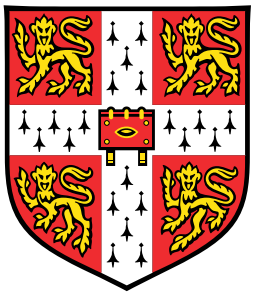
Data of Your Heart: Screening for Atrial Fibrillation



MEng Project Report

Jordan Smith
JS2432

Supervisor: Dr Elena Punskaya

Division F Engineering Department
University of Cambridge, 2022
United Kingdom

June 2022

Disclaimers

Student Disclaimer

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Author: Jordan Elliot Smith
Signed:

Abstract

An abstract.

Contents

1 Introduction

1.1 The Problem

Atrial Fibrillation (AF) is a common abnormal heart rhythm that is associated with a five-fold increase in stroke risk [4]. After being recommended for examination at a hospital, a patient has an ElectroCardiogram (ECG) taken, and may be diagnosed with AF if the signifying features of AF are detected in this ECG taken during monitoring. If a patient is diagnosed with AF, medication can be administered to reduce the stroke risk. Currently, however, patients only have their ECG taken if they are symptomatic with AF, such as experiencing heart palpitations, but a significant proportion of AF patients will be asymptomatic, therefore not referred for examination.

Furthermore, some patients will only experience AF episodes, which are short lasted episodes of AF which occur at varying frequencies, rather than consistently showing signs of AF. These episodes are often not detected during ECG monitoring in hospital, and therefore the diagnosis is missed despite the risks of AF still being present.

AF is identified through irregular-irregular RR intervals and an absence of P waves in an Electrocardiogram (ECG), see Figure 1. This ECG is taken at a single time point during a hospital visit, therefore it is likely that an AF episode is missed, especially for patients with low AF burden.

Figure 1: Electrocardiogram (ECG) recording with Zenicor device [5], showing Atrial Fibrillation (AF) (a) and sinus rhythm (b).

Screening offers a method to identify these asymptomatic, low AF burden patients who are otherwise neglected by current methods, but still at risk. AF burden is a measure of the proportion of time a patient experiences AF episodes, and it increases as the age of the patient increases [6], meaning the risk of AF continuously increases and should be monitored. The SAFER study [7] is looking into using the Zenicor 1 lead ECG device for patients to use at home, unassisted, to take 30-second samples 4 times a day over 3 weeks, at low cost. This significantly increases the chances of detecting AF episodes in low AF burden patients, and therefore enables the correct treatment or monitoring of these at risk patients.

However, this generates a significant number of samples, with 84 per patient, which need to be checked by a Cardiologist. Currently, the Cardiologist searches through all of these samples and if at least one sample out of 84 shows AF signs, the diagnosis is made. With only one AF sample needing to be found, this process can be considerably more efficient if this AF sample is the first one seen by the cardiologist. This is the motivation for this project.

1.2 The Challenge

While increasing the time frame of the ECG, screening using a 1 Lead ECG has issues of increased noise and decreased coverage of the heart's electrical activity.

The increased noise arises from multiple environmental factors, including user error. A 12 lead ECG taken in hospital is generated with trained professionals using sophisticated equipment, whereas a 1 lead ECG will be self administered. The resulting generated samples will be susceptible to contaminated contact points, movement of the patient, possible external signals such as from nearby electronic devices. This leads to a signal which is noisy, has baseline wander, and will have high frequency components not seen in 12 lead ECG.

Furthermore, the decreased coverage of the hearts electrical activity is due to the 12 Lead ECG taking signals from multiple angles across and through the heart, as seen in Figure 2, whereas a 1 lead ECG only detects transverse signals through the heart, labelled as "1" in Figure 2. This leads to significant losses in the signals that can be detected. In particular, "p" waves which would be found in a 12 lead ECG may be completely undetectable in a 1 lead ECG, which could incorrectly be used to diagnose AF when Atrial Depolarisation, the process which causes the p waves, is actually occurring.

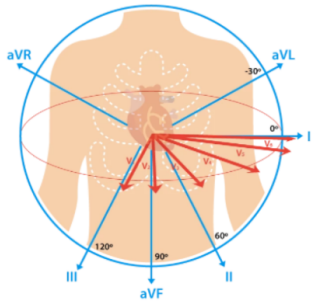


Figure 2: 12 lead ECG signal locations [8]

Another consideration for this project is the presence of other heart arrhythmia in the samples, such as Bradycardia, Tachycardia and Heart Block. These can easily be confused with AF, and therefore the distinction between them is important.

1.3 The Objective

With the vast amounts of data the Cardiologists will need to analyse with the SAFER trial, there is a need for a more efficient review process. Knowing that if a Cardiologist finds a sample which shows AF signs, they will diagnose the corresponding patient with AF regardless of the rest of the samples. This means if there is a sample that contains AF, significant time and resources are saved if it is the first one seen by the cardiologist. Automated algorithms can be used to identify abnormal ECGs to be reviewed manually [9], but these lead to approximately 35 ECGs being reviewed to identify each pathological ECG.

The aim for this project, therefore, is to find a model which generates a representative probability of an ECG sample exhibiting AF, to help order the samples which increases the likelihood that the AF samples appear first.

2 Literature Review

2.1 Background

A highly active area of research, classifying ECGs for AF or Sinus (normal) Rhythm (SR) has seen numerous academic competitions, with the largest scale being the Physionet Challenges [10] [1] [2]. Through these competitions, many approaches to this classification have been developed and implemented for use on 1 lead ECG data. Traditional approaches involve more hand crafted feature detection, with approaches using Signals Processing techniques to detect the different features of the ECG signal and then running deterministic algorithms on these feature properties.

2.2 Traditional Approaches

One particular, more traditional approach, is the one proposed by Dr Jie Lian et al. in their paper "A Simple Method to Detect Atrial Fibrillation using RR Intervals" [11], classifies ECGs using only the RR intervals of the ECG. It takes the RR intervals, calculates the difference in RR intervals at each step, and plots these two properties with a grid segmentation of the values to count the proportion of grids which are populated by at least 1 point in a scan of given time length. This method was tested on the MIT-BIH Atrial Fibrillation database [3] which is comprised of data gathered using a Holter device. A Holter device produces significantly different ECG samples to a 1 Lead handheld ECG device. The presence of 'extra peaks' which would be identified in a noisy samples could reduce the accuracy of this method.

2.3 Modern Machine Learning Techniques

In the past five years more focus has been placed on Machine Learning (ML) techniques, and especially Deep Learning (DL), to be used for the task of ECG classification. The ability of these methods to function well on noisy ECGs has been especially beneficial. The DL methods increasingly display the capability to learn the waveform shapes, both small and large scale features that correspond to AF or SR, and in some proposals the ability to distinguish between AF and other heart arrhythmia, along with identifying samples that are too noisy. This is especially useful for the application to the SAFER study, where the expectation is that plenty of noisy samples will be produced by the Zenicor device due to incorrect operation, and the presence of other heart arrhythmia is likely, therefore a model which distinguishes between these classifications is very useful.

In 2017, one of the aforementioned Physionet challenges [1] took aim at classification of noisier Lead 1 ECG samples, with data collected from small handheld devices, only differing from the SAFER trial data in that they are variable length, with some less than 10 seconds and other longer than a minute. The challenge asked for methods to classify samples between "Normal" "Atrial Fibrillation" "Other heart arrhythmia" and "Noisy" categories, measuring the performance as the average F1 score over each category. Two of the winning submissions to this challenge were investigated in depth.

1. "ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks" [12]. This method combined the use of traditional feature detection from Statistics, Signals Processing and Medicine, with modern DL methods that learn features through data-centric approach. This proposal method had a high accuracy with an F1 score of 0.83. This method was not pursued for the project, however, due to most of the methodology being omitted from the report submission. Therefore applying this method from scratch, without assistance from trained Cardiologists, is beyond the scope of this project.

2. "Robust ECG Signal Classification for Detection of Atrial Fibrillation Using a Novel Neural Network" [13], applied a more DL centred approach, with a novel form of the famous ResNet [14] neural network, using residual connections, being applied for the ECG classification. This method is particularly interesting because not only did it generate exceptional results in the challenge, with a near winning F1 score of 0.82, but it also had a complete architecture and algorithm explanation included in the report.

3. The paper, Identification of patients with atrial fibrillation: A big data exploratory analysis of the UK Biobank [15], analysed the performance of 10 ML techniques with some being classical ML approaches using Support Vector Machines, and others being a combination of classical ML with DL approaches. On the subset of the UK Biobank dataset the combination approach proved to be the most effective at ECG classification. This approach, however, also included the used of expert features assumed beyond the scope of this project.

3 Methodology

3.1 The Data

The data used for this project came from the Physionet [10] for most of the development and research of methods, with the application of these chosen methods on the SAFER datasets used for results to prove the methods effectiveness for application in the SAFER study. The Physionet datasets are open source, and readily available online. The SAFER datasets is used for testing of the models, with the evaluation of performance being most important for this data.

3.1.1 Physionet Challenge 2017 Dataset [1]

This open source dataset contains Lead 1 samples taken on the AliveCor device, which has similar properties of ECGs provided as that of the Zenicor device. The samples were taken at 300 Hz, then band passed by the AliveCor device before being labelled as follows:

1. "A" which corresponds to AF.
2. "N" which corresponds to Normal (Sinus) Rhythm.
3. "O" which corresponds to Other heard arrhythmia such as heart block.
4. "∼" which corresponds to Noisy sample.

This dataset has samples with similar noise properties, heart signal coverage and sampling rates as those provided in the SAFER datasets, yet the lengths of the samples vary from 9 seconds to over 60 seconds long. 8528 of these samples were available online.

3.1.2 Physionet Challenge 2020 Dataset [2]

The data for the 2020 challenge came from 4 open sources, all of which had 12 lead ECG data:

- CPSC Database and CPSC-Extra Database
- INCART Database
- PTB and PTB-XL Database
- The Georgia 12-lead ECG Challenge (G12EC) Database

The China Physiological Signal Challenge (CPSC) Database [16] was used in this project. It contained 6877 samples, each from 6 to 60 seconds long and taken with sampling frequency 500Hz. Due to these samples being taken by trained professionals in a hospital using 12 lead ECG equipment these samples are not as useful as the 1 lead data otherwise sourced, because they are less representative of the data expected to be seen in the SAFER study. These samples are less noisy, and have different properties in waveform detected. The lead 1 data can be isolated from the other 11 lead data, and is still useful for general comparison of methods, with due consideration.

3.1.3 MIT-BIH Dataset [3]

This dataset contains "8 half-hour excerpts of two-channel ambulatory ECG recordings" [17].

3.1.4 SAFER Datasets

These datasets were from the SAFER Feasibility Study (ISRCTN 16939438) which assessed feasibility of delivering an AF screening program in Primary Care, approved by the London - Central Research Ethics Committee (REC ref: 18/LO/2066). Further data was sourced from the SAFER trail study. Participants aged 65 and over were screened for AF, each asked to record 30-second ECGs four times a day. ECGs were acquired between two thumbs using the Zenicor EKG-2 device (Zenicor Medical Systems AB), as shown in Figure 3.

Figure 3: The Zenicor EKG-2 device which acquires 1 Lead ECGs

This data comes in the form of 3 different datasets, created at different times in the SAFER study:

1. SAFER Feasibility study 1 dataset
2. SAFER Feasibility study 2 dataset
3. SAFER Trail dataset

The statistics of each dataset is shown in Table 1.

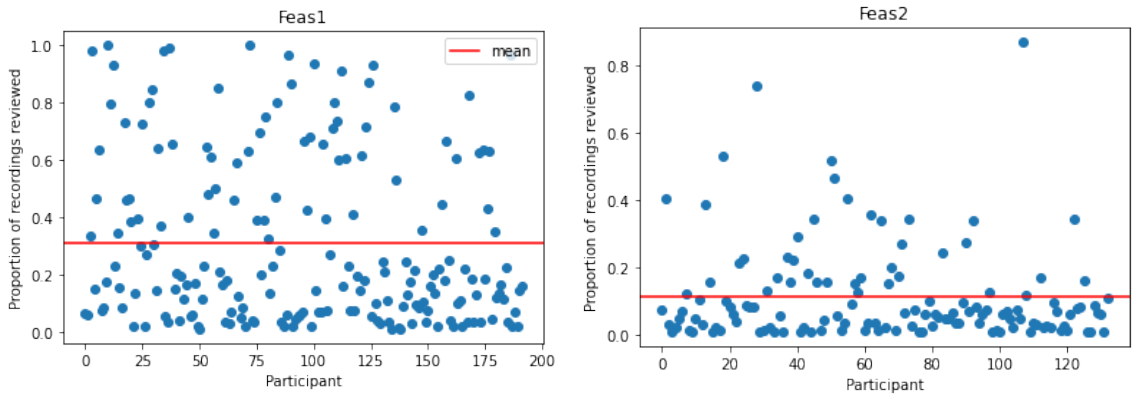


Figure 4: Proportions of recordings individually reviewed per participant in SAFER Feasibility studies 1 and 2

The primary data source for this project, the SAFER study [7] dataset contains 14,235 samples that have been individually reviewed by cardiologists, and over 300,000 samples in which the patient themselves has been diagnosed as having AF or not. In these samples, there will be numerous that are ground truth labelled as 'AF', yet do not actually display any signs of AF, because those individual samples are from a patient with low AF burden and therefore not all the samples from this patient will show AF signs.

Another consideration of the datasets is the skew towards 'Non-AF' samples. For the SAFER datasets there was less than 20% of samples labelled "AF". For the Physionet 2017 dataset only 9%. Results need to be framed with this in mind.

Finally, the presence of samples showing the aforementioned other heart arrhythmia is another consideration in each dataset. These can be easily confused with AF if the model is incorrectly trained and does not recognise the correct features. With these arrhythmia often present in the general population, it is very important for the model to be able to distinguish between them, and as such the careful consideration of how to choose the model and which classifications it should output will be needed.

	Dataset		
	Feas1	Feas2	Trial
No. recordings	162515	23259	272945
No. participants	2141	288	3338
Duration of screening	1,2,4	3	3
Study period	March 2019 - November 2019	October 2020 - January 2021	May 2021 - January 2022
Proportion of participants diagnosed as AF	65 (3.0%)	10 (3.5 %)	89 (2.7 %)
Proportion of participants diagnosed as no pathology	2027 (94.7 %)	277 (96.2 %)	3241 (97.1 %)
Proportion of participants diagnosed as other arrhythmia	40 (1.9%)	N/A	8 (0.2 %)
% high quality recordings	157781 (97.1 %)	22679 (97.5 %)	264476 (96.9 %)
Medium No. recordings per patient	61	83	81 (77-84)
No. participants reviewed by at least 1 Cardiology team	190 (8.9 %)	288 (100.0 %)	3338 (100.0 %)
No. recordings reviewed by at least 1 Cardiology team	4494 (2.8 %)	1241 (5.3 %)	8500 (3.1 %)
No. recordings labelled AF	137 (3 %)	16 (1 %)	1416 (17 %)
No. recordings labelled no pathology	1418	758 (3.3 %)	N/A
No. recordings labelled other arrhythmia	31	2 (0.0 %)	N/A
No. recordings labelled poor quality	416 (0.3 %)	465 (2.0 %)	7084 (2.6 %)
No. recordings labelled undecided	13	22018 (94.7 %)	264445 (96.9 %)

Table 1: Statistics of SAFER datasets

3.2 The Novel Neural Network (NNN) Approach

Two approaches were tested in depth. Predominantly, the DL approach using the "Novel Neural Network" [13] was examined. This approach was proposed during the Physionet 2017 challenge. Furthermore, a more simplistic approach using "RR dRR intervals" [11] was investigated in order to trial a low computational expense method and compare the results to the NNN approach.

Starting with the NNN model, which was developed for the Physionet challenge 2017 dataset which aimed at classifying 1 Lead ECG samples as stated in section 3.1.1. An inherently data driven method with all feature detection methods, parameters and classification likelihoods learned through training the model, this method relied on a large dataset provided.

3.2.1 Network Architecture

The neural network developed by Xiong et.al. [13] can be broken down into 16 convolutional blocks, each with skip connections from previous blocks and each block being down-sampled from the previous layers output. The final convolutional block is the input to a fully connected layer which feeds into a softmax followed by classification. The convolutional blocks contain the following components, see Figure 5:

1. Input from previous layer added to skip connection output passed through max pooling layer
2. Batch Normalisation
3. ReLU activation
4. Dropout
5. 1D Convolution (15×1 kernel)
6. 1D Average Pooling

The training process of such a deep network was carefully designed by the proponents of this method to increase efficiency and performance. The ADAM optimisation algorithm used for finding parameter values utilised gradients calculated from backpropogation at each step, and for deep neural networks these gradients have been shown to stagnate, the vanishing gradients problem. Skip connections using "residual" data from previous layers has been shown to reduce this effect [14] by ensuring non-zero gradients are calculated, especially in conjunction with batch normalisation. Batch normalisation normalises the outputs from each layers over the batch currently being trained on, by taking estimates of the mean and variance of these outputs for each batch and zero normalising with unit variance. This further increases training efficiency. Also, in training the use of dropout by "turning off" nodes in a stochastic manner reduced the model's ability to overfit to the data provided. Finally, ReLU activation was used to further increase the models ability to fit to the data by enabling non linear decision boundaries, with the increased efficiency that comes with this simple activation function.

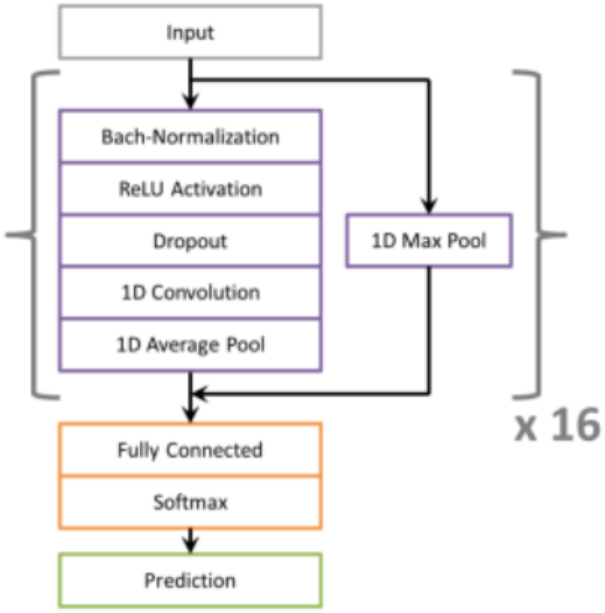


Figure 5: Novel Neural Network architecture with skip connections [13]

Pooling was used throughout this model. This developed the models ability to learn features at different scales with the convolutional layers 15×1 kernel being trained for each different scale allowing this hierarchical learning. This was implemented through average pooling in each convolutional block with max pooling used for the skip connections to ensure dimensional continuity. This allows the feature vector, which is the output of the final convolutional layer, to be a representative vector of features across many scales, trainable through a large number of labelled data samples, that can then pass through a fully connected layer which learns to classify the input image by learning the corresponding patterns in this feature vector for each class. The output of this densely connected layer is a 4×1 vector which, once passed through the softmax layer, is a probability assigned to the input image for being in each of the 4 classes. The end result is a classification made by taking the class with highest predicted probability.

3.2.2 Training Data and Parameters

Input to model is 5 second sample, prediction per sample

input is lead 1 relatively noisy data

parameters trained in 2017 challegne, these taken for this project

Model

The Novel Neural Network (NNN) approach [13] chosen has architecture as shown in Figure 5. This network was applied to samples split into 5-second sections, with each section having a classification prediction made for it, and then the classification that appeared the most over all of these split samples is the one chosen as the classification for the entire sample. Although this method was initially applied to a classification problem, the model will be adapted to produce a probability, rather than classification, of AF. There are multiple options for achieving this, with the simplest taking the proportion of the sample splits which

are classified as AF as the likelihood of the sample showing AF. A more likely option to achieve better, more continuous, results would be to examine the output of the softmax layer, and through experimentation find a suitable way of combining these outputs into a confidence of AF.

The original NNN model produced for the Physionet challenge 2017 was programmed to classify between 'AF', 'N' (Sinus rhythm), 'O' (Other heart arrhythmias) or ' ' (Noisy sample). The work of this project is only interested in 'AF' or 'Not AF', therefore this has been adapted accordingly. Currently 'O' and 'N' are combined, and '∼' are not of interest other than to state that the model was not confident for samples labelled '∼'.

3.3 RR dRR Intervals Approach

3.3.1 RR dRR Calculation

The RR values are counted by first detecting the R peak locations. The R peak is the location of the peak of QRS complex found in an ECG, as seen in figure 6.

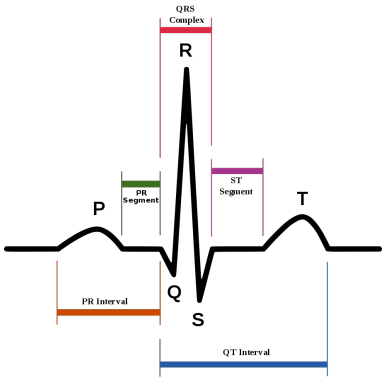


Figure 6: ECG wave locations [18]

3.3.2 Grid Counting

3.3.3 Threshold and Optimisation

3.4 Research Timeline

3.4.1 Preliminary Research

Initially, after briefing with Dr Peter Charlton from the SAFER team, the background of the research and aims for the project were outlined. This resulted in the project objective of prioritisation of ECGs based on a predicted probability of AF being present in the ECG sample. This needed to be scalable to the rest of the SAFER trail data. An understanding of the dataset, its limitations and properties was developed at this stage.

A literature review was carried out, with multiple significant methods being analysed before narrowing down to a few methods considered most appropriate. Following this, other data sources were researched to assist with development and research while awaiting complete access to the SAFER data.

3.4.2 NNN Approach Experiments on Physionet Datasets

After establishing this method to be most appropriate for the project aims, the tooling and implementation code was developed to use it. Taken in its pretrained state, trained on Physionet 2017 dataset by the proponents of the method, this method was adapted with its input and preprocessing adapted to allow its application to other datasets. Variables between datasets included sampling frequency, noise levels, baseline wonder and corresponding band-pass filtering and sample lengths.

Furthermore, the tooling was built for working with the predictions made by the model, and analysing these results for assessment of the models effectiveness. Results were then derived for the classification accuracy of the model on the Physionet 2017 dataset and then Physionet 2020 dataset. After proving the models effectiveness, the output of the model was adapted to provide, for a given ECG recording, a probability of AF rather than a prediction of the class it falls into. This enabled this models use for the project aims.

3.4.3 RR dRR Approach Experiments

In order to evaluate the effectiveness of the NNN method, comparison to a more simplistic traditional method was needed. The RR dRR approach was implemented from scratch for this reason.

After implementing the method, experiments were carried out on optimisation of the parameters the method requires through different cost functions and degrees of freedom used. Different grid sizes was experimented with, as well as the cost function used for optimisation of the threshold used as the decision boundary. Two cost functions were experimented with, one hand crafted explored in Subsection **??** as well as one based of the F1 score of the output to enable best comparison with the NNN approach.

After optimisation on 2017 dataset, this method was tested on the Physionet 2017 and Physionet 2020 datasets, and showed the more advanced NNN approach to be more effective especially on the noisier samples taken from Lead 1 ECGs. This method was also trailed on the MIT-BIH dataset but time limitations in the project preveneted extensive analysis on this data.

3.4.4 NNN Aproach Experiments on SAFER Datasets

Investigation into Feas1 and Feas2 properties of manyally reviewed smaples Trail on Feas1 results satisfactory with probabilites being as aimed for

3.5 Result Evalutation Techniques

3.5.1 F1 Score

The F1 score for a classification task is a function of the Recall and Precision of predictions made. Precision, Recall and the F1 score are defined below, in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$\text{precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + 0.5(FP + FN)} \tag{3}$$

3.5.2

4 Results

4.1 NNN Approach Testing

4.1.1 Tested on Physionet Challenge 2017 [1] Dataset

As a quick confirmation of the feasibility of the method, the NNN model was tested on the Physionet 2017 dataset on which it was trained. Unsurprisingly, the classification results were very good, see Table 2 for results. The mean F1 score of 84.25 is slightly higher than the competition score of 82, which is to be expected because the model had already seen this data, and so higher accuracy is expected than those found running the model on the test dataset from the competition which it hadn't seen before.

These results were promising due to both their accuracy, but also the model had clearly not overfit to the train dataset due to no large increase in accuracy when testing on the train dataset. This observation, therefore, showed the suitability of extending the model to SAFER data.

With these promising results confirming the model functioned as expected, the model was further investigated on the Physionet 2020 dataset.

Classification	F1 Score	Support
A	0.88	743
N	0.93	5070
O	0.84	2464
~	0.72	286
Accuracy avg	0.84	
Weighted avg	0.89	

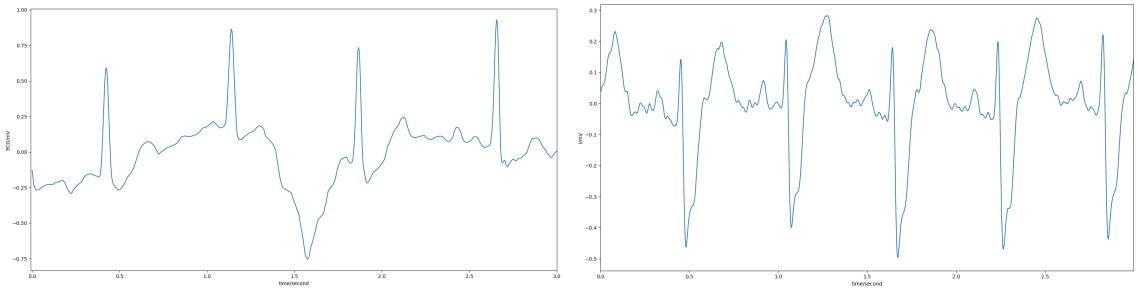
Table 2: Results of NNN model tested on Physionet challenge 2017 dataset

4.1.2 Tested on Physionet Challenge 2020 Dataset

The next step was to apply this model to data that it had not seen before. The CPSC Database [16] explained in Section 3.1.2 was used, with the Lead 1 data being isolated from the entire 12 Lead sample. Testing on this data which was collected in a hospital by trained professionals is not fully representative of the model performance for the application to data collected using the Zenicor device, for the SAFER trial, due to different ECG characteristics, yet is still a useful tool for validation purposes. See Figure 7 for a comparison between a typical 1 Lead handheld device recording and a Lead 1 recording taken using 12 lead ECG in a hospital.

The classification results of the model on this dataset are shown in Table 3, with only the classifications of 'A' or 'N' of interest. These results were gathered from all samples in the CPSC database which were either labelled as "AF" or "Normal", or labelled by the model as one of these results. No "Other" heart arrhythmia labelled or predicted data was included, or noisy samples, which explains the very high F1 scores.

Firstly, the model clearly performed very well on this dataset with F1 score results significantly better than those from the test on the models own train dataset. This is initially



(a) Physionet challenge 2017 sample

(b) CPSC sample

Figure 7: Samples from the two open-source datasets, showing the benefit of the ECG taken at hospital with more clearly defined p, qrs, and t waves (7b), and the drift and noise sometimes found in samples taken from self administered 1 lead ECGs (7a).

surprising, but is explained by the lower noise present in the samples from the CPSC database. This means it is easier for the model to depict the features corresponding to AF or Normal recordings. This lower noise is due to the higher quality equipment and training.

Secondly, the model is notably more accurate at distinguishing between AF or Normal samples, with significantly better results when noisy samples and other heart arrhythmia are present. This is an encouraging result, which suggests one of the key challenges of the project hereon would be getting the model to distinguish between other heart arrhythmia and AF.

Finally, the key result from this experiment was that the model performs well on data it has not seen before. This is an essential for application of the model to unforeseen data in the SAFER trail.

Classification	F1 Score	Support
A	0.98	918
N	0.90	1221
Accuracy avg	0.94	
Weighted avg	0.93	

Table 3: Results of NNN model from Physionet challenge 2017 tested on lead 1 CPSC data [16]

4.2 Preprocessing for SAFER data Experiments

The next step for application of the NNN method to SAFER data was to find the necessary preprocessing steps for this data, to ensure it looks as similar as possible to the data the model had seen before and been trained on in the PhysioNet 2017 challenge. Various experimentation was carried out looking at 5 key components of the signal that needed processing:

1. Baseline wander
2. High frequency noise

3. Scale normalisation

4. Sample frequency consistency

5. Sample length consistency

Baseline wander was removed from the signal using Asymmetric Least Squares smoothing [19] which was implemented using the code found in Appendix section 5.1.1. This resulted in the signal transformation seen in Figure 8.

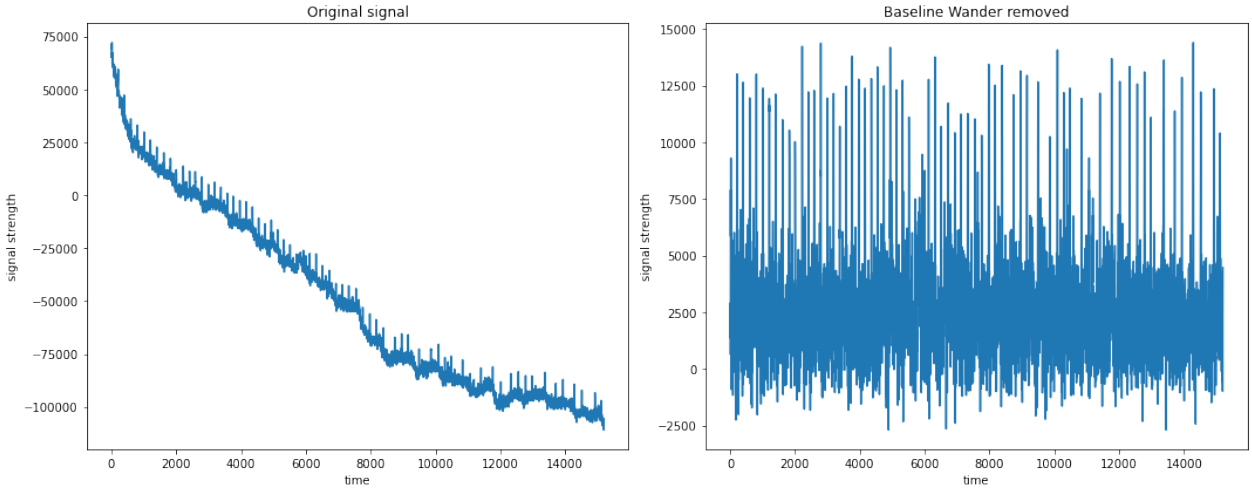


Figure 8: Signal difference removing the baseline wander initially present in SAFER data

The next step was to reduce the high frequency noise components of the signal, which was done using a Butterworth low pass filter. This was implemented using the code found in Appendix section 5.1.2.

The parameters of the filters were experimented to find those most suitable to transform SAFER data again into samples reminiscing those of Physionet 2017 Dataset. Figure 9 shows the resulting samples for different sampling frequencies. These cutoff sampling frequencies are in terms of the discrete time-steps. Therefore, note that frequencies have to be re-scaled in terms of the sampling frequencies of the Zenicor device to return values in Hz.

It was decided that a low pass filter cutoff frequency of 1.58 was best suited for the preprocessing.

By scale normalising the sample, using the code found in Appendix 5.1.3, the signal was transformed to ensure it was bound by signal strength ± 1. The resulting signal can be seen over different timescales in Figure 10 superimposed on a typical sample from the PhysioNet challenge 2017. Clearly these samples are largely very similar in terms of the scale of features, detail of signal, and sharpness and noise levels.

The final steps to prepocess the data from here was to downsample it from the 500Hz Zenicord device sampling frequency to the 300Hz that the model is expecting, using the `numpy.interp` function. Then, the samples were zero padded to ensure dimensional consistency. The resulting sample from this process is now ready for inference of AF probability using the NNN model.

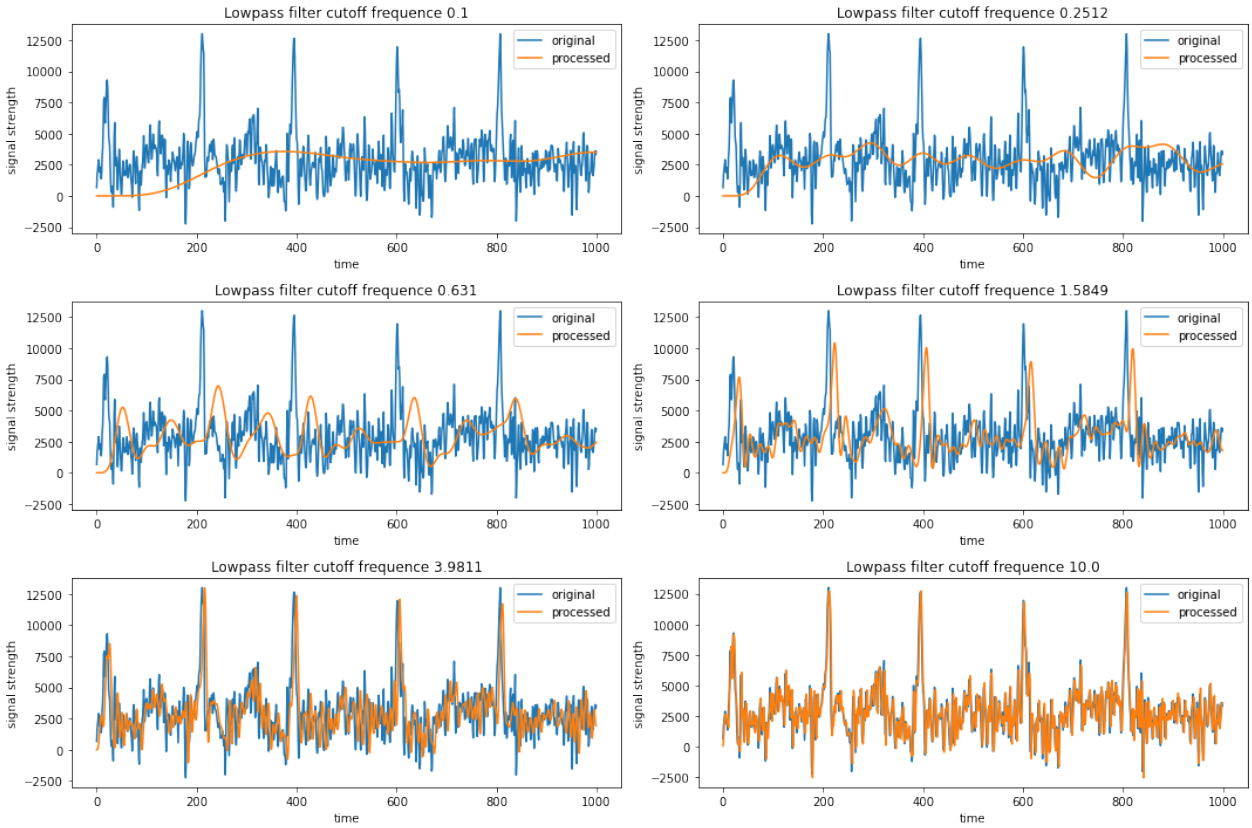


Figure 9: Resulting samples after removal of baseline wander and low pass filtered at various frequencies

4.2.1 Tested on SAFER Feasibility Study 1 Data

After preprocessing the SAFER Feasibility Study 1 dataset, the NNN was used to predict the probabilities of AF appearing in the sample. Figure 11 shows the probabilities of AF assigned for each sample in the dataset, with the samples being divided into groups corresponding to the label the Cardiologists gave them. These colour coded sections also have the mean values $\pm$ 1 standard deviation plotted to assist with analysis.

The first promising result is that the predicted probabilities of AF is highest for the samples which do contain AF, according to Cardiologist. This is seen by the mean of this section being highest, see red section, at around 0.35 which is significantly higher than those from other classes. Unfortunately, however, the predicted probabilities of AF for some samples in this section were very low, with a clear group around 0.05. This indicates that there is a high likelihood this model will under-predict the probability of an AF sample containing AF.

Another encouraging result is that for Normal samples, as labelled by the Cardiologist, the NNN predicted likelihood of AF is very low. The mean values, see the green section, of around 0.05 was clearly lower than all other classes. The plot shows that only a handful of these samples had a high AF predicted likelihood, meaning that a prioritisation method based off these probabilities was unlikely to put Normal samples first. Furthermore, the

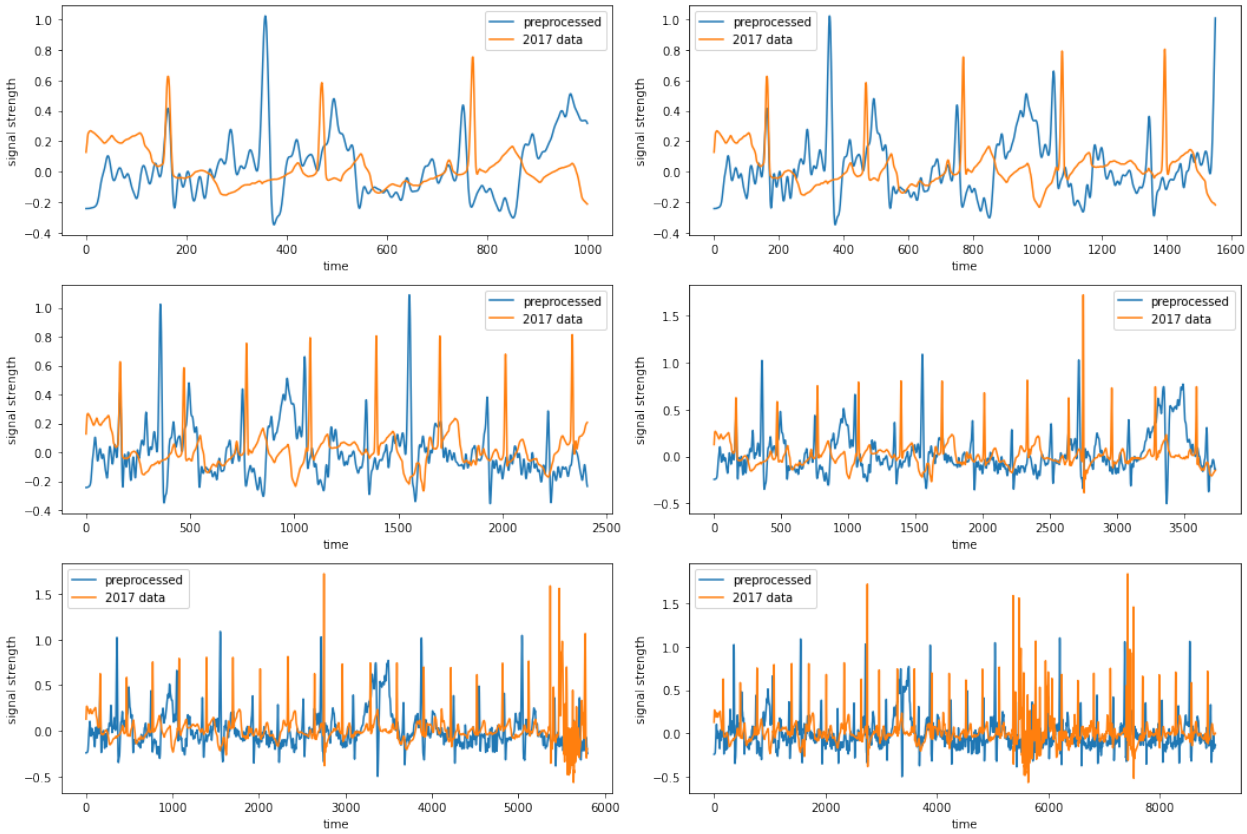


Figure 10: Processed SAFER sample comparison to sample from Physionet 2017 dataset, over different scales

mean probability predictions of AF for all other classes but normal are significantly higher than the mean value for Normal, which means that a sample which maybe contains AF, or contains another heart arrhythmia, is likely to be prioritised over a sample which contains no pathology. These results lead to an alternative direction to approach the problem, by including the NNN predicted probabilities of Other heart arrhythmia in the results.

Figure 12 shows the resulting NNN probability predictions of Other for the same samples as in Figure 11. This plot rules out the use of Other probabilities being included in the probability assigned to a sample for containing AF, with a weighted sum of the two values being used, because a Normal sample would have a higher increase in probability than an AF sample. This would be counter productive.

4.2.2 Tested on SAFER Feasibility Study 2 Data

4.2.3 Tested on SAFER Trail

4.3 RR dRR Approach Testing

4.3.1 Tested on Physionet 2017 Data

4.3.2 Tested on SAFER Feas Data

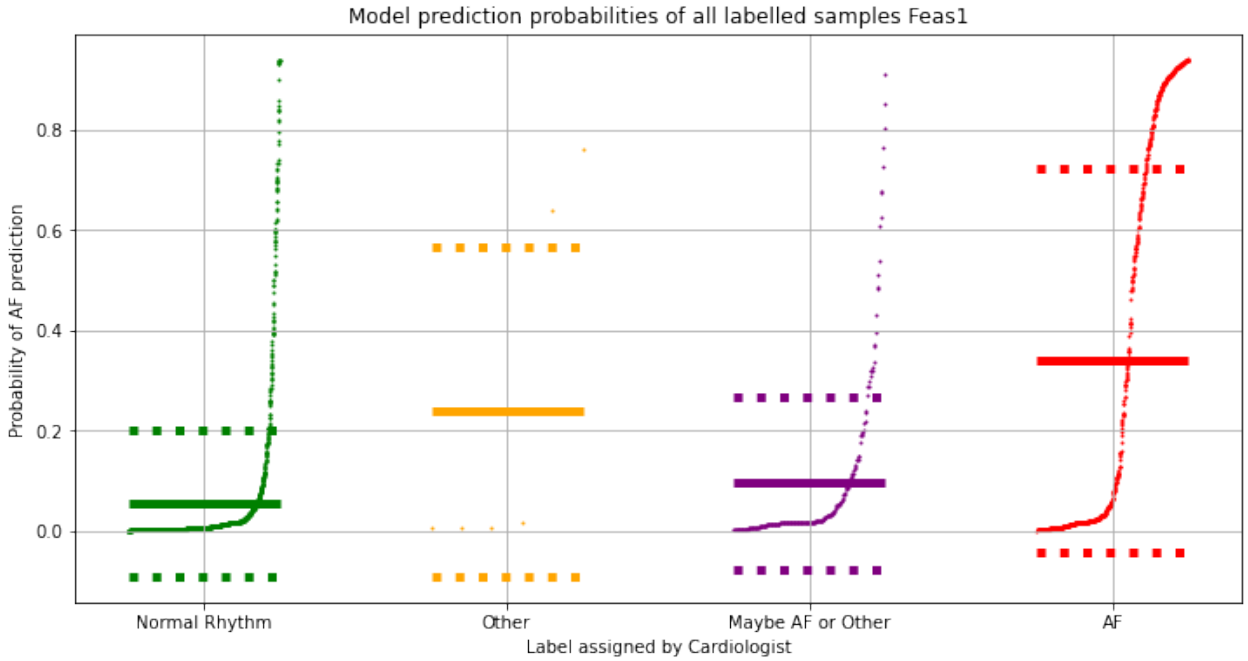


Figure 11: Probability of AF predicted by model for all labelled data in SAFER Feasibility study 1, separated by label assigned by cardiologist, with mean ± standard deviation plotted.

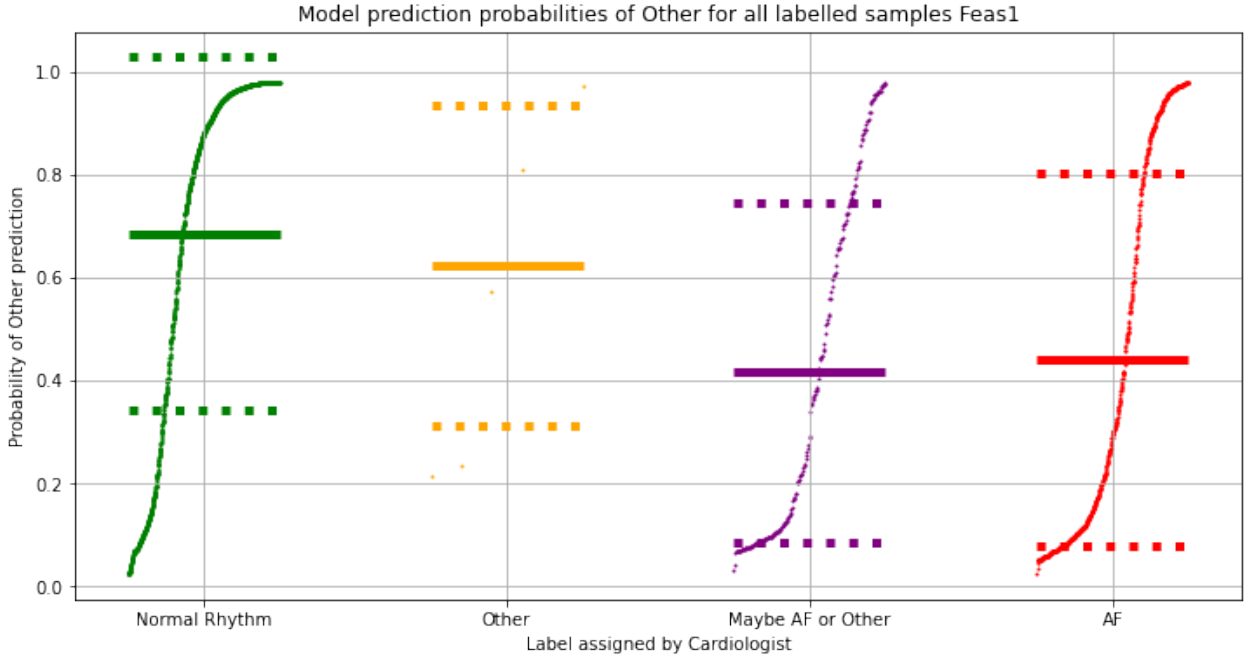


Figure 12: Probability of Other heart arrhythmia predicted by model for all labelled data in SAFER Feasibility study 1, separated by label assigned by cardiologist, with mean ± standard deviation plotted.

5 Conclusion

References

[1] D. Clifford Gari, Liu Chengyu, Moody Benjamin, Lehman Li-wei, Silva Ikaro, Johnson Alistair, and Mark Roger. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017.

[2] Alday Erick Andres Perez, Gu Annie, Shah Amit, Liu Chengyu, Sharma Ashish, Seyedi Salman, Rad Ali Bahrami, Reyna Matthew, and Clifford Gari D. Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020.

[3] George Moody and Roger Mark. Mit-bih atrial fibrillation database, 11 2000.

[4] NHS. Complications - atrial fibrillation, 5 2021.

[5] Sara Schukraft, Marco Mancinetti, Daniel Hayoz, Yannick Faucherre, Stéphane Cook, Diego Arroyo, and Serban Puricel. Handheld ecg tracking of in-hospital atrial fibrillation the hecto-af trial clinical study protocol. *Trials*, 20:92, 12 2019.

[6] Isabelle C. Van Gelder, Jeff S. Healey, Harry J.G.M. Crijns, Jia Wang, Stefan H. Hohnloser, Michael R. Gold, Alessandro Capucci, Chu-Pak Lau, Carlos A. Morillo, Anne H. Hobbelt, Michiel Rienstra, and Stuart J. Connolly. Duration of device-detected subclinical atrial fibrillation and occurrence of stroke in assert. *European Heart Journal*, 38:1339–1344, 5 2017.

[7] SAFER. The safer trial screening for atrial fibrillation with ecg to reduce stroke.

[8] 12-lead ecg placement guide with illustrations.

[9] Emma Svennberg, Martin Stridh, Johan Engdahl, Faris Al-Khalili, Leif Friberg, Viveka Frykman, and Mårten Rosenqvist. Safe automatic one-lead electrocardiogram analysis in screening for atrial fibrillation. *Europace*, 19:1449–1453, 2017.

[10] Physionet.

[11] Jie Lian, Lian Wang, and Dirk Muessig. A simple method to detect atrial fibrillation using rr intervals. *American Journal of Cardiology*, 107:1494–1497, 5 2011.

[12] Shenda Hong, Meng Wu, Yuxi Zhou, Qingyun Wang, Junyuan Shang, Hongyan Li, and Junqing Xie. Encase: An ensemble classifier for ecg classification using expert features and deep neural networks. volume 44, pages 1–4. IEEE Computer Society, 2017.

[13] Zhaohan Xiong, Martin K. Stiles, and Jichao Zhao. Robust ecg signal classification for detection of atrial fibrillation using a novel neural network. volume 44, pages 1–4. IEEE Computer Society, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 12 2015.

[15] Julien Oster, Jemma C. Hopewell, Klemen Ziberna, Rohan Wijesurendra, Christian F. Camm, Barbara Casadei, and Lionel Tarassenko. Identification of patients with atrial fibrillation: A big data exploratory analysis of the uk biobank. *Physiological Measurement*, 41, 2020.

[16] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8:1368–1373, 8 2018.

[17] Mit-bih arrhythmia database.

[18] Anthony. Wikipedia.

[19] Paul Eilers and Hans Boelens. Baseline correction with asymmetric least squares smoothing. *Unpubl. Manuscr*, 11 2005.

Appendix

5.1 Python Code

5.1.1 Asymetric least squares smoothing

Code taken from https://stackoverflow.com/questions/29156532/python-baseline-correction-library

```
from scipy.sparse import csc_matrix, spdiags
from scipy.sparse.linalg import spsolve
import numpy as np

def baseline_als(y, lam=1e6, p=0.01, niter=10):
    L = len(y)
    D = csc_matrix(np.diff(np.eye(L), 2))
    w = np.ones(L)
    for i in np.arange(niter):
        W = spdiags(w, 0, L, L)
        Z = W + lam * D.dot(D.transpose())
        z = spsolve(Z, w*y)
        w = p * (y > z) + (1-p) * (y < z)
    return z
```

Listing 1: Asymetric least squares smoothing code

5.1.2 Butterworth low pass filter

Code taken from https://stackoverflow.com/questions/25191620/creating-lowpass-filter-in-scipy-understanding-methods-and-units

```
from scipy.signal import butter, lfilter, freqz

def butter_lowpass(cutoff, fs, order=5):
    return butter(order, cutoff, fs=fs, btype='low', analog=False)

def butter_lowpass_filter(data, cutoff, fs, order=5):
    b, a = butter_lowpass(cutoff, fs, order=order)
    y = lfilter(b, a, data)
    return y
```

Listing 2: Butterworth LPF code

5.1.3 Scale normalisation

This is the code used after removal of baseline wander and LPF, which results in the `signal_smoothed`.

```
signal_offset = signal_smoothed - np.min(signal_smoothed)
signal_squashed = 2 * signal_offset / np.max(signal_offset)
signal_normalised = signal_squashed - np.mean(signal_squashed)
```

Listing 3: Scale normalisation code