B3 Group Design Project

Predicting Atrial Fibrillation in Real-Time: Harnessing Ensemble Machine Learning Techniques for Wearable Applications

Hengyu Wang, Aaron Chen, Owen Douglas, Nishen Menerapitiya

Supervisors:

Prof. Alison Noble, Prof. Vicente Grau, Prof. Konstantinos Kamnitsas, Prof. Jens Rittscher

Department of Engineering Science

University of Oxford

Trinity 2023

Contents

List of Abbreviations

AF	*Atrial Fibrillation*
AUC	*Area Under Curve*
CEA	*Cost-Effectiveness Analysis*
CNN	*Convolutional Neural Network*
CQC	*Care Quality Commission*
CSTM	*Cohort State Transition Model*
DOAC	*Direct-acting Oral Anticoagulant*
ECG	*Electrocardiogram*
EHR	*Electronic Health Record*
FDA	*U.S. Food and Drug Administration*
FN	*False Negative*
FP	*False Positive*
GDPR	*General Data Protection Regulation*
ICER	*Incremental Cost-Effectiveness Ratio*
LSTM	*Long Short-Term Memory*
MHRA	*Medicines and Healthcare products Regulatory Agency*
NASEM	*National Academies of Science, Engineering, and Medicine*
NHS	*National Health Service*
NICE	*National Institute for Health and Care Excellence*
NSR	*Normal Sinus Rhythm*
PiP	*Pill-in-the-Pocket*
PPG	*Photoplethysmography*
PPV	*Positive Predictive Value*
PSA	*Probabilistic Sensitivity Analysis*
PxAF	*Paroxysmal Atrial Fibrillation*
QALY	*Quality-Adjusted Life Year*
RNN	*Recurrent Neural Network*
ROC	*Receiver Operating Characteristic*
SaMD	*Software as a Medical Device*
TN	*True Negative*
TP	*True Positive*
UI	*User Interface*

Chapter 1

Introduction

Atrial Fibrillation is a condition causing the irregular beat of the heart and is one of the most common types of cardiac arrhythmia. With 1.4 million people in England carrying the burden of AF, 425,000 of them are undiagnosed and untreated [1]. The greatest mortality risk of AF is from the increased risk of stroke, with AF increasing the likelihood of stroke five-fold. In 2020, the NHS spent £2 B in direct costs due to AF, with 58% spent on primary admission and 15% on prescription and monitoring [2]. Furthermore, there is no current population screening for AF and the self-diagnosis of AF is unreliable. Therefore, there is a need for a cost-effective patient pathway to screen and diagnose AF.

The project had two primary aims: to determine whether there was an accurate method to detect asymptomatic AF in the wider populations and a way of providing better AF treatment. Through preliminary research, including an interview with a clinician, the specific unmet needs in current AF detection and treatment were determined and solutions to these needs were developed.

The proposed solution used an ensemble deep learning methodology with ECG and photoplethysmography (PPG) sensors in the Apple Watch Series 4 or above, to notify users of potential cases of AF. The model was trained and tested with the MIT-BIH Arrhythmia Database, ultimately attaining an impressive AUC score of 0.99. The solution gives the NHS the opportunity to detect asymptomatic cases of AF and can aid in the medication of using a Pill-in-the-Pocket methodology.

Two clinical validation studies, the AFib-AI Study and the AI-PiP Study, were designed in accordance with UK and EU regulations to evaluate the performance and the feasibility of the Pill-in-the-Pocket strategy guided by our algorithm.

Finally, an impact assessment was performed to determine the anticipated effects of the proposed design. Through a cost-effectiveness analysis, the economic viability of the PiP treatment pathway was compared to the current pathway and it was found to be the better treatment strategy. The expected outcomes for key stakeholders were analysed to establish the main benefits and drawbacks of the design. The assessment also highlighted important considerations for implementing the solution in practice and future research requirements.

Chapter 2

Background & Need

2.1 Background {Nishen Menerapitiyage Don}

Atrial fibrillation (AF) is a condition characterised by the irregular and rapid beating of the atrial chambers of the heart. Normally during an episode of AFib, the heartbeat can exceed 140 beats per minute.[3] Global studies show that around 33 million people worldwide live with the burden of AFib. The prevalence of AFib among young people under the age of 40 is very low at less than 0.5%.[4] However, the prevalence of AFib, increases drastically as the age of the population analysed increases. For example, for people above 80 years or more the prevalence of AFib is around 10%. Figure 2.1 shows the prevalence of AFib against age, demonstrating the increasing trend in prevalence as people age.

Furthermore, according to Staerkk et al.[6], based on the 70-year-long Farmington Heart Study in the USA the lifetime risk of an individual developing AFib was 23.4%. However, this lifetime risk depends on elevating risk factors such as alcohol consumption, smoking, body mass index and other factors. The lifetime risk of AF increases significantly the more elevating risk factors a patient possesses. For example, the lifetime risk of an individual with at least 1 elevated risk factor was more likely to get AFib.

AFib itself can be managed if detected early, and many people live healthy lives while living with AFib, however, the greatest risk from AFib is the heightened risk of stroke, cardiovascular disease, congestive heart failure and ischemic heart disease. These diseases can cause mortality, therefore, there is a heightened risk of mortality for individuals with the burden of AFib. According to a systematic

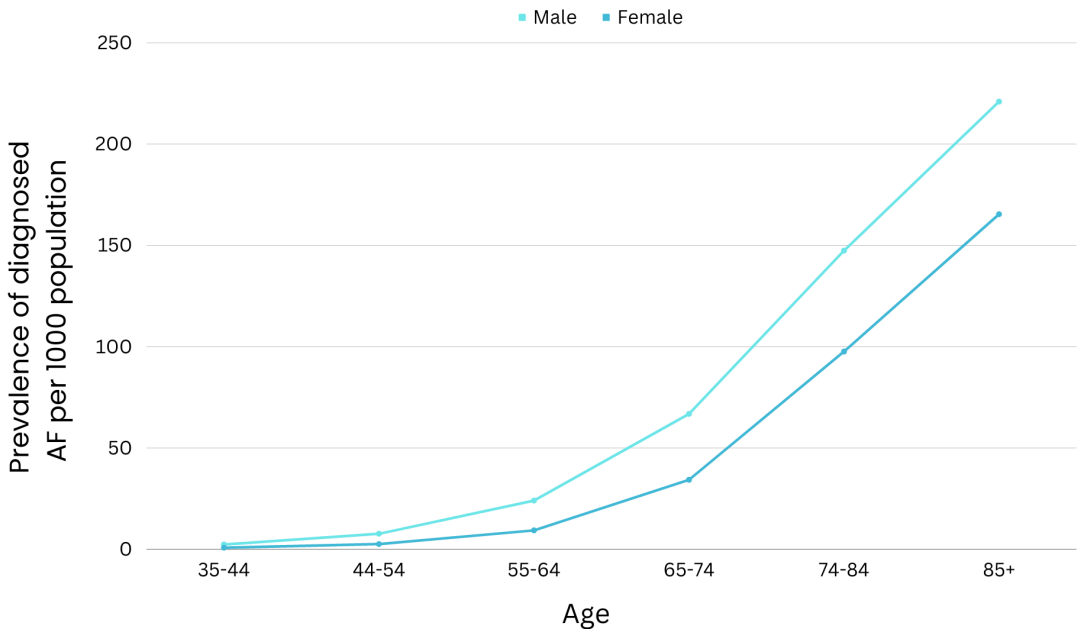


Figure 2.1: Prevalence of AFib increases with higher age groups in the UK, data from Adderley et al. [5]

review of 104 eligible cohort studies involving 9,686,513 participants by Odutayo et al.[7], AFib was demonstrated to be associated with an increased risk of mortality. The mortality through cardiovascular disease was 2.6 events per 1000 participant years with a median follow-up of 4.9 years. The mortality through stroke was 3.6 events per 1000 participants for a median follow-up of 4.2 years. The mortality through ischaemic heart disease was 1.4 events per 1000 participant years for a median follow-up of 4.1 years. The mortality through ischaemic heart disease was 11.1 events per 1000 participants for a median follow-up of 5.4 years.

Table 2.1 summarises their findings of the mortality frequency for AFib.

Label	**Mortality events per 1000 participant years**	**Median follow up (years)**
cardiovascular disease	2.6	4.9
stroke	3.6	4.2
ischaemic heart disease	1.4	4.1
Congestive heart failure	11.1	5.4

Table 2.1: Mortality Rates due to AFib, data from Odutayo et al.[7]

Therefore, it is evident that AFib can have a huge impact on people's lives and there is a clear need

for better screening and effective treatment for people living with AFib.

In the UK, AF is one of the most common cardiac arrhythmias with an overall population prevalence of 2.5%. Around 1.4 million people in England live with AF and 425 000 were estimated to be undiagnosed and untreated.[1]

In this study, we will focus on the UK and specifically the National Health Service (NHS) of the UK and the burden of AF on the NHS. In 2020, alone AFib is estimated to have directly costed the NHS on average £2 billion, which is on average 1% of NHS expenditure, spent on managing one disease. In the next two decades the expenditure is predicted to be increased on average 4% [2] with figure 2.2 depicting the predicted trend of increase in expenditure.

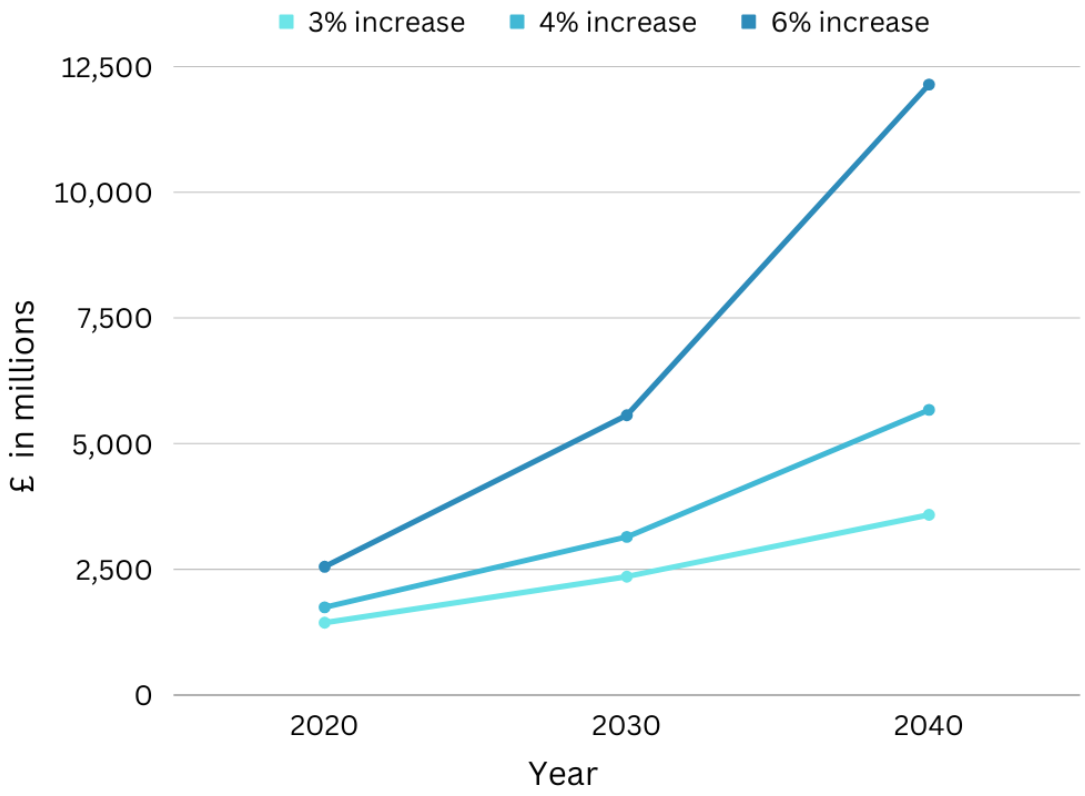


Figure 2.2: Projected Direct Costs of Managing AFib is predicted to increase over the next 2 decades, data from Burdett et al.[2]

Figure 2.3 depicts the largest costs associated with managing AFib, and it is evident that primary admission costs are the single largest cost factor.

Therefore, if over the next two decades, we could implement a solution to automatically streamline the process of detecting AFib, there could be a significant financial benefit for the NHS. Furthermore, such a process would make it easier to identify individuals with untreated AFib and ensure that the future burden of such individuals is reduced through early intervention.

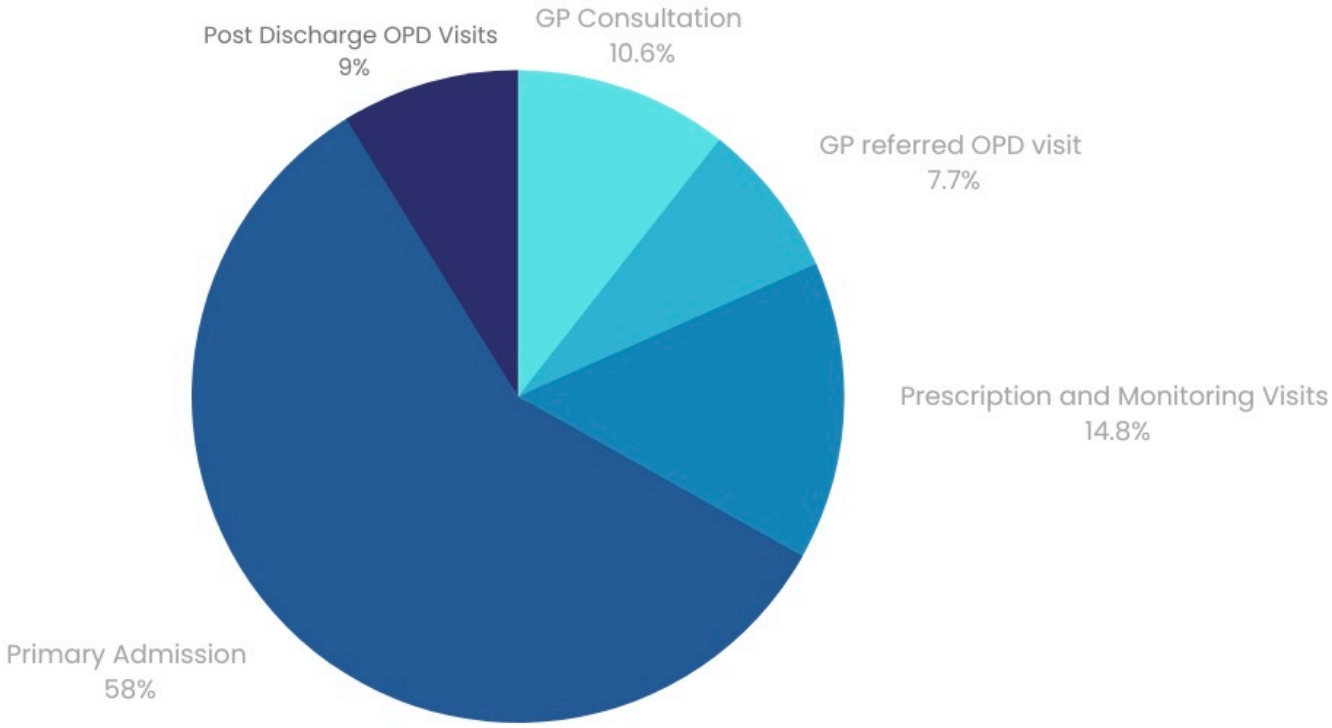


Figure 2.3: Direct costs associated with managing AF broken down into sections, data from Burdett et al.[2]

2.2 Identifying Needs {Owen Douglas}

To identify the unmet needs in the treatment of AF, a study was conducted drawing inspiration from the Stanford Biodesign process [8]. Following an interview with Alex Sharp, a clinician at Oxford University Hospitals, a needs statement was then formed. This helped to establish the existing challenges within the current treatment pathway and provided a foundation for innovating solutions to address them.

2.2.1 Research

The primary form of research was conducted via the interview. The objective was to gain a thorough understanding of the current clinical methods of AF detection, diagnosis and treatment pathways, and to determine opportunities for improvement. It was also critical to determine the current burden of AF on healthcare systems from a clinician's perspective, as this could provide further insight into the effects of potential solutions.

Based on the analysis of the interview, it was determined that the primary focus of treatment was to reduce the risk of stroke by prescribing anticoagulants to patients. Since strokes are a significant burden to both patients and the NHS, in terms of cost and quality of life, patients are started on anticoagulation treatment immediately [9].

Clinical research has suggested that an initial test for AF prior to a GP confirmation with a 12-lead ECG is expected to be more cost-effective than relying on 12-lead ECG testing alone [10]. Currently, individuals who are at risk of AF are required to check their pulse manually for irregularities. This method is difficult to perform accurately and may not be effective for cases of infrequent AF episodes. As approximately one-third of AF cases are asymptomatic [11], the risk of stroke could be reduced by utilising improved methods of detection.

Anticoagulants taken daily have an associated bleeding risk, which can be significant. The interviewee mentioned the promising Pill-in-the-Pocket (PiP) approach to anticoagulation prophylaxis. In this methodology a patient would only need to take anticoagulants when required, therefore reducing their exposure to the increased risk of major bleeds. The PiP technique is currently in its research stage, as it requires accurate detection algorithms and technology to ensure it is safe and effective.

2.2.2 Needs Statement

The results of the preliminary research and interview analysis were collated and formed into specific identified needs.

- Need 1: A way to accurately detect more cases of AF in the wider population, especially asymptomatic cases.
- Need 2: A way to administer Pill-in-the-Pocket anticoagulation to patients diagnosed with AF that is cost-effective.

A solution to the first need would reduce the risk of stroke in patients with asymptomatic AF and decrease the burden of stroke on the NHS. It would require a suitable accurate detection method that is simple to use and easily scaled. The second need aimed to reduce the risk of major bleeds and improve patients' quality of life. It would require consistent detection of AF episodes and a method to inform patients to take anticoagulant medication.

2.2.3 Stakeholders

The anticipated key stakeholders considered during the design and evaluation of the product were:

- *Patients* - It was essential for the design of solutions to the unmet needs to consider the effects on patients. Important points to analyse included quality of life, resulting risks of stroke and major bleeds and ease of use.

- *Clinicians* - Forming an essential component of the treatment pathway, the impact of the solutions on clinicians needed to be assessed. Ease-of-use was essential to ensure they were willing to adopt new technology.
- *Healthcare Systems* - A major aspect of the identified needs was the requirement for solutions to be cost-effective. This was to ensure solutions to the needs were economically viable for hospitals to implement. Therefore healthcare systems were a major stakeholder in the design.
- *Society* - The design of solutions needed to evaluate societal impacts to determine any ethical and social implications. The public's perception could play a crucial role in the adoption of new technology.

2.3 Technical Approach {Hengyu Wang}

In our research, we put forth an innovative, ensemble-based neural network methodology for classifying ECG signals. The process is initiated with the utilization of signal conditioning techniques, serving to normalize the ECG signals. Subsequently, we employed a wavelet denoiser, a signal processing tool, to filter out any intrusive noises while maintaining the integrity of the original signal shape and characteristics.

Signal processing was further utilized in the extraction of a variety of one-dimensional frequency domain features. These included instantaneous frequency, spectral entropy, a modified periodogram, and Welch's power spectral density. The extracted time-frequency features provided key information utilized in the subsequent modelling process.

Moreover, we transformed ECG time series data into two-dimensional scalograms using continuous wavelet transforms, another essential step in signal processing. Scalograms provide a visual representation of how these features evolve over time, offering a different perspective for the convolutional neural network models to learn from.

Our ensemble model incorporated the strengths of five different deep learning models, each of which had been either trained or fine-tuned specifically for this application. These models included a bi-directional LSTM (which was trained on the four time-frequency features), pretrained GoogLeNet, SqueezeNet, MobileNet-v2, and NASNet-Mobile (which were all fine-tuned on scalograms).

The integration of these diverse models within an ensemble approach demonstrated superior performance when compared to the individual models. This improvement was manifested as an en-

hancement in the classification accuracy. Notably, the ensemble model maintained a quick execution time on wearable devices, achieving a favourable balance between high accuracy and computational efficiency. This equilibrium is particularly vital for machine learning applications within embedded systems.

Overall, our model presents a compelling option for mobile applications, where the need for real-time processing is paired with a requirement for low computational overhead.

2.4 Anticipated Impact {Aaron Chen}

The impact of our product we envisioned spam through all stages of the patient care pathway, from early detection to continuous monitoring and patient-oriented long-term management. In this section, we will use the Cycle of Care concept to assess the solution and product's potential impact. The Cycle of Care is a fundamental framework used in healthcare to emphasize and promote a comprehensive, personalised, and patient-centred health journey for all patients. We will discuss the impact we anticipated in the four separate stages of the cycle of care, Prevention, Diagnosis, Treatment, and Management.

Anticipated Impact on Prevention

The non-invasive nature of the wearable device and its user-friendly design encourage users to check their heart rhythms regularly. The continuous heart rate monitoring module would also provide users with increasing awareness of potential AF episodes. The aim of the solution is not only to provide an alternative option to a population screening of AF but also to educate the population about AF to encourage a healthier lifestyle. The accompanied app will also be equipped with regular exercise notifications as well as advice for healthy diets. Consequently, this will lead to a reduction in the risk of AF and many heart diseases.

Anticipated Impact on Diagnosis

Our solution of continuous irregular heart rate monitoring can spot and alert any episode of AF, even for asymptomatic patients. The high accuracy and precision enable early detection of AF while minimising false positives. This leads to a more efficient diagnosis process and reduces costs for walk-in clinical pulse readings, and ultimately, reduces the number of undiagnosed cases in the population.

Anticipated Impact on Treatment

Our pill-in-the-pocket anticoagulation strategy, guided by our algorithm, provides an alternative option for AF-diagnosed patients with a low risk of stroke. By shortening the duration of anticoagulants, the strategy can effectively safeguard patients from experiencing severe side effects like minor and major bleeding.

This new strategy enables a more personalised and patient-oriented treatment plan which greatly improves the patient outcomes of consultation. As explained in Pendleton's consultation model [12], patients are more likely to comply and follow treatment plans when they are involved in the treatment decision-making process. In addition, a more personalised treatment plan encourages patients to pay more attention to comparing with a general plan or advice. As a result, our solution can provide more effective treatment plans while increasing patient adherence.

Anticipated Impact on Management

Continuous data collection allows healthcare providers to track patients' situations more regularly for a faster decision-making process when a change in the treatment plan is required. The regular follow-up process can also potentially be simplified by virtual meetings with data collected through our device. The pill-in-the-pocket strategy also enables a smoother transition in the treatment plan as patients who no longer require anticoagulant treatments with fewer episodes would already be pausing anticoagulant intake. Ultimately, our product and solution can improve long-term AF management with more accessible data and easier change in treatment plans.

Conclusion

Our product and solution of early detection with wearable devices and a pill-in-the-pocket strategy have the potential to make a significant impact on AF management. Our product aims to improve the patients' AF pathway by encouraging prevention, allowing early detection, providing patient-oriented treatment plans, and improving long-term management.

Chapter 3

Design {Nishen Menerapitiyage Don}

3.1 Literature Review

3.1.1 Research Solutions

During preliminary research, we investigated 5 different deep learning methodologies, currently used solely for classifying AFib ECG data, and general neutral net methodologies with the capability to analyse a broader scope of data. For each, the feasibility of the methodology for our proposed solution, computational cost of implementation of the methodology, and proven track record was analysed.

3.1.1.1 ECG Raw Data Based Classification

Bi-directional Long Short-Term Memory (LSTM) network

A bi-directional LSTM network is a recurrent neural network capable of learning order dependence. This network takes raw ECG data as input and using the deep learning layers gives a final classification layer output of the final class prediction. Such models have been used in a wide range of applications from speech recognition to pattern classification. Furthermore, there is a precedent for using LSTM models for atrial fibrillation diagnosis. Le Sun et al.[13] developed a recurrent neural network (RNN) composed of stacked LSTM for AFib prediction. They showed a 92% accuracy and 92% f-score for AFib prediction. Therefore, using LSTM can provide a high-accuracy AFib detection model directly using the raw data recorded by the wearable.

3.1.1.2 ECG Image Based Classification

GoogLeNet

GoogLeNet is a 22-layer deep convulsion neural network architecture proposed by Szegedy et al.[14] developed for the classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14).

The solution is carefully designed such that a derivative solution of GoogLeNet for a given depth and width of the network, is optimised to minimize computational budget. Therefore, with a carefully optimised network, this can be utilised for an onboard wearable device ECG image processing.

This model was not specifically designed to analyse ECG images and is a general-purpose image classification methodology. However, as the volume of ECG images required to train the model is minimal compared to other larger, less optimised models, it is anticipated that the computational and financial cost of implementing a derivative solution of GoogLeNet solely for ECG image classification would be relatively low.

Furthermore, the literature has implementations of GoogLeNet for ECG Rhythm analysis. For example, Kim et al. [15] produced a model to classify ECG rhythms into normal sinus rhythms, premature ventricular contraction, atrial ventricular contraction, and right/left bundle branch block arrhythmia using a GoogLeNet Deep Neutral Network architecture. The accuracy of the classifying model was on average 95.94% with a maximum positive predictive value of 95.7% between a normal sinus rhythm against the other three arrhythmias.

Therefore, this derivative implementation of GoogLeNet provides compelling evidence that GoogLeNet is a viable model to implement for our proposed solution given our constraints of reliability and stability.

SqueezeNet

SqueezeNet is a lightweight deep convolutional neural network (CNN) architecture introduced by Iandola et al.[16] which achieves an Alex Net [17] level accuracy with 50 times fewer parameters with less than 0.5MB model size. This low-parameter CNN offers more efficient distributed training of the model and lower overhead when exporting the model. Furthermore, this model can be deployed to analyse a video stream in real-time with less than 10MB of memory on board an embedded system. This low resource, high classification accuracy is achieved using a Fire module. This model was not specifically designed to analyse ECG images and is a general-purpose image classification

methodology. Despite this, the model is a great candidate for our proposed solution as this model is lightweight in terms of onboard wearable computation and the amount of computational and financial resources to train the model. Furthermore, as the overall size of the model is small, it can prove to be power efficient when the model is used in real-time on wearables.

Furthermore, the literature has implementations of ECG image-based Arrhythmia detection using SqueezeNet with a track record of success. For example, Rahman et al.[18] had implemented an ECG Arrhythmia classification based on AlexNet and SqueezeNet with an accuracy of 98.8% and 90.08% respectively. Also, He et al.[19] developed LiteNet, a neutral Network optimised for detecting Arrhythmia for resource-constrained mobile devices. LiteNet is a derivative of SqueezeNet and GoogLeNet. LiteNet consumes twice as less processing power and is four-fold faster compared to AlexNet with an accuracy of 97.87%. Therefore, these prior research implementations, provide compelling evidence that SqueezeNet is a viable model to implement for our proposed solution given our constraints of limited computational and power requirements.

MobileNet-v2

MobileNetV2 is a lightweight and efficient deep learning architecture designed for mobile and embedded computer vision applications by Sandler et al.[20] This model uses an inverted residual structure with linear bottlenecks thus improving the flow of information between layers and reduces the number of parameters necessary without sacrificing accuracy.

Although this model is not specifically designed to analyse ECG images, it is an embedded system classification methodology. Therefore, rendering the model useful for low latency and low resource utilization applications. To that extent, Adelazez et al.[21] developed a derivative model based on MobileNet -v2 CCN for the classification of AFib using a compressive sensing technique and spectrogram transformation reducing the overall memory intensity of training the model. The model had an Area under the curve (AUC) of 0.87, and 0.78 for uncompressed and 50 % compressed data respectively.

Therefore, it is suitable for our proposed solution, as there is a proven track record of success, and it would enable us to train the models with limited computational power and memory bandwidth with relatively high accuracy.

NASNet-Mobile

NASNet-Mobile is deep learning neural network architecture designed specifically for mobile-embedded

computer vision applications by Zoph et al.[22]. It is based on Neural Architecture Search (NAS), an automated approach to finding the best neural network architecture for a given task.

NAS uses reinforcement learning and evolutionary algorithms to search for optimal architecture configurations. Therefore, has proven to show a high accuracy in classifying images. Furthermore, the implementation minimises the model size and latency, allowing the use of the model in low-resource environments. Therefore, it would be able to give a high accuracy of real-time ECG classification while maintaining a low memory footprint on our proposed solutions wearable device.

Although this model is not specifically designed to analyse ECG images, it is a mobile embedded system classification methodology, therefore has a proven record of use in low-resource environments.

3.1.2 Market Solutions

3.1.2.1 Apple Watch OS AFib Detection

Apple Watch OS, AFib detection notification in Apple Health is a first-party Atrial Fibrillation detection built into the Apple Watch Ecosystem. According to Apple Inc.[23], technical documentation, the implementation uses two detection pathways. Photoplethysmography (PPG) based Arrhythmia detection for arrhythmia warning notification and a user-initiated ECG-based detection in the form of the ECG app.

The PPG sensor-based arrhythmia detection works based on the calculation of the Heart rate and heart rate variability made on average every 2 to 4 hours. This data is converted into a plot of a tachogram using a proprietary algorithm. This tachogram then allows for the detection of irregular arrhythmia, when a given classification threshold is met. This implementation had a 78.9% accuracy distinguishing AFib from other arrhythmias and a 98.2% accuracy distinguishing all arrhythmias compare to a normal pulse. This implementation allows users to understand the probability of them having AFib, but it is not a diagnostic tool.

The ECG App-based user-initiated detection relies on the user taking an ECG using the app. Given that the user took an ECG, it will be analysed using a proprietary algorithm, classifying the ECG as Safe Rate (SR), AFib, poor recording or inconclusive. This algorithm classification achieved a 98.3% sensitivity and 99.6% specificity. This implementation is CE certified for all users above 22 years, therefore providing the ECG app medical validity to diagnose arrhythmia based on the ECG data.

3.1.3 Current Diagnosis Pathway for AFib

According to the NHS Clinical Pathways [24] in the current diagnosis pathway for suspicion of AF, the patient is required to manually check their pulse by counting the number of beats through touch or using a wearable device. Afterwards, when the patient seeks medical attention, they will be assessed for suspected stroke, if positive they will be diagnosed based on the stroke diagnosis pathway. If the patient needs specialist services, there will be emergency referrals at this stage. If none of the conditions above was encountered, the patients history is taken, a physical examination is conducted, and a 12 lead ECG is performed.

If the ECG confirms AFib or atrial flutter, the symptom burden is assessed using a European Heart Rhythm Association (EHRA) score. If Persistent AFib is likely the patient will be considered to be put on anticoagulation or referred to cardiological management of persistent AFib.

If ECG shows a sinus rhythm, and paroxysmal AF is suspected, the patient is considered for differential diagnosis, and put on ambulatory monitoring. If Paroxysmal AFib is likely the patient will be considered to be put on anticoagulation or referred to cardiological management of paroxysmal AFib.

Figure 3.1 shows a simplified version of the current suspected AFib pathway. While figure 7.3 in Appendix depicts the full current pathway.

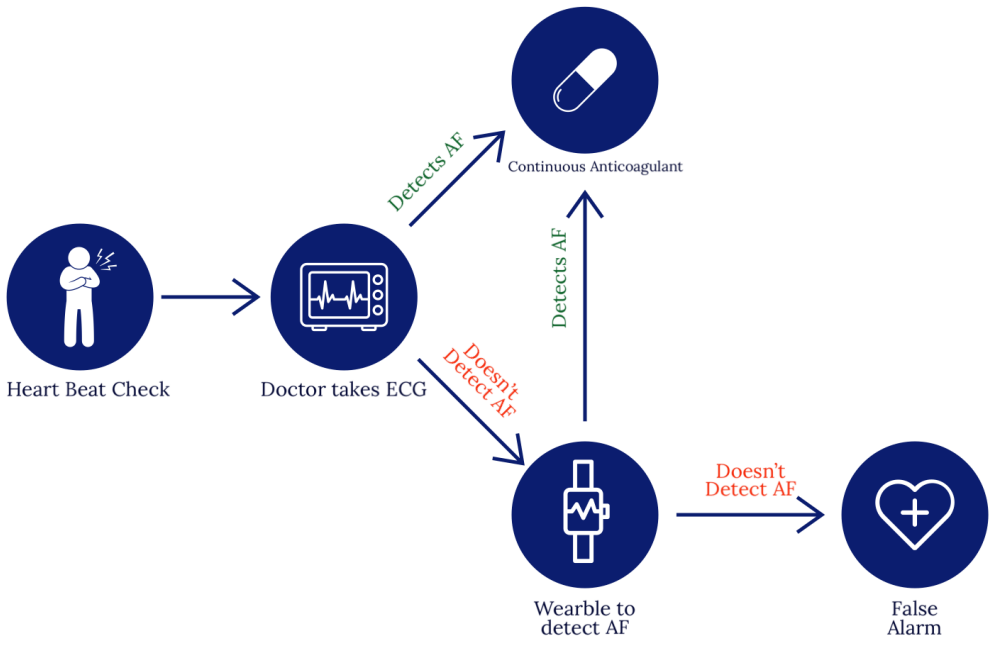


Figure 3.1: A simplified version of the current suspected AF patient pathway, highlighting the main stages.

However, this pathway is dependent on patients realising that they have AFib and seeking medical help. According to the British Heart Foundation [25] around 300,000 people in the UK live with undiagnosed AFib and are oblivious to the dangers of AFib. Thus, there is a great need to automatically detect AFib and create a better pathway to diagnose AFib.

3.1.4 Medication Pathway for AFib

After a diagnosis of AFib is confirmed by a medical professional based on their age, lifestyle, and medical history they will be recommended a medication pathway to prevent complications such as stroke arising due to AFib. There are currently two different medication pathways available to patients diagnosed with AFib.

3.1.4.1 Medication to control the rate of heartbeat

The goal of this pathway is to reduce the resting heart rate to less than 90 beats per minute. This is generally achieved through beta-blockers such as bisoprolol or atenolol or calcium channel blockers such as verapamil or diltiazem. Patients on Beta Blockers experience many side effects, according to Farzam et al.[26] these include hypotension, bradycardia, dizziness, fatigue nausea and constipation. Similarly, patients on calcium channel blockers have shown side effects such as low blood pressure, ankle swelling and heart failure.

3.1.4.2 Medication to reduce stroke

Due to the arhythmic nature of the heartbeats in patients with AFib, there is a high risk of blood clots forming in the heart chambers. The goal of this pathway is to reduce the chance of a stroke due to a blood clot using anticoagulation. However, patients on anticoagulation have a high risk of blood loss from bleeding. According to Shoeb et al.[27] the risk of major bleeding due to anticoagulation is 0.3% - 0.5% per year. Therefore, there is a serious risk to patients on this medication pathway.

3.1.4.3 Pill-in-the-Pocket (PiP)

As medications that are currently used to manage AFib, have side effects that reduce the quality-of-life of patients, the pill-in-the-pocket methodology aims to mitigate these side effects by asking patients to only take medications as soon as they realise, they have atrial fibrillation.

This approach only works if the patient can identify when an episode starts. According to Passman[28] even though this has been the main challenge of effectively implementing the pill in the pocket methodology, with the rise of wearable technologies and deep learning algorithms to detect

AFib ECG signals there is a great potential that the pill in the pocket could be an effective treatment pathway.

Two pilots have been carried out demonstrating the feasibility of the pill-in-the-pocket methodology. Rhythm Evaluation for Anticoagulation with Continuous Monitoring (REACT.COM) and Tailored Anticoagulation for non-continuous Atrial Fibrillation (TACTIC-AF).

REACT.COM was a trial among 59 patients with a mean age of 67 years monitoring their use of anticoagulation and adverse side effects due to anticoagulation. In this trial a 94% reduction in anticoagulation use was measured based on a 1-hour duration of anticoagulation reinitiation [29]

TACTIC-AF was a trial among 48 patients with a mean age of 71.3 years monitoring their use of anticoagulation and adverse side effects due to anticoagulation. In this trial, there was a 75% reduction of time on anticoagulation when patients who experience less than 6mins of AFib rhythms per day were asked to stop anticoagulation. [30]

According to Passman [28] during the two studies enrolling 96 patients with 112 patients years of follow-ups, had not observed stroke during the trials. Thus, providing clinical evidence that the PiP methodology is feasible as a potential medication pathway.

3.2 Available Datasets

3.2.1 Introduction

One of the biggest constraints of developing a reliable deep learning, approach to detect Atrial Fibrillation (AF) is the availability of data to train the model based on the selected specific hardware platform. Therefore, through analysing all available ECG datasets it was decided to use the PhysioNet 2017 Challenge dataset.[31] The PhysioNet 2017 Challenge dataset includes 12,186 single-lead ECG recordings, each lasting around 9 to 61 seconds as exemplified in figure 1. All ECG recordings in the data set have been resampled to a uniform sampling rate of 300 Hz. This enables us to consistently analyse and compare various algorithms. Furthermore, the dataset includes 8528 expert annotated ECGs in the training set. Each ECG in the training set is assigned 4 labels with each label containing an asymmetric distribution of no of recording as shown in table 3.1.

Label	No of recordings	Mean length of ECG (s)
Normal	5154	31.9
AF	771	31.6
Other Rhythm	2557	34.1
Noisy	46	27.1
Total	8528	32.5

Table 3.1: Physionet Dataset Labels and recordings depicting the asymmetry of the data labels

3.2.2 Feasibility

An ideal dataset would have been a labelled publicly available ECG dataset obtained from the single lead ECG on an Apple Watch Series 4 or above. However, the Physionet dataset provides single lead ECG, outputs of around 30 s obtained by AliveCor single channel ECG devices.[31] Therefore, it sufficiently micks the data characteristics of the ideal data set. Thus, enabling us to train the model based on the PhysioNet data and use Apple Watch ECG data as input data to accurately classify AFib. Furthermore, the volume of data provided is sufficient for training an accurate model using standard computational resources. Thus, the model resulting from this data can be locally run on a wearable, without resorting to cloud computation.

3.2.3 Limitations

However, one drawback of the dataset is the unequal distribution of the number of ECG signals in each of the labelled categories. This can lead to the majority label (in this case normal label) achieving a high classification accuracy compared to the other labels. Furthermore, it could lead to reduced generalization capabilities on the validation set. Moreover, 8528 training patterns dont offer much flexibility with the methodology employed. A larger dataset would have provided higher accuracy. Furthermore, as the data is fully anonymised without access to aggregated demographics such as age and gender distribution, it is hard to assess the extent of the validity of the data to the demographics of the UK. However, the advantages of this publicly available dataset overweight its limitations.

3.3 Proposed Design

Our proposed solution is to use an ensemble deep learning algorithm running on a wearable device sensor, to continuously detect AF. This would alter the current patient pathway shown in figure 3.1 to

our proposed patient pathway shown in figure 3.2. The proposed solution pathway has 3 stages with each stage employing a unique technology to achieve this.

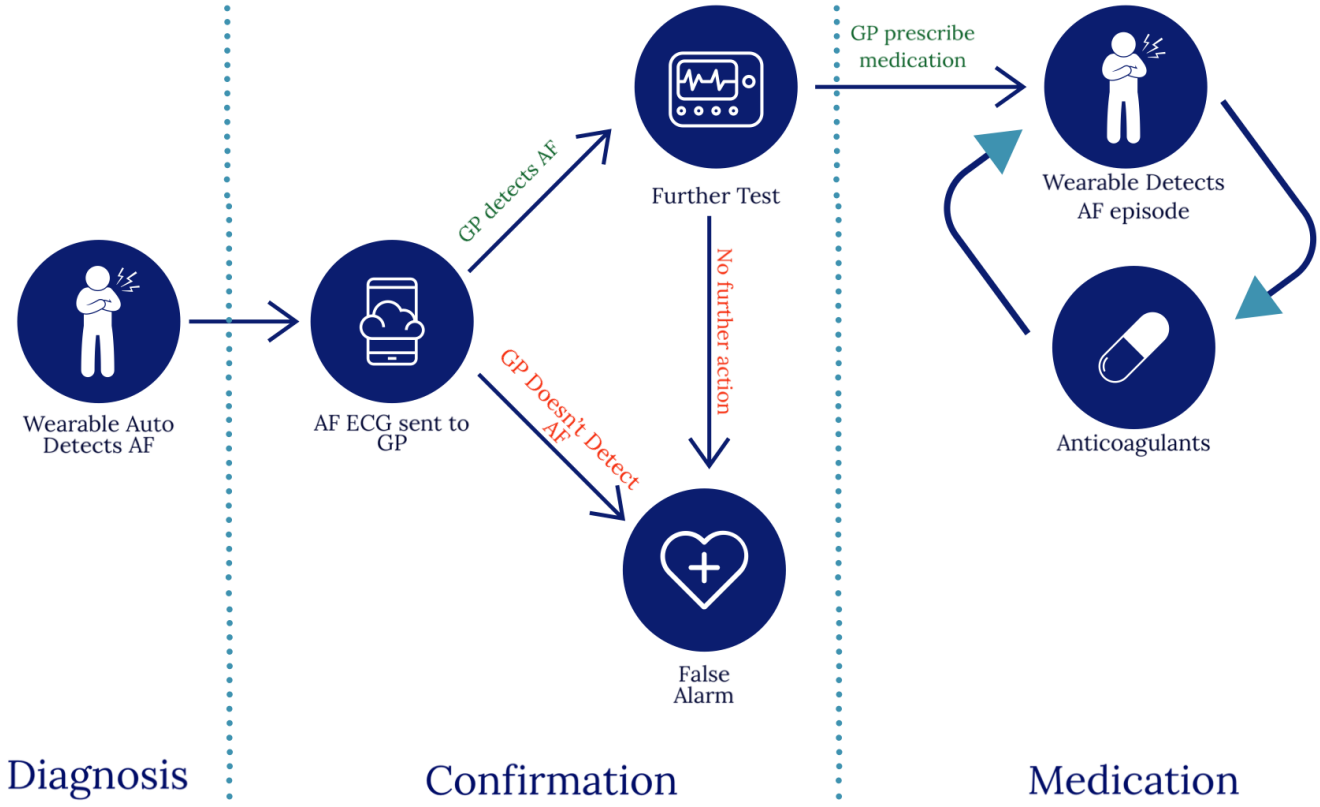


Figure 3.2: The proposed solution patient pathway depicting the three stages of the pathway and outcomes in each stage

3.3.1 Diagnosis

In the diagnosis stage, the users of the wearable app will be continuously monitored for irregular rhythm. When an irregular rhythm is detected, the user will be prompted to take an ECG using the wearable by holding the crown of the device. Figure 3.3,3.4 depicts the user interface the user will be presented when an irregular rhythm is detected in light mode and dark mode respectively. When the wearable detects that the user viewed the alert, it will transition to the user interface depicted in Figure 3.5,3.6 for light and dark modes respectively. A minimalist user interface was designed as it is anticipated that many users will be older than the general population, thus ease of use was a key priority. Furthermore, the user is given the option to ignore the warning by pressing the red cross in the corner. This was placed in the left-hand side corner as it is unlikely that a user will accidentally cancel the warning.

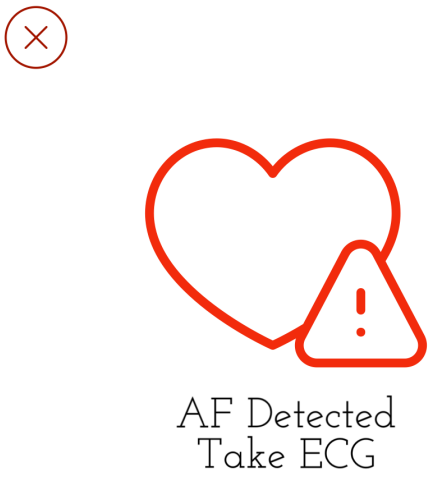


Figure 3.3: AF detection notification UI, light mode.

Figure 3.4: AF detection notification UI, dark mode.

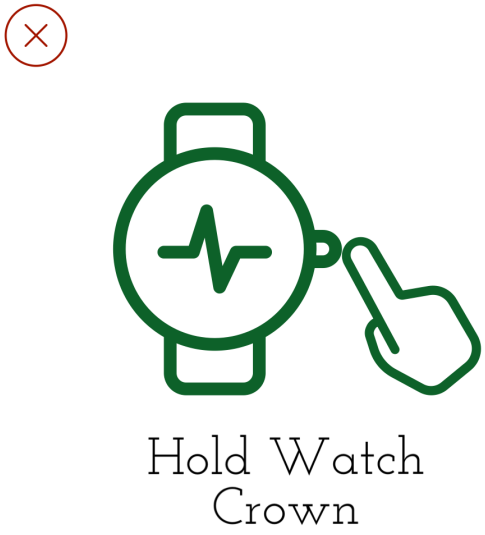


Figure 3.5: ECG instruction UI, light mode.

Figure 3.6: ECG instruction UI, dark mode.

3.3.1.1 Technology

We decided on using existing wearables to implement our solution. This was proven to be effective in getting wide user adoption. Furthermore, in the coming decade, it is hypothesised that there will be a surge in the adoption of commercial wearables. From a large array of wearables available on the market, it was decided to implement the solution on Apple Watch Series 4 and above. This was because those models of watches have the required ECG sensor integrated, the developer-friendly tools in the WatchOS wearable ecosystems and the largest market penetration standing at 26% of all wearables in the third quarter of 2022.[32] The continuous irregular heart rhythm detection is

going to be implemented using the Photoplethysmography (PPG) sensor Apple HealthKit library for WatchOS. This algorithm provides a single pulse reading rather than a full ECG of the user to make a prediction of irregular heart rhythm. This library has a true positive 78.9% at detecting AF and a 98.2% true positive rate detecting AF and other clinically relevant arrhythmias.[23] This approach was taken as this library, uses the PPG heart rate monitor of the wearable. This sensor is allowed to run in the background on WatchOS and doesnt require user input, unlike the ECG sensor on the wearable. Furthermore, as we are detecting mild or undetected AF, a fully automated no-user system, was not required as the user is capable of providing input. Therefore, the approach taken optimises for the limited amount of onboard processing and battery life with our target user needs. Given that an irregular pulse was detected, our implemented app will notify the user to take an ECG using the wearable ECG sensor. The ECG reading implementation will use the HKelectrocardiogram HealthKit library on WatchOS. This ECG reading module was implemented as it was the officially tested standard implementation on the Apple Developer Kit. This reading module outcome when paired with Apples proprietary processing methodologies has a sensitivity of 98% and specificity of 99% of identifying AFib.[23] Thus, indicating the validity of the ECG reading module.

3.3.2 Conformation

As our approach is patient-oriented giving power to the patient, in this stage the user is notified of the outcome of the algorithm and asked for consent for the data to be shared with the GP. If the user consents to this, their data will be securely communicated to the GP using a cellular network or using the companion phone app for non-cellular models. Afterwards, the GP can inspect the ECG along with the algorithm classification. The GP can then decide to invite the patient for further testing using a 12-lead ECG or discard the detection as a false positive. This decision is notified to the patient, using the user-facing app. Then it is the patients choice to visit the GP if they are given the opportunity to do so. Figure 3.7,3.8 depicts the user interface when the patient is notified of the positive outcome for AF for light and dark modes respectively.

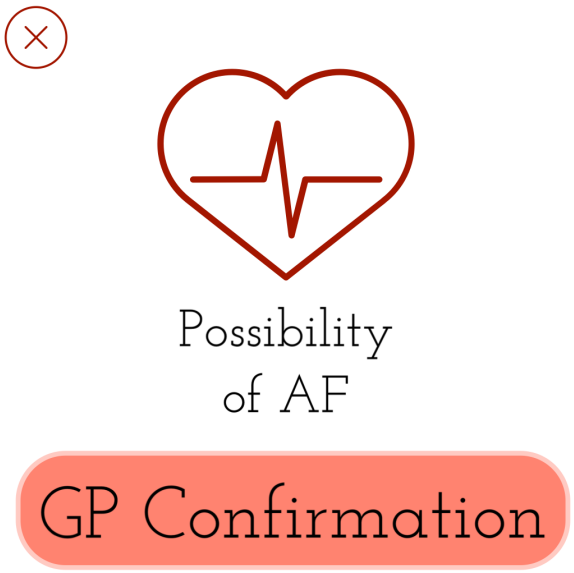


Figure 3.7: Possibility of AF UI,light mode.

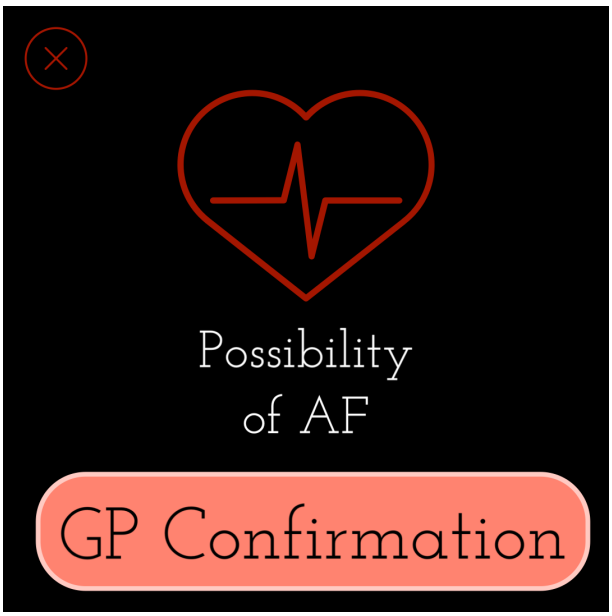


Figure 3.8: Possibility of AF UI, dark mode.

Figure 3.9,3.10 depicts the user interface when the patient is notified of the negative outcome for AF in light and dark modes respectively.

These user interfaces are transitioned from the user interface depicted in Figure 3.5,3.6.

Figure 3.9: No AF detected UI, light mode.

Figure 3.10: No AF detected UI, dark mode.

When the ECG recording was faulty or inconclusive, it will transition to the user interface depicted in Figure 3.11, 3.12, confirmation in this user interface will cause the transition to the user interface in Figure 3.5,3.6.

Figure 3.11: AF Unconfirmed UI, light mode.

Figure 3.12: AF Unconfirmed UI, dark mode.

Figure 3.13 shows the user interface when the GP has invited the user for further testing. The user is given a choice to book an appointment with the GP on the companion app. The help icon, in the right-hand corner, provides more information about AFib and what to expect during the appointment.

Figure 3.13: Companion app UI, providing the user the opportunity to book further test appointment

3.3.2.1 Technology

At the point where user consent occurs, the data is securely stored on Apple Electronic Health Records. This record is then synced with the patients GP who has integrated Apple Digital Health Records into the system. As this technology is based on Fast Healthcare Interoperability Resources (FHIR) it will be securely stored on NHS servers along with the full patient Electronic Health Record while maintaining interoperability with legacy systems.

The GP side client will then query FHIR for updates and the GP will be notified if there are any users ECGs that needed to be reviewed. After the GP decides, the update will be sent to the user-side app, where the user can view and respond to the GP decision.

In the NHS, Apple Electronic Health Records are currently implemented at Oxford University Hospital (OUH) and Milton Keynes University Hospital.[33] [34] A systematic meta-analysis of the implementa-

tion of EHR was conducted by Campanella et al.[35]. The report showed improved quality of healthcare, increasing time efficiency and guideline adherence and reducing medication errors in health service implementing EHRs. Furthermore, the report shows a reduction of documentation time difference with a mean percentage difference of -22.4% at a 95% confidence interval and a lower overall medication error with an overall risk ratio of 0.46. Moreover, it is anticipated that the wider NHS will adopt the Apple EHR system in the coming decade. Therefore, implementing our solution through the Apple EHR system can ensure that implementing our solution would be frictionless and future-proof. Also, Campanella et al.'s work[35] provides evidance that it can lead to faster GP diagnostic confirmation with lower errors with minimal bureaucracy and financial cost to the NHS. Likewise, on the user end, it allows them to seamlessly integrate their health records with their wearable health data generated through our solution with strong security and privacy measures and protocols built into Apple EHR.

3.3.3 Medication

During this stage, the doctor discusses with the patient about their medication plan. If the patient is referred to continuous anticoagulation, our proposed pathway gives the patient and the GP the opportunity to only take anticoagulation when required called the pill in the pocket, stroke prevention pathway. In this pathway, the patient will only take anticoagulation when there is an AF episode. This is achieved by the app continuing to function in the background and notifying the patient when to take an ECG and suggesting the patient to take anticoagulation if AF was detected. A summary of these recordings will be securely stored on Apple Electronic Health Records, for the GP to inspect at the next visit.

3.4 Regulatory Requirements

The regulatory environment for the proposed solutions depends on the action performed by each key component. Therefore, the regulatory environment for the solution is analysed for the 3 key stages

3.4.1 All stages

The MRHA regulates medical devices across the UK, and software used for diagnostic purposes is regulated as a medical device according to section 2 Part I of the Medical Device Regulation 2002(as amended)(MDR 2002).[36]

Therefore, as our proposed solution uses a machine learning algorithm to provide diagnostic infor-

mation, we will have to comply with MRHA regulations for medical devices. Therefore, to comply with the regulation, documentation will be filled under the general medical devices (Part II of the UK MDR 2002) to receive a UKCA and CE mark for the machine learning software and user-facing app. [36]

The proposed solution not only includes our machine algorithm and user-facing app but also external developer modules from Apple Health Kit.

The external modules used in the solution, HKelectrocardiogram, from Apple HealthKit ECG modules, are CE marked for people over age 22 within the Apple ECG App. These external developer modules were implemented to ensure that external modules used will not affect MRHA compliance.

These steps taken will ensure that the solution is compliant with MRHA regulations and provide healthcare providers and users the trust in our solution.

3.4.2 Diagnosis

The diagnosis stage of the solution involves using the algorithm for public screening of Atrial Fibrillation; therefore, the solution was built in adherence to the quality standards and guidelines set out by Public Health England (PHE) and the UK National Screening Committee (UK NSC).[37]

Our solution is patient-oriented providing the user with the opportunity to report the outcomes of our machine learning app. Therefore, aligning ourselves with the population screening, ethics, and guidelines of informed personal choice.

Furthermore, our solution is simple, has no harm to the user, and can be validated. We provide the option to the user for further diagnosis with a primary healthcare provider upon a positive result. The data for the algorithm has been tailored for the UK, population and it was made sure the diagnosis is suitable for the test. Therefore, implementing the solution such that criteria set for appraising viable, effective, and appropriate population screening programmes according to the UK NSC are met.

3.4.3 Conformation

The confirmation stage involves access to personal data, automated processing of data, storage of data, and transference of data to third parties.

Therefore under Article 4 of the General Data Protection Regulation (GDPR)[38], our solutions implementing entity is considered a data controller and data processor.

Article 9 of GDPR prohibits the use of health data without the explicit consent of the user. In our

solution, we explicitly ask for the informed consent of the user before measuring any ECG and other physiological information.

Furthermore Article 22 of GDPR, explicitly mentions that decisions cannot be made solely based on automated processing. Our solution pathway always requires the GPs clinical advice before proceeding with any medical test or intervention.

Moreover, before sending the algorithm output label data and ECG to the GP we ensure that consent from the user is obtained, therefore, ensuring that the patient is always in control of the data.

Our solution ensures that we dont store data on our servers, therefore, we dont have to take any action towards compliance with storage of data. However, it is our duty to ensure that the data is stored safely, within the UK or the EU. Therefore, we are using a secure GDPR-compliant storage implementation, Apple Digital Health Records, to store and transmit patient data.

Additionally, using Apple Digital Health Records will ensure that the data produced by our app is portable and the user is free to use any other competitor app of their choosing. Therefore, complying with Article 20 data subjects' Right to data portability.

Also, starting from the trial phase of the solution, it is intended to set up a data privacy officer, to make sure all data requests from users are handled, and compliant with regulations set by the Information Commissioners Office and Article 37 of GDPR.

Therefore, the design of our solution ensures that data privacy is maintained at the highest level, and it complies with all relevant GDPR regulations.

3.4.4 Medication

The proposed solution in the medication stage implementation performs the AF detection independent of a clinician, which is a regulated activity under the Health and Social Care Act 2008 (Regulated Activities) Regulation 2014.[39]

Therefore, the machine learning algorithm will need to be registered with the Care Quality Commission (CQC). This would require resources to be invested for compliance for use in the pill in a pocket phase of the solution, however, the public screening trials can go ahead without registration as there is complete clinician oversight.

This would subject our solution to CQC assessment frameworks for healthcare services[40] and the performance of the solution will be rated for the public. This rating will increase trust and credibility in

our solution whilst also ensuring that there is a motivation for us to comply with legislation and follow recommendations provided by the CQC.

3.5 Impact Assessment

The impact of the solution is measured using a Markov cohort state analysis for the population detection phase and medication Pill in the Pocket stroke prevention pathway. For each analysis, a Markov state transition life cycle of 1 year was used. This enables the proposed solution to accurately capture the patients outcomes with sufficient accuracy. For this analysis in each of the proposed solution phases, a cost analysis, disease progression analysis, and quality of life analysis are conducted. This will enable us to analyse the value added by the proposed solution to the NHS, the users, and the wider society. Furthermore, this will provide us with a starting point for the financial value of the proposed solution and its ability to penetrate the market.

Furthermore, a probability sensitivity analysis is carried out to account for the uncertainties in the true values of bias parameters in the Markov cohort state analysis. Moreover, this allows us to simulate the effects of adjusting bias to assess the impact of the solution beyond the NHS healthcare system analysed.

Finally, a stakeholder analysis is conducted for each phase of the solution pathway identifying key health, financial and structural outcomes. The stakeholder analysis methodology by Smith L.W.[41] is employed in carrying out this analysis, identifying all external and internal stakeholders. Moreover, stakeholder project impact diagrams estimation priorities are employed to identify barriers towards implementing the proposed solution. Also, the participation matrix will be analysed to identify the level of engagement of each stakeholder with our solution and to optimise the delivery of the solution to individual stakeholders. The detailed analysis and mathematical modelling of the impact assessment are discussed later in Chapter 5.

Chapter 4

Implementation {Hengyu Wang}

4.1 Introduction

The introduction of a groundbreaking real-time QRS detection algorithm [42] by J. Pan and W. J. Tompkins in 1985 initiated the beginning of the machine learning-based era for detecting arrhythmias in electrocardiograms (ECGs). Recently, machine learning approaches have shown tremendous potential in automatically classifying various cardiac abnormalities. In this research, we presented a robust method for atrial fibrillation (AFib) prediction from electrocardiograms, combining signal processing, time-frequency feature engineering, transfer learning, and ensemble methods. The PhysioNet 2017 Challenge dataset [31], collected by MIT and Boston's Beth Israel Hospital, served as the basis for our experimental evaluation. Our methodology consists of the subsequent phases:

1. Applying Butterworth or Wavelet Denoiser for noise reduction;
2. Developing a baseline model using a bi-directional LSTM trained on raw ECG data (input size: 1×9000);
3. Extracting time-frequency (TF) attributes and employing the same Bi-directional Long Short-Term Memory Network (Bi-LSTM) with these TF characteristics (input size: 4×255);
4. Generating scalograms and fine-tuning pre-trained CNN models through transfer learning;
5. Implementing an ensemble-based consensus voting strategy as our final model.

The proposed final model demonstrated a significant improvement in performance relative to the

baseline model, achieving greatly elevated levels of accuracy, sensitivity, and specificity.

4.2 Data

The MIT-BIH (Massachusetts Institute of Technology-Beth Israel Hospital) Arrhythmia Database is a renowned and widely utilized resource for cardiac arrhythmia research within the biomedical field, which is publicly accessible through https://physionet.org/challenge/2017. The PhysioNet 2017 Challenge dataset includes 12,186 single-lead ECG recordings, each lasting around 30 to 60 seconds. All ECG recordings have been resampled to a uniform sampling rate of 300 Hz, enabling consistent analysis and comparison among various algorithms. Expert annotations label 718 ECGs as atrial fibrillation and 4,937 as normal instances. This unequal data distribution can present multiple challenges during model training and evaluation phases. Machine learning algorithms might favor the majority class to achieve high classification accuracy on the training set, potentially leading to overfitting and reduced generalization capabilities on the validation set (as indicated by a high false-negative rate in our research, with the majority being normal). To mitigate this issue, we used cost-sensitive training and dataset resampling methods, which we will explore further in later sections.

The histogram displayed in Figure 4.1 reveals that most signals consist of 9,000 samples. During training, the **trainNetwork()** function in MATLAB [43] separates the dataset into mini-batches, then truncates or pads signals to achieve uniform length. Excessive padding or truncation could adversely affect the network's performance, as the model might incorrectly interpret a signal due to the presence of additional or missing data. To reduce the impact of excessive padding or truncation, we ensure ECG signals are segmented to maintain a uniform length of 9,000 samples. Signals shorter than 9,000 samples were excluded. For signals longer than 9,000 samples, we divided them into as many 9,000-sample fragments as possible and disregarded any remaining samples. For instance, an ECG signal with 18,100 samples would be separated into two 9,000-sample fragments, and the remaining 100 samples would be discarded.

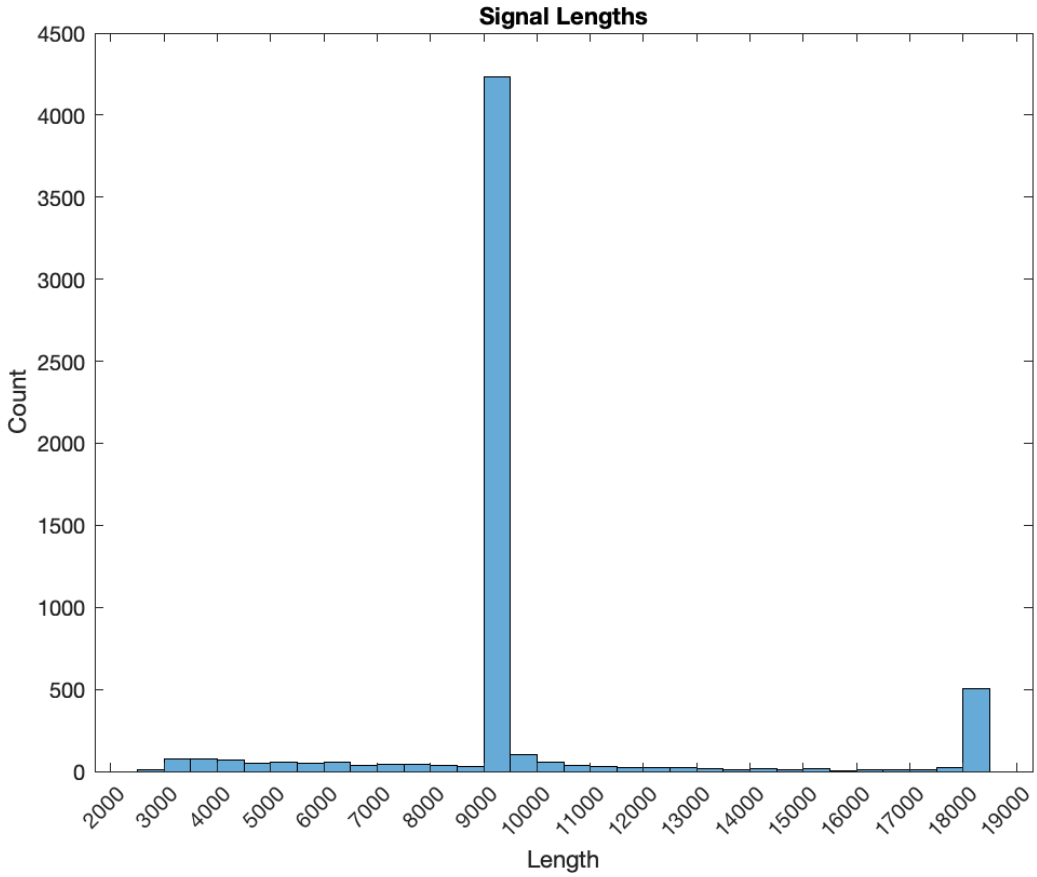


Figure 4.1: Distribution of ECG signal lengths.

Before exploring various machine learning algorithms, let's examine a fragment of one ECG signal from each of the two classes using a selected set of examples (AFib: A0008, normal: N0200). Through visual comparison, we can make the following observations from figure 4.8:

- Irregular rhythm: Contrary to the sinus rhythm, which displays regular intervals between consecutive beats, atrial fibrillation features irregular and unpredictable R-R intervals.
- Missing P waves: Rather than the uniform and smooth P waves observed in a normal sinus rhythm, the ECG trace of an individual with AF might indicate an absence of P waves.
- Slender QRS complex: Atrial fibrillation ECG's QRS complex morphology is usually normal, but when the ventricular rate is too fast and intraventricular conduction differences occur, the QRS complex appears deformed and widened.
- Tachycardia: Atrial fibrillation is frequently linked to an elevated heart rate (shorter period).

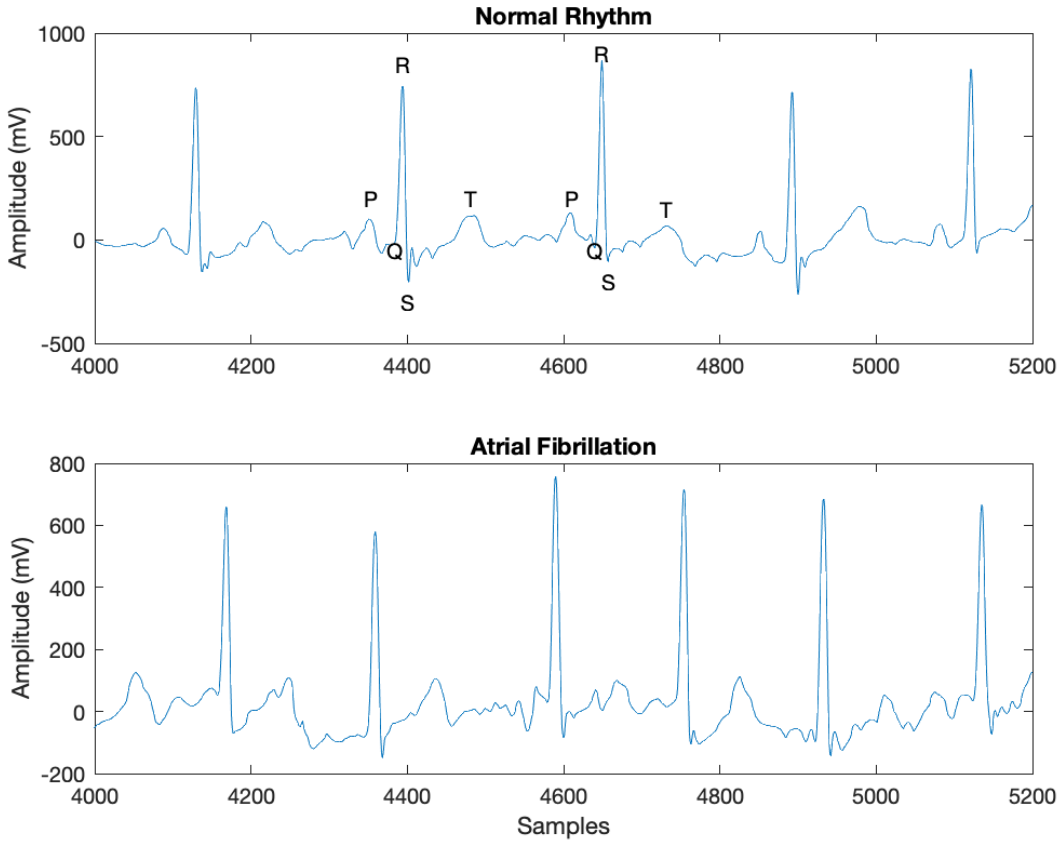


Figure 4.2: Segments of ECG data from each class. The P, Q, R, S, T points are labelled on the normal rhythm example.

In 2021, a study [44] led by Dimitris Bertsimas at the MIT Operations Research Center focused on extracting 880 features from time-domain representations, guided by recommendations from medical professionals. In contrast, our research approach relies on signal processing knowledge from electrical engineering courses rather than medical expertise. Initially, we trained an end-to-end bidirectional Long Short-Term Memory (LSTM) network using raw ECG data. Following this, we extracted frequency-domain features through signal processing methods, spectrograms, and scalograms. Next, we used the RGB images of scalograms to fine-tune several pre-trained large-scale convolutional neural networks. Ultimately, we implemented a consensus-based mechanism for making final predictions.

4.3 Probabilistic Modelling

Our problem involves binary classification, so the target function is logistic function. The probability that a sample x represents atrial fibrillation (1 or positive) is given by:

$$p(x) = \frac{1}{1 + \exp\left(-\theta^\top x\right)}, \tag{4.1}$$

where x is the vector embedding of an ECG signal from the second last layer of neural networks. (i,e: before the final classification layer).

Our objective is to minimize the objective function, which is the cross-entropy loss between the true labels $q(x)$ and the predicted probabilities $p(x)$:

$$Cross - Entropy\ Loss(p, q) = -\frac{1}{n} \sum_{x} q(x) \log p(x) \tag{4.2}$$

Nonetheless, given the imbalanced nature of the dataset, we implemented two strategies to tackle this issue: dataset resampling and cost-sensitive training.

4.3.1 Dataset Resampling

Data resampling techniques aim to achieve a balanced class distribution by either increasing the number of examples in the underrepresented class (referred to as oversampling) or decreasing the number of examples in the overrepresented class (known as undersampling). In our study using bidirectional LSTM, we equalized the number of normal and AFib signals by eliminating excess examples in the normal dataset. This decision was made due to computational constraints, as all the models in this research were trained on a single CPU of a MacBook Pro.

4.3.2 Cost-sensitive Training

In the forthcoming sections where we introduce our research involving the application of transfer learning and pre-trained convolutional neural networks, we utilized the class-weight method [45], which is a cost-sensitive training technique that tackles data imbalance by allocating distinct weights to each class. This method adjusts the entropy cost layer of learning algorithms to account for the data category imbalance, placing greater emphasis on the minority class during training.

For a binary classification problem with two classesclass 0 (majority, normal sinus rhythm) and class 1 (minority, AFib rhythm)the weights for each class can be determined as follows:

$$w_0 = \frac{n}{2 \cdot n_0} = 0.8731$$

$$w_1 = \frac{n}{2 \cdot n_1} = 0.1269$$

Here, n denotes the total number of examples, while n_0 and n_1 represent the number of samples in the normal and AFib classes, respectively. These weights are then integrated into the learning algorithms' cross-entropy loss function. Consequently, the cross-entropy loss detailed in the earlier subsection is now modified to the weighted loss function, expressed as:

$$\text{Weighted Cross-Entropy Loss} = -\frac{1}{n}\sum_{i=1}^{n}\left[w_{y_i}\cdot y_i\cdot\log\left(\hat{y_i}\right)+w_{1-y_i}\cdot\left(1-y_i\right)\cdot\log\left(1-\hat{y_i}\right)\right] \tag{4.3}$$

In this equation, y_i represents the groud-truth label of the i-th instance, $\hat{y_i}$ is the predicted probability of being AFib, and w_{y_i} denotes the weight associated with the true class of the i-th instance.

By incorporating the class weights into the cross-entropy loss objective function, the learning algorithms become more attuned to the minority class (AFib), allowing the model to better address data imbalance.

4.4 Evaluation Metrics

Throughout our study, we employed the following metrics to assess various machine learning algorithms: test set accuracy, precision, recall, F1-score, confusion matrix, and receiver operating characteristic (ROC) curves, according to this tutorial [46] by Tom Fawcett.

A *confusion matrix* is a table that summarizes a classification algorithm's performance by contrasting the predicted and the ground truth labels. The confusion matrix for our task, a binary classification problem in essence, is presented below in Table 4.1:

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Table 4.1: Confusion Matrix

Precision quantifies the percentage of accurate positive predictions amongst all positive predictions made, whereas *recall* (a.k.a. *sensitivity*) calculates the percentage of accurate positive predictions amongst all ground true positive examples. Their definitions are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.5}$$

Nonetheless, precision and recall alone offer only a partial perspective on a model's performance. This is where the *F1 score* plays a crucial role. As the harmonic mean of the two metrics, the *F1 score* serves as a more resilient metric when handling imbalanced datasets, such as those found in our MIT-BIH Arrhythmia Database.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.6}$$

The *Receiver Operating Characteristic (ROC)* curve is a visual representation that shows the efficacy of a binary classification system. It is constructed by plotting the true positive rate, also known as sensitivity, on the vertical axis and the false positive rate, equivalent to (1 - specificity), on the horizontal axis. This is done for various decision thresholds, demonstrating the trade-off between sensitivity and specificity for different thresholds.

The *Area Under the Curve* (AUC) score signifies the classifier's overall performance, calculated as the area beneath the ROC curve. An AUC score of unity denotes an ideal classifier, while a score of 0.5 suggests performance equivalent to random guesses. Therefore, a larger AUC indicates better classification capabilities. To elucidate this concept, an exemplar ROC curve is given in Figure 4.3.

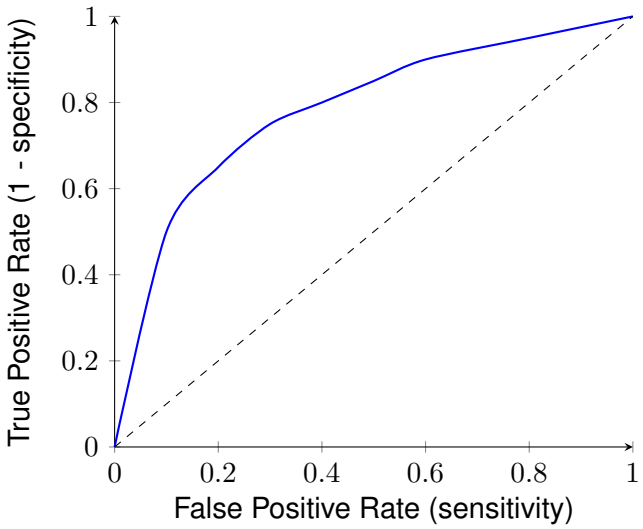


Figure 4.3: Exemplar Receiver Operating Characteristic Curve

4.5 Signal Pre-processing

4.5.1 Z-score normalization

Signal conditioning and pre-processing play crucial roles in cleaning and preparing data for subsequent analysis. They are essential due to the following factors:

- **Baseline Uniformity**: Variations in baseline levels can occur in ECG recordings due to electrode placement, individual patient differences, and recording conditions. Detrending the data eliminates baseline drift, ensuring a uniform starting point for all signals.
- **Improved Comparability**: In the context of machine learning models, it is necessary to confirm that multi-dimensional features have comparable orders of magnitude. Z-score normalization facilitates this by converting the data into a common scale, potentially enhancing performance and yielding more precise results.

The initial signal processing technique, Z-score normalization [47], removes the mean (μ) of the dataset to be zero and makes the standard deviation (σ) be unity. This is accomplished through two primary steps: detrending and standardization.

Detrending involves subtracting the mean from each data point. This is done by computing the empirical mean of the dataset and subtracting it from each data point. The formula for detrending is:

$$\hat{\mu} = \frac{1}{m}\sum_{j=1}^{m} x_j \tag{4.7}$$

$$x'_j = x_j - \hat{\mu} \tag{4.8}$$

In this context, x_j represents the j-th sample of an ECG sequence data. The length of the ECG sequence denoted as m, is 9000. Post-detrending, the sequence data is normalized by dividing it by the empirical standard deviation, symbolized as $\hat{\sigma}$:

$$\hat{\sigma}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \hat{\mu})^2 \tag{4.9}$$

$$z_j = \frac{x'_j}{\sqrt{\hat{\sigma}^2}} \tag{4.10}$$

Figure 4.4 illustrates the comparison between the original ECG signals and the signals after Z-score normalization.

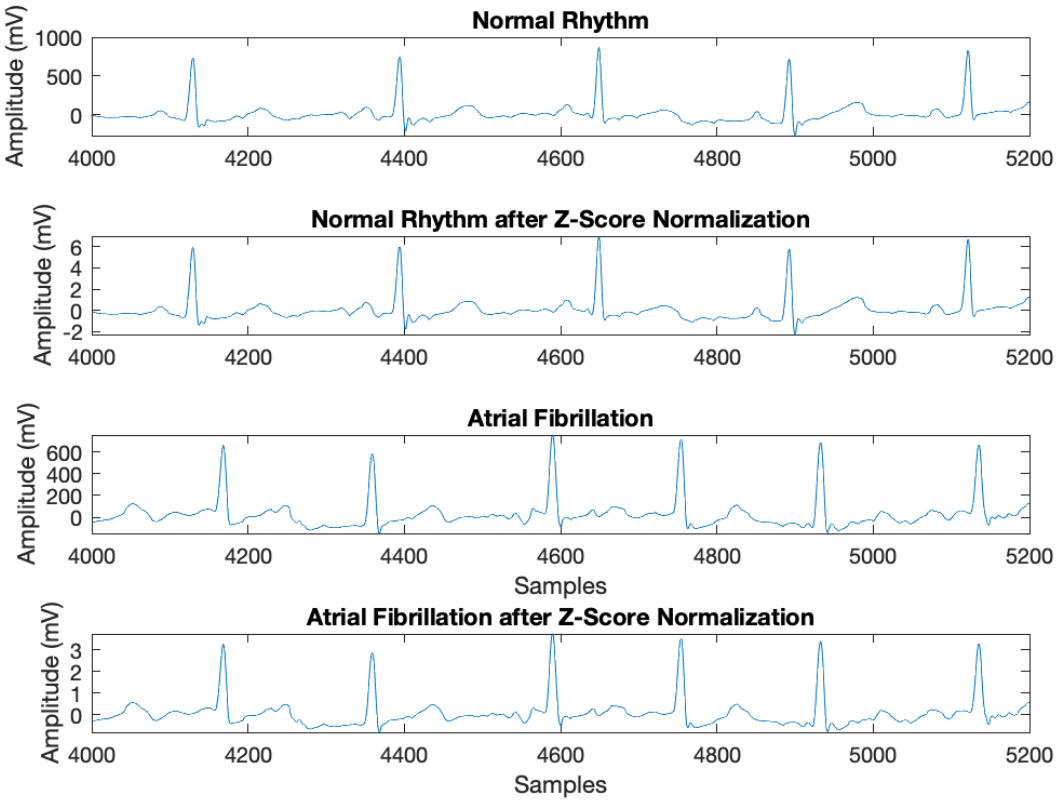


Figure 4.4: Segments of original and standardized ECG data from each class.

After applying both detrending and standardization to transform the original ECG data points, the next step in pre-processing is filtering.

4.5.2 Filtering

During ECG recording, several factors can affect signal quality, such as patient movement and electronic component noise. As a result, filtering the initial recording is crucial for noise reduction in the data. The first step we took was to apply two fourth-order Butterworth filters (a band-pass filter) with the high-pass cut-off frequency of 0.5Hz and the low-pass cut-off frequency of 30Hz.

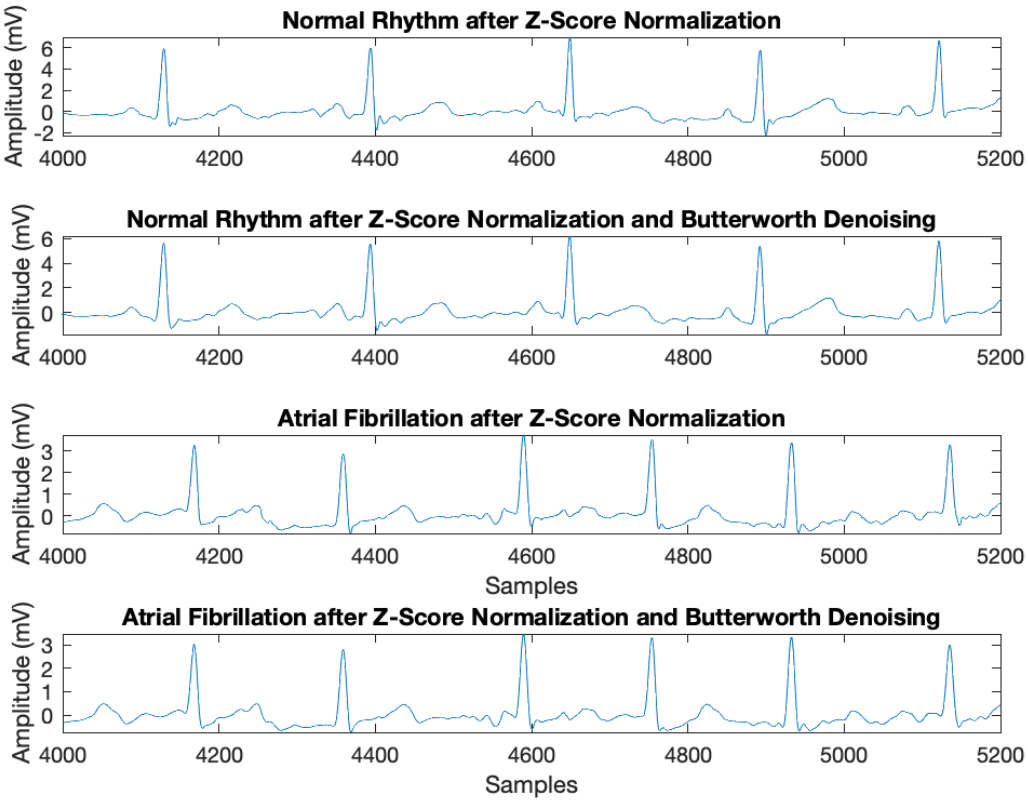


Figure 4.5: ECG data segments from each class (original and Butterworth filtered).

As depicted in Figure 4.5, the noise reduction effect of the two fourth-order Butterworth filters appears minimal. This can be due to the MIT-BIH database already undergoing basic filtering operations, reducing the impact of further noise reduction.

We then applied wavelet denoising [48] as an additional filtering technique. This signal processing method uses wavelet transformations to remove noise from a signal. The procedure involves decomposing the original signal into various frequency components, isolating the noise from the desired signal, and then reconstructing a clean signal by retaining only the relevant components.

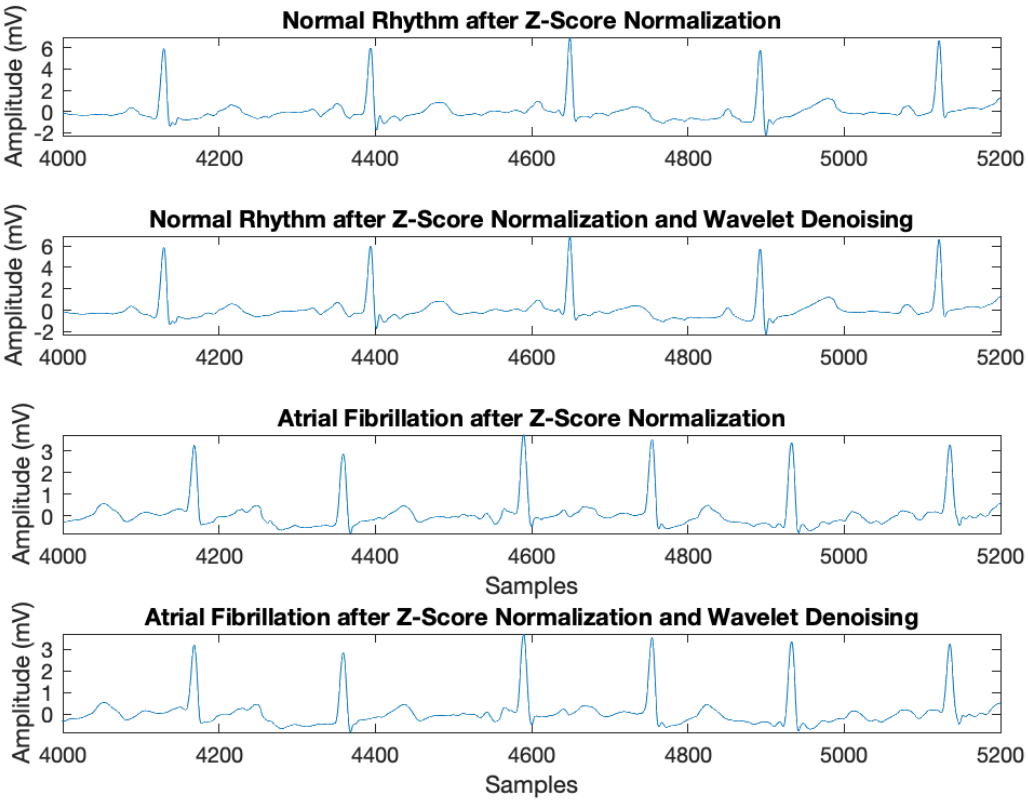


Figure 4.6: Segments of original and standardized ECG data from each class.

As shown in Figure 4.6, we found that wavelet denoising demonstrates an excellent ability to preserve crucial ECG features, such as the QRS complex and T wave, while effectively removing small ripple noise. This performance surpasses that of the Butterworth filter. The wavelet denoising technique is well-adapted to the non-stationary nature of ECG signals [49], as it allows for localized noise reduction across both time and frequency domains. Conversely, band-pass filters operate globally on the signal. Furthermore, wavelet denoising exhibits enhanced robustness against various noise types and is less dependent on specific filter design choices.

4.6 Recurrent Neural Networks Model: Bi-LSTM

4.6.1 Baseline Approach

In our initial study, we used the method from the MATLAB blog [50] for classifying ECG signals using a bi-directional Long Short-Term Memory (LSTM) network [51]. The network takes raw ECG data as input, with each data point consisting of 9000 samples. Our bi-directional LSTM network architecture includes the following layers:

- Sequence input layer with a feature dimension of 1.
- Bi-directional LSTM layer with 100 hidden units.
- Fully connected layer with 2 output units.
- Softmax layer for normalizing class probabilities.
- Classification layer for final class prediction.

The architecture of the bi-directional LSTM network is illustrated in the diagram 4.7 below:

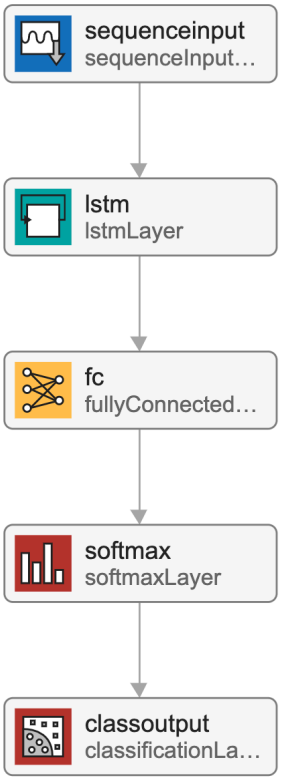


Figure 4.7: Bi-directional LSTM network architecture for ECG signal classification.

We formed a training set by randomly sampling 80% of instances from both the AFib and normal

datasets, with the remaining 20% constituting the test set. To ensure a balanced dataset and shorten training time, we down-sampled the normal samples, resulting in a training set with 574 AFib and 574 normal instances. The test set comprised 141 AFib and 141 normal instances. The training progress plot is shown in Figure 4.8 below.

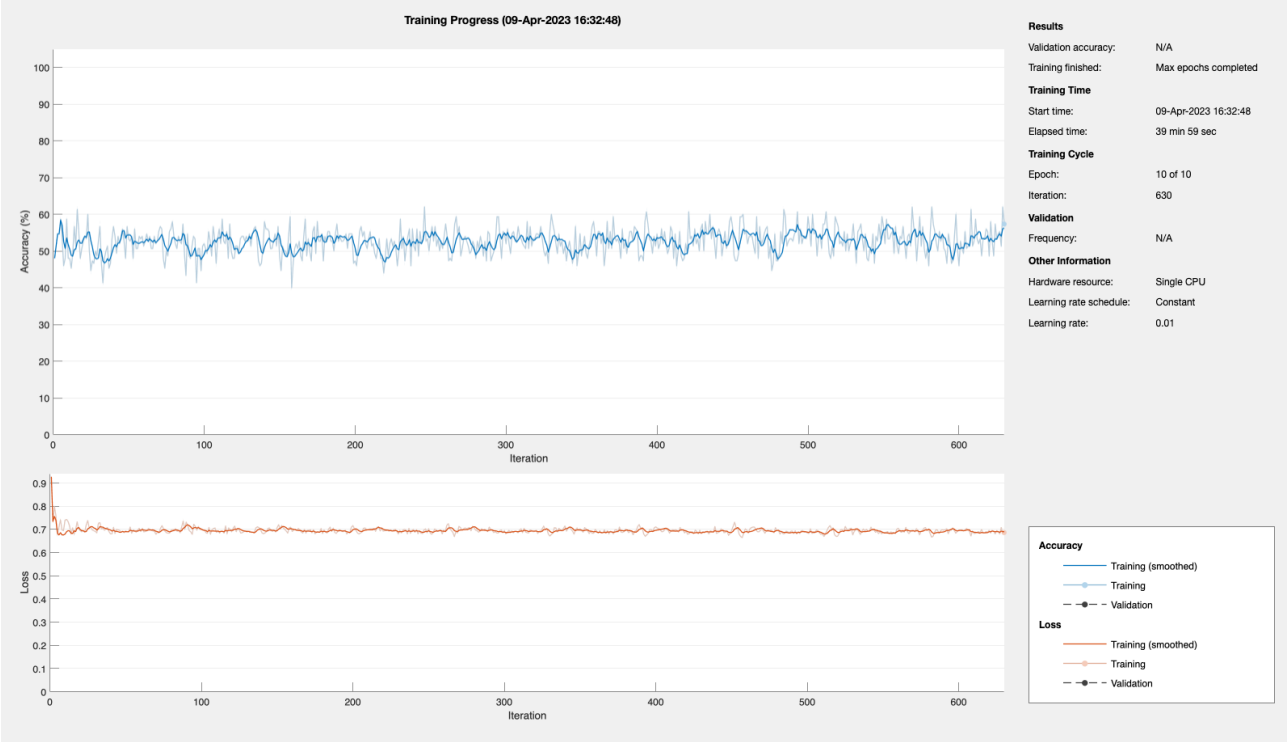


Figure 4.8: Training progress of Bi-directional LSTM network on raw data.

The top subplot displays the training accuracies of each mini-batch. This metric fluctuates between roughly 50% and 60%. After 10 epochs, the training phase lasted for several hours. As demonstrated in the following Figure 4.9 and Figure 4.10, the precision reaches 52.02%, the recall achieves 73.05%, and the F1 score amounts to 60.77%. With an AUC of 0.59, the performance of the initial model is just slightly better than random guessing.

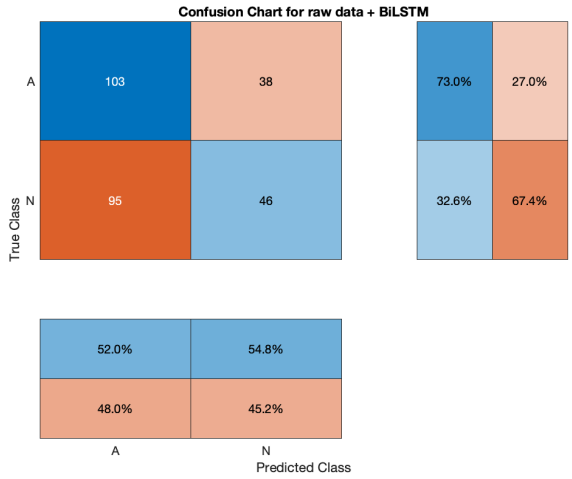


Figure 4.9: Confusion matrix of baseline model

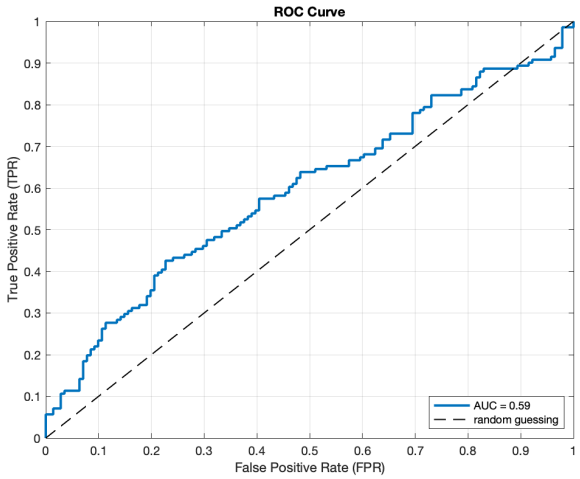


Figure 4.10: ROC of baseline model

Given these results, it is evident that the model requires further refinement.

4.6.2 Model Improvement with Extracted Time-frequency Features

Feature engineering is an important step in the machine learning pipeline that can greatly impact the performance and accuracy of the model. In this section, we explored four distinct frequency-domain features that can be derived from **any** time-series signals: Instantaneous frequency, Spectral entropy, adjusted periodogram, and Welch's power spectral density.

A spectrogram, as shown in Figure 4.11, is a graphical representation that illustrates the evolving frequencies and amplitudes of a signal over a period. This is achieved by segmenting the signal into smaller portions and computing the frequency spectrum for each section. The resulting spectra are organized in a two-dimensional chart, with time represented along the x-axis and frequency displayed on the y-axis.

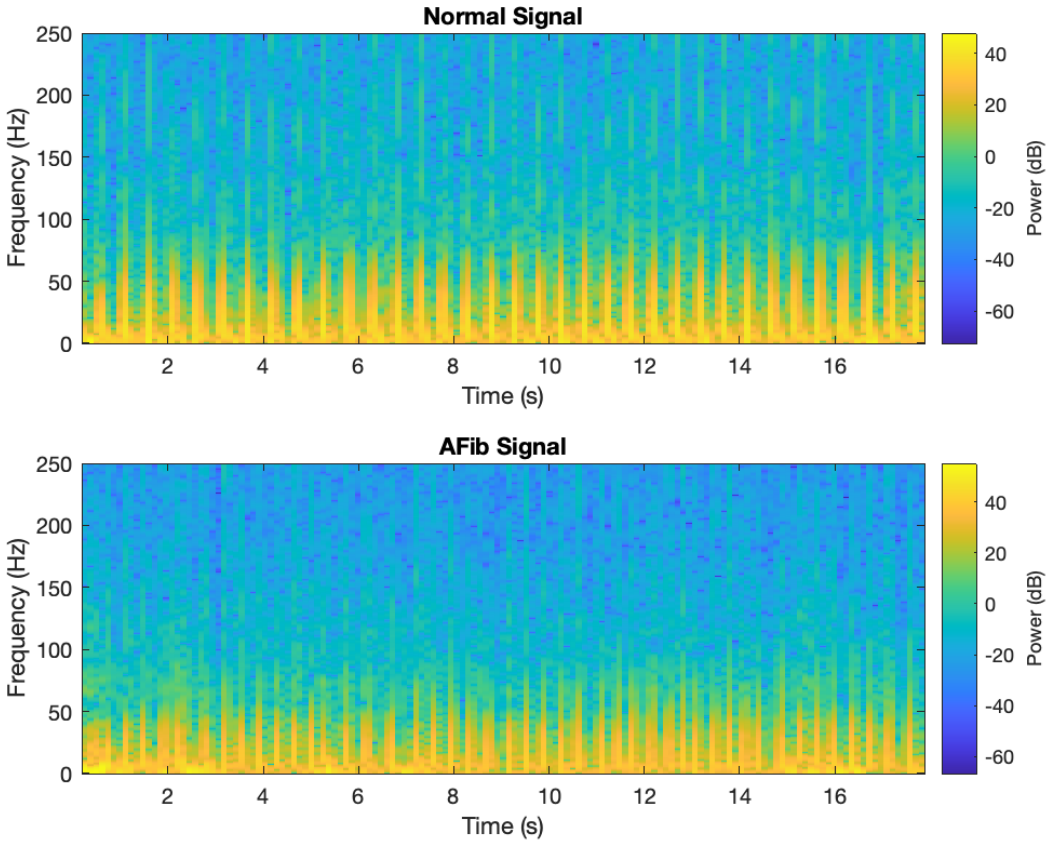


Figure 4.11: Spectrogram of selected normal and AFib examples.

Given that this example employs a Bi-LSTM rather than a CNN, it is essential to modify the methodology for processing one-dimensional signals. In this case, one-dimensional time-frequency (TF) moments are utilized as input features for the bi-directional LSTM model.

- Instantaneous frequency [52] [53]: a measure of how quickly the phase of a signal is changing over time.
- Spectral entropy [54]: a measure of the randomness or unpredictability of the spectral content of a signal.

The subsequent Figure 4.12 and Figure 4.13 depict the instantaneous frequency and spectral entropy of the two selected ECG examples for illustrative purposes.

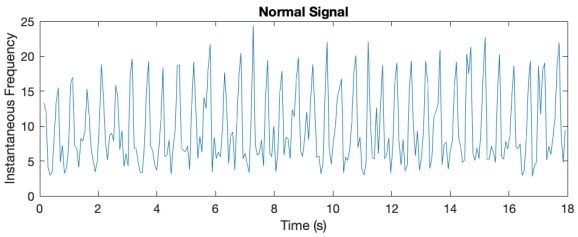


Figure 4.12: Instantaneous frequency of selected ECG examples.

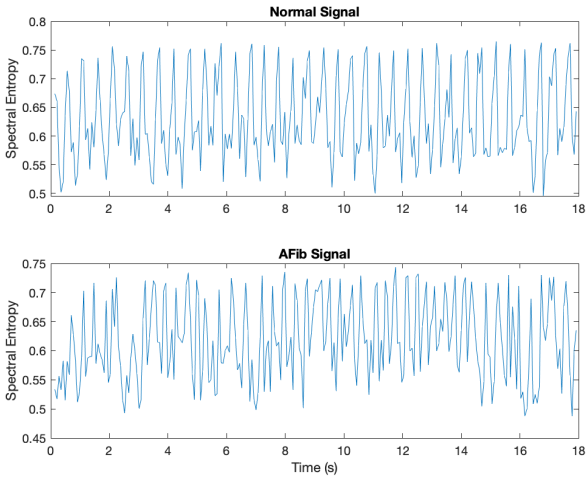


Figure 4.13: Spectral entropy of selected ECG examples.

Moreover, both Welch's method [55] and the modified periodogram [56] have commonly used techniques for estimating the power spectral density (PSD) of a signal. The modified periodogram is a non-parametric approach that computes the discrete Fourier transform (DFT) of a signal by fast Fourier transform (FFT), resulting in an estimate of its PSD. However, the periodogram can suffer from high variance and is not always a consistent estimator of the true PSD. To address this limitation, Welch's PSD estimation method divides a signal into overlapping sections and calculates the modified periodogram for each section. This approach can reduce the periodogram's variance by averaging over multiple segments, leading to a more reliable estimate of the PSD. As shown in the following Figure 4.14, we can observe several key differences between normal and AFib examples:

1. In a standard electrocardiogram (ECG) signal, the power spectral density (PSD) estimation often reveals distinct peaks at low frequencies, representing the heart rate, with considerably higher amplitude than other peaks in the PSD. Additional minor peaks may be present at higher heart rate harmonics due to respiration and various physiological processes.

2. On the other hand, the PSD estimation of ECG signals captured during atrial fibrillation (AF) might display a variety of frequencies with similar amplitudes. This characteristic signifies the chaotic and unpredictable electrical behavior in the heart during atrial fibrillation, which deviates from the regular sinusoidal pattern observed in healthy heart activity.

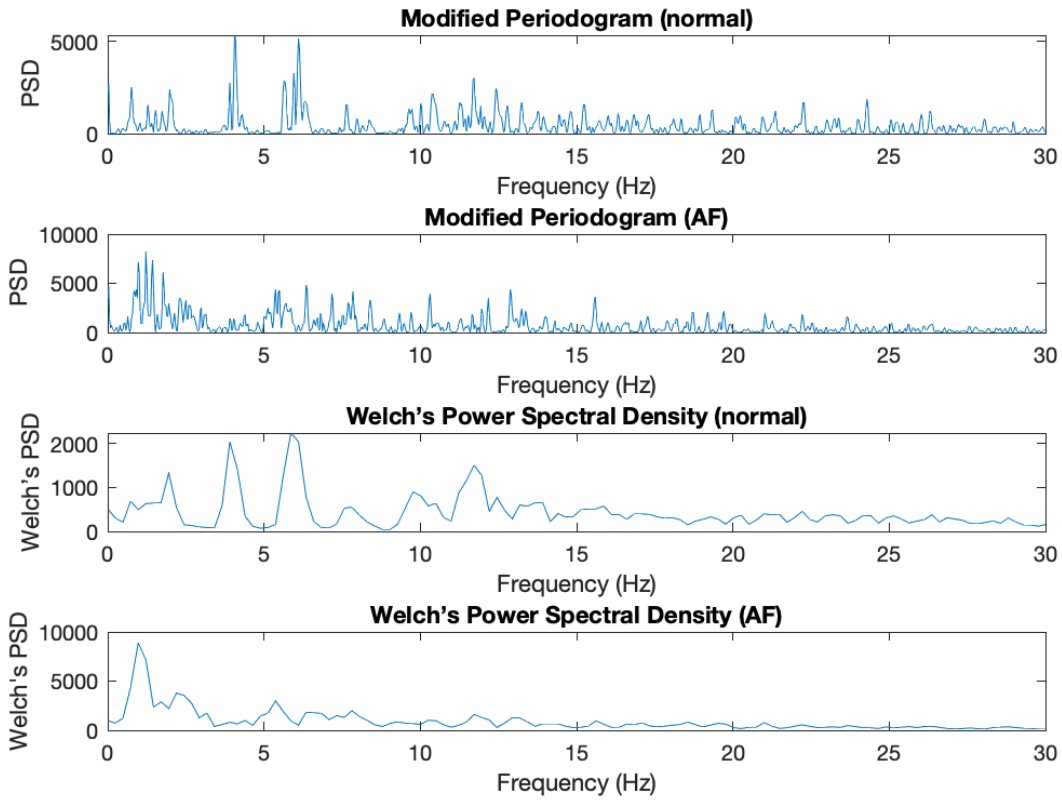


Figure 4.14: Periodogram and Welch's PSD Estimation of selected examples.

Now, switch our discussion back to machine learning. During the training phase, the four features are concatenated, with each one being equal in length (consisting of 255 samples). The sole alteration to the Bi-LSTM architecture is the input sequence's dimension, which has increased from 1 to 4. The training process is now significantly faster, as illustrated in Figure 4.15. This improvement in speed can be attributed to the shorter signal length in the updated model255, compared to the baseline model's 9000.

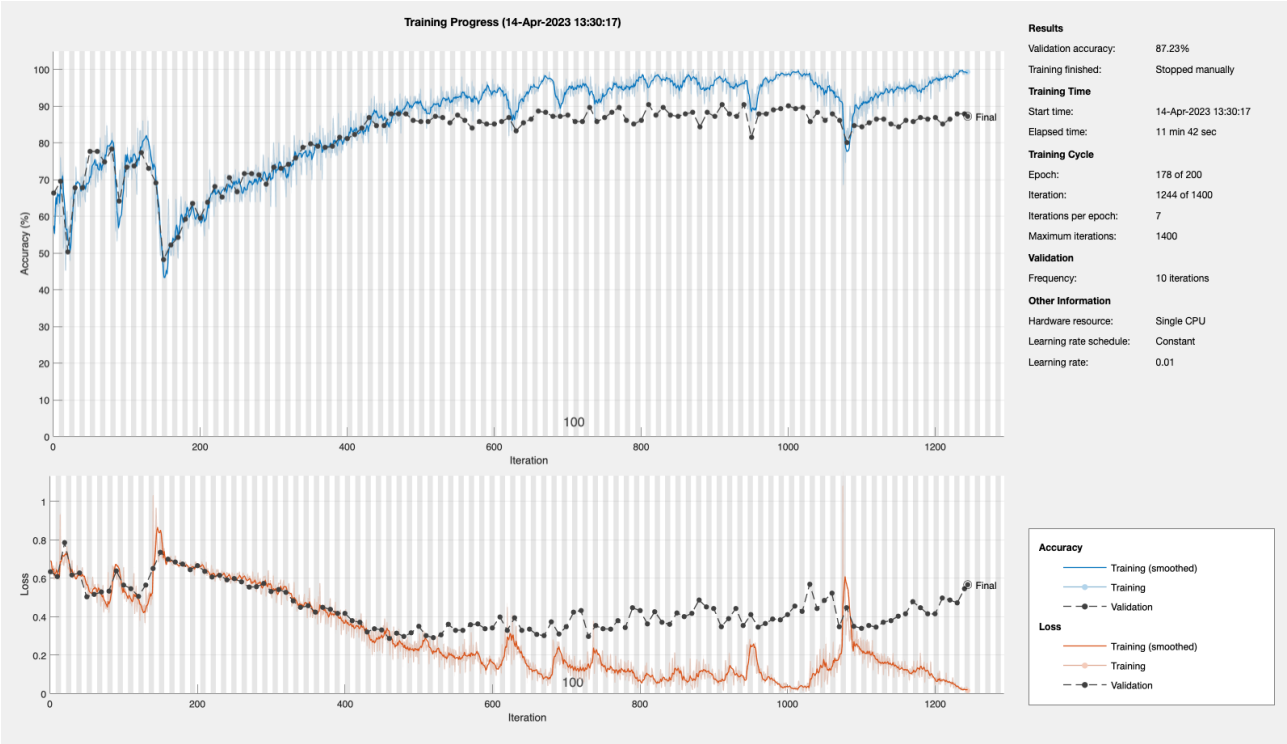


Figure 4.15: Training progress of Bi-directional LSTM network on four time-frequency features.

To further evaluate the impact of the feature extraction method, let's review the performance metrics of the updated model. Based on Figures 4.16 and Figure 4.17 below, the precision is 87.5%, the recall is 89.4%, and the F1 score is 88.4%. Additionally, the AUC value of 0.96 suggests that the extracted features have greatly improved the baseline model's performance.

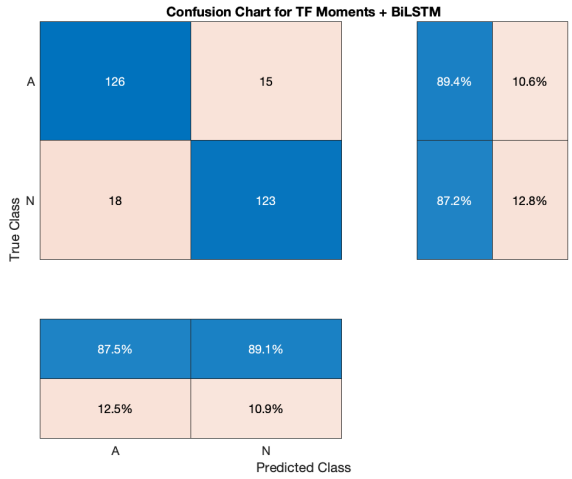


Figure 4.16: Confusion matrix of improved Bi-LSTM model.

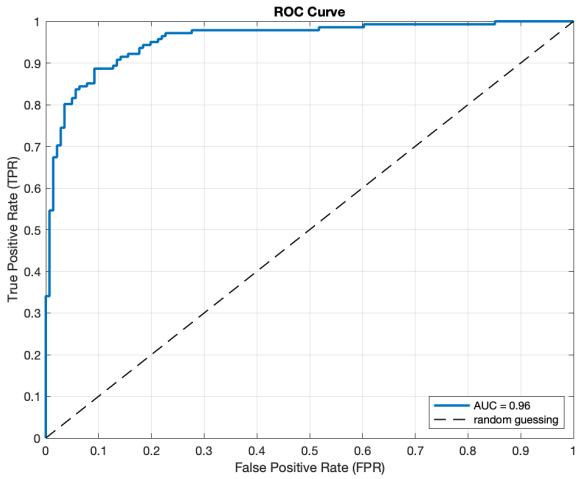


Figure 4.17: ROC of improved Bi-LSTM model.

4.7 Transfer Learning & Convolutional Neural Network Models

By leveraging pre-trained image classification neural networks, we can harness their powerful feature extraction capabilities to tackle new tasks. These neural networks have been trained using the ImageNet database. With experience in processing over a million images, these networks can distinguish between 1000 different object classes, ranging from everyday items like keyboards and coffee mugs to various animal species.

Transfer learning, the technique of applying a pre-trained neural network to new tasks, offers numerous benefits. First, it saves both time and computational resources, as training a neural network from scratch can be resource-intensive and time-consuming. Second, it takes advantage of the knowledge already gained by the pre-trained network, which may generalize well to the new task. This is particularly helpful when the new task has limited data available for training.

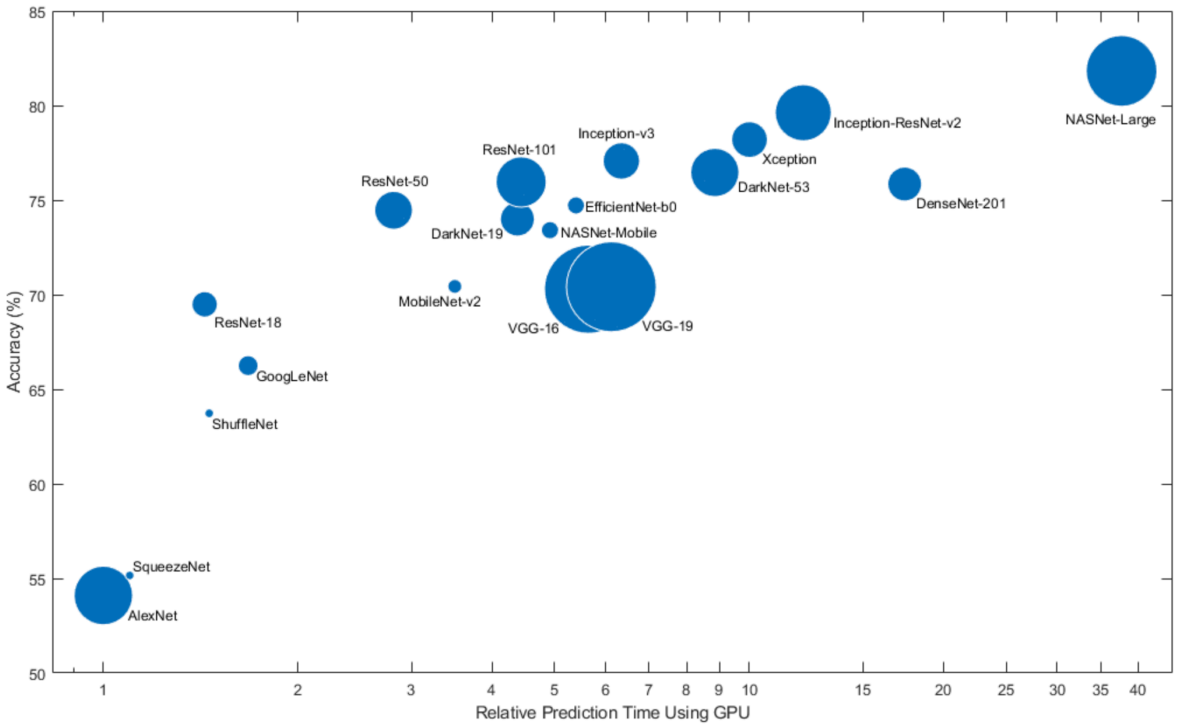


Figure 4.18: Illustration of transfer learning on pre-trained neural networks [57].

In our case, there are only 5655 examples available, while even the smallest pre-trained neural network, SqueezeNet, has 1.24 million parameters. Transfer learning can help overcome this data scarcity issue by leveraging the pre-trained network's existing knowledge, leading to better performance and more efficient training compared to starting from scratch.

In order to apply transfer learning, we followed these steps:

1. **Create scalograms:** We computed a Continuous Wavelet Transform (CWT) filter bank [58]. A scalogram, represented as a 224x224 RGB image, is derived from the absolute value of the CWT coefficients of an ECG signal.

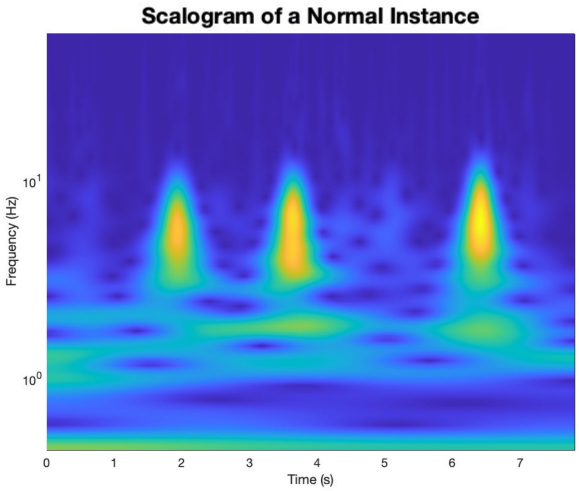


Figure 4.19: Scalogram of a normal instance.

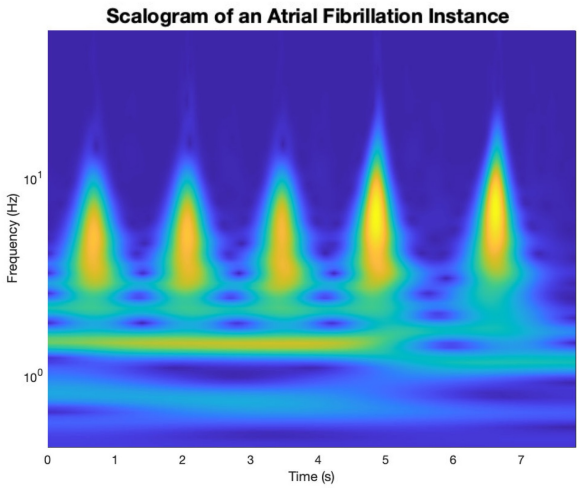


Figure 4.20: Scalogram of an atrial fibrillation instance.

2. **Select pre-trained neural architectures:** We selected SqueezeNet, GoogLeNet, ResNet101, and VGG19 as our models of choice, taking into consideration the balance between accuracy and computational speed required for integration with wearable devices. The sizes of the circles in Figure 4.18 represent the number of parameters for each respective model, with larger circles indicating more parameters.

3. **Apply regularization:** To prevent overfitting and enhance the networks' generalization capabilities, we incorporated a dropout layer with a probability of 0.6.

4. **Modify network structure:** Since the pre-trained network was designed for a specific classification task (ImageNet Large Scale Visual Recognition Challenge), we replaced the output layer with a customized layer tailored to our new task. Instead of downsampling the majority class (normal rhythm), we utilized a customized **weighted-cross-entropy** loss function as the final classification layer.

5. **Fine-tuning** We performed fine-tuning to optimize the networks for our ECG classification task.

4.7.1 GoogLeNet

GoogleNet [59], alternatively known as Inception, is a deep convolutional neural network architecture proposed by Szegedy et al. in their 2014 paper, *Expanding Convolutions for Improved Visual Recognition*. This architecture excelled in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014. The fundamental innovation in GoogleNet is the Inception module, which substantially enhances the network's depth and width while maintaining manageable computational costs. The architecture after our modifications is visualised in Appendix Figrue 7.1.

The training progress, confusion chart, and Receiver operating characteristic (ROC) curve are plotted below.

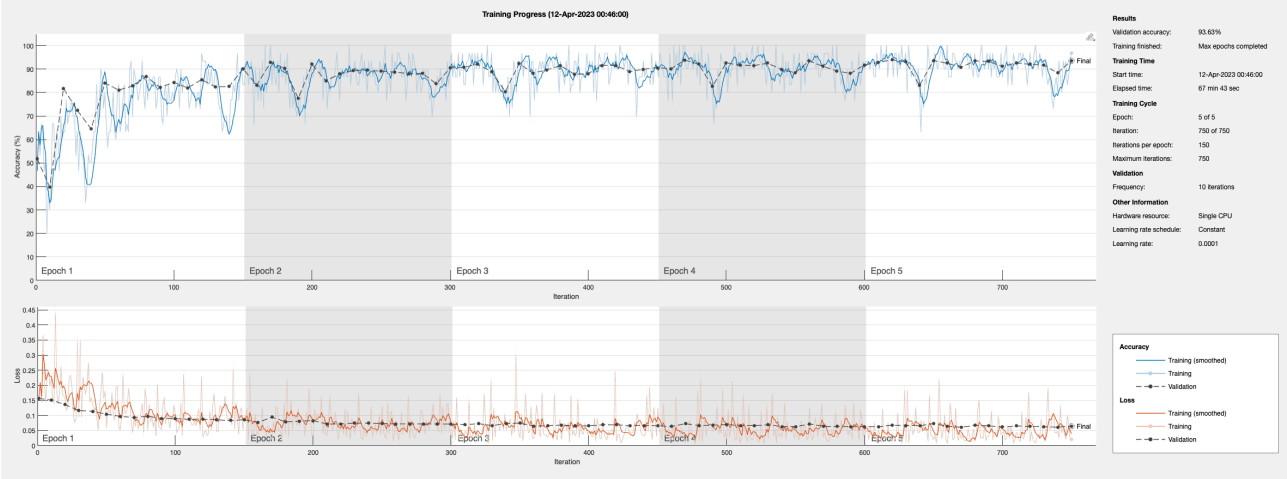


Figure 4.21: Fine-tuning progress of customized GoogLeNet.

Based on the figures presented below 4.22 and 4.23, the updated model achieved an accuracy of 71.7%, a recall of 82.6%, an F1 score of 76.8%, and an AUC value of 0.95.

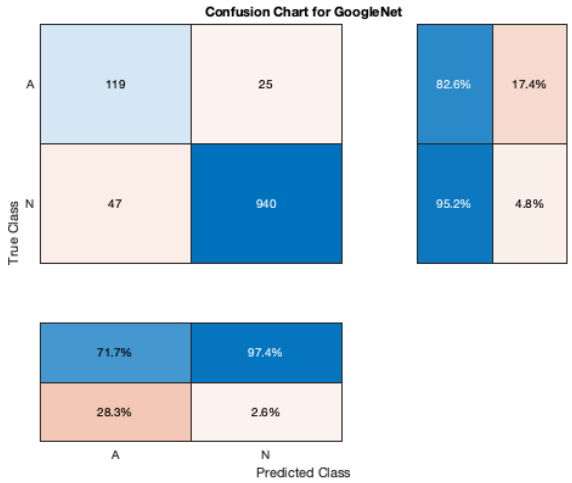


Figure 4.22: Confusion matrix of customized GoogLeNet.

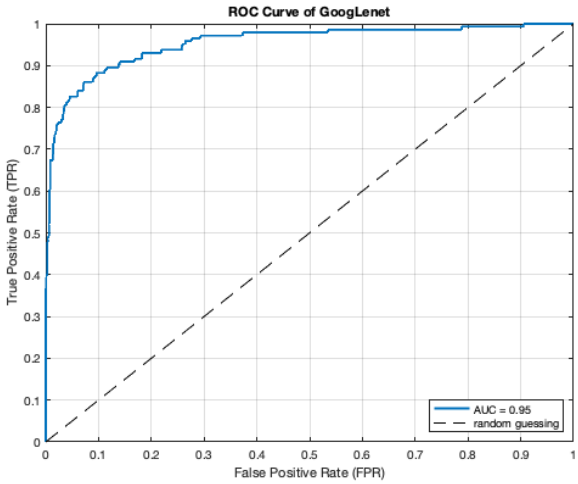


Figure 4.23: ROC of customized GoogLeNet.

4.7.2 SqueezeNet

SqueezeNet [60] is a lightweight deep convolutional neural network architecture introduced by Iandola et al. in their 2016 paper titled *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*. The primary goal of SqueezeNet is to achieve high classification accuracy while minimizing the model size and computation complexity through the use of the Fire module. The architecture after our modifications is visualised in Appendix Figure 7.2.

The training progress is plotted in Figure 4.24 below.

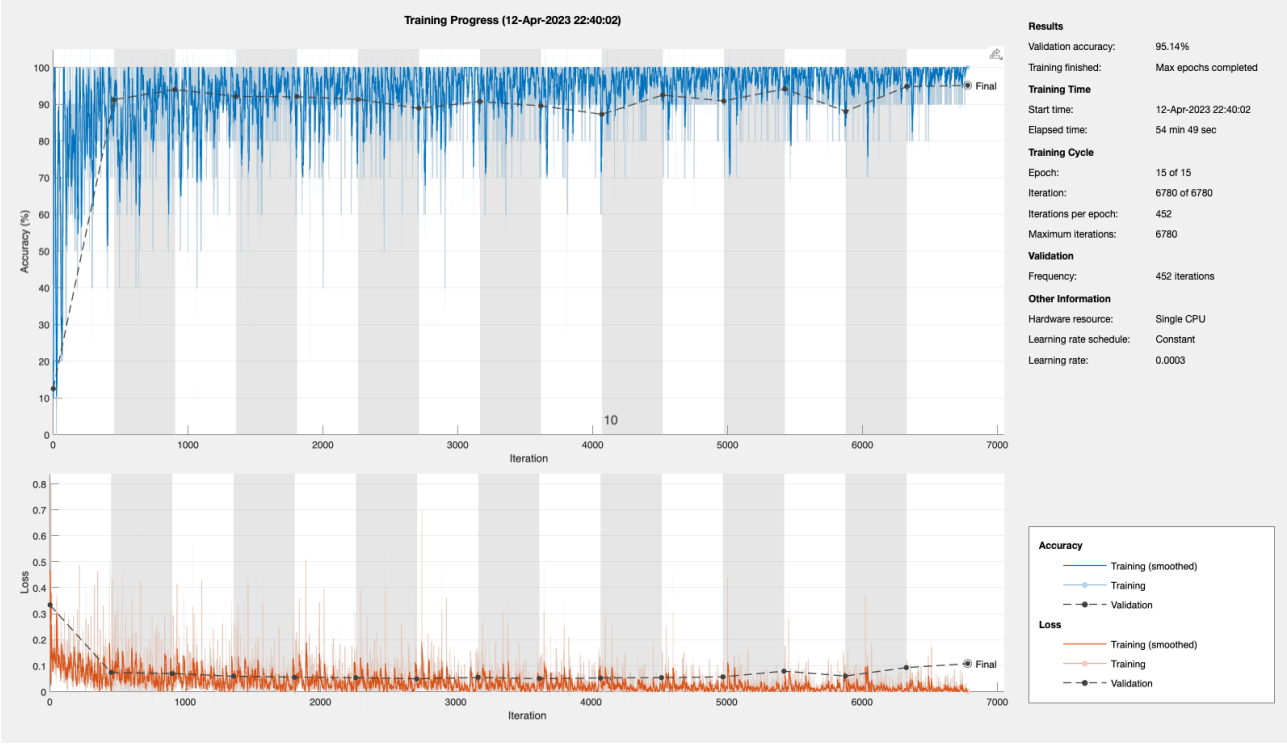


Figure 4.24: Fine-tuning progress of customized SqueezeNet.

As shown in Figure 4.24, the accuracy of the updated SqueezeNet model on the testing set is 95.14%. Furthermore, Figure 4.25 and Figure 4.26 indicate that the model achieved a 79.9% accuracy, an 82.6% recall, an F1 score of 81.2%, and an AUC of 0.97.

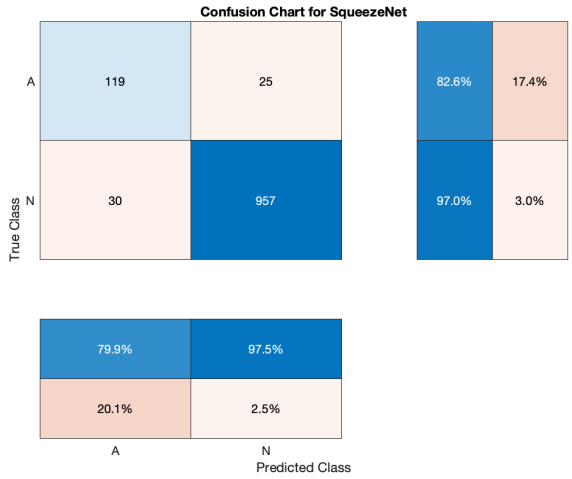


Figure 4.25: Confusion matrix of customized SqueezeNet.

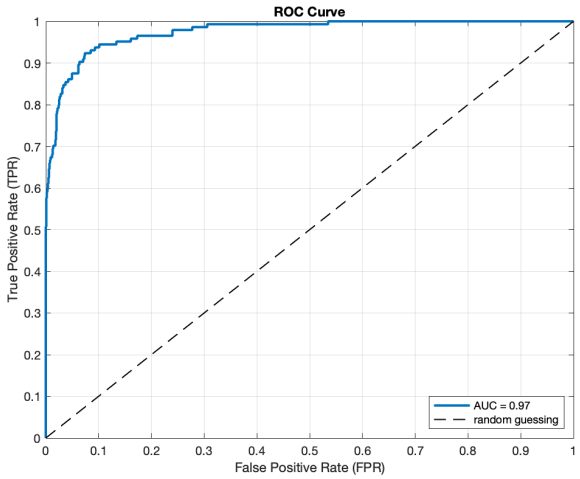


Figure 4.26: ROC of customized SqueezeNet.

4.7.3 MobileNet-v2

MobileNetV2 [61] is a lightweight and efficient deep learning architecture designed for mobile computer vision applications, making it ideal for deployment on devices where computational resources may be constrained, such as wearables in our study. It was introduced by researchers at Google in 2018 as a successor to the original MobileNet architecture.

One of the key innovations in MobileNetV2 is the introduction of the inverted residual structure with linear bottlenecks. This design choice improves the flow of information between layers and reduces the number of parameters without sacrificing accuracy. The architecture also employs a width multiplier and a resolution multiplier to allow developers to create models with different trade-offs between accuracy, size, and latency, depending on the specific requirements of their application.

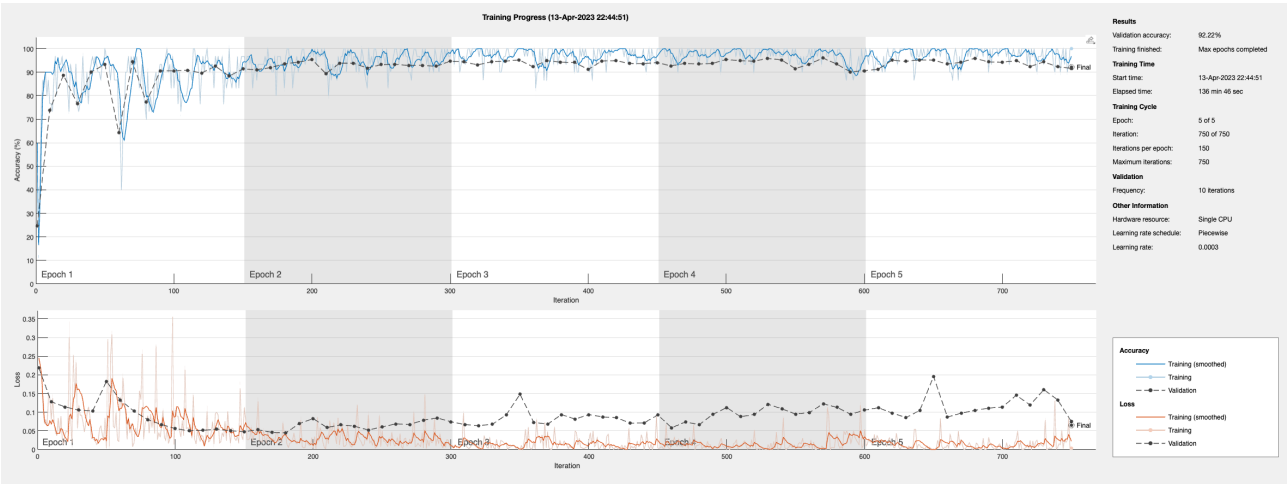


Figure 4.27: Fine-tuning progress of customized MobileNetV2.

Based on the evaluation results shown in figure 4.27, the updated model achieved a prediction accuracy of 92.22% on the testing set. In addition, the figures 4.28 and 4.29 indicate that the model had a 63.6% accuracy, a recall of 91.0%, an F1 score of 74.9%, and an AUC value of 0.97.

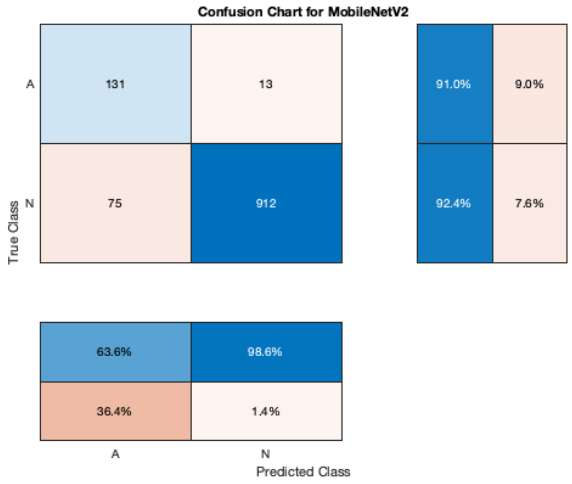


Figure 4.28: Confusion matrix of customized MobileNetV2.

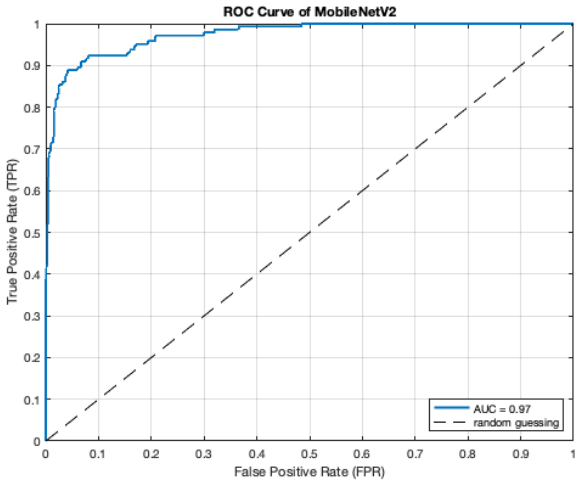


Figure 4.29: ROC of customized MobileNetV2.

4.7.4 NASNet-Mobile

NasNetMobile [62] is another deep learning neural network architecture designed specifically for embedded vision applications. It is a variant of the original NasNet architecture, introduced by Google Brain in 2017.

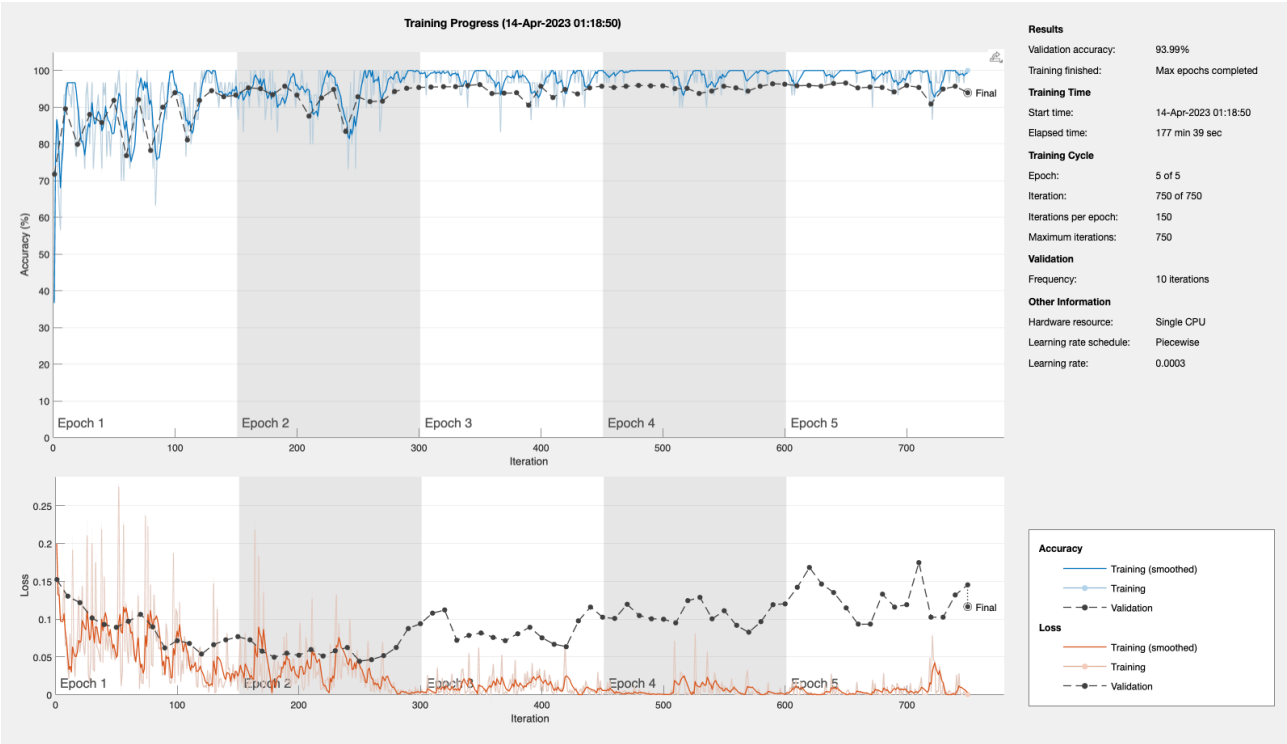


Figure 4.30: Fine-tuning progress of customized NasNetMobile.

NasNetMobile has achieved high accuracy while maintaining a small model size. This makes it an attractive choice for mobile applications. As shown in Figure 4.30, The prediction accuracy on the testing set is 93.99%. As demonstrated in the following Figure 4.31 and Figure 4.32, the accuracy is 71.3%, the recall is 88.2%, the F1 score is 78.9%, and the AUC is 0.97.

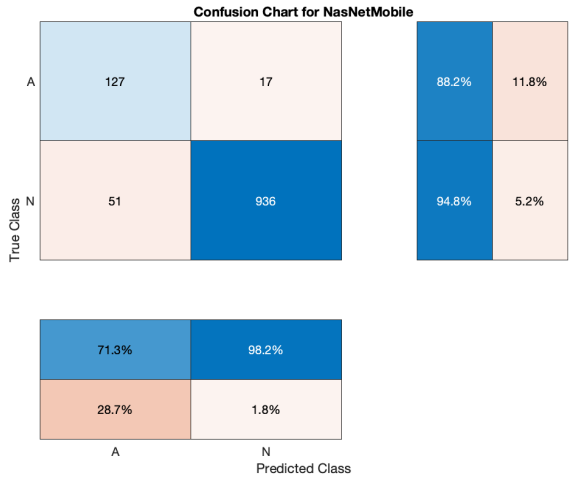


Figure 4.31: Confusion matrix of customized NasNetMobile.

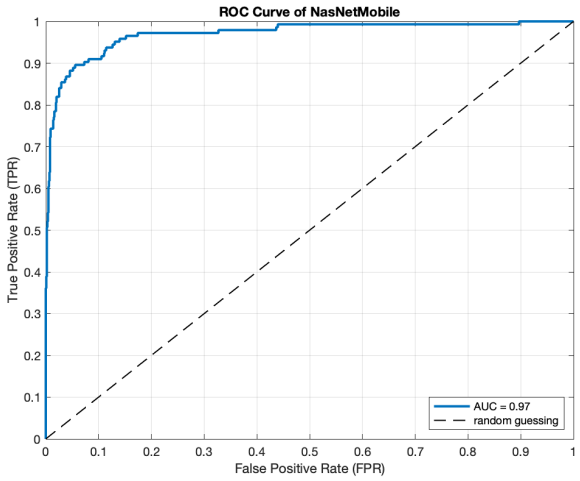


Figure 4.32: ROC of customized NasNetMobile.

4.7.5 Exploring Activation Layers

Understanding the features learned by convolutional neural networks (CNNs) is crucial for interpreting their behavior and performance. In this section, we analysed the activations in various layers of the SqueezeNet architecture, being inspired by the MATLAB blog MATLAB [63]. By comparing the areas of activation with the original image, we are able to discern the features learned by SqueezeNet. Our findings indicated that simple features, such as color and edges, can be learned in the initial layers, while more complex features, such as blobs, are learned in deeper layers.

A convolutional neural network is structured with multiple layers, each containing numerous 2-D arrays called channels. To investigate the features learned, we passed the image A_1.jpg through the network and analyzed the output weights (Figure 4.33) and activations (Figure 4.34) of the first convolutional layer, denoted as 'conv1'.

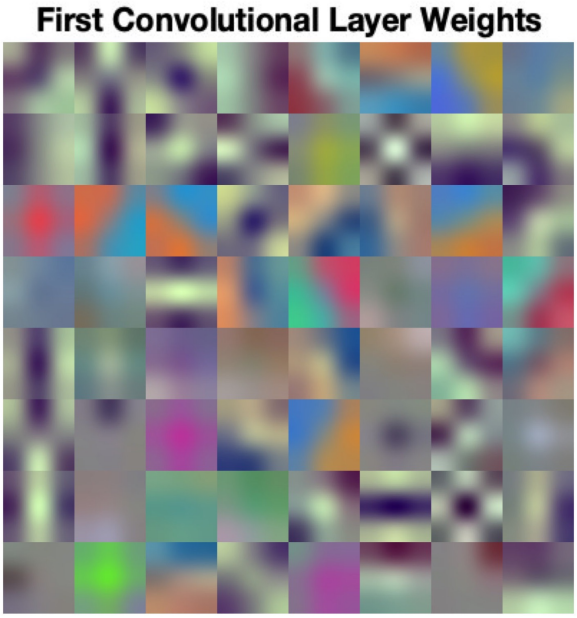


Figure 4.33: Weights of first convolutional layer.

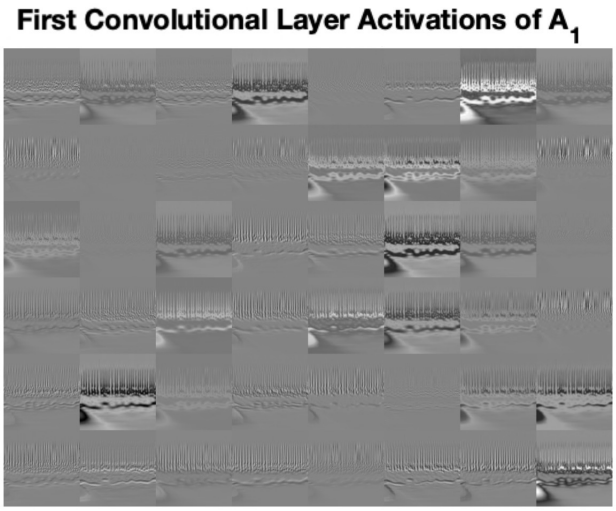


Figure 4.34: Activations of first convolutional layer.

As depicted in the activation map (Figure 4.34), each element corresponds to the output of a channel in the 'conv1' layer. White pixels denote high positive activations, whereas black pixels signal strong negative activations [64]. Channels that appear mostly gray have weak activations in response to the input image. The spatial placement of a pixel within the channel activation aligns with its corresponding location in the original image. We proceeded to identify the channel exhibiting the highest activation, as demonstrated in Figure 4.35.

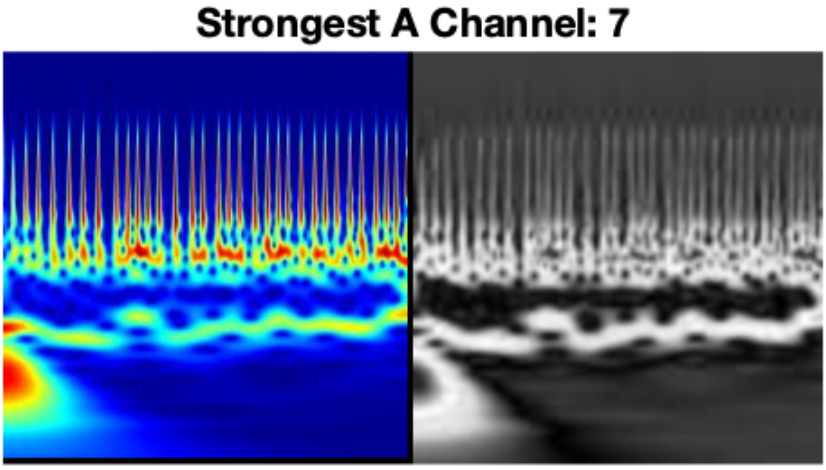


Figure 4.35: Original Scalogram Image and the strongest activation channel in 'conv1'.

It is evident that the channel with the highest activation responds to light green pixels. This observation can be made by noting that the whiter pixels within the channel correspond to the light green regions in the original scalogram image.

In general, convolutional neural networks are designed to identify features such as color and edges within their initial convolutional layer. As the network delves into deeper layers, it progressively learns to detect more complex features. Subsequent layers achieve this by integrating the features learned in earlier layers. To further understand this process, we examined the 'fire8-relu_squeeze1x1' layer using a similar approach as we did for the 'conv1' layer, as illustrated in Figure 4.36.

In contrast to the original scalogram image, the activations within the 'fire8-relu_squeeze1x1' layer effectively highlight regions of the image that exhibit prominent, aggregated, and larger-scale features.

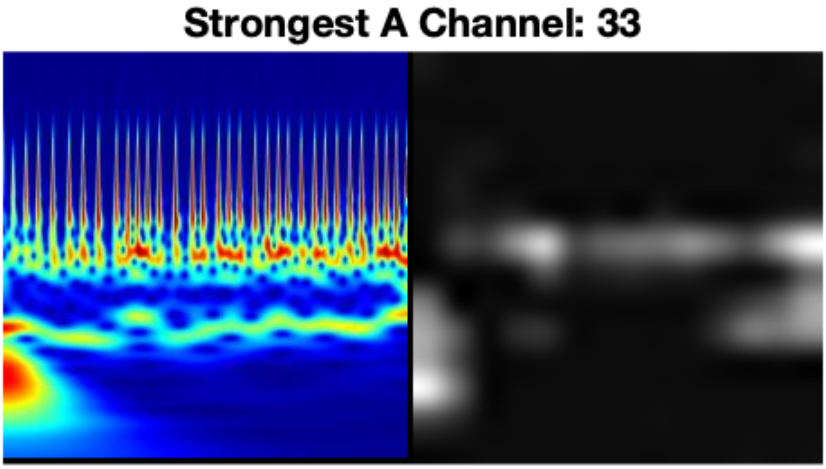


Figure 4.36: Original scalogram Image and the strongest activation channel in 'fire8-relu_squeeze1x1'.

4.7.6 Summary

To summarize our progress, we have performed fine-tuning on four pretrained convolutional neural networks and achieved highly promising results in terms of both specificity and sensitivity. The following Table 4.2 presents the prediction agreements among the four models:

Based on the successful performance of the individual models, we have been inspired to take advantage of their collective strengths by utilizing an **ensemble** technique. This approach can allow us to combine the predictions of the four models to further improve the overall accuracy and robustness of

	Model			
	GoogLeNet	SqueezeNet	MobileNet-v2	NASNet-Mobile
GoogLeNet	-	0.9602	0.9363	0.9416
SqueezeNet	0.9602	-	0.9266	0.9372
MobileNet-v2	0.9363	0.9266	-	0.9240
NASNet-Mobile	0.9416	0.9372	0.9240	-

Table 4.2: Prediction Agreements between Models

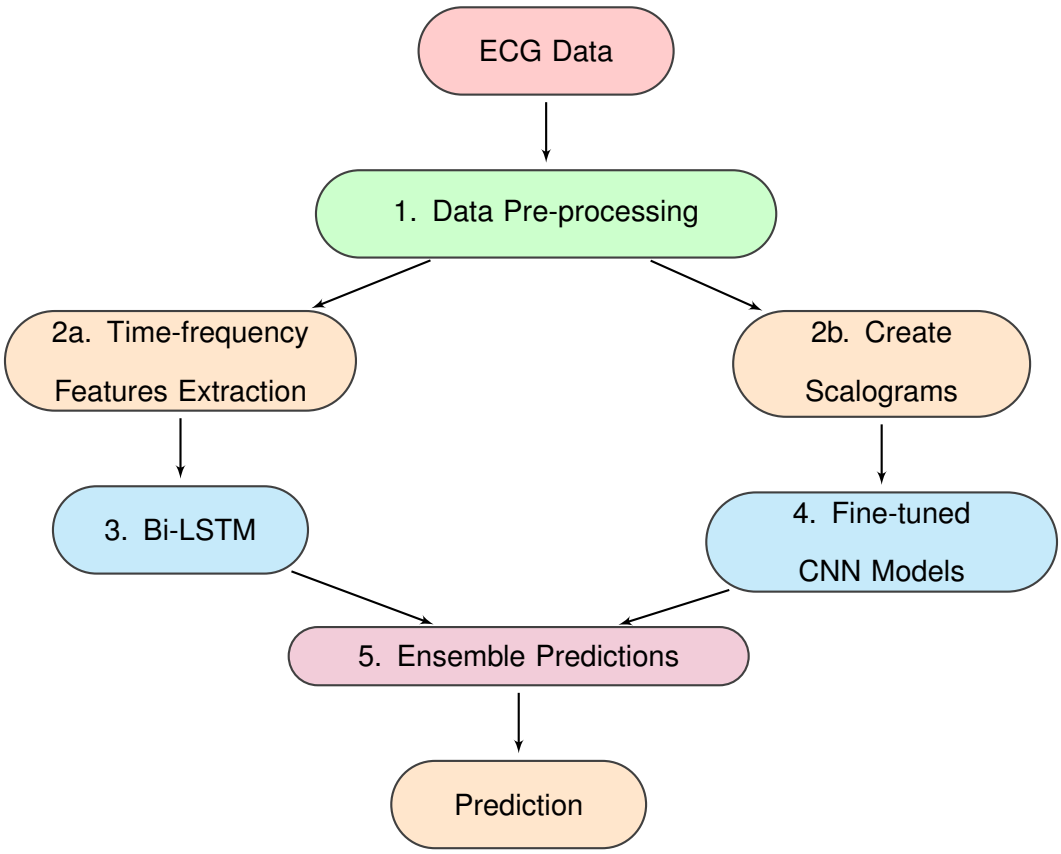
the machine learning system.

Final Model: Ensemble and Voting Mechanism

The final proposed model in our study is an ensemble approach [65] that integrates the capabilities of five distinct models that have been trained or fine-tuned: bi-directional LSTM on time-frequency features, GoogLeNet, SqueezeNet, MobileNet-v2, and NASNet-Mobile. Ensemble neural networks enhance performance and decrease errors by merging predictions from multiple base models, making them a widely used approach in machine learning for obtaining superior outcomes. Key benefits of ensemble neural networks are as follows:

- **Enhanced generalization:** Ensemble techniques average out biases and minimize overfitting through the integration of multiple models' outputs, resulting in improved generalization for unseen data.
- **Greater robustness:** Ensemble networks offer increased resilience against noise and data outliers due to the reduced sensitivity of the combined model to individual base model errors.
- **Lower variance:** By aggregating predictions from several base models, ensemble methods effectively diminish overall variance, yielding more consistent and dependable results.

We present the following steps for any input ECG signals as our model:



1. **Data preprocessing:** Initially, preprocess the input ECG signals through z-score normalization and Wavelet denoising.

2. **Feature extraction:** Subsequently, compute instantaneous frequency, spectral entropy, modified periodogram, and Welch's power spectral density (PSD), resulting in a 4x255 feature matrix.

3. **Bi-LSTM model:** The feature matrix is fed into the trained Bi-LSTM, generating a predicted label, l_1.

4. **Scalogram computation:** We calculate the scalogram of the ECG signal and derive an RGB image.

5. **Convolutional neural network models:** The RGB image is input into the fine-tuned GoogLeNet, SqueezeNet, MobileNet-v2, and NASNet-Mobile models, producing predicted labels l_2, l_3, l_4, and l_5.

6. **Majority voting:** The ultimate output is determined through a majority voting system, reminiscent of democratic decision-making processes in society. This ensemble-based methodology has been shown below to achieve a high area under the curve (AUC) value, signifying exceptional prediction performance.

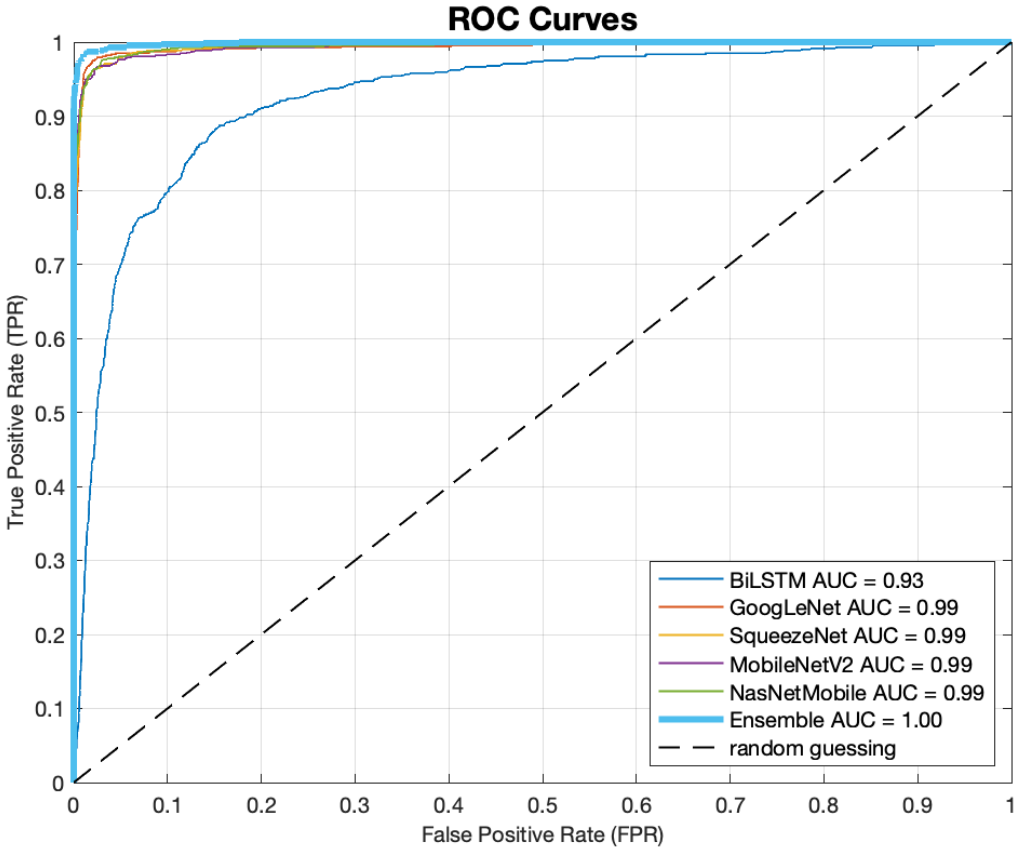


Figure 4.37: ROC curves of various models and final ensemble model

After conducting tests on the entire dataset, we proceeded to evaluate the performance of both the individual and ensemble models by visualizing their corresponding Receiver Operating Characteristic (ROC) curves, as presented in Figure 4.37. As anticipated, the ensemble model exhibited markedly superior performance when juxtaposed with the individual models, hence yielding robust predictions. Remarkably, the area under the curve (AUC) for the ensemble model approached unity, further affirming its strong predictive ability. To better comprehend the results, we have summarized all the evaluation metrics for each model in Table 4.3 below.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Bi-LSTM	87.6039%	50.7%	82.6%	62.8%	0.9262
GoogLeNet	97.9487%	87.9%	97.2%	92.3%	0.9929
SqueezeNet	98.0371%	90.1%	95.0%	92.5%	0.9945
MobileNetV2	94.9779%	72.4%	97.6%	81.3%	0.9943
NasNetMobile	97.0999%	82.9%	97.2%	89.5%	0.9941
Ensemble	98.8329%	93.4%	97.8%	95.5%	0.9988

Table 4.3: Evaluation metrics for various machine learning models

4.8 Demonstration

This section aims to showcase the integration of technical implementation with our AFib AI product, illustrating how the advanced signal processing and machine learning algorithms work cohesively to deliver an efficient and user-friendly solution for detecting atrial fibrillation. All essential resources, including the comprehensive MIT-BIH Arrhythmia Database, pre-processed features and images, detailed coding of our models, and output figures, are safely stored in our Google Drive, which can be accessed via this link. Moreover, the model parameters derived from the training process have been saved in the "model_parameters" folder for future use. These parameters are designed to be redeployed in our wearable devices, facilitating real-time atrial fibrillation forecasting.

Specifically, once a 9000 sample ECG signal is inputted, the function **ensemble_models()** in our codes will output the predicted probabilities and predicted labels given by each of the five models and the final predicted probability and label.

```
%% Demonstration
user_input = 742; % an example ECG input, enter any integer between 1 and 5655.
signal = Signals{user_input};
label = Labels(user_input);

[pred,probs,l1,probs1,l2,probs2,l3,probs3,l4,probs4,l5,probs5] =
ensemble_models(signal,net2,trainedGN,trainedSN,trainedMobileNetV2,
fineTunedNasNetMobile);

disp(['The final prediction by majority: ',pred])
disp(['The ground truth: ',label])
```

Here, “net2,trainedGN,trainedSN,trainedMobileNetV2,fineTunedNasNetMobile” are the saved model parameters of Bi-LSTM, GoogLeNet, SqueezeNet, MobileNetV2 and NasNetMobile.

The command window returns:

```
The final prediction by ensemble: N
The ground truth: {'N'}
```

This demonstrates an instance of accurate classification for the normal class.

Chapter 5

Validation Study {Aaron Chen}

5.1 Introduction

Validation is one of the most crucial steps in developing a medical product or intervention. It provides evidence of the safety and effectiveness of the product to the investigators, as well as invaluable insights into the product's potential market performance and any unforeseen adverse events. As the FDA puts it, "Validation means confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled" [66]. Therefore, a good validation study design is critical to provide accurate and unbiased results to the investigator to support the product as fit-for-purpose. In this chapter, we proposed two separate trial designs for our algorithm and intervention to validate the safety and effectiveness legally and ethically.

5.2 Regulations

In the UK, there are several regulatory validation requirements for a product to be legally marketed and sold, including the Clinical Investigation Notification (CIN) or Clinical Investigation Approval (CIA) from the MHRA. The regulations have changed significantly after Brexit. For example, the CE mark (Conformité Européene) is no longer recognised; instead, a UK-specific certificate, the UKCA (UK Conformity Assessed Marking), was set to replace CE starting in July 2023. The EU MDR (Medical Devices Regulation), EU 2017/745 will cease to be in action by 30 June 2023 in the UK [67]. Due to the product design aims for both the UK and EU markets, we would require to undergo both UK

and EU conformity assessments. Therefore, the following Clinical Validation study designs were in accordance with the UK and EU regulations.

5.3 Design

Our product can be separated into two parts, Atrial Fibrillation (AF) early detection and Pill-in-the-Pocket Anticoagulation guided by the detection algorithm. Since the Pill-in-the-Pocket intervention was a build-up on the performance of the AF detection algorithm, two separate clinical trials were proposed. The first trial, The AFib-AI Study, aims to verify the accuracy and precision of the AF detection algorithm claimed in the Technical section in its intended settings. The trial also plans to test how patients would react to a positive detection, as the increase of the population's awareness of AF is one of the main objectives of the early detection algorithm. The second trial, AFib-AI guided Pill-in-the-Pocket Anticoagulation Study (AI-PiP), builds up on the first trial to test the feasibility of the Pill-in-the-Pocket intervention with Oral Anticoagulation treatments for patients with moderate stroke risks.

The V3 framework, which separates Analytical and Clinical Validations, was adopted in the design of the trial in order to provide a complete picture of the interventions' quality, safety and effectiveness. The V3 framework is now commonly used for Biometric Monitoring Technologies (BioMeTs) and digital health devices in the United States, as requested and suggested by institutes and articles such as BEST [68], SaMD [69], FDA [70], and NASEM [71]. Analytical Validations in the V3 framework focus on the algorithm's performance and ability to measure and detect the targeted conditions [72]. This was already validated with the result matrix available in the Implementation chapter above. Clinical Validation, on the other hand, emphasises assessing the clinical utility in real-world clinical settings and would be the focus of the two clinical trials in this chapter.

We aimed to design both trials as pragmatically as possible to best reflect the effect of the intervention in clinical settings. This would give the policymakers and sponsors a better representation of its impact and provide transparency in design decision-making. The analysis was done by evaluating the trials with the PRECIS-2 (PRagmatic Explanatory Continuum Indicator Summary-2) tool before the design process. It helps trialists make better design decisions that are consistent with the intended purpose of the intervention and identify potential challenges [73]. PRECIS-2 separates the trial into nine domains; each scored on a 5-point Likert scale from 1 being very Explanatory to 5 being very Pragmatic. The PRECIS-2 results are presented in each trial's 'Introduction' section below.

5.4 Trial 1 - The AFib-AI Study

5.4.1 Introduction

An AF detection algorithm on wearable devices is no new territory, with the big technology companies each releasing their own products backed with strongly positive trial results. The Apple Watch and its Apple Heart Study, for example, recorded a specificity of 99.3 per cent and sensitivity of 98.5 per cent for their ECG 2.0 algorithm [74]. Our ensemble algorithm has shown similar performance in the analytical study in table 4.3; however, accuracy alone is insufficient. To be clinically validated, the algorithm must show high precision, or Positive Predictive Value (PPV), and high consistency in its performance. A high precision ensured that an early detection algorithm of a condition with high prevalence would not overload the existing primary care system with a high number of false positive cases. AF, with an estimated 13.1% cases undiagnosed in the population, would pose a huge threat to the existing NHS system if faced with a high false positive rate [75]. In addition, consistency in the algorithm performance is also crucial in clinical settings to give users, clinicians, and investigators confidence in the detection result. The AFib-AI study was therefore designed to measure and evaluate the two objectives in clinical settings.

The AFib-AI Study is a Prospective, Single Arm, embedded Pragmatic Clinical Trial with an overarching objective of evaluating the ability of the AFib-AI detection algorithm. The performance is compared with the gold standard of the clinicians' diagnosis using a 12-lead ambulatory ECG monitor to measure precision and consistency. The pragmatism was shown with the PRECIS-2 Wheel in figure 5.1. The detailed rationale behind the PRECIS-2 scoring can be found in Appendix, Table 7.1. The wheel has shown a high level of pragmatism in both Recruitment and Flexibility; however, due to the study's objective being to test accuracy and precision, the trial outcomes would be explanatory.

5.4.2 Objectives

Primary Objective: To measure the proportion of participants with notifications of AF confirmed by ambulatory 12-lead ECG diagnosis.

Secondary Objective: To estimate the number of appointments participants attended following an AF detection and notification.

The primary objective aims to measure the precision and consistency as described above. The secondary objective, on the other hand, aims to evaluate the potential increase in awareness of AF

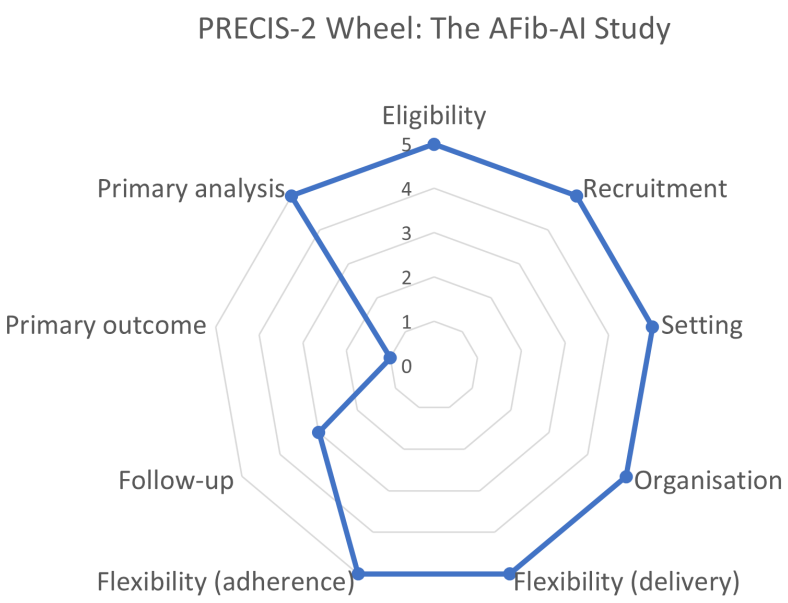


Figure 5.1: Pragmatism of the AFib-AI Study inferred from the PRECIS-2 Wheel.

with the device by assessing participants' adherence in the trial. This is done by calculating the proportion of participants who complied with the study protocol and attended the appointment after a positive detection.

5.4.3 Endpoints

The primary endpoints were designed to address the primary objective. The Precision, or PPV, is calculated by taking the ratio between the number of True Positive cases and the number of Predicted Positive cases, which would be our target for the primary endpoints. The True Positives are cases when the positive algorithm detection agrees with the gold standard 12-lead ECG. Predicted Positive cases can be identified by setting the endpoints when the algorithm detects AF. Consistency is then evaluated by comparing the width of the confidence interval of the precision measured. The complete analysis can be found in the 'Statistical Approach' section.

The secondary endpoints aimed to measure the adherence of the participants. The Apple Heart Study has shown an extremely high proportion of dropout participants, with nearly 44% trials incomplete [74]. It is, therefore, essential to collect data from all participants, including incomplete trials, for analysis to avoid attribution bias. Attribution bias arises when the dropout participants have systematic differences from the participants who finished [76]. Collecting and analysing the dropout participants' data can reveal the reasons for dropout and identify potential sources of bias if the reasons for dropout are related to the study and the intervention. At the same time, this is also a good

representation of how much users would trust and use the device on a daily basis when the device is put on the market.

Primary Endpoint:

1. Confirmed AF diagnosis with an ambulatory 12-lead ECG following AF detection by the algorithm.
2. AF detected with greater than 1 minute ECG recording.

Secondary Endpoint:

1. Self-reported appointment with a health care provider within one month following an AF detection.
2. GP report of an appointment in Electronic Health Record (EHR) within one month following an AF detection.
3. Number of ECG recordings taken within one week following an irregular heartbeat notification.

Tertiary Endpoint:

1. Initiation of therapies for AF, such as oral anticoagulation, rate controlling and anti-arrhythmic therapy.
2. Irregular Heartbeat duration.
3. Self-reported time interval between irregular heart rate notification and ECG monitor measurement.

5.4.4 Study Design

The study was separated into two different phases, The Phase III and Phase IV AFib-AI Trials. The Phase III Trial aims to estimate the Precision value, while the Phase IV trial design focuses on evaluating the consistency of the algorithm. The two phases will share the same design, only with different statistical approaches to the Primary Objective, which gives rise to a significant difference in the required sample size. The Phase III trial requires 4,250 participants, while the Phase IV trial needs at least 50,400 participants with conservative estimates of the dropout rate. Detailed sample size calculation can be found in the 'Statistical Consideration' section. The separate phases enable the investigators to spot potential biases or unforeseen adverse events. This allows changes to be made to the more extensive Phase IV study protocol before the commencement of the trial.

Recruitment of the trial will be conducted entirely virtually due to the Apple ecosystem through the app and email invitations. Participants will be required to agree and sign the study consent form to allow data collection from EHR as well as to complete a questionnaire for baseline data collection.

After enrolment, the participants would be placed in a waiting period of 3 months. During the waiting period, the PPG sensor would continuously check for irregular heart rates. If no irregular heart rate is detected during the waiting period, the participant will finish the trial process with an end-of-study survey for user experience.

For participants with an irregular heart rate notification, they will be required to take an ECG recording for at least one minute for AF detection. Similar to the previous step, if no AF is detected from the AFib-AI algorithm, the participants will return back to the waiting period for the next irregular heart rate notification. On the other hand, if AF is detected, a GP appointment will be arranged, by the app or manually, for a proper 12-lead ECG diagnosis and treatment if diagnosed positive. The complete study design is presented below in figure 5.2.

5.4.5 Study Population

Due to the rising popularity of the Apple Watch and the high number of existing users, the enrolment process was designed to be completely virtual with text and email invitations for Apple users. Learning from both the Apple Heart Study and Fitbit Study, no incentive was required to acquire a large sample size. Participants were willing to test the new algorithm as it does not require any change in lifestyle or habit.

The Pragmatic nature of the trial also suggested loose eligibility criteria with high flexibility in delivery and adherence. However, some basic inclusion and exclusion criteria are still in place; Apple Watch users with age over 22 years old live in the UK. We will exclude patients with a past history of AF or Atrial Flutter in the EHR as well as patients who are currently on anticoagulants or anti-arrhythmia therapy. The complete eligibility criteria are shown below.

5.4.5.1 Inclusion Criteria

Patients must meet all of the following criteria:

- Possession of the following at the time of enrollment, ascertained from automatic software/ device pairing check:
  - iPhone (iOS version 16.1 and above).
  - Apple Watch (Series 4 and above).
- Age > 22 years ago.
- Current resident of the United Kingdom and the EU.
- Proficient in written and spoken English.
- Valid phone number and email address.

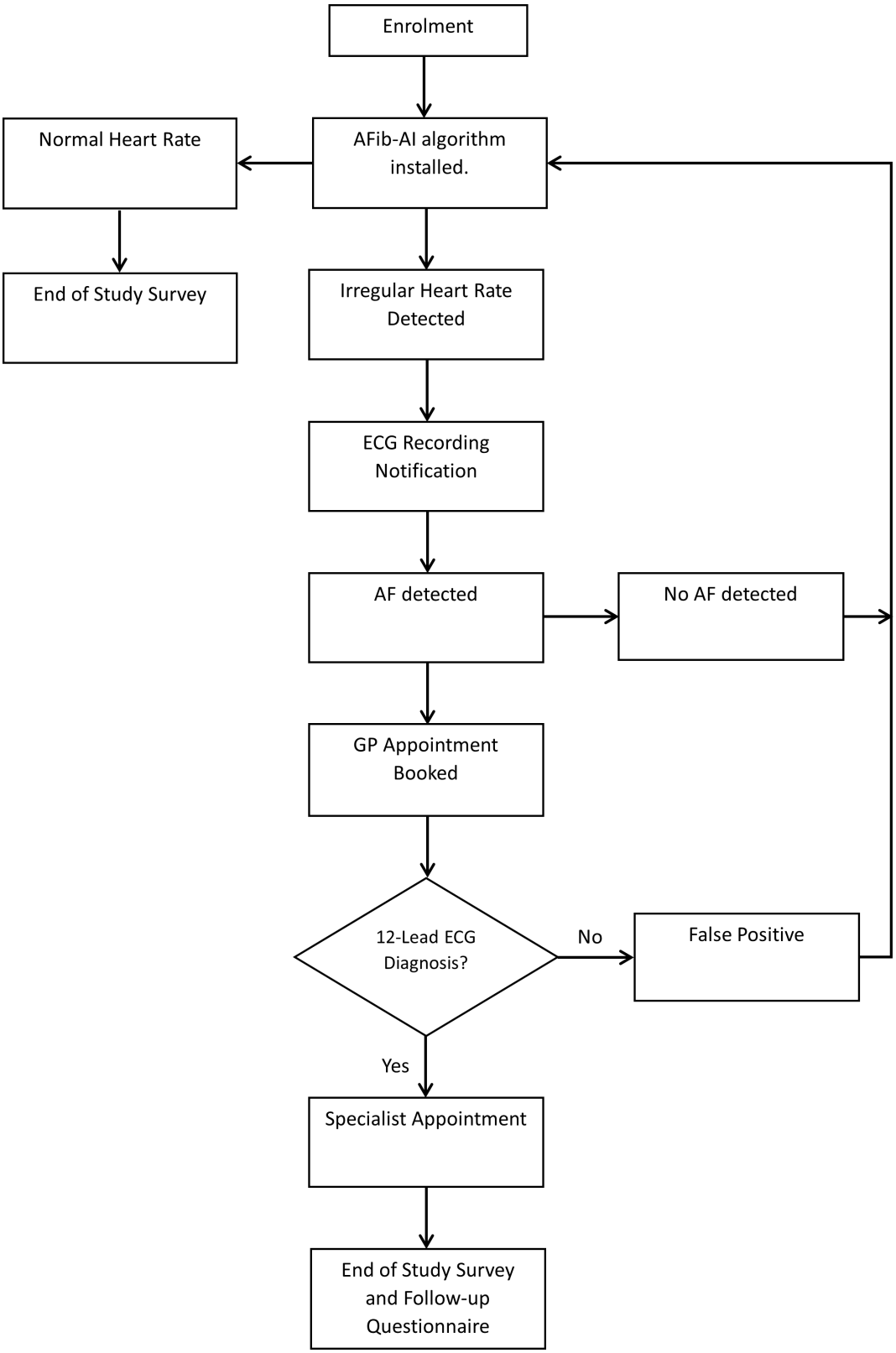


Figure 5.2: Overview of the AFib-AI Study Design.

5.4.5.2 Exclusion Criteria

We will exclude patients with:

- History of Atrial Fibrillation diagnosed by a 12-lead ECG at the time of consent.
- History of Atrial Flutter at the time of consent.
- Currently on anticoagulation therapy.
- Patient with pacemakers or implanted defibrillators.

5.4.6 Statistical Considerations

5.4.6.1 Analysis Sets

The Full Analysis Set (FAS) will include data from all enrolled participants regardless of their adherence.

The Detection Analysis Set (DAS) will consist of all participants who have AF detected by the algorithm with at least 1 minute of ECG signals recorded. The participants in this set would contribute to evaluating the secondary objective.

The ECG Analysis Set (EAS) will consist of participants who are diagnosed with AF with a 12-lead ECG following a positive detection from the algorithm. The primary objective will be evaluated with data collected from participants in this analysis set.

5.4.6.2 Statistical Approaches

The primary objective of the first phase requires the estimation of the algorithm's precision. In this phase, we designed the hypothesis test to compare the resulting precision with the gold standard of 0.75. The value of 0.75 is the gold standard of an AF detection device and is commonly used in studies such as the Apple Heart Study and the Fitbit Study. The detail of the hypothesis are:

Null Hypothesis: The Precision of the AFib-AI algorithm is lower than the gold standard of 0.75.

Alternative Hypothesis: The Precision of the AFib-AI algorithm is greater than 0.75.

We will design the study to ensure an 80% power with a Type I error of less than 0.025. We will conclude that the algorithm's precision is promising if the null hypothesis is rejected and the alternative hypothesis is accepted.

The primary objective of the second phase was to evaluate the consistency of the algorithm's precision after we received positive results from the first phase. The consistency of the algorithm's

precision is ensured by assessing the width of the confidence interval. We have designed the decision rule to conclude that the algorithm is consistent if the 97.5% confidence interval around the estimated precision has a width less than 0.1 or a margin of error (MOE) of 0.05 or smaller.

5.4.6.3 Minimum Sample Size Calculation

The minimum sample sizes required for the trial's two phases were determined by a statistical formula and the nQuery clinical trial design platform.

For the Phase III trial, we anticipate 85 participants with positive ECG detection from AFib-AI from the EAS. Among the 85 participants, we expect 72 to have AF diagnosed and agree with the algorithm detection. The value was obtained by designing an exact a priori proportional binomial test, viewing precision as a proportion. The expected proportion is set to be the gold standard of 0.75, with a conservative estimated proportion of our algorithm of 0.9. The test above has an actual power of 0.9577 and an actual Type I error of 2.197%. With an estimated AF prevalence of 2%, our trial would require at least 4,250 participants assuming full adherence. However, this assumption would not be valid as similar studies, such as the Apple Heart Study [74] and the Fitbit Study, have both shown a significant number of dropouts at nearly 44%. In our design, we assumed a conservative dropout rate of 50% to accommodate the extra follow-ups and appointments required instead of a completely virtual process. This corresponds to the 8,500 participants required overall.

For the Phase IV Trial, the minimum sample size needed for estimating the proportion equation is used. For a target confidence interval of $(1-\alpha)$ and a Margin of Error, d, specified, the probability of the difference between the sample mean and the population mean can be found with equation 5.1, with N referring to the population size, n referring to the sample size, and σ^2 representing the population variance.

$$P\left(\left|\overline{x}-\mu\right|\right) > z_{\alpha/2}\cdot\sqrt{\frac{N-n}{N}\cdot\frac{\sigma^2}{n}} = \alpha \tag{5.1}$$

From this, we can find an equation for the sample size, n, in terms of the MOE, d, as it is equal to half of the Confidence interval width. The population variance σ^2 was estimated by Bessel's Correction of unbiased sample variance and substituted into the equation below to obtain an equation for sample size with only one unknown, N.

$$
\begin{aligned}
z_{\alpha/2} \cdot \sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}} &= d \\
n &= \frac{1}{\frac{d^2}{Z^2_{\alpha/2}/\cdot\sigma^2} + \frac{1}{N}} \\
\sigma^2 &= \frac{N}{N-1} \cdot p \cdot (1-p) \\
n &= \frac{N \cdot p \cdot (1-p)}{(N-1)\frac{d^2}{Z^2_{\alpha/2}} + p \cdot (1-p)}
\end{aligned}
\tag{5.2}
$$

By assuming a large population size so that $(N-1) >> p \cdot (1-p)$ and $\frac{N}{N-1} \approx 1$, the minimum sample size for estimating proportion equation can be found 5.3.

$$
n \approx \frac{Z^2_{\alpha/2} \cdot p \cdot (1-p)}{d^2} \tag{5.3}
$$

Therefore, with a MOE of 0.05, a 97.5% Confidence Interval and a conservative estimated proportion p of 0.7, the minimum number of participants with an AF detection from AFib-AI was calculated to be 424. However, in this trial design, we aimed for the most conservative selection by setting the proportion as 0.5 to maximise the sample size. Therefore, the conservative minimum number of participants with an AF detection we selected was 504. With the same estimation of 2% prevalence and 50% dropout rate, the trial would require 50,400 participants.

5.4.7 Limitations

In this section, we will identify potential challenges and limitations to the study and propose measures to mitigate the effect on the trial result.

5.4.7.1 Bias

Selection Bias may arise as the trial only includes the population already owning an Apple Watch Series 4 or above. The study population may not be a good representation of the general population. The potential solution to the selection bias is to include the population without an Apple Watch and provide them for the duration of the study. This could also be used as an incentive for the study. However, this would also significantly increase the study's cost if we require a large sample size.

Healthy User Bias is the other bias due to the requirement of owning an Apple Watch to participate in the trial. Population with a smart wearable device may generally be more aware of their health and are more health-conscious. This may lead to an underestimation of the effectiveness. The potential

solution is the same as selection bias. In addition, we can make the device more affordable to prevent a high cost.

Reporting bias is when participants are under-reporting unhealthy behaviours during self-reporting. As some data are collected from self-reporting and surveys, we will also collect both baseline data and GP appointment records from patients' EHR to mitigate reporting biases.

Gold Standard Bias was minimised by comparing the algorithm detection to the 12-lead ECG diagnosis instead of another similar algorithm. The gold standard of 12-lead ECG diagnosis would have much better accuracy and consistency.

Attribution Bias was identified as the most significant potential challenge of the trial, as we expect a high dropout rate. Attribution bias arises when systematic differences exist between people who dropped out and those who completed the trial. To mitigate the bias, we separated the trial into two phases to check for the bias. We also implemented an additional dropout group analysis to find potential reasons for dropouts.

5.4.8 Data Handling

5.4.8.1 Analysis of the dropout group

To mitigate and minimise the effect of Attribution Bias, we designed additional measures to analyse the dropout group. To ensure generalisability, initial guesses of the reasons behind the lack of adherence and appointment absences were developed before the trial. The baseline data of the dropout group would also be analysed to find common characteristics that would match our initial guesses. The end-of-study survey would be used to compare to confirm the initial guesses.

We also aim to reduce the dropout rate with more retention approaches such as text and email reminders or app notifications. However, we need to good balance between effective reminders and annoyance. We aim to investigate further for more effective retention using digital means as suggested by multiple trialists [77].

5.4.8.2 Duplicated Participant Identification

Conducting the recruitment process virtually would greatly simplify and reduce the overall cost. On the other hand, virtual recruitment makes it hard to identify and avoid duplicate participants. For example, if a participant changed the email address, phone number, or Apple Account, we would analyse the same participant twice and potentially overestimate the result.

To avoid duplicate participants, extra identification data such as patient ID from the EHR, the NHS number, and the Device MAC address are collected. Baseline details such as email address, full name, and date of birth would also be collected during enrolment with consent.

Data, by themselves, are not enough. A simple algorithm will be implemented to identify potential duplicated participants. The algorithm will check for duplicate information across participants and assign a weight based on the features in common. If the sum of the weight reaches a certain threshold, we would identify the two participants as a duplicate pair and exclude one from the EAS for statistical analysis.

With the extra Data collection and the duplication identification algorithm, we can spot and eliminate the potential duplicate pairs before statistical analysis, providing us with more accurate outcomes. In this trial, we do expect some duplicate participants. However, we assumed that it would not exceed 5% of the overall sample size as there is no direct incentive for this trial. In the AHS, a similar study with no direct incentive, a duplicated amount of 4.4% was identified out of all 438,635 participants [74].

5.5 Trial 2 - AFib-AI guided Pill-in-the-Pocket Anti-coagulation Study (AI-PiP)

5.5.1 Introduction

The 'Pill-in-the-Pocket' strategy is not a novel intervention in AF management. NHS currently suggests a 'pill-in-the-pocket' pathway only for patients with paroxysmal AF guided by implantable cardiac monitors (ICM). The REACT.COM pilot study, for example, showed the effectiveness of this strategy guided by ICMs with a 94% reduction in anticoagulants [78]. However, the pilot study only had a sample size of 59 and did not have the statistical power to prove the safety of the strategy. In addition, the invasiveness of ICMs may intimidate patients due to the perceived dangers, including infections and discomforts at the implantation site. The AI-PiP was therefore proposed and designed to assess the safety of the 'pill-in-the-pocket' strategy as well as its effectiveness when it is guided by non-invasive wearable devices.

A pilot study was not required as iCARE-AF, a similar pilot study of pill-in-the-pocket strategy guided by Holter monitor, has shown positive results and a warrant was granted to more extensive studies [79].

The AI-PiP Study will be a non-inferiority randomised controlled trial with a 1-to-1 ratio. The design was again guided by the PRECIS-2 tool with the resulting wheel shown in figure 5.3. The rationale behind each domain score is attached in Appendix Table 7.2.

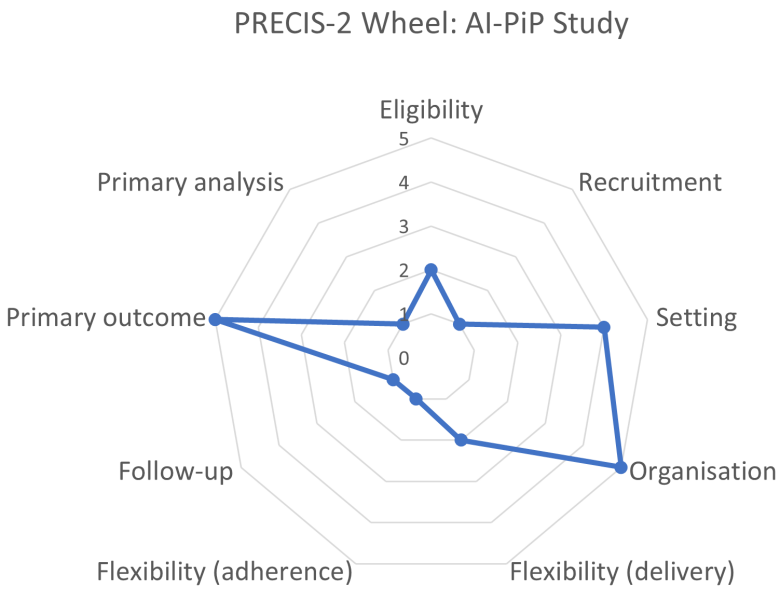


Figure 5.3: Relatively Explanatory design suggested from the AI-PiP Study PRECIS-2 Wheel analysis.

The PRECIS-2 wheel has suggested an explanatory trial with pragmatism in the primary outcome, organisation and settings. The explanatory nature was expected as the trial would require strict eligibility criteria as well as high adherence for participants.

5.5.2 Objectives

Primary Objective: To evaluate the safety and feasibility of the 'Pill-in-the-Pocket' anticoagulation intervention guided by the AFib-AI algorithm.

Secondary Objective: To measure the amount of time reduced in patients' duration on anticoagulants compared with continuous oral anticoagulants pathway.

The overarching objective of the trial is to evaluate the safety and effectiveness of the pill-in-the-pocket strategy guided by the AFib-AI algorithm. The objectives will be evaluated with a non-inferiority trial by comparing with the usual continuous treatment.

5.5.3 Endpoints

To address the objectives and the hypothesis test, the primary endpoints were set to measure the rate of stroke and bleeding. The stroke rate in the treatment group will be compared with the gold standard stroke rate in the controlled group for the hypothesis testing. The secondary endpoints will be used to measure the reduction in the number of days on oral anticoagulants in order to address the secondary objective. Any cardiovascular deaths and survival rates were also included in the secondary endpoint to check for adverse events and withdrawals in the trial.

Primary Endpoint: [12 months time frame]

1. Incidence of Stroke or Arterial Embolism.
2. Incidence of Bleeding.

Secondary Endpoint: [12 months time frame]

1. Number of days on oral anticoagulants.
2. Cardiovascular caused Deaths.
3. Overall Survival rate.
4. Duration of AF detected.

5.5.4 Study Design

After enrolment, the participants will be split randomly at a one-to-one ratio into the control group and the treatment group. The control group participants will undergo the usual care of continuous anticoagulants suggested by the cardiologists. Regular follow-ups will occur once every three months to check for adverse events and AF progression. Patients with AF who progressed to permanent AF would no longer be eligible for the study and would have to be included as part of the withdrawal group. The treatment group will undergo the 'pill-in-the-pocket' strategy presented in figure 5.4. There will be continuous AF monitoring guided by the PPG module and the AFib-AI algorithm to detect an episode of AF for more than 1 hour within a 24-hour period. If such an episode of AF is detected and confirmed with the ECG algorithm, the patient would be suggested to take the oral anticoagulant for a minimum of 30 days. If no episode is detected for 30 days consecutively, the patients can stop the anticoagulant until an episode is detected. The participants are required to log the dates when starting or stopping anticoagulants on our AFib-AI app calendar. The treatment group will also have regular follow-ups once a month to check for AF progression or any increasing stroke risk.

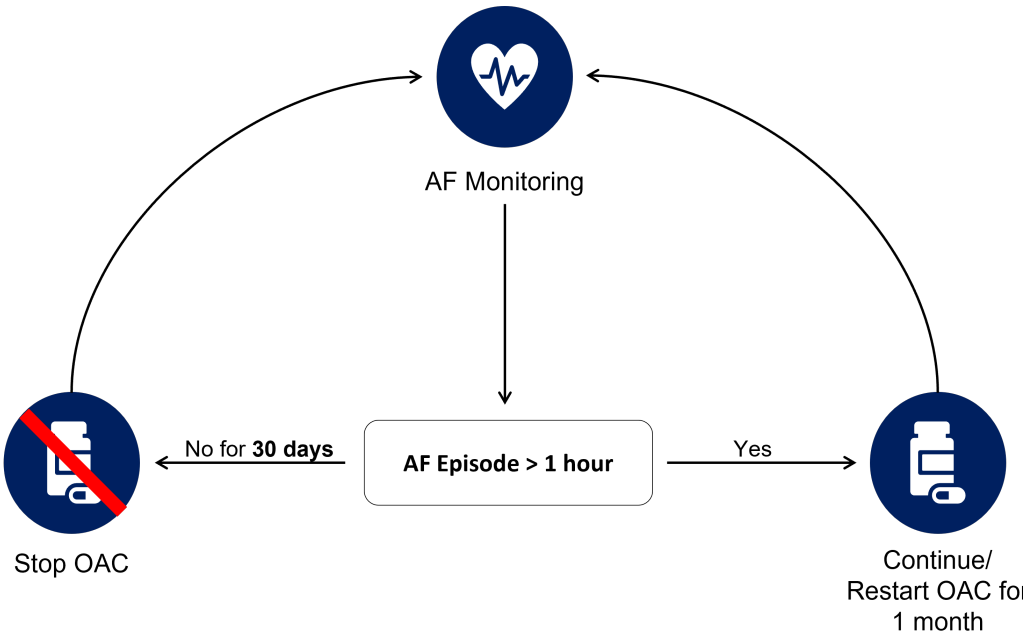


Figure 5.4: Overview of AI-PiP Study Treatment Group Design.

5.5.5 Study Population

The study population for the trial would be patients with AF and a moderate risk of stroke. This is indicated by the CHA_2DS_2-VASc Score, commonly used by cardiologists to determine the risk of stroke and treatment pathways. For this trial, we will include patients with a CHA_2DS_2-VASc score of 1-3, which correspond to a 1.3%-3.2% rate of stroke per year [80]. Patients with a score of 1-3 are also the population whom cardiologists would consider recommending the pill-in-the-pocket strategy. For a high risk of stroke, the pill-in-the-pocket strategy would not be safe or sufficient to prevent stroke. We are also excluding patients with prior stroke or TIA for the same reason of high risk with insufficient prevention from the strategy. The full inclusion and exclusion criteria is presented in Table 5.1.

5.5.6 Statistical Consideration

5.5.6.1 Analysis Sets

The Full Analysis Set (FAS) will include data from all enrolled patients regardless of their adherence.

The Comply Analysis Set (CAS) will include data from all participants who complied and finished the 24-month trial. Participants from this analysis set would be used to evaluate the primary and secondary objectives.

The Risk Analysis Set (RAS) will consist of participants who have a stroke or bleeding or any cardio-

Inclusion Criteria	Exclusion Criteria
• Possession of the following at the time of enrollment, ascertained from automatic software/ device pairing check: – iPhone (iOS version 16.1). – Apple Watch Series 4 and above. • Age > 22 years old. • Current resident of the United Kingdom and the EU. • Proficient in written and spoken English. • Valid phone number and email address. • Had at least one episode of paroxysmal or persistent AF documented by a 12-lead ECG. • Had a CHA_2DS_2-VASc score of 1-3 without prior stroke or transient ischaemic attack (TIA). • Currently on oral anticoagulation treatment for a minimum of 30 days.	• Patient during pregnancy. • Patient with permanent AF. • Patient with contraindications to anticoagulation. • Patient with prosthetic heart valves, prior stroke or TIA.

Table 5.1: Inclusion and Exclusion Criteria

vascular event during the trial. Data from this analysis set would contribute to the primary objective.

5.5.6.2 Statistical Approaches

The non-inferiority RCT trial will be evaluated by the following hypothesis testing:

Null Hypothesis: The Pill-in-the-Pocket treatment proposed is inferior to the standard continuous treatment in reducing patients' risk of stroke and bleeding events.

Alternative Hypothesis: The Pill-in-the-Pocket treatment proposed is not inferior to the standard continuous treatment in reducing patients' risk of stroke and bleeding events.

For the test, we aim for a minimum of 80% power and a 0.05 significance level for the one-tailed test. This allows a smaller sample size while ensuring low Type I and Type II errors. The test would compare the rate of stroke from the two groups with a non-inferiority margin of 0.5%. The margin was designed to get a good balance between a low sample size and a higher accuracy for the test. With the margin, in the worse case where the mean rate of stroke for pill-in-the-pocket is indeed 0.5% higher than the usual pathway, it still would not affect the CHA_2DS_2-VASc score of the patient. A CHA_2DS_2-VASc score of 4 refers to a rate of stroke of 4.0%, which is 0.7% higher than the rate at a score below.

5.5.6.3 Minimum Sample Size Calculation

The Minimum sample size required for the trial was calculated by both nQuery and statistical formula to select the most conservative number. The statistical formula is given in equation 5.4, with p_1 and p_2 referring to the expected control and treatment group stroke rate, respectively. In this case, we expect the same stroke rate so that $p_1 = p_2 = 0.010$. Δ is the non-inferiority difference with a value of 0.005 or 0.5%, as presented above.

$$n = \frac{(p_1(1-p_1) + p_2(1-p_2)) \cdot (Z_{1-\alpha} + Z_{1-\beta})^2}{\Delta^2} \tag{5.4}$$

The Z-values for 0.95 and 0.80 are 1.644 and 0.84, respectively, from the normal distribution table. With the values and the equation, we can find a minimum sample size of 4887 per group, 9774 in total.

The sample size calculation was also done with nQuery with a PTE0-1 Non-Inferiority Test for two proportions. The input and result are shown in table 5.2.

Test Significance Level, α	0.050
Standard Proportion, p_1	0.010
Non-Inferiority Difference, Δ	0.005
Test Expected Proportion, p_2	0.010
Expected Difference, Δ_1	0.000
Power (%), $(1-\beta)$	80
Sample Size per Group, n	4897

Table 5.2: nQuery Non-Inferiority Test for Two Proportions details and the sample size calculated.

We can see that the two calculations' results are extremely similar, with only a 0.2% difference. This can be due to the inaccurate Z-value estimate for the 0.95 percentile point, as the values were taken with three decimal places. To be on the safe side, we would use the more significant value from nQuery, 4897 for each group, as our minimum sample size. We would require an overall sample size of 9794 for the two groups.

For a clinical study with high risk, we would require and expect a complete adherence, therefore a 0% dropout rate was assumed with a 5% withdrawal rate from AF progression according to a study by Ogawa et al. (2018) [81]. In conclusion, the study would require a minimum sample size of 10,310 participants.

5.5.7 Limitations

The trial was designed to assess safety under a highly restricted condition with a high requirement for adherence. The explanatory nature of the trial may lead to a less representative outcome in more generalised settings, as when the intervention is integrated, the patients will not be under the same restricted settings as the trial. The requirement for adherence is also challenging to ensure. However, we will implement digital notifications and participants' training before the trial to help patients comply.

The large sample size required may also cause trouble in recruitment. Due to the strict eligibility criteria, we would need a large number of clinical sites from multiple countries. From the REACT.COM study, the recruitment of 59 participants was completed after two years in both the US and Canada [78]. Therefore, to aim for a two-year recruitment process in our study, we designed the trial to be conducted in both the UK and EU with a target of 60 participants per clinical site over the two-year period. This would require an estimate of 172 clinical sites to complete the recruitment process.

The recruitment process is also limited to two years to avoid potential participant overlap and additional resources. As the trial time frame is 24 months, setting the recruitment duration to two years can prevent any participant overlap. No participant would finish the trial before the recruitment process ended.

5.5.7.1 Bias

Selection Bias may still arise in the randomised controlled trial when there are systematic differences between the participants in treatment and control groups. To mitigate this effect, we will implement a computer-generated randomisation sequence to ensure there is no systematic difference in baseline characteristics between the two groups.

To avoid detection bias, a standardised outcome assessment was designed to determine incidences of stroke and bleeding. We defined a bleeding event as an episode of bleeding resulting in a > 3 g/dL (grams per decilitre) decrease in haemoglobin or fatal bleeding within seven days [82]. Ischemic stroke was defined as "an episode of neurological dysfunction lasting $\geq$ 24 hours or until death caused by focal cerebral, spine, or retinal infarction based on pathological, imaging, or other objective evidence", according to the REACT.COM study [78] and the TACTIC-AF study [83]. TIA, similarly, was defined as "a transient episode of neurological dysfunction lasting $<$ 24 hours caused by focal brain, spinal cord, or retinal ischemia without acute infarction".

5.6 Conclusion

In this chapter, we proposed two clinical validation studies to address the performance of the AFib-AI algorithm and the safety of the 'pill-in-the-pocket' strategy guided by Apple Watch with our detection algorithm. We presented the study protocol of the trials designed and the statistical approaches to evaluate the trial objectives based on the UK and EU regulations and the Good Clinical Practice guidelines. The chapter aims to create a general structure of the two clinical trials required in order to get sponsors for the trials.

Further work may include a more detailed design of the trial process, such as interim checks, record-keeping protocols, and the required training process. Risk assessments are also encouraged to be evaluated with clinicians to identify and mitigate potential risks to participants.

It is worth noting that the regulations regarding BioMeTs may undergo changes in the near future due to recent legislative changes for clinical trial consultations in 2022 [84]. The legislative change proposal included an improved notification scheme for low-intervention trials, which aims for a much faster approval process and less paperwork. The proposed low-intervention AFib-AI trial would benefit from the improved notification scheme as there is already extensive clinical experience with the Apple Watch AF detection technology with a similar safety profile.

Chapter 6

Impact Assessment {Owen Douglas}

6.1 Introduction

Having designed and validated the product, the next step was to determine its cost-effectiveness and feasibility as a clinical treatment pathway. In this impact assessment, a study was conducted into the economic and clinical benefits of the proposed PiP strategy compared to the current Continuous anticoagulation pathway. Then, the results' robustness to uncertainty and limitations to the model were assessed. Finally, the study evaluated the impact of the product from the perspectives of key stakeholders identified: patients, clinicians, healthcare systems and society. The associated benefits and drawbacks were identified and analysed for each stakeholder. This would help to minimise any potential negative impacts that could arise in the development and implementation of the product in practice.

6.2 Cost Effectiveness Analysis

6.2.1 Objective

Through a cost-effectiveness analysis (CEA) study, the current AF treatment pathway was compared with the proposed PiP alternative in order to provide evidence of economic feasibility. The Quality-Adjusted Life Year (QALY) was utilised as the preferred metric to assess the effectiveness of each treatment since it is a widely accepted and used measure in health economics. The underlying assumption of the QALY metric is that one year of life lived in perfect health is valued at 1 QALY,

whereas a year lived in less than perfect health is valued at less than 1 QALY [85], i.e.

$$1\,Year\,of\,Life \times 1\,Utility = 1\,QALY \tag{6.1}$$

The associated utility values can be estimated and used within the study. Using QALYs and costs, a composite £ per QALY metric was formed to compare the cost per unit health outcome gained by the proposed treatment strategies.

6.2.2 Methodology

6.2.2.1 Model Framework

Following a similar process outlined in [86], a cohort state-transition model (CSTM) was adopted to simulate the progression of a hypothetical cohort transitioning through various health states over a given time horizon. The CSTM is a stochastic model that represents Markovian processes, hence commonly known as a Markov model. The fundamental feature of the model is the imposition of the Markovian property, whereby the prediction of the patient's prognosis is solely dependent on their current state rather than previous states. Each state within the model may feature varying levels of quality of life and associated costs and therefore it is possible to attribute distinct utilities and costs to each state. This modelling approach enables the estimation of metrics such as the expected utility and resource costs for each state and the overall quality-adjusted life expectancy [87]. By running the simulation for various strategies, the computed values could be used to compare their cost-effectiveness and viability as clinical treatment pathways. The study was conducted from the perspective of the NHS.

The model constructed for the analysis was a three-state stationary Markov chain, as shown in Figure 6.1. The health states were the same for each treatment strategy and were defined as follows:

- *Paroxysmal AF* - Patients entered the model in this state having been diagnosed with paroxysmal AF, defined as recurrent "episodes of AF that terminate within seven days" [88]. The study assumed that these patients had a CHA_2DS_2-VASc score of 2 at the time of diagnosis, and their treatment was determined by the strategy adopted.

- *Post-stroke AF* - Upon transitioning to this state it was assumed that patients had suffered a stroke, resulting in a raised CHA_2DS_2-VASc score (≥ 4) [89]. As the score had significantly increased, the decision to anticoagulate became independent of the temporal pattern of AF [9] and therefore classification of the type of AF was not required in this state.

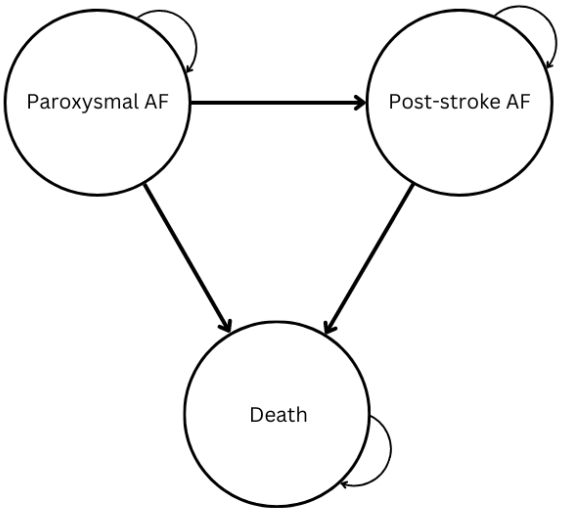


Figure 6.1: The three-state Markov model, with each arrow representing possible state transitions. Death was the absorbent state since no individual could move back to either of the previous states, as indicated by the unidirectional arrows.

- *Death* - This was the absorbent state in the model as once patients transitioned to this state they could not leave. Transitions to this state were possible from either of the other two states (with corresponding probabilities), as indicated by the directional arrows in Figure 6.1.

The transition probabilities represented the probability of individuals in the cohort moving between or remaining in the same health state in a given cycle [86]. In the study, these probabilities (along with state costs and utilities) remained constant over time, since a time-independent model had been assumed. The model could then be represented as a Markov chain $\boldsymbol{X_t}$ over a finite state space $S = \{Paroxysmal\ AF,\ Post\text{-}stroke\ AF,\ Death\}$, with a probability matrix

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \tag{6.2}$$

where each element

$$p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i) \quad \text{for } i, j \in S \tag{6.3}$$

The state at time t was the value of X_t [90] and the indices of the entries corresponded to the indices of states in S. Costs of treatments at each state were strategy-dependent, as were the utilities.

Utilities are numbers between zero and one that represent the quality of life an individual experiences in a specific health state and treatment strategy. They are used in the calculation of QALYs, as in Equation 6.1. The expected cost per year was used to evaluate the cost of bleeding events, again with the probabilities dependent on the strategy employed. The values of these transition probabilities, costs and utilities are further discussed in Section 6.2.2.3.

Finally, the simulations used a cycle length of one year, since most parameter values available in the literature were annualised, and a time horizon of ages 65 to 100, as stroke risk and anticoagulation treatment are reviewed when AF patients reach 65 [91]. Both costs and QALYs were discounted at a rate of 3.5%, as recommended by the National Institute for Health and Care Excellence (NICE) [92].

6.2.2.2 Treatment Strategies

The study considered three different treatment strategies: Base, Continuous anticoagulation and PiP anticoagulation.

The Base strategy provided a baseline to which both the Continuous and PiP anticoagulation strategies could be compared. It represented the case where no anticoagulation treatment was administered to patients. By comparing the outcomes and costs associated with the Base strategy to those of the Continuous and PiP anticoagulation strategies, the relative clinical and economic impact of these interventions could be determined.

Continuous anticoagulation reflected the current treatment recommended under NICE guideline NG196 [91]. A patient's risk of stroke is evaluated using the CHA_2DS_2-VASc score and if deemed significant (≥ 2) they are offered direct-acting oral anticoagulants (DOAC) to reduce this risk. Patients are also considered for anticoagulation with a CHA_2DS_2-VASc score of 1. Although highly effective at the prevention and treatment of thromboembolism, anticoagulants are linked with notable bleeding risks [93]. A patient's bleeding risk is calculated using the ORBIT score and is taken into account when prescribing anticoagulants. Apixaban, dabigatran, and rivaroxaban are among the recommended DOACs [91]. Although these drugs demonstrate comparable effectiveness, apixaban is potentially associated with a lower risk of major bleeding [94], and so this was the drug selected for the model. In both the paroxysmal AF and post-stroke AF health states, the model assumed that apixaban was taken twice daily, as recommended by NICE [95].

Finally, the PiP anticoagulation strategy represented the proposed treatment pathway. In this strategy, patients in the paroxysmal AF health state took DOACs only when an AF episode was detected

by the wearable device. The aim of this approach was to maintain the benefits of anticoagulant prophylaxis, whilst reducing exposure to the associated bleeding risks. During this study, it was assumed that patients in the post-AF state would continuously take DOACs, similar to the Continuous anticoagulation strategy. This was because their CHA_2DS_2-VASc scores had increased due to the occurrence of stroke, and the existing research on the safety of the PiP approach has primarily focused on patients with low CHA_2DS_2-VASc scores, as observed in previous studies [96].

6.2.2.3 Parameters

The values for the transition probabilities, costs and utilities used in the analysis were all taken or inferred from various studies and papers. In some papers, risks of events, such as bleeding or stroke, were expressed as annual rates and these needed to be converted into probabilities for use in the model. The study followed the method described in [86], where for an annual rate μ, the annual probability could be expressed as

$$p = 1 - e^{-\mu} \tag{6.4}$$

Assuming the rate of an event was constant over a year, Equation 6.4 suggested that the time until the next event occurred was exponentially distributed [86]. For simplicity, it was assumed that the transition probabilities were the same for all strategies. The parameter values, along with their sources, are displayed in Table 6.1. Some costs and utilities were common to all strategies, whereas others were strategy-specific. For all strategies, it was assumed that once transitioning to the Death state, patients remained there and the costs and utilities of being in that state were zero.

Some comments about the parameters:

- *Rates* - All values for event rates were sampled as percentages and then divided by 100 to convert into a probability.
- *Post-stroke to Paroxysmal AF* - Patients could not return to the Paroxysmal AF state from the Post-stroke AF state.
- *Bleeding event* - The cost was calculated as the average of the "Other major bleeds" in [100], composed of gastrointestinal bleeds, non-gastrointestinal related and clinically relevant non-major bleeds.
- *Annual apixaban DOAC (Continuous)* - This was the annualised drug cost of £2.20 per day.
- *Reduction in PiP DOAC time* - The reduction in time a patient was required to take DOACs in

Parameter	Value	Distribution	Source
Number of cycles	35	-	Model
Number of health states	3	-	Model
Annual cost discount rate	3.5%	-	NICE 2020 [92]
Annual QALY discount rate	3.5%	-	NICE 2020 [92]
Annual constant transition rates			
Mortality with Paroxysmal AF	0.0352	Gamma(1.377, 2.557)	Shafeeq & Tran 2014 [97]
Bleeding event Continuous	0.032	Gamma(1.138, 2.813)	Estimate
Bleeding event PiP	0.01	Gamma(0.111, 9)	Estimate
Annual constant transition probabilities			
Paroxysmal AF to Post-stroke	0.0053	Beta(2.639, 423.062)	Kaplan et al. 2019 [98]
Post-stroke to Paroxysmal AF	0	-	Assumed
Post-stroke to Death	0.25	Beta(248, 745)	Saborido et al. 2010 [99]
Costs			
Bleeding event	£2191.84	Gamma(4.804, 456.238)	Dorian et al. 2014 [100]
Annual apixaban DOAC (Continuous)	£803.00	Gamma(2321.294, 0.346)	NICE TA275 [95]
Reduction in PiP DOAC time	0.843	Beta(10.314, 1.921)	[101, 102]
Wearable device	£419.00	Gamma(1644.137, 0.255)	Apple Inc. 2023 [103]
Annual cost of stroke	£3061.20	Normal(3061.2, 666.67)	Youman et al. 2003 [104]
Death	£0.00	-	Assumed
Utilities			
Annual probability of recurring AF event	0.303	Beta(1.346, 3.096)	Healey et al. 2017 [105]
Normal Sinus Rhythm (PiP and Continuous)	0.89	Beta(26.482, 4.838)	Saborido et al. 2010 [99]
AF event	0.71	Beta(43.195, 14.398)	Saborido et al. 2010 [99]
Loss of utility due to bleeding event	0.1	Beta(0.8, 7.2)	Estimated
Post-stroke	0.74	Beta(13.498, 4.742)	Saborido et al. 2010 [99]
Death	0	-	Assumed
Base Paroxysmal AF	0.4	Beta(21.222, 31.832)	Estimated
Base Post-stroke	0.3	Beta(56.457, 131.732)	Estimated

Table 6.1: The values, distributions and sources of the parameters used in the model. Values were either assumed from the model, estimated or derived from various guidance, studies and papers.

Strategy	Cost (£)	QALYs	Incr. Costs (£)	Incr. QALYs	ICER (£ / QALY)	Status
Base	687.68	4.89				ND
PiP	2795.51	9.36	2107.83	4.47	471.16	ND
Continuous	11382.73	7.47				D

Table 6.2: The base-case cost-effectiveness results, showing the incremental (Incr.) costs and QALYS, and the incremental cost-effectiveness ratios (ICER). The results show that the PiP strategy was the most cost-effective compared to the others.

the PiP strategy compared to Continuous. The reduction was averaged from the sources listed.

- *Wearable device* - The cost of the standard Apple Watch Series 8. This was the latest model at the time of the study and was compatible with the specifications in 3.3.
- *NSR (PiP and Continuous)* - It was assumed that patients diagnosed with Paroxysmal AF would experience a normal sinus rhythm outside of recurrent AF events.

6.2.3 Results

6.2.3.1 Base-case Analysis

The results of the cohort simulation are presented in Table 6.2. The PiP strategy did not result in the lowest cost, however it was more clinically effective (QALYs) compared to the other strategies. The Base strategy cost was the lowest due to the absence of anticoagulation treatment provided in this pathway. The £ per QALY metric outlined in Section 6.2.1 was calculated as an incremental cost-effectiveness ratio (ICER). This was computed by dividing the incremental costs by the incremental effect (QALY), with "incremental" referring to the relative difference to the base strategy [106]. The Base strategy had no associated ICER value since it was the baseline other strategies were compared against. The Continuous strategy was an economically dominated strategy, as indicated by its "D" Status in Table 6.2. This meant that the other strategies were both clinically more beneficial and more cost-efficient [107] (Non-Dominated), and so it was not necessary to calculate the ICER for this strategy. Overall, the PiP strategy appeared to be the most cost-effective according to the base-case analysis.

6.2.3.2 Probabilistic Sensitivity Analysis

The values of the model parameters, including probabilities, rates, costs and utilities, were all derived from different studies and trials which naturally introduced uncertainty into the calculations. This uncertainty may have been caused by a variety of factors, such as estimation error, differences in results reported across studies or random error [108]. To account for this a probabilistic sensitivity

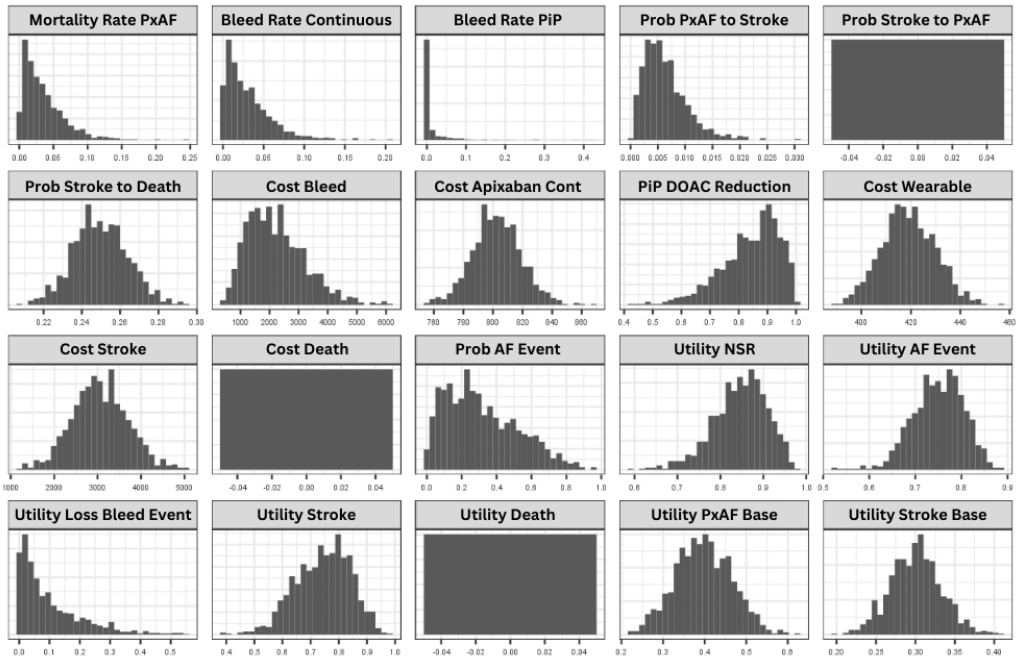


Figure 6.2: The distributions of the parameters used in the PSA, where *PxAF*, *Prob* and *Cont* and are short for Paroxysmal AF, Probability and Continuous respectively. They consisted of Gamma for transition rates and costs, Beta for probabilities and utilities and Normal for the cost of stroke. Parameters with greyed-out distributions imply that no distribution was used, due to them being fully known or assumed.

analysis (PSA) was conducted, whereby each parameter was modelled as a random variable and sampled from its assumed distribution. With a new set of parameter values, we re-ran the cohort simulation and repeated this process 1000 times. The parameter distributions are described in Table 6.1 and visualised in Figure 6.2. Standard distributions were used to model parameter uncertainties, with appropriately wide tails to account for the large variance in potential values. Specifically, the Gamma distribution was used for transition rates and costs, and the Beta distribution for probabilities and utilities, since it is defined on the interval $[0, 1]$ [86]. A Normal distribution was used for the annual cost of stroke since it better captured its distribution than the others. The mean of each distribution was the value of the parameter found or inferred from the literature. No distribution was used for the *Prob Stroke to PxAF* transition probability since it would be impossible for a patient to return to a pre-stroke state after having a stroke, i.e. reversing a stroke. Therefore the probability of this transition was zero with zero variance. Similarly, no distributions were used for the cost and utility of being in the Death state, since it was assumed these were both zero.

To conduct the PSA, 1000 different parameter sets were generated and used in each separate cohort simulation. The results of the PSA are shown in Figure 6.3, a cost-effectiveness scatter plot, where each point represents the result of an individual simulation. The 95% confidence regions for each strategy are shown by the dotted ellipses. From the plot, it can be seen that the Base strategy

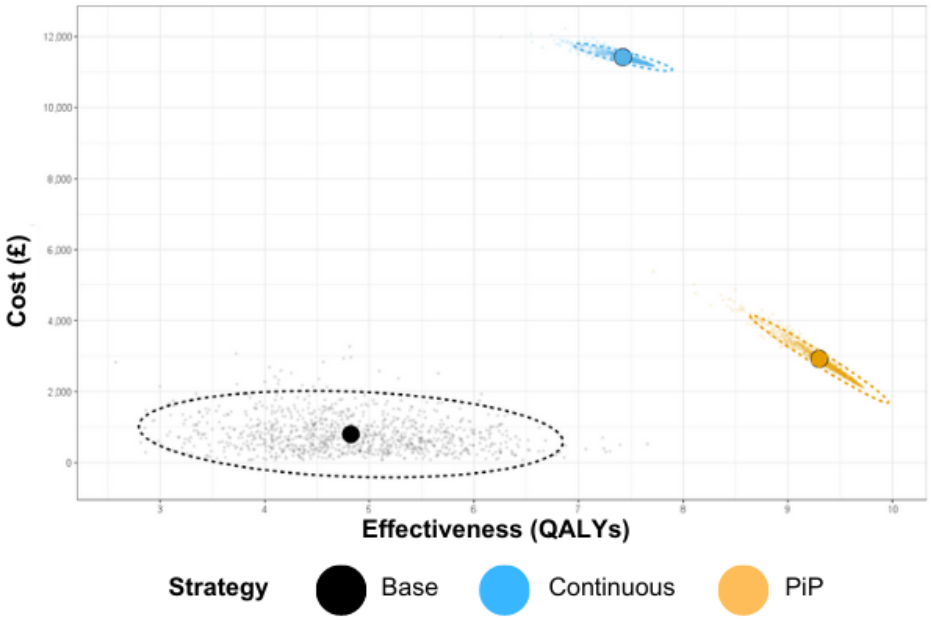


Figure 6.3: The results of the PSA in determining the effects of uncertainty on the cost-effectiveness calculations. The 95% confidence regions are shown as dotted ellipses. The PiP strategy remained the most cost-effective compared to the Base strategy after the PSA simulations.

contained most of the uncertainty, as shown by its larger confidence region. This easily allowed the comparison of the Continuous and PiP anticoagulation strategies relative to the Base. The PiP strategy (shown in yellow) remained the most cost-effective, due to its higher QALY amount and lower cost compared to the Continuous (blue) and Base (black) strategies.

6.2.4 Evaluation

6.2.4.1 Summary

To summarise, a CEA was performed to compare the clinical and economic outcomes and feasibility of three proposed treatment strategies described in Section 6.2.2.2. The base-case analysis determined that PiP was the dominant strategy out of PiP and Continuous anticoagulation, relative to the Base. To account for parameter uncertainty, a PSA was conducted to assess the robustness of the calculations. Overall, the PiP treatment strategy was determined to be the most cost-effective and dominant out of the proposed strategies. Even when accounting for parameter uncertainty, the PiP strategy remained the best choice. Although only a simple model, the result provides evidence for the economic feasibility of the product and the PiP treatment pathway.

6.2.4.2 Model Limitations & Improvements

Although the model effectively demonstrated the benefits of using the PiP treatment strategy, there were some model limitations to be aware of and room for further improvement. These are summarised below:

- *Limited Research and Data* - Since the PiP method of anticoagulation treatment is still in the research stage, the data on event rates, hazard ratios, etc is limited and specific to certain trials and studies. Parameter values were inferred from this data and so any fundamental error or lack of generalisability would have reflected in the results, even when accounting for uncertainty with the PSA. The ongoing REACT-AF trial aims to further study the PiP strategy for preventing stroke in AF patients with wearable devices [109]. As it is due to last seven years, the trial can help to provide more data specific to this treatment pathway, which can be used in further cost-effectiveness studies to provide a more accurate analysis.

- *Health States* - The model only contained three health states for simplicity, but more could be included in the future for additional complexity. With access to more data, other processes involved with the treatment could be modelled, for example, the progression of paroxysmal AF to persistent or permanent AF. However, this returns to the previous issue of having data attributed to specific studies and the resulting discrepancies in reported values. In the progression-of-AF example, the European Heart Survey recorded an annual progression rate of 15% in patients with AF, whereas another study with Japanese patients found a progression rate of just 6% per year [110]. It is clear that when modelling more health states, additional data is required to form better estimates of parameters.

- *Time Dependence* - The model assumed time independence, that is the values of parameters did not change over the course of the simulation. For a more complex model, temporal variations in certain parameter values could be factored in. For example, a patient's risk of stroke increases with age according to the CHA_2DS_2-VASc score [89]. These changes could potentially affect the final outcome of the simulation.

- *Parameter Correlations* - To make results more robust, it may be beneficial to explore any potential correlations between parameters. This is especially important when performing the PSA since parameter values were independently sampled from their distributions.

- *Other Costs* - Finally, other costs that may arise from different treatment pathways were likely

not accounted for in the model. There may have been obscure indirect costs to the healthcare system not incorporated into the model, which may have affected the results. Further analysis of costs and more research could be undertaken to build a better picture of the total costs for each strategy. Another point to consider regarding costs is that any past cost data should be inflated to current values for a more accurate cost-effectiveness estimate.

In practice, the results from the trials, designed in Chapter 5, would be fed into the cost-effectiveness analysis. Since the trials would provide relevant data specific to the product and proposed PiP pathway, more accurate simulation outcomes could be obtained. The model drew some inspiration from a PiP CEA study by Saborido et al., 2010 [99], however this focused on antiarrhythmic drugs instead of anticoagulants. Based on the preliminary research, there appeared to be a limited number of CEA studies done on the PiP strategy for anticoagulants specifically. However with the completion of the REACT-AF trial, there could be sufficient data available to conduct a further in-depth cost-effectiveness study.

6.3 Patients

6.3.1 Health Effects & Quality of Life

6.3.1.1 Improved AF Detection

As identified in Section 2.2, there was an inherent need for widespread AF detection for accurate diagnosis and early treatment of this common arrhythmia. Coupled with a wearable ECG device, the designed algorithm was capable of detecting AF early and with high accuracy, therefore providing the potential to improve patient outcomes and quality of life. An early diagnosis and initiation of treatment may prevent or limit the progression of AF [111], reducing the risk of further associated complications. Approximately one-third of AF patients are initially asymptomatic, yet their risk of stroke and other complications appears to be comparable to that of patients with symptomatic AF [11, 112]. Therefore the product could aid the detection of more asymptomatic cases, giving patients earlier access to treatment and care and potentially reducing future complications.

Patients can benefit from improved management of AF by detecting and diagnosing it early. With advice from clinicians, patients can make more informed decisions about any dietary or lifestyle adjustments to limit the risk and severity of AF symptoms, leading to an overall better quality of life. The additional awareness of the condition could help to reduce anxiety about any potentially undiagnosed AF cases in individuals and provide peace of mind, since the condition has been detected and treated

early.

6.3.1.2 Reduced Stroke Risk

One of the main objectives of AF treatment is to limit the risk of ischaemic stroke. This is because not only does AF increase your risk of stroke five-fold [113], but AF-related stroke increases the risk of further medical complications and death compared to stroke in patients without AF [114]. The designed product aimed to use the PiP treatment pathway to administer anticoagulation prophylaxis using DOACs such as apixaban, as described in Section 6.2.2.2. It is widely known that strokes are a detriment to a patient's quality of life and so by providing earlier AF detection, the product can help to reduce this risk.

6.3.1.3 Reduced Bleeding Risk

A considerable side effect of anticoagulation treatment is the increased risk of major bleeding since they are used to prevent ischaemic stroke by decreasing the body's ability to form blood clots. Currently, a patient being considered for anticoagulants has their risk of stroke weighed against their risk of bleeding, calculated using the CHA_2DS_2-VASc and ORBIT scores respectively [91]. If a patient is deemed suitable for anticoagulation treatment, they are prescribed DOACs to take daily. The proposed treatment pathway aimed to combine the AI detection algorithm with a wearable device to utilise the PiP method. Patients would only need to take DOACs when an AF event had been detected, therefore reducing the time required for a patient to be taking the drugs (up to 94% in one study [102]), especially for paroxysmal AF. This was intended to reduce the risk of major bleeds in anticoagulated patients and improve their overall quality of life, although this will be studied further in the REACT-AF trial.

6.3.2 Usability & Accessibility

6.3.2.1 Comparison with Existing Detection Devices

The product was designed to use a wearable device for real-time AF detection, however there are other devices currently used to monitor patient ECGs.

- *Holter Monitor* - These devices are worn for periods of up to 48 hours [115] and record a patient's ECG signals during this time. Since recordings are confined to this time window, Holter monitors are mostly useful for patients who experience frequent AF symptoms and so were unlikely to be the best choice for PiP treatment. The main advantage of the Holter monitor

is that it utilises multiple leads to take ECG measurements, which can improve the received signal quality. However, the downside is that the longer distances between electrodes can make readings more susceptible to signal artefact from body movements. This, coupled with its bulky size and many wires, made the Holter monitor not suitable as an everyday wearable device for PiP treatment.

- *Patch Recorders* - Devices such as the iRhythm Zio patch [116] have demonstrated even better performance than the Holter monitor [117]. Patch recorders are also lighter and have a much lower profile, so are better suited to everyday wear with limited disruption. Whilst patch recorders were more suitable for PiP treatment than Holter monitors, they have a lifespan of only 14 days, meaning that patches would need to be replaced regularly. This limited their use as a PiP AF monitor.

- *Implantable Loop Recorders* - Whilst loop recorders have the ability to monitor your heart for up to three years [118], these devices are much more invasive than the previous two, as they require the device to be implanted under a patient's skin. For long-term PiP AF detection, the repeated implant and removal of the devices may be uncomfortable for patients.

For PiP AF treatment, the proposed solution was to use a wearable device, such as an Apple Watch, to record and detect AF episodes. Using these devices does not require invasive surgery, is comfortable to wear and its lifespan is likely around three to four years (based on the Apple Watch [119]). These benefits made wearables more appropriate for the PiP anticoagulation pathway than the other devices discussed, however they are not perfect and had some potential drawbacks, as discussed in the next section.

6.3.2.2 Potential Drawbacks of Wearable Devices

Wearable devices, such as the Apple Watch, are not without their drawbacks and it was important to identify potential limitations to their use in practice. Since the algorithm was designed to detect AF in real time, it is vital that patients wear their devices as much as possible to fully benefit from PiP treatment. Therefore, it would be vital to ensure that recommended devices have a satisfactory level of comfort and convenience so that it does not become a burden to patients' everyday lives. It is expected that device manufacturers, such as Apple, account for this in the design of their products, however further research is needed to ensure patient satisfaction.

In a patient questionnaire study about their perspectives on PiP anticoagulation, a slightly lower

number of patients (48%) indicated they would use wearable devices for PiP anticoagulation than implantable monitors (53.7%) [120]. Apart from clinical reasons, the main concern regarded the "reliability of [wearable] monitoring technologies". However, the algorithm has demonstrated promising results and so in practice more information must be provided to patients to gain their confidence in using wearable devices for PiP treatment.

Another potential limitation of wearable devices is the dependence on ensuring the battery remains charged and patients remembering to wear the device. This is a clear drawback since there is a chance a patient could experience an AF episode when the device is on charge or when the patient has forgotten to wear it. In the PiP strategy, this could mean that the patient misses their signal to take a DOAC dose. More research is needed to explore the health implications of this and any potential solutions.

6.3.2.3 Older Population

As discussed at the end of Section 6.2.2.1, anticoagulation treatment for AF is mostly prevalent in the older population. This is because their risk of stroke is generally higher, as indicated by the CHA_2DS_2-VASc score [89]. Therefore, the product must be easy and simple to use. One way of doing this would be to design and distribute a questionnaire to current AF patients about if and how they currently use wearable devices, any concerns they have regarding the PiP process and their willingness to adopt a detection algorithm to monitor their condition. Also, it was important to design an intuitive and straightforward user interface (UI), as displayed in Chapter 3 to ensure the product is as effortless to use as possible. A small trial could then be designed to assess the usability of the UI in older patients.

6.4 Clinicians

6.4.1 Clinical Workflow

6.4.1.1 Primary Care

Since the algorithm could accurately detect asymptomatic episodes of AF, a greater number of otherwise undetected AF cases would be expected. Whilst clearly there is a benefit to patients in detecting this condition, as discussed in the previous section, the resulting workload on primary care (GPs) would somewhat increase. With the numbers of GPs already struggling to keep up with demand [121], an increased workload due to more AF cases could add even more pressure to primary care.

However during the needs research interview, it was explained that this increased workload would be acceptable since more patients would be receiving the required treatment.

Furthermore, it was expected that the PiP treatment pathway would require fewer in-person appointments with GPs since AF events could be reliably detected and monitored using the product outside of the clinic. Therefore, after an initial GP appointment to confirm an AF diagnosis and start the patient on PiP anticoagulation, it was assumed that fewer GP visits would be required. Consequently, this would supposedly reduce the workload in primary care, freeing GPs to tend to other patients with other conditions. Of course, if the product were to be introduced into clinical practice this process would need to be fully determined and standardised.

6.4.1.2 Secondary Care

As more cases are presented to GPs, more referrals would be made to secondary care specialists, such as cardiologists and electrophysiologists. Currently, humans are still required to make a formal diagnosis of AF and determine a patient's suitability for anticoagulation treatment. Therefore with more cases detected, it is possible that secondary care could also experience a higher workload processing these referrals. Already there is a backlog in NHS secondary care [122], and adding to this could have poor consequences. In the interview, it was mentioned that although adding more work to an already stressed system could cause problems further down the pathway, the way the current system works is that this would then be the next problem to address. This implies that if the use of our product caused non-major problems elsewhere then it would generally still be acceptable, given that more patients would be getting access to the required treatment.

6.4.2 Adoption of AI

6.4.2.1 Current Stance on AI

With a global annual market growth rate of 37% [123], there is clearly an increasing prevalence of AI in healthcare applications. However, the successful implementation of AI methods into daily clinical practice remains a significant challenge [124]. In the NHS, medical advice is always provided by clinicians and the final care and treatment decisions lie with the patient, not an AI system. Clinicians are allowed to use AI technology to assist in diagnosis or to help reach decisions about which treatments to recommend [125], but a fully-automated pathway is currently not allowed. Whether this becomes practice depends on further research and the actions of the regulators.

From the interview, it was determined that in general clinicians are willing to adopt AI technology to

assist them with their daily responsibilities. The product could help to streamline the detection and diagnosis of AF and through the PiP treatment pathway, reduce the number of in-person appointments required, further decreasing the time spent by clinicians managing these patients. A crucial point raised however, was that if a proposed AI method reduces treatment efficiency, slows down a clinician in their day-to-day tasks or requires extra training, there would be reluctance from clinicians to adopt it, given that their time is already significantly limited.

6.4.2.2 Changes to Treatment Pathway & Training

It was critical that the product did not impede or delay clinicians in their routine responsibilities. If it did so in any way, there would be a risk of push-back to adoption by clinicians. In the proposed pathway, no major changes to current care were anticipated. The PiP method was expected to reduce the number of in-person appointments required since the management of the anticoagulation treatment would mostly be performed by patients themselves with the aid of a wearable device and the detection algorithm. ECG measurements and other data could be sent to clinicians remotely and they could perform a similar analysis of the patient's current condition. Apart from this, it was assumed that the rest of the treatment would remain mostly unchanged. Further studies, such as clinician questionnaires, should be done to examine this before implementing the product in clinical practice.

Additionally, it was expected that no extra major training would be required for clinicians to use the product. They may need some time to gain familiarity with the system, which would also be an opportune time to gather any feedback or suggested improvements, but likely would not require a significant amount of training. As part of the development, it would be wise to distribute a "beta" version of the system to gather feedback from clinicians before releasing the final product. Overall, a positive impact on clinicians was predicted and from the interview, it seemed that they would be willing to adopt such a system, as long as the design ensures it does not hinder them in their routine responsibilities.

6.5 Healthcare Systems

6.5.1 Economic Impact

6.5.1.1 Costs

As identified in the cost-effectiveness analysis in Section 6.2, the proposed PiP treatment pathway is likely to be more cost-effective than the current Continuous anticoagulation strategy. The fundamental purpose of anticoagulation treatment is to avoid strokes in patients with AF, as represented by the 'A' in the Atrial Fibrillation Better Care (ABC) pathway endorsed by the European Society of Cardiology [9]. Hospitalisations due to AF-related stroke and other heart conditions constitute the majority of direct AF costs to the NHS (direct costs of AF are predicted to be between 0.9% and 1.6% of total NHS expenditure) [126] and so by detecting more cases of AF earlier, the product could help to reduce this economic burden on the NHS. The additional costs of the increased workload discussed in Section 6.4.1 would be massively outweighed by the savings in preventing more potential strokes.

The costs of bleeding events are also significant and are comparable to the cost of stroke [126]. The proposed PiP strategy has been shown to reduce the DOAC treatment duration by as much as 94% [102] and the CEA model assumed this translated into a reduction in bleeding risk, resulting in fewer hospitalisations. With DOACs not needing to be taken as regularly as during the Continuous strategy, there would be a reduction in medication costs. Therefore the PiP method coupled with the designed AF detection algorithm could further reduce AF costs to the NHS.

6.5.1.2 Funding

For patients to have access to the designed detection algorithm a suitable wearable device is needed, many of which can be expensive. It is unclear who would cover this cost in practice, either the healthcare system (NHS) or the patient themselves. Whether the NHS funds these devices would depend on the amount of funding available and whether it is economically viable to do so. The ability of a patient's local hospital to pay for wearable devices would depend on its allocated budget, which has the potential to cause equitable access problems as further discussed in Section 6.6.2.1. If patients are required to pay for devices it could severely limit the reach of improved AF detection and the benefits of PiP treatment. Similarly, this could cause issues where a patient's access to better healthcare is dependent on their economic status. Further research is needed into determining how willing patients from different regions of the UK are to purchase wearable devices to access the PiP AF treatment.

The trials designed in Chapter 5 would also require funding to implement them. This would be an important factor to consider in practice and more analysis is needed to determine whether trial sponsors would be willing to provide funding. Other sources of trial funding could include charities, universities and other research institutions.

6.5.2 Delivery of Care

One of the main assumptions of the PiP strategy was the remote patient-centred monitoring of AF. There would be a shift in how care would be delivered to patients, from in-person appointments to home-based management. Regular primary care appointments can be advantageous since patients can easily share updates on their treatment and any symptoms experienced. However they could consume valuable GP resources, which could be better spent on other patients/conditions, especially for the management of paroxysmal AF. Remote monitoring using the designed detection algorithm and wearable device could reduce the number of appointments required, since relatively infrequent episodes of AF could be managed using the PiP method. Conversely, there is the possibility for patients to use the system incorrectly and improperly follow the PiP treatment, which could result in poor health consequences. Therefore to facilitate this change in the delivery of care, adequate information on the PiP treatment pathway must be made available for patients considering it. Further studies are also needed to identify how healthcare systems would support the integration of remote patient PiP treatment with the other aspects of the AF treatment pathway.

6.6 Society

6.6.1 Broader Impact on Society

6.6.1.1 Adoption of AI in Society

Due to an increased use of AI in assisting diagnosis and recommending treatment, it was crucial to analyse society's perception of these AI technologies. How society as a whole views AI in healthcare could have the power to determine whether the product is feasible in practice. The AF detection algorithm used deep learning methods to identify and notify the user of AF episodes, and as such it is an example of an AI tool used to assist in diagnosis.

One study of patient perspectives on AI in healthcare found that patients generally expressed a positive outlook on the ability of AI to improve healthcare [127]. However, they did show scepticism about some aspects pertaining to AI technology, most notably the accuracy and transparency of AI

methods. 91.5% of patients in [127] expressed concern about the misdiagnosis of AI systems and its potential implications. The designed AF detection algorithm however demonstrated a high accuracy and so this would need to be communicated effectively to patients. The technology would only be used as a tool to assist clinicians with diagnosis, not automate it, and so this human factor should reassure most patients. An additional area of concern was whether patients would be informed if AI was used in their diagnosis, with a significant total of 95.8% of patients deeming transparency very or somewhat important. The wearable AF detection and PiP treatment were specifically designed to be patient-centred and therefore patients would be completely aware of when and how AI would be used in their diagnosis and treatment management. Lastly, patients were more likely to be uncomfortable with receiving a diagnosis from an AI algorithm that lacked the ability to explain its decision-making process. Explainable AI is essential in a healthcare setting to build trust with both patients and clinicians. This issue however extends to the broader field of AI applications and future developments could prevent the deep learning method in the design from becoming a "black box", increasing the transparency and accountability of diagnosis.

6.6.1.2 Patient Self-Management

The PiP strategy utilised more patient autonomy in managing their own health. This shift towards patient self-management in AF treatment provides the possibility for a positive impact on patients' health and their behaviour towards living a healthy lifestyle [128]. Since patients would be more aware of AF, its symptoms and its health implications, they would have a greater incentive to self-manage their condition by making better lifestyle choices to mitigate the risks, especially of stroke. However with the PiP treatment pathway, it would be critical to ensure that patients get access to the same level and quality of care as other non-remote pathways.

6.6.1.3 Global Impact

Looking beyond the UK, implementing the product in other countries would require more analysis of each country individually. Many factors need to be taken into account, such as any cultural differences affecting healthcare, varying data privacy laws and relative levels of healthcare funding. One of the most important aspects to consider would be a country's laws surrounding trial processes and gaining approval for medical technology. This would likely be the main limiting factor in introducing the PiP technology to other countries and healthcare systems. Whilst the algorithm and PiP treatment pathway would not necessarily be limited to the UK, further in-depth and country-specific analysis would be needed to launch the technology globally.

6.6.2 Ethics

6.6.2.1 Equitable Access

It is critical and morally right to ensure everyone has access to an equal level of care regardless of socioeconomic status. To ensure equitable access to the early AF detection algorithm and PiP treatment, the potential effects of the economic disparity across the UK needed to be explored. Research by QualityWatch found that people residing in more deprived areas of England received a "worse quality of NHS care and poorer health outcomes" than those living in more affluent areas [129]. Consequently, there exists the potential for patients living in poorer areas to not have access to early AF detection and PiP treatment. This could cause large divides in local population health between poorer and wealthier areas, especially in cases of AF. The CEA study in Section 6.2 showed that the PiP treatment strategy is likely to be more cost-efficient than the current Continuous anticoagulation pathway and so it could be economically viable for hospitals in more deprived regions to implement it. This does not however account for any costs in swapping to the new pathway. A limiting factor could be the cost of the wearable devices themselves, especially if they were to be patient-funded. Compared to patients in wealthier areas, patients living in low-income areas may experience difficulty in affording expensive wearable devices. Extra funding opportunities should be explored to ensure all patients have equitable access to the AF detection and management technology, regardless of their socioeconomic status.

6.6.2.2 Algorithmic Bias

The data used to train the detection algorithm was the PhysioNet Challenge 2017 dataset [31]. The demographic attributes, such as age, sex, race, etc, of the ECG recordings were not clear and so there was a possibility of accidentally introducing bias into the algorithm, based on these factors. It is crucial to be aware of any biases in the detection algorithm so that measures can be made to correct them - the development of explainable AI models could help with this. These biases could disproportionately affect underrepresented groups and produce erroneous outputs in practice. Bias has been shown to occur even in simple heart disease prediction methods, which have been used extensively in routine medical practice [130]. To reduce any bias, more ECG recordings should be collected to capture a larger spread in human demographics and data point labels should be carefully chosen [131]. Facilitating data sharing between hospitals globally could further help generalise the detection algorithm, although there may be associated data privacy and security issues [132].

6.7 Impact Assessment Conclusion

The Impact Assessment aimed to determine the cost-effectiveness of the proposed AF detection system and PiP treatment, whilst also conducting an in-depth analysis of the potential impacts on key stakeholders. The objective of the CEA was to evaluate the relative cost-effectiveness, measured in č per QALYs, of the suggested PiP and the current Continuous anticoagulation strategies. A Markov CSTM was successfully used to implement the simulation and determine the economic viability of the new PiP system, fulfilling the objective. The results showed that the PiP pathway was likely more cost-efficient than the current Continuous pathway, providing compelling evidence for the feasibility of the designed solution. A PSA was performed to assess the robustness of the results to the identified uncertainty in model parameters. It was found that the PiP strategy remained more cost-effective than the Continuous strategy relative to the Base, regardless of the uncertainty. Model strengths and limitations were also discussed to contextualise the results and provide a basis for further studies.

The potential impacts of the product on key stakeholders were analysed to determine considerations or further research which may be needed before implementing it in practice. The greatest impact on patients was the vastly improved quality of life due to not only a reduced risk of ischaemic stroke but also a reduced major bleeding risk from PiP anticoagulation. However, more research is needed to determine how willing patients would be to continuously wear wearable devices and the significance of any health effects resulting from issues such as forgetting to wear them or not wearing them during sleep. For clinicians it was found that an increased workload could be expected due to more AF cases being detected, however this was deemed to be acceptable since more patients would be receiving personalised treatment. In general, clinicians were willing to adopt AI technology in diagnosis and treatment assistance provided it did not impede their daily practices. Before going into production, it would be beneficial to release a "beta" or prototype version of the system to gather feedback from clinicians and refine its usability.

Regarding the healthcare systems, the study determined that costs associated with implementing the new PiP system were significantly outweighed by the economic burden of stroke and major bleeds, providing further justification for the superior cost-effectiveness of the product. However, further studies and considerations need to be explored to ensure that patients on the PiP pathway receive the same quality of care as they would with in-person appointments with GPs. Also, more information about the adoption of AI in healthcare needs to be provided to patients and the public to improve accountability and transparency. Finally, a variety of funding opportunities should be considered to

ensure equitable access to the improved early AF detection and PiP treatment pathway.

Overall, the study provided strong evidence for the cost-effectiveness of the proposed solution and determined that there would be an overall benefit to the key stakeholders. Before the implementation of the product, more research would be needed to ensure the highlighted limitations are resolved.

Chapter 7

Conclusion

The study presented a design of an AF detection algorithm running on an Apple Watch to identify asymptomatic AF in the broader population accurately. It was also applied to facilitate Pill-in-the-Pocket anticoagulation medication pathway. These two product designs satisfied the unmet needs in the AF treatment pathway, identified during the research stage of the project.

The technical design introduced a novel ensemble-based neural network methodology for the classification of ECG signals. This technique underscored the advantages of combining multiple models, which resulted in superior performance compared to single-model approaches. In particular, our model exhibited exemplary performance on the MIT-BIH Arrhythmia Database, with an accuracy of 98.8%, a precision of 93.4%, and a recall rate of 97.8%. A notable F1 score of 95.5% highlighted a well-balanced harmony between precision and recall. Furthermore, the model achieved an exceptional Area Under the Curve (AUC) score of 0.99, illustrating its near-perfect ability to differentiate between the atrial fibrillation and normal classes. Impressively, our model surpassed the diagnostic capabilities of seasoned physicians, suggesting its promising potential as a valuable tool to facilitate faster and more accurate arrhythmia diagnoses, ultimately contributing to improved patient outcomes. Utilizing ensemble neural networks led to several advantages, including improved generalization, heightened resilience to noise, and reduced prediction variance. These merits emphasize the potential application of ensemble techniques across various medical and healthcare domains.

The two proposed validation study designs provide a good foundation for us to look for sponsors and plan the clinical trial application process. The studies aim to evaluate the performance of our algorithm and the feasibility of the algorithm-guided pill-in-the-pocket anticoagulation strategy. With

promising results from the trials, the solution will be granted the UKCA marking and CE marking to be commercially available and integrated with wearable devices.

The impact assessment provided compelling evidence for the economic feasibility of the product. The PiP treatment pathway was determined to be more cost-effective than the current Continuous anticoagulation pathway. Finally, the stakeholder analysis determined an overall benefit to the key stakeholders. The most significant advantages of the designed product were the reduced burden of stroke in patients and healthcare systems, and the reduced risk of major bleeds in PiP anticoagulation treatment. Both of these drastically reduce the associated costs to the healthcare system and improve patients' quality of life.

The design of the solution currently depends on the implementation of the Apple Electronic Health Record System in the wider NHS. If this system isn't widely implemented in the coming years, the user experience of our solution would be degraded. This is because user data stored locally on Apple Electronic Health Records will need to be transmitted through an alternative protocol compatible with NHS systems. However, as the FIHR protocol that Apple EHR is based on is interoperable, our code will only need minimal intervention to make it compatible with any future or legacy NHS EHR system.

Furthermore, our design implementation is built on the WatchOS developer platform. If the adoption of WatchOS declines, our solution could be obsolete as a method of population screening. Even though this is unlikely, this possibility suggests that we need to implement our design on multiple platforms. As our design is versatile, there is a higher likelihood that we are able to port our design to other platforms. This would be a further extension of our work.

Potential model limitations for the CEA were highlighted and discussed in Section 6.2.4.2. The suggested improvements could further improve the ICER estimates and fully represent the whole costs to healthcare systems, however this requires the availability of more data on PiP anticoagulation. Some limitations in the stakeholder analysis arose from the uncertainty of PiP treatment in practice, as it is still in ongoing research.

The encouraging results of our research motivate further investigation into variants of ensemble techniques. Future work may involve incorporating more advanced denoising filtering techniques [133], additional base models, exploring alternative ensemble strategies, and assessing the performance of our proposed method on an expanded range of datasets or medical conditions.

Further works for the validation studies are also required before the CTA (Clinical Trial Authorisation) submission and the start of the trial including sponsorship confirmation, risk assessment, ethics

assessments, and R&D consultation. We would also consider designing interim checks to provide a higher flexibility to the trials given the large sample size required, and reduce potential biases during the trial process.

To ensure a positive impact on patients, further studies need to be conducted to determine whether patients would receive the same quality of care with remote PiP treatment as they would with regular GP appointments. Other investigations, including surveys and questionnaires, are needed to gather perspectives from patients and clinicians on this technology and their willingness to adopt it.

Bibliography

[1] National Institute for Health and Care Excellence. *Atrial Fibrillation - Background Information*. Last revised in March 2023. 2023.

[2] Paul Burdett and Gregory Y H Lip. "Atrial fibrillation in the UK: predicting costs of an emerging epidemic recognizing and forecasting the cost drivers of atrial fibrillation-related costs". In: *European Heart Journal - Quality of Care and Clinical Outcomes* 8.2 (Dec. 2020), pp. 187–194. ISSN: 2058-5225. DOI: 10.1093/ehjqcco/qcaa093.

[3] NHS Inform. *Atrial fibrillation*. [Online; accessed 15-May-2023]. 2023.

[4] Neeraj Aggarwal et al. "Atrial Fibrillation in the Young: A Neurologist's Nightmare". In: *Neurol Res Int* 2015 (2015), p. 374352. DOI: 10.1155/2015/374352.

[5] Nicola J Adderley et al. "Prevalence and treatment of atrial fibrillation in UK general practice from 2000 to 2016". In: *Heart* 105.1 (2019), pp. 27–33.

[6] Laila Staerk et al. "Lifetime risk of atrial fibrillation according to optimal, borderline, or elevated levels of risk factors: cohort study based on longitudinal data from the Framingham Heart Study". In: *BMJ* 361 (2018). ISSN: 0959-8138. DOI: 10.1136/bmj.k1453.

[7] Ayodele Odutayo et al. "Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis". In: *BMJ* 354 (2016). DOI: 10.1136/bmj.i4482.

[8] P Yock et al. *Biodesign: The Process of Innovating Medical Technologies (2nd ed.)* Cambridge University Press, 2015.

[9] G Hindricks et al. "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC". In: *European Heart Journal* 42.5 (2020), pp. 373–498.

[10] S King et al. "Evidence Summary for Screening for Atrial Fibrillation in Adults". In: *UK National Screening Committee* (2019).

[11] D Sgreccia et al. "Comparing Outcomes in Asymptomatic and Symptomatic Atrial Fibrillation: A Systematic Review and Meta-Analysis of 81,462 Patients". In: *Journal of clinical medicine* 10.17 (2021).

[12] David Pendleton et al. *The New Consultation: Developing doctor-patient communication*. Oxford University Press, Apr. 2003. ISBN: 9780192632883. DOI: 10.1093/med/9780192632883.001.0001.

[13] Lei Sun et al. "A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs". In: *Health Inf Sci Syst* 8.1 (2020), p. 19. DOI: 10.1007/s13755-020-00103-x.

[14] Christian Szegedy et al. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV].

[15] Jeong-Hwan Kim et al. "Assessment of Electrocardiogram Rhythms by GoogLeNet Deep Neural Network Architecture". In: *Journal of Healthcare Engineering* 2019 (2019). Article ID 2826901, p. 2826901. DOI: 10.1155/2019/2826901.

[16] Forrest N. Iandola et al. *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*. 2016. arXiv: 1602.07360 [cs.CV].

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.

cc / paper _ files / paper / 2012 / file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[18] Atif Ullah Rahman et al. “ECG Classification for Detecting ECG Arrhythmia Empowered with Deep Learning Approaches”. In: *Comput Intell Neurosci* 2022 (2022), p. 6852845. DOI: 10.1155/2022/6852845.

[19] Zhihui He et al. “LiteNet: Lightweight Neural Network for Detecting Arrhythmias at Resource-Constrained Mobile Devices”. In: *Sensors (Basel)* 18.4 (2018), p. 1229. DOI: 10.3390/s18041229.

[20] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV].

[21] Mohamed Abdelazez, Sreeraman Rajan, and Antoni D Chan. “Transfer learning for detection of atrial fibrillation in deterministic compressive sensed ECG”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2020, pp. 5398–5401. DOI: 10.1109/EMBC44109.2020.9176324.

[22] Barret Zoph et al. *Learning Transferable Architectures for Scalable Image Recognition*. 2018. arXiv: 1707.07012 [cs.CV].

[23] Apple Inc. *Apple*. 2020. URL: https://www.apple.com/healthcare/docs/site/Apple_Watch_Arrhythmia_Detection.pdf.

[24] UK National Health Service. *Atrial Fibrillation (AF) suspected*. https://clinicalpathways.org.uk/sites/default/files/guidance/Cardiology/atrial-fibrillation-af-suspected.pdf. Reviewed: December 2021, Accessed on: April 29, 2023. 2018.

[25] British Heart Foundation. *Atrial fibrillation: finding the missing 300,000*. https://www.bhf.org.uk/for-professionals/healthcare-professionals/blog/2019/atrial-fibrillation-finding-the-missing-300000. Accessed on: April 29, 2023. 2019.

[26] K Farzam and A Jan. “Beta Blockers”. In: *StatPearls* (2022). [Updated 2022 Dec 27]; Accessed 2023 Apr 28.

[27] Marwa Shoeb and Margaret C Fang. “Assessing bleeding risk in patients taking anticoagulants”. In: *J Thromb Thrombolysis* 35.3 (2013), pp. 312–319. DOI: 10.1007/s11239-013-0899-7.

[28] Rod Passman. ““Pill-in-Pocket" Anticoagulation for Atrial Fibrillation: Fiction, Fact, or Foolish?” In: *Circulation* 143.22 (2021), pp. 2211–2213. DOI: 10.1161/CIRCULATIONAHA.121.053170.

[29] Rod Passman et al. “Targeted anticoagulation for atrial fibrillation guided by continuous rhythm assessment with an insertable cardiac monitor: The rhythm evaluation for anticoagulation with continuous monitoring (REACT.COM) pilot study”. en. In: *J. Cardiovasc. Electrophysiol.* 27.3 (Mar. 2016), pp. 264–270.

[30] Jonathan W Waks et al. “Intermittent anticoagulation guided by continuous atrial fibrillation burden monitoring using dual-chamber pacemakers and implantable cardioverter-defibrillators: Results from the Tailored Anticoagulation for Non-Continuous Atrial Fibrillation (TACTIC-AF) pilot study”. en. In: *Heart Rhythm* 15.11 (Nov. 2018), pp. 1601–1607.

[31] PhysioNet. *AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge, 2017*. Accessed: 2023-04-23. 2017. URL: https://physionet.org/challenge/2017/.

[32] Counterpoint Research. *Quarterly smartwatch unit shipment share worldwide from 2018 to 2022, by vendor*. Statista. Accessed: April 29, 2023. 2022.

[33] Oxford University Hospital NHS Foundation Trust. *Apples Health Records On iPhone Now Available to OUH Patients*. 2020. URL: https://www.ouh.nhs.uk/news/article.aspx?id=1390.

[34] Milton Keynes University Hospital NHS Foundation Trust. *Milton Keynes University Hospital offers Patients Health Records on iPhone*. 2020. URL: https://www.mkuh.nhs.uk/news/milton-keynes-university-hospital-offers-patients-health-records-on-iphone.

[35] Paolo Campanella et al. “The impact of electronic health records on healthcare quality: a systematic review and meta-analysis”. In: *European Journal of Public Health* 26.1 (July 2015), pp. 60–64. ISSN: 1101-1262. DOI: 10.1093/eurpub/ckv122. eprint: https://academic.oup.com/eurpub/article-pdf/26/1/60/7472041/ckv122.pdf.

[36] *The Medical Devices Regulations 2002*. SI 2002/618. 2002.

[37] Public Health England. *Population screening: our approach to screening standards*. Accessed on April 29, 2023. PHE publications gateway number: GW-586. 2019.

[38] European Union. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016.

[39] *The Health and Social Care Act 2008 (Regulated Activities) Regulations 2014*. https://www.legislation.gov.uk/uksi/2014/2936/contents/made. Accessed on April 29, 2023. 2014.

[40] Care Quality Commission. *Healthcare services: Key lines of enquiry, prompts and characteristics*. https://www.cqc.org.uk/sites/default/files/20180628%20Healthcare%20services%20KLOEs%20prompts%20and%20characteristics%20FINAL.pdf. Accessed on April 29, 2023. 2018.

[41] L.W. Smith. “Stakeholder analysis: a pivotal practice of successful projects”. In: *Project Management Institute Annual Seminars & Symposium*. Project Management Institute. Houston, TX, 2000.

[42] Jiapu Pan and Willis J. Tompkins. “A Real-Time QRS Detection Algorithm”. In: *IEEE Transactions on Biomedical Engineering* BME-32.3 (1985), pp. 230–236. DOI: 10.1109/TBME.1985.325532.

[43] MathWorks. *trainNetwork*. Accessed: 2023-04-23. 2023. URL: https://uk.mathworks.com/help/deeplearning/ref/trainnetwork.html#d124e196242.

[44] Dimitris Bertsimas, Luca Mingardi, and Bartolomeo Stellato. “Machine Learning for Real-Time Heart Disease Prediction”. In: *IEEE Journal of Biomedical and Health Informatics* 25.9 (2021), pp. 3627–3637. DOI: 10.1109/JBHI.2021.3066347.

[45] B. Zadrozny, J. Langford, and N. Abe. “Cost-sensitive learning by cost-proportionate example weighting”. In: *Third IEEE International Conference on Data Mining*. 2003, pp. 435–442. DOI: 10.1109/ICDM.2003.1250950.

[46] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[47] MathWorks. *zscore*. Accessed: 2023-04-23. 2023. URL: https://uk.mathworks.com/help/stats/zscore.html.

[48] MathWorks. *wdenoise*. Accessed: 2023-04-23. 2023. URL: https://uk.mathworks.com/help/wavelet/ref/wdenoise.html.

[49] Mohamed Hammad et al. “Deep Learning Models for Arrhythmia Detection in IoT Healthcare Applications”. In: *Computers and Electrical Engineering* 100 (2022), p. 108011. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2022.108011. URL: https://www.sciencedirect.com/science/article/pii/S0045790622002786.

[50] MathWorks. *Classify ECG Signals Using Long Short-Term Memory Networks*. Accessed: 2023-04-23. 2023. URL: https://uk.mathworks.com/help/deeplearning/ug/classify-ecg-signals-using-long-short-term-memory-networks.html.

[51] Mike Schuster and Kuldip Paliwal. “Bidirectional recurrent neural networks”. In: *Signal Processing, IEEE Transactions on* 45 (Dec. 1997), pp. 2673 –2681. DOI: 10.1109/78.650093.

[52] B. Boashash. “Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals”. In: *Proceedings of the IEEE* 80.4 (1992), pp. 520–538. DOI: 10.1109/5.135376.

[53] B. Boashash. “Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications”. In: *Proceedings of the IEEE* 80.4 (1992), pp. 540–568. DOI: 10.1109/5.135378.

[54] Y N Pan, J Chen, and X L Li. “Spectral entropy: A complementary index for rolling element bearing performance degradation assessment”. In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 223.5 (2009), pp. 1223–1231. DOI: 10.1243/09544062JMES1224. eprint: https://doi.org/10.1243/09544062JMES1224. URL: https://doi.org/10.1243/09544062JMES1224.

[55] P. Welch. “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms”. In: *IEEE Transactions on Audio and Electroacoustics* 15.2 (1967), pp. 70–73. DOI: 10.1109/TAU.1967.1161901.

[56] Wikipedia. *Periodogram*. Accessed: 2023-04-23. 2023. URL: https://en.wikipedia.org/wiki/Periodogram.

[57] MathWorks. *Comparing model size, speed, and accuracy for popular pretrained networks*. Accessed: 2023-04-23. 2023. URL: https://uk.mathworks.com/discovery/transfer-learning.html.

[58] Wikipedia. *Continuous wavelet transform*. Accessed: 2023-04-23. 2023. URL: https://en.wikipedia.org/wiki/Continuous_wavelet_transform.

[59] Christian Szegedy et al. “Going deeper with convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[60] Forrest Iandola et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size”. In: (Feb. 2016).

[61] Mark Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[62] Barret Zoph et al. “Learning Transferable Architectures for Scalable Image Recognition”. In: June 2018, pp. 8697–8710. DOI: 10.1109/CVPR.2018.00907.

[63] MathWorks. *Visualize Activations of a Convolutional Neural Network*. Accessed: 2023-04-23. 2023. URL: https://uk.mathworks.com/help/deeplearning/ug/visualize-activations-of-a-convolutional-neural-network.html.

[64] Jichi Chen et al. “Convolutional neural network with transfer learning approach for detection of unfavorable driving state using phase coherence image”. In: *Expert Systems with Applications* 187 (2022), p. 116016. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.116016. URL: https://www.sciencedirect.com/science/article/pii/S0957417421013634.

[65] Lior Rokach. “Ensemble-based classifiers”. In: *Artif. Intell. Rev.* 33 (Feb. 2010), pp. 1–39. DOI: 10.1007/s10462-009-9124-7.

[66] U.S. Food and Drug Administration (FDA). *CFR - Code of Federal Regulations Title 21*. 2023.

[67] UK Government. *Regulating medical devices in the UK*. 2022.

[68] FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. Food and Drug Administration (US), 2016.

[69] IMDRF SaMD Working Group. *Software as a Medical Device (SaMD): Key Definitions*. Tech. rep. International Medical Device Regulators Forum, 2017.

[70] Office of the Commissioner. *FDA in brief: FDA seeks public feedback on biomarker and study endpoint glossary*. FDA. 2019.

[71] National Academies of Sciences, Engineering, and Medicine. *An Evidence Framework for Genetic Testing*. The National Academies Press, 2017. DOI: 10.17226/24632.

[72] Jennifer C. Goldsack, Andrea Coravos, Job P. Bakker, et al. “Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs)”. In: *npj Digital Medicine* 3 (2020), p. 55. DOI: 10.1038/s41746-020-0260-4.

[73] Kirsty Loudon et al. “The PRECIS-2 tool: designing trials that are fit for purpose”. In: *BMJ* 350 (2015). DOI: 10.1136/bmj.h2147.

[74] Apple Inc. *Using Apple Watch for Arrhythmia Detection*. Internal Study. Apple Inc., 2020.

[75] Mintu P. Turakhia et al. “Estimated prevalence of undiagnosed atrial fibrillation in the United States”. In: *PLOS ONE* 13.4 (Apr. 2018), pp. 1–11. DOI: 10.1371/journal.pone.0195088.

[76] Julian P.T. Higgins and Sally Green, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: John Wiley & Sons, 2008.

[77] Shaun Treweek and Matthias Briel. “Digital tools for trial recruitment and retentionplenty of tools but rigorous evaluation is in short supply”. In: *Trials* 21 (2020), p. 476. DOI: 10.1186/s13063-020-04361-8.

[78] Rod Passman et al. “Targeted Anticoagulation for Atrial Fibrillation Guided by Continuous Rhythm Assessment With an Insertable Cardiac Monitor: The Rhythm Evaluation for Anticoagulation With Continuous Monitoring (REACT.COM) Pilot Study”. en. In: *J Cardiovasc Electrophysiol* 27.3 (Nov. 2015), pp. 264–270.

[79] Stavros Stavrakis et al. “Intermittent vs. Continuous Anticoagulation theRapy in patiEnts with Atrial Fibrillation (iCARE-AF): a randomized pilot study”. en. In: *J Interv Card Electrophysiol* 48.1 (Oct. 2016), pp. 51–60.

[80] Andrea Gažová et al. “Predictive value of CHA2DS2-VASc scores regarding the risk of stroke and all-cause mortality in patients with atrial fibrillation (CONSORT compliant)”. In: *Medicine (Baltimore)* 98.31 (2019), e16560. DOI: 10.1097/MD.0000000000016560.

[81] Hisashi Ogawa et al. “Progression From Paroxysmal to Sustained Atrial Fibrillation Is Associated With Increased Adverse Events”. In: *Stroke* 49.10 (2018), pp. 2301–2308. DOI: 10.1161/STROKEAHA.118.021396.

[82] Roxana Mehran et al. “Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium”. In: *Circulation* 123.23 (2011), pp. 2736–2747.

[83] Jonathan W. Waks et al. “Intermittent anticoagulation guided by continuous atrial fibrillation burden monitoring using dual-chamber pacemakers and implantable cardioverter-defibrillators: Results from the Tailored Anticoagulation for Non-Continuous Atrial Fibrillation (TACTIC-AF) pilot study”. In: *Heart Rhythm* 15.11 (2018), pp. 1601–1607. ISSN: 1547-5271. DOI: https://doi.org/10.1016/j.hrthm.2018.06.027.

[84] *Consultation on proposals for legislative changes for clinical trials*. https://www.gov.uk/government/consultations/consultation-on-proposals-for-legislative-changes-for-clinical-trials. Accessed: 2023-03-21. gov.uk.

[85] L Prieto and J A Sacristán. “Problems and solutions in calculating quality-adjusted life years (QALYs)”. In: *Health and quality of life outcomes* 1.80 (2003).

[86] A-E Fernando et al. “An Introductory Tutorial on Cohort State-Transition Models in R Using a Cost-Effectiveness Analysis Example”. In: *Medical Decision Making* 43.1 (2023), pp. 3–20.

[87] M Hunink et al. *Decision Making in Health and Medicine: Integrating Evidence and Values (2nd ed.)* Cambridge University Press, 2014. Chap. 10.

[88] N Gahungu et al. “Paroxysmal atrial fibrillation”. In: *BMJ* 375 (2021). DOI: 10.1136/bmj-2021-058568.

[89] G Lip. *CHA_2DS_2-VASc Score for Atrial Fibrillation Stroke Risk*.

[90] University of Auckland Department of Statistics. *Stochastic Processes Course Notes*.

[91] The National Institute for Health and Care Excellence. *Atrial fibrillation: diagnosis and management*. 2021.

[92] The National Institute for Health and Care Excellence. *CHTE methods review: Discounting*. 2020.

[93] M Shoeb and M C Fang. “Assessing bleeding risk in patients taking anticoagulants”. In: *Journal of thrombosis and thrombolysis* 35.3 (2013), pp. 312–319.

[94] P A Noseworthy et al. “Direct Comparison of Dabigatran, Rivaroxaban, and Apixaban for Effectiveness and Safety in Nonvalvular Atrial Fibrillation”. In: *Chest* 150.6 (2016).

[95] The National Institute for Health and Care Excellence. *Apixaban for preventing stroke and systemic embolism in people with nonvalvular atrial fibrillation*. 2021.

[96] A Briosa e Gala et al. “Pill-in-the-pocket Oral Anticoagulation Guided by Daily Rhythm Monitoring for Stroke Prevention in Patients with AF: A Systematic Review and Meta-analysis”. In: *Arrhythmia Electrophysiology Review 2023;12:e05* (2023).

[97] H Shafeeq and Tran T H. “New oral anticoagulants for atrial fibrillation: are they worth the risk?” In: *P T : a peer-reviewed journal for formulary management* 39.1 (2014), pp. 54–64.

[98] R M Kaplan et al. “Stroke Risk as a Function of Atrial Fibrillation Duration and CHA2DS2-VASc Score”. In: *Circulation* 140.20 (2019).

[99] C M Saborido et al. “Systematic review and cost-effectiveness evaluation of pill-in-the-pocket strategy for paroxysmal atrial fibrillation compared to episodic in-hospital treatment or continuous antiarrhythmic drug therapy”. In: *Health Technology Assessment* 14.31 (2010).

[100] P Dorian et al. “Cost-effectiveness of apixaban vs. current standard of care for stroke prevention in patients with atrial fibrillation”. In: *European heart journal* 35 (28 2014), pp. 1897–1906.

[101] J W Waks et al. “Intermittent anticoagulation guided by continuous atrial fibrillation burden monitoring using dual-chamber pacemakers and implantable cardioverter-defibrillators: Results from the Tailored Anticoagulation for Non-Continuous Atrial Fibrillation (TACTIC-AF) pilot study”. In: *Heart Rhythm* 15 (11 2018), pp. 1601–1607.

[102] R Passman et al. “Targeted Anticoagulation for Atrial Fibrillation Guided by Continuous Rhythm Assessment With an Insertable Cardiac Monitor: The Rhythm Evaluation for Anticoagulation With Continuous Monitoring (REACT.COM) Pilot Study”. In: *J Cardiovasc Electrophysiol* 27 (3 2016), pp. 264–270.

[103] Apple Inc. *Apple Watch Series 8 Cost*. 2023.

[104] P Youman et al. “The economic burden of stroke in the United Kingdom”. In: *Pharmacoeconomics* 21 Suppl 1 (2003).

[105] J S Healey et al. “Subclinical Atrial Fibrillation in Older Patients”. In: *Circulation* 136 (14 2017), pp. 1276–1283.

[106] York Health Economics Consortium. *Incremental Cost-Effectiveness Ratio (ICER) [online]*. 2016.

[107] D J Cohen and M R Reynolds. “Interpreting the results of cost-effectiveness studies”. In: *Journal of the American College of Cardiology* 52.25 (2008), pp. 2119–2126.

[108] K Bambha and W R Kim. “Cost-effectiveness analysis and incremental cost-effectiveness ratios: uses and pitfalls”. In: *European Journal of Gastroenterology Hepatology* 16.6 (2004), pp. 519–526.

[109] Northwestern Now. *Can Apple Watch reduce patients reliance on blood thinners?* 2022.

[110] H Ogawa et al. “Progression From Paroxysmal to Sustained Atrial Fibrillation Is Associated With Increased Adverse Events”. In: *Stroke* 49.10 (2018).

[111] S Nattel et al. “Early management of atrial fibrillation to prevent cardiovascular complications”. In: *European Heart Journal* 35.22 (2014), pp. 1448–1456. DOI: 10.1093/eurheartj/ehu028.

[112] A García-Fernández, V Roldán, and F Marín. “Strategies for prediction and early detection of atrial fibrillation: present and future”. In: *EP Europace* 19.4 (2016), pp. 515–517.

[113] C Steger et al. “Stroke patients with atrial fibrillation have a worse prognosis than patients without: data from the Austrian Stroke registry”. In: *European Heart Journal* 25.19 (2004), pp. 17341740.

[114] J Tracz et al. “Long-Term Outcomes after Stroke in Patients with Atrial Fibrillation: A Single Center Study”. In: *International journal of environmental research and public health* 20 (4 2023), p. 3491.

[115] Mayo Clinic. *Holter monitor*. 2022.

[116] *Zio by iRhythm*. 2023.

[117] P M Barrett et al. “Comparison of 24-hour Holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring”. In: *The American journal of medicine* 127.1 (2014).

[118] British Heart Foundation. *Implantable Loop Recorders*. 2021.

[119] A Truly. “How Long Does An Apple Watch Last? Here Are Some Life-Extending Tips”. In: (2022).

[120] A Briosa E Gala et al. “Patients perspective on a pill-in-the-pocket oral anticoagulation as an alternative stroke prevention strategy in atrial fibrillation”. In: *EP Europace* 24 Suppl 1 (2022).

[121] British Medical Association. *Pressures in general practice data analysis*. 2023.

[122] British Medical Association. *NHS backlog data analysis*. 2023.

[123] C Stewart. *Artificial intelligence (AI) in healthcare market size worldwide from 2021 to 2030*. 2023.

[124] T Davenport and R Kalakota. “The potential for artificial intelligence in healthcare”. In: *Future healthcare journal* 6.2 (2019), pp. 94–98.

[125] NHS England - Transformation Directorate. *Artificial Intelligence*. 2023.

[126] P Burdett and G Y H Lip. “Atrial fibrillation in the UK: predicting costs of an emerging epidemic recognizing and forecasting the cost drivers of atrial fibrillation-related costs”. In: *European Heart Journal - Quality of Care and Clinical Outcomes* 8.2 (2022), pp. 187194.

[127] D Khullar et al. “Perspectives of Patients About Artificial Intelligence in Health Care”. In: *JAMA network open* 5.5 (2022).

[128] P A Grady and L L Gough. “Self-management: a comprehensive approach to management of chronic conditions”. In: *American journal of public health* 104.8 (2014).

[129] Nuffield Trust. *Poorest get worse quality of NHS care in England, new research finds*. 2020.

[130] K J Igoe. “Algorithmic Bias in Health Care Exacerbates Social Inequities How to Prevent It”. In: *Harvard T.H. Chan School of Public Health* (2021).

[131] Obermeyer Z et al. “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366.6464 (2019), pp. 447453.

[132] N Norori et al. “Addressing bias in big data and AI for health care: A call for open science”. In: *Patterns* 2.10 (2021), p. 100347.

[133] Leslie Casas et al. *Adversarial Signal Denoising with Encoder-Decoder Networks*. 2020. arXiv: 1812.08555 [cs.LG].

Appendix: Modified GoogLeNet Architecture

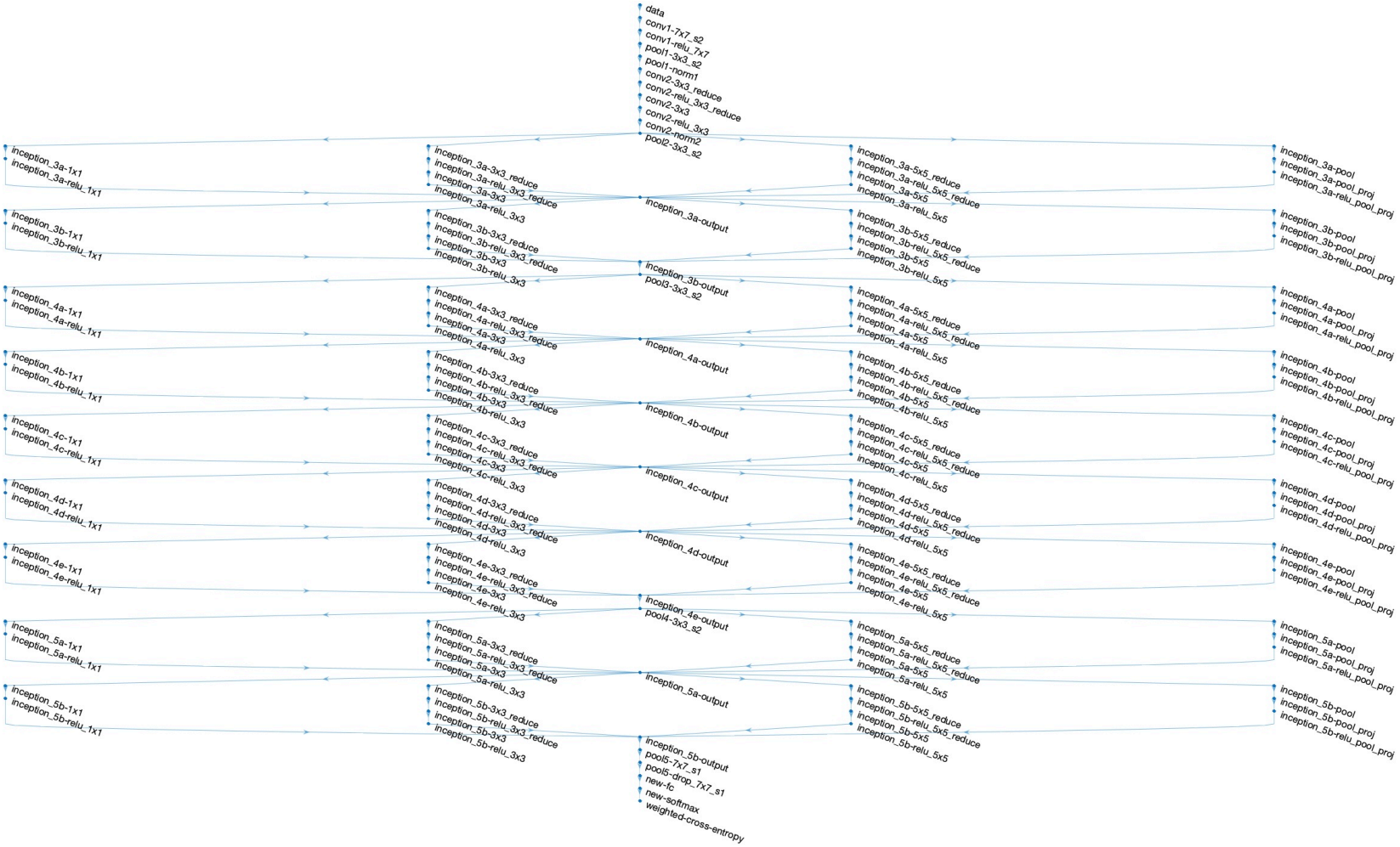


Figure 7.1: Customized GoogLeNet architecture.

Appendix: Modified SqueezeNet Architecture

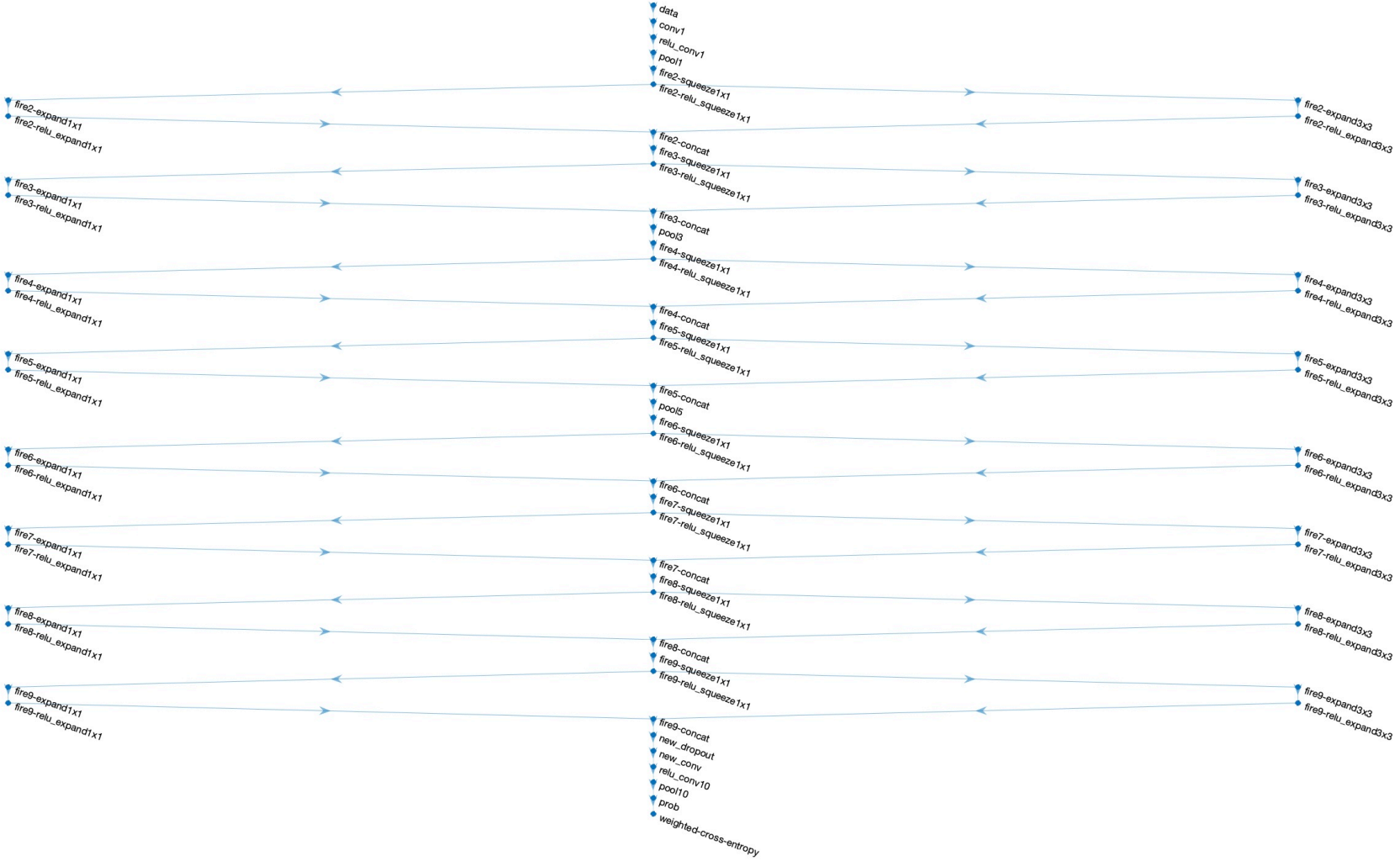


Figure 7.2: Customized SqueezeNet architecture.

Appendix: Current Patient Pathway

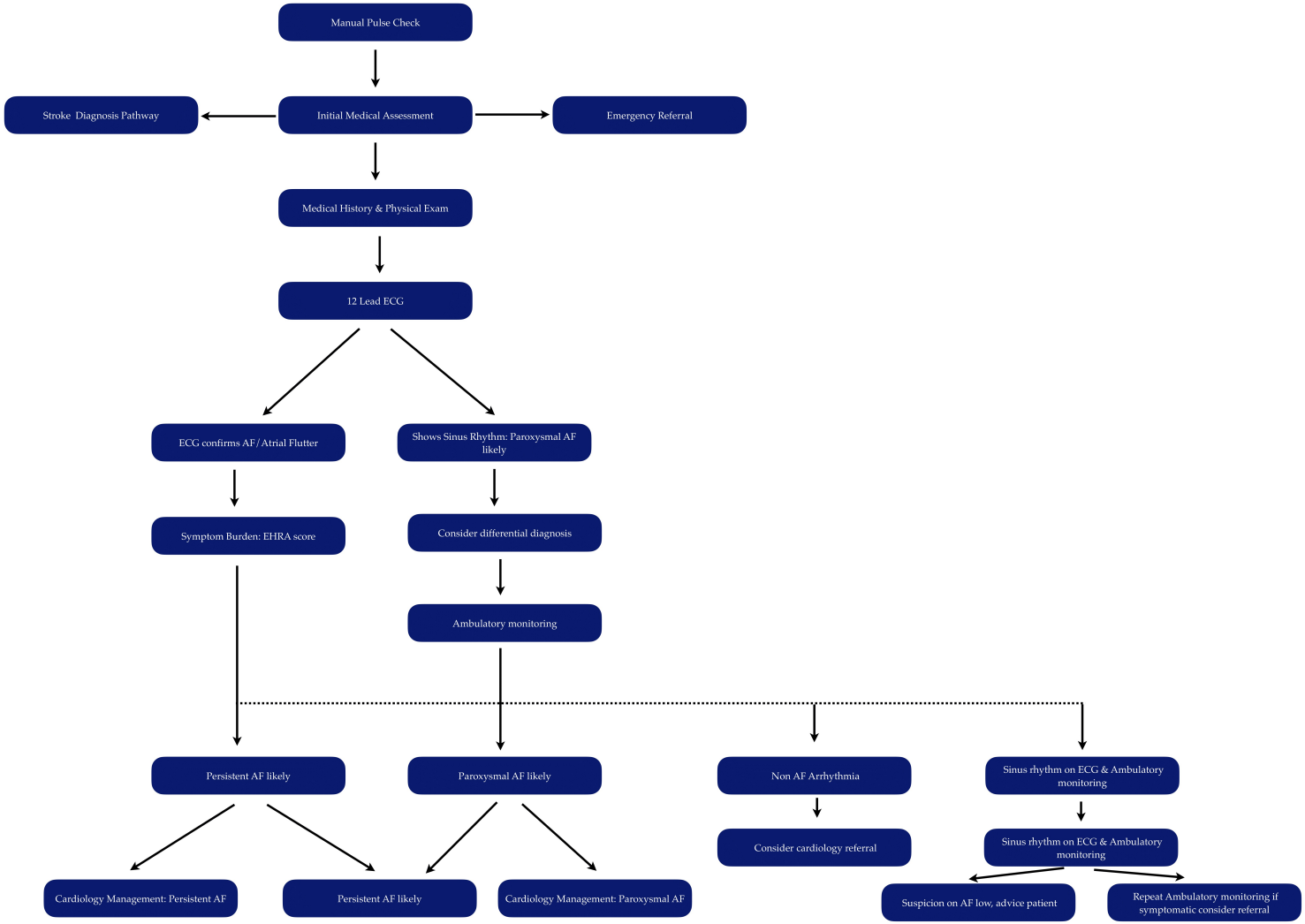


Figure 7.3: Current Patient Pathway

Appendix: PRECIS-2 Evaluation for The AFib-AI Study

PRECIS-2 Domains	Score	Rationale
Eligibility	5	Inclusion and Exclusion Criteria are expected to be simple. Anyone over 22 and a compatible wearable device should be allowed as a participant.
Recruitment	5	Past trials have shown high enthusiasm even without incentives. Due to the Apple ecosystem, virtue recruitment is valid.
Setting	5	Identical settings to usual care setting. Participants do not require any change in habit or lifestyle.
Organisation	5	Identical organisation to usual care with no additional training required for staff and GP. Participants may need to understand how to use the device after a warning. This can be done with a short video on the app.
Flexibility (delivery)	5	Highly flexible in delivery with no additional resources required. However, the participants would require to own and purchase the device beforehand at their own expense.
Flexibility (adherence)	5	Participants free to not comply with the warning. There will be no action to encourage compliance. Compliance is also measured in secondary endpoints with the number of participants attending GP appointments with positive AF detection.
Follow-up	3	Follow up after one month with an additional online survey after three months, while no follow-up is required in usual care. Data Collection is required to log the booked appointment with GP individually in a calendar.
Primary outcome	1	Primary outcome is not relevant to participants. The aim is to evaluate the accuracy and precision of the algorithm.
Primary analysis	5	All participants' data will be accepted for analysis, including patients who did not comply.

Table 7.1: PRECIS-2 Evaluation with Rationale for the AFib-AI Study.

Appendix: PRECIS-2 Evaluation for AI-PiP Study

PRECIS-2 Domains	Score	Rationale
Eligibility	2	Eligibility criteria are similar but more aggressive than usual care by allowing patients with persistent AF to participate. Strict exclusion criteria for low stroke risks with CHA_2DS_2-VASc of 1-3.
Recruitment	1	Recruitment will be acquired through cardiologist appointments for regular checkup patients and would require invitation emails or letters for eligible patients. A large sample size is required, so we might also need incentives.
Setting	4	Settings will not change compared with usual care. However, there will be inconvenience in wearing the device at all times.
Organisation	5	Identical organisation to usual care. Participants would require to be able to identify and understand warnings to take OAC.
Flexibility (delivery)	2	There will be strict protocols to ensure accurate and continuous heart rate signals. Requires participants to wear the device at all times with measures to improve compliance.
Flexibility (adherence)	1	Participants are required to comply with the trial instructions, including wearing the device at all times and following the advice promptly. There will be extra measures to ensure compliance through notifications if there is missing data and video guides about the intervention
Follow-up	1	There will be regular follow-ups to check for AF development and reassess risks of stroke (CHA_2DS_2-VASc Score).
Primary outcome	5	Primary outcomes are incidences of stroke and bleeding. They are directly related to the patients as this trial aims to test the intervention's safety.
Primary analysis	1	Data for participants without compliance are not accepted. Due to the seriousness of the intervention, we will assume full compliance.

Table 7.2: PRECIS-2 Evaluation with Rationale for the AI-PiP Study.