Data of Your Heart: Screening for Atrial Fibrillation

Technical Milestone Report

IIB Student: Jordan Smith

Project Supervisor: Dr Elena Punskaya

SAFER team representative: Dr Peter Charlton

January 18, 2022

1 Summary

1.1 The problem

Atrial Fibrillation (AF) is a common abnormal heart rhythm that is associated with a five-fold increase in stroke risk [1]. It is identified through irregular-irregular RR intervals and lack of P waves in an Electrocardiogram (ECG), see Figure 1. If a patient is diagnosed with AF, medication can be administered to reduce this stroke risk. Currently, however, patients only have their heart monitored if they are symptomatic with AF, such as feeling their heart having palpitations, yet a significant proportion of AF patients will be asymptomatic, therefore not referred for examination. For some patients, AF episodes, which are short lasted episodes of AF which occur at varying frequencies, will not actually be detected by the ECG taken in hospital. The ECG is taken at a single time point, therefore likely that it misses the AF episode, especially for patients with low AF burden.

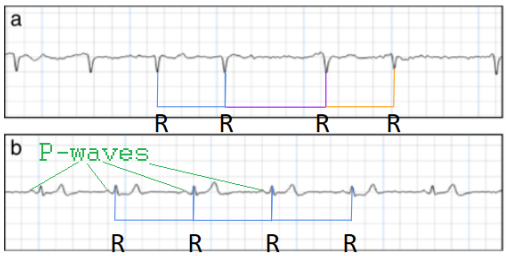


Figure 1: Electrocardiogram (ECG) recording with Zenicor device [2], showing atrial fibrillation (AF) (a) and sinus rhythm (b).

Screening offers a method to identify these asymptomatic, low AF burden patients who are otherwise missed by current methods, but still at risk. 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. This significantly increases the chances of detecting AF episodes in low AF burden patients. However, this generates a significant number of samples, 84 per patient, which needs to be checked by a Cardiologist. Because only one sample, out of 84, needs to show an AF episode for the patient to be diagnosed as having AF, significant time and cost is saved if the first sample the Cardiologist sees is one that shows AF signs if any do at all. This is the motivation for this project.

There are some key difficulties in this approach of using self-administered 1 lead ECGs such as the Zenicor device, instead of 12 Lead ECGs in a hospital. Firstly, noise in the samples due to user error is significantly more likely, with the incorrect use of the device leading to poor contact between the patient and electrodes, for example. Also, due to the direction of the current a 1 Lead ECG reads through the heart, there is a high likelihood that p waves will not be detected, even if Atrial depolarisation, the process which causes the p waves, is occurring. Therefore, a lack of p waves present in these samples is not significant evidence to suggest AF, because the equipment could be to blame.

Another consideration for this project is the presence of other heart arrhythmias in the samples, with some being easily confused with AF such as Bradycardia, Tachycardia and Heart Block.

1.2 The objective

To develop a systematic way to order the samples by likelihood of exhibiting AF symptoms, in order to speed up the review process for the Cardiologists in the SAFER trial.

1.3 Background research into approaches

Classifying ECGs for AF or Sinus Rhythm (SR) is a big area of research, with many approaches already developed through numerous competitions, such as Physionet challenges. The application of Machine Learning, and in particular deep learning, has for the last 5 years received a lot of attention due to the ability of these methods to function very well on noisy ECGs.

One approach used by Dr Jie Lian et al. in their paper "A Simple Method to Detect Atrial Fibrillation using RR Intervals" [4], neatly produces accurate classifications using only the RR intervals of the ECG. However, this method was tested on the MIT-BIH AF database [5], which ... doesn't deal well with noise.

*** why is this method, applied to the MIT-BIH database, not suitable for SAFER data? Do Ambulatory ECG recorders produce different looking ECGs to the Zenicor device?

The ML methods mentioned before, such as those developed for the Physionet Challenge 2017, were aimed more at these noisier Lead 1 samples that are expected from small handheld devices. The Physionet 2017 challenge asked for algorithms to classify samples between "Normal" "AF" "Other" and "Noisy" categories and measured the performance as the average F1 score over each category. Out of the many successful methods submitted to this challenge, a couple of the approaches stood out as particularly interesting, both with very high

performance:

• "ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks" [6], which had a high joint winning score of 0.83 from a combination of DL methods and expert features which could be learned with more traditional ML techniques. However, most of the methodology was not included in the report, and therefore applying this method from scratch, without assistance from trained Cardiologists, is deemed above the scope of this project.

• "Robust ECG Signal Classification for Detection of Atrial Fibrillation Using a Novel Neural Network" [7], was of particular interest due to its ease of implementation, with the entirety of the algorithm being stated in the report. It also still had a high score of 0.82. After considering many other possible approaches, from different sources, this was deemed most appropriate.

2 Methodology

2.1 The Data

The primary data source for this project is the dataset produced in the SAFER study [3], which contains over 9988 samples that have been individually reviewed by cardiologists, and growing as the study continues. As well as this, these datasets contain around 175000 samples in which the patient themselves has been diagnosed as having AF or not. In these samples, there will be numerous that are ground truth labelled as 'AF', yet do not actually display any signs of AF, because those individual samples are from a patient with low AF burden and therefore not all the samples from this patient will show AF signs.

Another source of data is 2 open-source datasets, the Physionet Challenge 2017 dataset (8528 samples) which is Lead 1 ECG data, and the China Physiological Signal Challenge (CPSC) Database [8] (6877 samples) which is 12 lead ECG data. The 12 lead data is not as useful, even though the lead 1 data can be isolated, because it is taken from ECGs produced in a hospital, therefore, is less noisy. Augmentation to add noise, and bandpass filtering, in order to simulate data produced by a handheld device is needed.

Furthermore, another consideration of the datasets is the skew towards 'Non-AF' samples. For the SAFER dataset, there is an average of 12%, of samples labelled, labelled as 'AF'. For the Physionet 2017 dataset only 9%. Care is taken to prevent the model from overfitting to the 'Non-AF' end of the probability spectrum.

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

2.2 Novel Neural Network approach

The Novel Neural Network (NNN) approach [7] chosen has architecture as shown in Figure 2. This network was applied to samples split into 5-second sections, with each section having

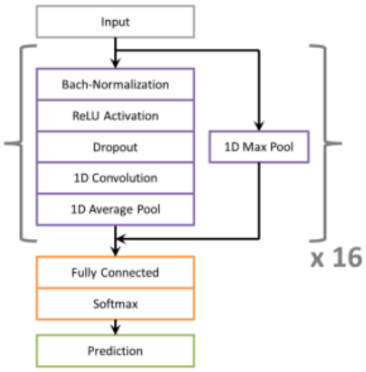


Figure 2: Proposed network architecture for the Novel Neural Network approach with skip connections [7]

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

Table 1: Results of NNN model from Physionet challenge 2017 tested on its train dataset for quick validation

a classification prediction made for it, and then the classification that appeared the most over all of these split samples is the one chosen as the classification for the entire sample. Although this method was initially applied to a classification problem, the model will be adapted to produce a probability, rather than classification, of âAFâ. There are multiple options for achieving this, with the simplest taking the proportion of the sample splits which are classified as âAFâ as the likelihood of the sample showing AF. A more likely option to achieve better, more continuous, results would be to examine the output of the softmax layer, and through experimentation find a suitable way of combining these outputs into a âconfidenceâ of AF.

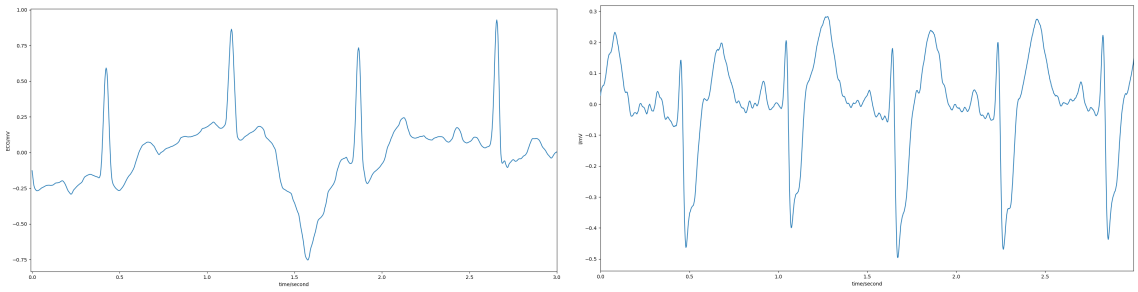
3 Results

3.1 Tested on Physionet Challenge 2017 dataset

In order to check that the NNN model worked it was tested on the dataset on which it was trained, see Table 1 for results. Results were promising, with values confirming the model functioned as expected. 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.

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

Table 2: Results of NNN model from Physionet challenge 2017 tested on lead 1 CPSC data [8]



(a) Physionet challenge 2017 sample

(b) CPSC sample

Figure 3: Samples from the two open source datasets, showing the benefit of the ECG taken at hospital with more clearly defined p, qrs, and t waves (3b), and the drift and noise sometimes found in samples taken from self administered 1 lead ECGs (3a).

3.2 Tested on CPSC database [8] Lead 1 data

The next step was to apply this model to data that it had not seen before, the CPSC database. 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 but is still a useful tool for validation purposes, see Figure 3 for comparison. Results are shown in Table 2, with only the classifications of 'A' or 'N' of interest. This model actually performs significantly better on this dataset than when tested on its own train dataset, which is initially surprising. This is explained by the lower noise present in the samples from the CPSC database, due to the higher quality equipment, leading to the model performing very well at recognising key features, because they are more clear.

4 Plan for project

Most of the tooling for the data pipeline has been developed for the open-source datasets, with minor adjustments needed to incorporate the SAFER datasets. Functionality for visualisation and analysis of results has been developed.

Currently, the model is being built from scratch in Tensorflow, based on the NNN architecture, for experimentation and application to SAFER data. A decision on how to categorise the other heart arrhythmia present in samples is going to take some experimentation, but currently for speed of implementation purposes these samples are being ignored. Should

'other' be a separate category, as in the original application in Physionet challenge 2017, or should 'other' and 'N' be bunched together as 'Non-AF''?

Future tasks once this is complete:

• Produce a performance metric that most suitably reflects the objective of the project.

• Once this metric is produced, the model will be adapted to produce a probability, rather than classification, of 'AF'.

• Splitting the SAFER data into train and test datasets in order to develop this model for application to SAFER trial.

• There will be further experiments on the sample splitting technique mentioned in section 2.2. Would splitting the samples according to RR intervals, with padding needed, instead of into 5-second sections lead to better results? Is this splitting technique even necessary?

• Experimenting with preprocessing of the data, filtering and augmentation for increasing the train dataset size.

References

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

[2] Sara Schukraft, Marco Mancinetti, Daniel Hayoz, Yannick Faucherre, StÃ©phane 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.