UNIVERSITY OF CAMBRIDGE
Department of Engineering

Cardiac Arrhythmia Screening Using Feature-Based and Deep Learning Methodologies for UK Biobank data

Author Name: Hiral Radia
Supervisor: Elena Punskaya
Date: 27th May 2020

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

Signed  *date* 27/05/2020

Abstract

Cardiovascular disease was responsible for 1 in 4 premature deaths in the UK in 2019. Cases usually originate as a result of damaged heart tissue, and their only symptom is often a cardiac arrhythmia (an irregular heart rhythm). Cardiac rhythms are recorded using an electrocardiogram (or ECG) signal, which monitor the activity of the heart by measuring the variation in the potential difference between two points in the body over time. By analysing the ECG signal, the presence of cardiac arrhythmias can be detected.

In this project, the development of an automated screening system for cardiac arrhythmias was investigated. Both ensemble-based methods and methods using 1-dimensional neural networks were investigated, in order to compare their relative merits as part of a screening system. The system would enable readily-available devices, which record Lead I ECG signals, to be used in healthcare. The costs typically associated with recording and screening using 12 lead ECG signals can thus be reduced.

Three datasets were used when conducting this investigation: the PhysioNet Challenge 2017 dataset, the PhysioNet Challenge 2020 dataset and the UK Biobank ECG database. For each of these datasets, if readings from more than one ECG lead was available, only the first was used. The Challenge 2017 dataset permitted the use of deep learning due to its large size, but investigating the performance of these methods in detail was difficult due to labelling inadequacies in the dataset. The Challenge 2020 dataset was more specifically labelled, but its smaller size made it less suited to deep learning. As such, a random forest and manual feature extraction were used instead for this dataset. Lastly, the UK Biobank ECG database was used to simulate real-world testing of the models developed and trained using the two Challenge datasets. To accomplish this, the performances of the two models and the *CardioSoft* automated ECG diagnostic system (used by the UK Biobank ECG machines) were compared using the Biobank database.

The report introduces the motivation behind the project in Chapter 1, with Chapter 2 providing the technical background required to understand the rest of the project. Chapter 3 details the development of neural network models using the Challenge 2017 dataset, and Chapter 4 focuses on the development of a random forest model using the Challenge 2020 dataset. Chapter 5

summarises the simulation of real-world testing using the UK Biobank ECG database whilst Chapter 6 concludes all of the work done.

When using neural networks to classify the PhysioNet Challenge 2017 dataset, the deep ResNet architecture and dilational neural network used performed better than the shallow convolutional neural networks and LSTM-based networks. For a binary 'normal'/'abnormal' classification of the Challenge 2017 records, the deep ResNet architecture achieved a test AUROC score of 0.875 - an impressive score when working with a challenging, noisy dataset.

By using a random forest to classify the signals found in the PhysioNet Challenge 2020 dataset, an AUROC score of 0.967 was achieved, and feature selection and hyperparameter optimisation were found to improve the performance of the model. The ResNet model previously trained with Challenge 2017 dataset was subsequently found to achieve an AUROC score of 0.758 when used to classify the signals in the Challenge 2020 dataset. Pre-processing and noise removal made a significant difference to the model's performance when being tested using a different dataset to the one on which it was trained (more than doubling the AUROC score achieved). By using the output of the ResNet model as an input feature for the random forest, the random forest model's AUROC was improved to 0.971.

When the models were tested on the Biobank dataset, random samples of records for which the screening processes developed were not in agreement were investigated. The *CardioSoft* automated diagnosis system performed poorly when detecting a variety of arrhythmias, whilst the ResNet model performed poorly when detecting atrioventricular blocks, and the random forest model performed worse when detecting ectopic beats. The random forest model achieved a better estimated true positive rate than the other diagnosis methods, performing better than the *CardioSoft* automated diagnosis system despite its use of 1 ECG lead rather than 12.

Contents

Chapter 1: Introduction

According to the NHS website, cardiovascular disease was responsible for 1 in 4 premature deaths in the UK in 2019 [1]. The development of technology to allow the treatment of cardiovascular disease to be improved therefore represents an opportunity to significantly reduce the number of premature deaths in the UK. Although the treatment of many cardiovascular diseases has been refined in recent years, the detection of these conditions remains a significant issue.

The rise in readily-available devices which record Lead I ECG signals means that ECG signals can be obtained at a far lower cost, but each patient must ultimately be diagnosed by a cardiologist in order to receive treatment. This project aims to develop an automated ECG screening system, to reduce the number of samples that cardiologists are expected to investigate.

In order to develop the screening system, data to train and test the models is essential. This project used data from the PhysioNet [2] Computing in Cardiology (CinC) Challenge 2017, and 2020 datasets to develop binary classification models, allowing unseen ECG signals to be labelled as 'normal' or 'abnormal'. These models were applied to the $37,207$ ECG signals in the UK Biobank ECG database [3], in order to simulate real-world testing using large volumes of data. As such, the screening system was developed to be used at scale, with a focus on producing a sufficiently generalised model to classify ECG signals recorded using a range of hardware.

The ECG machines used to record the signals in the UK Biobank ECG database natively support *CardioSoft* automated diagnosis software. This software is proprietary and has very little supporting documentation, so it is difficult to know how it was developed. The project aims to address this by comparing the *CardioSoft* automated diagnosis software's performance to that of the models developed when testing is performed using the UK Biobank ECG database. As the *CardioSoft* automated diagnosis software uses a 12 Lead ECG signal when performing diagnoses, allowing a far more holistic view of the heart, an additional challenge is presented when trying to produce a functional screening process using just a Lead I ECG.

The project hence aims to use machine learning to develop an automated screening system for cardiac arrhythmias using a Lead I ECG signal.

Whilst doing this, the answers to 3 questions were also investigated:

1. How do the techniques used vary depending on the training data available?

2. Are the models produced general enough to be used on a different dataset?

3. How does the model compare to the *CardioSoft* software currently used by UK Biobank to help with the diagnosis of arrhythmias?

Most literature focuses on the training and testing of ECG arrhythmia classification algorithms, either by using a dataset from one source throughout or by combining multiple datasets before any models were trained and tested. In most cases, longer ECG recordings were separated into many, shorter segments, before splitting the dataset into training and test data. Due to the limited variation in a single patient's ECG signal over short time periods, if a single patient's ECG recordings are present in both the training and test datasets, it can falsely inflate the accuracy of the algorithms used. This is because near-identical records are present in both the training and test datasets, so performances cannot be replicated when performing real-world testing [4].

This report uses a different source of data to train the models to that used to simulate real-world testing. This allows an assessment of the generality of the models produced as well as a realistic view of the models' accuracy. It was ensured that patients that had ECG recordings present in the training dataset did not have recordings present in the test dataset for any of the testing performed in the report.

To date, no work has been done to investigate the detection of cardiac arrhythmias using the UK Biobank ECG database (to the best of our knowledge), so all work focusing on the use of Biobank data is entirely unique. This includes the work done to devise a system to investigate the performance of the models produced without the provision of ground truth labels.

Throughout this report, various facets of the development of a cardiac screening system are explored. Chapter 2 details the technical background, whilst summarising the preprocessing and noise removal techniques used throughout the rest of the report. Chapter 3 uses the PhysioNet CinC Challenge 2017 dataset to train and test neural network based classification, and Chapter 4 uses the Challenge 2020 dataset to train and test a random forest classification model. Chapter 5 explores the real-world testing performed using the UK Biobank ECG database, and Chapter 6 concludes the work done during this project.

Chapter 2: Technical Background

2.1 What is an ECG?

An electrocardiogram (or ECG) is a way to monitor the behaviour of the heart by measuring the variation in the potential difference across the heart, using sensors attached to the skin. ECG signals are used to detect cardiac arrhythmias (abnormal heart rhythms), allowing the diagnosis of cardiovascular disease.

Both the placement and number of the electrodes used to measure the ECG affect the clarity with which particular symptoms can be seen; between 1 and 15 signals are used in clinical settings to diagnose arrythmias. For the purposes of this report, the term ECG refers to a Lead I ECG, measuring the potential difference between the left and right arms. This choice was made beacuse Lead I ECG signals can be recorded by the patient using low cost devices, rather than needing an ECG technician and an ECG machine, removing the high cost of labour and capital associated with recording ECG signals. The use of Lead I ECG signals presents a significant challenge, as Lead II ECG signals (measured between the right arm and left leg) provide a more comprehensive view of cardiovascular behaviour, whilst the other leads that are included in widely-used 12 lead ECG recordings provide vital supporting information when detecting arrhythmias [5].

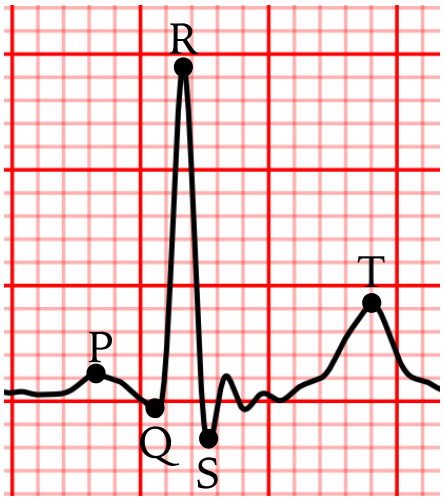


Figure 2.1: A diagram showing the P, Q, R, S and T waves on a heartbeat from a Lead I ECG.

A healthy ECG signal is composed of a periodic set of features, as shown in Figure 2.1. These features are a result of the depolarisation and repolarisation of cardiac muscles, with feature intervals and shapes providing an insight into the behaviour of the cardiac muscles.

This report concentrates on 6 of the most common arrythmias - atrial fibrillation, premature ventricular complexes, premature atrial complexes, atrioventricular blocks, non-sinus tachycardia and non-sinus bradycardia. The effects and symptoms of these arrythmias are detailed below.

Atrial Fibrillation (AF)

- Lack of P wave.
- 'Irregularly irregular' interval between R waves.
- Fibrillation (oscillatory) waves between waves, rather than a flat baseline.

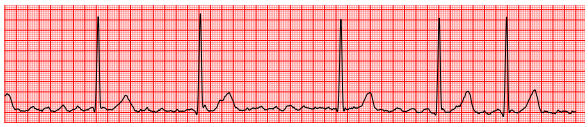


Figure 2.2: A Lead I ECG trace showing symptoms of atrial fibrillation.

Premature Ventricular Complexes (PVC)

- An ectopic beat (a beat occuring prematurely - beat 4 in Figure 2.4).
- The ectopic beat is wider than a normal beat.

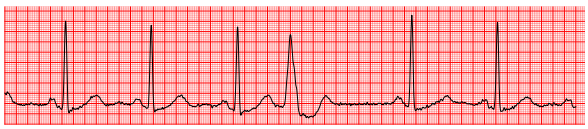


Figure 2.3: A Lead I ECG trace showing symptoms of premature ventricular complexes.

Premature Atrial Complexes (PAC)

- An ectopic beat (a beat occuring prematurely - beat 4 in Figure 2.4).
- The ectopic beat is narrower than a normal beat.

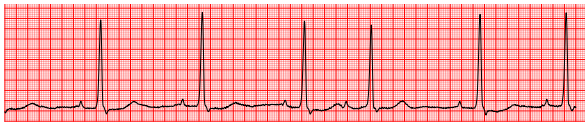


Figure 2.4: A Lead I ECG trace showing symptoms of premature atrial complexes.

Atrioventricular Block (AVB)

- The interval between the P and R waves is prolonged (greater than 0.2 seconds).

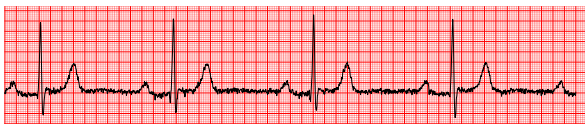


Figure 2.5: A Lead I ECG trace showing symptoms of an atrioventricular block.

Tachycardia

- The heart rate is higher than 100bpm (R peaks are too close together).

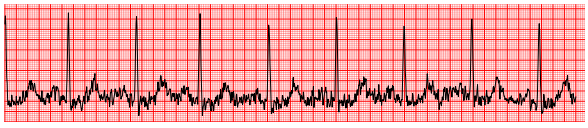


Figure 2.6: A Lead I ECG trace showing symptoms of tachycardia.

Bradycardia

- The heart rate is lower than 60bpm (R peaks are to widely spaced).

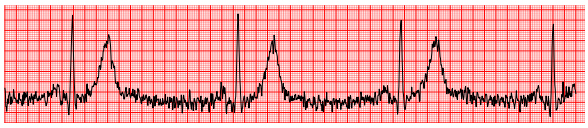


Figure 2.7: A Lead I ECG trace showing symptoms of bradycardia.

2.2 Simulation of Real-World Testing

In order to show that a medical system works, it is important to perform real-world testing. This involves the use of the technology developed through the project with real-world data in large volumes, in order to assess its performance and ease of use. An application to obtain ECG recordings from the UK Biobank [3] was made and approved, allowing access to a large-volume dataset taken in clinical settings to simulate the real-world testing of the automated screening system developed.

The use of the Biobank dataset presents its own unique set of challenges. The records in the dataset have not been individually analysed by an ECG technician, so any excessively noisy ECG recordings remain in the dataset. Similarly, the testing of models using the dataset is made more complex as the ECG signals have not been labelled by a cardiologist to provide a ground truth with which the models' outputs can be compared. The patients from whom the Biobank ECG recordings were taken were not selected solely to facilitate the development of a cardiovascular disease screening system, so the resulting dataset is very unbalanced. This means that, even if a ground truth could be obtained, many standard measures of model performance (such as the 'F1 Score') do not give a holistic view of the model's successes and flaws. Lastly, as the source of the Biobank dataset is different to that of the training dataset, the nature of the datasets (the presence of different sources of noise, the sampling rate, the amplitude of the signals etc.) must be aligned to ensure that the both datasets are comparable. The pre-processing allows the generality of the models developed to be increased [6].

2.3 Model Scoring

In order to be able to easily compare the two models produced, a set of standardised metrics will be used when testing them, and a standardised set of visualisations produced. This section describes these visualisations and metrics.

2.3.1 Result Statistics

Definitions of true positive, false positive, true negative and false negative rates are given in Equations 2.1 to 2.4.

$$\texttt{True Positive Rate} = \frac{\sum \text{Correctly Diagnosed by Model}}{\sum \text{Patient has Arrhythmia}} \tag{2.1}$$

$$\texttt{False Negative Rate} = 1 - \texttt{True Positive Rate} \tag{2.2}$$

$$\texttt{True Negative Rate} = \frac{\sum \text{Correctly Labelled as Healthy by Model}}{\sum \text{Patient is Healthy}} \tag{2.3}$$

$$\texttt{False Positive Rate} = 1 - \texttt{True Negative Rate} \tag{2.4}$$

For the purposes of a screening system, it is important to refer all patients with arrythmias to cardiologists, whilst referring as few healthy patients as possible. As such, the true positive rate must be maximised, whilst keeping the false positive rate as low as possible.

2.3.2 Reciever Operating Characteristic Curve

The receiver operating characteristic (ROC) curve will be used in order to visualise the model performance. The false positive and true positive rates are plotted as the threshold used to obtain labels from the probabilities output by the model are varied, hence indicating how large the true positive rate can be made whilst keeping the false positive rate as low as possible. An example ROC curve is given in Figure 2.8 below.

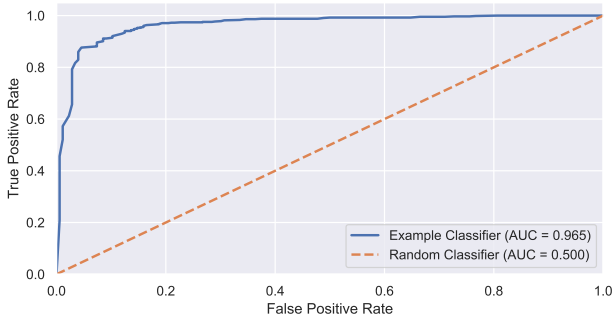


Figure 2.8: An example ROC curve, showing an example classifier and a baseline curve (based on a random uniform distribution between 0 and 1).

Figure 2.8 shows that the example classifer performs far better than the baseline curve, as it is able to achieve a higher true positive rate whilst maintaining a lower false positive rate for all non-trivial values of the threshold used.

2.3.3 Area Under the Curve

On a plot of an ROC curve, the area under the curve (AUC) gives the probability that a randomly chosen positively labelled record will be assigned a higher probability (hence being deemed more likely to be positive by the model) than a randomly chosen negatively labelled record. This metric allows models to be compared using a single number, but loses much of the information associated with ROC curves. This is because it is no longer possible to see how the proportion of false positive and false negative samples varies with the value of the threshold used. As such, ROC curves are used where possible in this report, but at times AUC statistics are used for conciseness.

2.4 Pre-Processing

Pre-processing ensures that the datasets used are directly comparable. Bandpass filtering is generally avoided here as it distorts the features shown in Figure 2.1, potentially resulting in misdiagnoses [7]. As such, bandpass filtering stage is only used in this project if any of the datasets were bandpassed before they were obtained. Figure 2.9 overleaf shows the generalised pre-processing method used.

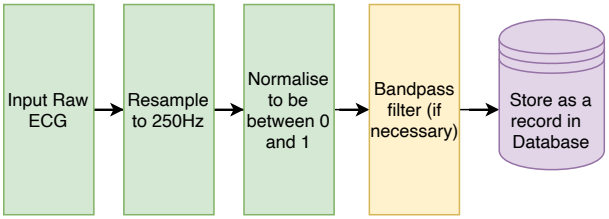


Figure 2.9: A flowchart showing the pre-processing used for this project.

Pre-processing increases the generality of the models developed, allowing different datasets to be used for training and testing. A graph showing the performance of a convolutional neural network model with and without preprocessing is shown in Figure 2.10 (see Chapter 3 for an in-depth explanation of the model used). The Challenge 2017 dataset was used to train the model and the Challenge 2020 dataset was used to test it.

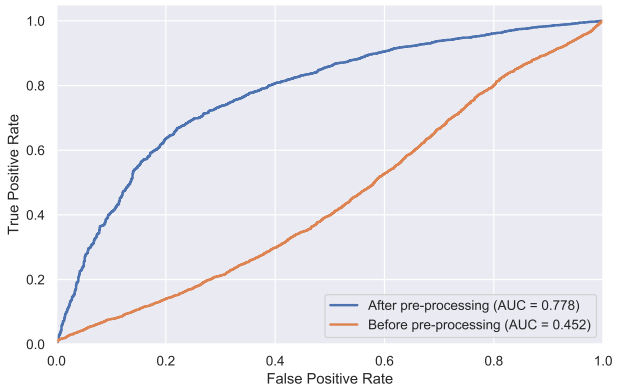


Figure 2.10: Receiver operating characteristic curves showing the performance of a CNN (trained on one dataset and tested on another) with and without pre-processing.

Figure 2.10 shows that pre-processing improves the performance of the classifier, as the receiver operating characteristic (ROC) curve after pre-processing has a greater true positive rate than the before pre-processing ROC curve for all false positive rates.

2.5 Noise Characterisation and Removal

In this section the sources of noise present in the ECG recordings, and their corresponding removal methods, are detailed. Noise removal is used to ensure that models do not underfit (due to noise) by increasing the prominence of the underlying ECG signal.

2.5.1 Characterisation

The main noise found in ECGs has 3 sources:

1. **Wandering baseline:** patients' muscle movements (including breathing) result in low frequency fluctuations in the baseline of the ECG. This means that peak magnitudes can be incorrectly measured.

2. **Electromyographic (EMG) noise:** shivering and muscle tremors result in the presence of stochastic high frequency noise in the signal. This can be modelled as additive Gaussian white noise [8], the variance of which depends on the muscle contraction that produces it.

3. **Electrode motion artefacts:** the stretching of the skin around the electrode as well as poor electrode adhesion result in artefacts similar to wandering baseline noise. These artefacts have a frequency between 1 Hz and 50 Hz, which is similar to that of the medical features to be detected. As a result, linear filtering to remove the noise in this range will distort the key medical features.

2.5.2 Noise Removal

Baseline Wander Removal

For the use of both random forest and neural network based classifiers, baseline wander removal was important to prevent overfitting. This is because the calibration and level of noise varied between records, possibly resulting in correlations between the level of noise and the label. This was particularly important when using multiple datasets, as data measured using different hardware has a different noise profile.

Initially, ensemble empirical mode decomposition (EEMD) [9] was used to remove the baseline wander from the full signal. The residual was removed after the EEMD was performed, and up to 3 of the lowest frequency intrinsic mode funtions (IMFs) were subsequently removed. Figure 2.11 below shows the process used to accomplish this.

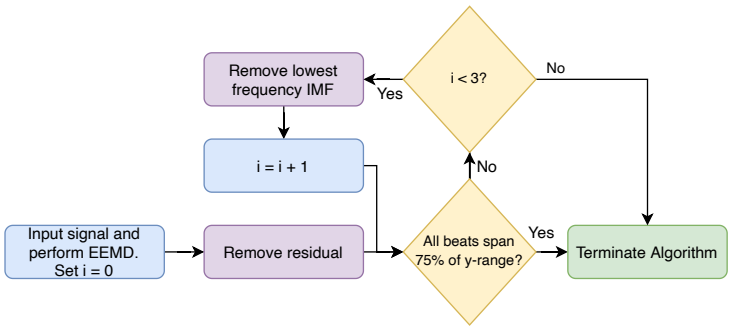
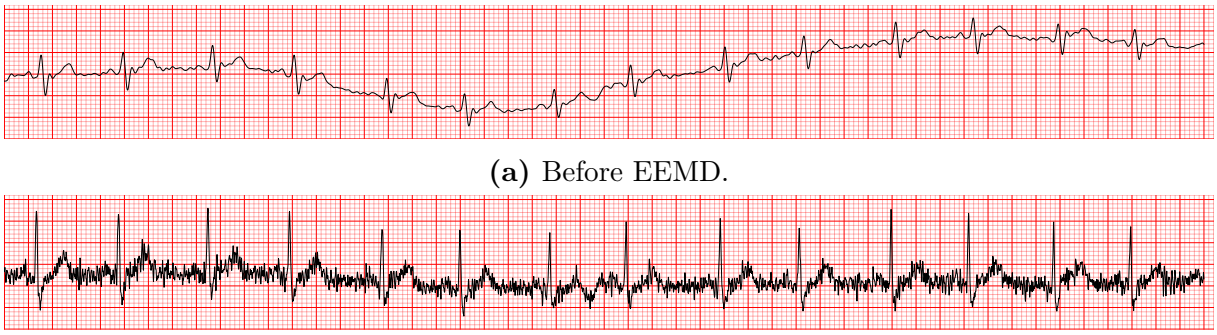


Figure 2.11: A flowchart showing the baseline wander removal algorithm used.

This process allowed a constant baseline to be established, at the cost of adding a small amount of Gaussian white noise. The number of trials for EEMD had to be low to make the process less computationally intense, but with more trials the amount of Gaussian white noise added can be reduced. The effect of EEMD on the signal can be seen in Figure 2.12.



(a) Before EEMD.

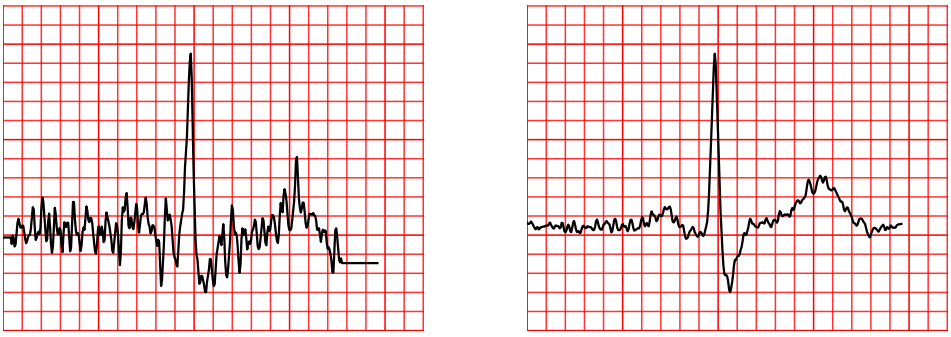
(b) After EEMD.

Figure 2.12: A figure showing the full ECG of Biobank patient 3035182 [3] before and after ensemble empirical mode decomposition was used to remove baseline wander. A flatter baseline as well as the increased Gaussian white noise can be seen in Figure 2.12b, compared to Figure 2.12a.

Mean Beat Extraction

It was required to obtain a representative beat with minimal noise, to be used with feature extraction. As such, a mean beat was extracted for each ECG recording to reduce the effect of additive Gaussian white noise after baseline wander removal.

The Hamilton-Tompkins [10] algorithm was used to detect the R peaks, after which the beats were separated by slicing the trace midway between the R Peaks detected. This point was chosen because the PR and RT (the intervals lying between consecutive features seen in Figure 2.1) segments of the ECG both span less than half of the RR interval, so there is no risk of cutting the beat short. If any beats are cut short the screening model used will, correctly, mark the trace as abnormal.



(a) Beat 6. **(b)** Mean beat.

Figure 2.13: A figure showing the 6th and mean beats from the ECG of Biobank patient 3035182 [3] after ensemble empirical mode decomposition was used to remove baseline wander. The Gaussian white noise observed is reduced when comparing Figure 2.13b to Figure 2.13a. There were not enough beats in the signal to remove all of the Gaussian white noise, but the effect that taking the mean beat has on the noise is still clear, nonetheless.

Each beat was padded by repeating its initial and final value to ensure the beats were all an equal length, and that the sample corresponding to the R peak was exactly in the middle of the set of samples. The mean value of each sample point was then calculated over all of the beats, in order to reduce the effect of additive Gaussian noise due to the EMG noise and EEMD. The use of a mean beat allowed the P and T waves to be more reliably detected in the presence of high EMG noise, as the PR and RT intervals tend not to vary during short ECG signals [11].

The overall effect of the noise removal on the signal can be shown mathematically as follows:

Let $x_{i,j}$ represent the ith sample of the jth beat of the true ECG signal, and $d_{i,j}$ represent similarly represent the additive noise due to the EMG and EEMD. This means that the noisy signal can be written as:

$$y_{i,j} = x_{i,j} + d_{i,j}. \tag{2.5}$$

We know that EMG noise and the noise added by EEMG can both be modelled as additive Gaussian white noise, so $d_{i,j} \sim \mathcal{N}(0, \sigma^2)$. The value of the σ depends on the strength of the noise-inducing muscle contractions, the skin's conductivity and the EEMD implementation. This means that the expectation of y (as seen in Equation 2.5) over the set of beats J can be taken to give:

$$\begin{aligned} \bar{y}_i &= \mathbb{E}\left[y_{i,j} \mid j \in J\right] \\ &= \mathbb{E}\left[x_{i,j} + d_{i,j} \mid j \in J\right] \\ &= \bar{x}_i + \mathbb{E}\left[d_{i,j} \mid j \in J\right] \\ &= \bar{x}_i + \sum_{j \in J} d_{i,j}, \end{aligned} \tag{2.6}$$

where $\bar{x}_i$ is the mean value of the ith sample over all of the beats of the true ECG and $\bar{y}_i$ similarly represents the noisy ECG. If $|J|$ approaches infinity, the weak law of large numbers can be used along with the zero-mean nature of the noise to give:

$$\lim_{|J| \to \infty} \left\{P\left(|\bar{y}_i - \bar{x}_i| > \epsilon\right)\right\} = 0, \forall \epsilon > 0. \tag{2.7}$$

Equation 2.7 shows that, if there is a sufficiently large number of beats, the noise in the signal processed becomes arbitrarily small when using this noise removal algorithm. As the Biobank ECG samples are only 10 seconds long (and hence on average contain 6 to 10 beats), the number of beats is not large enough to remove all noise. However, the noise removal still significantly improves the detection of medical features.

2.5.3 Effect of Noise Removal

In order to demonstrate the effect of noise removal on the screening process, a random forest model (see Chapter 4 for a description of the data and features used) was trained and tested with and without noise removal. The improvement in the performance of the model can be seen below in Figure 2.14, as the area under the receiver operating curve is increased after noise removal (see Section 2.3.2).

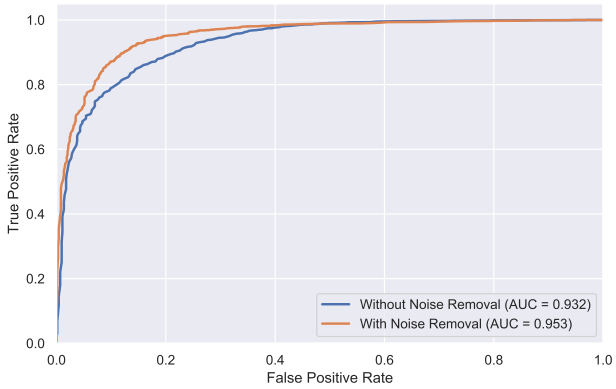


Figure 2.14: ROC curves showing the performance of a Random Forest model with and without noise removal.

Chapter 3: Use of Neural Networks when Screening For Atrial Fibrillation and Other Arrhythmias

In this section, a variety of convolutional neural network and long short term memory models were used to classify the ECG signals from the PhysioNet CinC Challenge 2017 dataset [1]. The data was labelled as '*normal*', '*atrial fibrillation*', '*other arrhythmia*' and '*noisy*', allowing the development of a screening system for atrial fibrillation and a selection of other arryhthmias. A lack of more specific labels (or documentation) makes it difficult to analyse which specific conditions can be accurately diagnosed, but the dataset does allow a 'normal'/'abnormal' screening system to be developed for the conditions present.

3.1 Data

The dataset consists of 8528 Lead I ECG recordings lasting between 30 and 60 seconds, and labelled as '*normal*', '*atrial fibrillation*', '*other arrhythmia*' or '*noisy*'. The number of samples with each diagnosis is shown in Table 3.1 below.

[1] As the test data was never released by PhysioNet, the models produced in this project cannot be compared to the challenge submissions. The performance of the best model in this project when tested, however, suggests that the models produced are comparable to the top-performing entries.

Table 3.1: A table showing the number of samples for each condition in the Physionet CinC Challenge 2017 training dataset.

Label	Number of Samples
Normal	5154
Atrial Fibrillation	771
Other Arrhythmia	2557
Noisy	46
Total	8528

The recordings in this dataset were taken using a KardiaMobile device and were transmitted to a smartphone using a microphone, after which they were bandpass filtered. This process produces a lot of additive noise, and smooths the medical features that are used to diagnose arrhythmias.

3.2 Approach

The noise and bandpass filtering of the signal makes manual feature extraction incredibly difficult, hence reducing the effectiveness of classification methods that require this (such as random forests). The large dataset volume allowed deep neural networks to be used to perform classification, as there were enough training dataset points to train the classifier without overfitting [12].

Initially 3 different CNN architectures were used, allowing the ECG to be considered as a single 1-dimensional image. These architectures had varying depths and layer hyperparameters, which allowed the analysis of the effect that this had on the performance of the network.

Following this, 4 architectures that used LSTM layers were evaluated. This enabled the ECG to be considered as a time series of 1-dimensional 'beat images'. These 4 architectures also had varying depths, but some also provided the RR interval as an input alongside the series of beats, allowing the analysis of the effect of explicitly providing expert features to the model.

A 80/10/20 train/validation/test split was used to determine the best-performing CNN and LSTM architectures. Each model was trained, with validation data being used after each epoch in order to ensure that the model did not overfit. The records used were '*normal*', '*atrial fibrillation*' and '*other arrhythmia*' (as seen in Table 3.1), so a categorical cross-entropy loss function was used when training (see Section 3.3.3). Noisy readings were ignored, as their detection is outside the scope of this project. The Keras '*Adam*' optimiser [13] was used to optimise parameters, to take advantage of its per-parameter learning rate, and gradient history memory. The models' performance was measured each epoch using

the mean area under the ROC curve over all of the labels (see Section 2.3.3). The models were subsequently tested and the ROC curves plotted for both '*atrial fibrillation*' and '*other arrhythmia*' labels.

3.3 Theoretical Background

Neural networks are designed to model abstractions in data using processing layers of artificial neurons, with each neuron transmitting a function of the signals it receives to subsequent neurons. Combining artificial neurons in neural networks allows non linear functions to be approximated, such as a function to map raw time series data (for example a sampled ECG signal) to a set of labels. The ability of neural networks to approximate these functions without any explicitly defined feature extraction makes them a useful tool when performing classification and regression tasks. Their performance is greatly influenced by the volume of training data available, the extent to which the training data is representative of the test data, and the architecture of the neural network. The more neurons added to a neural network, the more complex the functions that it is able to approximate are. However, more training data and computing power are required to exploit this.

3.3.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of deep neural network (neural networks with many hidden layers between their input and output layer). They are composed of a series of convolutional layers (often with a non-linear function applied to their outputs), followed by one or more fully connected layers (which may also have non-linear functions applied to their output). Although they are generally used for image classification, CNNs are becoming more widely adopted in a diverse range of fields. A range of layers are used adapt the network, in order to tailor it to its particular application. For the purposes of this report only 1 dimensional CNNs will be considered, as ECG signals are 1 dimensional.

Convolutional Layers

Convolutional layers are the most common layers in convolutional neural networks. The constituent neurons output a function of the weighted sum of input signals (as seen in Equation 3.1). The functions used ($\phi(\cdot)$) are known as activation functions, and allow neurons to exhibit non-linear behaviour. In this project, rectified linear unit (ReLU) functions are used as convolutional layer activation functions ($\phi(x) = \max(0, x)$) to avoid vanishing gradients when training, and to avoid the computationally expensive operations used in other activation functions.

$$y(\mathbf{x}) = \phi\left(\sum_{i=1}^{N} w_{N-i} x_{j+i} + b\right) \tag{3.1}$$

Convolutional layers are translationally equivariant, so information about relative position is preserved. This is important when working with data for which the relative position of features is important, such as ECG signal.

The kernel size of a layer refers to the size of the filter that is convolved over the input elements, whilst the stride refers to how far the filter moves between each step convolutions. The number of kernels determines the number of filters used, and hence the dimensionality of the vector output by the neuron. The kernel size, stride and number of kernels of a convolutional layer can be varied.

Fully Connected Layers

Fully connected layers effectively perform a weighted sum of inputs, and add a bias term. Their general formula is given by Equation 3.2. Due to their less general form, they have more parameters than that of an equivalent convolutional layer.

$$y_j(\mathbf{x}) = \phi\left(\mathbf{w}_j^T\mathbf{x} + b_j\right) \tag{3.2}$$

Fully connected layers are used with a softmax activation function ($\sigma(\cdot)$), see Equation 3.3) as the last layer in a neural network when performing classification. This produces a probability distribution over the number of classes, K.

$$\sigma_j(z) = \frac{\exp(z_j)}{\sum_{i=1}^{K}\exp z_i} \tag{3.3}$$

Max Pool Layers

Max pool layers are used to introduce local translational invariance in the network [14], and to decrease computational complexity. They effectively subsample the signal, by taking the maximum value within a set of points. The use of max pool layers hence ensures that both high and low frequency signal components are considered [15]. The kernel size determines the size of the set of points considered, and the stride varies how separated the sets of points are (and hence whether there is any overlap).

Dropout Layers

Dropout layers are used to prevent overfitting, hence maintaining generality in the neural network. This is done by randomly setting the edges between layers in a neural network to zero, thus forcing redundancy to be developed when training the network. Dropout layers are often added between fully connected layers as they tend to have the greatest number of parameters and are hence more prone to overfitting. However, literature suggests that

they can be effective when used between convolutional layers as well (by effectively adding noise to the network when training) [16].

Batch Normalisation Layers

Batch normalisation standardises the input to a network or layer in the network for each mini batch (a subdivision of the data) when training. It is widely acknowledged to improve the performance and stability of neural networks, but a consensus as to why this is the case has not been reached [17, 18].

Skip Connections

Skip connections are a key feature of ResNet architectures. They ensure that adding more layers to a neural network does not degrade its accuracy (as skipped layers can be reduced to an identity layer if contributing to the degradation), whilst speeding up learning by eliminating singularities (resulting from the linear dependence, permutation symmetry and consistent deactivation of nodes) [19].

3.3.2 Long Short-Term Memory Networks

Long short-term memory (LSTM) networks are recurrent neural network architectures with memory cells capable of 'storing' data. This allows them to process long sequences of data, so they can be be applied to ECG classification, by separating the ECG signal into a time series of beats. Hidden layers can be used between the input and LSTM layers in a neural network to effectively perform feature extraction. By doing this, the LSTM is forced to consider the ECG signal as a cardiologist would: by analysing the features in a series of beats, relative to the features shown in the rest of the beats.

3.3.3 Loss Functions

Loss functions quantify the prediction error of a neural network. They are differentiable, to allow the use of gradient-based optimisation methods when training the network. This project will use the categorical cross-entropy function, as the model developed is for single-label classification. The formula for this can be seen in Equation 3.4 below, where $\hat{y}$ is the output of the CNN, y is the ground truth, M is the number of training points and N is the number of classes.

$$L(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N} y_{ij} \log \hat{y}_{ij} \tag{3.4}$$

3.4 Screening Using Convolutional Neural Networks

In order to use the ECG signals as convolutional neural network inputs, the signals were all resampled at 250Hz and divided into 10 second segments. Any segments that were less than 10 seconds long were discarded. Ensemble empirical mode decomposition (see

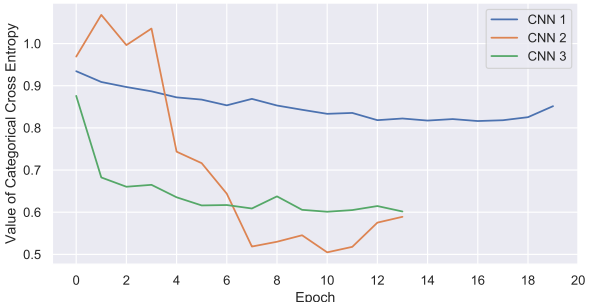
Section 2.5.2) was used to remove any baseline wander and DC bias in the signal before it was used as an input for the CNN.

3.4.1 Architectures Used

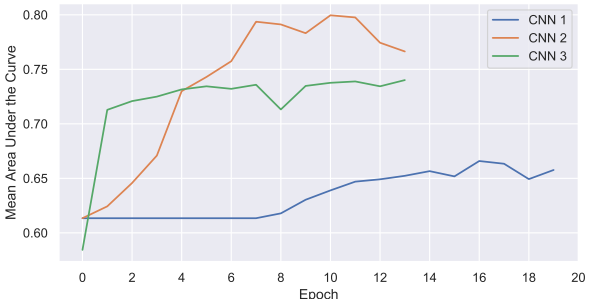
Figure 3.1 on page 18 shows the overall structure of the 3 CNN architectures used. CNN Model 1 in 3.1a is a shallow CNN with just 2 convolutional layers, whilst Model 2 in Figure 3.1b has a much deeper ResNet architecture [20], which makes use of multiple max pool layers to effectively repeatedly subsample the input. This allows the use of deeper neural networks to be evaluated when classifying ECGs, with the max pool layers allowing both higher and lower frequency features to be identified using identically sized convolutional kernels. Lastly, Model 3 in Figure 3.1c uses varying strides and kernel sizes to identify different frequency features, allowing an alternative approach to the use of max pooling to be evaluated (the exact values of the hyperparameters used can be found in the original paper by Yıldırım et al. [21]).

3.4.2 Results and Discussion

The value of the categorical cross-entopy function, and the value of the mean AUC (over all of the target labels) after each epoch for each CNN model are plotted below in Figure 3.2.



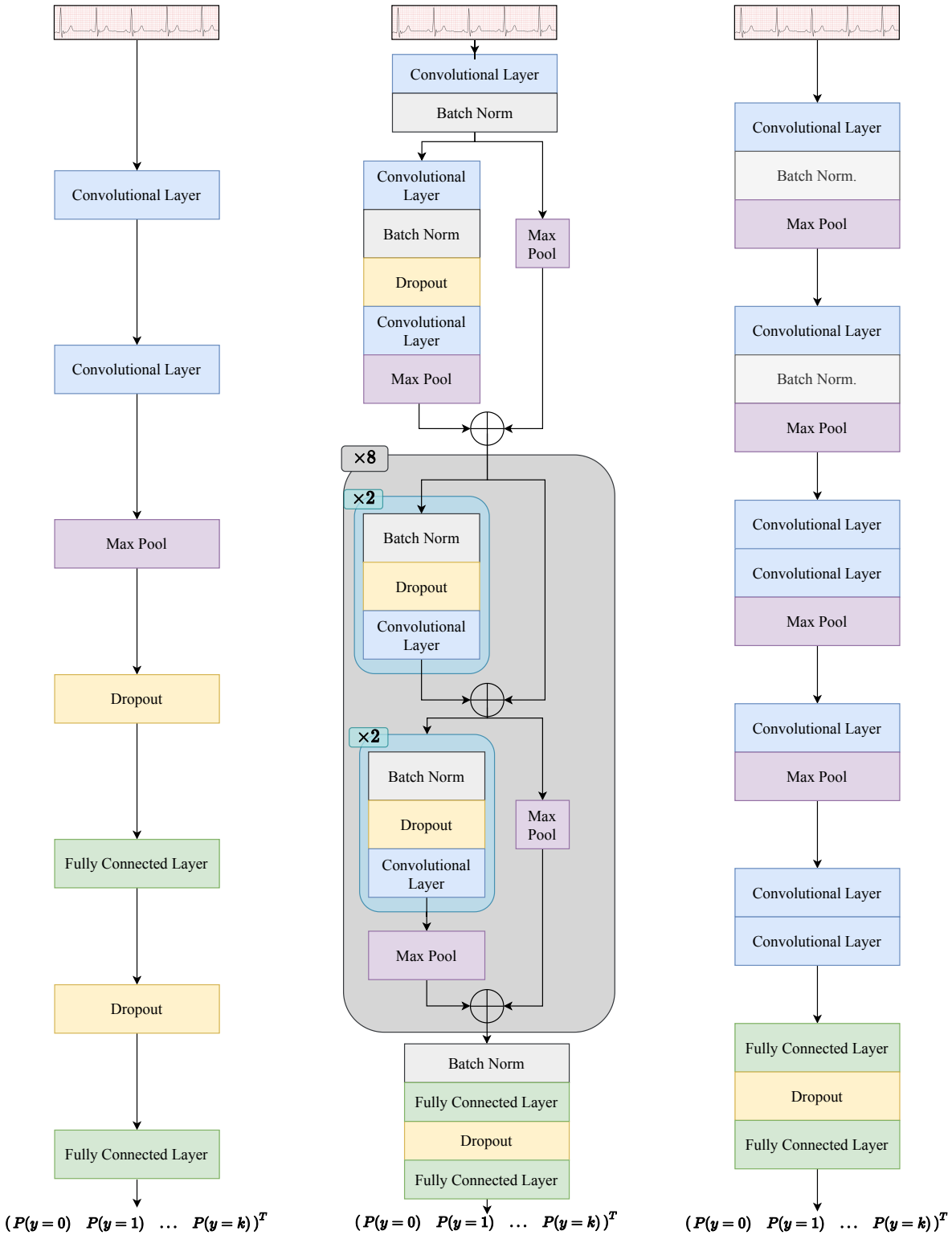
(a) Categorical cross entropy per epoch when training.



(b) Mean AUC per epoch when training.

Figure 3.2: A figure showing the value of the categorical cross entropy and the mean AUC at the end of each epoch when training each of the CNN models.

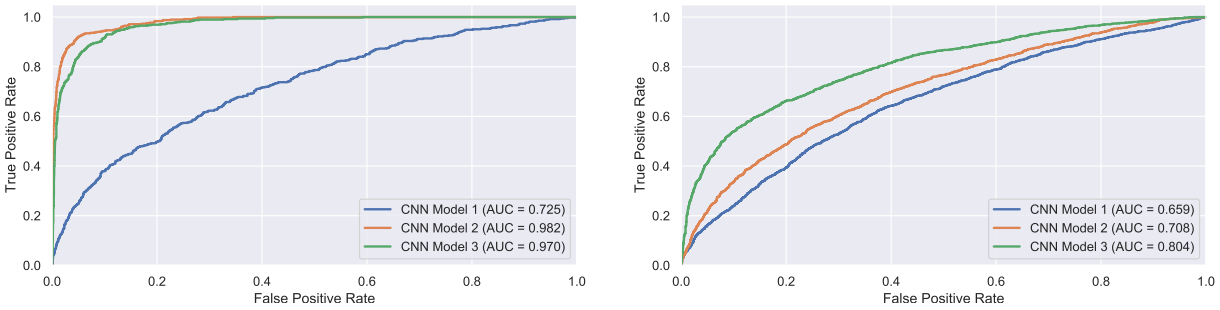
Looking at Figure 3.2, it can be seen that CNN Model 1 (the generalised, shallow network) takes the greatest number of epochs to be optimised and achieves the worst performance. Models 2 and 3 both achieve their minimum cross entropy after 10 epochs, with Model 2 (the deep ResNet model) performing slightly better than Model 3 (the dilational neural network) after 6 epochs when classifying the validation data.



(a) Model 1. **(b)** Model 2. **(c)** Model 3.

Figure 3.1: The 3 CNN models used.

Figure 3.3 shows the receiver operating characteristics for each model.



(a) ROC for '*atrial fibrillation*' traces. **(b)** ROC for '*other arrhythmia*' traces.

Figure 3.3: A figure showing the ROC curves for each of the CNN models when performing a 'normal'/'abnormal' classification.

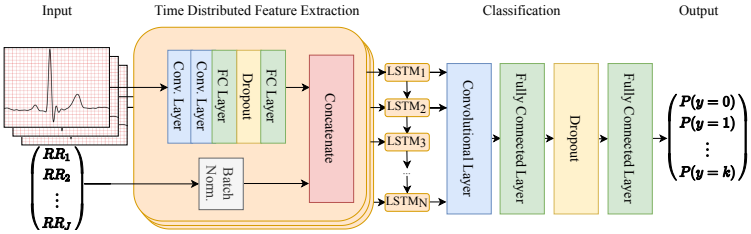
Figure 3.3a shows that CNN Models 2 and 3 perform far better on the atrial fibrillation records in the test data than Model 1, with Model 2 performing best. This indicates that deeper models, which force the consideration of both high and low frequency components, perform the classification of atrial fibrillation better than simple models. Figure 3.3b shows that CNN Model 3 performs the best by far when classifying 'other' traces, whilst Model 2 performs worse and Model 1 worse still. Model 2 may be overfitting, as the diverse set of other arrhythmias are far harder to fit with any generality given the small volume of data provided, whilst Model 1 is likely to be underfitting.

3.5 Screening Using LSTM Networks

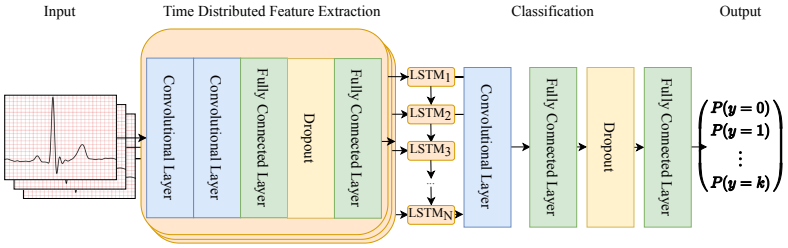
In order to use the ECG signals with LSTM architectures, each signal was separated into its constituent beats. The Hamilton-Tompkins algorithm [10] was used to detect R peaks, and by subsequently slicing the signal midway between these R peaks. Each beat was then resampled to be 256 samples long in order to make each beat a consistent length. A tensor of ECG beat samples was used as the input for the networks.

3.5.1 Architectures Used

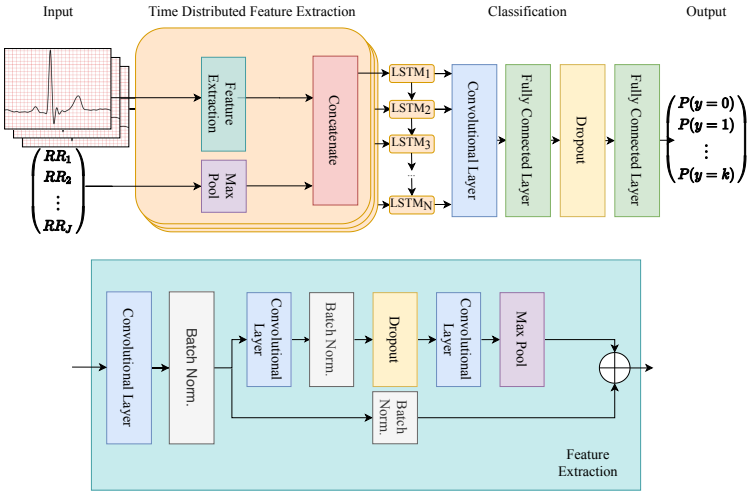
Figure 3.4 shows the 4 LSTM architectures used. The model in Figure 3.4a shows a simple LSTM model with shallow, convolutional 'feature extraction' and RR intervals input, whilst the model in Figure 3.4b does the same without RR intervals input and the model in Figure 3.4c uses a deeper 'feature extraction'. The model in Figure 3.4d allows the use of Max Pool layers in the feature extraction stage to be evaluated. Comparison of the first two architectures allows the evaluation of the network's ability to classify based on the shape of the ECG beats, without knowledge of the RR intervals. Meanwhile, comparison of the first and third architectures allows the evaluation of the performance of deeper, more complex LSTM-based models. Comparison of the first and fourth models allows the analysis of the effect of the use of max pool layers on the model's performance.



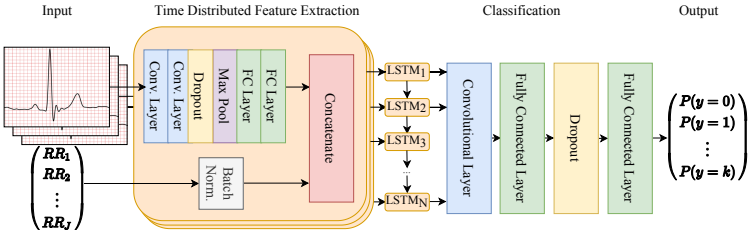
(a) LSTM Architecture 1.



(b) LSTM Architecture 2.



(c) LSTM Architecture 3.

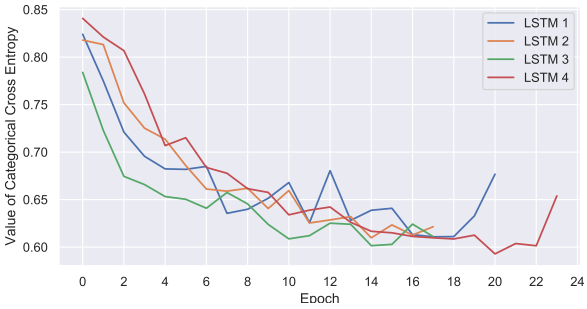


(d) LSTM Architecture 4.

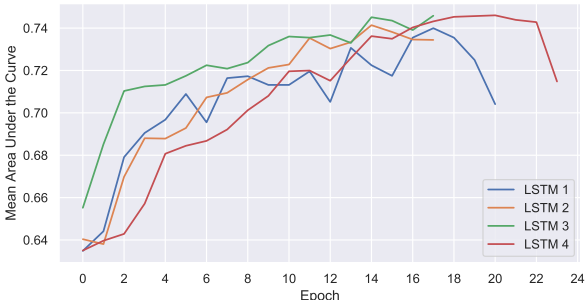
Figure 3.4: The 4 LSTM architectures used to classify the ECG traces.

3.5.2 Results and Discussion

The value of the categorical cross-entopy function, and the value of the mean AUC (over all of the target labels) after each epoch for each LSTM model are plotted in Figure 3.5.



(a) Categorical cross entropy per epoch when training the LSTM models.

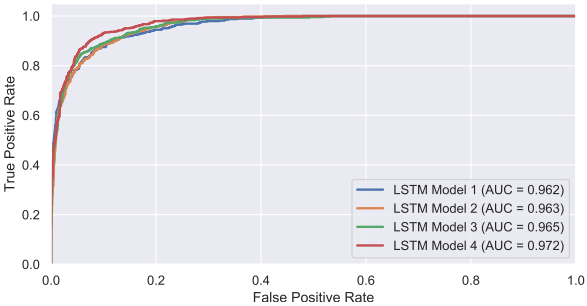


(b) Mean AUC per epoch when training the LSTM models.

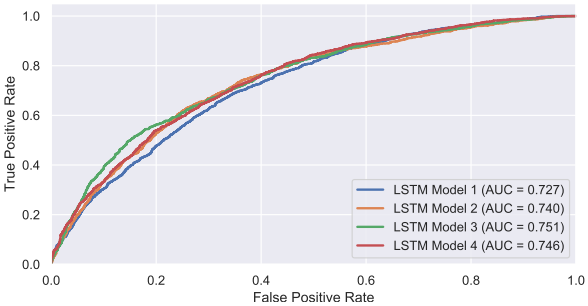
Figure 3.5: A figure showing the value of the categorical cross entropy and the mean AUC at the end of each epoch when training the LSTM models.

Figure 3.5 above shows that the 4 LSTM models have fairly similar categorical cross entropy and AUC values when training. The training of LSTM Models 2 and 3 finishes before that of Models 1 and 4. A comparison between Models 1 and 2 indicates that the input of RR intervals alongside the beats results in a slower rate of convergence. Comparing Models 1 and 4 suggests that the inclusion of a max pool layer results in a small increase in time for the model to converge when training (despite the resultant decrease in the number of parameters). The results for LSTM Model 3 show that a small increase in the depth of the network does not mean an increase in the number of epochs required for the model to converge when training.

Figure 3.6 shows the receiver operating characteristics for each LSTM model when testing.



(a) ROC for '*atrial fibrillation*' traces.



(b) ROC for '*abnormal*' traces.

Figure 3.6: A figure showing the ROC curves for each of the LSTM models when performing a 'normal'/'abnormal' classification.

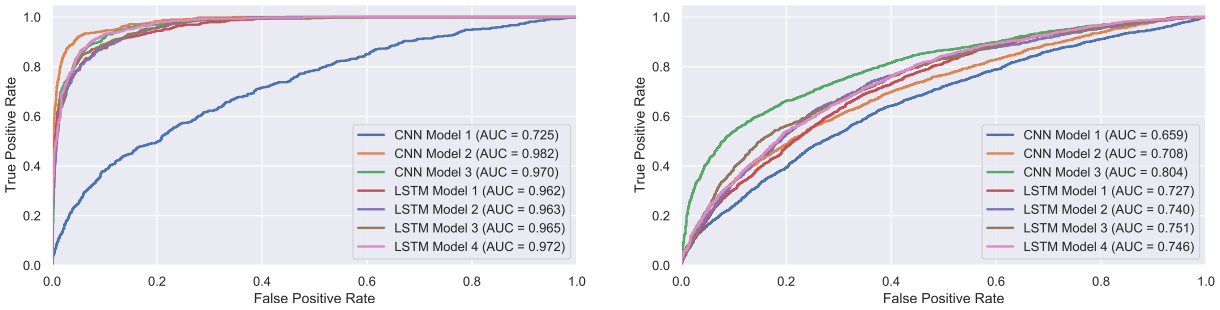
It can be seen in Figure 3.6a that, although all of the LSTM models have very similar ROC curves, LSTM Model 4 performs the best. This is likely due to the local translational invariance introduced by the max pool layer, which avoids overfitting based on

the small changes in the intervals between ECG features. The slight improvement in the performance of Model 2 compared to that of Model 1 suggests that the explicit provision of RR intervals is detrimental to performance, rather than beneficial.

Figure 3.6b shows that Model 3 performs the best when classifying abnormal traces, suggesting that a deeper network (allowing more complex feature extraction) is useful when detecting a wider range of conditions. The poor peformance of Model 2 shows that the explicit provision of RR intervals is detrimental to the performance of the network, whilst Model 4's improved performance when compare to that of Model 1 shows that the use of max pool layers to remove the translational equivariance is beneficial.

3.6 Model Comparison

The receiver operating characteristic curves for all of the neural network models for both '*atrial fibrillation*' and '*other arrhythmia*' are shown below in Figure 3.7.



(a) ROC for '*atrial fibrillation*' traces. **(b)** ROC for '*other arrhythmia*' traces.

Figure 3.7: A figure showing the ROC curves for each of the neural network models by label when performing a 'normal'/'abnormal' classification.

The results show that although all of the neural networks performed similarly well when detecting atrial fibrillation, their performance was far worse when detecting records containing other arrhythmias. This is likely to be due to the variety of conditions considered to be other arrhythmias; with a relatively small volume of data per condition, it is difficult to avoid underfitting models. The lack of specific labels for each arrhythmia in the dataset means that the model's performance cannot be analysed for each arrhythmia (which would make the models' flaws to clearer), but the results suggest that atrial fibrillation can be easily screened for, even in noisy datasets, using neural networks. When screening for other arrhythmias CNN Model 3 is by far the best-performing model, but to achieve just an 80% true positive rate, a 40% false positive rate must be tolerated. This performance is far too poor for a viable screening method, so a further investigation using more techniques and datasets is important.

When simplifying the diagnosis to a 'normal'/'abnormal' (where abnormal represents atrial fibrillation or other arrhythmia) diagnosis, the ROC curves in Figure 3.8 below are produced (using the same method as before).

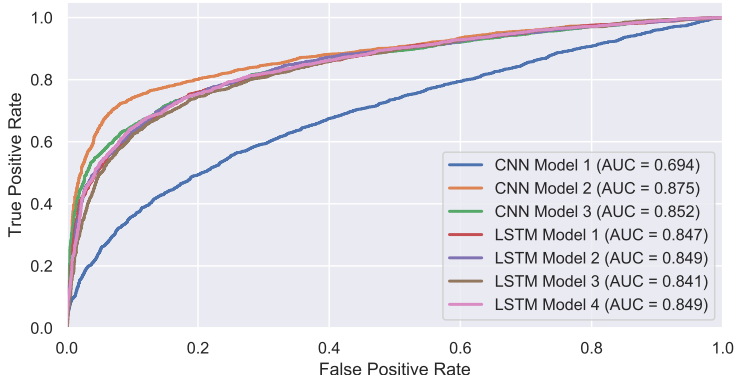


Figure 3.8: A figure showing the ROC curves for each of the neural network models when performing a 'normal'/'abnormal' classification.

Figure 3.8 shows that CNN Model 2 performs best by far, whilst CNN Model 1 performs worst. The performances of the rest of the models are fairly similar. This suggests that, although CNN Model 3 is more suitable for a more specific screening system, CNN Model 2 is the best model for binary screening systems. As such, CNN Model 2 is most suitable for the development of a binary 'normal'/'abnormal' screening system.

Chapter 4: Use of Machine Learning Techniques when Screening for a Variety of Cardiac Arrhythmias

This chapter details the work done using the PhysioNet Challenge 2020 dataset; this ranges from the development of a random forest model to the testing of a CNN model trained in Chapter 3. The dataset was chosen as it provides specific labels for each arrhythmia present in the traces, allowing a more specific screening system to be produced. It also enabled the performance of a binary 'normal'/'abnormal' screening system to be assessed for each condition.

4.1 Data

The Physionet CinC Challenge 2020 dataset consists of 4154 12 Lead ECG recordings. As this project's primary goal is the development of an ECG screening method, only the first lead was used when using this dataset. Each sample was labelled as '*normal*', or with one or more of a set of overlapping diagnoses. These, along with the number of instances of each diagnosis are shown in Table 4.1.

Table 4.1: A table showing the number of samples for each condition in the Physionet CinC Challenge 2020 training dataset (N.B. Due to the overlapping diagnoses, the **Number of Samples** column does not sum to the total number of records).

Label	**Number of Samples**
Normal	918
Atrial Fibrillation	1221
Atrioventricular Block	722
Premature Ventricular Complexes	700
Premature Atrial Complexes	616
Total	4154

4.2 Approach

Due to the small dataset size, deep learning methods similar to those used in Chapter 3 were deemed to be unsuitable. This is because the number of parameters in the simplest model used ($\sim 7.96 \times 10^6$) was many orders of magnitude greater than the number of labelled records, so the model would be prone to overfitting.

In order to replicate a cardiologist's performance on this dataset, a random forest was chosen to be the best model. This is because both the cardiologist's process and the implementation of a random forest model involve feature extraction and a subsequent decision making process. Random forests have been used in literature on a variety of other datasets [22, 23], so the features and method used in these papers were taken as a starting point when implementing the model.

The random forest used in this section performs a binary classification, with the output labelling whether the input ECG is *normal* or *abnormal* (defined as showing evidence of one or more of the overlapping diagnoses), using a set of features extracted from the ECG.

4.3 Theoretical Background

Random Forest (RF) classifiers are an ensemble method that output the empirical distribution of the labels predicted by a parallel set of decision trees. They allow the information about the classification provided by a variety of features to be combined when performing the decision-making process. Individual decision trees are prone to overfitting when aiming to achieve high accuracy classification, because increasing the depth of decision trees reduces bias whilst increasing the variance. RF classifiers avoid this problem by taking the modal classification given by a set of decision trees generated in parallel. Using the central limit theorem (as many decision trees are combined in the random forest) [24], it can be shown that the expected decision error will tend to the irreducible error as long as the bias is already arbitrarily small and the variance tends to 0. This can be seen in

Equation 4.1, where N (the number of trees) goes to infinity, $\hat{f}(x)$ is the output of the RF, $\hat{f}_i(x)$ is the output of a single decision tree in the RF and σ is the irreducible error.

$$\begin{aligned}\mathbb{E}\left[\left(y-\hat{f}(x)\right)^2\right] &= \text{Bias}(\hat{f}(x))^2 + \text{Var}(\hat{f}(x)) + \sigma^2 \\ &= \lim_{N\to\infty}\frac{1}{N}\text{Var}(\hat{f}_i(x)) + \sigma^2 \\ &= \sigma^2\end{aligned} \tag{4.1}$$

When using RF classifiers, the input features used have a large influence on the performance of the classifier. The information about the labels provided by the features must be maximised whilst avoiding features with high correlation, in order to achieve the best possible performance. Provision of suitable input features can be achieved through intelligent feature extraction (using medical recommendations alongside features suggested in literature) and by subsequently using feature selection to remove correlated features.

4.4 Feature Extraction

After using the noise removal detailed in Section 2.5, a variety of features were extracted from the ECG. The importance of each feature was then quantified using the mutual information between the feature and the label, allowing the success of the feature extraction to be verified. Feature selection was subsequently used to maximise the performance of the random forest.

4.4.1 Feature Extraction Methods

Hamilton-Tompkins

The Hamilton-Tompkins algorithm [10] can be used to detect R peaks in an ECG signal. It involves filtering the ECG, followed by differentiating and squaring the signal to highlight the R peaks (because R peaks are the sharpest prominent peaks in the signal, so they have the largest gradient even after the filtering). After this, the energy in each peak is evaluated and thresholded to ensure that the peak is an R peak.

Medical Feature Detection

The detection of P, Q, S and T wave peaks was achieved by detecting all the relevant turning points (maxima for P and T, minima for Q and S) in the mean beat (see Section 2.5.2) within a given range of the R peak. The turning point with the greatest topographic prominence ("the minimum vertical distance one must descend from a peak in order to climb a higher peak" [25], see Figure 4.1) was then determined to be the wave peak. This method relies on adequate noise removal ensuring that the magnitude of the noise peaks does not exceed that of the features to be detected.

If the P or T wave are not detected, the R peak is used as the P or T wave location. Similarly, if the Q or S wave are not detected, an arbitrary point 0.06 seconds in the correct direction (a point just outside the 'normal' range) is used instead.

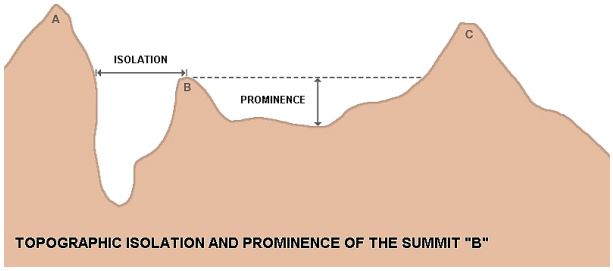


Figure 4.1: A diagram showing the topographic prominence of the central peak [26].

Trapezium Rule

The approximation of an integral by splitting the area into a series of trapezia.

Baseline Detection

The baseline was taken to be the mean of the start and end samples of the mean beat after noise removal.

Beat Variance

When taking the mean beat, there is an associated variance which allows the consistency of the beats to be quantified. By taking the integral of this, the effect of beat variation and noise on the mean beat can be considered. This integral can be evaluated over 3 separate sections, as shown in Figure 4.2.

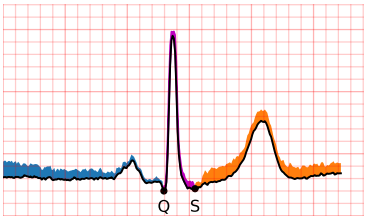


Figure 4.2: A plot of the mean beat of Biobank patient 3934036 [3], showing the beat variance split in to 3 sections (blue, magenta and orange).

Instantaneous Heart Rate (IHR)

The Instantaneous Heart Rate is calculated by applying a simple moving average filter with $n = 3$ to $\frac{60 \times \texttt{sample rate}}{\texttt{RR Interval}}$.

Discrete Wavelet Transform (DWT) Analysis

A multi-level discrete wavelet transform is applied to the ECG using a Daubechies 4-tap wavelet, allowing the signal to effectively be filtered using different frequency bands. This was used to extract frequency domain features, using an approach similar to Alickovic and Subasi [22]. The mean value of each of the DWT outputs was evaluated, as well as the mean of the square of the outputs. These, alongside the ratio between the maximum amplitudes of neighbouring outputs were used as input features for the Random Forest.

4.4.2 Feature Descriptions

A description of the random forest features extracted, as well as the method used to detect them, are given in Table 4.2. Methods are described in full in Section 4.4.1.

Table 4.2: A table detailing the random forest features used in this project, and their extraction methods.

Number	Name	Description	Method
1	RR Mean	Mean interval between R peaks.	Hamilton-Tompkins
2	RR Variance	Variance of interval between R peaks.	Hamilton-Tompkins
3	Normalised RR Variance	$\frac{\text{RR Mean}}{\text{RR Variance}}$	Hamilton-Tompkins
4	PR interval	Interval between P and R peaks.	Medical Feature Detection
5	RT interval	Interval between R and T peaks.	Medical Feature detection
6	R:P	Ratio between R and P peaks.	Feature detection
7	QS interval	Interval between Q and S peaks.	Medical Feature Detection
8	R:T	Ratio between R and T peaks.	Medical Feature detection
9	QRS area	Area under the QRS complex.	Medical Feature detection, Trapezium rule
10	Max − Baseline	The difference between the maximum value found in the beat and the baseline.	Baseline Detection
11	Baseline − Min	The difference between the baseline and the minimum value found in the beat.	Baseline Detection
12	Beat Variance 1	The area under the beat variance curve from the start of the beat to the Q peak.	Beat Variance and Trapezium Rule

continued on the next page

Number	Name	Description	Method
13	Beat Variance 2	The area under the beat variance curve from the Q peak to the S peak.	Beat Variance and Trapezium Rule
14	Beat Variance 3	The area under the beat variance curve from the S peak to the end of the beat.	Beat Variance and Trapezium Rule
15	Mean IHR	Mean instantaneous heart rate.	Instantaneous Heart Rate
16	IHR Variance	Instantaneous heart rate variance.	Instantaneous Heart Rate
17-27	Mean Beat DWT	Calculations using the Discrete Wavelet Transform with the mean beat and a depth of 4.	Discrete Wavelet Transform Analysis
17-27	Whole Trace DWT	Calculations using the Discrete Wavelet Transform with the whole trace and a depth of 6.	Discrete Wavelet Transform Analysis

Feature Significance

The feature significance (information about the label provided by each feature) is quantified using the mutual information (a measure of the decrease in uncertainty about one variable when the other is known) between the feature and the target label. The mutual information is a quantity between 0 and 1, with a score close to 1 indicating that a feature can be used as a strong estimator of the label. A scikit-learn [27] implementation of a k-nearest neighbour algorithm to estimate mutual information [28] was used. This was implemented for a binary classification between normal and abnormal, and for a binary classification between normal and each condition in the dataset, as shown in Figure 4.3.

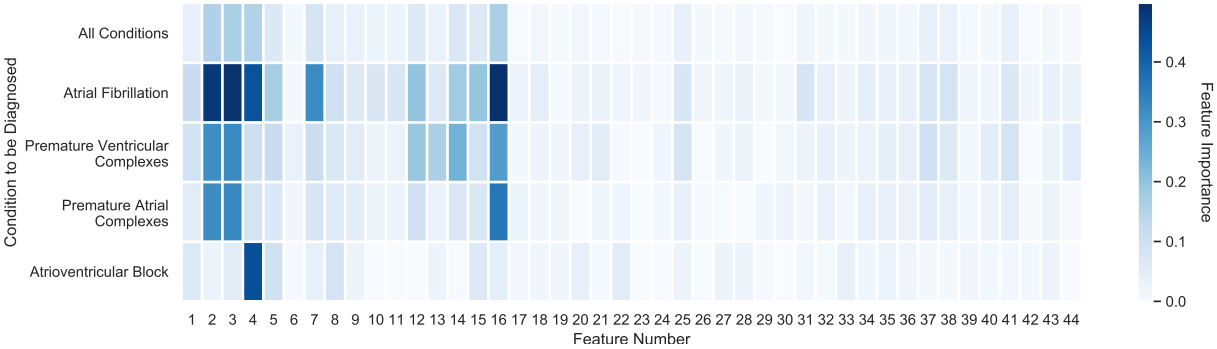


Figure 4.3: A plot showing the mutual information between feature and label for the binary classification between 'normal' and the given conditions.

Figure 4.3 shows that mutual information varies by the condition to be diagnosed. This was to be expected as each condition produces different feature variations in the ECG signal.

A large RR variance and P wave supression are indicative of **atrial fibrillation** (see Figure 4.4), so the significance of features 2, 3, 4 and 16 is higher.

Figure 4.4: A plot from the Challenge 2020 dataset showing the variable RR interval and P wave supression that characterise **atrial fibrillation**.

Premature ventricular complexes (see Figure 4.5) have an irregular RR interval and a single wider QRS complex, so features 2, 3, 12, 13, 14 and 16 are more significant.

Figure 4.5: A plot from the Challenge 2020 dataset exhibiting the irregular RR interval and irregular beat shape (seen in the QRS complex) associated with **premature ventricular complexes**.

Premature atrial complexes (see Figure 4.6) also have an irregular RR interval and a single narrower QRS complex, so features 2, 3, 12 and 16 are more significant.

Figure 4.6: A plot from the Challenge 2020 dataset exhibiting the irregular RR interval and irregular beat shape (seen in the fourth beat) associated with **premature atrial complexes**.

Atrioventricular blocks (see Figure 4.7) are characterised by extended PR intervals, so feature 4 is most significant.

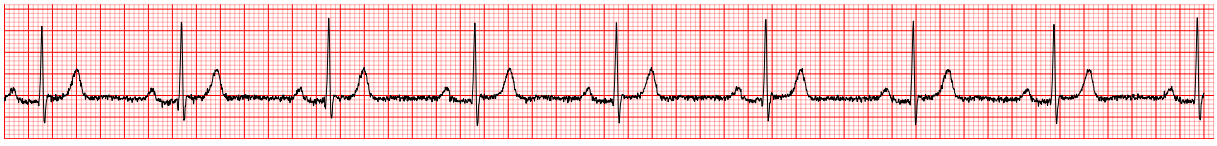


Figure 4.7: A plot from the Challenge 2020 dataset showing the subtly extended PR interval associated with **atrioventricular blocks**.

4.4.3 Feature Selection

The suggestion in literature [29] that the performance of a random forest can be improved by performing feature selection is investigated in this section. Figure 4.8 shows the effect that the number of features used has on the accuracy. This was implemented using recursive feature extraction [30], involving repeatedly training a random forest and removing the feature that was found to be least useful when training the random forest.

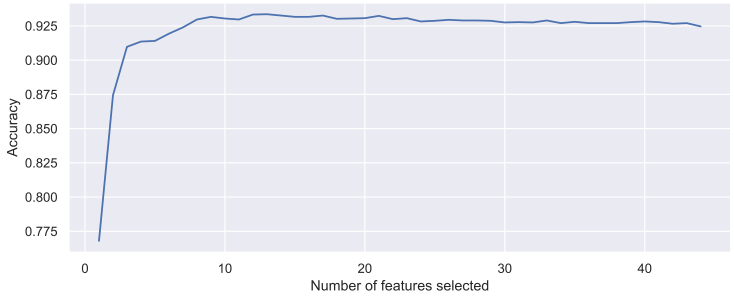


Figure 4.8: A figure showing the accuracy of the binary ('normal'/'abnormal') classification using the random forest after each step when using recursive feature extraction.

Figure 4.8 shows that the accuracy of the classifier increases as the number of features increases, from 2 to 13 features, after which the accuracy decreases slightly. Table 4.3 shows the features used by the 13 feature classifier, as well as their mutual information ranking (as seen in Figure 4.3). It can be seen that the optimal feature set does not consist of the 13 features with the highest mutual information. This is because conditional mutual information (the information about the label provided by a feature, given the value of another feature) must be considered, in order to assess which combinations of features provide the most information about the label [31].

Table 4.3: A table showing the feature, its mutual information and its mutual information rank (where a rank of 1 denotes the most important feature) for the 14 constituent features in the best performing feature set.

Feature	Mutual Information	Mutual Information Rank
RR Mean	0.0402	11
RR Variance	0.1592	4
Normalised RR Variance	0.1664	1
PR Interval	0.1607	3
R:P	0.0804	6
QS Interval	0.0830	5
R:T	0.0426	10
Beat Variance 1	0.0607	9
Beat Variance 2	0.0337	16
Beat Variance 3	0.0705	7
Mean IHR	0.0646	8
IHR Variance	0.1657	2
Mean Beat DWT Feature 6	0.0024	43

4.5 Model Training and Testing

A simple random forest binary classifier was trained to classify whether an input sample was '*normal*' or '*abnormal*'. '*Normal*' ECG signals were defined as not showing the presence of any arrhythmias (using the dataset labels provided), whilst '*abnormal*' ECG signals were oppositely defined. The features used were the 14 features identified using recursive feature elimination (see Table 4.3 above). Initially, the hyperparameters for each random forest model were optimised using a grid search and 5 fold cross validation. The model was tested (using an 80/20 Train/Test split) before and after hyperparameter optimisation, and the relevant ROC curves were plotted.

4.5.1 Hyperparameter Optimisation

Table 4.4 shows the hyperparameters used in the grid search when selecting an optimal set of hyperparameters. The scoring function for the grid search was chosen to be the area under the ROC curve (see section 2.3.3 on page 7 for more detail). The training time for the model was sufficiently short (due to the small dataset) that it could be ignored when selecting the optimal set of hyperparameters.

Table 4.4: A table showing the random forest hyperparameters used in the random search.

Hyperparameter	Value 1	Value 2	Value 3	Value 4	Value 5
Number of Estimators	500	800	1500	2500	5000
Maximum Depth	`None`	10	20	30	40
Minimum Samples Split	2	5	10	15	20
Minimum Samples Leaf	1	2	5	10	15

Table 4.5 shows the optimal hyperparameters found using recursive feature elimination. The minimum samples split, maximum samples leaf and maximum depth restriction regulate the decision trees in the random forest to ensure that they do not overfit to individual records, but still fit specifically enough to accurately classify signals. The increased number of estimators, on the other hand, allows the random forest variance to be minimised.

Table 4.5: A table showing the default and optimised hyperparameters when a grid search with 5 fold cross validation was used to improve the normal/abnormal classification using a random forest.

Hyperparameter	Default Value	Optimum Value
Number of Estimators	100	500
Maximum Depth	`None`	20
Minimum Samples Split	2	2
Minimum Samples Leaf	1	2

The ROC curves in Figure 4.9 show the results when the models were tested with an 80/20 Train/Test split before and after hyperparameter optimisation. The optimisation of hyperparameters slightly increases the AUC by increasing the true positive rate slightly when the false positive rate is low.

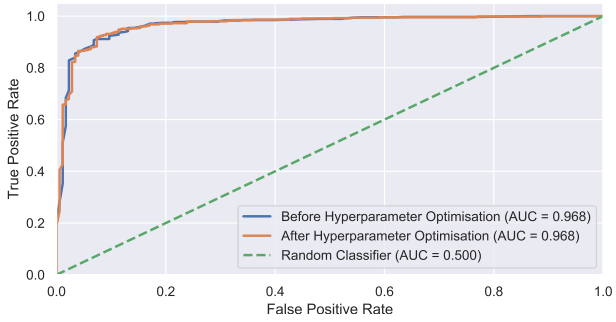


Figure 4.9: A figure showing the effect of hyperparameter optimisation on the ROC curve.

4.6 Multi-Class Diagnosis

The model was subsequently adapted to perform a separate binary classification for each condition. This involved the training of 4 parallel random forest models, each of which provides a binary label to indicate whether an input signal shows the presence of a specific condition (atrial fibrillation, premature ventricular complexes, prematue atrial complexes and atrioventricular block). The model is illustrated in Figure 4.10.

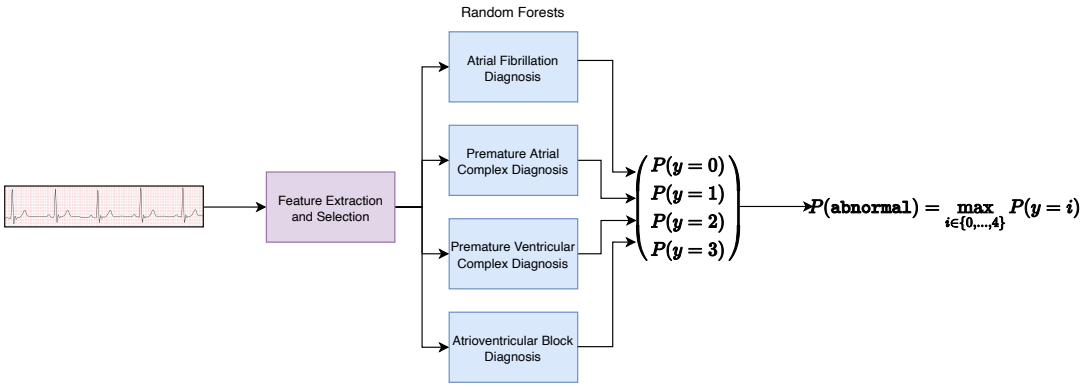


Figure 4.10: A diagram of the model used for multi-class diagnosis, composed of 4 parallel random forests.

To train and test this model, the hyperparameters for each of the random forest models were optimised using a grid search and 5-fold cross validation in order to maximise the AUC. The hyperparameter values used for the grid search are given in Table 4.4 on Page 31, and the optimal hyperparameters found are given in Table 4.6. The hyperparameters differ, as conditions for which the input features have a reduced mutual importance will need deeper decision trees, and increased feature variance for particular conditions must be offset by increasing the number of estimators.

Table 4.6: A table showing the default and optimised hyperparameters when a grid search was used to improve classification of Atrial Fibrillation (AF), Premature Atrial Complexes (PAC), Premature Ventricular Complexes (PVC) and AV Block (AVB).

Hyperparameter	AF	PAC	PVC	AVB
Number of Estimators	800	500	500	1500
Maximum Depth	30	40	40	`None`
Minimum Samples Split	2	2	2	2
Minimum Samples Leaf	1	1	1	1

Following the above hyperparameter optimisation, feature selection was performed, as in 4.4.3. The result of this is shown in Table 4.7.

Table 4.7: A table showing the number of features selected for each condition using recursive feature elimination.

Condition	AF	PAC	PVC	AVB
Number of Features Selected	8	18	14	39

The number of features selected varies with the condition as the features and labels have different mutual information values. Conditions with an increased number of significant features have a smaller feature selection, whilst conditions with fewer significant features will need more features. This can be empirically assessed by counting the number of 'dark' squares for each condition in Figure 4.3 and comparing this with the number of features selected for that condition in Table 4.7.

The model was then trained and tested with an 80/20 train/test split in order to produce the ROC curves for each of the 4 conditions found in the dataset. The curves and AUC scores indicate that the diagnosis of atrial fibrillation can be done most successfully with this model, followed by that of atrioventricular blocks. The diagnosis of premature atrial complexes is performed with a similar success rate to the diagnosis of premature ventricular complexes, which is lower than that of the other two conditions.

By referring back to Figure 4.3 on page 28, we can see that the mutual information provided by the most important features for atrial fibrillation diagnosis is far higher, and hence more information about the labels can be inferred from the features. Atrioventricular blocks have a single important feature, whilst premature atrial and premature ventricular complexes have several moderately important features. The implication here is that having fewer, stronger features (atrioventricular block) is more useful than having more, less important features (premature ventricular/atrial complexes) in this mutual information range but this is impossible to assert given the small sample size.

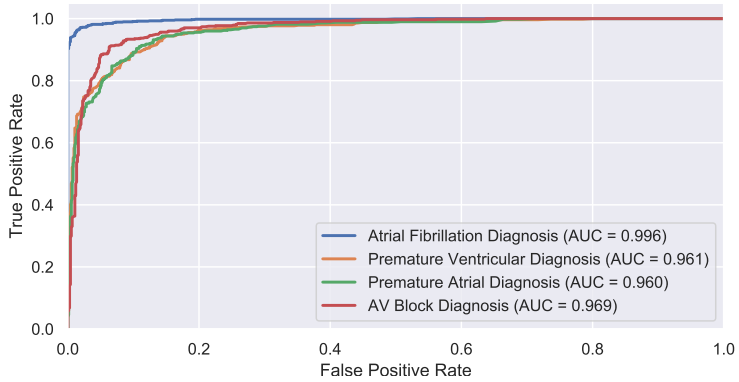


Figure 4.11: A graph of the ROC curves of each of the 4 random forests for the individual conditions they are trained to classify.

In order to directly compare the use of the multiclass model to that of the simple model as a screening system, the maximum probability of diagnosis over the four random forest classifiers in the multiclass classifier was used as the overall diagnosis probability, to provide the most inclusive diagnosis system possible. The ROC curve for the multiclass model could hence be plotted alongside that of the simple model, as seen in Figure 4.12.

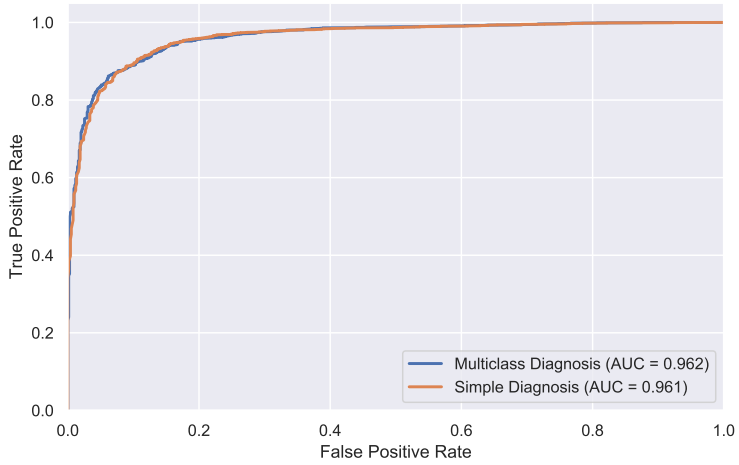


Figure 4.12: A graph of the ROC curves of the multiclass model and the simple model.

Figure 4.12 shows that the performance of the multiclass classifier is fairly similar to that of the simple model, but at small false positive rates the performance of the simple model is better. At large true positive rates, however, the multiclass model is better. This is because the method of combining the probabilities output by the models (simply using the maximum probability) results in a more inclusive diagnosis for the multiclass model — an advantageous property for a screening method.

4.7 Use of Previously Trained CNN Model

The Challenge 2020 data was subsequently classified as 'normal'/'abnormal' using CNN model 2 (the ResNet model in Section 3.4) after training and validating using the Challenge 2017 dataset. This was done to test the best performing neural network model's generality.

In order to be used with the model trained with Challenge 2017 data, the Challenge 2020 dataset was bandpass filtered, between 0.135Hz and 54.6Hz, using a 5th order Butterworth filter. Figure 4.13 shows the ROC curve produced for each condition when the filtered signals were passed through the CNN.

The Challenge 2017 dataset was not tested using the random forest model produced earlier in this section. This is because the signal processing required to detect features in the noisier Challenge 2017 dataset would take too long to implement — this is work that must be done in the future to improve the robustness of the feature extraction.

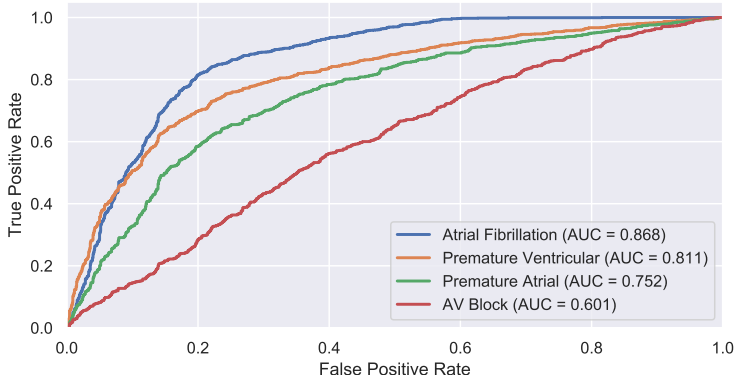
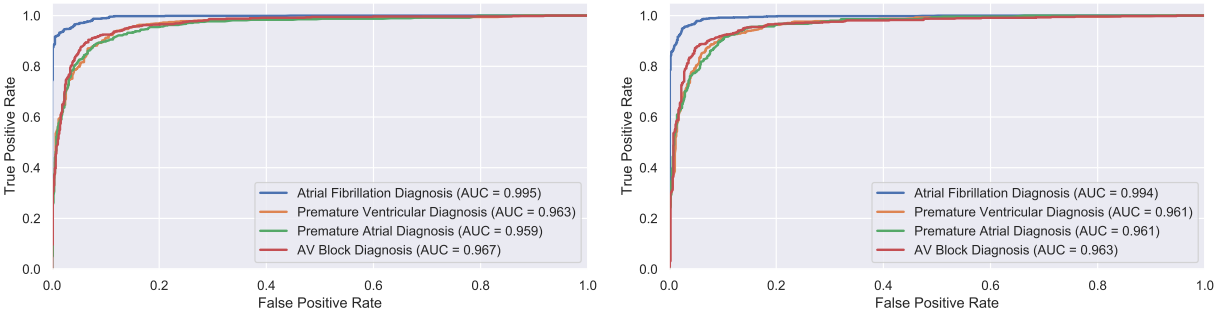


Figure 4.13: A graph of the ROC curves for each condition when testing the combined CNN model from Section 3.6 (trained with data from just the Challenge 2017 dataset) on the Challenge 2020 dataset.

It can be seen in Figure 4.13 that although the CNN model performs consistently well when classifying three of the conditions, it performs particularly poorly when classifying traces with an atrioventricular block. This suggests that there were not enough traces showing atrioventricular blocks in the Challenge 2017 dataset.

4.8 Combining Models

Lastly, a method combining the RF model with CNN model 1 from Section 4.7 was investigated. The CNN model detailed was trained and validated using Challenge 2017 data. The probabilities output when each of the Challenge 2020 records were passed through the CNN were used as an additional input feature when training the Random Forest model detailed in Section 4.6. The ROC curves with and without this additional feature are shown in Figure 4.14.



(a) Not including CNN output. **(b)** Including CNN output.

Figure 4.14: The ROC curves for the RF model when performing a binary ('normal'/'abnormal') diagnosis with and without the CNN output included as an input feature.

It can be seen that the performance of the random forest improves for all conditions when the CNN output is included as an input feature. This indicates that, despite the CNN's relatively poor performance when classifying samples by itself, that the performance of the random forest can be enhanced by including other models in the input features. The results could hence be further improved by including the output from better models as input features to the random forest, as they may provide increased mutual information about the labels. This improvement would be bounded when adding the output from more than one model, as the output of the models included will likely be correlated (the effect of which is discussed in Section 4.4.3).

The addition of a CNN-based classifier as a fifth classifier in the multiclass model (shown in Figure 4.10 on page 32) was considered as an alternative approach to combining models. This approach was rejected due to the large false positive rate for each of the conditions when the CNN model was tested on the Challenge 2020 dataset (as seen in Figure 4.13). As the overall label used for the multiclass model is 'abnormal' if any of the parallel classifiers report the input signal to be 'abnormal', the large false positive rate will be propagated to the overall classification if the CNN-based classifier is included.

Chapter 5: Screening UK Biobank ECG Database

In this chapter, a model from each of Chapters 3 and 4 was applied to the Biobank dataset, allowing real-world testing to be simulated. This allowed a realistic assessment of the performance of each of the models, as well as with the diagnosis system currently used by the ECG machine software - the *CardioSoft* ECG diagnosis system. The testing was done to evaluate the performance of the models when producing a binary 'normal'/'abnormal' label for each record in the dataset.

5.1 Data

The Biobank dataset consists of 37, 207 10 second 12 Lead ECG samples, with all patients being between 40 and 69 years of age when the recordings were taken. As this project focuses on the development of an automated screening system, only Lead I was used by the two models developed (however, the *CardioSoft* ECG diagnosis system uses 12 Leads). The patients' ECG recordings were taken for all participants in the Biobank study, not necessarily due to the observation of cardiovascular disease symptoms, so it provides a more realistic reflection of the data expected when using a screening system. The dataset has not been labelled by a cardiologist, so the only diagnostic information provided comes from the output from the *CardioSoft* ECG diagnosis system.

The number of records labelled by the *CardioSoft* system as showing each relevant condition is shown in Table 5.1 below. Relevant conditions for this investigation were those found in the training datasets used, as well as any that could be thresholded using their RR intervals (due to their more explicitly defined clinical diagnoses and the existence of reliable R peak detection algorithms [10, 32]). The labels with relevant conditions were the only records used from this dataset for the purposes of this project.

Table 5.1: A table showing the number of samples for each condition in the Biobank dataset, as labelled by the CardioSoft ECG diagnosis system (N.B. the total is not equal to the sum of the number of samples due to overlapping labels).

Label	**Number of Samples**
Normal	16299
Atrial Fibrillation	361
Premature Ventricular Complexes	1041
Premature Atrial Complexes	716
Atrioventricular Block	1930
Tachycardia	72
Bradycardia	16102
Total	34784

The Biobank ECG recordings were taken by ECG technicians in a hospitals around the UK, using an ECG machine. Each recording was taken once, and may exhibit a large amount of noise. The different recording method and preprocessing (compared to the Challenge 2017 dataset), and the different level and sources of noise in the ECG signals (compared to both the Challenge 2017 and 2020 datasets) made preprocessing and noise removal (see Sections 2.4 and 2.5) particularly important. Excessively noisy samples were manually removed when encountered because the automated detection of excessive noise (although vital for the development of an automated screening system) was outside the scope of this project.

5.2 Measuring Performance

Firstly, the *CardioSoft* diagnosis system's performance when the detection of tachcardia and bradycardia was analysed. This was done by thresholding the minimum and maximum heart rates according to clinical diagnosis guidelines. The heart rates were found by obtaining the RR intervals (by taking the difference between R peaks, as found using the Hamilton-Tompkins algorithm [10]), and by subsequently calculating the moving average of $\frac{60 \times \texttt{sample rate}}{\texttt{RR Interval}}$.

Next, the performance of CNN model 2 (Section 3.4) and the combined RF model (Section 4.6), when detecting the rest of the conditions given in Table 5.1, was compared to that of the *CardioSoft* diagnosis system. In the absence of a ground truth (and due to the large volume of the Biobank dataset), a random subset of the records for which the three diagnosis methods' labels diverged were manually labelled. This allowed an estimation of the accuracy of each of the screening methods when performing the 'normal'/'abnormal' binary classification over the whole population. The venn diagram in Figure 5.1 below shows the labelling convention — 50 records from each subset (001 to 111) will be manually labelled, so that only samples labelled as abnormal by one of the diagnosis methods are considered.

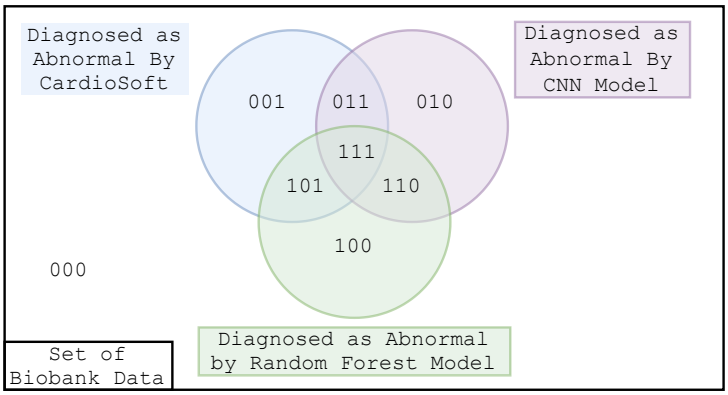


Figure 5.1: A venn diagram showing the 3 sets of records labelled as abnormal by each model, and the labelling convention used. Records in each of the three larger subsets (each represented by a circle and a corresponding binary bit) were labelled as being abnormal by the diagnosis method indicated. Overlapping subsets (011, 101, 110 and 111) were labelled by multiple methods of diagnosis.

When labelling samples using the models, the probabilities of abnormality output by the model were thresholded (to obtain labels) using the value that maximised the weighted sum of the true negative and positive rates ($\texttt{True Negative Rate} + 1.5 \times \texttt{True Positive Rate}$), in order to bias the model labels towards high true positive rates. A venn diagram of the records labelled as abnormal by each of the two models, as well as the *CardioSoft* diagnosis system, was produced (as seen in Figure 5.1). A random subset of 50 records was

then taken from each subset other than 000, to ensure that number of samples inspected was manageable yet still meaningful. These subsets were inspected to diagnose any abnormality and subsquently labelled. It was assumed that the samples were representitive of the full subset, hence allowing the performance of each model to be compared.

5.3 Results and Discussion

5.3.1 Bradycardia and Tachycardia

If the heart rate in the signal is below 60bpm, it is labelled as showing bradycardia [33] (some sources consider 50bpm to be the threshold, but for a screening method the most inclusive criterion is most appropriate). The signal is labelled as showing tachycardia if its heart rate exceeds 100bpm. When implementing these diagnoses, due to the potential presence of irregular heart rhythms, a simple moving average with $n = 3$ was used to smooth changes the heart rate. The moving average filter is a more consistent implementation of the '6 second method' used in cardiology [34]. These criteria were applied to the Biobank dataset, after using the Hamilton-Tompkins algorithm [10] to detect R peaks, and will be referred to as the **MA Algorithm**.

The records labelled by the 'MA Algorithm' and the *CardioSoft* automated diagnosis software as showing evidence of tachycardia are shown in Figure 5.2.

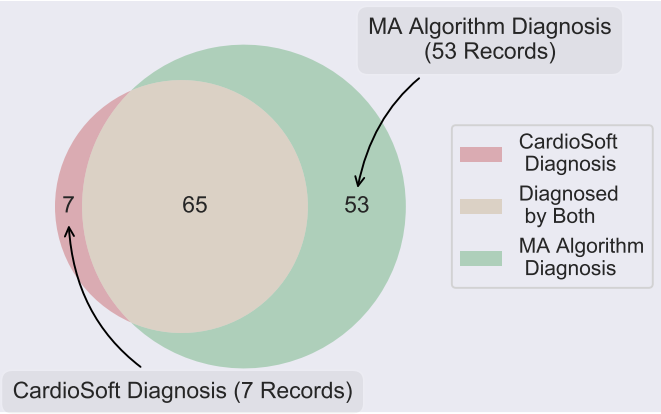


Figure 5.2: A venn diagram showing the number of records labelled as showing tachycardia by the MA Algorithm and *CardioSoft* diagnosis software.

Table 5.2 shows the proportion of the records labelled as showing tachycardia by just one diagnosis method that are correctly labelled, and the proportion that are mislabelled.

Table 5.2: A table showing the empirical distribution of the true labels of the traces labelled as showing tachycardia by just one diagnosis method.

Method	Number of Samples	Correctly labelled	Mislabelled
CardioSoft Only	7	0.28	0.72
MA Algorithm Only	53	0.94	0.06

It was found that most of the records labelled as showing tachycardia by *CardioSoft* and not by the MA algorithm were mislabelled (an example is given in Figure 5.4), whilst most of the labels were correct for records labelled by the MA algorithm and not by *CardioSoft* (an example of which can be seen in Figure 5.3).

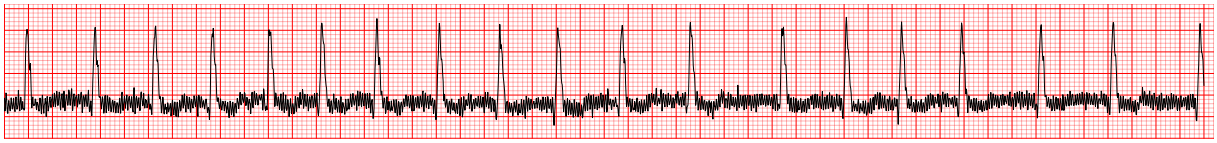


Figure 5.3: A trace labelled as having tachycardia by the MA algorithm and not by the *CardioSoft* diagnosis system, with a heart rate range from 100 to 132bpm.

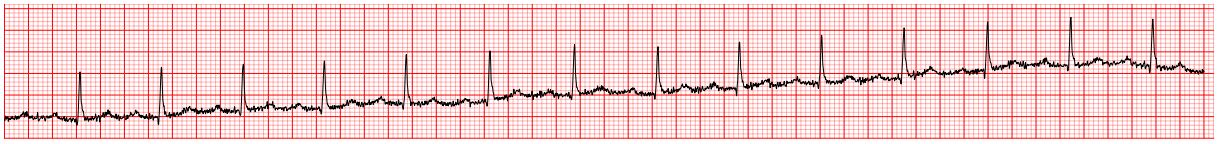


Figure 5.4: A trace mislabelled as having tachycardia by the *CardioSoft* diagnosis system, with a heart rate ranging from 85 to 89bpm.

Similarly, Figure 5.5 and Table 5.3 can be produced to evaluate the two methods for bradycardia diagnosis. Samples of 50 traces were taken from each of the subsets where it was only labelled as showing bradycardia by either *CardioSoft* or the MA Algorithm.

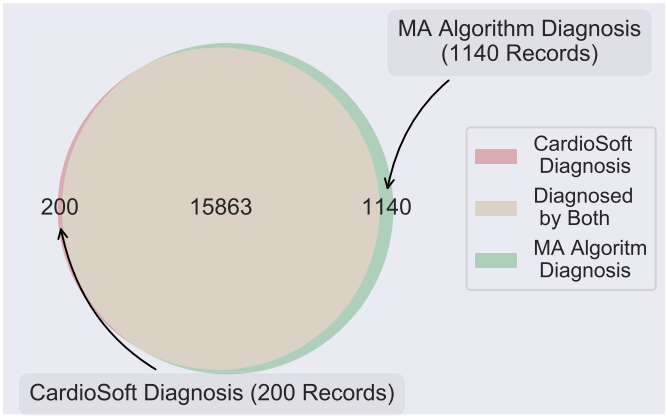


Figure 5.5: A venn diagram showing the number of records labelled as showing bradycardia by the MA Algorithm and *CardioSoft* diagnosis software.

Table 5.3: A table showing the empirical distribution of the true labels of random samples of 50 traces labelled as having bradycardia by just one diagnosis method.

Method	Number of Samples	Correctly Labelled	Mislabelled
CardioSoft Only	50 (Random Sample)	0.18	0.82
MA Algorithm Only	50 (Random Sample)	0.74	0.26

It can be seen that the MA algorithm was far more accurate than the *CardioSoft* diagnostic system. In particular, signals with inverted QRS complexes were found to be mislabelled by the *CardioSoft* diagnosis system, and correctly labelled using the MA algorithm (an example is shown in Figure 5.6). This inversion is either due to the presence of another arrhythmia, or due to the leads being the wrong way round when the ECG was recorded, but the recordings would expect to be labelled as showing bradycardia by a cardiologist, nonetheless.

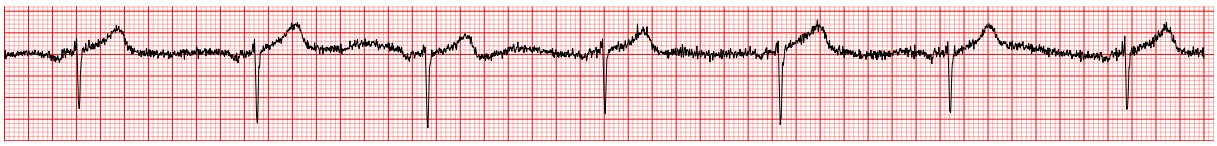


Figure 5.6: A trace with inverted QRS complexes showing bradycardia (due to a heart rate of 42bpm), labelled as 'normal' by the *CardioSoft* diagnostic system.

5.3.2 Overall Abnormality Detection

Initially, the ability of the combined random forest model (Section 4.6) and CNN model 2 (the ResNet model in Section 3.4) to detect abnormalities was tested, using the records labelled by *CardioSoft* as abnormal (other than those labelled as showing tachycardia or bradycardia, which were detected using the MA Algorithm). This was done to ensure that the models were able to replicate the true positive rate of the *CardioSoft* diagnosis system, despite the reduced number of ECG leads used. Subsequently, samples labelled by *Cardiosoft* as 'normal' were also screened using the two models, to detect any abnormal samples mislabelled by *Cardiosoft*. The models' ability to achieve a better true positive rate than *CardioSoft* whilst maintaining a low false positive rate was therefore able to be estimated.

The models were firstly used to provide a binary normal/abnormal label for each of the records labelled as showing atrial fibrillation, premature ventricular complexes, premature atrial complexes or atrioventricular block by *CardioSoft*. The result of this can be seen in Figure 5.7 overleaf.

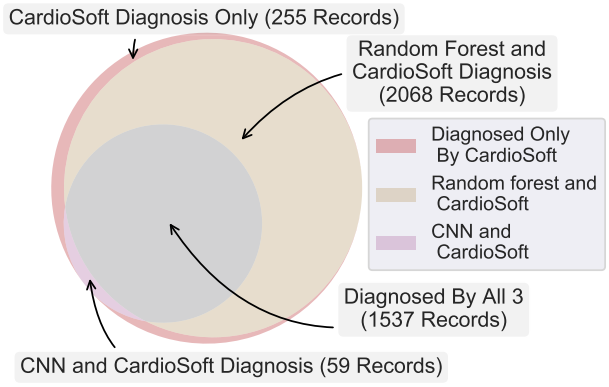


Figure 5.7: A venn diagram showing the number of records labelled as 'abnormal' by each model, when the records labelled as 'atrial fibrillation', 'premature atrial complexes', 'premature ventricular complexes' and 'atrioventricular block' by *CardioSoft* were analysed using the RF and CNN models. N.B. the external border is not present, due to the lack of samples labelled by *CardioSoft* as normal.

Each subset in which the three diagnosis methods did not agree were investigated, to provide an estimate of the true and false positive rates in the subsets. The results of this are shown in Table 5.4, whilst Table 5.5 investigates the true diagnoses of the abnormal records that were not labelled as such by all three diagnosis methods. This allows a more in-depth analysis of the strengths and weaknesses of each of the diagnosis methods.

Table 5.4: A table showing the empirical distribution of the correct and incorrect labelling of random samples of 50 traces from the subsets given (see Figure 5.7).

True Label	**Subset**		
	001 CS Only	**011 RF & CS**	**101 CNN & CS**
Correct	0.18	1	0.36
Incorrect	0.82	0	0.64

Table 5.5: A table showing the empirical distribution of the true diagnoses of the correctly labelled 'Abnormal' records from the samples in Table 5.4.

True Label	**Subset**		
	001 CS Only	**011 RF & CS**	**101 CNN & CS**
Atrial Fibrillation	0	0	0
Premature Atrial Complexes	0.11	0.04	0.56
Premature Ventricular Complexes	0.45	0.08	0.44
Atrioventricular Block	0.11	0.88	0
Other Arrhythmia	0.11	0	0

Looking at the results for the records labelled as abnormal by just *CardioSoft* (Subset 001) in Table 5.4, it is clear that most of them are false positives. The results from the records labelled by *CardioSoft* and the RF model as 'abnormal', but not by the CNN model (Subset 011) from both Tables 5.4 and 5.5 show that the CNN model misses many 'abnormal' samples, particularly when atrioventricular blocks are present. This would agree with the conclusions drawn when the CNN model was used to classify the samples from the Challenge 2020 dataset in Section 4.7. Finally, the results from the records labelled as abnormal by the CNN model and *CardioSoft* but not the RF model (Subset 101) show that, whilst the majority of the records labelled as abnormal were false positives, the RF model performs poorly when needing to detect ectopic beats (to detect premature atrial complexes and premature ventricular complexes). Despite this shortcoming, the RF model is seen to perform well when attempting to duplicate the 'abnormal' labels that the *CardioSoft* system produced, with the RF model being correct in the majority of cases when the labels produced by the diagnosis methods differ.

The models were subsequently applied to the samples labelled by *CardioSoft* as 'normal'. This was done to assess the ability of the models to **detect abnormal samples missed by *CardioSoft***. The result of this can be seen in Figure 5.8.

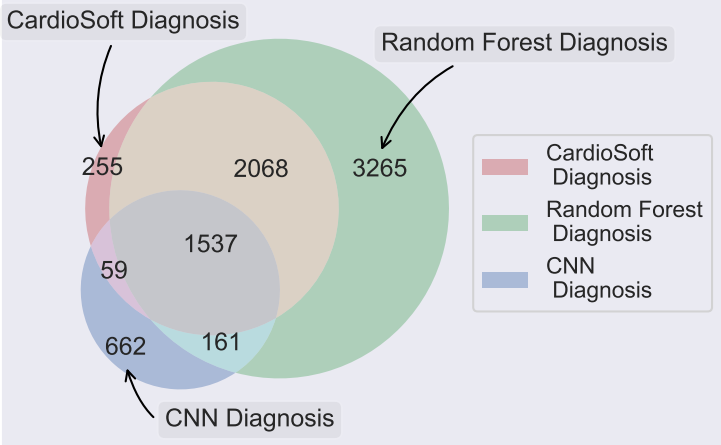


Figure 5.8: A venn diagram showing the number of records labelled as 'abnormal' by each model, when the records labelled as 'normal', 'atrial fibrillation', 'premature atrial complexes', 'premature ventricular complexes' and 'atrioventricular block' by *CardioSoft* were analysed using the RF and CNN models.

Tables 5.6 and 5.7 show the results when random samples of 50 records from each subset (001 to 111) (see Figure 5.1 for the labelling convention) were manually labelled.

Table 5.6: A table showing the empirical distribution of the correct and incorrect abnormality labelling of random samples of 50 traces from subsets 001 to 111 (see Figure 5.1 for the labelling convention).

True Label	Subset						
	001 CS	010 RF	011 RF & CS	100 CNN	101 CNN & CS	110 CNN & RF	111 All 3
Correct	0.18	0.68	1	0.36	0.36	0.66	1
Incorrect	0.82	0.32	0	0.64	0.64	0.34	0

Table 5.7: A table showing the empirical distribution of the true diagnoses of the records correctly labelled 'Abnormal' from the samples in Table 5.6.

True Label	Subset						
	001 CS	010 RF	011 RF & CS	100 CNN	101 CNN & CS	110 CNN & RF	111 All 3
Atrial Fibrillation	0	0.04	0	0	0	0	0.16
Prem. Atrial Complexes	0.11	0.18	0.04	0	0.56	0.04	0.24
Prem. Ventricular Complexes	0.45	0.07	0.08	0	0.44	0	0.28
Atrioventricular Block	0.11	0.04	0.88	0	0	0.13	0.12
Other Arrhythmia	0.11	0.67	0	1	0	0.83	0.20

Whilst the results from Subsets 001, 011 and 101 are the same as those discussed earlier, the results from Subsets 010, 100 and 110 can be used to extend conclusions about the models produced.

The results in Table 5.6 from the records labelled as abnormal by just the Random Forest (RF) model (Subset 010) contributes to the evidence that the it has a low false positive rate, whilst Table 5.7 shows that it is better than the other two models at detecting other arrhythmias.

The results of the records labelled as abnormal only by the CNN model (Subset 100) suggest that it has a higher false positive rate than the RF model. The results also suggest that the abnormal traces that only the CNN model detects also show evidence of other arrhythmias.

The results of the records labelled as abnormal by the CNN and RF models, but not the *CardioSoft* diagnosis system (Subset 110) suggest that, although the majority of these abnormal labels are correct, the false positive rate is higher than when only the CNN labelled them as abnormal (Subset 100). Despite this, by looking at the size and false positive rate of the subset where all three methods labelled the samples as abnormal (Subset 111), it is clear that combining the CNN and Random Forest methods would allow a far smaller false positive rate. This would result in a very low true positive rate, however, so it is not viable when developing a screening method.

Statistical analysis of the results in Table 5.6 allows the calculation of summary statistics shown in Table 5.8 for the random forest and CNN models, and the *CardioSoft* diagnosis software. To calculate these statistics, it was assumed that the random samples used are representative of the full subset (in order to estimate the total number of abnormal samples), and that all the abnormal samples have been found. This allows an estimate of the true positive rate (TPR), and the positive predicted value (PPV) ($\frac{\sum \texttt{True Positive}}{\sum \texttt{Labelled as Abnormal by Model}}$). Of the two assumptions made, the assumption that the random sample is representitive of the full subset is the only one that could affect the result — if this experiment were repeated, the use of a fully labelled test dataset would avoid the need for this assumption.

Table 5.8: A table showing the estimated Positive Predictive Value (PPV) and True Positive Rate (TPR) for each diagnosis method.

Statistic	***CardioSoft***	**Random Forest**	**CNN**
Estimated TPR	0.59	0.95	0.31
Estimated PPV	0.94	0.84	0.79

Table 5.8 shows that the *CardioSoft* diagnosis system and random forest model are both better than the CNN model. Although the random forest model has a lower positive predictive value than the *CardioSoft* diagnosis system, it correctly labels a far larger proportion of the abnormal traces (as evidenced by its higher true positive rate). As such, it can be said that the random forest system is far better suited to the production of a screening system than the *CardioSoft* diagnosis system, despite the fact that it uses just one ECG lead.

Chapter 6: Conclusion

Throughout the project, algorithms using Lead I electrocardiogram (ECG) signals to develop a screening system for cardiovascular disease were investigated. This would allow ECG signals input with any abnormalities to be highlighted, so that resource allocation in cardiology can be based on symptoms rather than statistics. To develop these algorithms, 3 datasets were used — a large dataset (the PhysioNet Challenge 2017 dataset) was used to investigate the feasibility of using deep learning based approaches, Due to labelling deficiencies in the Challenge 2017 dataset, a smaller dataset (the PhysioNet Challenge 2020) dataset was used to train a random forest model. The second dataset enabled the accuracy of the screening system produced to be quantified by each specific condition detected. Finally, a large unlabelled dataset (the UK Biobank ECG database) was used to simulate real-world testing of the screening system.

It was found that a diverse range of machine learning algorithms were suitable for use in cardiovascular disease screening. These ranged from deep convolution neural networks to neural networks employing Long Short Term Models, and ensemble methods. The steps taken to use an ensemble method (the random forest algorithm) closely resembled the methodology employed by cardiologists when classifying ECG signals, involving feature extraction followed by a decision-making stage. Conversely, neural network-based approaches allowed feature extraction and decision-making to occur simultaneously.

The presence of noise, and the pre-processing methodology used, influence the ease with which feature extraction, and hence ensemble methods such as random forests, can be used. Conversely, the size of the dataset influenced the suitability of deep learning based algorithms; with a small dataset, neural networks with lots of parameters are prone to overfitting. The depth of the neural network used must be sufficiently large, due to the complex nature of the function that the network is approximating. As a result the size of the dataset directly influences the suitability of the models used.

The models developed were general enough to label a different (test) dataset to that used when developing the model, as long as signals were adequately pre-processed and the noise was removed. The performance of the CNN model (trained using the Challenge 2017 dataset) was worse when used to classify records from the Challenge 2020 dataset rather than the test data from the Challenge 2017 dataset. This shows that, although an adequate level of performance can be achieved when training with a dataset from a different source, the performance of the classification algorithm is better when the data used to train it is representative of the test data.

When the random forest (RF) model was compared to the *CardioSoft* software used by UK Biobank, it was found that, although the RF model produced proportionally more false positives than the *CardioSoft* software, it had a far higher true positive rate. Meanwhile, the CNN model used was estimated to have a proportionally higher false positive rate and a lower true positive rate than the *CardioSoft* software. The random forest is hence more suitable than the *CardioSoft* automated diagnosis software for automated screening of arrhythmias despite its use of 1 ECG lead rather than 12. This is because a high true positive rate is vital when developing a screening system, and its false positive rate was not seen to be much higher than that of the *CardioSoft* system, when positive predicted values were used as an indicator.

The CNN was seen to be far worse than the other two diagnosis systems, due to its inability to detect atrioventricular blocks. The random forest model performed slightly worse when detecting arrhythmias defined by the presence of ectopic beats — future work should investigate whether this is due to data limitations, or whether it was an inherent flaw of the model.

In summary, random forest and deep learning models were both found to be promising when developing an ECG screening system — evidence suggests that random forest models are more appropriate when features can be accurately extracted, whilst deep CNN models were found to be best when accurate feature extraction cannot be guaranteed. The use of pre-processing was vital to ensure that the models produced were generalised enough to be used to classify data from datasets other than the one used when training the model. The random forest model's performance was better than that of the *CardioSoft* ECG diagnosis system, although work must be done to improve ectopic beat detection.

This report recommends that further work focuses on the effect that the expansion and diversification of datasets has on the performance of the models produced. Obtaining labels for the full Biobank dataset would allow more concrete judgements on the relative merits of the models developed, whilst the investigation of other feature-based classification methods such as support vector machines would allow the suitability of the random forest algorithm to be compared to other feature-based approaches. The use of more sophisticated signal processing methods when extracting features for use with the random forest model would allow the model to cope with noisier inputs. Although CNN and random forest models have been shown to be promising when developing cardiovascular disease screening systems, both have limited interpretability. In order to facilitate the adoption of the screening method suggested in healthcare, this is a shortcoming that must be addressed in further work.

Appendix A: Risk Assessment Retrospective

The risk assessment completed before the project started asserted that the project would be low-risk. Some ergonomic hazards such as eye strain, back ache and headaches were identified. Regular breaks were taken, and correct posture was maintained, to prevent harm as a result of these hazards. This was successful, with no ill effects arising during the lifecycle of the project. As a result, no changes would be made to the risk assessment if a similar project were undertaken.

Appendix B: COVID-19 Disruption

For Chapter 5, it was previously intended to get doctors to label a significant portion of the UK Biobank ECG Database between 02/05/2020 and 16/05/2020. This would have allowed concrete solutions to be drawn about the models developed. Instead I diagnosed 510 ECG signals myself in order to assess the performance of the models, requiring a significant time investment.

Appendix C: Electronic resources

A logbook detailing the work involved can be found at:

`docs.google.com/spreadsheets/d/1PLtI9VOvkAH5RK3lmi33n_wQi5EOkfclyabibQy-Owc`

A GitHub repository containing all of the code written for this project can be found at:
`github.com/hiralradia/4th-year-project`

Please contact the author of this paper to be granted access to the repository.

Bibliography

[1] "Cardiovascular disease." [Online]. Available: https://www.england.nhs.uk/ourwork/clinical-policy/cvd/

[2] A. L. Goldberger and H. E. Stanley, "PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13).

[3] "UK Biobank," www.ukbiobank.ac.uk/resources/, 2010, accessed: 2010-09-30.

[4] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ecg morphology and heartbeat interval features," *IEEE transactions on biomedical engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.

[5] S. Meek and F. Morris, "ABC of clinical electrocardiography.Introduction. I-Leads, rate, rhythm, and cardiac axis," *BMJ (Clinical research ed.)*, vol. 324, no. 7334, pp. 415–418, Feb. 2002, publisher: BMJ. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11850377

[6] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *Nuclear Science, IEEE Transactions on*, vol. 44, pp. 1464 – 1468, 07 1997.

[7] F. Buendía-Fuentes, M. A. Arnau-Vives, A. Arnau-Vives, Y. Jiménez-Jiménez, J. Rueda-Soriano, E. Zorio-Grima, A. Osa-Sáez, L. V. Martínez-Dolz, L. Almenar-Bonet, and M. A. Palencia-Pérez, "High-Bandpass Filters in Electrocardiography: Source of Error in the Interpretation of the ST Segment," *ISRN Cardiology*, vol. 2012, p. 706217, Jun. 2012, publisher: International Scholarly Research Network. [Online]. Available: https://doi.org/10.5402/2012/706217

[8] K. Nazarpour, A. H. Al-Timemy, G. Bugmann, and A. Jackson, "A note on the probability distribution function of the surface electromyogram signal," *Brain research bulletin*, vol. 90, pp. 88–91, 2013.

[9] K.-M. Chang, "Arrhythmia ECG noise reduction by ensemble empirical mode decomposition," *Sensors (Basel, Switzerland)*, vol. 10, no. 6, pp. 6063–6080, 2010,

edition: 2010/06/17 Publisher: Molecular Diversity Preservation International (MDPI). [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22219702

[10] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database," *IEEE transactions on biomedical engineering*, no. 12, pp. 1157–1165, 1986.

[11] P. Puech, R. Grolleau, and C. Guimond, *Incidence of Different Types of A-V Block and their Localization by his Bundle Recordings*. Dordrecht: Springer Netherlands, 1978, pp. 467–484. [Online]. Available: https://doi.org/10.1007/978-94-009-9726-4_26

[12] P. M. Radiuk, "Impact of training set batch size on the performance of convolutional neural networks for diverse datasets," *Information Technology and Management Science*, vol. 20, no. 1, pp. 20–24, 2017.

[13] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[14] E. Kauderer-Abrams, "Quantifying translation-invariance in convolutional neural networks," 12 2017.

[15] Y.-L. Kavukcuoglu and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in neural information processing systems*, 2010, pp. 1090–1098.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] J. Kohler, H. Daneshmand, A. Lucchi, M. Zhou, K. Neymeyr, and T. Hofmann, "Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization," 2018.

[18] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" 2018.

[19] A. E. Orhan, "Skip connections as effective symmetry-breaking," *CoRR*, vol. abs/1701.09175, 2017.

[20] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *arXiv:1707.01836*, 2017.

[21] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ecg signals," *Computers in biology and medicine*, vol. 102, pp. 411–420, 2018.

[22] E. Alickovic and A. Subasi, "Medical Decision Support System for Diagnosis of Heart Arrhythmia using DWT and Random Forests Classifier," *Journal of Medical Systems*, vol. 40, no. 4, p. 108, Apr. 2016.

[23] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Computer methods and programs in biomedicine*, vol. 130, pp. 54–64, 2016.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001.

[25] "Topographic prominence," May 2017. [Online]. Available: http://www.andrewkirmse.com/prominence

[26] W. Commons, "File:topographic isolation and prominence.jpg — wikimedia commons, the free media repository," 2018. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:Topographic_isolation_and_prominence.jpg&oldid=317656198

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.69.066138

[29] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, 10 2013.

[30] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan 2002. [Online]. Available: https://doi.org/10.1023/A:1012487302797

[31] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, p. 307, 08 2008.

[32] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, March 1985.

[33] D. Da Costa, W. J. Brady, and J. Edhouse, "Bradycardias and atrioventricular conduction block," *BMJ (Clinical research ed.)*, vol. 324, no. 7336, pp. 535–538, Mar. 2002, publisher: BMJ. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11872557

[34] B. Aehlert, *Pocket Reference for ECGs Made Easy - E-Book*. Elsevier Health Sciences, 2012. [Online]. Available: https://books.google.co.uk/books?id=2UEaMx4mrSoC