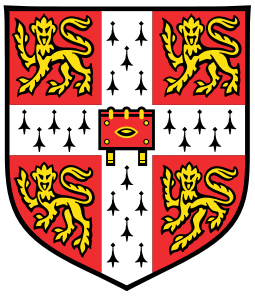
Data of Your Heart: Screening for Atrial Fibrillation



Jordan Smith

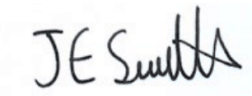
Supervisor: Dr Elena Punskaya

Division F Engineering Department
Christ's College, University of Cambridge

1 June 2022

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Signed:



Abstract

Data of Your Heart: Screening for Atrial Fibrillation
Jordan Smith, Christ's College
1 June 2022

Motive - Atrial Fibrillation (AF) is the most common heart arrhythmia, with an 11.4% prevalence in those over 65s [1], and causes a five-fold increase in stroke risk [2]. The stroke risk for AF patients can be reduced by anticoagulation, but this requires an AF diagnosis first. Currently, symptomatic AF patients are referred to 12 Lead Electrocardiogram (ECG) monitoring of the heart, taken in the hospital, if referred by the GP. However, asymptomatic AF patients are still at increased risk of stroke, but these patients go largely undiagnosed. AF screening offers a method to diagnose low AF burden patients otherwise missed by current practices, through the use of self-administered ECGs using a 1 lead ECG monitor such as the Zenicor device. The SAFER study [3] is looking into the feasibility of this method of screening, with patients taking four 30 second samples a day for a screening period of 3 weeks.

This generates a large number of ECG samples, with 84 samples per participant. However, for an AF diagnosis to be made, only 1 sample showing AF needs to be found and reviewed by a Cardiologist.

This project offers a prioritisation method for increasing the likelihood that the first samples seen by the Cardiologist show AF signs, where possible, so an AF diagnosis can be made with fewer manual reviews. With diagnosis still being made by the Cardiologist reviewing ECG samples, this method only aids the process and increases the efficiency in a safe way, increasing the scalability of the screening program.

Method - Multiple methods for classifying ECGs were analysed, some being traditional Machine Learning (ML) methods using features extracted from ECGs, others being data-driven Deep Learning (DL) approaches, and a few being a combination of the two. The method outlined in the paper "Robust ECG signal classification for detection of atrial fibrillation using a novel neural network" [4], which was proposed in the PhysioNet Challenge 2017 [5], was chosen due to its proven ability at classifying ECGs as AF or not, with it being a runner up in the competition. A variant of the famous ResNet [6] deep neural network model, this purely data-driven approach learned the necessary features for AF ECG classification without the detection of expert features needed in other traditional approaches.

This model was adapted from its original application of classifying 1 lead ECGs, taken using the AliveCor device, as either "AF", "Normal", "Other" heart arrhythmia, or "Noisy. It was used to infer a probability of a sample, from the

SAFER data, showing signs of AF. This enables prioritisation of samples for review. Experiments were carried out on the necessary model adaptations, the preprocessing of SAFER data, and inference processes followed by the utilisation of inferred class probabilities to produce a viable system to classify ECGs from the SAFER study for manual review.

Two different applications of the resulting methods looked at both the ordering of ECG samples on a per participant basis and on an entire dataset basis, leading to different opportunities for streamlining the screening review process.

Results - The resulting method lead to an estimated 91% reduction in reviews needed for each participant with AF when the model was applied on a per-participant basis. With 11.4% of over 65s having AF, this estimated reduction in the workload of the entire process is 10%. For participants without AF, no reduction in workload can be achieved when prioritisation of samples is taken on a per-participant basis.

However, when the prioritisation of ECGs is done over the entire dataset, it was observed that over two-thirds of the study's participants who were diagnosed with AF can be found with only 2 reviews per diagnosis needed. With each ECG review taking 20 seconds [7], this can be extended to say that every 40 seconds an AF diagnosis could be made, or 90 per hour of a Cardiologists time. This is a significant increase in the efficiency of the review process, and dramatically increases the scalability of an AF screening process.

Deliverable - This project takes an in-depth analysis of the classification methods available for ECG analysis and offers a viable option for the prioritisation of ECGs for screening review. Complete with a description of all preprocessing methods, interpretation of the model's outputs, and utilisation of the probabilities to most effectively reduce the workload for Cardiologists, it also offers direction for future work on the method.

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 [2]. 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 the 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 [10], showing Atrial Fibrillation (AF) (a) and Normal rhythm (b)

Screening offers a method to identify these asymptomatic, low AF burden patients who are otherwise neglected by current methods, but are 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 [11], meaning the risk of AF continuously increases and should be monitored. The SAFER study [3] 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 a hospital is generated with trained professionals using sophisticated equipment, whereas a 1 lead ECG will be self-administered. The resulting samples generated are susceptible to contaminated contact points, movement of the patient and external signals 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 heart's 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 occurring.

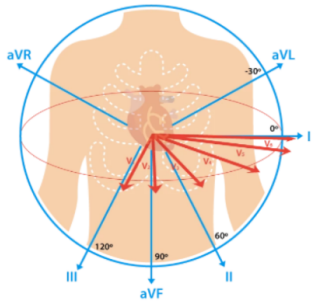


Figure 2: 12 lead ECG signal locations [12]

Another consideration for this project is the presence of other heart arrhythmias in the data, such as Bradycardia, Tachycardia and Heart Block, which cannot be confused with AF. Some heart arrhythmia such as Atrial Flutter can be easily confused with AF, see Figure 3. Furthermore, some other heart arrhythmia can be present in a sample at the same time as AF being present, and the model would need to be able to detect AF even in these cases. Careful preprocessing is needed to avoid the loss of essential information for these distinctions.

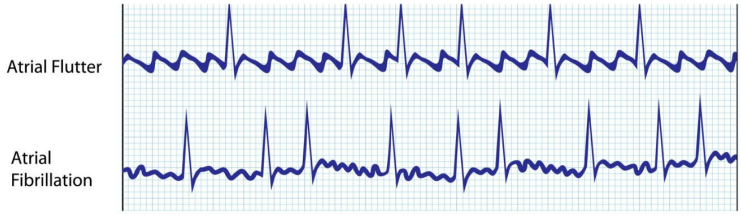


Figure 3: Atrial Fibrillation vs Atrial Flutter, sourced from Dr AFib [13]

1.3 The Objective

With the vast amounts of data the Cardiologists will need to analyse with the SAFER trial, a more efficient review process is needed. 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 for a participant in the study who has a low AF burden, therefore at least one of the samples from this participant will show AF, and significant time and resources are saved if one of these samples showing AF is the first one seen by the Cardiologist. An AF diagnosis is made then, with all other samples not needing to be reviewed, saving the Cardiologist time and therefore costs. Automated algorithms can be used to identify abnormal ECGs to be reviewed manually [7], but these identify around 35 samples for each sample showing any pathology.

The project aim is to find a model which generates a representative probability of an ECG sample exhibiting AF, to order samples by, therefore increasing the likelihood that the AF samples appear first. This is done by distinguishing samples between 2 classes, AF and Non-AF. All other heart arrhythmias are bunched into this Non-AF class.

A further extension is to threshold the samples by this probability and omits low AF probability samples from analysis by the cardiologist.

2 Literature Review

2.1 Background

A highly active area of research, classifying ECGs for AF or Sinus (normal) Rhythm has seen numerous academic competitions, with the largest being the PhysioNet Challenges. Through these competitions, many approaches to this classification have been developed and implemented for use on 1 lead ECG data. Traditional approaches involve more handcrafted 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" [14], which 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 a given time length. This method was tested on the MIT-BIH Atrial Fibrillation database [9] 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 detection of 'extra peaks' in an ECG taken with the Zenicor device reduces 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. DL methods increasingly display the capability to learn the waveform shapes, both small and large scale features that correspond to AF or Sinus Rhythm (SR), and in some proposals the ability to distinguish between AF and other heart arrhythmias, 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 took aim at the classification of noisier Lead 1 ECG samples. The data was 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 others 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, the ENCASE method [15] and the Novel Neural Network (NNN) method [4]. The following 3 papers state these methods, and the final one is an exploration of many different approaches.

1. "ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks" [15]. This method combined the use of traditional feature detection from Statistics, Signals Processing and Medicine, with modern DL methods that learn features through a data-centric approach. This proposed method had a high accuracy with an F1 score of 0.83. This method was not pursued in the project, however, due to most of the methodology being omitted from the report submission. Therefore applying this method from scratch would involve extensive research into the expert features from the Medicine, Statistical and Signals processing fields in order to implement it effectively.

2. "Robust ECG Signal Classification for Detection of Atrial Fibrillation Using a Novel Neural Network" [4], applied a more DL centred approach, with a novel form of the famous ResNet [6] 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 [16], 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 approach. 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 use of expert features assumed beyond the scope of this project.

After analysing these methods, the NNN approach was the one chosen to continue with. Based on the state of the art ResNet model, with proven effectiveness in image classification, the NNN had high accuracy in the PhysioNet 2017 challenge which proved it has been successfully adapted for use with time series, sequence, data. Furthermore, the ease of implementation, with the pre-trained model being provided, led to quicker testing and experiments with the model. Finally, being a DL method, the NNN learned the necessary features from the ECG without being defined in code, therefore implementation required less research and experimentation than methods based on expert features.

The project aim of prioritising ECGs is such that a model which is less accurate, but easier to implement, is favourable. Incorrect classification is low risk due to diagnosis still being made by the Cardiologist, and as such the NNN method which has very high accuracy, although slightly lower than methods based on expert features, is appropriate and was chosen.

3 Methodology

3.1 The Data

The data used for this project came from the Physionet [17] 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 method's effectiveness for application in the SAFER study. The Physionet datasets are open source, and readily available online. The SAFER datasets are used for testing the models, with the evaluation of performance being the most important for this data.

3.1.1 PhysioNet Challenge 2017 Dataset [5]

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" corresponds to AF.
2. "N" corresponds to Normal (Sinus) Rhythm.
3. "O" corresponds to Other heard arrhythmia such as heart block.
4. "∼" 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 [8]

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 [18] was used in this project. It contained 6877 samples, each from 6 to 60 seconds long and taken with a 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 the waveform detected. The Lead 1 data can be isolated from the other 11 lead data, and is still useful for a general comparison of methods, with due consideration.

3.1.3 MIT-BIH Dataset [9]

This dataset contains "8 half-hour excerpts of two-channel ambulatory ECG recordings" [19], generated using a Holter device. A sample from a Holter device produces has different characteristics from a sample from a 1 lead handheld ECG device due to the different locations of the electrodes on the body. ECGs from a handheld device are on one thumb from each hand, whereas for a Holter device the electrodes are across the chest, meaning the signals that the device detects are from different stages in the cardiac cycle. This dataset was mainly used for research into the RR dRR method.

3.1.4 SAFER Datasets

The key datasets for method evaluation were the SAFER Feasibility Study (ISRCTN 16939438) datasets. The assessment of the feasibility of delivering an AF screening program in Primary Care, approved by the London - Central Research Ethics Committee (REC ref: 18/LO/2066), was the motivation for this dataset to be created. 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 4.

Figure 4: The Zenicor EKG-2 device which acquires 1 lead ECGs

All SAFER datasets were generated using this device, with 3 different datasets being created at different times during the study using slightly different methods. The datasets are divided as follows, with the statistics of each seen in Table 1.

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

The combination of all these datasets contains 14,235 samples that have been individually reviewed by cardiologists and over 300,000 samples in which the participant themselves has been diagnosed as having AF or not. An important distinction is made between labels applied to a participant and a label provided for an individual sample. The label applied to the participant, which is fundamentally a diagnosis of the patient, with either "AF" or not being of interest to this project, could be incorrectly extended as a label for all samples from this participant. However, this would lead to contradictory results, or unsuccessful training of any model using these extrapolated labels because many of the individual samples from

a participant which is ground truth labelled as 'AF' may not actually display any signs of AF, despite those individual samples being from a patient with low AF burden and therefore not all the samples from this patient will show AF signs. Therefore when using SAFER in this project, only samples which were individually reviewed by a cardiologist were used for generating results and analysis.

The below plots show the proportions of each participant's samples which were individually reviewed, for all the participants that had at least 1 of their samples individually reviewed by the Cardiologist. Clearly, Feasibility study 1 had a much higher proportion of each participant's samples which were individually reviewed, with a mean proportion of over 0.3 whereas for Feasibility study 2 the value was near 0.1. This shows this dataset to be more useful for evaluation of the model's performance at prioritising ECGs for manual review, to decrease the number of samples which have to be individually reviewed, because the performance metric, defined in Section 3.4.2, calculates time savings on a per participant basis and takes the average overall participants which can be used for this calculation. Clearly, the more samples per patient that are manually reviewed, the more representative this metric calculation will be.

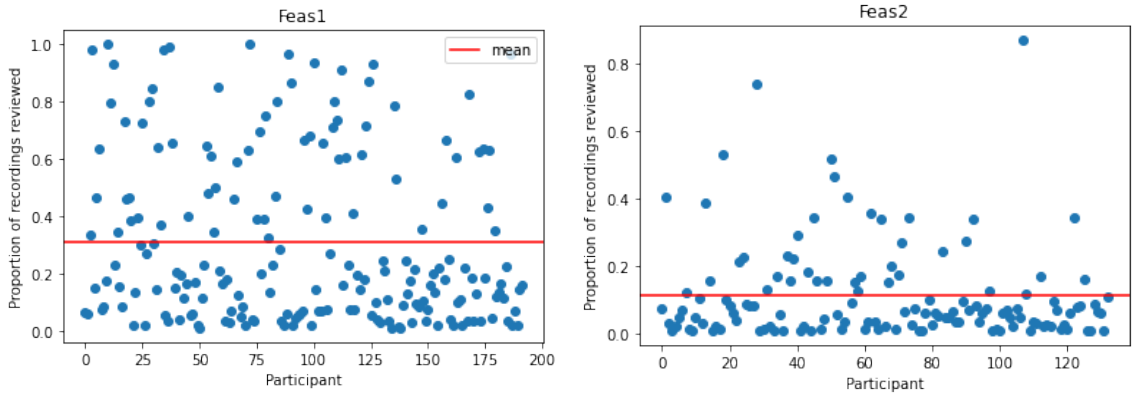


Figure 5: Proportions of recordings individually reviewed per participant in SAFER Feasibility studies 1 and 2

Another consideration for these datasets is the skew towards 'Non-AF' samples. For the SAFER datasets, less than 20% of samples labelled "AF", which is less severe than the skew in the PhysioNet Challenge 2017 dataset with only 9% but has to be considered when gathering results. If a model was to be trained on this dataset, there are only 153 samples which were labelled as "AF" in the combined Feasibility Study datasets, which is insufficient for training a Deep Neural Network for recognition of these samples. Augmentation and other techniques would have to be used to increase the training examples, and likely the model would not extend in performance well to new data. The trail dataset, however, with 1416 samples labelled as "AF" would be a lot more suited for the task of training a DL method. However, this dataset was not made available to this project in time for this experiment to be carried out.

Finally, the presence of samples showing other heart arrhythmias, such as those mentioned in Section 1.2, is another consideration in these datasets. With many arrhythmias often present in the general population, it is very important for the model to be able to distinguish

between these and AF, and as such the careful consideration of how to choose the model and which classifications it should recognise was 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

3.1.5 SAFER Feasibility 1 Study Dataset

The most appropriate dataset for use when generating results of the NNN effectiveness at prioritising ECGs based on their AF likelihood, the SAFER Feasibility study 1 dataset has a high proportion of samples individually annotated by the cardiologist for each participant. This dataset was the first to be generated in the SAFER trial and saw participants take 4 samples using the Zenicor device per day, over a period of 1, 2 or 4 weeks. This study went on up until the COVID-19 pandemic started to affect the Study, and hereafter the approach to data collection changed.

The process of getting the labels for the dataset started with the samples taken by the Zenicor device. These samples were then collected and the entire dataset of collected samples was run through an algorithm provided by Zenicor which identified "abnormal" ECGs. These "abnormal" ECGs were then analysed quickly by a nurse before deciding whether to be passed on to a Cardiologist. The samples that then made it to the Cardiologist were reviewed, with options of classes labelled as in Table 2.

This trial included 2141 participants, with 190 of these participants being reviewed by at least 1 Cardiology team, and 65 of these being diagnosed with AF.

From these 2141 participants, a total of 162515 samples were taken with 4494 of these samples being individually reviewed by at least 1 Cardiology team. A median of 61 samples was taken per patient, with an upper quartile of 111 and a lower quartile of 53. This led to 137 samples labelled as AF, 1418 labelled as No pathology, or Normal, and 31 labelled with Other heart arrhythmias. Other heart arrhythmia included "Heart Block" and "Ventricular

Diagnosis Type	Diagnosis Code
Heart Block (HB)	1
Atrial Fibrillation (AF)	2
Cannot exclude pathology	3
No pathology	4
Screening failure	5
Undecided	6
Ventricular Tachycardia (VT)	7
Disagreement	-1

Table 2: Table of label codes chosen by Cardiologist from SAFER study

Tachycardia" as seen in Table 2. Also, the Diagnosis type "Cannot exclude pathology", where pathology means AF, VT or HB, was labelled as "Maybe AF" in all subsequent results.

97.2% of the samples in this dataset were not reviewed by a Cardiology team, with the label "Undecided" provided for these samples. Because these 157965 that were not reviewed provided no meaningful information to the project, they were ignored when generating results, see Sections 4.2 and 4.3. Furthermore, the samples labelled as "Screening Failure" and "Disagreement" were also ignored because they do not contribute meaningfully to the results.

3.2 The Novel Neural Network (NNN) Approach

Two approaches were tested in-depth. Predominantly, the DL approach using the "Novel Neural Network" [4] was examined. This approach was proposed during the PhysioNet 2017 challenge. Furthermore, a more simplistic approach using "RR dRR intervals" [14] was investigated in order to trial a low computational expense method and compare the results to the NNN approach. However, it did not prove useful and therefore is not explained in this report.

Starting with the NNN model, which was developed for the PhysioNet challenge 2017 dataset which aimed at classifying 1 lead ECG sample 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 to learn to classify between the four target classes "AF", "Normal", "Other" heart arrhythmia or "Noisy".

The model's ability to distinguish between "AF" and "Other" is very important because the SAFER study is only screening for AF, not any other heart arrhythmia.

3.2.1 Network Architecture

The neural network developed by Xiong et.al. [4] can be broken down into 16 convolutional blocks, each with skip connections from previous blocks and each block being down-sampled from the output of the previous layer. The output of the ith convolutional layer, before addition of the skip connection, is convolution of the 15×1 kernel $\mathbf{w}^{(i)}$ with trainable parameters $\{w_k^{(i)}\}_{k=1}^{15}$ with the input to that layer $y^{(i-1)}$.

$$\begin{aligned} y_j^{(i)} &= \mathbf{w}^{(i)} \cdot [y_{j-7}^{(i-1)}, ..., y_j^{(i-1)}, ..., y_{j+7}^{(i-1)}]^T \\ &= w_1^{(i)} y_{j-7}^{(i-1)} + ... + w_8^{(i)} y_j^{(i-1)} + ... + w_1 5^{(i)} y_{j+7}^{(i-1)} \\ &= \sum_{k=1}^{15} w_k^{(i)} \cdot y_{j+k-7}^{(i-1)} \end{aligned} \tag{1}$$

Which for the layer i, which has a J dimensional output vector, will be of the form stated below. Zero padding is used on the previous output layer, $\mathbf{y}^{(i-1)}$ to enable the kernel to be centred on each point from this previous layer.

$$\mathbf{y}^{(i)} = \begin{bmatrix} \sum_{k=1}^{15} w_k^{(i)} \cdot y_{1+k-7}^{(i-1)} \\ \sum_{k=1}^{15} w_k^{(i)} \cdot y_{2+k-7}^{(i-1)} \\ ... \\ \sum_{k=1}^{15} w_k^{(i)} \cdot y_{J+k-7}^{(i-1)} \end{bmatrix} \tag{2}$$

The final convolutional block is the input to a fully connected layer which feeds into a softmax followed by classification. The diagram of model architecture is seen in Figure 6, which shows the convolutional blocks contain the following components:

1. Input from the 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

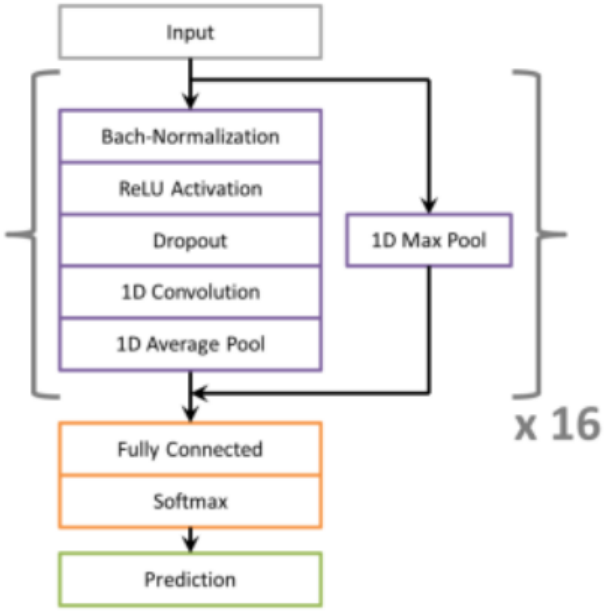


Figure 6: Novel Neural Network architecture with skip connections [4]

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 backpropagation 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 reduces this effect [6] by ensuring non-zero gradients are calculated, especially in conjunction with batch normalisation. This can be explained by examining the formula for the output of a neuron with a skip connection, which is a simple addition of the output of the convolutional layer with the max pooled output from the previous layer, $\mathbf{y}_{mp}^{(i-1)}$.

$$\mathbf{y}^{(i)} = \begin{bmatrix} \sum_{k=1}^{15} w_k^{(i)} \cdot y_{1+k-7}^{(i-1)} \\ \sum_{k=1}^{15} w_k^{(i)} \cdot y_{2+k-7}^{(i-1)} \\ \cdots \\ \sum_{k=1}^{15} w_k^{(i)} \cdot y_{J+k-7}^{(i-1)} \end{bmatrix} + \mathbf{y}_{mp}^{(i-1)} \tag{3}$$

This will lead to the component of the backpropagation algorithm, when differentiating the Loss function L by the input to the network $\mathbf{x}$, which is always non-zero, as seen by taking the differential of the output of this layer with respect to the input to this layer.

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{x}} &= \quad ... \quad \cdot \quad \frac{\partial \mathbf{y}^{(i)}}{\partial \mathbf{y}^{(i-1)}} \quad \cdot \quad ... \\ &= \quad ... \quad \cdot \quad \left(\frac{\partial}{\partial \mathbf{y}^{(i-1)}} \left(\begin{bmatrix} \sum_{k=1}^{15} w_k^{(i)} \cdot y_{1+k-7}^{(i-1)} \\ \sum_{k=1}^{15} w_k^{(i)} \cdot y_{2+k-7}^{(i-1)} \\ ... \\ \sum_{k=1}^{15} w_k^{(i)} \cdot y_{J+k-7}^{(i-1)} \end{bmatrix} \right) + \mathbf{1} \right) \quad \cdot \quad ...\end{aligned} \tag{4}$$

Clearly, the $+\mathbf{1}$ will ensure non-zero gradients are calculated, and therefore the vanishing gradients problem is significantly reduced.

Batch normalisation normalises the outputs from each layer 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 of these outputs. 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 model's ability to fit the data by enabling non-linear decision boundaries, with the increased efficiency that comes with this simple activation function.

Pooling was used throughout this model. This developed the model's 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 the highest predicted probability.

3.2.2 Training Data and Resulting Parameters

This model was trained on the data from the PhysioNet 2017 Challenge by the proponents of the method, as described in Section 3.2.1. This dataset, as mentioned in Section 3.1.1, has features similar to that of the SAFER dataset, and as such the resulting model parameters were assumed to be appropriate for application to SAFER data, following the appropriate tests.

For training and drawing inference from new data the samples are initially split into 5-second sections, each being fed through the model and subsequent classification prediction made. The mean value of probabilities of each class over these different section classifications is referred to, and the class with the highest probability is the predicted class the model will choose.

Although this method was initially applied to a classification problem, the model has been adapted to produce a probability, rather than classification, of AF. This was achieved through

examining the output of the softmax layer, with the weighted average of these softmax layers over each section taken as this probability.

Furthermore, the original NNN model was trained to classify between the classes stated in Section 3.1.1. However, this project was not concerned with any class but 'AF'.

3.2.3 Preprocessing of SAFER data for NNN

For the application of the NNN method to SAFER data, preprocessing was needed to ensure the signals looked as similar as possible to the data the model had seen before and been trained on in the PhysioNet 2017 challenge, obtained using the AliveCor device. Various experimentation was carried out looking at 5 key components of the signal that needed processing:

1. Sample frequency consistency
2. Baseline wander
3. High-frequency noise
4. Scale normalisation
5. Sample length consistency

Firstly, the signals were down-sampled from the 500Hz Zenicor device sampling frequency to the 300Hz that the model is expecting from the AliveCor device, using the `numpy.interp` function.

Next, as an alternative to high pass filtering, to remove the baseline wander low-frequency components, Asymmetric Least Squares smoothing [20] was used, which was implemented using the code found in the Appendix section 8.3.3. This resulted in the signal transformation seen in Figure 7. This was found to be more suitable for use with the NNN but did use more computation than a simple bandpass filter, and as such if applications do not allow this expense, this method should be replaced.

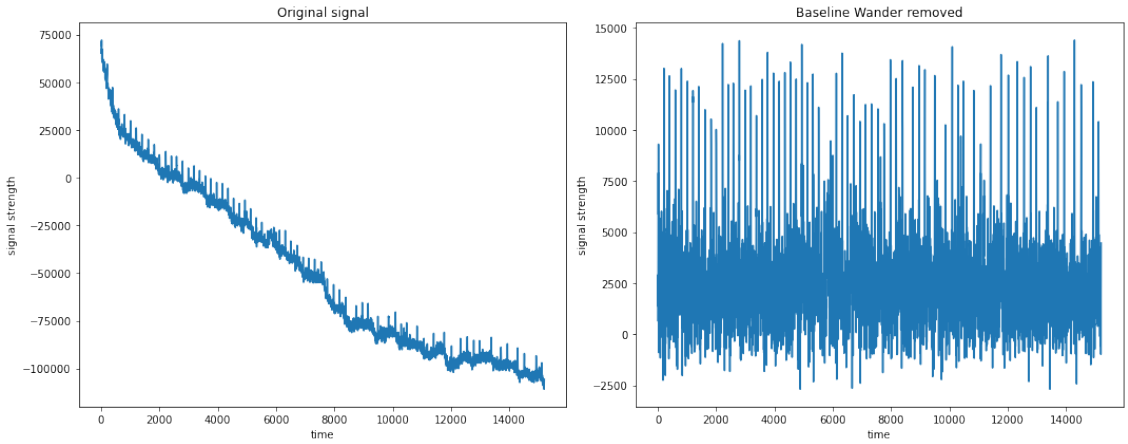


Figure 7: Removing the baseline wander initially present in SAFER data

The next step was to reduce the high-frequency noise components of the signal, which was done using a Butterworth low pass filter. A Butterworth filter was chosen over a Chebyshev filter due to its flat passband, and chosen over a Bessel filter due to its steeper cutoff of frequencies outside of the band, see Appendix section 8.2. The Butterworth filter was implemented using the code found in the Appendix section 8.3.2.

The parameters of the filters were experimented with to find those most suitable to transform SAFER data again into samples reminiscing those of PhysioNet 2017 Dataset. Taking into account that the AliveCor device used to generate the data for the PhysioNet 2017 challenge used bandpass filtering to eliminate frequency components outside of the band 1-40Hz, the low pass cutoff frequency was experimented with to find the most appropriate values to reform the samples for easiest reading by the NNN, around these values. It was important to remove the high-frequency noise without reducing the QRS complex amplitude, which was difficult considering the significant contribution that high-frequency components of the signal make to the QRS complex waveform, being a waveform with higher frequency components than the P or T wave. A low pass filter cutoff frequency of 15.8 Hz was used.

By scale normalising the sample, using the code found in Appendix 8.3.1, the signal was transformed to ensure it was bound by signal strength ± 1. The resulting signal can be seen over different timescales in Figure 8 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.

Finally, 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.

3.3 Research Timeline

3.3.1 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 trial data. An understanding of the dataset, its limitations and its properties were 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.

3.3.2 NNN Approach Experiments on PhysioNet Data

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 wander and corresponding bandpass filtering and sample lengths.

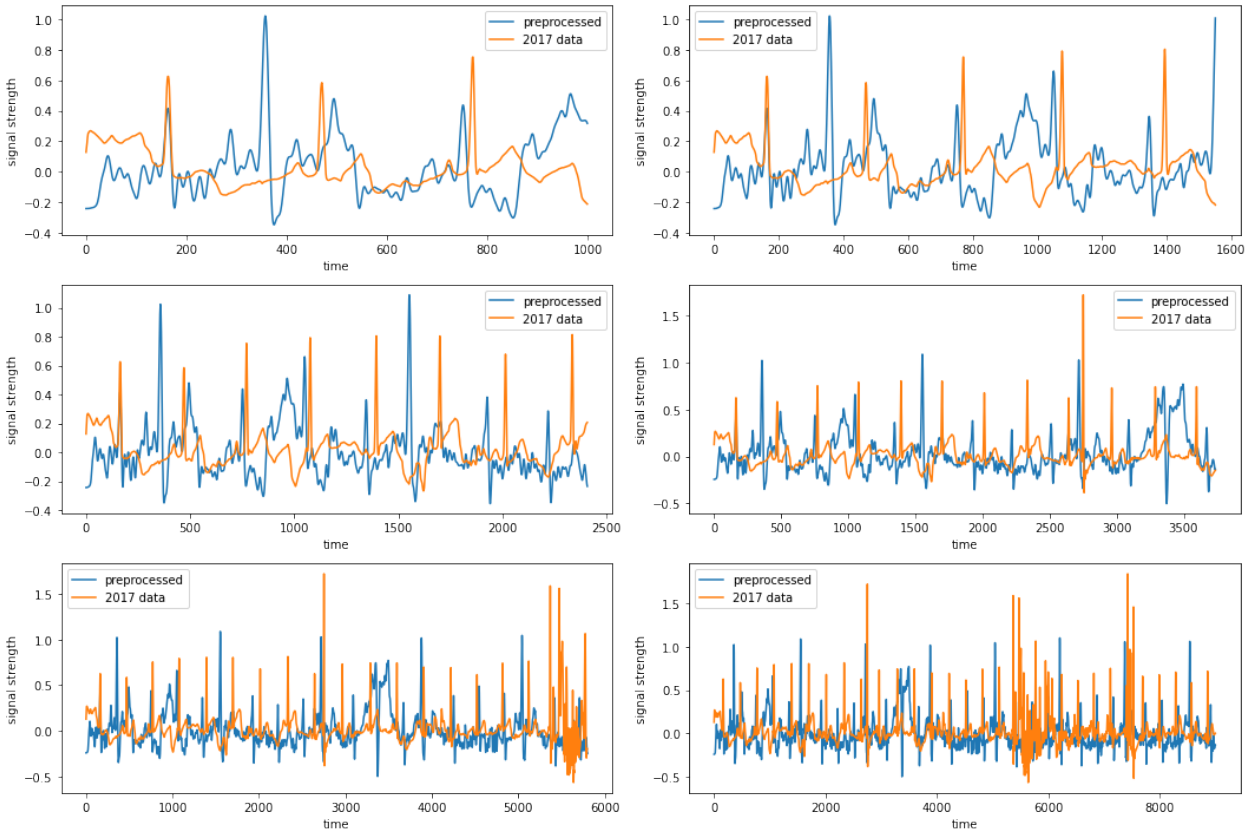


Figure 8: Processed SAFER sample comparison to sample from PhysioNet 2017 dataset, over different scales

Furthermore, the tooling was built for working with the predictions made by the model and analysing these results for assessment of the model's 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 model's 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 model use for the project aims.

3.3.3 RR dRR Approach Experiments

In order to evaluate the effectiveness of the NNN method, a 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 were experimented with, as well as the cost function used for the optimisation of the threshold used as the decision boundary, one based on the F1 score of the output to enable the 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 prevented extensive analysis of this data.

3.3.4 NNN Approach Preprocessing Experiments

Before applying NNN model to the raw SAFER data, preprocessing was developed to reform the data into a form that the model could recognise and work with, with the methods used being outlined in Section 3.2.3. This process was iterative, with several different avenues explored and resulting failures and successes being used to direct the resulting methods used.

3.3.5 NNN Approach Application to SAFER data Experiments

After preprocessing methods were decided on, the model was applied to the SAFER Feasibility study 1 dataset and the resulting predictions were analysed. The AF probabilities predicted by the model were used in multiple different ways to analyse the performance of the model.

3.4 Result Evaluation Techniques

3.4.1 F1 Score for Classification

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).

$$\text{precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + 0.5(FP + FN)} \tag{7}$$

3.4.2 Efficiency Gain Metric

For analysing the effectiveness of a model at reducing the Cardiologist workload by prioritising the ECGs by predicted AF probability, the percentage increase in efficiency of this method was calculated as follows. This first involved dividing the Cardiologist labels into the binary classification "Conclusive AF Diagnosis", and "No Conclusive AF Diagnosis", with the latter of the two classes being "Heart Block", "Cannot exclude pathology", "No pathology" and "Ventricular Tachycardia" as seen in Table 2.

$$\text{Percentage Efficiency Increase} = \frac{1}{N} \sum_{i \in S} \frac{n_n^{(i)} - a_0^{(i)}}{n_n^{(i)}} \cdot 100 \tag{8}$$

Where $n_n^{(i)}$ corresponds to the number of "No conclusive AF diagnosis" samples for the patient i, and $a_0^{(i)}$ corresponds to the index of the first "Conclusive AF diagnosis" sample

from patient i, after these samples are ordered by predicted AF probability. S is a set of all participant IDs for which at least one sample from the corresponding participant was labelled "Conclusive AF diagnosis" by the cardiologist. N is simply the number of patients' IDs in this set S.

With the index value $a_0^{(i)}$ starting from 0, this metric will produce a value of 100% if the AF samples always appear first for each participant, and a value of 0% if the AF samples always appear last. This makes an appropriate performance metric because the aim of this project is to reduce the number of ECGs that need to be reviewed by the Cardiologist, therefore reducing time and costs, and once a single AF sample has been found for a participant it is not necessary to review any of their other samples as the AF diagnosis is made at this point. Therefore, if the AF sample appears first there is a huge increase in efficiency, yet if it appears last there is no increase in efficiency. Any samples that were not reviewed by at least 1 Cardiology team were not included in this calculation, because they provide no meaningful information.

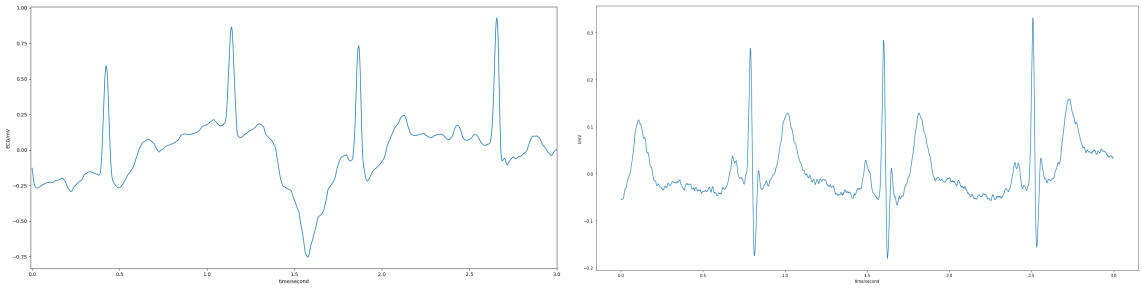
This efficiency gain metric reflects the time that is saved with the method used, for the Cardiologists reviewing samples from a participant whose samples contain at least 1 AF sample. This efficiency gain metric does not apply to any participants with no AF samples, because no time can be saved when every sample needs to be reviewed for AF to be ruled out.

4 Results

4.1 NNN Tested on PhysioNet Challenge 2020 Dataset

4.1.1 Experiment setup

After the initial implementation of the NNN on PhysioNet 2017 data, for validation, with promising results, see Appendix Section 8.4.1, the next step was to apply this model to data that it had not seen before. The CPSC Database [18] explained in Section 3.1.2 was used, which was a database of labelled 12 Lead ECG recordings. To create meaningful results, the Lead 1 data was isolated from the entire 12 Lead sample, and this created the labelled dataset for testing the model. 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. The Lead 1 isolated samples were the closest in waveform shape and features of the heart signal that are detected to the samples taken by a handheld 1 lead ECG device, such as the AliveCor and Zenicor devices. See Figure 9 for a comparison between a typical 1 lead handheld device recording and an isolated Lead 1 recording taken using a 12 lead ECG in a hospital.



(a) PhysioNet challenge 2017 sample

(b) CPSC sample

Figure 9: Samples showing the benefit of the ECG taken at the hospital with more clearly defined P, QRS, and T waves (9b), and the baseline wander found in samples from self administered 1 lead ECGs (9a).

In order to test the model's ability to distinguish between AF and non-AF, it was decided to split up the CPSC dataset for generating the results. With a vast number of different heart arrhythmia being labelled in this trial, many of which were a combination of AF and other heart arrhythmia detected in the same sample, it was decided to take only samples labelled as "AF" or "Normal" from this dataset for analysis.

4.1.2 F1 scores

The precision, recall and F1 scores from the classification results of the model on this dataset are shown below. These were used to compare the model's performance on this dataset to that of its known performance in the PhysioNet 2017 challenge, which used F1 scores for measuring performance. These values are calculated as in equation 7 where for the "A" row a positive is an AF label, and a negative is any other label, and for the "N" row a positive is a Normal label, and a negative is any other label. If the model predicts a sample to be "Other" or "Noisy", which are two of the four classes it has been trained to recognise, these values are included as a negative for both "A" and "N" rows,

$$\text{precision} = 0.9980 \tag{9}$$

$$\text{recall} = 0.8149 \tag{10}$$

Classification	F1 Score	Support
A	0.897	918
N	0.889	1221
F1 weighted average	0.892	

Table 3: Results of NNN model from PhysioNet challenge 2017 tested on Lead 1 CPSC data

4.1.3 Confusion matrix

The confusion matrix, Sensitivity, Specificity and Accuracy for "AF" predictions from these results are found below, which offer more insight into the model's performance on this dataset.

		Actual values	
		Positive	Negative
Predicted Values	Positive	995	2
	Negative	226	916

Table 4: Confusion matrix of NNN model predictions on Lead 1 CPSC data

$$\text{Sensitivity (recall)} = 0.8149 \tag{11}$$

$$\text{Specificity} = 0.9978 \tag{12}$$

$$\text{Accuracy} = 0.8934 \tag{13}$$

4.1.4 Observations

The first conclusion to be made from these results is that the model performed very well on this unforeseen dataset, with F1 scores of around 0.9 being significantly higher than those from the test of the NNN on its own train dataset, see Appendix Section 8.4.1, of around 0.84 to 0.88. This is initially surprising because it would be expected that a Deep Neural Network would perform best on data that it has already seen and been trained on than data which is new to it. However, this result is explained by the quality of the samples in the CPSC dataset, with significantly less baseline wander being present in the samples from the CPSC database than from the PhysioNet 2017 challenge dataset, as seen in Figure 9. Clearly, the preprocessing of the PhysioNet dataset, which uses bandpass filtering, could be improved with perhaps a more suitable method for removing the baseline wander of the signal. Furthermore, the bandpass filtered PhysioNet data has significantly less high-frequency noise than the samples from the CPSC dataset, which would be a cause for concern for worse results of the NNN predictions on the CPSC dataset, yet the opposite is observed. This suggests the model's increased sensitivity to low frequency, over high frequency, noise. In conclusion, the model places most weight in its predictions on the features it detects at a larger than smaller scale. It can ignore high-frequency noise if the waveform is still discernible. This helped aid in the decision for preprocessing of SAFER data for predictions of "AF" probability, with a more sophisticated method than band-pass filtering used to remove baseline wander.

Secondly, the model seems to be more accurate at distinguishing between AF and Normal samples, rather than AF and Other heart arrhythmia samples. The absence of "Other" samples in this dataset could be the reason for such good results, with the misinterpreting of another heart arrhythmia as AF by the model not being possible. This would naturally lead to an increase in precision, with the likelihood of false positives decreasing, and is observed with the low value of 2 seen in the confusion matrix. However, analysis of the 226 false negatives reveals 202 of them to be the model labelling an AF sample as "Other", and only 24 of them are the model labelling an AF sample as "Normal", therefore showing the model to have more difficulty spotting the difference between an AF sample and a sample showing a different heart arrhythmia.

This is encouraging, however, because the current methods used in the SAFER trial for identifying "abnormal" ECGs lead to roughly 35 samples being identified for every 1 that shows any pathology. Therefore, the ability of the model to identify these 34 Normal samples out of 35 is already a significant efficiency increase. However, this conclusion 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, with an accuracy of around 89%. This is essential for the application of the model to unforeseen data in the SAFER trail.

4.2 NNN Predictions of AF Probability on SAFER Feas1 data

4.2.1 Experiment setup

All samples from the SAFER Feasibility Study 1 dataset were first preprocessed as explained in Section 3.2.3. The NNN was then used to predict the probabilities of AF appearing in these preprocessed samples. Ideally, the predictions of AF probability will be high for samples diagnosed as AF by a cardiology team, but low for samples not labelled as conclusive AF diagnosis. These predictions were analysed using visualisations to find the most appropriate way of ordering the samples. With the ultimate aim of binary classifying the likelihood of samples being "AF" or "Non-AF", the resulting NNN predicted probabilities for the 4 classes "AF" "Normal" "Other" and "∼" were experimented with to find the most appropriate use of these probabilities for the task. The proposal metrics to calculate the AF probabilities experimented with include:

1. NNN predicted AF probability
2. A combination of NNN predicted AF probability and Other heart arrhythmia probability
3. 1 - NNN predicted probability of Normal sample

The visualisation in Figure 10 shows the probabilities of AF predicted, on the y axis, for each sample in the dataset which had been manually reviewed, on the x-axis, with the samples being divided into groups corresponding to the label the Cardiologists assigned to them. These colour coded groups are divided up to enable the easy visualisation of the probabilities the model applies to samples from each of these groups, and help with the choice of probabilities to be used for assigning to a sample for prioritisation methods. The mean probability values ± 1 standard deviation of these probabilities, for each of these categories, are plotted to assist with analysis. The solid lines are the mean values, the dotted lines are these mean values ± 1 standard deviation. Figure 11 shows the histograms of AF and Normal samples, from the same data as in Figure 10, for closer examination.

4.2.2 NNN predicted AF probability method

The first promising result from Figure 10 is that the predicted probability of "AF" is highest for the samples which do contain AF, according to the Cardiologist. This is seen by the mean of this section being highest, see the red section, at around 0.4 which is significantly higher than those from the other classes. Moreover, the histogram in Figure 11b shows that there is a much larger concentration at the higher end of the spectrum for AF samples than for Normal samples. Unfortunately, however, the predicted probabilities of "AF" for some samples in this section were very low, as seen in this histogram with a spike in probabilities around 0-0.05 probability. This indicates that there is a high likelihood this model will under-predict the probability of an AF sample containing AF, which reduces the likelihood of a threshold method, mentioned in Section 1.3, being found from these probabilities which will be safe for use. It is assumed that too many samples containing AF would go unreviewed by this method.

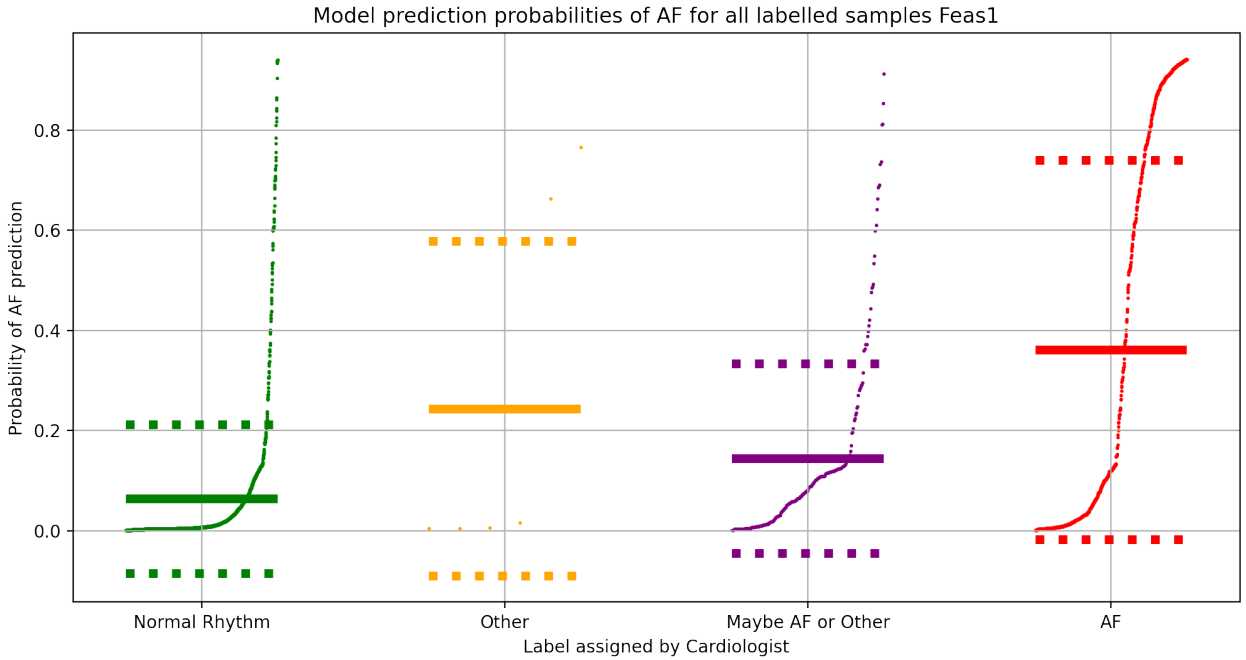
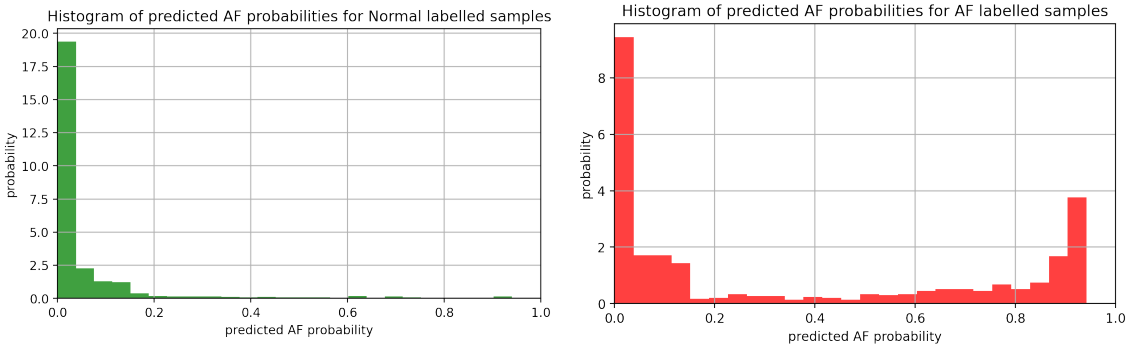


Figure 10: Probability of "AF" predicted by the model for all labelled data in SAFER Feasibility study 1, separated by label assigned by the cardiologist, with mean ± standard deviation plotted.



(a) Normal samples

(b) AF samples

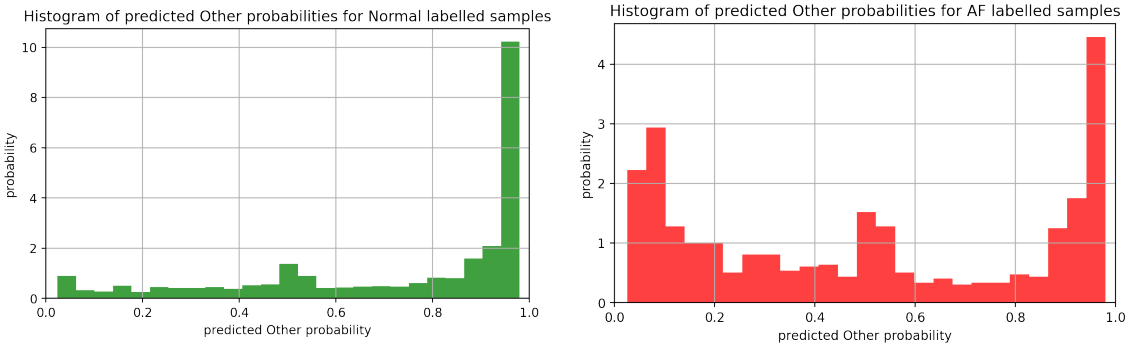
Figure 11: Histograms of the probability of "AF" predicted by the model for all Normal (11a) and AF (11b) labelled data in SAFER Feasibility study 1

Another encouraging result is that for Normal samples, as labelled by the Cardiologist, the NNN predicted likelihood of AF is very low. The mean value, see the green section, of around 0.05 was clearly lower than all other classes, and this is especially clear in the histogram shown in Figure 11a, where there is a concentration of probabilities around 0-0.05. The visualisation shows that only a handful of these samples had a high AF predicted likelihood, meaning that a prioritisation method based on NNN predicted "AF" probabilities was unlikely to put Normal samples first. Furthermore, the 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. The model's clear ability to distinguish between any heart arrhythmia samples and Normal samples suggest an alternative direction to approach the problem, by including the NNN predicted probabilities of "Other" heart arrhythmia in the values to order the samples for prioritisation.

4.2.3 NNN predicted AF probability ∪ NNN predicted Other probability method

In order to investigate the use of a probability for prioritisation which is the weighted sum of probabilities assigned to AF and Other heart arrhythmias, the histograms in Figure 12, show the NNN predicted probability of "Other" heart arrhythmia for the same samples as in Figure 11, were produced. These histograms rule out the use of "Other" probabilities being included in the probability assigned to a sample containing AF because a Normal sample would have a higher increase in probability than an AF sample. This is seen in the concentration of probabilities in Figure 12a around 1, whereas in Figure 12b the values are concentrated over lower probabilities, which would be counterproductive when the aim was to increase the probabilities assigned to these red, AF, samples.



(a) Normal samples

(b) AF samples

Figure 12: Histograms of the probability of "Other" predicted by the model for all Normal (12a) and AF (12b) labelled data in SAFER Feasibility study 1

4.2.4 1 - NNN predicted Normal probability method

The last metric to investigate was the NNN predicted "Normal" probability, in particular its inverse, for the potential use of this value in the ordering of samples for Cardiologist review. This is effectively showing the cardiologist the "least normal" samples first, or most abnormal, regardless of the abnormality. If proven useful, the NNN predicted probability of a sample being "Normal" would be used in a weighted addition of this probability with the NNN predicted "AF" probability, to increase the accuracy of the resulting probability. This proposal would order samples by a new value:

```
pred_AF_prob = NNN_probability_prediction_of_AF(sample)
inv_pred_Normal_prob = 1 - NNN_probability_prediction_of_Normal(sample)
sort_value(sample) = pred_AF_prob +  inv_pred_Normal_prob
```

Figure 13 is a plot of the 1 - NNN predicted probabilities of "Normal" for all samples from SAFER Feasibility study 1. This is effectively the NNN predicted probability of a sample not being Normal. Encouragingly, AF samples were assigned the highest probability of not being "Normal", which shows the model to be effective at distinguishing between Normal and AF samples, on SAFER data. However, this is only a small margin of difference with the mean predicted probability of a Normal sample not being "Normal" at around 0.8, whereas the mean prediction probability of an AF sample not being "Normal" was around 0.9, which is very close and arguably this difference should not be recognised as the model effectively distinguishing between these two classes. The difference is even more difficult to discern than in the histograms seen in Figure 14. Yet, the aim of this project is not to be as close to 100% accuracy as possible, but to provide a suitable method for ordering samples in increasing AF likelihood, and therefore this difference, however small, may be useful to use. The addition of the inverse of this probability could reduce the number of AF samples, as labelled by the cardiologist, having a low predicted probability assigned, which was the issue seen in Figure 10.

The low value of around 0.2 NNN predicted "Normal" probability for Normal samples is far from ideal, with an ideal model returning values closer to 1. One explanation could be explained on the preprocessing methods used. The methods chosen in Section 3.2.3 prioritised the model's ability to assign high predicted "AF" probability to AF samples, not for Normal classifications. This is clearly an opportunity for future work to improve performance.

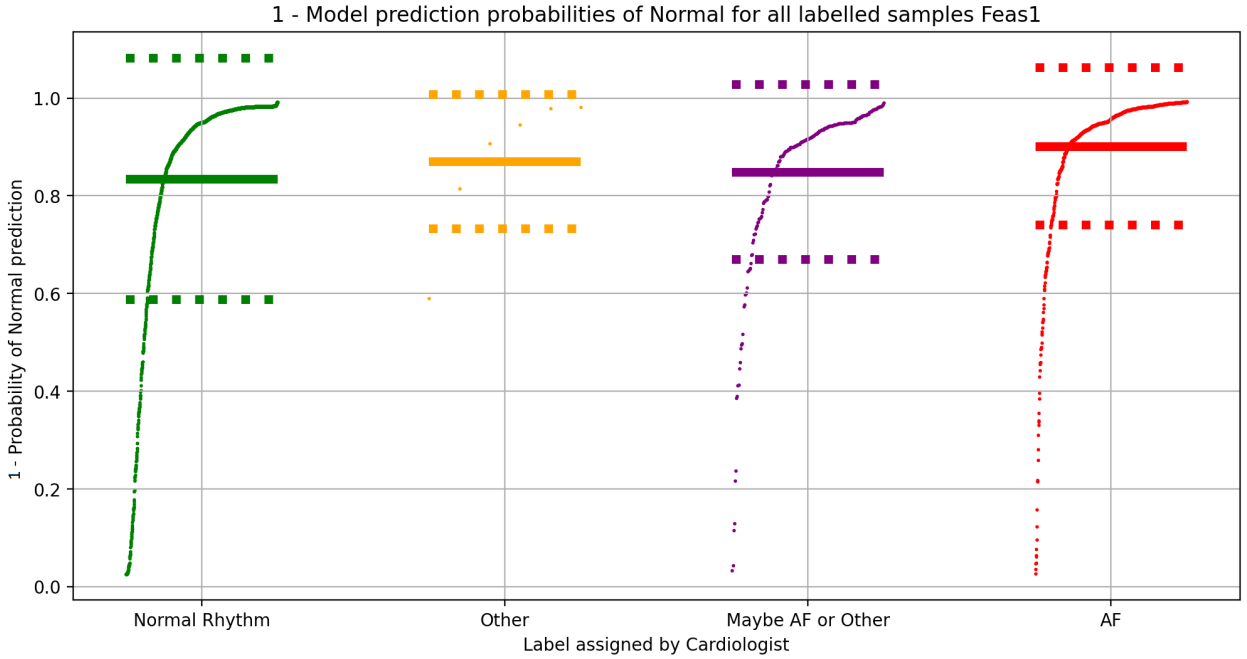
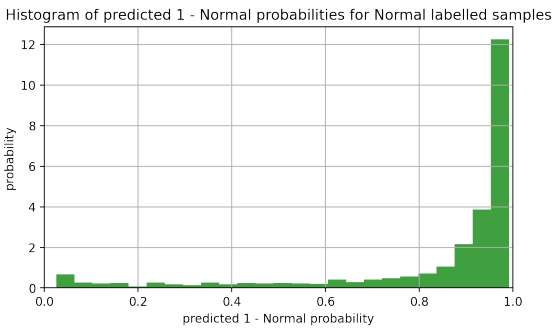
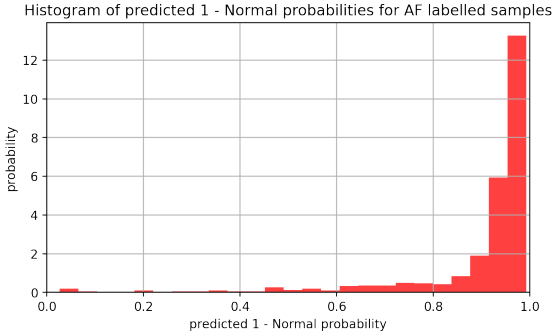


Figure 13: Probability of "Normal" sample predicted by the model for all labelled data in SAFER Feasibility study 1, separated by label assigned by a Cardiologist, with mean ± standard deviation plotted.



(a) Normal samples



(b) AF samples

Figure 14: Histograms of 1 - Probability of "Normal" predicted by the model for all Normal (14a) and AF (14b) labelled data in SAFER Feasibility study 1

4.3 Ordering SAFER samples using NNN on a per-patient basis

4.3.1 Experiment setup

After establishing the methods to test for prioritising the ECG samples, the implementation of these methods and the resulting efficiency increases were analysed. The 3 different values to use for ordering the samples were as follows.

1. NNN predicted "AF" probability
2. 1 - NNN "Normal" probability
3. Addition of the above two values

Although the first method is obviously the most promising because it utilises the NNN as it was trained to be used, it was not tested in a vacuum. All three of these methods were tested and corresponding values of efficiency increase, the metric defined in Section 3.4.2, were calculated. A visualisation was developed in order to analyse the subsequent ordering of samples using these values. Figure 15 shows the visualisation for samples ordered using the NNN predicted "AF" probability. All manually labelled samples are taken, with each being grouped by the participant that provided the samples. These participants were ordered on the y axis by the number of their samples that were individually reviewed by a Cardiologist team. These ordered participants are found along the y axis, with the x-axis corresponding to each sample from these participants. Colour coding was used to visualise the labels given to each sample by the Cardiologist teams. The aim of this project is to distinguish between samples where a conclusive AF diagnosis was made by the Cardiologist, "AF", or any other diagnosis was made, which are explained in Table 2, the colours to focus on in this visualisation are red for "AF", or any other colour for not "AF". However, the other sections were not bunched into one colour and label group because it is useful to see how the prioritisation method works with these less clear diagnosed samples, "Maybe AF", which correspond to samples labelled as "Cannot exclude pathology".

4.3.2 NNN predicted AF probability method

The first method to test, using the NNN predicted "AF" probability, produced the ordering as seen in Figure 15. Attention should be drawn to the locations where a participant has only a minority of samples labelled as AF. This is because the participants with all AF samples, or None, do not provide any insight into the model's performance. It is clear that for the participants with this lower AF burden, or lower proportion of samples labelled AF, the AF samples are more likely to appear near the left of the plot. This shows these samples are correctly being prioritised ahead of Normal, or any other classified samples, and would be the first samples to be reviewed by a Cardiologist team. This is a very promising result for the NNN method and is achieving the aim of the project. The same is also true of "Cannot exclude pathology" samples, labelled purple, which also mostly appear to the left of the plot. Yet AF samples generally appear before "Cannot exclude pathology", therefore the model is correctly becoming unsure of samples which the Cardiologists were not sure of themselves,

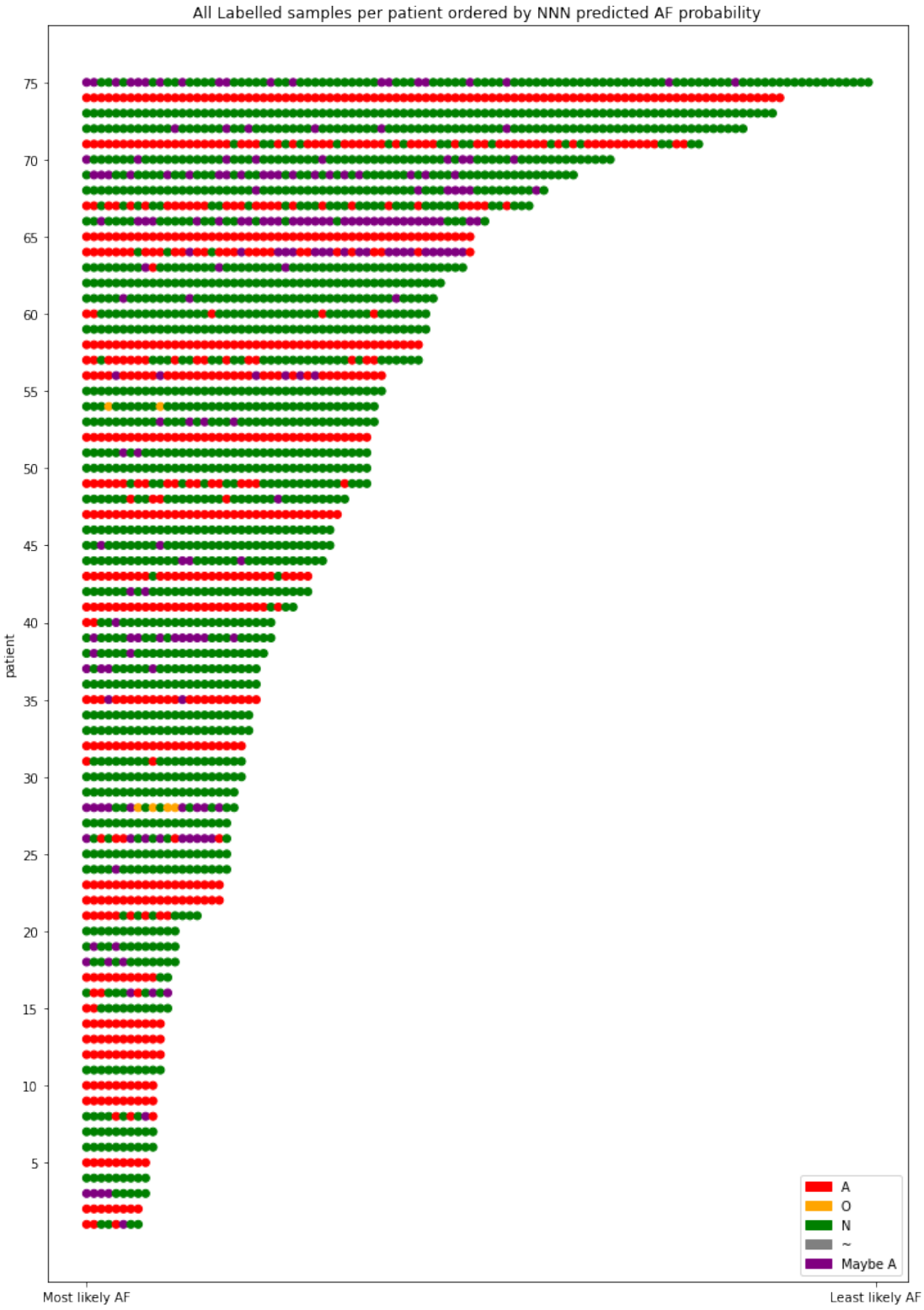


Figure 15: All labelled samples for 75 patients in Feas1 dataset, ordered by NNN predicted AF probability colour coded by the Cardiologist label

but is still putting them at a higher priority than the Normal samples, another promising result.

It is known that a patient's AF burden increases over time [21]. Also, the greater the AF burden the higher the risk of stroke and therefore the greater the importance of detecting AF in high AF burden patients. It is seen in Figure 15 that the higher the AF burden, or the greater the proportion of AF samples per participant, the higher the likelihood that the first sample from this participant is an AF sample. With the AF burden risks in mind, this property means that the model becomes more reliable as the risk to the participant, of misdiagnosis, becomes more severe. This increases the safety of this model being considered for a thresholding method, although this would need significantly more testing and validation.

The final step in this method was analysis was to calculate the prospective efficiency gain from Equation 8.

$$\text{Percentage Efficiency Increase} = 91\% \tag{14}$$

A projected 91% reduction in Cardiologist workload is significantly higher than the corresponding values from using more low level, non-DL, expert feature detection methods, with one recent paper stating a corresponding result of 74% [22].

4.3.3 1 - NNN predicted AF probability method

Following this result, the next method to test was using 1 - NNN predicted probability of "Normal" sample to prioritise the same ECGs. The visualisation of this ordering is seen in Figure 16. Interestingly, this method also ordered the samples in an appropriate way with AF samples appearing more towards the left of the plot. However, when the same efficiency gain calculation is made, the reduction in workload is 81%, which is lower than the 91% of simply using the NNN predicted "AF" probability.

4.3.4 NNN predicted AF probability ∪ 1 - NNN predicted AF probability method

Finally, the last method tested was using the addition of NNN predicted "AF" probability and 1 - the NNN predicted "Normal" probability. This unfortunately did not lead to better results, with the efficiency gain calculation being 88%, somewhere in between the values obtained using these methods individually.

Testing these other 2 methods validated using the NNN predicted "AF" probability alone as the value to order samples by.

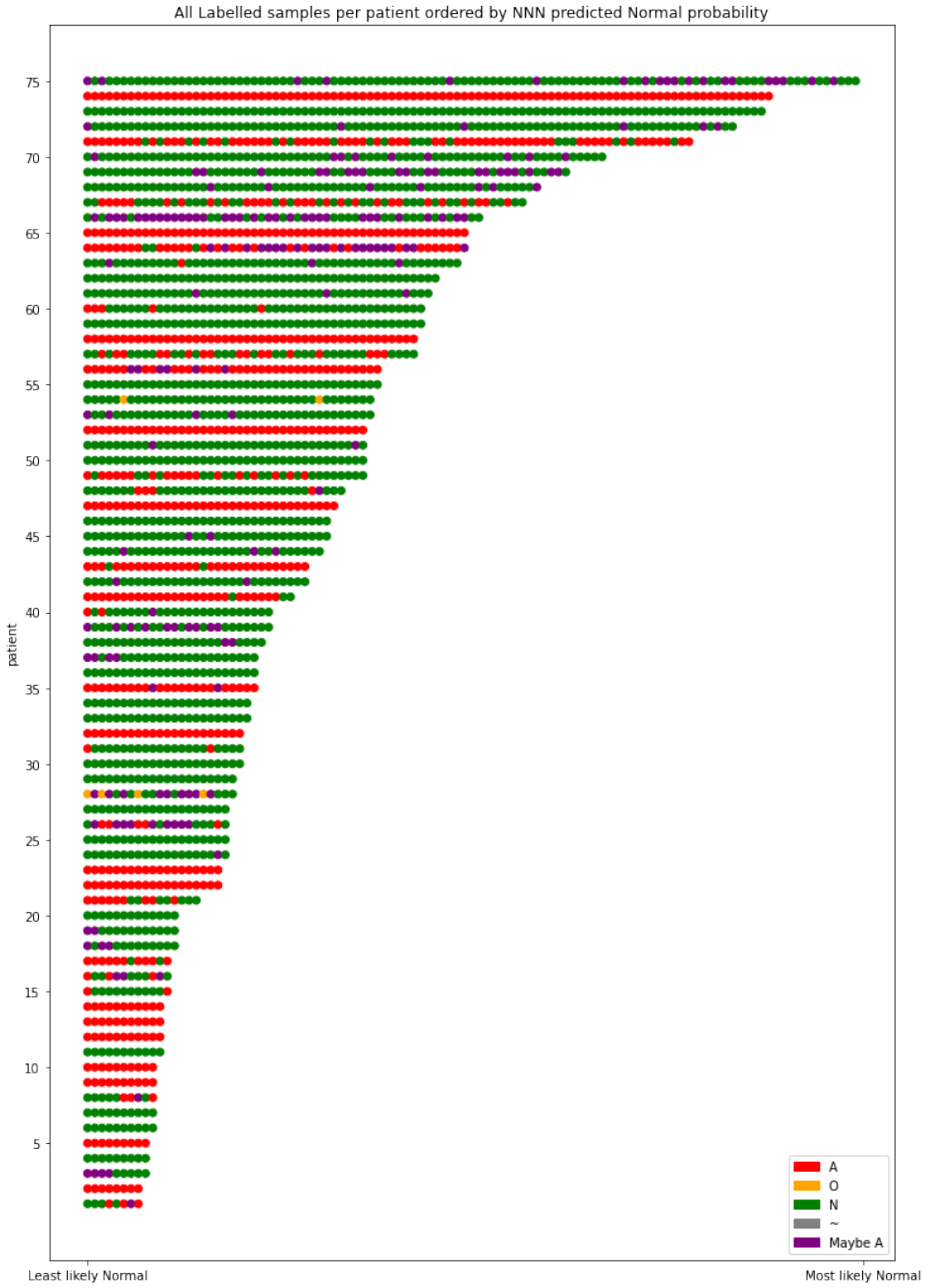


Figure 16: All labelled samples for 75 patients in Feas1 dataset, ordered inversely by NNN predicted Normal probability colour coded by Cardiologist label

4.4 Ordering the entire SAFER dataset using NNN

4.4.1 Experiment setup

One of the key challenges of screening for AF is the large number of samples that will have to be reviewed, and most of them will be irrelevant for diagnosis of AF because a Non-AF sample is insufficient to rule out an AF diagnosis. Only an AF sample confirms a diagnosis, that of AF. Therefore, one way a prioritisation model could be used is to prioritise the entire dataset collected during a screening process to quickly make diagnoses of patients. Ideally, all AF samples will appear first, and as each is reviewed and diagnosed as AF, all other samples from the diagnosed participant can be removed from the workload because the diagnosis has been made for these participants. This could drastically reduce the workload for Cardiologists, therefore the cost of the screening process, and increase its feasibility and quality of it.

In order to test the projected efficiency gains by ordering the entire dataset, the NNN model was used to prioritise the samples in the SAFER Feasibility study 1 dataset. Then the minimum number of manual reviews needed to make an AF diagnosis for a given number of participants in the study, a maximum of 65, was calculated. This was done by sequentially taking each sample, referring to the label provided by the Cardiologist, and if it was AF this participant is "diagnosed" therefore all samples from this participant is removed from the remainder of the samples for review. This is continued, with results of the number of "reviews" taken before each diagnosis being recorded at each "diagnosis". The ideal model would find all 65 AF patients from this dataset using only 65 samples.

4.4.2 Review count per diagnosis projection

Working within the limits of the dataset provided, the results of this method are found in Table 5, which shows the projected number of reviews needed for the number of patients then diagnosed with AF. In the ideal world, with a model which has 100% accuracy at prioritising ECG samples, these values will be equal. The worst-case scenario is that it takes all of the 4494 samples, which were manually reviewed, to find an AF sample for every single patient who had one.

No. participant AF diagnosis made	No. manual reviews required
5	5
10	15
20	26
30	37
40	70

Table 5: Results for the estimated number of samples needed to be manually reviewed for a given number of patient AF diagnoses, after prioritisation of samples using NNN

From Table 5 it is shown that the first 5 samples seen by the cardiologists would have resulted in 5 AF diagnoses. The first 15 would have resulted in 10. Importantly, however, a middle ground value of 70 sample reviews were needed to diagnose 40 participants with AF, which is roughly two-thirds of the dataset of 65 patients diagnosed with AF. If this is to be extended to a much larger dataset, such as the SAFER Trail dataset, it is assumed

that two-thirds of the Trail's participants, who have AF, can be diagnosed with an average of fewer than 2 samples needing review per patient.

4.4.3 Contextual comparison

This is a much bigger workload reduction for the Cardiologist than simply ordering the samples per participant, and then reviewing all participants in random order. This previously mentioned method will save 91% of the workload for any AF patients, but studies have shown that only around 11.4% of over 65s will have AF [1], therefore the workload saving can be estimated as 0.91×0.114, or 10.4%. With this new approach, with only 2 reviews per participant to find most of the AF patients in the study, this workload reduction can be reduced far more than this.

With the estimate of each ECG review taking 20 seconds [7], this means it will take less than 40 seconds for a Cardiologist to diagnose a participant with AF, every 40 seconds, with the possibility of 90 AF diagnoses being made for each hour of a Cardiologist's time. This is a very promising result, and increases the scalability of AF screening.

5 Discussion

5.1 Significance

These results show this purely DL approach which learned the key features for classification of ECGs rather than using more low level, expert features programmed using traditional methods from the fields of medicine, statistics and signals processing, was very effective at recognising the likelihood of AF being present in a sample. This purely data-driven method can be applied to SAFER data when the correct preprocessing is used, and using the softmax output from the NNN a reliable prediction can be made on the probability of AF for a sample. This is shown through the promising results of correlation between NNN predicted AF probability when a sample is labelled as AF by a Cardiologist, as seen in Figure 10.

These probabilities can then be used for ordering the samples, as seen in Figure 15, with a reliable outcome of AF samples appearing first, or near the start, for participants who had at least 1 AF sample. Although most AF samples did appear as the first sample with this ordering, in a few cases a non-AF sample appeared before an AF sample. In these cases, however, it was observed that the first AF sample would be not far behind, usually within a few samples from the first one. Therefore, the increase in workload for these cases was still minimal.

With this ordering used, time can be saved for the Cardiologist to review the samples in the SAFER study. With a Cardiologist's time is expensive, any increase in the efficiency of the review process is valuable. If this method is used, calculations for the ordering used in Figure 15 resulted in a reduction in workload for these participants of 91%, or a 10% reduction when accounting for 11.4% of the target age groups having AF [1].

However, further gains in efficiency could be made. Firstly, if the entire dataset was ordered by the NNN predicted probability, it is estimated that over two-thirds of AF patients can be correctly diagnosed with only an average of 2 samples needing review per participant. With the estimate of each ECG review taking 20 seconds [7], it is extended that 1 hour of a Cardiologists time could lead to the correct AF diagnosis of 90 patients.

Importantly, these prioritisation methods offer a way to reduce the workload of the Cardiologist without detriment to the safety of the diagnosis process. No diagnoses are made, and missed, by the method, it only aids the process by decreasing the number of reviews needed for the correct diagnoses to be made.

5.2 Limitations

One limitation of this project was the availability of individually labelled data. Only a small proportion of the dataset had been individually reviewed by a Cardiologist, with the majority of the data being unreviewed, unlabelled samples that at best had a participant level diagnosis which was often not appropriate for use on the individual samples. Not all samples from a patient diagnosed with AF would show signs of AF, with low AF burden patients having most samples not showing AF despite the patient level diagnosis being AF.

For a larger dataset, such as the SAFER Trail dataset which contains nearly double the number of labelled samples, and considerably more AF labelled samples, see Table 1, the calculated results would be more representative of the model's ability. This dataset was

unfortunately not made available to this project at this stage.

5.3 Future Work

With the promise of this method shown in this report, future work to take it further would start with the retraining of the model to include training examples from the SAFER study. Including these values in the training of the NNN would aid its ability to detect the necessary features in the samples obtained using the Zenicor device, not just with the PhysioNet 2017 dataset obtained using the AliveCor device. Also, the model could be trained with a cost function such as binary cross-entropy, with the focus of this project more on probabilities than just classifications, and the confidence of the model in the prediction being made is important. Also, the binary cross-entropy function would be suitable for the two options "AF" or "not AF" as opposed to recognising noisy samples, and other heart arrhythmias.

Moreover, this purely data-driven DL approach could be improved by including the expert features and applying more traditional ML methods in conjunction with the NNN approach. One paper from the SAFER study team has proven the effectiveness of prioritisation of ECG method solely using information from R peak locations [22]. This could be jointly implemented with the NNN, and analysis of the change in model performance may show an increase in performance. Furthermore, studies have shown that a combination of expert features methods, using traditional ML techniques such as Support Vector Machines, and DL approaches have led to better performance than these methods individually [16].

If these alterations make significant gains to the accuracy of the model, a threshold method could be considered for use with the predictions of AF this model would produce. The efficiency gains and safety of this approach would have to be considered, as it is assumed at this stage a thresholding method would leave too many AF patients undiagnosed.

Finally, this method needs to be tested on the SAFER Trail data, with this being the application of the prioritisation method, and the continuous testing and analysing of the model's performance on this data as more of it is produced in the study.

6 Conclusion

6.1 NNN as a model for predicting AF probabilities

This project showed the Novel Neural Network proposed for the PhysioNet Challenge 2017 [4], trained on the corresponding dataset for this challenge, can be effectively applied to the SAFER study data. When modified to provide a predicted probability of AF, the model correctly assigns the highest probabilities to the "AF diagnosis" samples and assigns low probabilities to the "no conclusive AF diagnosis" samples.

6.2 Effectiveness of prioritising ECGs

Using the proposed method of prioritising ECGs for review led to a review time efficiency increase of 91% for participants with AF. The review time efficiency increase for the entire population was estimated as 10.4% because most participants will not have AF, therefore the 91% increase in efficiency only applies to participants who have AF. If the entire dataset was ordered using this method, however, the efficiency increase is much larger. 1 hour of a Cardiologists time could lead to the correct AF diagnosis of 90 patients.

This significantly increases the scalability of an AF screening process. By reducing the time needed by the Cardiologist per AF diagnosis, this prioritisation method will reduce the costs, and increase the quality of the process without sacrificing safety, with diagnosis still being made by the Cardiologist review of ECG samples.

6.3 Preprocessing importance with SAFER data

The preprocessing methods had to be carefully chosen to ensure successful inference of AF probabilities from the NNN model on SAFER data. Most emphasis was placed on the removal of baseline wander with the model proving more susceptible to baseline wander than high-frequency noise, leading to a high chance of the model labelling a sample as "noisy" rather than useful predictions. This baseline wander sensitivity showed the NNN had learned to detect larger-scale features, of an order of tens to hundreds of time steps, rather than placing much weight on smaller-scale features. It could detect AF with even a large amount of high-frequency noise, as long as the waveform was discernible. However, the model did perform best when this high-frequency noise was removed, with more ideal probability predictions being inferred, therefore a low pass filter was still used.

6.4 AF screening promise

With AF screening offering a viable option for the large scale diagnosis of patients with AF, the quantity of data that will be generated during the SAFER study, and other studies, will need adequate methods for sorting. It will be impossible to review all samples, and alongside the filtering methods already developed, prioritisation of ECGs by AF probability as proposed in this report offers a viable option for reducing this workload.

7 Acknowledgments

This project was undertaken with valuable assistance and support from Collegues, friends and family. In particular the author would like to thank:

- Dr Elena Punskaya for her invaluable insight, advice, critique and guidance along the entire process from project start to report submission.
- Dr Peter Charlton from the SAFER Study for assistance with all matters relating to the Study, the data and interpretation of key medical concepts, and the sharing of useful literature for further insight into the research field.

References

[1] Chris Wilkinson, Andrew Clegg, Oliver Todd, Kenneth Rockwood, Mohammad E. Yadegarfar, Chris P. Gale, and Marlous Hall. Atrial fibrillation and oral anticoagulation in older people with frailty: a nationwide primary care electronic health records cohort study. *Age and ageing*, 50:772–779, 5 2021.

[2] NHS. Complications - atrial fibrillation. `https://www.nhs.uk/conditions/atrial-fibrillation/complications`, 5 2021.

[3] SAFER. The safer trial screening for atrial fibrillation with ecg to reduce stroke. `https://www.safer.phpc.cam.ac.uk`.

[4] 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.

[5] 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. `https://physionet.org/content/challenge-2017/1.0.0/`, 2017.

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

[7] 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.

[8] 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. `https://physionet.org/content/challenge-2020/1.0.0/`.

[9] George Moody and Roger Mark. Mit-bih atrial fibrillation database. `https://physionet.org/content/afdb/1.0.0`, 11 2000.

[10] 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.

[11] 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.

[12] 12-lead ecg placement guide with illustrations. `https://www.cablesandsensors.com`.

[13] Percy F. Morales. Atrial flutter: Symptoms, causes, and treatment. https://drafib.com/blog/atrial-flutter.

[14] 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.

[15] 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.

[16] 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.

[17] Physionet. https://physionet.org/.

[18] 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.

[19] Mit-bih arrhythmia database. https://physionet.org/content/mitdb/1.0.0/.

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

[21] 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.

[22] Mary Adeniji, James Brimicombe, Martin R. Cowie, Andrew Dymond, Hannah Clair Lindén, Gregory Y. H. Lip, Jonathan Mant, Madhumitha Pandiaraja, Kate Williams, and Peter H. Charlton. Prioritising electrocardiograms for manual review to improve the efficiency of atrial fibrillation screening.

8 Appendix

8.1 GitHub link

The GitHub repository for this project can be found here:
`https://github.com/js2432/Data_of_your_Heart_11B_Project/tree/NovelNN`

8.2 Low pass filter choice

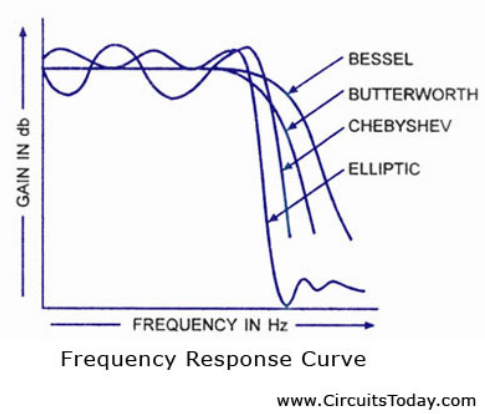


Figure 17: Comparison of LPF

8.3 Python Code

8.3.1 Scale normalisation

This is the code used after the 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 1: Scale normalisation code

8.3.2 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

8.3.3 Asymmetric 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 3: Asymetric least squares smoothing code

8.4 Further Results

8.4.1 NNN Tested on PhysioNet Challenge 2017 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 6 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 overfitted 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
$\sim$	0.72	286
Accuracy avg	0.84	
Weighted avg	0.89	

Table 6: Results of NNN model tested on PhysioNet challenge 2017 dataset

8.5 Relevant Function definitions

For reference, the equations for calculating the Accuracy, Sensitivity and Specificity of a classification model.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{17}$$

8.6 Risk assessment retrospective

The risk assessment performed at the start of this project identified minimal risks to health. The risk associated with computer working was repetitive strain injury. Adequate working environment was chosen to minimise these risks. These would be the same risks and precautions taken if the project was repeated.