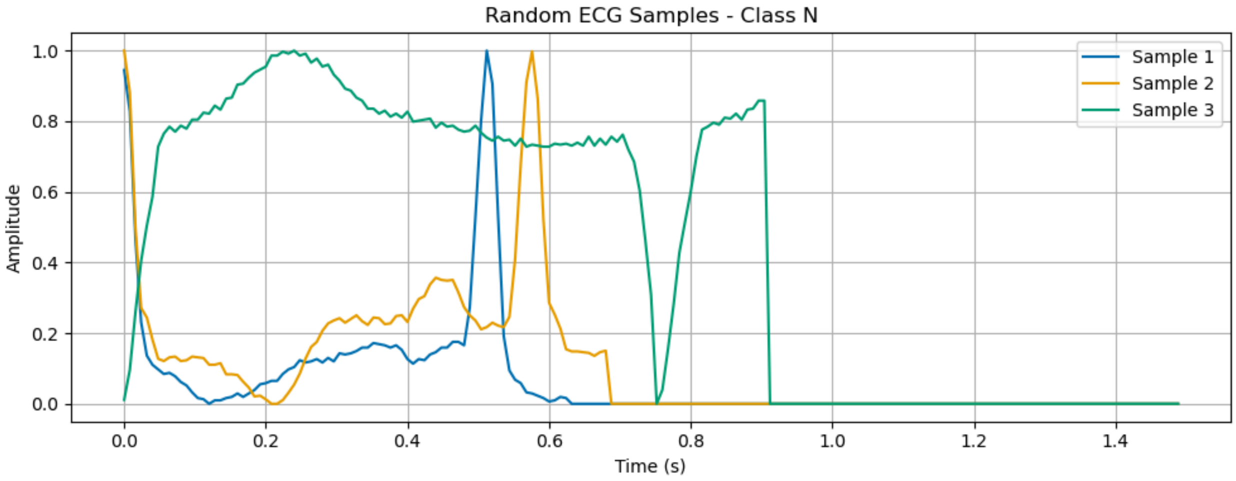
All the samples are cropped, downsampled and padded with zeros if necessary to the fixed dimension of 188. *—--dataset provider*

2. Class Distribution

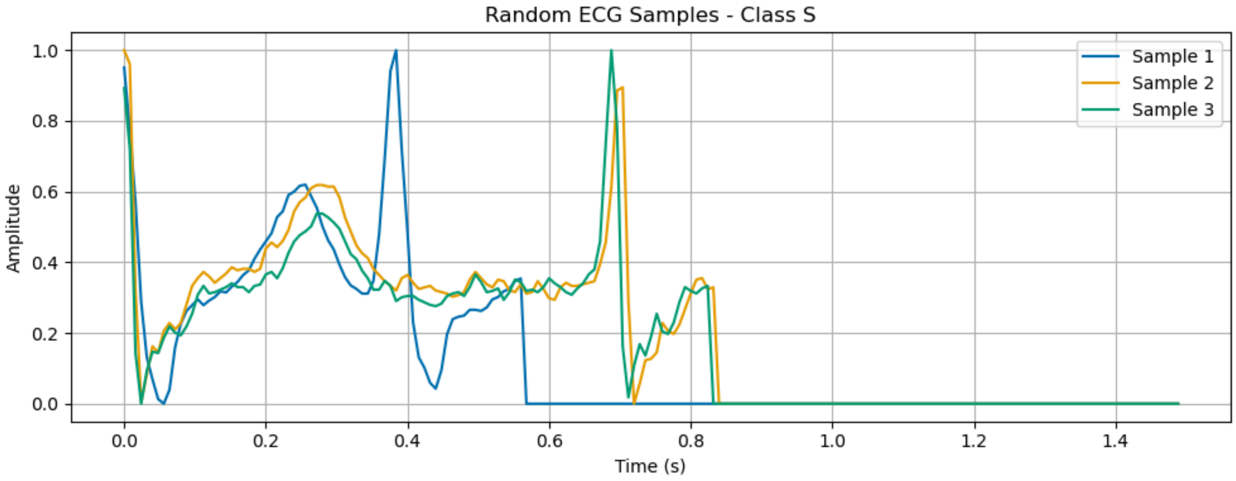


The number of samples in the Normal beat category is extremely high, while the Supraventricular Ectopic Beat and Fusion Beat categories have relatively few samples. This imbalance may negatively impact the classification accuracy for these underrepresented classes.

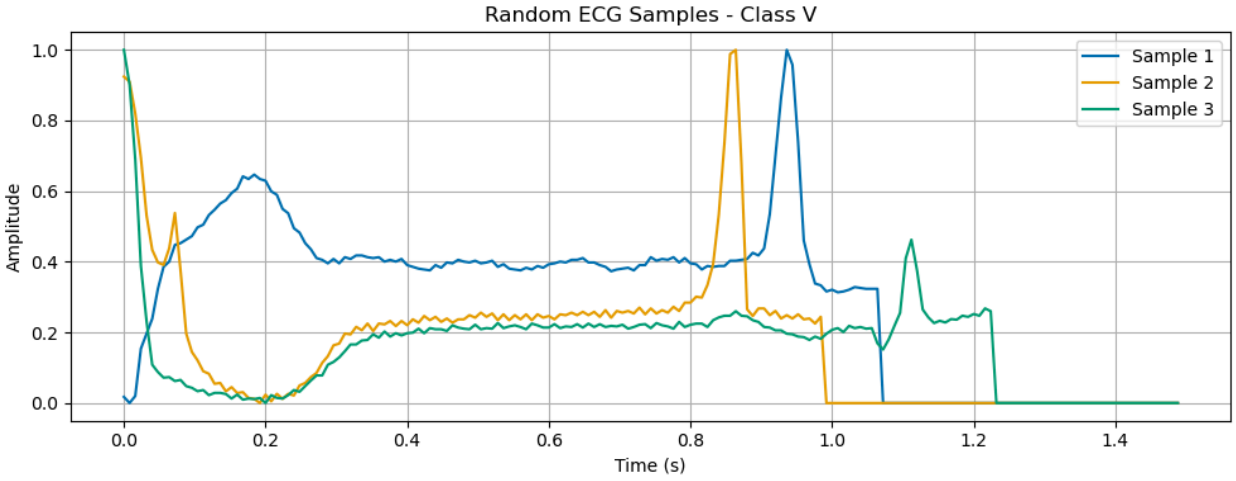
3. Random Signal Visualization



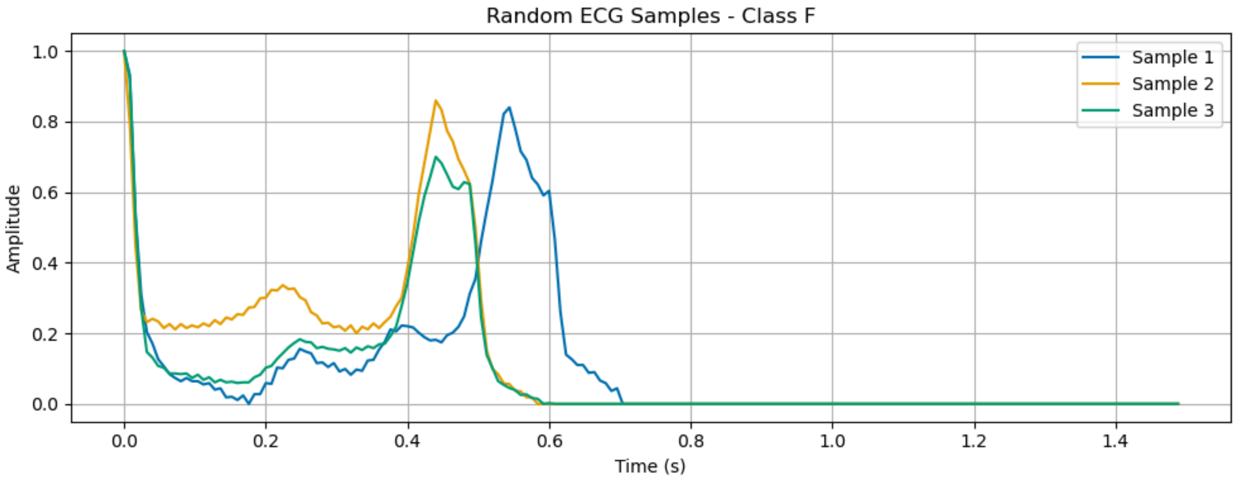
Random ECG Samples - Class N

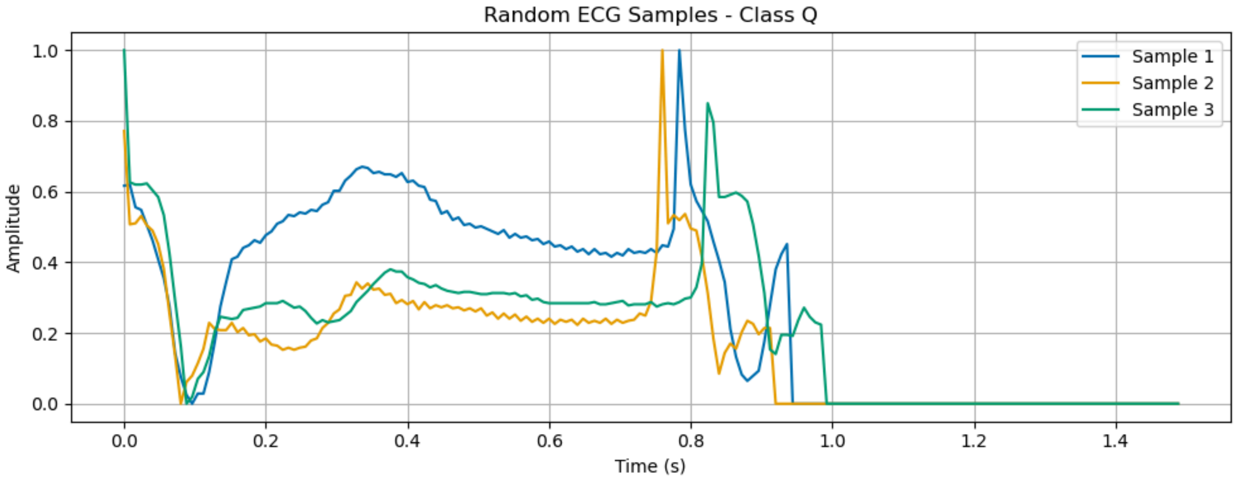


Random ECG Samples - Class S

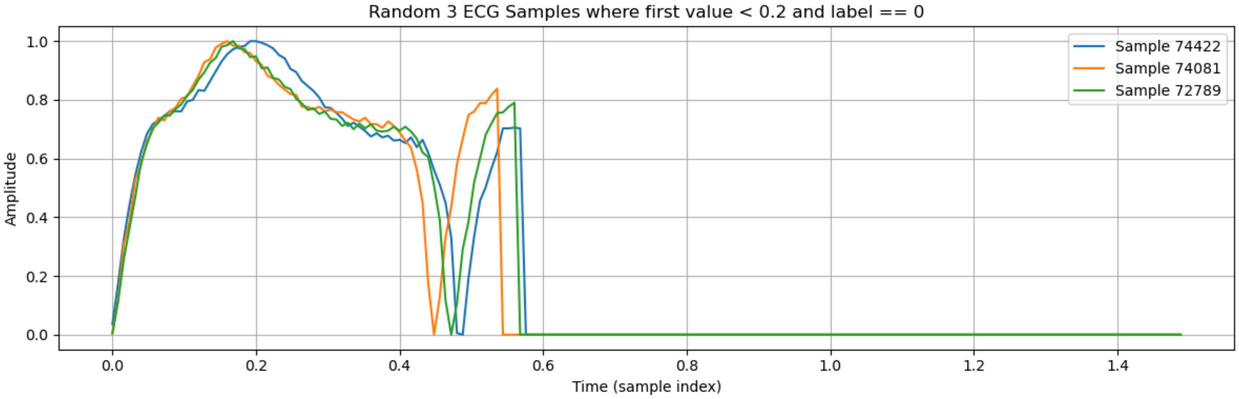


Random ECG Samples - Class V

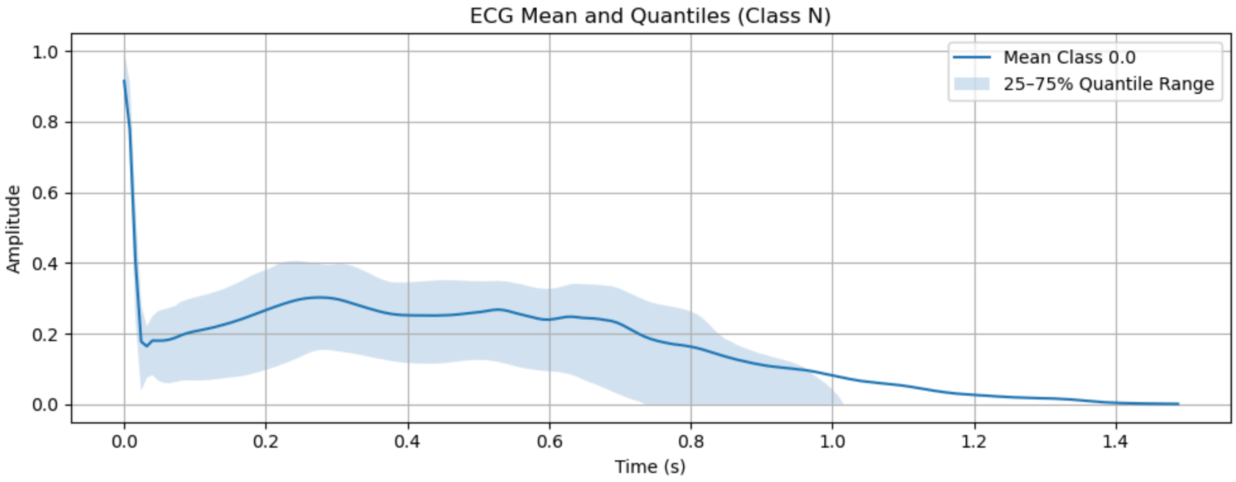




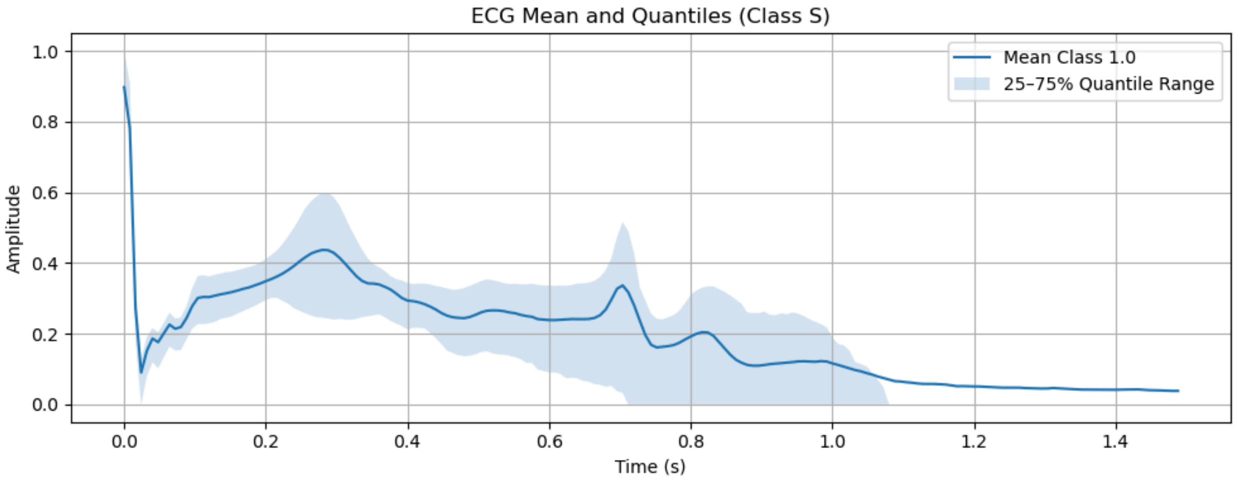
I observed that among the three randomly selected waveforms from class N, one appears to be a mirrored version around the line x = 0.5, with the end of the signal padded with zeros. Due to a lack of medical expertise, I cannot determine whether this is an issue with the data itself or a result of the cropping process. This mirrored pattern is noticeably present in both class N and class S in the training set — with 3116 samples in class N and 93 samples in class S, slightly under 5%. In the absence of expert input, I chose not to modify these samples for now, as excessive processing could lead to information loss.



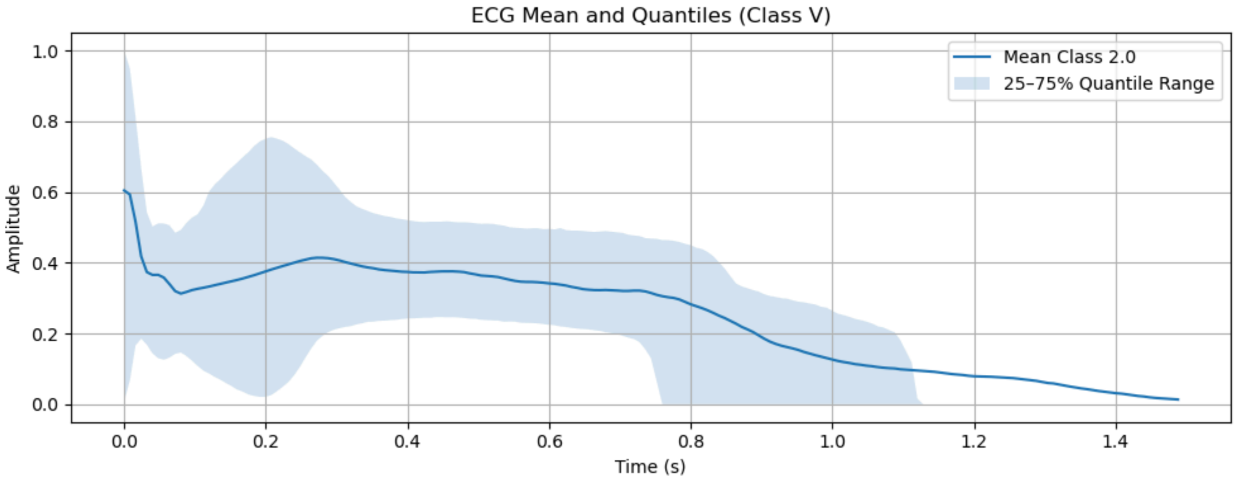
4. Signal Characteristics



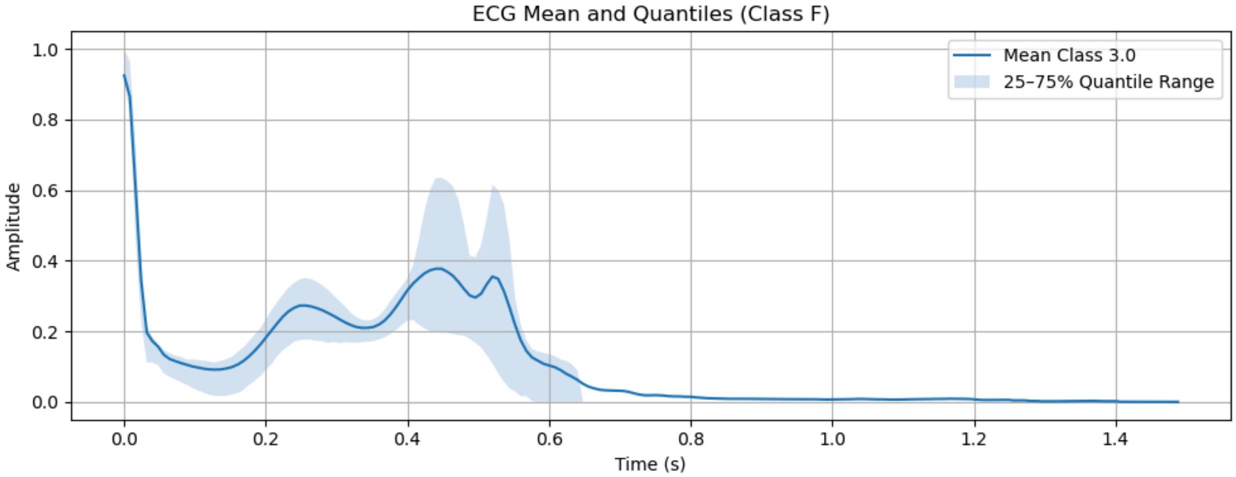
ECG Mean and Quantiles (Class N)

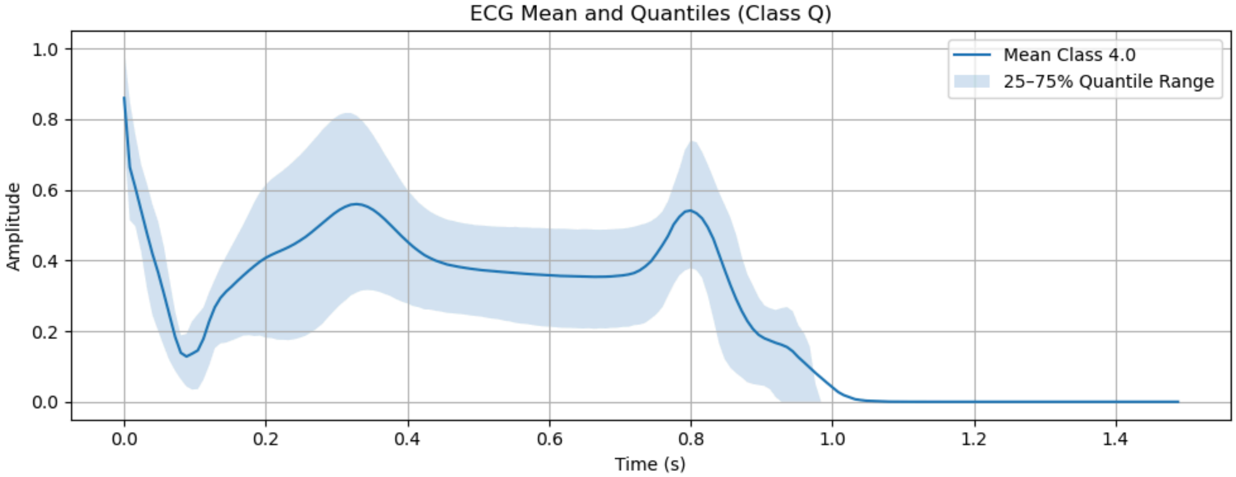


ECG Mean and Quantiles (Class S)



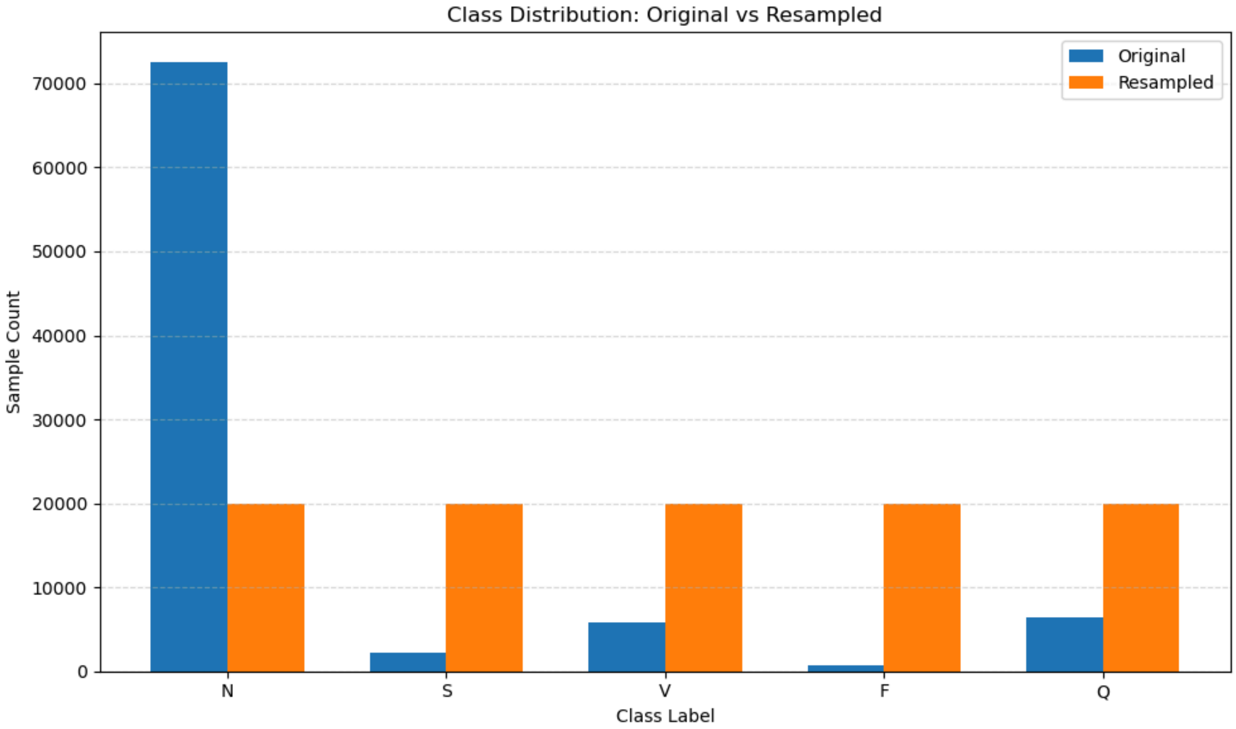
ECG Mean and Quantiles (Class V)


ECG Mean and Quantiles (Class F)
Mean Class 3.0
25–75% Quantile Range
Amplitude
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Time (s)


ECG Mean and Quantiles (Class Q)
Mean Class 4.0
25–75% Quantile Range
Amplitude
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Time (s)

5. Signal processing

I use RandomUnderSampler to under sample class N, and use SMOTETomek to over sample class S/V/F/Q. It shall reduce the imbalance impact.



6. Challenges

a. The S and F samples are extremely low, the resampled data is up to 30 times of the original one.
b. There are less than 5% patterns that we cannot tell if it came from wrong processing before.