

Data cleaning report

Classifying Arrhythmia in Fighter Pilots



March 14, 2025

Group 3

[Company address]

Contents

[1. Introduction 2](#_Toc192884728)

[2. Dataset Analysis 3](#_Toc192884729)

[Dataset Content and Collection Method: 3](#_Toc192884730)

[Data Format 3](#_Toc192884731)

[Features and Classes 3](#_Toc192884732)

[Quantitative and Qualitative Analysis: 4](#_Toc192884733)

[3. Dataset Visualization 5](#_Toc192884734)

[Class Distribution Plot 5](#_Toc192884735)

[ECG Waveform Samples 5](#_Toc192884736)

[4. Cleaning of Dataset 5](#_Toc192884737)

[Noise Reduction (Bandpass Filtering) 5](#_Toc192884738)

[Normalization: 5](#_Toc192884739)

[Segmentation and Automatic Skipping of Incomplete Windows: 5](#_Toc192884740)

[Appendices 6](#_Toc192884741)

# 1. Introduction

This project investigates unexplained physiological events (UPEs) that F‑35 pilots occasionally experience during high-stress, high-G manoeuvres. These events, which might feel like arrhythmic heartbeats or other physiological anomalies, raise the question of whether they stem from genuine cardiac irregularities or false alarms. Our goal is to develop an arrhythmia classifier that can reliably flag potentially dangerous heart rhythms in real time, even when data is recorded from different sensor types or under varied conditions.

Because there is limited real-world data showing pilots with confirmed arrhythmic episodes in flight, we train and test our methods primarily on standard datasets collected in controlled environments. Ensuring the classifier can adapt to new domains—such as different ECG sensors or novel flight conditions—remains essential. Previous experiments found that while certain traditional algorithms (like Random Forests) can excel on the dataset they are trained on, they often fail to generalize.

This report describes our **data cleaning** process, which serves as the foundation for developing a well-generalized arrhythmia detection model. By filtering noise, normalizing signals, and ensuring data integrity, we aim to produce a dataset that best captures essential heartbeat characteristics, thereby increasing our model’s reliability across diverse real-world scenarios.

# 2. Dataset Analysis

## Dataset Content and Collection Method:

Our team analysed the dataset (preprocessed\_ecg.csv), which contains 2272 ECG records. Each record comprises 216 numerical features and one categorical label specifying the ECG beat type: Normal ('N'), Arrhythmia class 'A', and Arrhythmia class 'V'. Each feature represents ECG waveform amplitude at a specific timestamp and has been normalized for consistency. The dataset originates from the MIT-BIH Arrhythmia Database, a reputable source of clinically validated ECG recordings annotated by medical experts, chosen by our team and client for its reliability and wide acceptance in the medical community.

The ECG signals in the MIT-BIH Arrhythmia Database were recorded using standard Holter monitors. The data was collected from inpatients at Boston’s Beth Israel Hospital between 1975 and 1979. Each patient was monitored continuously, and heartbeat annotations were manually reviewed and corrected by medical experts. The data was digitized and stored in a structured format for further analysis.

### Data Format

The dataset consists of:

* **Signal files (.dat):** Contain raw ECG waveform data, stored as binary files.
* **Annotation files (.atr):** Contain labels indicating heartbeat classes and rhythm events, provided as text-based numerical markers.
* **Header files (.hea):** Provide metadata about each recording, including patient ID, sampling frequency, gain factors, and lead information.

### Features and Classes

The dataset includes a variety of heartbeat types, each labelled according to the Association for the Advancement of Medical Instrumentation (AAMI) standards. The main heartbeat classes include:

* **Normal (N):** Regular heartbeats.
* **Supraventricular ectopic beats (S):** Abnormal beats originating above the ventricles.
* **Ventricular ectopic beats (V):** Premature beats originating from the ventricles.
* **Fusion beats (F):** Combined normal and ventricular beats.
* **Unknown beats (Q):** Beats that do not fall into predefined categories.

Annotations also include additional markers for rhythm changes, signal quality indicators, and other relevant cardiac events, such as:

* **Atrial fibrillation (AFIB)**
* **Atrial flutter (AFL)**
* **Ventricular tachycardia (VT)**
* **Paced beats (P)**

## Quantitative and Qualitative Analysis:

Our comprehensive analysis yielded the following insights:

* **Number of Records:** 2272
* **Number of Features:** 216 numerical amplitude measurements per record
* **Feature Types:** 216 numeric (float64), 1 categorical ('label')
* **Range of Values:** ECG amplitudes range approximately from -0.27 to 1.24.
* **Classes and Annotation:**
  + Normal ('N'): 2238 samples (~98.5%)
  + Arrhythmia ('A'): 33 samples (~1.45%)
  + Arrhythmia ('V'): 1 sample (<0.1%)

Our analysis highlighted substantial class imbalance, which we identified as a critical factor impacting subsequent modelling.

# 3. Dataset Visualization

Our team performed detailed visual analyses, including:

* Class Distribution Plot**:** Clearly illustrates the significant class imbalance, emphasizing the dominance of normal beats ('N') compared to arrhythmia beats ('A', 'V').

A graph showing a number of signals

AI-generated content may be incorrect.

Figure 1

Figure 1 highlights the substantial imbalance within our ECG dataset, with the class "N" (Normal) clearly dominating at approximately 98.5% of the total samples. In contrast, arrhythmia classes "A" and particularly "V" are significantly underrepresented. This imbalance poses potential challenges for training accurate and unbiased predictive models, making data balancing techniques highly recommended.

* ECG Waveform Samples**:** We visualized sample waveforms from each class, highlighting typical shapes and characteristics.

A graph of a signal

AI-generated content may be incorrect.

Figure 2

Figure 2 demonstrates distinctive characteristics associated with ventricular arrhythmias. The waveform displays irregular signal amplitudes and atypical morphological features compared to normal ECG patterns. Recognizing these specific anomalies visually assists our team in accurately identifying and handling such outliers during modeling and preprocessing stages.

Since our dataset is designed to classify arrhythmia vs. no arrhythmia, calculating the median or mean ECG values is not particularly meaningful. Unlike continuous numerical datasets where central tendency metrics provide insights into data distribution, ECG signals are inherently categorical in their classification (normal vs. arrhythmic beats). Additionally, identifying outliers is out of scope for this analysis, as our focus is not on analysing rare arrhythmias but rather on ensuring the dataset effectively distinguishes between normal and arrhythmic beats.

# 4. Cleaning of Dataset

## Noise Reduction (Bandpass Filtering)

* The ECG signals were passed through a bandpass filter (0.5–40 Hz) to remove low-frequency drift and high-frequency noise. This ensures the core heartbeat frequencies are preserved while eliminating much of the irrelevant “static” or artifact in the signal.

## Normalization:

* After filtering, the code normalizes each signal by subtracting its mean and dividing by its standard deviation. This step brings all heartbeats onto a consistent scale, making later analysis or model training more reliable.

## Segmentation and Automatic Skipping of Incomplete Windows:

* The code extracts a portion of the signal around each R-peak (200 ms before to 400 ms after). Only segments fully within the recording are kept.
* No manual datapoint removal or outlier removal was performed beyond skipping segments that would exceed the recording boundaries.

### Data Augmentation Considerations

Given the severe class imbalance in our dataset (98.5% normal beats vs. 1.45% arrhythmia class A and <0.1% class V), data augmentation could be a useful technique to improve model performance. However, augmentation techniques such as SMOTE or synthetic ECG generation were not applied in this stage.

Potential augmentation strategies include:

* **Synthetic ECG data generation** using generative models like GANs or variational autoencoders to create realistic arrhythmic beats.
* **Time-warping or jittering** of existing arrhythmia samples to create variations in the dataset.
* **Oversampling** minority classes to ensure a more balanced distribution during training.

Since our focus in this stage was on cleaning the raw ECG signals, augmentation was not included in this pipeline but remains an important step for improving classification accuracy. Future iterations could incorporate augmentation techniques to address the imbalance issue and ensure better model generalization.