Towards broader application of deep learning methods to the automated analysis of electrocardiograms

Dr Rob Brisk, MBBCh, MRCP(London)

Faculty of Computing, Engineering and the Built Environment of Ulster University

Thesis submitted for the degree of Doctor of Philosophy

September 2022

(I can confirm that the word count of this Thesis is less than 100,000 words)

Table of Contents

CHAPTER 4: DEEP LEARNING TO AUTOMATICALLY INTERPRET IMAGES OF THE ELECTROCARDIOGRAM: DO WE NEED THE RAW SAMPLES?

CHAPTER 5: WASP-ECG: A WAVE SEGMENTATION PRETRAINING TOOLKIT FOR ELECTROCARDIOGRAM ANALYSIS

APPENDIX 1: PERSONAL ECG DEVICES: HOW WILL HEALTHCARE SYSTEMS COPE? A SINGLE CENTRE CASE STUDY

APPENDIX 2: AI TO ENHANCE INTERACTIVE SIMULATION-BASED TRAINING IN RESUSCITATION MEDICINE

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my supervisors, Prof Raymond Bond, Dr David McEneaney, Prof Dewar Finlay and Prof James McLaughlin. As first academic supervisor, Prof Raymond Bond's advice, patience and encouragement have been invaluable throughout this work. Without the unwavering support of Dr David McEneaney since I joined the Southern Trust Cardiology Department in 2016, I would have had neither the opportunity nor the resolve to undertake this PhD project.

I would also like to thank the co-authors of the various publications that have resulted from this thesis period, from whom I've learned a huge amount.

This PhD was undertaken as part of a scholarship from the Eastern Corridor Medical Engineering Centre that is supported by the European Union's INTERREG VA Programme, managed by the Special EU Programmes Body (SEUPB). It was also supported by the Craigavon Cardiac Care Association, whom I wish to thank for their active support of cardiovascular research in Northern Ireland over the last 50 years.

This thesis is dedicated to my wife, Maeve.

Abstract

Introduction: Automated analysis of the electrocardiogram (ECG) is one of the most impactful forms of computerised clinical decision support. Recent advanced in deep learning (DL) techniques have shown the potential to overcome historical limitations of manually engineered ECG analysis. However, DL-based ECG analysis is a nascent technology and questions remain about its limitations and usability.

Methods: To investigate the limits of DL's capacity for detecting acute myocardial infarction (AMI), a DL classifier was trained to detect early onset of coronary artery occlusion. To investigate DL's ability to broaden access to ECG analysis through high-quality interpretation of ECG images, a DL classifier was trained to detect atrial fibrillation (AF) from images of ambulatory ECG recordings. To evaluate a novel approach to reducing the volume of training data required to train DL-based ECG analysers, a system was developed to simulate ECG signals and corresponding wave segmentation masks. DL models were pretrained using this synthetic data, fine-tuned to detect AMI and AF from real ECGs, then compared again non-pretrained models.

Results: A DL model was unable to detect hyperacute coronary occlusion better than a random chance classifier. Performance appeared better in an earlier iteration of the experiment, but this appeared to be due to data leakage. A DL model detected AF from ECG images with equivalent accuracy to raw ECG samples. Pretraining with synthetic ECG data reduced the need for training on real ECGs to achieve comparable accuracy and provided a potential mechanism for clinician confidence calibration.

Conclusion: DL requires large volumes of training data and suffers from a "black box" effect. DL can broaden access to automated ECG analysis through high quality interpretation of ECG image data. Wave segmentation pretraining reduces the need for training data and provides a potential mechanism for confidence calibration. This may ameliorate the black box phenomenon.

[299 words]

Note on Access to Contents

Abbreviations

Abbreviation	Term
2D	2-dimensional
5FCV	5-fold cross validation
ACTCO	Acute complete thrombotic coronary occlusion
AF	Atrial fibrillation
AI	Artificial intelligence
ALS	Advanced life support
AMI	Acute myocardial infarction
ANN	Artificial neural network
AU	Activation unit
AUROC	Area under the receiver operating characteristic curve
AV node	Atrio-ventricular node
CNN	Convolutional neural network
CP	Cardiac physiologist
CPU	Central processing unit
CV	Cross validation
DL	Deep learning
ECG	Electrocardiogram
ED	Emergency department
GNN	Graph neural network
GP	General practitioner
GPU	Graphics processing unit
HPC	High performance computing
ILR	Implantable loop recorder
IT	Information technology
LAD	Left anterior descening artery
LCX	Left circumflex artery
LSTM	Long-short term memory
MHA	Multi-headed attention
ML	Machine learning
MLP	Multi-layer perceptron
NICE	National Institute for Clinical Excellence
NLP	Natural language processing
NSR	Normal sinus rhythm
NSTEMI	Non-ST elevation myocardial infarction
PCA	Principle component analysis

PCI	Peripheral component interface
PINN	Physics informed neural network
PIRL	Pretext invariant representation learning
PPCI	Primary percutaneous coronary intervention
PPV	Positive predictive value
QALY	Quality adjusted life year
RC	Routine care
RCA	Right coronary artery
RDMA	Remote direct memory access
ReLU	Rectified linear activation unit
RL	Representation learning
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SA node	Sino-atrial node
SHSCT	Southern health and social care trust
SNR	Signal-to-noise ratio
STE	ST segment elevation
STEMI	ST elevation myocardial infarction
VCG	Vectorcardiogram
WaSP	Wave segmentation pretraining
WCT	Wilson Central Terminal

Chapter 1:

Introduction and thesis summary

1.1 Introduction

Computerised electrocardiogram (ECG) analysis was one of the first major applications of automated diagnostics in healthcare, having been an active field of research and development for over 60 years [1]. However, recent breakthroughs in artificial intelligence (AI) based on deep learning (DL) technology have allowed computerised ECG researchers to overcome some important long-standing challenges, with two particularly notable studies carried out in 2018 marking something of a watershed in the evolution of the field [2, 3]. (These will be discussed further in the next chapter.)

Given how recently these developments have occurred, many unanswered questions remain about the limitations of AI for ECG interpretation, and about how best to translate this technology into safe and effective clinical applications. This thesis aims to address some of these research questions over the following five chapters.

1.2 Thesis outline and research questions

Chapter 2 presents a more extensive review of the topic of AI for ECG analysis. It is broken into two parts.

The first is a narrative review of the topic as a whole, including a summary of previous work in the field. This identifies three key research questions:

1. Can AI help us detect myocardial ischaemia ("heart attacks", in lay terms) earlier than was previously possible?
2. Can DL improve the performance ECG image analysis applications (where non-AI methods have historically faced significant challenges)?
3. How can ECG AI be 'democratised' to better cater for population groups and disease cohorts where there is a relatively paucity of labelled data?

Following on from the first research question, the second part of chapter 2 presents a systematic literature review of DL for ischaemia detection.

Chapter 3 presents original research that tests the hypothesis that DL methods can allow for the hyperacute diagnosis of acute myocardial infarction (AMI). In the event, the results fail to reject the null hypothesis. However, they highlight the challenge of identifying confounding data features learned by DL models in the clinical setting. This theme is explored throughout chapters 4 and 5.

Chapter 4 moves towards a more translational focus and presents original research investigating the possibility that DL methods can improve results in the area of ECG image analysis. This is a field of study whose value has been widely noted, but where conventional methods have failed to achieve results comparable with raw sample analysis. The results of this

chapter support the hypothesis that DL can improve the performance of ECG image analysis applications. Chapter 4 also outlines the advantages of ECG image analysis in terms of interpretability, which builds on the discussion within the previous chapter of the risks of DL models leveraging confounding data features.

Chapter 5 focuses primarily on representation learning (RL) as a means for pretraining DL models that can be 'fine-tuned' for a range of downstream ECG tasks with relatively small labelled datasets. RL is proposed as a means to more easily bring the advantages of DL to applications where data paucity may otherwise have been a barrier. This chapter also investigates a different approach to ECG image analysis, and proposes mechanisms by which DL applications could be made more interpretable for clinicians.

Chapter 6 recaps the conclusions of the previous chapters, proposes areas of future research within the field of ECG AI, makes a note of related works undertaken during the period of this thesis, and suggests how the work presented in this thesis may relate to the wider field of medical AI.

1.3 Research outputs

1.3.1 Journal papers

- **Brisk, R**., Bond, R. R., Finlay, D., McLaughlin, J. A., Piadlo, A. J., & McEneaney, D. J. WaSP-ECG: A wave segmentation pretraining toolkit for electrocardiogram analysis. Frontiers in Physiology, 2022;13:ePub (https://doi.org/10.3389/fphys.2022.760000)
- **Brisk, R**., Bond, R. R., Finlay, D., McLaughlin, J., Piadlo, A., Leslie, S. J., . . . Warren, S. The effect of confounding data features on a deep learning algorithm to predict complete coronary occlusion in a retrospective observational setting. European Heart Journal-Digital Health, 2021;2(1):127-134.
- **Brisk, R**., Bond, R. R., Banks, E., Piadlo, A., Finlay, D., McLaughlin, J., & David, M. Deep learning to automatically interpret images of the electrocardiogram: Do we need the raw samples? Journal of Electrocardiology, 2019;57:65-69.

1.3.2 Conference contributions

1.3.2.1 Papers

- **Brisk, R**., Bond, R., Finlay, D., & McEneaney, D. Personal ECG devices: How will healthcare systems cope? A single centre case study. Proceedings of the 2019 Computing in Cardiology conference, 2019;1-4.

1.3.2.2 Posters

- **Brisk, R**., Bond, R., Finlay, D., McLaughlin, J., Jasinska-Piadlo, A., Jennings, M., & McEneaney, D. Neural networks for ischaemia detection: Revolution or red herring? A systematic review and meta-analysis. 45th International Society for Computerized Electrocardiology meeting, Sep 2020 (virtual).
- **Brisk, R**., Bond, R., Liu, J., Finlay, D., McLaughlin, J., & McEneaney, D. (2018). AI to enhance interactive simulation-based training in resuscitation medicine. British HCI Conference, Jul 2018 (Belfast, UK).

1.3.2.3 Oral presentations

- Artificial intelligence: the future of automated ECG analysis? STAFF Symposium, Sep 2019 (Les Diablerets, Switzerland).
- AI for ECG interpretation: can we overcome the black box effect? 44th International Society for Computerized Electrocardiology meeting, Apr 2019 (Atlantic Beach, Florida, USA).

1.3.3 Co-authored journal papers

- Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., **Brisk, R**., Campbell, P., & McEneaneny, D. What can machines learn about heart failure? A systematic literature review. International Journal of Data Science and Analytics, 2021;1-21.
- Finlay, D., Bond, R., Jennings, M., McCausland, C., Guldenring, D., Kennedy, A., Biglarbegi, P., Al-Zaiti, S. S., **Brisk, R**., McLaughlin, J. Overview of featurization techniques used in traditional versus emerging deep learning-based algorithms for automated interpretation of the 12-lead ECG. Journal of Electrocardiology, 2021;69:7-11.
- Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., **Brisk, R**., Campbell, P., Browne, F., & McEneaneny, D. Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset. International Journal of Data Science and Analytics, 2020;1-18.

1.3.4 Co-authored conference papers

- Jennings, M. R., Biglarbeigi, P., Bond, R. R., **Brisk, R**., Güldenring, D., Kennedy, A., McLaughlin, J., Finlay, D. D. Machine learning approach to assess the performance of patch based leads in the detection of ischaemic electrocardiogram changes. Proceedings of the 2020 Computing in Cardiology conference, 2020;1-4.
- Jennings, M. R., Rababah, A. S., Biglarbeigi, P., **Brisk, R**., Güldenring, D., Bond, R., McLaughlin, J., Finlay, D. D. Coefficients for the derivation of posterior and right sided chest leads from the 12-lead electrocardiogram. Proceedings of the 2020 Computing in Cardiology conference, 2020;1-4.

- Bond, R. R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., **Brisk, R.**, Boger, T., Adel, T. Human centered artificial intelligence: Weaving UX into algorithmic decision making. Proceedings of RoCHI 2019, 2019;2-9.

1.4 References

[1] Macfarlane PW, Kennedy J. Automated ECG Interpretation—A brief history from high expectations to deepest networks. Hearts. 2021;2:433-48.

[2] Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25:65-9.

[3] Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet. 2019;394:861-7.

Chapter 2:

Background literature

2.1 Chapter structure

This chapter is split into two parts: a narrative review of AI for ECG analysis, followed by a systematic review and meta-analysis of AI for ischaemia detection.

2.2 Introduction to the narrative review of AI for ECG analysis

Unlike in the aviation industry, which has seen substantial success in mitigating the fallibility of individuals in life-critical settings, human error remains a major driver of adverse outcomes in modern healthcare [1, 2]. Acute cardiology is a particularly high-stakes specialty, where simple mistakes sometimes have disastrous consequences [3]. Although it is generally accepted that medical errors are an inevitability of complex healthcare environments, it has also been shown that effective use of information technology can substantially reduce this risk [4, 5].

Historically, the impact of healthcare technology has been limited by computers' inability to undertake some particularly challenging tasks, such as reliably detecting important diagnostic features in radiology images, or making sense of natural language to detect symptoms from a patient history [6, 7] The advent of modern machine learning (ML)-based AI has caused a substantial re-think of what is possible with computers in healthcare [8]. However, what some are calling the healthcare 'AI revolution' remains in its infancy [9]. Many questions remain, both about what is possible and what is advisable [10].

A particularly appropriate lens for exploring the potential of AI to reduce the burden of human error in healthcare is the ECG. It is one of the most important diagnostic tools in modern medicine, but also frequently misinterpreted [11]. The rest of this introductory chapter will cover the following points:

- An overview of the principles of electrocardiography
- An introduction to computerised ECG analysis
- An overview of AI and its relevance to ECG analysis
- A note on key open research questions in AI for ECG analysis

2.3 Electrocardiography basics

2.3.1 Origins of electrocardiography

In 1924, Willem Einthoven won the Nobel Prize in Medicine for discovering the ECG [12]. An ECG comprises a 2-dimensional (2D) recording of the electrical activity of the heart, with time on the X axis and voltage on the Y axis. Einthoven used a device called a 'string

galvanometer', though modern devices use more sophisticated and practical recording methods [13].

2.3.2 Naming conventions of the ECG

Einthoven's first ECG recording consisted of just two deflections, representing ventricular contractions, which he named A and B. When refinement of the technique exposed additional deflections that represented atrial contractions, he began to use letters from the middle of the alphabet, starting with P. This is supposed to have been inspired by Descartes' naming convention for the points on a curve. When Eindhoven finally produced an ECG waveform that it still recognised today as the gold standard, he renamed all the waves beginning with P. Hence, the naming convention shown in Figure 2.1 [14].

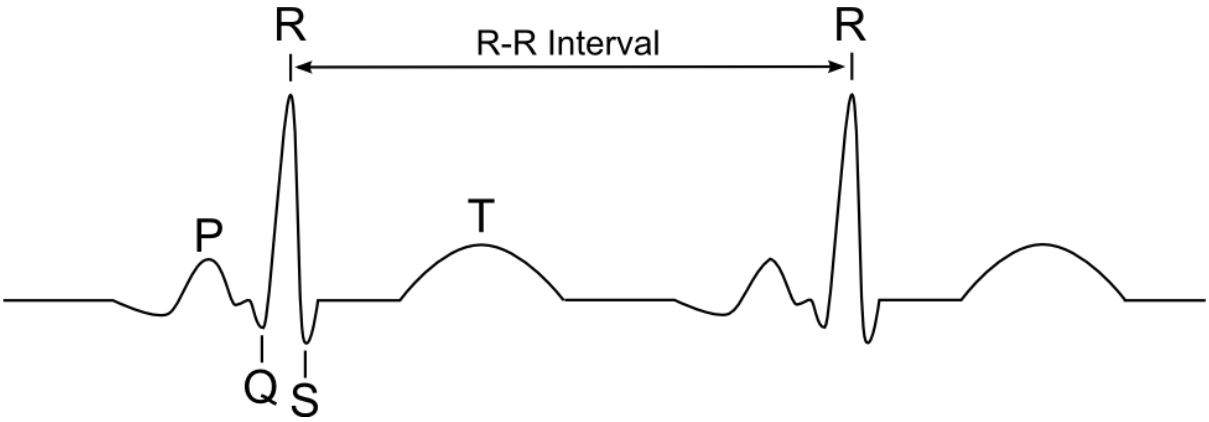


Figure 2.1 - The principal waves of the ECG. (Image from Wikimedia Commons under Creative Commons license, no author permissions required.)

2.3.3 Relationship between ECG waves and heartbeats

As Galvani described in the late 1700s, based on his famous experiments with frogs' legs, muscle contractions can be induced by electrical impulses [15]. This is the basic principle upon which the heart muscle operates [16].

In a healthy heart, an electrical impulse originating from the sino-atrial (SA) node propagates through the atria to the atrio-ventricular (AV) node (see Figure 2.2). This causes the atria, which contain blood returning from venous circulation, to contract and pre-fill the ventricles. This electrical activity manifests as the P wave on an ECG.

After a brief pause, during which the AV node 'holds' the electrical impulse, it is propagated into the ventricles. The ventricles, which are larger and more muscular than the atria, contract and eject blood into arterial circulation. This manifests as the QRS complex on an ECG. Finally, the ventricles repolarise ready for the next heartbeat. This manifests as a T wave.

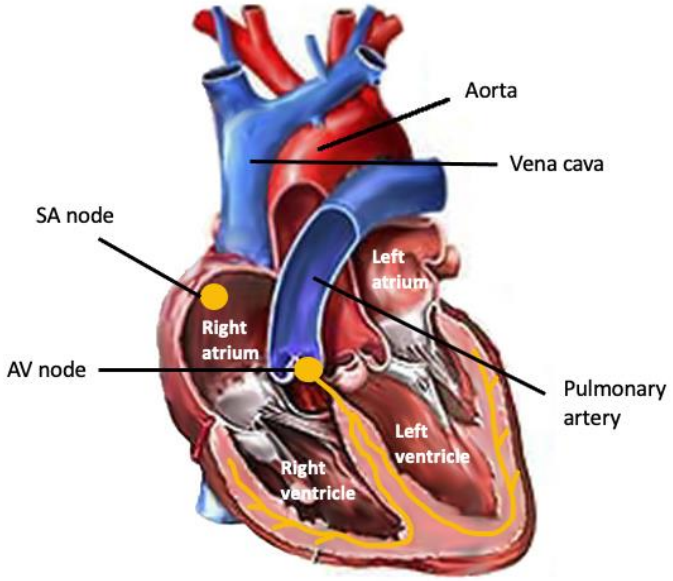


Figure 1.2 - Electrical pathways in the heart. (Image adapted from Wikimedia Commons under Creative Commons license, no author permissions required.)

2.3.4 The ECG in cardiac diseases

The heart's role in maintaining the health of an individual is relatively simple: it acts as a unidirectional pump, maintaining circulating blood pressure. Broadly, there are three types of heart disease: vascular, electrical and mechanical. Vascular heart disease is the leading cause of death globally [17]. Atrial fibrillation (AF), which is a disease of the electrical conduction system of the heart, is a common cause of stroke [18]. Primary mechanical disorders of the heart are relatively less common than "heart failure" caused by vascular and electrical problems, and are often congenital [19]. However, primary mechanical heart disease caused by rheumatic fever remains a serious problem in the developing world [20].

The ECG is the definitive diagnostic test for most primary electrical diseases of the heart. Mechanical and cardiovascular diseases generally require additional testing modalities to diagnose with complete confidence. However, the strong correlation between electrical activity and mechanical / cardiovascular problems of the heart means the ECG is frequently employed as a cheap, accessible, non-invasive first line investigation for all cardiac issues. For this reason, the ECG remains critical to modern cardiology [16].

2.3.5 Types of ECG

There are different types of ECG recording. The most common is the 12 lead ECG. This comprises 10 physical electrodes attached to the patient: one on each limb and six chest (or 'precordial') leads as shown in Figure 2.3. The limb leads are used to infer the "Wilson Central Terminal" (WCT) [21]. The WCT is then used to derive unipolar recordings for each ECG lead, with the exception of the lower right limb (this is the neutral lead).

The resultant 9-lead ECG is sometimes used in clinical practice, but an additional three leads are usually calculated to create the standard 12-lead ECG. This gives a "view" of the electrical activity of the heart seen from 12 different perspectives, which clinicians are expected to mentally reconstruct into a 3D electrical model of the heart to localise regional abnormalities (such as myocardial ischaemia in the region of a single coronary vessel) [16].

A 12-lead ECG is usually recorded over the course of 10 seconds to give a snapshot view of electrical activity of the heart. In cases where intermittent electrical issues are suspected, continuous ECG monitoring is preferable. For practical reasons (to allow patients to move around freely), this usually comprises fewer leads: often three, but sometimes as just one. This results in a very limited ability to detect localised abnormalities, so pauci-lead monitoring is not usually suitable for detecting regional ischaemic events [22].

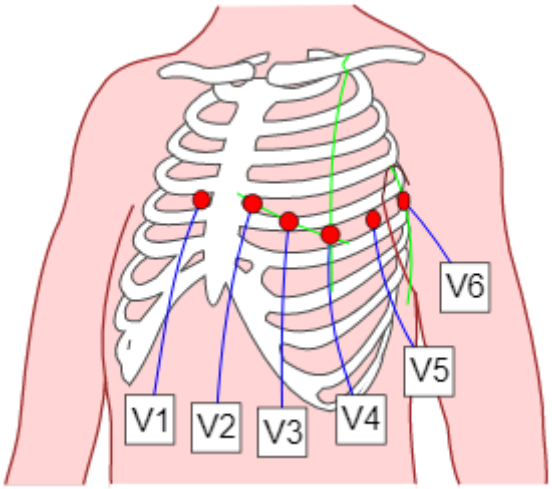


Figure 2.2 - Placement of the precordial ECG lead. (Image from Wikimedia Commons under Creative Commons license, no author permissions required.)

Conversely, body surface potential mapping uses much larger numbers of electrodes to build high resolution 3D maps of electrocardiac activity. This has been an active line of research for several decades but does not play a major role in many clinical pathways at present [23].

2.4 Computerised ECG interpretation

ECG interpretation is notoriously challenging for clinicians, especially non-cardiologists [24]. Attempts to automate ECG analysis using modern computers started over 60 years ago [25]. Many applications emerging from this field have been shown to improve clinician

interpretation of ECGs [26]. However, there remain significant shortcomings with conventional rule-based approaches [27].

A major challenge in automated ECG interpretation is feature extraction. This generally requires pre-processing steps to boost the signal-to-noise ratio (SNR). Noise types in ECG signals include baseline wander, motion artifact, non-cardiac muscle activity, powerline interference and electrode contact noise. A range of filtering techniques is employed to deal with these [28]. Once a reasonably clean signal is obtained, signal processing techniques can be used to detect the key ECG waves and many additional derived features [29].

A detailed review of conventional filtering and signal processing techniques used in ECG processing is beyond the scope of this discussion. From an AI practitioner's perspective, the salient property of these rule-based (i.e. non-machine learned) approaches is that they fall consistently short of the ability of a human expert to discern key ECG features and arrive at relevant diagnoses [27]. This is proving not to be the case for ML approaches, particularly those based on DL [30].

A possible explanation for this phenomenon is that rule-based approaches can only use logic that is fully expounded by human experts (via computer programming languages). For a cardiologist, on the other hand, ECG feature extraction is a question of detecting patterns from complex visual input. This type of task is known to employ subconscious pathways. Therefore, even if most humans are extremely adept at making sense of visual stimuli, they cannot clearly articulate how they do it [31].

Conversely, ML algorithms learn by exposure rather than explicit instruction. They can capture logic processes that defy clear explanation [32]. Hence, ML-based AI is driving breakthroughs in a range of tasks that involve medical image processing, and also in the domain of ECG interpretation [33, 34].

2.5 AI basics

There is no widely accepted definition for AI, in part because there is no universally agreed definition for 'natural' intelligence [35]. At the time of writing, AI is commonly used to refer to applications that are based on ML principles [36]. It will be used in this context hereafter.

ML algorithms are statistical models whose parameterisation is determined by automated trial-and-error searching rather than classical statistical methods [36]. These parameter searches are not generally performed using 'brute force' methods, which would render optimal solutions computationally intractable. Rather, each update is targeted using an optimisation algorithm, generally based on the principle of gradient descent [37].

At a high level, gradient descent involves a guided 'walk' through the model's parameter space. At each step on this walk, the model predicts an output for a given input. A loss is calculated

for the prediction using a loss function (sometimes called a cost function). The loss function must be differentiable, because its gradient is used to determine both the direction and magnitude of the next step in the walk. The update function for the model parameters will endeavour to take the model 'downhill' in parameter space, meaning that the model parameters are updated in a direction that aims to reduce the loss at each step [32]. See Figure 2.4 for a visual illustration.

2.6 Machine versus deep learning

There are many types of ML algorithm, ranging from relatively simple logistic regression models to multi-trillion parameter 'transformers'. Over the course of the 21st century, a subtype of ML algorithm based on the concept of 'artificial neural networks' (ANNs) has emerged as particularly potent. Modern ANNs generally comprise many sequential layers, giving them the appearance of depth in visual depictions of network architecture. Hence, this type of AI application has become known as 'deep learning' [38]. DL tends to have a substantially diminished reliance on manually engineered features than other ML approaches. Instead, DL is often employed in an 'end-to-end' fashion. whereby raw data is mapped directly to a desired endpoint [32]. This has significant pros and cons, which will be discussed later.

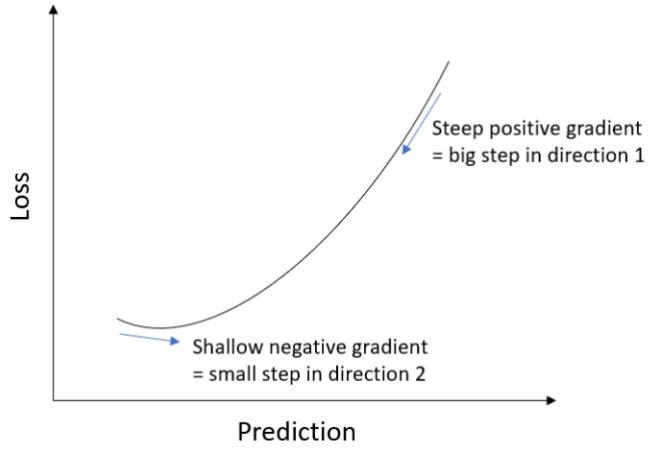


Figure 2.3 - parameter updates guided by gradient descent

The discrepancy between the performance of modern DL applications versus any other type of ML for suitable (usually particularly challenging) tasks is so profound that they are frequently referred to as entirely separate entities. Thus, the use of the term ML in modern literature often implies that the application in question does *not* leverage DL [39].

2.7 The ImageNet milestone and the rise of CNNs

Though DL is viewed as a modern phenomenon, its fundamental building block – the artificial neuron – was proposed in the 1940s by McCulloch and Pitts [40]. Over the following decades, optimism about the potential of ANNs waxed and waned, creating periods that are sometimes

referred to as ‘AI winters’ and ‘AI springs’ [41]. The breakthrough that is often cited as cementing DL’s current position at the cutting edge of modern AI occurred during the 2012 ‘ImageNet’ annual computer vision competition [42].

ImageNet was designed to be a gold standard benchmark for detecting highly abstract semantic features from photographic images: i.e. a bird or a bicycle, as opposed to rudimentary geometric shapes like squares and triangles. Prior to the 2012 event, human-level performance was widely considered beyond the capabilities of any computer software; although it was clear that attaining this level of performance would open the door to breakthroughs such as autonomous vehicles, image searching, etc. [43].

In the 2012 competition, Krizhevsky et al. used a particular variant of ANN, called a deep convolutional neural network (CNN), to surpass the previous state of the art by a compelling margin and approach human accuracy [44]. The CNN architecture they employed was first proposed in a 1998 paper by LeCun et al., who demonstrated that this paradigm was more efficient than the conventional multi-layer perceptron (MLP) ANN for recognising hand-written characters (see Figure 2.5) [45]. However, Krizhevsky et al.’s model was an exceptionally deep implementation of this type of AI model.

Given the speed of processing chips at the time, the size of the model should have made it impractically slow to train. However, the team solved this issue by using a graphics processing unit (GPU) rather than a conventional central processing unit (CPU) to run the training [44]. The key difference is that GPUs leverage highly parallel computation, whereas CPUs perform operations serially. The mathematics of DL revolve heavily around matrix operations, which are highly parallelisable. Hence, parallel computing can affect very substantial performance gains with DL applications. Since 2012, other parallel processors that focus specifically on ML have been developed, and they are collectively known as ‘AI accelerators’. However, GPUs remain dominant in this market and have played a key role in driving the current ‘AI Spring’ [46].

The ImageNet moment marked the start of a new era in DL: the combination of more efficient ANN architectures and parallel computing allowed AI models to scale to unprecedented size and complexity. Other types of DL algorithm have become popular since 2012, but the CNN has played a key role in ECG analysis and will be discussed in more detail below.

2.8 How CNNs work

In a conventional MLP, each 'activation unit' (AU) within a layer of the network is connected to all the AUs in the preceding and proceeding layers by trainable weights, as shown in Figure 2.5. The activation units comprise nonlinear functions. The 'tanh' function was once popular, but it has been shown that the more computational efficient rectified linear activation unit (ReLU) is equally effective for most tasks [47]. The nonlinearity between linear layer weights allows the network to model complex logical operations. Stacking these 'densely connected' layers can create powerful AI models [32].

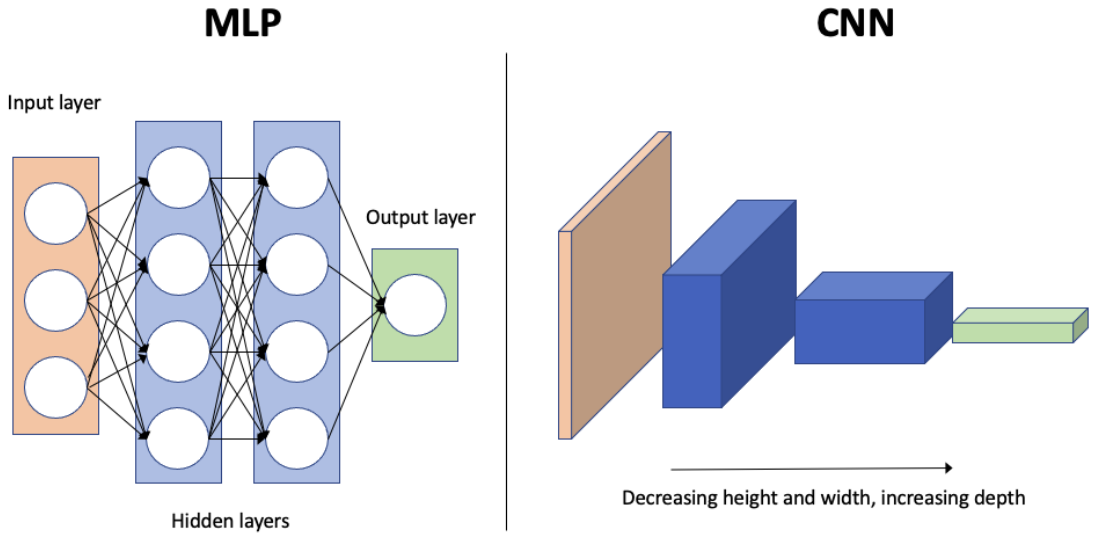


Figure 2.4 - the MLP architecture compared with a CNN

A major downside to MLP's is the computational intensity. Ignoring the layer bias, which is a single parameter applied uniformly to all the weights in a given layer of the network, the number of trainable weights for each additional layer L is D_L x D_{L-1}, where D is the dimension, or number of AUs, in the layer. Each time the MLP makes a prediction during the training process (known as a 'forward pass', given the flow of data through the network is unidirectional), the relative contribution of each layer weight to the overall error is calculated by taking partial derivatives of the loss with respect to each weight. The weights are then updated, using the principles of gradient descent, during 'backpropagation'. Computational requirements of forward- and backpropagation increase exponentially with additional trainable layers [48].

By contrast, a CNN employs layer weights more sparsely, but more efficiently, through the use of trainable convolutional filters. These filters comprise numerical matrices that are multiplied with an input matrix to detect local patterns within the input data. Filters are passed over the input data using a sliding window approach. Each filter looks for a single pattern, but multiple "stacked" filters are generally used at each layer of the CNN to detect more complex, composite patterns.

Figure 2.6 gives a simple but hopefully intuitive example of how two filters, each detecting a straight line in the vertical or horizontal plane, can be used to detect a more complex feature (a

square) in a black and white image. By learning different filter values within a multi-layer network using gradient descent-based approaches, modern CNNs can detect extremely

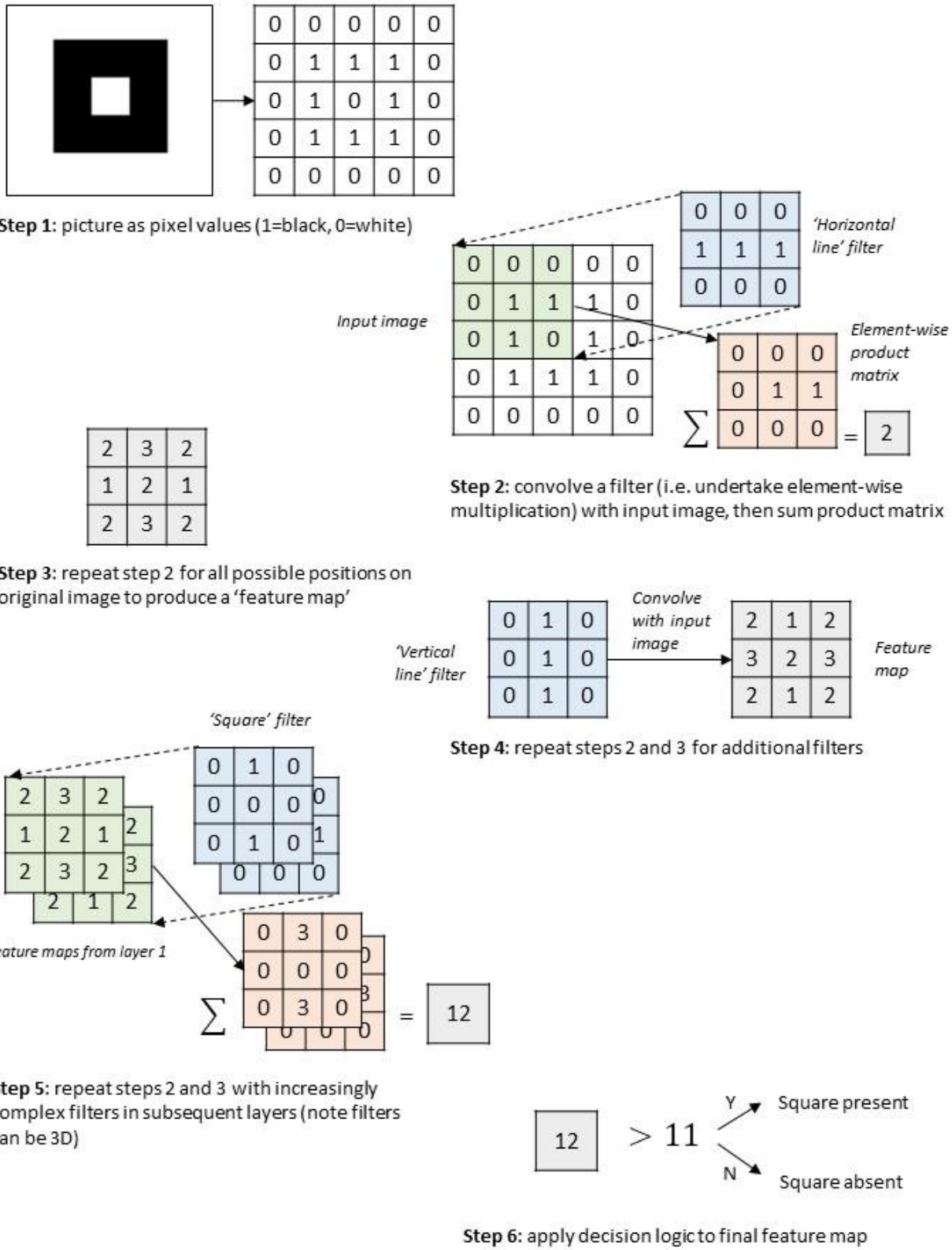


Figure 2.5 - a basic image analysis workflow using convolutional filters. In the case of a CNN, filter values would be learned during the training process.

complex patterns within high resolution data [49].

It is important to note that, while imaging tasks provide the archetypal use-case for a CNN, many other data types (including single- or higher-dimensional data) can be used at input for CNNs. A significant limitation of CNNs is the poor ability to detect patterns that are defined by a relationship between distant elements of the input data, as each filter can only detect patterns that occur within its receptive field. Techniques such as residual layer connections have helped to ameliorate this limitation. But the fact remains that the switch from a MLP to a CNN is ultimately a trade-off between the ability to model more complex and distant relationships within the input data versus a substantial increase in computational efficiency [50].

2.9 Progress over the last decade

Since 2012, there have been substantial advances in DL practice. As noted above, residual layer connections, in addition to several other architectural improvements, have continued to drive progress in CNNs. For much of the 2010s, recurrent neural networks (RNNs), along with long-short-term memory (LSTM) networks, showed promise in sequence processing. Natural language processing (NLP) is among the chief applications within this field [51]. Since a seminal paper from scientists at Google in 2018, however, RNNs and LSTMs have largely been superseded by 'transformer' networks, which leverage multi-headed attention (MHA) [52, 53]. At the time of writing (early 2022), transformers are also beginning to achieve state-of-the-art performance in computer vision.

2.9.1 Transformers and the implications for the AI ecosystem

The largest transformer models in production run to over a trillion parameters [54, 55]. There is currently a consensus that 'bigger is better', as long as you have enough data to feed larger models [56]. However, models in excess of a billion parameters usually require very specific infrastructure to train. They do not fit within the memory of a single GPU and must be 'sharded' across multiple GPUs to train using an approach known as 'model parallelism'. Very large models are sharded across multiple GPUs within multiple compute nodes in a cluster.

The GPUs within a model parallel training pipeline must function effectively as a single compute unit, which requires very high bandwidth GPU-GPU communication. Conventional high-performance computing (HPC) facilities that are equipped with GPUs generally have the accelerators connected via peripheral component interface (PCI) ports. To share data within a server (or 'node'), GPUs must communicate via the PCI bus, which is relatively slow. To share data between GPUs spread across multiple nodes, GPUs must communicate via both the PCI bus and ethernet switches, which is slower still [57].

Within HPC facilities that cater specifically for training large AI models, each node tends to have a high GPU-to-CPU ratio. GPUs within each node are networked directly with high speed, multi-lane connections. Nodes are connected with ultra-fast networks that allow remote direct memory access (RDMA) between GPUs [58]. This architecture looks very different from traditional HPC, which can cause challenges for information technology (IT) teams. It also tends to be very expensive. This risks creating a barrier to the democratisation of modern AI [59]. Approaches to partially address this challenge will be discussed in subsequent chapters on experimental work.

2.9.2 Graph neural networks

Graph neural networks (GNNs) have many design features in common with transformers and are becoming popular in key domains such as molecular simulation [60]. GNNs can also leverage the principles of convolution (graph convolutional neural networks, or GCNNs), and can be architected in such a way that they incorporate the principles of Newtonian or even quantum physics (physics informed neural networks, or PINNs) [61].

The field is evolving quickly. At the time that the work presented in the following chapters was commenced, CNNs were emerging as a promising approach in the domain of AI-enabled ECG processing [62]. Therefore, newer techniques such as transformers are not explored in the original research works. However, they will be further discussed in the final chapter.

2.10 DL for ECG analysis and areas for further investigation

The potential value of ANNs in ECG analysis has been discussed for over two decades [63]. Prior to 2017, the focus of work in this area was almost exclusively on ‘shallow’ ANNs, with no comprehensive evaluation of DL in this area [30]. From 2017 to 2019, there were a number of works published in this field, summarised in a systematic review by Ebrahimi and Zahra in 2020 [64]. DL for AF detection saw a particularly marked surge in publications during 2018 [65-77]. This may have been partly driven by the 2017 Computing in Cardiology challenge, which made a large dataset of single-lead ECG recordings available and tasked participants with developing automated AF detectors [78]. The substantial majority of the studies on DL for AF detection published in 2018 employed CNNs [65-68, 70-73, 75-77].

However, two seminal papers in 2019 claimed to the push the bounds of what was previously considered possible in the field of automated AF detection. The first was an article by Hannun et al. claiming ‘cardiologist-level’ detection of AF from ambulatory ECG signals [30]. The second was a paper by Attia et al. claiming that their algorithm was able to diagnose incipient AF from ECGs showing normal sinus rhythm (NSR) [79]. This latter claim was particularly noteworthy because this is not something that is general considered possible for human cardiologists; thus, the paper was effectively claiming super-human performance.

2.10.1 Area for investigation 1: DL for ischaemia detection

Following the surge in work on DL for arrhythmia detection, and particularly in light of the notable breakthroughs from Hannun and Attia's groups, it appeared likely that progress in the domain of DL arrhythmia detection would continue. However, the ECG is also an essential first-line investigation for the detection of AMI (or 'heart attack' in lay terms). In this setting, ECG findings are instrumental in diagnosing the underlying aetiology of the AMI and determining whether a patient will benefit from emergency endovascular surgery known as primary percutaneous coronary intervention (PPCI). If there is ECG evidence of a transmural ('full thickness') AMI, generally caused by total or near-total thromboembolic occlusion of a major coronary artery, primary (emergency) PPCI is potentially lifesaving. If the ECG is mis-diagnosed and the patient misses out on this procedure, the consequences can be catastrophic [80]. As will be discussed in the next chapter, which presents a systematic literature review of works on DL for ischaemia detection, the application of DL to this highly impactful problem had been poorly investigated at the time this PhD project was commenced. Therefore, the following research question was identified as the focus for chapter 3:

1. Will AI-enabled ECG analysis allow us to detect hyperacute myocardial ischaemia? (Or, in lay terms: can we detect heart attacks earlier using AI?)

2.10.2 Area for investigation 2: Democratising ECG AI

The second part of this project is addressed to the topic of how the impact DL-based ECG analysis can be maximised from a patient perspective. In 2018, prominent figures from the National Institutes of Health, the Radiological Society of North America and the American College of Radiology published a position paper on translational research in medical AI. They highlighted the mismatch between progress being made at the foundational research level and progress at the clinical level. They called for more translational research to address barriers to widespread implementation of AI at the point of care [81].

As noted at the outset of this chapter, the primary motivation for the work presented in this thesis is to examine the potential of AI to improve patient care. ECG analysis is proposed as an appropriate lens through which to undertake this investigation, but with the intention of relating the findings of the ECG-specific research chapters back to broader challenges within the world of medical AI in the concluding chapter. Thus, in addition to investigating the potential utility of DL for ischaemia detection in a controlled research setting, it was also felt to be important to investigate research questions whose primary focus is translational. On this basis, two key translational research questions were selected to be the focus of chapters 4 and 5:

1. Can DL improve the performance ECG image analysis applications, where non-AI methods have historically faced significant challenges?

The rationale for investigating this topic is that in many centres, ECG data is still stored as paper printouts rather than digital files containing the raw sample readings. Most computerised ECG analysers rely on raw samples, which limits the extent to which emerging, state-of-the-art methods can be applied to retrospective medical records. Historical medical records are often associated with rich multi-modal data and clinical outcomes, which makes them highly valuable for developing and testing prognostic models, risk scores, etc. [82]. The development of state-of-the-art ECG image analysers could open the door to retrospective observational studies that could fuel early stage research, particularly in the nascent field of multi-modal clinical AI [83].

2. How can DL be 'democratised' to better cater for population groups and disease cohorts where there is a relatively paucity of labelled data?

This is a topic with broad application in the field of clinical AI, as DL methods tend to rely on large volumes of labelled data relative to other approaches [32]. This can limit the extent to which DL-based applications benefit groups who are under-represented in the digital world.

2.11 Introduction to the systematic literature review of AI for ischaemia detection

As noted above, early detection of AMI is a central problem of modern cardiology. The advent of PPCI has transformed the prognosis of patients suffering AMI due to acute thromboembolic occlusion of the coronary arteries [84]. However, PPCI is a time-critical intervention and a delay in diagnosis can be catastrophic [85]. The delineation of patients who are and are not likely to benefit from PPCI must be based on clinical data that is available at the time of presentation: most importantly, the presence of ischaemic chest pain and ST segment elevation (STE) on the ECG [86].

Using DL to obtain automated, expert-level ischaemia detection from ECG signals would be highly desirable, given that the missed case rate for ST elevation myocardial infarction (STEMI) in emergency departments is over 10% [87]. Current, rule-based automated ECG analysis performs poorly compared with experts [88].

The aim of chapter 2 was to give some relevant background on the key concepts underlying both ECG analysis and modern, DL-based AI. A narrative review of DL for ECG analysis was presented, which noted that the majority of published works at the time this thesis was commenced related to arrythmia detection. It was felt that relatively little work had been undertaken in the field of DL for ischaemia detection. The aim of this systematic literature review and meta-analysis is to take a more rigorous approach to establishing what work has

been done in the field to date, and to evaluate the likelihood that DL technology can improve upon existing methods.

2.12 Methods

2.12.1 Search strategy

The literature search was conducted according to the PRISMA framework [89]. The titles, abstracts and keywords of full-text articles on Medline, Scopus and Web-of-Science were searched using the following terms: ((myocardial infarction OR ischaemia) AND (neural network OR deep learning) AND (electrocardiogram OR ECG)). The searches were performed in November 2019.

2.12.2 Study selection

All research articles documenting the analysis of ECG signals via ANNs to detect myocardial ischaemia were included. Articles not pertaining to ANNs, or pertaining to ECG analysis for other purposes (e.g. arrhythmia detection, hyperkalaemia detection), were excluded.

Search terms were applied to full texts, but screening was conducted based upon titles and abstracts. Studies selected for inclusion were obtained in full for manual analysis.

2.12.3 Quality evaluation

The QUADAS 2 framework was used to assess the quality of the studies [90]. The outcomes of quality evaluation were noted but no study was excluded on the grounds of poor quality as this is a nascent field and all completed work was felt to be relevant for the purposes of this review.

2.12.4 Data extraction

Data was extracted according to a proforma developed during analysis of the first two studies and refined during the analysis of the next two. (See appendix.)

2.13 Results

2.13.1 Study selection

The search terms generated 36 results from Medline, 33 from Scopus and 127 from the Web of Science. Of these, 46 studies were deemed suitable for inclusion in this review [91-99]. Study characteristics are noted in Table 2.1.

2.13.2 Study quality

As assessed using the QUADAS 2 framework, there was high variability in the quality of the studies. In particular, patient selection was poorly defined in many cases (often because the studies were conducted using publicly available, anonymised ECG databases). Four of the studies took cardiologists' interpretation of ECG data (either ambulatory or exercise) as the reference standard. With the exception of Xiao et al., results of these studies were framed as "ischaemia detection", which makes the assumption that cardiologists' interpretation of ambulatory or exercise ECG data in isolation of other clinical variables correlates well with underlying myocardial ischaemia. One study (Neagoe et al.) did not explicitly define their reference standard, though they used an annotated ST-T database so the assumption is made that ST elevation annotated by domain experts was their target endpoint. Details regarding quality evaluation are given in Table 2.2.

2.13.3 Technology used

Only two studies used DL technology, defined as ANN architectures with multiple hidden layers or a combination of multiple neural networks. All other studies used ANNs consisting of an input layer, a single hidden layer and an output layer. In the majority of the studies, features were extracted using either rule based methods or dimensionality reduction with principal component analysis (PCA). Details in Table 2.1.

2.13.4 Heterogeneity and statistical analysis

There was significant variation in the approach to data acquisition, reference standards and statistical reporting of results. Only two studies prospectively collected patient data. Four studies used a definition of myocardial ischaemia that employed clinical variables independent of cardiologist-annotated ECG changes. One study used multi-modal clinical data (including ECG signals) as input to the ANN, whereas the other eight only used ECG signals. Seven studies reported sensitivity and specificity of the ANNs with respect to the reference standard. One only reported area under the receiver operator characteristic curve (AUROC). Another reported sensitivity and positive predictive value but not specificity. See Table 2.3 for details.

2.13.5 Data synthesis

Results of data synthesis are shown in Table 2.4. The salient result is that ANNs appear to perform better in detecting cardiologist-annotated ST segment elevation than in detecting more objectively defined myocardial ischaemia.

Table 2.1 Study characteristics

Study	No. subjects	M:F (%)	Age (mean)	Positive cases (%)	Location of data collection	ECG type	Evaluation approach	Technology used	Reference standard
Xue et al. 2004	Trained on 1358, evaluated on 1902	N/S	N/S	N/S	Rochester, MN, USA	12 lead	Prospective	Combination of rule-based model and 3 layer ANN to enhance clinician diagnosis (NB: ANN analyses were not evaluated independently of clinician analysis)	Composite definition AMI as per 1984 WHO criteria (below)
Dehnavi et al. 2011	70	53	55	86	Khomeinishahr, Iran	12 lead (exercise)	Retrospective (cross validation)	Feature extraction by PCA, with outcomes fed to a 3 layer ANN. Process repeated with vectorcardiograms (VCGs) to compare results.	Comparison with cardiologist evaluation of ischaemic changes within ECGs
Paploukas et al. 2002	Not clearly stated	N/S	N/S	N/S	Various in USA / Europe	2 lead (ambulatory)	Retrospective (hold out)	Rule-based analysis to detect the J point of each beat, then following 400mS fed to a 3 layer ANN	Comparison with cardiologist evaluation of ischaemic changes within ECGs
Neagoe et al. 2003	Trained on 20, evaluated on 20	N/S	N/S	N/S	Various in USA / Europe	Single lead (V5 only, ambulatory)	Retrospective (hold out)	Best model: PCA followed by a "fuzzy" ANN (4 layers)	ST elevation as annotated by human experts

Forberg et al. 2012	Trained on 3000, evaluated on 560	5500%	70	8	Lund, Sweden	12 lead	Prospective	Composite model of 25 ANNs (each 3 layers)	1) Evaluation of ST elevation against international criteria 2) Evaluation of need for PPCI against final outcome (patient had PPCI or not)
Baxt et al. 2002	2204	54	56	16	Philadelphia, PA, USA	12 lead	Retrospective (cross validation)	3 layer ANN with ECG *plus* multiple clinical features as input (including biomarkers where available)	Composite definition of AMI *plus* at least one 70% coronary lesion. For unstable angina: clinical history *plus* 70% coronary lesion or functional evidence of ischaemia or elevated creatinine kinase
Sbrollini et al. 2019	Trained on 241, evaluated on 241	N/S	N/S	17	Charleston, WV, USA and Lund, Sweden	Pairs of 12 lead ECGs, where positive examples contained one non-ischaemic and one ischaemic ECG	Retrospective (hold out)	3 layer ANN whose input features are derived from rule-based ECG analysis and measurement of differences between paired ECGs	Myocardial ischaemia assumed after 3 mins of complete balloon occlusion of a coronary artery
Xiao et al. 2018	Trained on 20, evaluated on 15	N/S	N/S	47	Various in USA / Europe	2 and 3 lead (ambulatory)	Retrospective (hold out)	Inception V3 (48 layer convolutional neural network with residual layer connections)	ST elevation as annotated by human experts

Maglaveras et al. 1998	Not clearly stated (90 records, unclear if from unique patients)	N/S	N/S	N/S	Various in USA / Europe	7 lead (ambulatory)	Retrospective	Rule-based analysis to detect the ST segment of each beat, then following 160mS fed to a 3 layer ANN	ST elevation as annotated by human experts

Abbreviations: M:F = male:female, N/S = not stated

Table 2.2 Quality evaluation and additional notes

Study	Quality notes	Additional notes
Xue et al. 2004	No demographic details available re subjects. Cannot evaluate bias. The composite AMI endpoint (including biomarkers, ECG analysis and other clinical variables) was felt to be a strength of the study.	3 layer ANN is not truly "deep" learning. The fact that computerised ECG analysis was only evaluated when used as an adjunct to clinician analysis makes an evaluation of the technology itself impossible.
Dehnavi et al. 2011	The assumption that cardiologists' analysis of exercise ECGs is a reliable indicator of underlying ischaemia is a major weakness of this study.	3 layer ANN is not truly "deep" learning. Due to the weakness of the endpoint, results of this study should be interpreted with caution.
Paploukas et al. 2002	The assumption that cardiologists' analysis of ambulatory, 3 lead ECGs is a reliable indicator of underlying ischaemia is a weakness of this study.	3 layer ANN is not truly "deep" learning. Limiting the ANN to the segment of ECG signal immediately following the J point potentially missed important ischaemia indicators in other parts of the ECG. Due to the weakness of the endpoint, results of this study should be interpreted with caution.
Neagoe et al. 2003	The failure to define how the "ischaemic" ECGs were acquired and diagnosed makes it impossible to evaluate the quality of this study.	The ANN was only exposed to the QRST segment of a single ECG lead, presumably to approximate ambulatory single lead recordings. However, without further information regarding the study design, the results of this study should be interpreted with extreme caution.
Forberg et al. 2012	This appears to be a robust study according to analysis using the QUADAS-2 framework.	Interestingly, the ANN identified 7 patients who met international ST elevation criteria on subsequent manual analysis, but who were not identified as having a STEMI by the on call cardiologist. The ANN thus outperformed the cardiologist, but in the event, none of these patients required PCI.

Baxt et al. 2002	This appears to be a robust study according to analysis using the QUADAS-2 framework.	This study employs a particularly strong end point, and the idea of a mixed-modality input vector for an ischaemic detection model is likely to be explored further by future research. However, such a model would be less useful for the rapid discrimination of patients who would and would not benefit from PPCI. Furthermore, this 3 layer ANN has been superseded by more sophisticated models in recent years and better results may be obtained by revisiting the data with new technology.
Sbrollini et al. 2019	Unreported demographic data precludes evaluation of potential bias in this study.	For a modern study, the technology employed was not particularly sophisticated. The authors refer to a "deep learning" approach, but with a 3 layer ANN this is questionable. There is no benchmark included for the complexity of the classification task on this particular dataset - comparison with human experts may add weight to the findings.
Xiao et al. 2018	Unreported demographic data precludes evaluation of potential bias in this study. Though the reference standard is cardiologists' interpretation of ST changes, this is not overtly assumed to denote underlying myocardial ischaemia.	This study uses a state-of-the-art deep learning algorithm and employs transfer learning to improve the efficiency of the training process, thus demonstrates a promising approach to deep learning based ischaemia detection. However, small study numbers, the absence of a composite endpoint and a poorly defined cohort limit its application at this stage.
Maglaveras et al. 1998	Unreported demographic data precludes evaluation of potential bias in this study. The assumption that cardiologists' analysis of ambulatory, 3 lead ECGs is a reliable indicator of underlying ischaemia is a weakness of this study.	This is the earliest example of ANN-based analysis of ECG signals for ischaemia detection within this review.

Table 2.3 Results of studies

Study	Results
Xue et al. 2004	Cardiologist reader: sensitivity increase from 29% to 37% with CAD. ED physician reader: sensitivity increase from 29% to 44% with CAD. Specificity unchanged at 99% for both.
Dehnavi et al. 2011	ECGs: sensitivity 60%, specificity 70% VCGs: sensitivity 70%, specificity 86%
Paploukas et al. 2002	Best model: sensitivity 91%, specificity 90%
Neagoe et al. 2003	Best model: sensitivity 100%, specificity 100%
Forberg et al. 2012	1) Evaluation of ST elevation against international criteria: sensitivity 95%, specificity 68% 2) Evaluation of need for PPCI against final outcome (patient had PPCI or not): sensitivity 97%, specificity 68%
Baxt et al. 2002	Patients presenting with chest pain either diagnosed as being ischaemic or non-cardiac: Sensitivity 80.9%, specificity 81.3%
Sbrollini et al. 2019	AUROC 0.83 Further breakdown of results not reported.
Xiao et al. 2018	Sensitivity 84.4%, specificity 84.9%, AUROC 0.89
Maglaveras et al. 1998	Sensitivity 91%, positive predictive value 83% (specificity not reported)

Table 2.4 Synthesis of data

	MI defined by features independent of ECG changes (n=4)	**MI defined by cardiologist annotation of ECG (n=5)**
	Mean +/- SD	Mean +/- SD
Sensitivity	64.7 +/- 28.9	88.1 +/- 11.6
Specificity	86.8 +/- 15.1	85.8 +/- 11.6

2.14 Discussion

The evidence base in this field remained small at the time of this systematic review, but contained some useful insights. Most importantly, the results reported by these studies suggested that DL based analysis of ECG signals is a viable approach to the detection of myocardial ischaemia. The study by Xue et al. was particularly noteworthy for having demonstrated an improvement in clinician analysis of ECGs when supported by an ANN based system.

The rapidly evolving nature of ML technology represents a challenge in this domain. The 2002 study by Baxt et al. benefits from robust design and a rich dataset. However, the ANN used to predict myocardial ischaemia from multi-modal input features is what is now referred to in the data science field as a "vanilla" network, denoting a simplistic architecture and shallow structure. Revisiting this dataset with a modern DL model may produce improved results.

The study by Sbrollini et al. explores the hypothesis that an ML model could learn the idiosyncratic morphology of a patient's ECG and use this to predict ischaemic changes unique to that individual. In fact, the scope of their study includes predicting altered cardiac status of heart failure patients using the same approach, and their results are reported with respect to both ischaemia and heart failure rather than each separately. Drawing firm conclusions regarding ischaemia detection in isolation is therefore difficult, but it provides some rationale for further work in this area. Coupled with evidence that reliable multi-lead ECG signals can be captured from wearable devices, this could form the basis of a method for hyperacute detection of myocardial ischaemia in the ambulatory setting [100].

Finally, five of these nine studies focussed exclusively on ST segment changes as the ECG hallmark of myocardial ischaemia. This may represent a common approach of clinical researchers towards working with ML tools: namely, restricting the input of ML models to data that has a proven causal link with the desired endpoint. In some instances, this is done intentionally to force a model to approximate the established human approach to a given task, which may be necessary for an AI tool to function within a wider clinical context [101]. When done by default, however, curating the input to a ML model can preclude the possibility of discovering novel features that can be used to improve upon the existing approach and even, in some instances, surpass expert human performance [102].

Sbrollini et al. trained their ANN on uncurated ECG signals and labelled each segment according to presence or absence of balloon occlusion of a coronary artery at the time of signal acquisition, rather than relying on expert analysis of the ECGs or a composite endpoint of which expert ECG analysis forms a significant part. Though one may question whether balloon occlusion of a vessel inevitably results in myocardial ischaemia (e.g. in the presence of

extensive collateralisation or non-viable myocardium), the strength of this approach lies in the fact that the ML model is not constrained by human curation of the input data nor reliant on human analysis for the reference standard. It is thus conceivable that a sophisticated ANN, by virtue of an enormous capacity for discerning subtle patterns within complex high-resolution data, may leverage previously unrecognised ECG features of myocardial ischaemia when faced with this task. Sbrollini et al. do not report any attempt to interrogate the internal logic of their ANN, nor to benchmark its performance against human experts faced with the same dataset, but this was felt to represent a promising line of enquiry, and informs the design of study presented in a subsequent chapter.

The heterogeneity of approaches and the questionable quality of some of the studies as evaluated by the QUADAS 2 framework should prompt significant caution in interpreting synthesised data from this review.

2.15 Conclusion

The conclusion drawn from the systematic review and meta-analysis presented in this chapter is that DL-based AMI detection from ECG signals is a field that both needs and warrants further work. The next chapter presents an original research study undertaken in this field.

2.16 References

[1] Edkins GD. A review of the benefits of aviation human factors training. Human factors and Aerospace safety. 2002;2:201-16.

[2] Hogan H, Healey F, Neale G, et al. Preventable deaths due to problems in care in english acute hospitals: A retrospective case record review study. BMJ Qual Saf. 2012;21:737-45.

[3] Amsterdam EA, Aman E. The patient with chest pain: Low risk, high stakes. JAMA internal medicine. 2014;174:553-4.

[4] Waterman AD, Garbutt J, Hazel E, et al. The emotional impact of medical errors on practicing physicians in the united states and canada. The Joint Commission Journal on Quality and Patient Safety. 2007;33:467-76.

[5] Bates DW, Cohen M, Leape LL, et al. Reducing the frequency of errors in medicine using information technology. Journal of the American Medical Informatics Association. 2001;8:299-308.

[6] Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nature Reviews Cancer. 2018;18:500-10.

[7] Kamath U, Liu J, Whitaker J. Deep learning for NLP and speech recognition. Springer; 2019.

[8] Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: Review, opportunities and challenges. Briefings in bioinformatics. 2018;19:1236-46.

[9] Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. The Lancet. 2021;397:1254.

[10] Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25:44-56.

[11] Southern WN, Arnsten JH. The effect of erroneous computer interpretation of ECGs on resident decision making. Medical Decision Making. 2009;29:372-6.

[12] Barold SS. Willem einthoven and the birth of clinical electrocardiography a hundred years ago. Cardiac electrophysiology review. 2003;7:99-104.

[13] Rivera-Ruiz M, Cajavilca C, Varon J. Einthoven's string galvanometer: The first electrocardiograph. Tex Heart Inst J. 2008;35:174-8.

[14] Hurst JW. Naming of the waves in the ECG, with a brief account of their genesis. Circulation. 1998;98:1937-42.

[15] Piccolino M. Luigi galvani and animal electricity: Two centuries after the foundation of electrophysiology. Trends Neurosci. 1997;20:443-8.

[16] Macfarlane PW, Van Oosterom A, Pahlm O, et al. Comprehensive electrocardiology. Springer Science & Business Media; 2010.

[17] **The top 10 causes of death**. 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[18] Mahajan R, Perera T, Elliott AD, et al. Subclinical device-detected atrial fibrillation and stroke risk: A systematic review and meta-analysis. Eur Heart J. 2018;39:1407-15.

[19] Cowie M, Mosterd A, Wood D, et al. The epidemiology of heart failure. Eur Heart J. 1997;18:208-25.

20] Stollerman GH. Rheumatic fever. Lancet. 1997;349:935-42.

[21] Bacharova L, Selvester RH, Engblom H, et al. Where is the central terminal located?: In search of understanding the use of the wilson central terminal for production of 9 of the standard 12 electrocardiogram leads. J Electrocardiol. 2005;38:119-27.

[22] Stern S, Tzivoni D, Stern Z. Diagnostic accuracy of ambulatory ECG monitoring in ischemic heart disease. Circulation. 1975;52:1045-9.

[23] Giffard-Roisin S, Jackson T, Fovargue L, et al. Noninvasive personalization of a cardiac electrophysiology model from body surface potential mapping. IEEE Transactions on Biomedical Engineering. 2016;64:2206-18.

[24] Salerno SM, Alguire PC, Waxman HS. Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence. Ann Intern Med. 2003;138:751-60.

[25] Macfarlane PW, Kennedy J. Automated ECG Interpretation—A brief history from high expectations to deepest networks. Hearts. 2021;2:433-48.

[26] Gregg RE, Deluca DC, Chien CS, et al. Automated serial ECG comparison improves computerized interpretation of 12-lead ECG. J Electrocardiol. 2012;45:561-5.

[27] Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: Benefits and limitations. J Am Coll Cardiol. 2017;70:1183-92.

28] Luo S, Johnston P. A review of electrocardiogram filtering. J Electrocardiol. 2010;43:486-96.

[29] Serhani MA, T El Kassabi H, Ismail H, et al. ECG monitoring systems: Review, architecture, processes, and key challenges. Sensors. 2020;20:1796.

[30] Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25:65-9.

[31] Hoshiyama M, Kakigi R, Watanabe S, et al. Brain responses for the subconscious recognition of faces. Neurosci Res. 2003;46:435-42.

[32] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

[33] Ting DS, Liu Y, Burlina P, et al. AI for medical imaging goes deep. Nat Med. 2018;24:539-40.

[34] Kashou AH, May AM, Noseworthy PA. Artificial intelligence-enabled ECG: A modern lens on an old technology. Curr Cardiol Rep. 2020;22:1-8.

[35] Legg S, Hutter M. A collection of definitions of intelligence. Frontiers in Artificial Intelligence and applications. 2007;157:17.

[36] Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care? J Arthroplasty. 2018;33:2358-61.

[37] Andrychowicz M, Denil M, Gomez S, et al. In: Learning to learn by gradient descent by gradient descent. Advances in neural information processing systems; ; 2016. p. 3981-9.

[38] Ruppert D. The elements of statistical learning: data mining, inference, and prediction. 2004.

[39] Zhang L, Tan J, Han D, et al. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017;22:1680-5.

[40] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115-33.

[41] Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of big Data–evolution, challenges and research agenda. Int J Inf Manage. 2019;48:63-71.

[42] Barash Y, Klang E. Automated quantitative assessment of oncological disease progression using deep learning. Ann Transl Med. 2019;7:S379.

[43] Bhatt D, Patel C, Talsania H, et al. CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics. 2021;10:2470.

[44] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25:1097-105.

[45] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278-324.

[46] Baji T. In: Evolution of the GPU device widely used in AI and massive parallel processing. 2018 IEEE 2nd electron devices technology and manufacturing conference (EDTM); IEEE; 2018. p. 7-9.

[47] Karlik B, Olgac AV. Performance analysis of various activation functions in generalized MLP architectures of neural networks. International Journal of Artificial Intelligence and Expert Systems. 2011;1:111-22.

[48] Ramchoun H, Idrissi MAJ, Ghanou Y, et al. Multilayer perceptron: Architecture optimization and training. Int.J.Interact.Multim.Artif.Intell. 2016;4:26-30.

[49] Sewak M, Karim MR, Pujari P. Practical convolutional neural networks: Implement advanced deep learning models using python. Packt Publishing Ltd; 2018.

[50] Lee H, Kwon H. Going deeper with contextual CNN for hyperspectral image classification. IEEE Trans Image Process. 2017;26:4843-55.

[51] Goldberg Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research. 2016;57:345-420.

[52] Vaswani A, Shazeer N, Parmar N, et al. In: Attention is all you need. Advances in neural information processing systems; ; 2017. p. 5998-6008.

[53] Wolf T, Chaumond J, Debut L, et al. In: Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations; ; 2020. p. 38-45.

[54] Ranftl R, Bochkovskiy A, Koltun V. In: Vision transformers for dense prediction. Proceedings of the IEEE/CVF international conference on computer vision; ; 2021. p. 12179-88.

[55] Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961. 2021.

[56] S. Rajbhandari, J. Rasley, O. Ruwase, et al. In: ZeRO: Memory optimizations toward training trillion parameter models. - SC20: International conference for high performance computing, networking, storage and analysis; ; 2020. p. 1-16.

[57] Brown WM, Nguyen TD, Fuentes-Cabrera M, et al. An evaluation of molecular dynamics performance on the hybrid cray XK6 supercomputer. Procedia Computer Science. 2012;9:186-95.

[58] Liu J, Wu J, Panda DK. High performance RDMA-based MPI implementation over InfiniBand. International Journal of Parallel Programming. 2004;32:167-98.

[59] Liu L, Liu X, Gao J, et al. Understanding the difficulty of training transformers. arXiv preprint arXiv:2004.08249. 2020.

[60] Mercado R, Rastemo T, Lindelöf E, et al. Graph networks for molecular design. Machine Learning: Science and Technology. 2021;2:025023.

[61] Qiao Z, Welborn M, Anandkumar A, et al. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. J Chem Phys. 2020;153:124111.

[62] Pyakillya B, Kazachenko N, Mikhailovsky N. In: Deep learning for ECG classification. Journal of physics: Conference series; IOP Publishing; 2017. p. 012004.

[63] Edenbrandt L, Devine B, Macfarlane PW. Neural networks for classification of ECG ST-T segments. J Electrocardiol. 1992;25:167-73.

[64] Ebrahimi Z, Loni M, Daneshtalab M, et al. A review on deep learning methods for ECG arrhythmia classification. Expert Systems with Applications: X. 2020;7:100033.

[65] Savalia S, Emamian V. Cardiac arrhythmia classification by multi-layer perceptron and convolution neural networks. Bioengineering. 2018;5:35.

[66] Rubin J, Parvaneh S, Rahman A, et al. Densely connected convolutional networks for detection of atrial fibrillation from short single-lead ECG recordings. J Electrocardiol. 2018;51:S18-21.

[67] Xia Y, Wulan N, Wang K, et al. Detecting atrial fibrillation by deep convolutional neural networks. Comput Biol Med. 2018;93:84-92.

[68] Zhao Z, Zhang Y, Deng Y, et al. ECG authentication system design incorporating a convolutional neural network and generalized S-transformation. Comput Biol Med. 2018;102:168-79.

[69] Faust O, Shenfield A, Kareem M, et al. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. Comput Biol Med. 2018;102:327-35.

[70] Shashikumar SP, Shah AJ, Clifford GD, et al. In: Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining; ; 2018. p. 715-23.

[71] Kamaleswaran R, Mahajan R, Akbilgic O. A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length. Physiol Meas. 2018;39:035006.

[72] X. Fan, Q. Yao, Y. Cai, et al. Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. IEEE Journal of Biomedical and Health Informatics. 2018;22(6):1744-53.

[73] Chen M, Wang G, Xie P, et al. In: Region aggregation network: Improving convolutional neural network for ecg characteristic detection. 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC); IEEE; 2018. p. 2559-62.

[74] Taji B, Chan AD, Shirmohammadi S. False alarm reduction in atrial fibrillation detection using deep belief networks. IEEE Transactions on Instrumentation and Measurement. 2017;67:1124-31.

[75] Poh MZ, Poh YC, Chan PH, et al. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. Heart. 2018;104:1921-8.

[76] Xu X, Wei S, Ma C, et al. Atrial fibrillation beat identification using the combination of modified frequency slice wavelet transform and convolutional neural networks. Journal of healthcare engineering. 2018;2018.

[77] Sodmann P, Vollmer M, Nath N, et al. A convolutional neural network for ECG annotation as the basis for classification of cardiac rhythms. Physiol Meas. 2018;39:104005.

[78] Clifford GD, Liu C, Moody B, et al. In: AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. 2017 computing in cardiology (CinC); IEEE; 2017. p. 1-4.

[79] Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet. 2019;394:861-7.

[80] Keeley EC, Boura JA, Grines CL. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: A quantitative review of 23 randomised trials. The Lancet. 2003;361:13-20.

[81] Allen Jr B, Seltzer SE, Langlotz CP, et al. A road map for translational research on artificial intelligence in medical imaging: From the 2018 national institutes of Health/RSNA/ACR/The academy workshop. Journal of the American College of Radiology. 2019;16:1179-89.

[82] Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. Journal of the American Medical Informatics Association. 2017;24:198-208.

[83] Santosh K. AI-driven tools for coronavirus outbreak: Need of active learning and cross-population train/test models on multitudinal/multimodal data. J Med Syst. 2020;44:1-5.

[84] Keeley EC, Boura JA, Grines CL. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials. Lancet. 2003;361(9351):13-20.

[85] Shahin M, Obeid S, Hamed L, et al. Occurrence and Impact of Time Delay to Primary Percutaneous Coronary Intervention in Patients With ST-Segment Elevation Myocardial Infarction. Cardiol Res. 2017;8(5):190-198.

[86] Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J. 2017;39(2):119-177.

[87] Yiadom MY, Baugh CW, Mcwade CM, et al. Performance of Emergency Department Screening Criteria for an Early ECG to Identify ST-Segment Elevation Myocardial Infarction. J Am Heart Assoc. 2017;6(3)

[88] Garvey JL, Zegre-hemsey J, Gregg R, Studnek JR. Electrocardiographic diagnosis of ST segment elevation myocardial infarction: An evaluation of three automated interpretation algorithms. J Electrocardiol. 2016;49(5):728-732.

[89] Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700.

[90] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.

[91] Xue J, Aufderheide T, Scott wright R, et al. Added value of new acute coronary syndrome computer algorithm for interpretation of prehospital electrocardiograms. J Electrocardiol. 2004;37 Suppl:233-9.

[92] Dehnavi AR, Farahabadi I, Rabbani H, Farahabadi A, Mahjoob MP, Dehnavi NR. Detection and classification of cardiac ischemia using vectorcardiogram signal via neural network. J Res Med Sci. 2011;16(2):136-42.

[93] Papaloukas C, Fotiadis DI, Likas A, Michalis LK. An ischemia detection method based on artificial neural networks. Artif Intell Med. 2002;24(2):167-78.

[94] Neagoe V, Iatan I, Grunwald S. A neuro-fuzzy approach to classification of ECG signals for ischemic heart disease diagnosis. AMIA Annu Symp Proc. 2003;494-8.

[95] Forberg JL, Khoshnood A, Green M, et al. An artificial neural network to safely reduce the number of ambulance ECGs transmitted for physician assessment in a system with prehospital detection of ST elevation myocardial infarction. Scand J Trauma Resusc Emerg Med. 2012;20:8.

[96] Baxt WG, Shofer FS, Sites FD, Hollander JE. A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain. Ann Emerg Med. 2002;40(6):575-83.

[97] Sbrollini A, De jongh MC, Ter haar CC, et al. Serial electrocardiography to detect newly emerging or aggravating cardiac pathology: a deep-learning approach. Biomed Eng Online. 2019;18(1):15.

[98] Xiao R, Xu Y, Pelter MM, Mortara DW, Hu X. A Deep Learning Approach to Examine Ischemic ST Changes in Ambulatory ECG Recordings. AMIA Jt Summits Transl Sci Proc. 2018;2017:256-262.

[99] Maglaveras N, Stamkopoulos T, Pappas C, Strintzis MG. An adaptive backpropagation neural network for real-time ischemia episodes detection: development and performance analysis using the European ST-T database. IEEE Trans Biomed Eng. 1998;45(7):805-13.

[100] Muhlestein JB, Le V, Albert D, et al. Smartphone ECG for evaluation of STEMI: results of the ST LEUIS Pilot Study. J Electrocardiol. 2015;48(2):249-59.

[101] De fauw J, Ledsam JR, Romera-paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342-1350.

[102] Ding Y, Sohn JH, Kawczynski MG, et al. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using F-FDG PET of the Brain. Radiology. 2019;290(2):456-464.

Chapter 3:

The effect of confounding data features on a DL algorithm to predict complete coronary occlusion in a retrospective observational setting

3.1 Introduction

The previous chapter presented a systematic literature review and meta-analysis of DL-based AMI detection from ECG signals. It was concluded that previous studies of ML for AMI detection had shown promise, but that further work was needed to evaluate the potential of DL in this domain. This chapter presents an experiment designed to evaluate the ability of a DL algorithm to detect hyper-acute AMI caused by acute complete thrombotic coronary occlusion (ACTCO): the pathophysiology underlying STEMI. To set the context for this work, additional detail on the history of STEMI and its relationship to non-STEMI (NSTEMI) is given below.

In 1918, Smith et al. ligated the coronary arteries of canine models [1] while recording ECGs. In healthy individuals, most regions of the myocardium derive their blood supply from just one of the three major coronary arteries (right coronary artery or RCA: left anterior descending or LAD; left circumflex or LCX) [2]. Thus, surgical ligation of a large coronary artery is expected to cause transmural (as opposed to partial thickness) ischaemia. In the canine models in Smith et al.'s experiment, it was noted that transmural ischaemia was reliably associated with STE on the ECG. Since then, STE has been associated with transmural AMI in humans, and has become the bedside test of choice for this condition [3]. STEMI in humans is most commonly caused by ACTCO [4].

Patients suffering from STEMI have been consistently shown to benefit from PPCI, whereas patients with NSTEMI have not [5, 6]. The decision to activate the PPCI pathway is, therefore, largely contingent upon the presence of STE plus a clinical presentation in keeping with AMI (usually chest pain) [3]. The limitation of this approach is that, while STE is very specific for ACTCO, its sensitivity may be as low as 50% [7].

There have been few large-scale studies evaluating alternative models for predicting which patients will benefit from primary PCI [8]. Furthermore, such attempts have principally focussed on extending urgent revascularisation to 'high risk' NSTEMIs, generally defined using a very small number of hand-crafted features (sometimes just two or three) and not incorporating ECG features [9, 10]. It could be argued that such low-dimensional feature representations poorly express the complex physiology of the patient with AMI, and that an approach incorporating more relevant features might be more effective.

In the domain of AF detection, DL models have been shown to match "expert level" performance in the context of ambulatory recordings [11]. This is the highest possible performance one could expect for a task where the gold standard diagnostic criteria are based on expert interpretation of ECG data. In the domain of AMI, on the other hand, it is possible to use composite definitions that do not rely on ECG criteria but incorporate biochemical and angiographic data [4]. Therefore, it is plausible that a DL model could not only match, but also outperform, existing gold standard ECG criteria.

The aim of this study was to establish whether a DL algorithm can detect ACTCO, as defined by angiographically-proven acute coronary occlusion, by leveraging more complex ECG features than a manual approach would allow.

3.2 Methods

3.2.1 Data acquisition

ECG signals were downloaded from the STAFF III database (Physionet) [12-14]. This contains a collection of ECGs taken from 104 patients undergoing prolonged intracoronary balloon inflation. The records consist of nine lead ECGs at 1000Hz (investigators can calculate the three augmented limb leads if they wish). 76 records contain baseline ECGs obtained in a relaxing room prior to transfer to theatre. The inflations lasted an average of 262 seconds, with 84 lasting in excess of five minutes. Annotations contain the time of balloon inflations and deflations, contrast injection times and anatomical position of the balloons.

STAFF III remains one of the most valuable datasets for groups studying the early ECG effects of prolonged, total coronary occlusion in humans. It is the only publicly available dataset that contains angiographically-proven acute coronary artery occlusion without pre-selecting subjects based on ECG criteria nor chest pain.

Basic demographic information from the 76 STAFF III subjects included as per the original inclusion criteria (described below) are shown in Table 3.1.

Table 3.1 (first iteration) – demographic details, including subgroups defined by anatomical location of balloon inflation. LMS = left main stem, LAD = left anterior descending, Diag = diagonal branch, LCx = left circumflex, RCA = right coronary artery

Patient characteristics	All patients	LMS	LAD	Diag	LCx	RCA
Male, n (%)	51 (67.1)	2 (100)	11 (52.4)	2 (100)	10 (62.5)	26 (74.3)
Female	25	0	10	0	6	9
Age, mean years (range)	60 (32-100)	62 (55-70)	61 (40-85)	53 (53,54)	65 (32-100)	58 (38,80)

3.2.2 Ethical considerations

No ethical issues were identified with this study, as it involved open data from an anonymised, publicly available database. This decision was ratified by the heads of research governance at two of the participating academic centres (Ulster University and Southern Health and Social Care Trust).

3.2.3 Inclusion / exclusion criteria

Initially, only records that included relaxing room ECGs were deemed eligible, as these were used as the non-ischaemic samples. Records where balloon inflations lasted less than 90 seconds were excluded as they contained insufficient ischaemic samples.

Several subjects underwent multiple inflations in different anatomical locations. Only data from the first inflation was used due to concerns that "hangover" electrical effects from previous inflations may confound results.

The study was executed and written up following completion of this initial protocol. However, following a conversation with a group who have worked extensively with the STAFF III database (including its creator), it was pointed out that the 28 patients excluded because they had no ECG from the relaxing room could be included if the beginning of their theatre ECG (taken prior to catheter insertion) was used as an alternative baseline.

It was decided that the experiment should be re-run with the inclusion criteria thus amended. It was also felt that standardising the baseline ECG acquisition by using pre-catheterisation theatre ECGs for all patients would be more methodologically sound.

3.2.4 Algorithm design

The model was a 34-layer CNN with residual connections culminating in a fully connected layer with a single, sigmoid-activated output node. Researchers from the Stanford Machine Learning Group have identified this architecture as being particularly well-suited to processing ECG signal data [11]. The model was initiated using weights from an AF detection task [15], on the assumption that many ECG features learned during arrhythmia analysis would improve generalisation in the setting of ischaemia detection. This is known as 'transfer learning' and can allow DL models to train for complex tasks on relatively small datasets [16].

During the training process, ECG signals were split into one second segments. Each ECG window was reshaped into a 9000 dimensional vector (9 leads x 1000Hz x 1 second). The loss was calculated using binary cross-entropy, where non-ischaemic samples were labelled 0, ischaemic traces 1.

3.2.5 Model evaluation

The model was evaluated using a 5-fold cross validation (CV) process, whereby each of 5 versions of the model were trained on data from 80% of the patients and tested on data from the remaining 20%. The experiment was subsequently repeated using a 10-fold CV process whereby data was split into 80% training, 10% validation and 10% test sets. This was to ensure the 5-fold CV process did not encourage overfitting.

Testing was undertaken using one 10 second trace for each patient taken from the baseline ECG (non-ischaemic examples) and one 10 second trace for each patient taken 60 seconds into

balloon occlusion of a coronary artery (positive examples). 10 seconds was chosen because it is the standard length of printed 12 lead ECGs used to diagnose STEMI and would facilitate a fair comparison with cardiologist-labelled benchmarks.

The input vector for the model comprised a tensor of shape [batch size, 10, 9000]. The final dimension comprised one second of samples for each of nine leads at 1000Hz concatenated into a 9000-dimensional vector (the augmented limb leads were not explicitly calculated for the model). The penultimate dimension represented the 10 seconds of the ECG.

3.2.6 Benchmarks

Three consultant cardiologists were given all of the test traces in a random order and asked to label them as showing either no signs of ischaemia, non-specific ischaemic changes or STE. These results were used as a basis for comparison with the DL model performance as described below.

3.2.7 Statistical analysis

The accuracy of each classifier was calculated by dividing the number of correct labels with the total number of ECGs labelled. The consensus opinion of the three cardiologists regarding both non-specific ischaemic changes and STE was taken to be the current gold standard in clinical practice. This was evaluated against the DL model's accuracy using the Chi-square test. For each classifier sensitivity, specificity, positive predictive value (PPV) and F1 score were calculated.

A receiver operating characteristic (ROC) curve was plotted for the DL model and AUROC calculated.

3.2.8 Interrogating the model

Attention heatmaps were generated using selective input masking. The fully trained model was shown each ECG in the test set with 50 millisecond (mS) segments "blanked out" (by substituting voltage values for zero). The greater the difference between the original prediction and the new prediction, the higher the value assigned to the masked part of the ECG on the heatmap. The process was repeated until a value had been assigned to each 50 mS window of each ECG.

3.3 Results

3.3.1 First iteration of the study using original inclusion and exclusion criteria

The results of ECG analysis by ST-elevation criteria (as defined by consensus opinion among the three cardiologists), individual analysis by each expert using a combination of both STEMI

criteria and non-specific ischaemic changes, consensus opinion among the experts using both STEMI criteria and non-specific ischaemic changes, and analysis by the DL model are shown in Figure 3.1. The DL model had both the highest accuracy (0.803) and the highest F1 score (0.814). Classification using the STEMI criteria produced the highest specificity (0.947). Cardiologist 3 achieved the highest sensitivity (0.842).

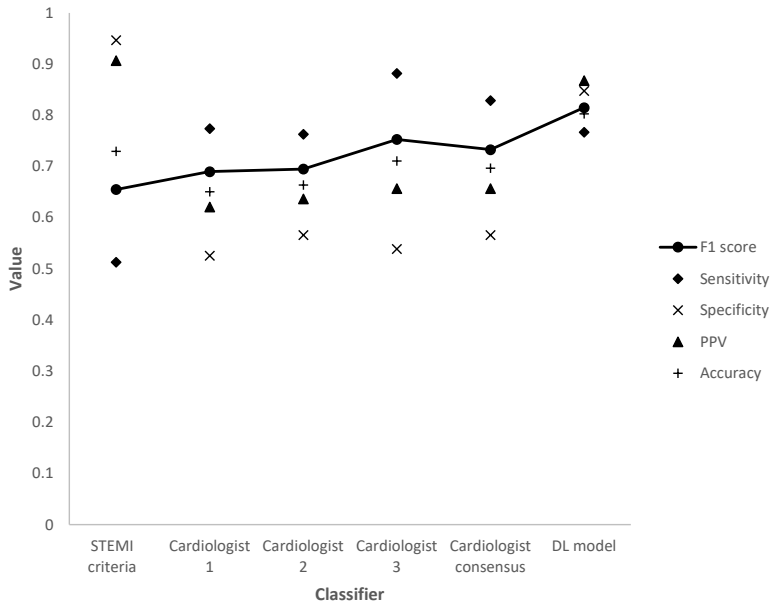


Figure 3.1 (first iteration) – performance metrics of each classifier across the whole dataset

The confusion matrices used to calculate these results are included in Table 3.3. As previously noted, the DL model's results were calculated by taking the mean results of each cycle of the 5-fold CV process. Confidence intervals (95%) for these results are shown in Figure 3.2.

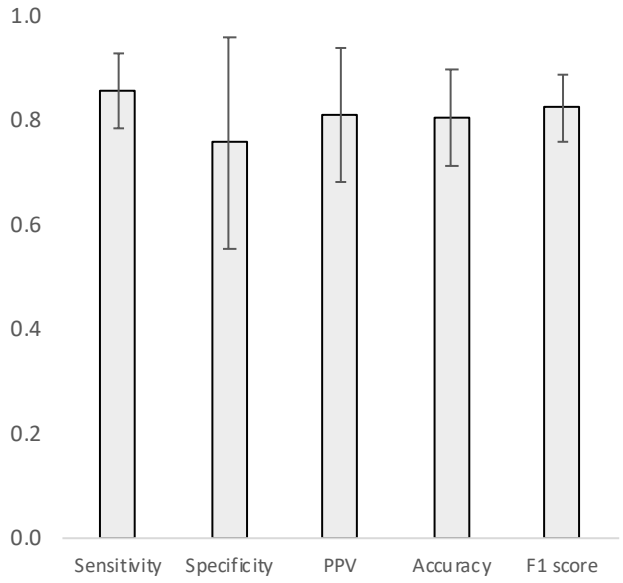


Figure 3.2 (first iteration) – results from the 5-fold cross validation process of the deep learning model across the whole dataset (averages and 95% confidence intervals)

Difference in accuracy between the DL model and the consensus cardiologist opinion for any type of ischaemic change was evaluated using the Chi-square test and found to be significant using a threshold of 0.05 (p=0.0469). Marginal homogeneity was evaluated using McNemar's test. Results are shown in Table 3.2.

Table 3.2 (first iterations) – classifier concordance calculated using McNemar's test. Statistically significant results (p < 0.05) in bold.

	STEMI	Cardiologist 1	Cardiologist 2	Cardiologist 3	DL model
STEMI	-	0.193	0.126	0.699	0.177
Cardiologist 1	0.193	-	0.856	0.238	**0.009**
Cardiologist 2	0.126	0.856	-	0.201	**0.004**
Cardiologist 3	0.699	0.238	0.201	-	0.065
DL model	0.177	**0.009**	**0.004**	0.065	-

Figure 3.3 shows ROC curve for the DL model. AUROC was 0.860. Results were reproducible using a 10-fold CV process as described in the methods section. Attention heatmaps appeared to show that the model was primarily focussing on the latter part of the QRS complex or the ST-T segment. See Figure 3.4 for an example.

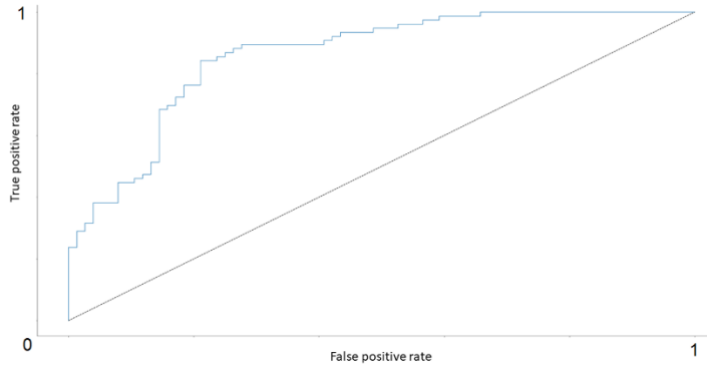


Figure 3.3 (first iteration) – ROC curve for the DL model (AUROC = 0.860). The dotted black line represents the ROC for a binary classifier based on random chance, where AUROC = 0.5

3.3.2 Second iteration of the study using amended inclusion and exclusion criteria

Following amendment of the inclusion criteria so that baseline samples were obtained from theatre ECGs, 99 patients were included in the second run of the experiment. The model was retrained using the same 5-fold CV process, the same data sampling methods and the same hyperparameters as the first run.

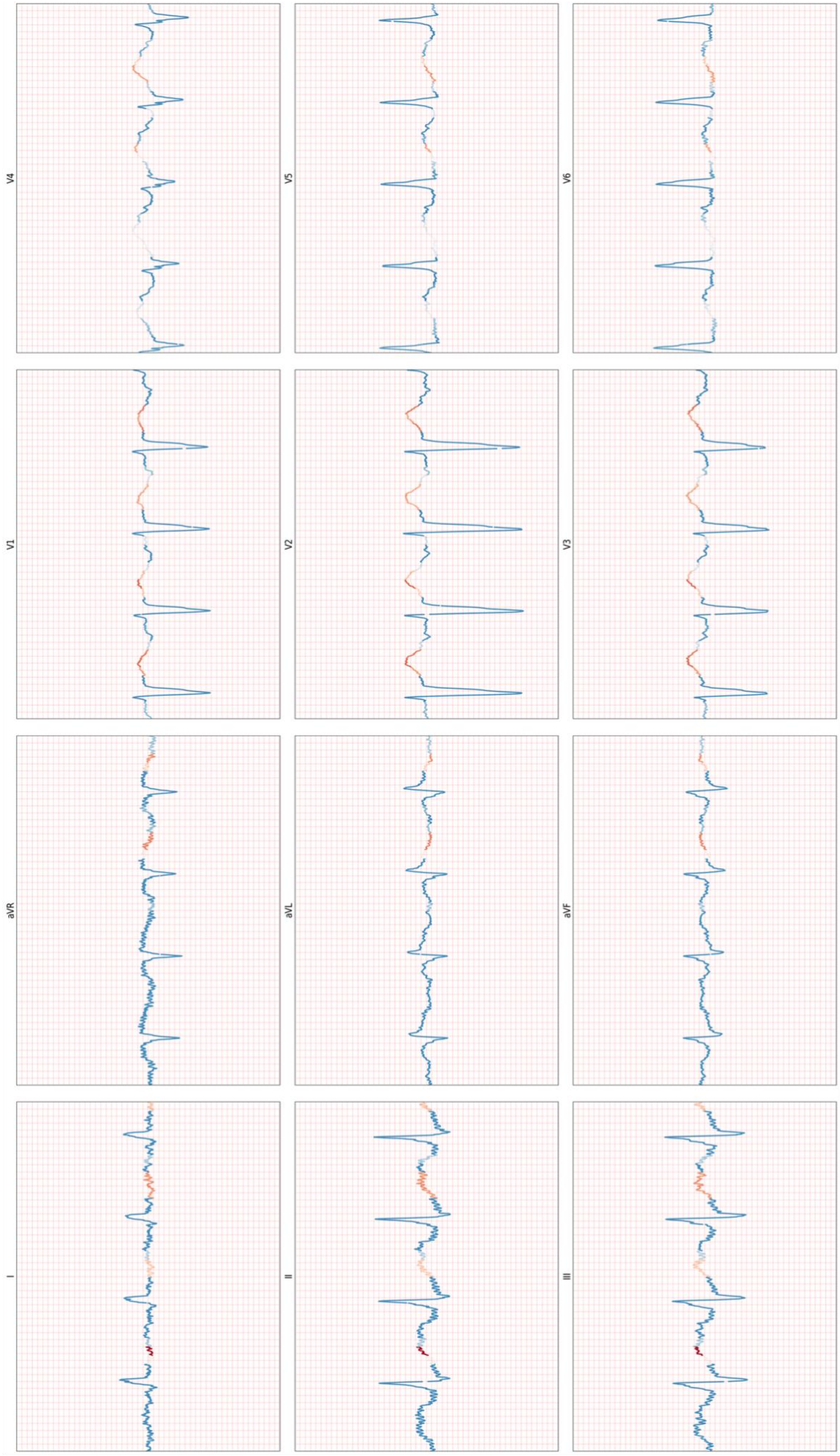


Figure 3.4 (first iteration) - an example heat map for an ischaemic example, obtained selectively masking input data to establish which parts of the ECG the model relies on most to make its prediction

Accuracy was 0.555 (standard deviation 0.08, 95% confidence interval 0.505 – 0.605). F1 score was 0.533 (standard deviation 0.17, 95% confidence interval 0.433 – 0.633). The experiment was repeated in case the stochastic nature of the DL approach has resulted in particularly poor results, but there was no change.

The results provide a case study of a DL model that, under certain conditions, may achieve high accuracy scores due to its ability to also exploit confounders and data leakages. This explains why the results in iteration 1 are superior to the results in iteration 2. The high performance in iteration 1 is likely due to the DL model detecting 'noise' as opposed to detecting ischaemia.

Table 3.3 – confusion matrices from the first iteration of the experiment

STEMI	Predicted: YES	Predicted: NO
Actual: YES	39	37
Actual: NO	4	72
Cardiologist 1		
Actual: YES	58	18
Actual: NO	33	43
Cardiologist 2		
Actual: YES	59	17
Actual: NO	36	40
Cardiologist 3		
Actual: YES	67	9
Actual: NO	35	41
DL model		
Actual: YES	66	10
Actual: NO	20	56

3.4 Discussion

This single centre, retrospective, observational study of 104 patients investigated the ability of a DL model to predict hyperacute myocardial ischaemia from ECG recordings. The first iteration, which obtained non-ischaemic samples from resting room ECGs, appeared to have an ability to detect ischaemia. The second iteration, which obtained non-ischaemic samples from inside theatres, did not. In the first iteration, the model appeared to outperform a panel of

three cardiologists with statistical significance. On the latter occasion, the model performed at the level of a random chance classifier.

The proposed explanation for the discrepancy in results is that the first model learned to associate background electrical noise in theatre with ischaemic samples during the first run of the experiment. Background electrical activity in cardiac theatres is known to manifest on ECGs (including noise in the 100Hz range from fluoroscopy) [17]. Given that the 'ischaemic' ECGs exhibited this noise, the algorithm was able to discriminate between ischaemia and non-ischaemia by simply detecting the noise in the 'ischaemic' ECGs. This is referred to as data leakage or a confounding factor.

During the second run, all samples were acquired in theatre and the model's true ability to discern causative (as opposed to purely correlative) links within the data was revealed. The hypothesis had been that transfer learning from an arrythmia detection task may allow the model to glean generalisable insights from a small dataset [18], but the results demonstrate that this was not the case.

This experiment is not the first study showcasing how DL models can leverage confounding factors within the data to produce spuriously high performance. A number of similar occurrences have been described in healthcare and other domains [19-22]. Deep learning is currently receiving much attention in the domain of automated ECG interpretation, as it is in the fields of cardiac imaging, coronary evaluation and heart failure [23]. It is, therefore, particularly important that the cardiology community be aware of its pitfalls as well as its strengths.

It is acknowledged that this was a highly speculative experiment at increased risk of spurious results due to a small study cohort and retrospective, observational setting [24]. It is also recognised that neither cross-validation nor any other approach to validation guarantees against such an outcome, and agree with recent calls for more ML and DL applications to be in evaluated prospective, multi-centre clinical trials [25-27]. However, it must be noted that even DL algorithms trained on huge datasets and extensively validated by world-leading technical experts can behave in surprising, unacceptable and sometimes catastrophic ways [28, 29]. In addition, such tools may not integrate well into current clinical practice, where transparency is highly prized [30, 31].

Based on these results, it is proposed that AI in the medical domain must always retain a degree of 'explainability' in order to facilitate human oversight and supervision. This does not necessarily require an exhaustive account of a DL model's logic, which is encoded by the state of millions of coefficients within a complex computing graph [16] and may be impossible to explain in human terms. Rather, it is proposed that the clinical community stipulate a set of

minimum requirements for what is determined to be acceptable transparency in future cardiac DL applications.

3.5 Conclusion

In summary, DL continues to show significant promise and has many potential applications in modern medical practice [32]. However, it remains a nascent technology. It was concluded from this study that future research was needed to develop mechanisms to allow clinicians to calibrate their confidence in DL applications before using them to inform clinical decision making.

Whilst this chapter uses raw ECG signals that would accessible for analysis, not all raw data from ECGs are easily accessed for signal analysis. However, given that all ECGs are presented visually as an image or PDF (for example), it is important to determine the value of machine learning when using ECG images as opposed to the raw ECG amplitude data. Hence, the next chapter focuses on DL for ECG image analysis. This topic was noted in chapter 2 to be an important component of democratising ECG AI, and the rationale for that statement will be presented shortly. However, DL for ECG image analysis also has important benefits for interpretable ECG AI. These will also be explained at the outset of the next chapter, which is proposed as an appropriate follow-on from the findings described above.

3.6 References

[1] Smith, F.M. The ligation of coronary arteries with electrocardiographic study. Archives of Internal Medicine, 1918;22(1):8-27.

[2] Seiler C. The human coronary collateral circulation. Heart. 2003;89:1352-7.

[3] Ibanez, B., James, S., Agewall, S., Antunes, M.J., Bucciarelli-Ducci, C., Bueno, H., Caforio, A.L., Crea, F., Goudevenos, J.A. and Halvorsen, S. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). European heart journal, 2017;39(2):119-177.

[4] Rott D, Leibowitz D. STEMI and NSTEMI are two distinct pathophysiological entities. Eur Heart J. 2007;28:2685-.

[5] Menown, I., MacKenzie, G. and Adgey, A. Optimizing the initial 12-lead electrocardiographic diagnosis of acute myocardial infarction. European heart journal, 2000;21(4):275-283.

[6] Cox, D.A., Stone, G.W., Grines, C.L., Stuckey, T., Zimetbaum, P.J., Tcheng, J.E., Turco, M., Garcia, E., Guagliumi, G. and Iwaoka, R.S. Comparative early and late outcomes after primary percutaneous coronary intervention in ST-segment elevation and non-ST-segment elevation acute myocardial infarction (from the CADILLAC trial). The American Journal of Cardiology, 2006;98(3):331-337.

[7] Pollehn, T., Brady, W.J., Perron, A.D. and Morris, F. The electrocardiographic differential diagnosis of ST segment depression. Emergency medicine journal, 2002;19(2):129-135.

[8] Banning, A.S. and Gershlick, A.H. Timing of intervention in non-ST segment elevation myocardial infarction. European Heart Journal Supplements, 2018;20(suppl. B):B10-B20.

[9] Badings, E.A., Dambrink, J.H., Tjeerdsma, G., Rasoul, S., Timmer, J.R. and Lok, D.J. Early or late intervention in high-risk non-ST-elevation acute coronary syndromes: results of the ELISA-3 trial. EuroIntervention: journal of EuroPCR in collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology, 2013;9(1):54-61.

[10] Mehta, S.R., Granger, C.B., Boden, W.E., Steg, P.G., Bassand, J., Faxon, D.P., Afzal, R., Chrolavicius, S., Jolly, S.S. and Widimsky, P. Early versus delayed invasive intervention in acute coronary syndromes. New England Journal of Medicine, 2009;360(21):2165-2175.

[11] Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P. and Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine, 2019;25(1):65.

[12] Martinez, J.P., Pahlm, O., Ringborn, M., Warren, S., Laguna, P. and Sornmo, L. The STAFF III database: ECGs recorded during acutely induced myocardial ischemia, Computing in Cardiology (CinC), 2017;IEEE:1-4.

[13] Pettersson, J., Carro, E., Edenbrandt, L., Maynard, C., Pahlm, O., Ringbord, M., Sornmo, L., Warren, S.G. and Wagner, G.S. Spatial, individual, and temporal variation of the high-frequency QRS amplitudes in the 12 standard electrocardiographic leads. American Heart Journal, 2000;139(2):352-358.

[14] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C. and Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation, 2000;101(23):e215-e220.

[15] Brisk, R., Bond, R., Banks, E., Piadlo, A., Finlay, D., McLaughlin, J. and McEneaney, D. Deep learning to automatically interpret images of the electrocardiogram: Do we need the raw samples? Journal of electrocardiology, 2019;57S:S65-S69.

[16] Goodfellow, I., Bengio, Y. and Courville, A., 2016. Deep learning. MIT press.

[17] Haar, C.C.T., Maan, A.C., Schalij, M.J. and Swenne, C.A. ST and ventricular gradient dynamics during percutaneous transluminal coronary angioplasty, Computing in Cardiology, 2012;341-344.

[18] Yu, X. and Aloimonos, Y. Attribute-Based Transfer Learning for Object Categorization with Zero/One Training Example, ECCV, 2010;127-140.

[19] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015;1721-1730.

[20] Barocas, S. and Selbst, A.D. Big data's disparate impact. Calif.L.Rev., 2016;104:671.

[21] Sweeney, L., 2013. Discrimination in online ad delivery. arXiv:2013;1301.6822.

[22] Saunders, J., Hunt, P. and Hollywood, J.S. Predictions put into practice: a quasi-experimental evaluation of Chicagoâ€™s predictive policing pilot. Journal of Experimental Criminology, 2016;12(3):347-371.

[23] Lopez-Jimenez, F., Attia, Z., Arruda-Olson, A.M., Carter, R., Chareonthaitawee, P., Jouni, H., Kapa, S., Lerman, A., Luong, C. and Medina-Inojosa, J.R. Artificial intelligence in cardiology: present and future, Mayo Clinic Proceedings, 2020;1015-1039.

[24] Bollen, C.W., Hoekstra, M.O. and Arets, H.. Pooling of studies in meta-analysis of observational research leads to precise but spurious results. Pediatrics, 2006;117(1):261-262.

[25] Liu, X., Faes, L., Calvert, M.J. and Denniston, A.K. Extension of the CONSORT and SPIRIT statements. The Lancet, 2019;394(10205):1225.

[26] Liu, X., Rivera, S.C., Faes, L., Ruffano, L.F.D., Yau, C., Keane, P.A., Ashrafian, H., Darzi, A., Vollmer, S.J. and Deeks, J. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nature Medicine 2019;25:1467-1468

[27] Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N. and Ho, T.B. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. Journal of medical Internet research, 2016;18(12):e323.

[28] Howard, A. and Borenstein, J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and Engineering Ethics, 2018;24(5):1521-1536.

[29] Stilgoe, J. Machine learning, social learning and the governance of self-driving cars. Social Studies of Science, 2018;48(1):25-56.

[30] Hirsh, J. and Guyatt, G. Clinical experts or methodologists to write clinical guidelines? The Lancet, 2009;374(9686):273-275.

[31] Norheim, O.F. Healthcare rationing—are additional criteria needed for assessing evidence based clinical practice guidelines? BMJ, 1999;319(7222):1426-1429.

[32] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., Depristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J. A guide to deep learning in healthcare. Nature medicine, 2019;25(1):24-29.

Chapter 4:

Using deep learning to interpret images of the electrocardiogram

4.1 Introduction

To date, the vast majority of research into DL-based ECG interpretation has focussed upon raw samples recorded directly from ECG hardware. Yet, there is an enormous body of historical ECG data worldwide that exists only in paper form, or as scanned images thereof [1]. These ECGs are often associated with medical records containing years of rich clinical information: echocardiograms, angiographic findings, cardiac biomarkers, morbidity and mortality endpoints, and so on. It has long been acknowledged that such data could provide a rich source of insights to inform the science of ECG interpretation. Furthermore, the printed ECG is the universal format. Accurate, computerised analysis of ECG images would overcome the difficulties arising from proprietary formats and algorithms, long cited by researchers in the field as a substantial hindrance [2]. ECG image analysis is, therefore, proposed as an important step towards democratising the application of ECG AI.

4.1.1 ECG images and interpretable DL

Perhaps even more importantly, given the conclusion of the previous chapter, ECG images are interpretable by clinicians. Raw samples are not, or at least they are almost impossible to interpret without graphing amplitudes with respect to time. Therefore, DL applications that are trained on ECG images may be more interpretable to human experts than applications trained on raw samples.

To elucidate this position with a more concrete example, one might consider a 10 second 12-lead ECG recorded at 150Hz. In many clinical settings, it is commonplace to print this ECG onto thermal paper using a standard 2.5 seconds-per-lead format, then digitize the printed ECG using a desktop scanner with resolution as low as 200 dots per inch (DPI). This can be done for the purposes of digital storage, or to send the digital image to another centre for review [3-5].

The process of printing and scanning the ECG incurs significant signal loss, both through the 7.5 seconds' worth of samples that are recorded but not printed for each lead, and through loss of resolution during digitization. Yet, the resulting digital image is widely considered adequate to inform major clinical decisions, such as whether to activate the PPCI pathway [6]. It seems reasonable to conclude, therefore, that the key diagnostic features are retained within the digital image.

Seen through this lens, converting raw samples into digital images can be considered a form of 'feature filtering'. In other words, some data features that are not considered diagnostically significant are disregarded, while the key diagnostic features are retained. Feature filtering has been employed by other researchers training clinical DL models, who posit that it can "*'guide' the model to 'look' into more informative regions*" [7]. These authors argue that training DL

models in this way reduces the likelihood that DL models will leverage confounding data features and produces more interpretable clinical applications. By the same logic, training DL models on ECG images rather than raw samples could ameliorate the 'black box' effect described in the previous chapter.

4.1.2 The diagnostic challenge of ECG images

The obvious drawback to training DL models on ECG images is that transposing raw ECG samples into an image creates a huge amount of non-ECG-related noise. For example, a digital ECG image that is 800 x 600 pixels with three colour channels contains over 1.4 million data points. Only a tiny fraction of these data points comprise the ECG signal; the rest are image background.

The risk here is not necessarily that the DL model will learn confounding data features within background image noise that will subsequently go unnoticed, as was the case with the experiment described in the previous chapter. ECG images are human readable, and features derived from image regions distant to the ECG lines may be readily identified using semantic masks (see next chapter), saliency mapping and similar feature localisation techniques [8]. Rather, the key risk is that the SNR within ECG images will prove too low for a DL model to learn diagnostically useful features and produce accurate results.

Much of the previous work on ECG image analysis using rule-based (i.e. non-ML/DL) algorithms concluded exactly this [2, 9, 10]. However, as discussed in previous chapters, novel DL models (particularly CNNs) have since proven to be extremely adept at handling low SNRs in the context of image data [9, 10]. The central hypothesis being tested by the experiment described below is that the application of CNNs to ECG images can overcome the increased diagnostic challenge associated with low SNR.

4.1.3 Evaluating CNNs as a potentially effective tool for ECG image analysis

The following study was designed to compare a CNN trained with ECG images against DL models trained with raw samples. AF detection was chosen as the diagnostic task for this comparison, as it is a common but clinically important problem with publicly available datasets and performance benchmarks that allow for comparative study. To increase the likelihood that the results of this experiment would generalise to a real-world setting, a dataset consisting of ambulatory ECGs was selected, as these usually contain more noise and environment artifact than recordings in a controlled clinical environment [11].

4.2 Methods

4.2.1 Data acquisition

The 2017 Physionet AF Challenge (PAFC) was identified as an appropriate dataset for this study. It contains 8528 single-lead ambulatory ECG signals, each of which has been labelled as showing NSR, AF, 'other', or 'noisy'. The training data and results from several approaches (both rule-based and DL-based) are available at https://physionet.org/challenge/2017/ [12].

4.2.2 Plotting ECGs to image files

To generate an image database for this study, all ECG recordings were plotted as RGB image files using a standard Python library (MatPlotLib). Original signals were recorded at 300Hz on AliveCor devices, thus a 300 pixels / second resolution would have been required to maintain full resolution. In fact, a target resolution of 150 pixels / second and 75 pixels / mV was chosen, as this corresponds to an ECG printed at 25mm/s and 10mm/mV then scanned using a low-resolution, 150DPI scanner. (Modern digital scanners are usually much higher resolution than this, but 150DPI scanners may still be found in developing health systems and it was felt to be an appropriate test of robustness of the computerised analysis pipeline.) Figure 4.1 shows an example ECG image generated by this process.

Figure 4.1 – an ECG image plotted from raw sample recordings from the PTB database.

4.2.3 Extrapolation of ECG signals from ECG images

A number of approaches to extrapolating ECG signals from ECG images have been explored over previous decades [2].**Error! Bookmark not defined.** In order to accommodate the unique characteristics of the ambulatory ECG dataset, a bespoke extrapolation method was developed based upon these approaches. The method consisted of scaling, thresholding, binarization and column-wise pixel searching. A thorough discussion of each of these techniques is provided by Waits and Soliman [2]. However, a visual summary is presented in Figure 4.2 above, and the link to the full code base is provided later in this chapter. As noted above, it was hypothesised that the CNN used to interpret the signals generated by this extrapolation method would be more robust to noise than most rule-based approaches. Therefore, some noise-filtering techniques used by other authors were omitted (e.g. median filtering and interpolation, which Ravichandran et al. (2013) applied to deal with the "salt-and-pepper" noise caused by thresholding) [9].

4.2.4 DL model

As discussed in previous chapters, state-of-the-art arrhythmia detection from ambulatory signals has been achieved using a 34-layer CNN with residual connections between layers,

Step 1: scale signal vector and plot against gridline background to create ECG image

Step 2: convert to greyscale values

Step 3: apply Otsu filter to create a binary image matrix where 0 = background, 1 = ECG line

```
extrapolated_signal_vector = new List;

For column X in binary_image_matrix:
     For row Y in binary_image_matrix:
          If binary_image_matrix[X,Y] == 0:
               Pass;
          Else:
               extrapolated_signal_vector.append(Y);
               Break;
```

Step 4: undertake column-wise pixel search to extrapolate the raw samples from the ECG image

Figure 4.1 – overview of signal extrapolation process from ECG images

developed by researchers at Stanford University [13]. This architecture was therefore selected for this experiment.

In order to streamline the training process, the model was initiated with pre-trained weights published by researchers at Oxford University, who had trained a model with the same architecture on the raw signals from the PAFC [14]. After some experimentation, the model architecture was modified slightly for handling image-derived data, with two fully connected layers each containing 512 nodes interposed between the final convolutional layer and the fully connected output layer (which contained four nodes, as this was a four-class problem). The weights of the additional fully connected layers of the model were randomly initialised.

4.2.5 Training and analysis

Model performance was evaluated on the entire dataset prior to any training. This was necessary to ensure the pre-trained weights obtained from the Oxford team did not cause the model to overfit the data.

The model was then trained and evaluated using a five-fold cross validation (5FCV) process with 80% of the data used for training and 20% for validation during each 5FCV cycle. During training, the weights of the latter six layers of the network (two fully connected layers and four convolutional layers) were progressively unfrozen. Each time a new layer was unfrozen, the model was trained until five epochs had passed without improvement in the validation accuracy.

5FCV was chosen because six of the top 10 scoring teams in the PAFC published results from 5FCV on the training set, so it was possible to make a direct comparison with their models. It should be noted that the 5FCV results were published within papers written by each individual team; the results from the collective scoreboard were based on a hidden test set. Therefore, none of the official competition results in were included in the analysis.

As in the competition itself, the single performance metric used to undertake a like-for-like comparison between models was the combined F1 score, which is the harmonic mean of the F1 score for each of the four categories.

4.3 Results

The model was evaluated on the full image-based dataset upon initialisation with pre-trained weights. The results were in keeping with random chance, with a combined F1 score of approximately 0.5.

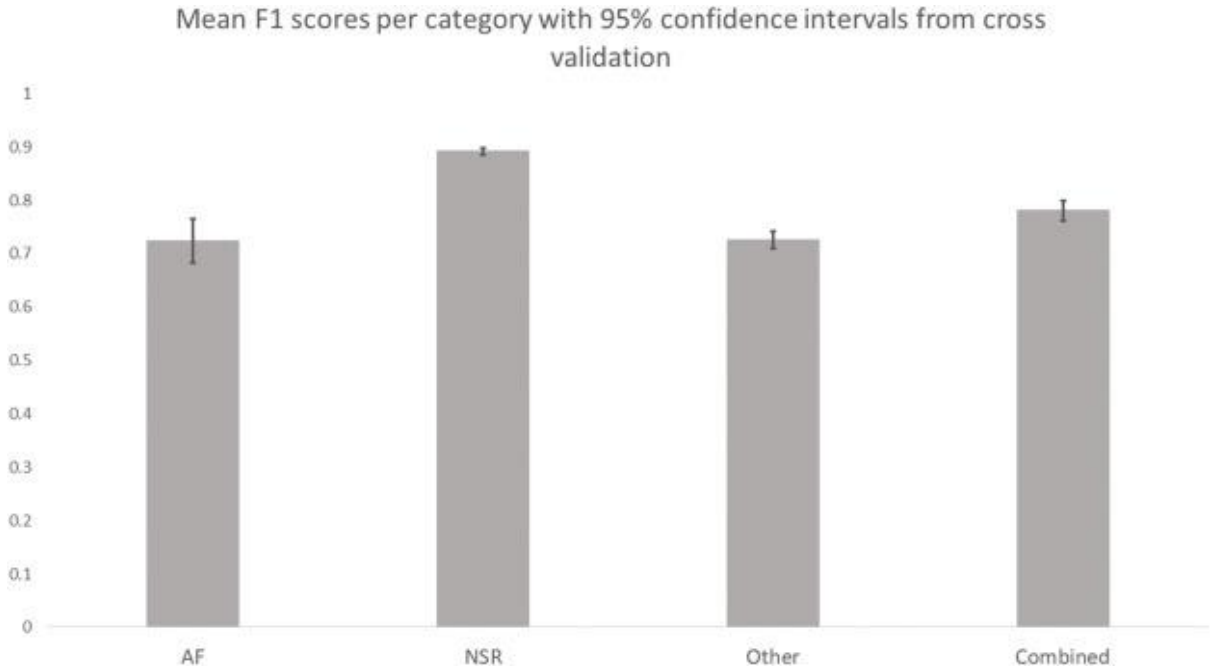


Figure 4.3 – F1 scores obtained by the trained CNN using the extrapolated ECG signals

Following training, the mean combined F1 score and 95% confidence interval across the five cycles of this process was 0.78 (+/- 0.02). The source code for this experiment is available at https://github.com/docbrisky/af-challenge. Figure 4.3 gives a visual report of the F1 score obtained for each of the four categories, plus error bars reflecting the 95% confidence interval across the 5FCV process.

Official scores from the 2017 AF Challenge were based on a hidden test set, which was not available at the time of this study. However, six of the top 10 competitors published 5FCV scores obtained on the training set, which is the same data used for this study. The mean combined F1 score of those six teams was 0.83. (See https://physionet.org/challenge/2017/papers/ for a full list of publications.)

The model produced by the Oxford University team whose weights were used for model initialisation obtained a combined F1 score of 0.72 at 5FCV.

4.4 Conclusion

The results produced by this study suggest that CNN-based arrhythmia detection from ambulatory ECG images can be undertaken without substantial loss of accuracy compared with raw sample analysis. This is despite the fact that (i) ambulatory ECGs generally contain more noise and movement artefact than recordings in a controlled environment, (ii) the ECG signals in this study were plotted into particularly low resolution images to simulate outdated hardware and (iii) several noise-filtering techniques were omitted from the signal extrapolation approach [15]. It is proposed that this represents a state-of-the-art result in terms of image-based ECG analysis.

This outcome indicates that there is value in using deep learning to automatically interpret images on the ECG. This would allow anyone with a smartphone camera to use the algorithm, which could democratise the use of ECG algorithms without the need to have access to raw ECG data.

4.5 Discussion

A recent paper in the Lancet provides an apt context for the relevance of this finding. By undertaking a retrospective analysis of over 600,000 ECGs from nearly 200,000 patients, Attia et al. (2019) used a CNN to predict incipient AF among patients currently in normal sinus rhythm with approximately 80% sensitivity and specificity [16]. In this case, the researchers were investigating a high-incidence endpoint (the development of AF) and were able to obtain sufficient digital ECG signals without needing to digitise historic ECG images. However, the obvious question arising from this study is whether patients deemed to be ‘at risk of future AF’

based on an ECG in normal sinus rhythm have a correspondingly increased lifetime risk of stroke, and whether they should therefore be prescribed oral anticoagulation. Pending a prospective study to answer this question, which may take many decades, it is likely to be beneficial to apply Attia et al.'s algorithm to historic ECGs that are already associated with a lifetime of follow-up data. Such ECGs will inevitably be images rather than digital signals, in which case the findings of this study would suggest that (i) signals generated by digitizing ECG images can be used to obtain reliable results from a DL model and (ii) weights obtained by training a CNN on raw signal data can be expected to transfer well to the task of analysing image-derived ECG data.

There are, however, important limitations to this study. Firstly, the ECG images were plotted directly from signal data, rather than being printed and scanned. They therefore contained minimal visual artefact and were unrotated. It is proposed that any additional artefact within printed and scanned ECGs compared with the direct-to-image ECGs would be easily overcome with established image processing techniques, and therefore that the printing and scanning of 8528 ECGs was unnecessary to produce meaningful results from this study. (See Figure 4.2 for an example ECG image used in this study.) Nevertheless, to confirm that the results obtained herein will transfer to printed and scanned ECGs, further work in this area should be undertaken.

Secondly, the pretrained weights used to initialise the convolutional layers of the network had, presumably, been exposed to all of the ECG examples in the Physionet Challenge, albeit in raw signal form. Although three fully-connected layers were appended to the network and randomly initialised, and the performance of the newly-formed network was then confirmed to be approximately equal to a random-chance classifier, there is nonetheless a risk that the early convolutional layers of the network have overfit the data. This may explain why the results obtained from this experiment were substantially better than those obtained by the model whose weights were used for initialisation, though it is proposed that the improvement is down to a greater level of data augmentation and the two additional, fully-connected layers. The only way to evaluate this would be to re-train the network from randomly initialised weights, though any drop in performance of the randomly initialised model could also be ascribed to the stochastic nature of the training process.

Nonetheless, it is proposed that the advent of DL-based ECG interpretation, and particularly its increased robustness to noise and resolution loss, should catalyse a renewed interest in high-quality, automated interpretation of image-based ECGs. In addition to the ability to apply cutting edge diagnostics to historical ECGs, as discussed on the previous page, the results of this study provide a rationale for proceeding to investigate ECG images as a form of feature-filtered data that will promote the development of explainable AI.

The next chapter presents further experimental work in the domain of ECG image analysis. There is a particular focus on mechanisms for interpretable ECG AI, and an investigation of methods for making any form of ECG AI (whether for raw samples or images) more applicable to data-poor areas. In our previous two studies we have shown the challenges (e.g. data leakage) and the value of deep learning with ECG images but there is a need to explore different approaches to 'explainable' ECG algorithms.

4.5 References

[1] Holkeri A, Eranti A, Kenttä TV, et al. Experiences in digitizing and digitally measuring a paper-based ECG archive. J Electrocardiol. 2018;51(1):74-81.

[2] Kligfield P. Overview of the ISCE ECG "genome project". J Electrocardiol. 2003;36 Suppl:163-5.

[3] Garcia TB. Introduction to 12-lead ECG: The art of interpretation. Jones & Bartlett Publishers; 2014.

[4] Garg DK, Thakur D, Sharma S, et al. ECG paper records digitization through image processing techniques. International Journal of Computer Applications. 2012;48:35-8.

[5] Lewis MC, Maiya M, Sampathila N. In: A novel method for the conversion of scanned electrocardiogram (ECG) image to digital signal. International conference on intelligent computing and applications; Springer; 2018. p. 363-73.

[6] Al-Zaiti SS, Shusterman V, Carey MG. Novel technical solutions for wireless ECG transmission & analysis in the age of the internet cloud. J Electrocardiol. 2013;46:540-5.

[7] Pintelas E, Liaskos M, Livieris IE, et al. A novel explainable image classification framework: Case study on skin cancer and plant disease prediction. Neural Computing and Applications. 2021;33:15171-89.

[8] Alqaraawi A, Schuessler M, Weiß P, et al. In: Evaluating saliency map explanations for convolutional neural networks: A user study. Proceedings of the 25th international conference on intelligent user interfaces; ; 2020. p. 275-85.

[9] Srivastava, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 2014;15(1):1929-1958.

[10] Borodinov N, Neumayer S, Kalinin SV, Ovchinnikova OS, Vasudevan RK, Jesse S. Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy. NPJ Computational Materials. 2019;5(25)

[11] Moeyersons J, Smets E, Morales J, et al. Artefact detection and quality assessment of ambulatory ECG signals. Comput Methods Programs Biomed. 2019;182:105050.

[12] Clifford GD, Liu C, Moody B, et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017. Comput Cardiol. 2017;44

[13] Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65-69.

[14] Andreotti F, Carr O, Pimentel MAF, Mahdi A, De Vos M. Comparing Feature-Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG. IEEE Proceedings of the Conference on Computing in Cardiology (CinC); Rennes, France. 2017

[15] Chae DH, Alem YF, Durrani S, Kennedy RA. Performance study of compressive sampling for ECG signal compression in noisy and varying sparsity acquisition. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; Vancouver, BC. 2013;1306-1309

[16] Attia ZI, Noseworthy PA, Lopez-jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394(10201):861-867

Chapter 5:

A wave segmentation pretraining toolkit for electrocardiogram analysis

5.1 Introduction

To recap on the key points from previous chapters, correct ECG interpretation is key to the diagnosis and treatment of myocardial infarction and life-threatening arrhythmias, among many other conditions [1]. Computerised ECG analysers have been in existence for over 50 years [2]. However, semantic interpretation of ECG data requires the identification of subtle patterns from a complex signal. It is challenging to describe this process in conventional computer code. AI can perform strongly in this field because it does not rely on the ability of human experts to expound process knowledge. AI-enabled analysis has led to state-of-the-art performance across a range of ECG interpretations tasks [3].

5.1.1 Types of AI for ECG interpretation

Machine learning-based AI refers to a set of automated statistical modelling techniques. AI models learn through trial and error. At each step of the learning process, the model makes a prediction. An error is calculated based on a loss function. A new set of model parameters is discerned using an optimisation function. Further steps are taken until some endpoint is reached [4].

As discussed in chapter 2, DL arose from the study of artificial neural networks ANNs [5]. ANNs are computational graphs comprising densely interconnected MLPs. They are inspired by the biological brain. The difference between 'classical' ML and DL is often summarised thus: ML techniques generally rely on prior processing of input data to extract key features using expert domain knowledge; DL techniques learn end-to-end processing, which includes feature extraction [6] In practice, this results in a trade-off: DL techniques are able to detect more complex patterns in higher dimensional data compared with ML approaches. They can also function with lower signal-to-noise ratios. However, this is at the cost of being less interpretable.

In the domain of ECG processing, it is the feature extraction step that presents the greatest challenge for conventional (non-AI) applications. As discussed in the previous chapter, this is particularly true of ECG images. However, feature extraction is also a major challenge for raw sample analysis. Filtering noise and other electrical artefact from ECG signals, then identifying key features such as the primary waves, has been a major research theme in automated ECG analysis for decades but is by no means a solved problem [7]. This limits the utility of ML algorithms, where knowledge-based feature extraction remains an important part of the pipeline [8]. It is here that DL algorithms can excel.

5.1.2 State of the art in DL for ECG interpretation

As a variant of ANNs, CNNs leverage large numbers of learnable convolutional filters to detect important signals within high-noise data [5]. They were developed primarily for semantic

analysis of real-world images, but the technology transfers well to ECG signals and has been applied to a broad range of clinical problems [9-12]. As already discussed, this includes a landmark 2019 study by Hannun et al. that claimed ‘cardiologist level’ diagnosis of atrial fibrillation, and even a study later that year by Attia et al. that described a DL algorithm able to detect incipient atrial fibrillation [13, 14]. In the context of raw sample analysis, it is likely that convolutional filters in the earlier layers of a CNN learn filtering methods to deal with common ECG noise and artifact, such as baseline wander, powerline interference and non-cardiac muscle activity. This reduces or negates the need for traditional filtering methods [15].

5.1.3 Possible future directions

As noted in the introductory chapter, transformer neural networks, or just ‘transformers’, emerged from the field of NLP. They use attention mechanisms to parallelise sequential data processing. Attention mechanisms can evaluate the relative importance of distant features within data, whereas CNNs have a limited capacity for this. This can be advantageous when relationships between non-local features are important, such as in multi-clause sentences or even entire documents [16]. However, it has recently been shown that transformer models can scale to sizes up to hundreds of billions of trainable parameters with a relatively linear improvement in performance [17]. The sheer power of these ‘mega-AI models’ means that they are beginning to attain state-of-the-art performance in domains where CNNs have traditionally dominated, such as image processing [18]. Transformers for ECG signal analysis is an active research area, and it may be that this is the place to look for the next wave of breakthroughs in this field [19, 20].

5.1.4 Current challenges

5.1.4a Data paucity

As AI models grow larger and more sophisticated, they need more data to maximise their learning potential. This is becoming particularly true as transformers are being applied in new domains. In the field of ECG analysis, this challenge is being actively addressed by the creation of large public datasets such as Physionet's PTB-XL [21]. However, in rarer ECG conditions, data paucity remains a bottleneck to training even small AI models. As noted in the last chapter, data paucity is also a challenge for ECG image analysis, where SNR is much lower than in raw signal format, and where more training data are needed to compensate for this [22].

5.1.4b Explainable AI

Elucidating the process logic encoded by networks comprising millions of parameters is extremely difficult. This is often referred to as the 'black box effect', which has given rise of a field of study known as ‘explainable AI’ [23]. The black box effect can make it difficult for humans to exercise oversight of an AI system's decision logic. Without this oversight, it is

difficult to be sure that AI models are not leveraging confounding data features and generating results through flawed logical processes. This was demonstrated during an experiment described in a previous chapter. Thus, it is challenging for clinicians to calibrate their confidence in the outputs of an AI systems. Confidence calibration is known to play a key role in ECG interpretation [24]. It was proposed in the previous chapter that ECG image analysis may have a key role to play in this area.

5.1.5 Related work

5.1.5a Overcoming data paucity

RL lessens the need for labelled training data. In RL, an AI model is trained for a task that forces it to learn useful 'latent representations' of the data without manually assigned labels. This can be hard to intuit for non-data scientists, and a full explanation is beyond the scope of this chapter. Interested readers are directed to a review by Bengio et al. (2013) [25].

RL is used for some of the most sophisticated AI models in existence today [16]. Models are pre-trained using RL and then fine-tuned for specific tasks using labelled training data, which is to say that they undergo a further training period for a specific task with constraints placed upon the rate at which they learn [26]. The constrained learning rate means that the fine-tuning period serves to refine the latent representations acquired during pretraining, rather than simply overwriting previous representations with new ones. This latter phenomenon is known as 'catastrophic forgetting' [27].

RL has been investigated in the domain of ECG interpretation by a small number of studies. A recent example is from Sarkar et al. in 2020 [28]. They tasked a model with identifying which augmentations had been applied to ECG signals, such as addition of Gaussian noise or signal flipping. This reduced the need for labelled data when fine-tuning for downstream tasks. However, this is a sparsely explored topic to date.

5.1.5b Explainable DL for ECG analysis

Several approaches to make DL-enabled ECG analysis more explainable have been investigated. A recent paper by Maweu et al. infers the relative importance of key ECG waves with respect to a DL model's output [18]. This approach of retrospectively interrogating trained models to infer logic processes is widely used. An experiment described in a previous chapter describes the use of one such technique known as saliency mapping. It was found that the outputs provided false reassurance, in that they appeared to show that the DL model was leveraging the ST segment to diagnose acute myocardial ischaemia. This supported the idea that the model was leveraging features in the input data known to relate closely to the target label, whereas it was later discovered that this was not the case [29].

A study by Jo et al. prioritises explainable outputs at the algorithm design stage [30]. They detect AF by using two linked AI models. One is for detecting the presence or absence of P waves. The other is for detecting regular or irregular R-R intervals. This follows the established decision logic of clinical experts and results in relatively interpretable outputs. It is unclear that this approach would generalise well to more complex diagnostic patterns.

5.1.6 Focus of the experimental work described in this chapter

Wave identification is a fundamental step for any ECG analysis by a human expert. Therefore, it was hypothesised that a DL model trained to segment key waves from ECG signals (either in raw sample or image format) could:

1. Hypothesis 1: Learn generalisable representations of ECG data and be fine-tuned for downstream tasks with relatively small labelled datasets. This may be particularly useful in ECG image analysis. Pretrained RL models could cater more effectively for rarer ECG diagnoses and promote the democratisation of AI.

2. Hypothesis 2: Be guided to learn features that are recognised by human experts as being diagnostically important. As noted in the last chapter, this 'feature filtering' approach has been used by other groups to develop interpretable DL applications [31].

   It was also hypothesised that human-readable wave segmentation masks (see Figure 5.2b) could provide an indication of both the nature and quality of features learned by the DL model. This could act as a mechanism for confidence calibration and help build trust among clinicians.

3. Hypothesis 3: Facilitate a choice between using DL technology as a feature extractor for explainable downstream analysis using rule-based algorithms, or using DL for end-to-end ECG analysis.

Testing these hypotheses would require a dataset of ECG traces where the individual waves had been accurately segmented, which would be used pretrain a DL model. However, manual segmentation of waves within 12-lead ECGs is extremely laborious and RL approaches generally require very large datasets. They usually circumvent the data labelling bottleneck by leveraging self- or semi-supervised methods [25].

Manual wave segmentation was not, therefore, felt to be practical for this experiment. This led to a further hypothesis, whose evaluation is proposed as the most significant contribution of this study to the field:

4. Hypothesis 4: Representations learned from pretraining on synthetic data and labels will transfer to downstream tasks using real ECG data.

5.2 Methods

5.2.1 Overview

The following approach was designed to test hypotheses (1-4) described above. Steps 2 – 7 below were repeated for raw sample and image formats. Steps 8 and 9 were only undertaken for the image-based experiment.

1. Develop an ECG and segmentation mask generator.
2. Train an AI model to predict segmentation masks using a synthetised dataset: referred to hereafter as Wave Segmentation Pretraining, or WaSP.
3. Predict segmentation masks for a database of real ECGs (for analysis at step 7).
4. Fine-tune the model for downstream diagnostic tasks using database of labelled real ECGs.
5. Re-initialise the model with pre-WaSP weights and train this model for downstream diagnostic tasks using database of labelled real ECGs.
6. Compare the results from steps 4 and 5 to test hypothesis (1).
7. Undertake a qualitative analysis of segmentation masks from step 3 to test hypothesis (2).
8. Develop and evaluate a rule-based diagnostic pipeline to evaluate hypothesis (3).
9. Train a 'mixed modality' model for diagnostic tasks, whereby an ECG signal is read back from the predicted segmentation mask and fed into a 1D AI model.

5.2.2 Terminology

A segmentation mask is a set of labels that overlays some input data, denoting the semantic category to which each datum belongs. In the case of image data, the segmentation mask has the same height and width as the pixel array of the original image. Where the original image contains colour channel values at each position in the pixel array, however, the segmentation mask contains an integer value. This value denotes the semantic class to which each pixel belongs. In the case of a single-class segmentation task – for example, segmenting human faces from photographs – all pixels belonging to a face will be represented by a 1, whereas all other pixels will be considered as background and will be assigned a 0. For the purposes of this experiment, the following target classes were defined:

0	Background
1	P wave

2	P-R interval
3	QRS complex
4	ST segment
5	T wave
6	T-P segment
7	T/P overlap

Downstream tasks can be any task for which a pretrained AI model is subsequently re-trained. In the case of this experiment, these tasks are described in the '*Fine-tuning for diagnostic classification*' section below.

5.2.3 Synthetic ECG generation

An application was developed to simulate 12-lead ECG signals. The Python programming language was used. The aim of the simulator development was to produce a broad spectrum of realistic ECG phenotypes. The parameters determining rhythm and morphology of ECGs were governed by pseudo-random number generation to ensure each ECG was unique. Random noise and baseline wander were added to each signal. Voltages were scaled randomly.

In effect, the simulator was a form of expert system informed by key works in the field such as **[1]**, in addition to the developer's own experience as a practising cardiologist **[**32**]**. The ECG signals were also plotted into 12-lead ECG images. Segmentation masks were generated for each ECG signal and image.

5.2.4 Wave Segmentation Pretraining (WaSP)

5.2.4a Model architecture

U-Net model architectures were used for ECG segmentation. The U-Net is a popular CNN-based architecture for image segmentation. It comprises two halves: an encoder and a decoder. The encoder abstracts high level features from the input image. The decoder generates a segmentation mask based on the encoder feature map [33]. See Figure 5.1a for a visual depiction.

The encoder used for each model was based on the SEResNet architecture [34]. This is one of many permutations of the 'vanilla' CNN. A full review of CNN types is beyond the scope of this work, though such reviews exist [35]. SEResNet was felt to represent a demonstrably performant architecture that would fit with the compute constraints of the experiment. The signal-based model used a 1D U-Net with a SEResNet encoder. The image-based model used a 2D U-Net with a SEResNet152 encoder.

The 1D models were initialised with random parameter values (commonly known as model weights). The 2D models were initialised with weights derived from real-world image classification training with the ImageNet database [36].

5.2.4b Training protocol

For the self-supervised pretraining, 32 000 ECGs and segmentation masks were synthesised. The segmentation models were trained during a single pass through the dataset (known as an epoch). A dice loss function was use with Jaccard smoothing [37]. Hyperparameters (parameters that control the training process, rather than forming part of the model itself) were manually tuned based on the training loss, training F1 score and a visual inspection of segmentation masks at the end of each training cycle.

An enhanced pretraining step was undertaken as an additional experiment. A further 12 000 ECGs and segmentation masks were synthesised. Each ECG showed either SR or AF. Each ECG also showed one of six morphological phenotypes: normal, left anterior hemiblock, left posterior hemiblock, high take-off, left bundle branch block or anterior ST-elevation. A classification head was added to the model encoder to predict the rhythm and morphological phenotype of each ECG. The model was simultaneously trained for both segmentation and classification using a multi-task learning approach.

5.2.5 Fine-tuning for diagnostic classification

The Physionet PTB-XL database was downloaded, along with the label files [21]. This is one of the largest publicly available repositories of labelled ECG signals, comprising 21,837 ECGs from 18,885 subjects. The labels for each ECG include one or more of 71 ECG-SCP statements, and each ECG is assigned one of five diagnostic super-classes. The raw samples were converted to Numpy arrays. They were also plotted into ECG images using the software developed for this experiment.

Two diagnostic classification tasks were undertaken: SR vs AF and normal morphology vs AMI. Following the same logic as the previous chapter, these diagnostic tasks were selected because they are both common but clinically impactful, with high quality labelled datasets and benchmarks publicly available. For each of these tasks, the signals were divided into training, validation and test sets using a 60:20:20 split. A hold-out test set approach was used.

To fine-tune the models for diagnostic classification, average pooling was applied to the output of the final convolutional filter of the U-Net encoder. Two densely connected layers were appended, and a sigmoid activation function applied to the output nodes. See Figure 5.1a for a visual representation.

5.2.6 Rule-based AF detector

A rule-based AF detector was designed to investigate the possibility of using a DL model for feature extraction, then passing the features into a fully explainable rule-based classifier. To create the rule-based AF detector, segmentation masks were predicted for ECG images using the pretrained 2D U-Net model. A rule-based algorithm was used to determine the locations of QRS complexes, based on clusters of pixels assigned to the QRS class. The standard deviation of the R-R intervals was calculated. The area approximately 250mS prior to each QRS complex was evaluated for the presence of a P wave, based on cluster of pixels assigned to the P wave class. See Figure 5.1b for a visualisation.

If the number of QRS complexes preceded by a P wave was less than threshold X, and the standard deviation of R-R intervals was greater than threshold Y, the ECG was classified as AF. Thresholds X and Y were set using a brute force search on the validation set, where the combination maximising the F1 score was selected.

5.2.7 Mixed modality model

Segmentation masks were predicted for ECG images using the pretrained 2D U-Net model. A rule-based algorithm was used to read back the ECG signal. This employed a grid search method described by this group in a previous paper [22]. The extrapolated ECG signal was fed into a 1D ResNet encoder which made a diagnostic prediction. See Figure 5.1c for a visualisation.

5.2.8 Analysis

Sensitivity, specificity, positive predictive value and F1 score were calculated with respect to the AF and MI classes. The F1 score was used as the primary metric for comparing the models and testing hypothesis (1). Training loss curves were plotted. No additional statistical analysis was undertaken.

This experiment resulted in three sets of results for each of the two diagnostic tasks for the raw samples dataset:

1. Results from the non-pretrained model
2. Results from the model pretrained using wave segmentation
3. Results from enhanced pretraining

For the ECG image dataset, three additional sets of results were produced:

4. Results from a model initiated with random weights (as opposed to ImageNet weights) without WaSP
5. Results from the mixed modality model
6. Results from the rule-based AF detector

Segmentation masks for selected ECGs from the PTB-XL dataset were predicted at the end of pretraining. Another set of masks were predicted after fine-tuning the models. The segmentation masks predicted by the raw samples model were transposed into images for manual inspection.

A small subset of ECG images were printed and either photographed or scanned. Segmentation masks were predicted using the pretrained 2D U-Net model to evaluate robustness to real-world image artifact.

5.3 Results

5.3.1 Data

5.3.1a ECG generator

The source code for the ECG generator can be found here: https://github.com/docbrisky/WaSP-ECG

5.3.1b Synthetic dataset

Case studies of ECGs and segmentation masks produced by the ECG generator can be seen in Figures 5.2a-b.

5.3.1c Real dataset

Characteristics of the PTB-XL database are described by Wagner et al. [21].

5.3.2 Segmentation pretraining

Case studies of predicted segmentation masks for real ECGs can be seen in Figures 5.3a-e. This includes predicted segmentation masks for ECG images that were printed and either photographed or scanned, some with additional artifact added. The models pretrained exclusively on synthetic data were felt to generalise well to real-world ECG data. Robustness to image artifact was variable.

5.3.3 Fine tuning

Loss curves for 1D and 2D models can be seen in Figure 5.4a-d. Across modalities, models with randomly initialised weights converged more slowly than pretrained models. Among the 2D models, those with weights derived from non-enhanced WaSP converged more slowly than either models that had undergone enhanced pretraining or models that were initialised with ImageNet-derived weights.

5.3.4 Diagnostic classification

The sensitivity, specificity, positive predictive value and F1 scores for the diagnostic tasks can be seen in Tables 5.1-5.4.

Tables 5.1-4: Test set results for the two diagnostic classification tasks. Highest and lowest scores for each set of results are highlighted in green and yellow, respectively.

AF detection 1D	Random weights	WASP	Enhanced WASP
Sensitivity	0.659	0.764	0.854
Specificity	0.989	0.997	0.991
PPV	0.723	0.913	0.808
F1	0.689	0.832	0.830

MI detection 1D	Random weights	WASP	Enhanced WASP
Sensitivity	0.906	0.956	0.910
Specificity	0.840	0.831	0.915
PPV	0.763	0.763	0.858
F1	0.828	0.848	0.883

AF detection 2D	Random weights	Imagenet weights	WASP
Sensitivity	0.000	0.799	0.396
Specificity	1.000	0.988	0.992
PPV	0.000	0.846	0.809
F1	0.000	0.821	0.531
	Enhanced WASP	Mixed modality model	Rule-based model
Sensitivity	0.801	0.657	0.531
Specificity	0.948	0.968	0.958
PPV	0.897	0.921	0.519
F1	0.846	0.767	0.525

MI detection 2D	Random weights	Imagenet weights	WASP
Sensitivity	0.000	0.824	0.514
Specificity	1.000	0.951	0.965
PPV	0.000	0.905	0.891
F1	0.000	0.863	0.652
	Enhanced WASP	Mixed modality model	
Sensitivity	0.848	0.668	
Specificity	0.935	0.950	
PPV	0.881	0.884	
F1	0.864	0.761	

For the ECG images, the models that underwent enhanced pretraining achieved the highest F1 score in both AF and MI detection. For the raw samples, the enhanced pretrained model scored highest for MI detection. The unenhanced pretrained model scored highest for AF detection.

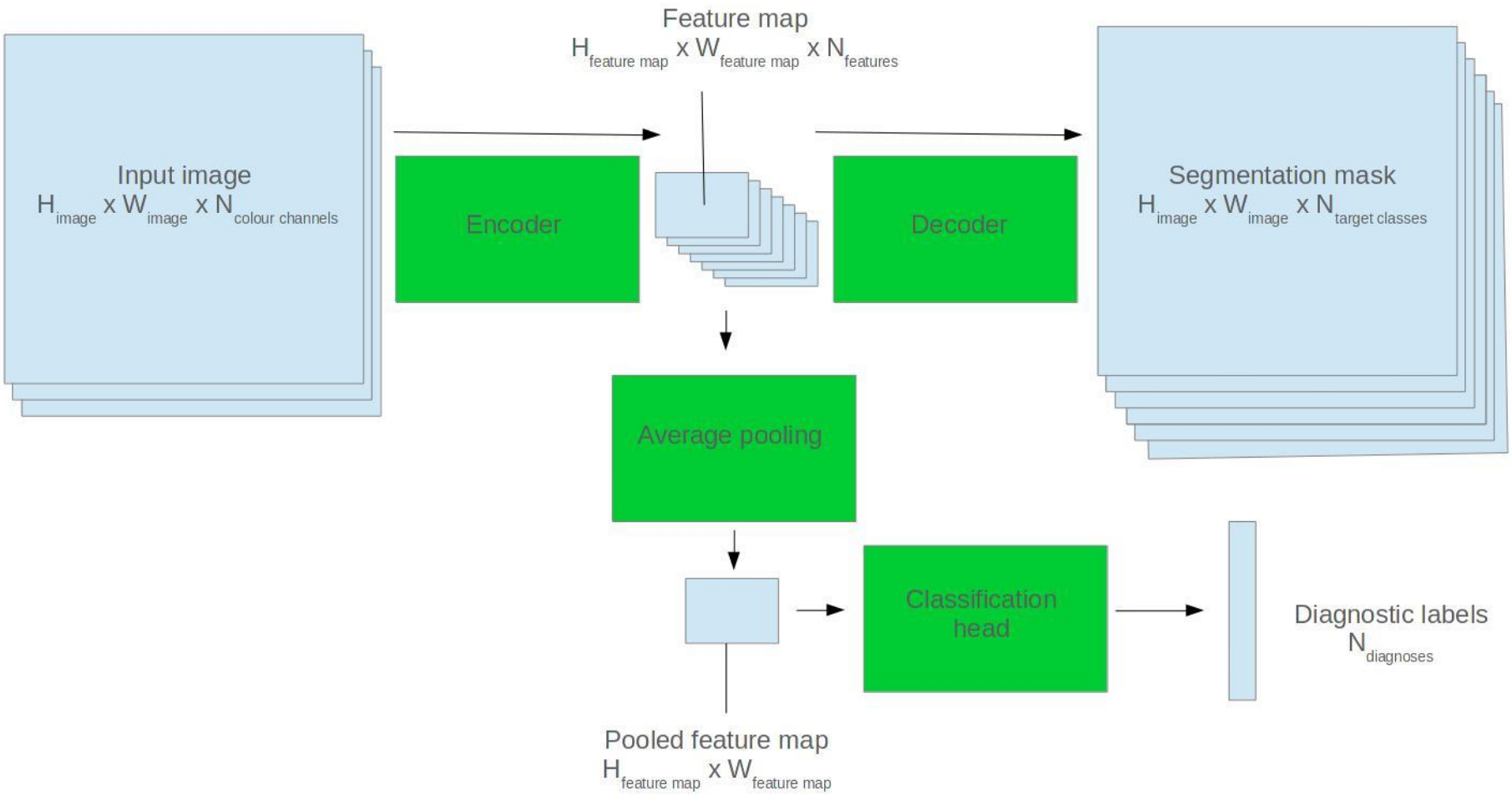


Figure 5.1a: Illustration of how a 2D U-Net model can be applied to segmentation and classification tasks

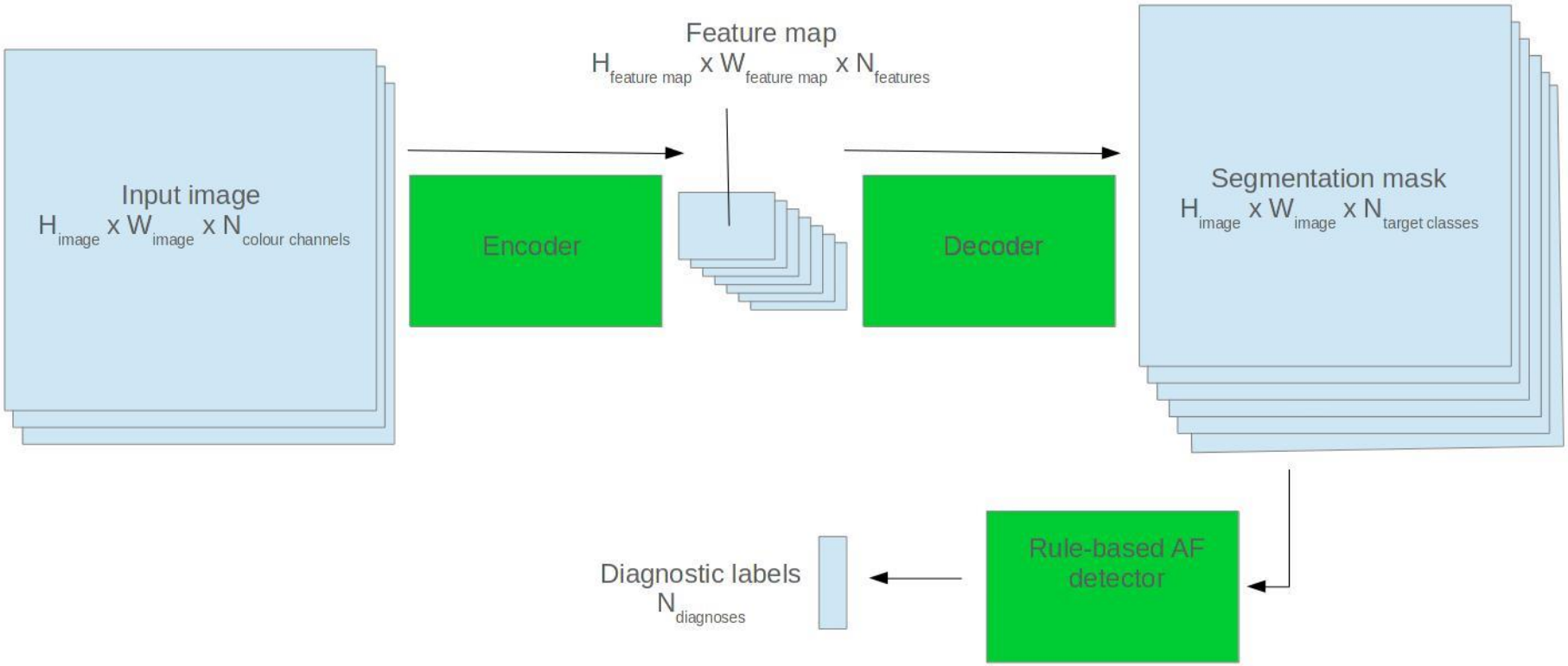


Figure 5.1b: Visualisation of the rule-based AF detector

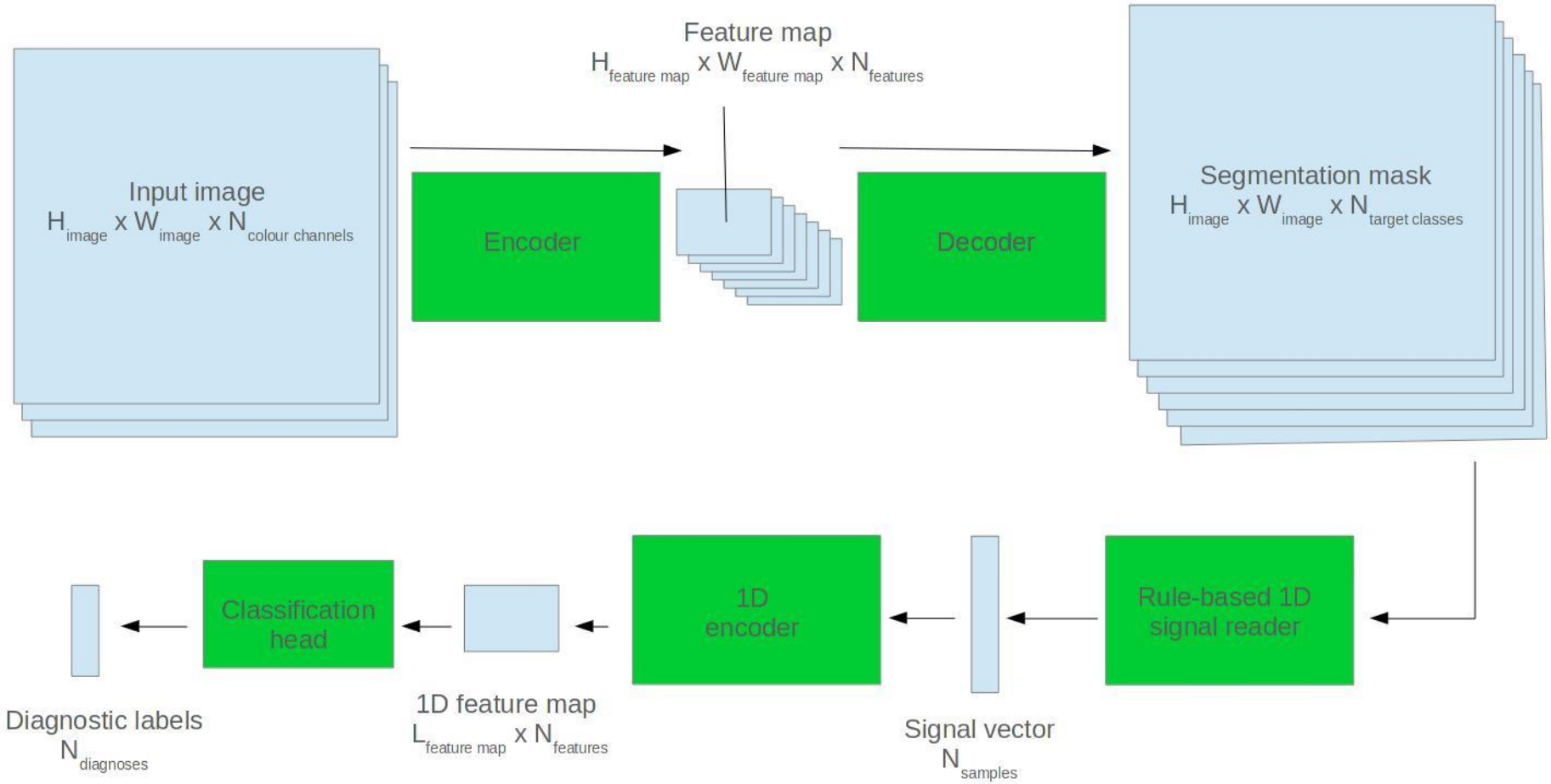


Figure 5.1c: Visualisation of the mixed modality analyser

Figure 5.2a: Synthetic ECG showing SR with anterior ST elevation

Figure 5.2b: The same image with the ground truth wave segmentation mask superimposed

Figure 5.3a: A segmentation mask for a randomly selected ECG signal from the PTB database. This mask was predicted for the raw ECG signal using a 1D U-Net. Both the raw signal and the segmentation mask were subsequently plotted into an image file. The model that predicted this mask was pretrained exclusively on synthetic ECG signals.

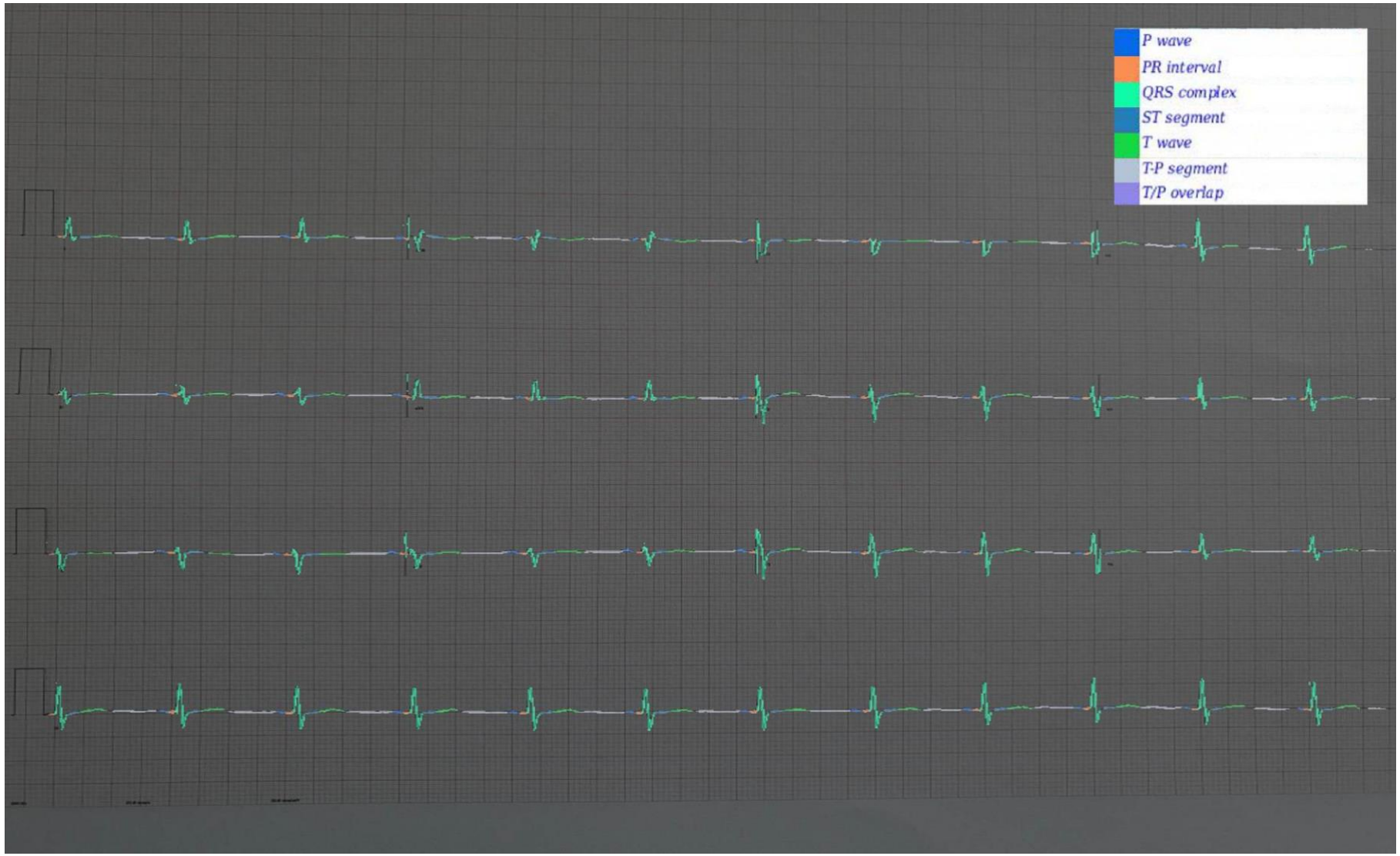


Figure 5.3b: Segmentation mask for a randomly selected PTB signal that was (i) plotted into an ECG image using the software developed for this experiment; (ii) printed using a standard desktop printer (HP Envy 4520 series); photographed using a Samsung Galaxy S10 mobile phone (flash off, bright daylight). The mask was then predicted by a 2D U-Net model that had been pretrained exclusively on synthetic data.

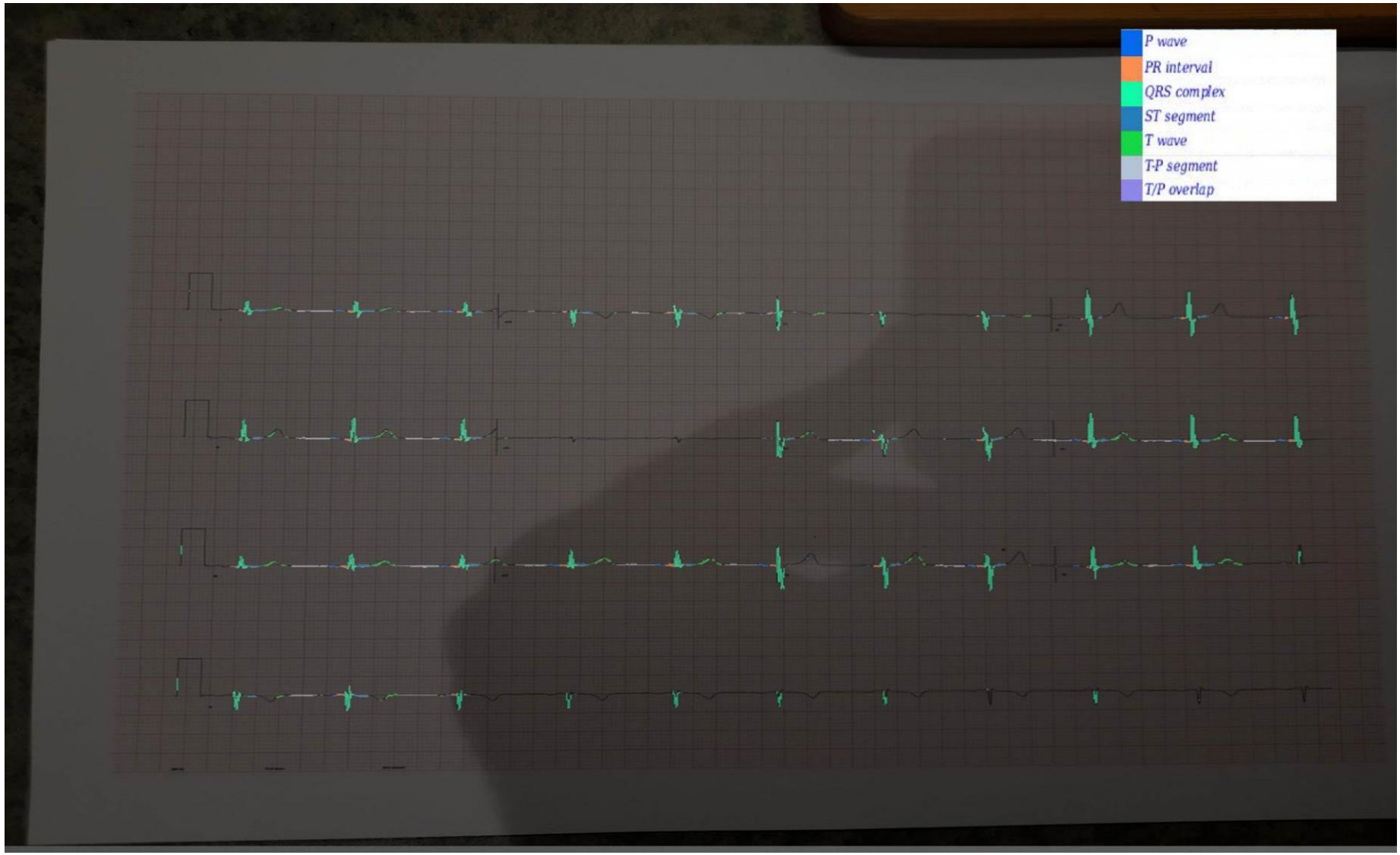


Figure 5.3c: This segmentation was produced using the same process as fig 5b, except that the photograph was taken in more challenging lighting conditions (at night, flash off, xenon strip lighting with shadows on image)

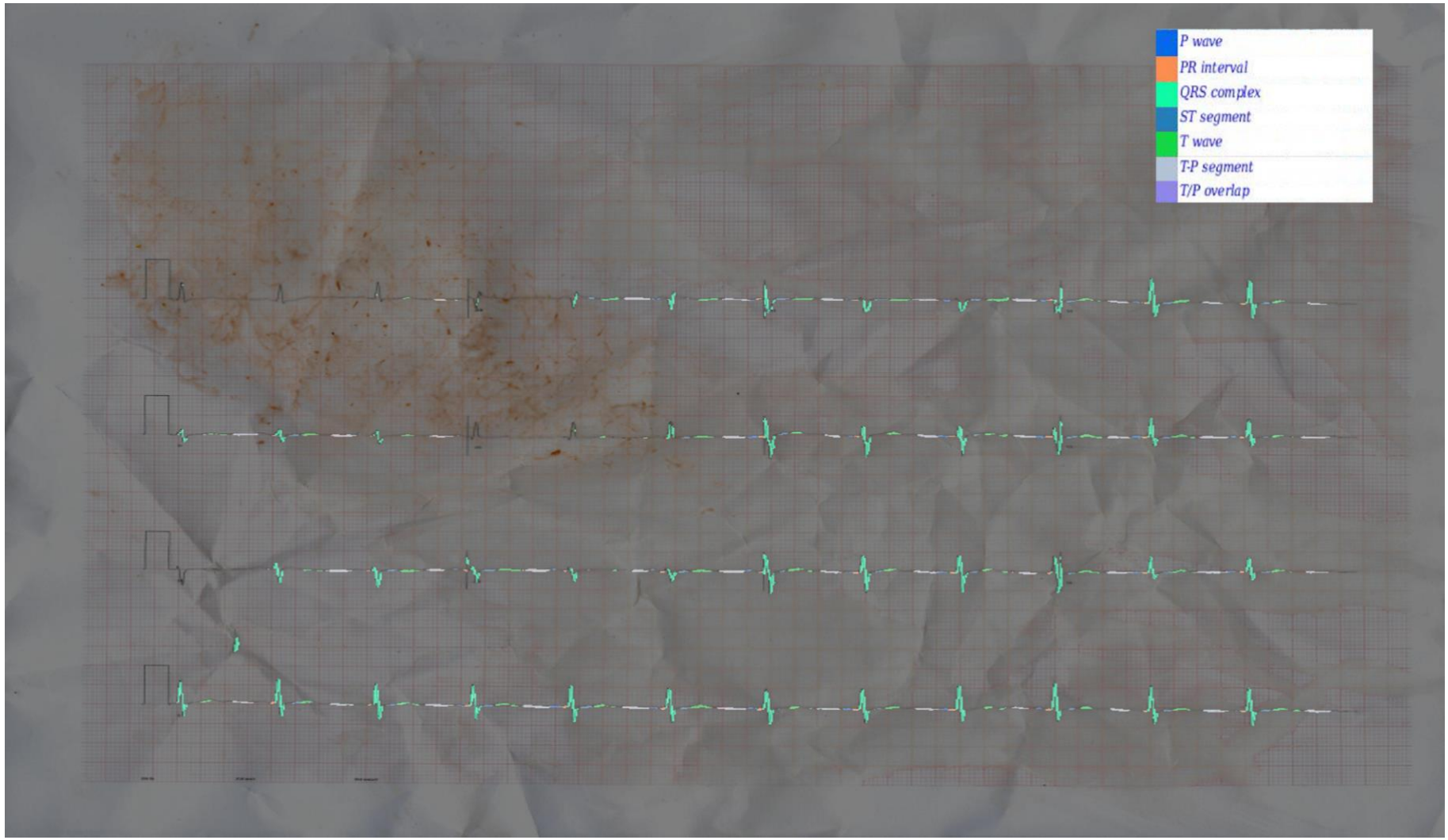


Figure 5.3d: This segmentation was produced using the same process as fig 5b, except that the printed ECG was (i) crumpled up; (ii) sprinkled with coffee; (iii) smeared with tomato sauce; (iv) scanned using an HP Envy 4520 desktop scanner (at 600DPI). This process was the result of a discussion about how to recreate a level of image artifact that might represent real-world clinical practice. Author DJM noted that he is regularly asked to review ECGs that have been stained with blood or coffee, and occasionally ECGs that have been thrown in the bin and subsequently retrieved.

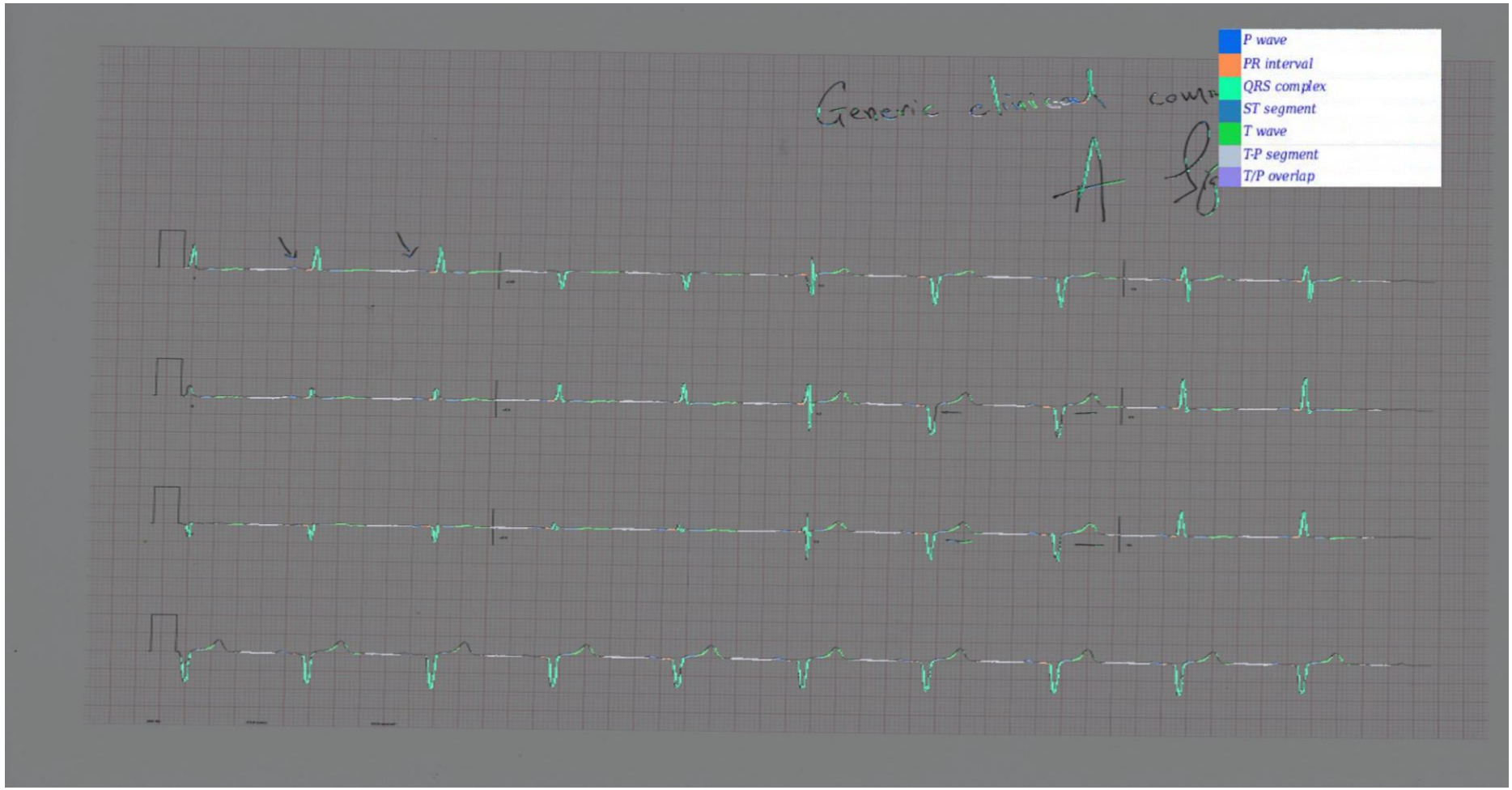
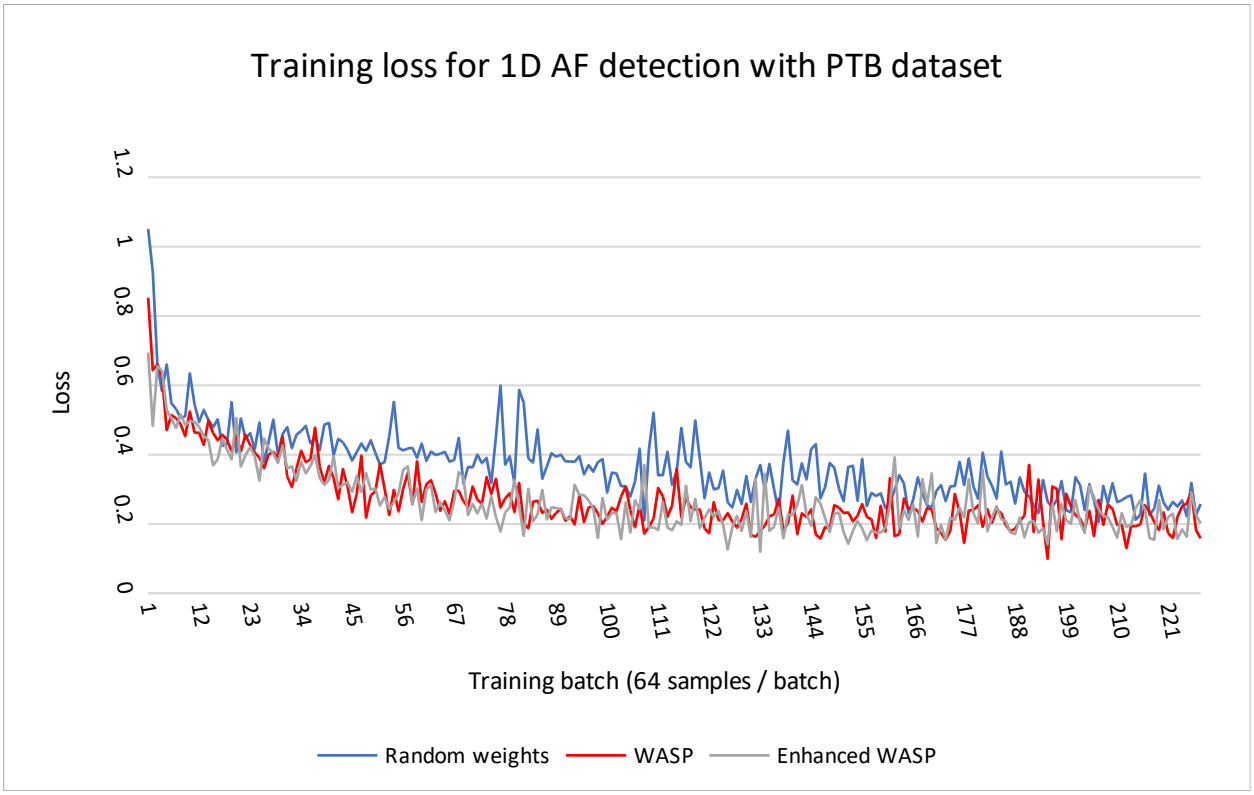


Figure 5.3e: This segmentation was produced using the same process as fig 5b, except that manual annotation artefact was added and the image was scanned using an HP Envy 4520 series scanned (at 600DPI).



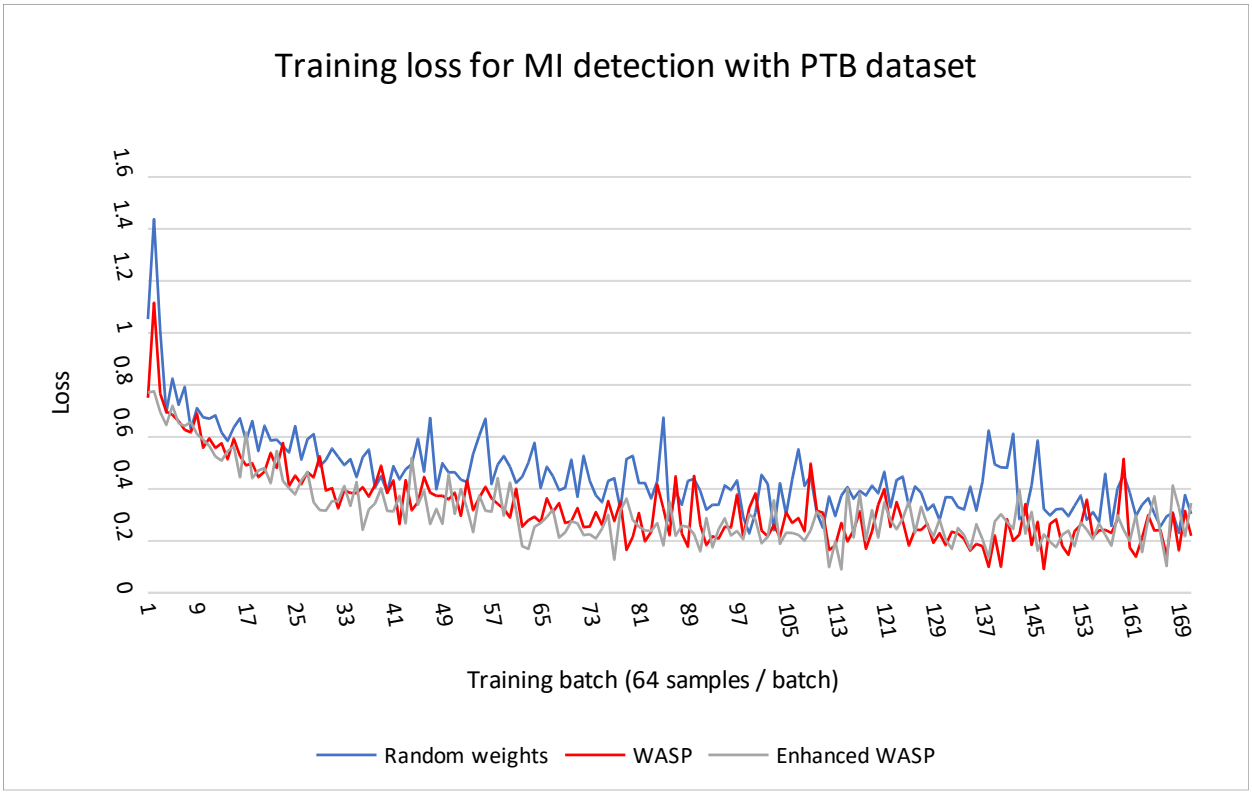
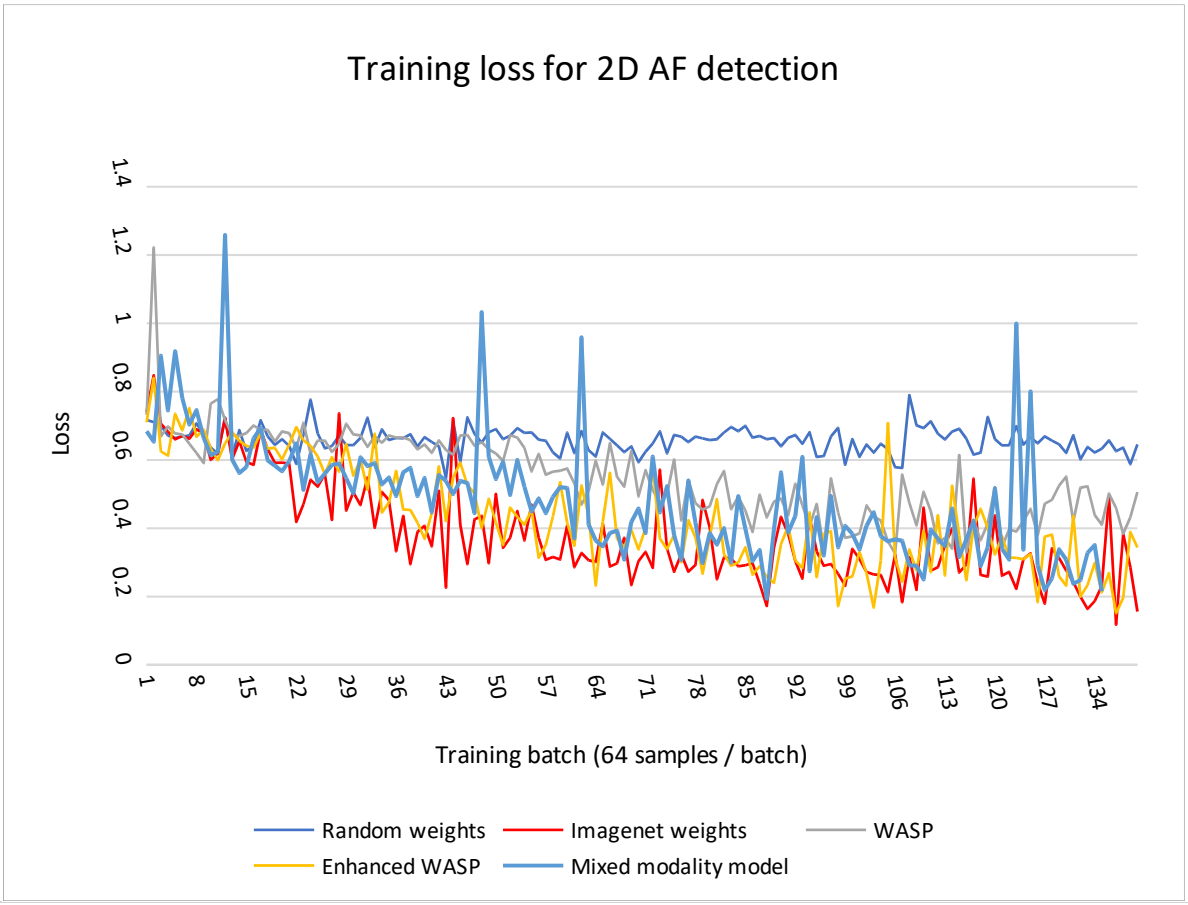


Figure 5.4a/b: Training losses for 1D models on diagnostic classification tasks with PTB ECGs. Each training run comprised a single epoch. WASP = WaSP



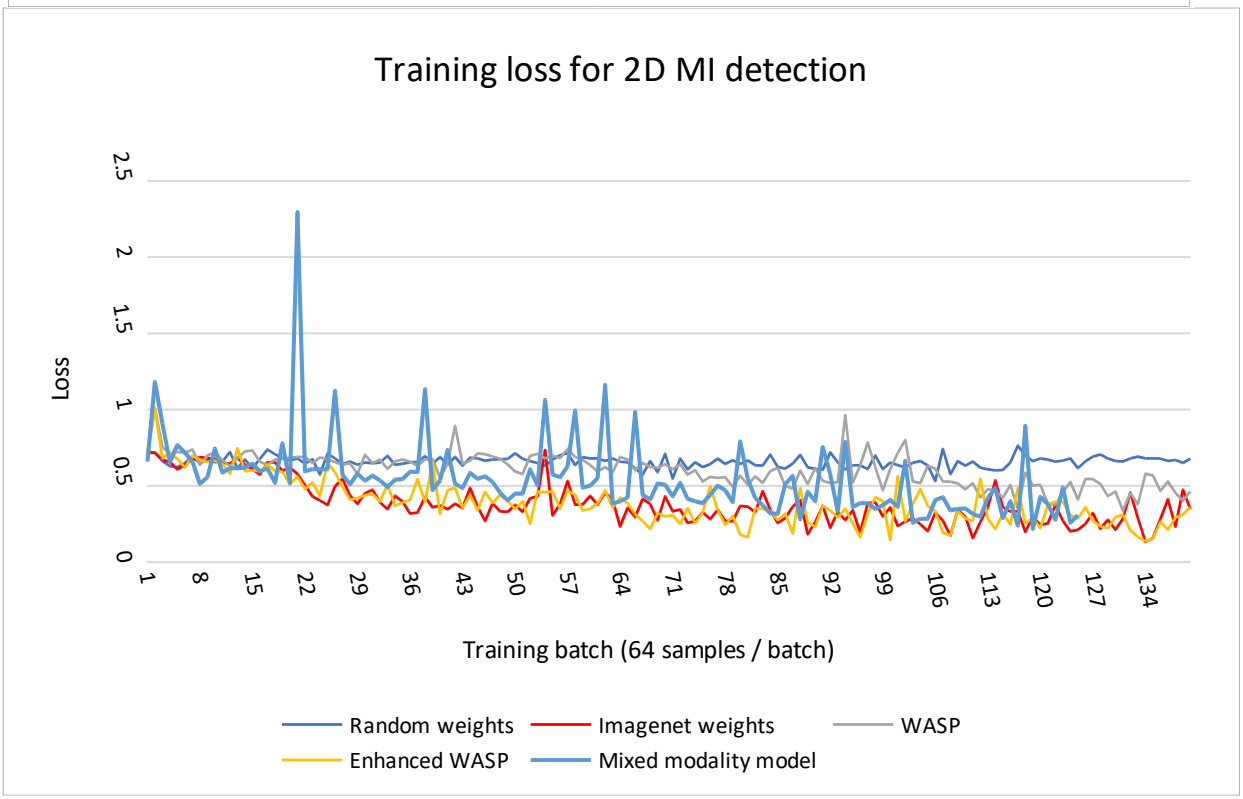


Figure 5.4c/d: Training losses for 2D models on diagnostic classification tasks with PTB ECGs.

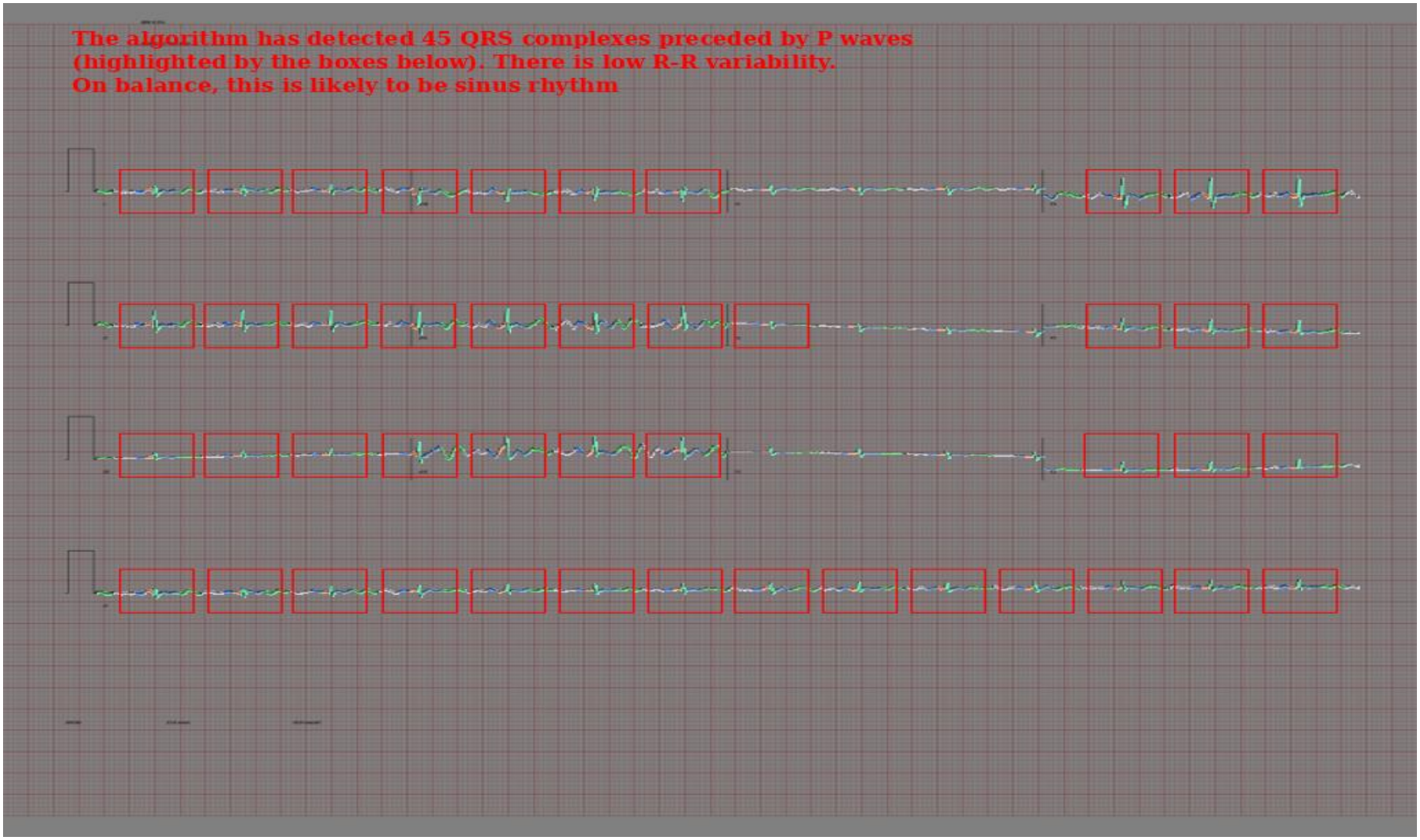


Figure 5.5: Output of the rule-based AF detection algorithm. The authors propose that this is highly explainable compared with end-to-end AI analysis.

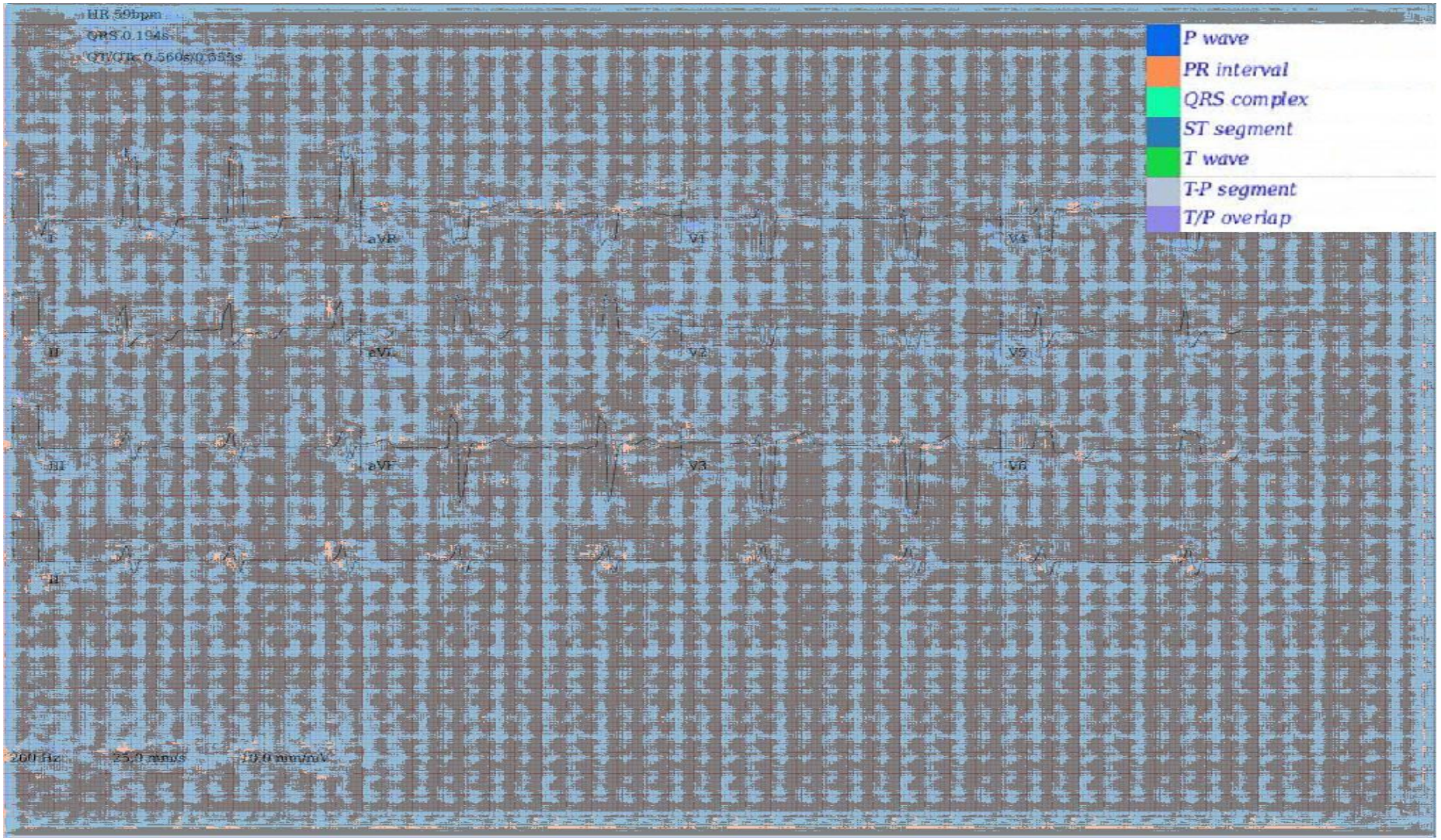


Figure 5.6: Segmentation mask produced by a 2D model initiated with ImageNet weights, but not having undergone any further training on ECG segmentation nor classification tasks.

For the ECG image set, the non-pretrained models predicted all samples as normal. Consequently, the sensitivity and positive predictive value for both models was zero. The rule-based AF detector scored lower than any pretrained model, although the F1 scores for the rule-based detector and the unenhanced pretrained model were close at 0.52 and 0.53, respectively. The mixed modality model outperformed the rule-based and unenhanced pretrained models, but underperformed the enhanced pretrained and ImageNet-trained models.

5.3.5 Confidence calibration and explainable outputs

In addition to Figures 5.3a-e, which show examples of segmentation masks, Figure 5.5 shows an example output from the rule-based AF classifier. Figure 5.6 shows a segmentation mask produced by a model that has been newly initialised with ImageNet weights. This model can be assumed to have no diagnostic capabilities with respect to ECG analysis. It is proposed that this segmentation mask would cause a clinician to place low confidence in the model's outputs, whereas the segmentation masks shown in Figures 5.3a-b may warrant relative high confidence. The segmentation masks in Figures 5.3c-e may alert the clinician to some issues caused by image artifact, and trigger additional caution when considering the model's final diagnostic output.

5.4 Conclusion

This study shows that WaSP using a synthetic dataset can improve training efficiency for downstream ECG tasks with real ECG data. The impact of pretraining was particularly marked with ECG image analysis. WaSP also enables meaningful intermediate output from the AI model.

The rule-based AF detection algorithm demonstrated a novel approach to ECG image analysis that benefits from advances in modern AI but is proposed to be highly explainable. Accuracy was limited but refinement of the technique may result in performance improvements.

Reading back signals from ECG image segmentation masks allowed a 1D classifier to detect both MI and AF with moderate accuracy. This shows that the SNR within the extrapolated data is high enough to facilitate some degree of downstream analysis. The motivation for investigating this is discussed under ‘future work’ below.

5.5 Discussion

5.5.1 Limitations

The diagnostic tasks chosen for this study are not representative of the spectrum of clinical ECG phenotypes encountered in real-world practice. The absolute results from these tasks add

little to the field; rather, it is intended that the relative results serve as an early evaluation of WaSP and of pretraining with synthetic ECG data. More work is needed to determine whether the findings of this study would generalise to a wider range of diagnostic problems.

One of the stated motivations for investigating WaSP was that it may facilitate clinician confidence calibration. The figures shown in this study may enable readers to begin forming their own conclusions on this matter. However, this hypothesis was not formally evaluated and can be considered unproven to date.

This study was undertaken in a retrospective observational setting. A single dataset was used for training, testing and validation. There is an increased risk of over-fitting a particular data distribution in this context. Results shown here may not generalise to other datasets or populations.

For the diagnostic classification evaluation, ECG images were plotted directly from the signals *in silico*. In a clinical setting, ECG images would be printed and either scanned or photographed. This would introduce image artifact that may alter the accuracy of downstream tasks, as illustrated in Figures 5.3d-e. For any future work aiming to establish whether the novel image-based techniques described here are useful for downstream clinical applications, it is likely that the full evaluation would need to be conducted with paper ECGs.

The rule-based AF detector was only evaluated with ECG images and not raw samples as this would have required a substantial re-write of the application, which was not felt to be warranted as there are already many rule-based AF detection algorithms for raw sample data.

5.5.2 Comparison with existing approaches

As discussed during the introduction section, approaches to both pretraining and explainable DL for ECG analysis have been explored by other groups. To the best of the author's knowledge, however, this is the first demonstration that pretraining with synthetic data is effective.

This has potentially significant implications for the fast-growing field of ECG AI. The increasing number of large public ECG databases like PTB-XL is helping to drive research in this field. However, such databases are finite and may be subject to bias: centres with the expertise and resources to produce such datasets tend to exist in more affluent global regions and may over-represent certain demographic groups; rare diseases and paediatric conditions are often under-represented in such biobanks [38]; studies from patients suffering with emerging diseases that may have cardiac involvement (e.g. COVID-19) may take some time to reach these datasets.

Knowledge-based engineering of synthetic datasets allows much greater control over the distribution of covariates-of-interest within the training data. This can help to counterbalance

bias and to increase the occurrence of rare but important features. It can also facilitate the creation of much larger datasets than would be possible using real patient data. Historically, supplementing labelled datasets with synthetic samples for task-specific training has been problematic [39]. For learning general representations during pretraining, however, it is proposed that a lower fidelity is acceptable: the model will learn additional or altered features that occur in real-world datasets during fine-tuning.

5.5.3 Additional points of interest

ECG image models pretrained on ImageNet performed significantly better than models initialised with random weights. This implies that some features learned from analysing photographs of real-world scenes transfer well to ECG analysis.

Performance worsened when the ImageNet-trained models underwent non-enhanced WaSP. A possible explanation for this is that WaSP caused catastrophic forgetting. It may be possible to overcome this issue by freezing early convolutional layers and reducing the learning rate [40].

Enhanced WaSP involved the addition of a diagnostic labelling task in addition to wave segmentation; the model was asked to output both types of label for each sample using an approach known as ‘multi-task learning’. This seemed to improve performance significantly compared with non-enhanced WaSP. The same black box nature of AI that was one of the motivating factors for this study makes it difficult to ascertain exactly why this was the case. However, the authors posit that the addition of a diagnostic label for the whole ECG forced the model to learn about relationship between more distant parts of the ECG (for example, the diagnosis of left bundle branch block requires that the model evaluate the QRS-T morphology in multiple leads simultaneously), whereas wave segmentation can be achieved by leveraging only very local parts of the data.

5.5.4 Relevance of this work to the wider field

As state-of-the-art AI models grow in size and complexity, more training data is required to capitalise on their increased pattern recognition capabilities [41]. In this study, WaSP expedited convergence during fine-tuning and produced higher results after a single training epoch. Therefore, WaSP can reduce the need for labelled training to produce equivalent results. This approach may allow larger AI model architectures to be used for ECG tasks where there would otherwise be insufficient labelled training data.

Explainable AI is an active research topic in healthcare [42]. Mechanisms by which clinicians can calibrate confidence or review decision logic may provide key to adoption of AI in practice. The work undertaken for this study may catalyse future research into segmentation masks as a mechanism for confidence calibration in ECG analysis, and mixed AI and rule-based analysis as a mechanism for explainable ECG image analysis.

The code base for this experiment has been published under a permissive open source license. The application has been named WaSP-ECG. The intention is to facilitate reproduction of results and accelerate future research in the field. The inclusion of Zero optimisation functionality in the code base [43] allows researchers to train larger models on their existing infrastructure than would have otherwise been possible, or to use higher resolution input data. This may allow researchers to extend existing AI techniques and improve model performance.

5.5.6 Future work

Only two rhythm types and six morphological phenotypes were simulated during the enhanced pretraining phase of this study. Given the performance improvement observed with enhanced pretraining over non-enhanced WaSP in the context of ECG images, it may be that a wider repertoire of simulated ECG phenotypes would further improve downstream performance.

The robustness of AI techniques to image artifact (see Figures 5.3d-e) was felt by the authors to be limited. The ability to photograph ECG images on a mobile phone and upload for cloud-based analysis is proposed to be a worthwhile goal, as it would decrease the dependence on hardware-bound analysers. This, in turn, would allow for more agile development of novel ECG applications and easier integration with multi-model clinical data, such as symptomatology, biochemical results, cardiac imaging, etc. There is an emerging body of evidence that fusing multi-modal data leads to improved performance of medical AI systems [44]. For this reason, investigating approaches to improve robustness to image artifact, challenging lighting conditions, etc. may be a valuable research avenue.

Evaluating WaSP for diagnostic tasks more representative of real-world clinical problems would be a key next step for the line of investigation presented in this study. The use of data from additional patient populations and evaluation of diagnostic capabilities in a prospective setting would help to establish the generalisability of the results presented here.

5.6 References

[1] Macfarlane PW, Van Oosterom A, Pahlm O, Kligfield P, Janse M, Camm J. Comprehensive electrocardiology. Springer Science & Business Media; 2010.

[2] Rautaharju PM. Eyewitness to history: Landmarks in the development of computerized electrocardiography. J Electrocardiol 2016;49(1):1-6.

[3] Kashou AH, May AM, Noseworthy PA. Artificial intelligence-enabled ECG: a modern lens on an old technology. Curr Cardiol Rep 2020;22(8):1-8.

[4] Svensén M, Bishop CM. Pattern recognition and machine learning. 2007.

[5] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

[6] Rusk N. Deep learning. Nature Methods 2016;13(1):35-35.

[7] Luo S, Johnston P. A review of electrocardiogram filtering. J Electrocardiol 2010;43(6):486-496.

[8] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature communications 2020;11(1):1-9.

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012;25:1097-1105.

[10] Makimoto H, Höckmann M, Lin T, Glöckner D, Gerguri S, Clasen L, et al. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. Scientific reports 2020;10(1):1-9.

[11] Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology 2021;18(7):465-478.

[12] Lopez-Jimenez, F., Attia, Z., Arruda-Olson, A. M., Carter, R., Chareonthaitawee, P., Jouni, H., Kapa, S., Lerman, A., Luong, C., Medina-Inojosa, J. R., Noseworthy, P. A., Pellikka, P. A., Redfield, M. M., Roger, V. L., Sandhu, G. S., Senecal, C., & Friedman, P. A. (2020). Artificial Intelligence in Cardiology: Present and Future. Mayo Clinic proceedings, 95(5), 1015–1039.

[13] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65-69.

[14] Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial

fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet 2019;394(10201):861-867.

[15] Arsene, C.T., Hankins, R. & Yin, H. 2019, "Deep learning models for denoising ECG signals", 2019 27th European Signal Processing Conference (EUSIPCO)IEEE, 2019;1.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA. pp. 6000–6010.

[17] Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digital Medicine 2021;4(1):1-3.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 2020.

[19] Yan, G., Liang, S., Zhang, Y., & Liu, F. (2019). Fusing transformer model with temporal features for ECG heartbeat classification. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM):898-905.

[20] Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vij, S., et al. (2020). A wide and deep transformer neural network for 12-lead ECG classification. 2020 Computing in Cardiology:1-4.

[21] Wagner P, Strodthoff N, Bousseljot R, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;7(1):1-15.

[22] Brisk R, Bond R, Banks E, Piadlo A, Finlay D, McLaughlin J, et al. Deep learning to automatically interpret images of the electrocardiogram: Do we need the raw samples? J Electrocardiol 2019 Nov - Dec;57S:S65-S69.

[23] Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K. Explainable AI: interpreting, explaining and visualizing deep learning. : Springer Nature; 2019.

[24] Bond RR, Novotny T, Andrsova I, Koc L, Sisakova M, Finlay D, et al. Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. J Electrocardiol 2018;51(6):S6-S11.

[25] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798-1828.

[26] Komodakis, N., & Gidaris, S. (2018). Unsupervised representation learning by predicting image rotations. International Conference on Learning Representations (ICLR)

[27] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A 2017 Mar 28;114(13):3521-3526.

[28] Sarkar, P., & Etemad, A. (2020). Self-supervised ECG representation learning for emotion recognition. IEEE Transactions on Affective Computing

[29] Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B. CEFEs: A CNN Explainable Framework for ECG Signals. Artif Intell Med 2021;115:102059.

[30] Jo Y, Cho Y, Lee SY, Kwon J, Kim K, Jeon K, et al. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. Int J Cardiol 2021;328:104-110.

[31] Pintelas E, Liaskos M, Livieris IE, et al. A novel explainable image classification framework: Case study on skin cancer and plant disease prediction. Neural Computing and Applications. 2021;33:15171-89.

[32] Jackson P. Introduction to expert systems. 1986.

[33] Ronneberger, O., Fischer, P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015;234-241.

[34] Hu, J., Shen, L., & Sun, G. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018;7132-7141.

[35] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 2020;53(8):5455-5516.

[36] Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009;248-255.

[37] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., et al. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019;92-100.

[38] Kim P, Milliken EL. Minority participation in biobanks: An essential key to progress. Biobanking. 2019:43-50.

[39] Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S. N., & Chellappa, R. Learning from synthetic data: Addressing domain shift for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018;3752-3761.

[40] Shmelkov, K., Schmid, C., & Alahari, K. Incremental learning of object detectors without catastrophic forgetting. Proceedings of the IEEE International Conference on Computer Vision, 2017;3400-3409.

[41] Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. JMIR medical informatics 2019;7(3):e14830.

[42] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making 2020;20(1):1-9.

[43] Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. Zero: Memory optimizations toward training trillion parameter models. SC20: International Conference for High Performance Computing, Networking, Storage and Analysis:2020;1-16.

[44] Huang S, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine 2020;3(1):1-9.

Chapter 6:

Discussion and conclusions

6.1 Introduction

At the outset of the thesis, modern AI was presented as a technology with the potential to reduce the burden of human error in healthcare. ECG analysis was proposed as an appropriate lens through which to investigate some of the salient issues within this field. Chapter 2 provided an overview of the ECG, AI, and DL for ECG analysis. Outstanding areas identified for further investigation included DL for ischaemia detection, and two key translational research questions. These latter topics were: the use of DL for ECG image analysis, and the application of AI to rarer diagnoses where there is a relatively paucity of labelled data. The interpretability of DL-based applications was raised as an important issue in chapter 3, and became a third translational theme during chapters 4 and 5.

Chapters 2-5 described investigations of specific aspects of:

- DL for AMI detection
- DL for ECG image analysis
- RL as a means to develop more explainable ECG AI, particularly for applications where there is a paucity of labelled data.

This final chapter aims to review the key conclusions of the presented in previous sections, propose how they may impact future research, and discuss their implications with respect to their broader field of AI in healthcare. A note will also be made of related work not detailed in the previous chapters.

6.2 Review of conclusions from previous chapters

6.2.1 DL for AMI detection

As previously noted, AMI detection could be a particularly impactful application for DL. STEMIs are the type of AMI where PPCI has proven morbidity and mortality benefits [1]. In real-world practice, the diagnosis of STEMI is often delayed or missed altogether, for a variety of reasons that largely stem from human fallibility exacerbated by complex clinical problems and environments [2]. It has been proposed that these are exactly the kind of circumstances where AI can help by automating complex diagnostic processes [3].

Chapter 2 presented a systematic literature review on the use of DL for AMI detection from ECG data. It concluded that relatively little work had been done in that field at the time of the review, although it should be noted that this is a fast moving research field and a number of works have been published in this vein since, such as [4-7].

Chapter 3 described an original research work that investigated an approach to hyperacute detection of AMI using DL. It was hypothesised that latent space representations learned from an arrhythmia detection task would transfer across to ischaemia detection and allow for

effective training with a small cohort. This ‘transfer learning’ approach follows very similar principles to RL, which was discussed extensively in chapter 5. On the initial iteration of the experiment, it appeared that the DL model was able to reliably detect hyperacute infarction. However, a further iteration produced results in keeping with a random chance classifier. It was concluded that the first model had learned features based on confounding elements within the data. Specifically, it was proposed that the model learned to detect background electrical noise associated with cardiac theatres, which bore a correlative rather than causative relationship to the experiment’s endpoint.

The contribution of this result to the body of research on DL for AMI detection was relatively small. However, it did serve as an illustration of the broader challenges and dangers around the use of DL methods in the clinical setting. In particular, it highlighted the difficulties posed by the relatively uninterpretable logic processes employed by end-to-end DL, where both the features and the decision logic are learned with no direct input from human experts. In the presence of a ‘data leak’, the DL model can learn features that allow it to produce compelling results using spurious logic. It can be very difficult for a human user of such an application to detect this phenomenon. The experience described in chapter 3 led to an increased emphasis on mechanisms to promote interpretability during subsequent chapters.

6.2.2 DL for ECG image analysis

Chapter 4 described an experiment designed to test the hypothesis that DL methods may allow an ECG image analysis application to attain results comparable with raw sample analysis. The PAFC dataset was selected because it is a large, high quality labelled dataset and also because a number of other groups have published the results of their raw sample applications on this data. An application was developed to transpose raw samples into ECG images. A DL-based application was then developed to infer the raw samples (or, more accurately, scaled and downsampled equivalents) from the ECG images, and to predict a diagnostic class for each ECG. This application was able to produce results comparable with raw sample-based applications developed by other groups.

The introduction to this chapter also outlined the role that ECG image analysis may play in making DL-based ECG analysers more interpretable. Namely, transposing raw samples to an image file provides a form of ‘feature filtering’. The focus of the rest of the chapter was on investigating the potential of DL methods to improve performance in ECG image analysis. However, having obtained results that suggested DL methods do indeed have a role to play in this field, the following chapter returned to this theme.

6.2.3 RL to help address data paucity

Chapter 5 introduced the concept of RL as a means to guide DL models to learn generalisable latent representations of data features using general purpose, often self-supervised, tasks. These

models can then be fine-tuned using smaller labelled datasets. In the domain of NLP, where transformer models can be pretrained on vast quantities of unlabelled data and learn very effective feature representations, the term ‘few shot learners’ [8]. This term refers to the fact that, following extensive RL-based pretraining, these models may only need a very small number of labelled data points to ‘get the hang’ of a new downstream task.

In other domains such as NLP and image processing, RL commonly employs a ‘fill in the blank’ task. In the case of masked language models, this means removing one or more words in a sentence or paragraph and tasking the model predicting the word that has been removed [9]. Alternatively, autoregressive language models are trained to predict multiple missing words at the end of a sentence [10]. Image transformers take a similar approach, whereby a patch of pixels are removed and must be predicted back in [11]. All of these methods allow for very large scale transforming without the need for any manual data labelling.

The regularly repeating nature of ECG signals does not lend itself well to this approach. For example, if one was to take a single-lead ECG recording and blank out a QRS complex, this complex would be identical to most others within the recording in the vast majority of cases. Thus, the task becomes trivial for a sufficiency powerful DL model. Other researchers have addressed this challenge by employing alternative self-supervised methods, whereby the ECG signals are augmented and the original recordings must be matched with their augmented counterparts [12]. This is a form of ‘pretext invariant’ RL (PIRL), which has been used effectively in a number of contexts [13]. The major downside to PIRL within the context of chapter 5 was it offers little in terms of mechanisms for making the DL model more interpretable. Hence, WaSP was proposed as a novel approach for RL in the context of ECG analysis.

The quantitative results of the experiment described in chapter 5 showed that WaSP was effective in the context of both ECGs as raw samples and for ECG image analysis for AF and AMI detection, particularly when combined with multi-task learning to predict both rhythm and morphology abnormalities (‘enhanced WaSP’). This is not to say that WaSP will work well for every downstream diagnostic task, but good performance on both arrhythmia detection and AMI detection does point to a certain level of generalisability.

Chapter 5 also proposed mechanisms to make the DL application more interpretable to clinicians and to enable confidence calibration. The wave segmentation mask was proposed as a way to sanity check the features learned and used by the DL model and minimise the risk of confounding data features being leveraged unnoticed. The ability to use the DL model for feature extraction and then to undertake downstream diagnostic classification using rule-based methods was demonstrated via the AF detection algorithm. However, a major limitation of the study was that these mechanisms were proposed but not evaluated.

6.2.4 Different methods for ECG image analysis

The use of synthetic data as part of the WaSP workflow facilitated the generation of a very large dataset. This allowed for an alternative approach to ECG image analysis, whereby the whole image was fed to a CNN. In chapter 4, conversely, the ECG image analyser relied on rule-based extrapolation of raw sample equivalents from the ECG image. This substantially reduced the dimensionality of the training data for the DL model, such that the model was effectively trained on 1D data. Dimensionality reduction is known to avoid issues such as overfitting on small datasets [14].

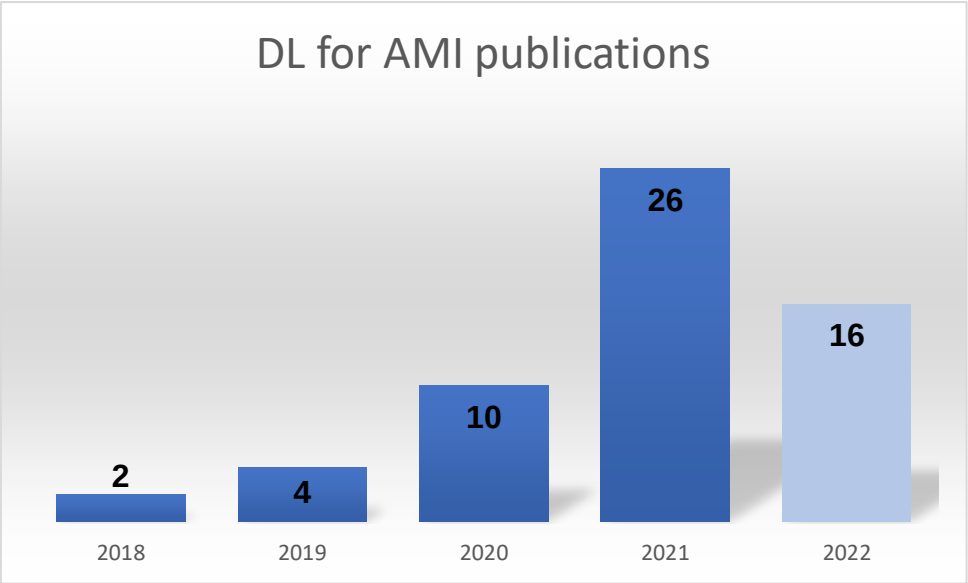
The other challenge to training the CNN directly on image files in the chapter 4 experiment was the inability to use transfer learning. ECG images produced from continuous single lead recordings have unusual dimensions (i.e. they are very wide and very short). Whereas the experiment described in chapter 5 initiated the 2D CNNs with weights derived from ImageNet pretraining, the unusual dimensions of the ambulatory ECG images in chapter 4 was felt to preclude this.

The possible advantage to training DL models directly with ECG images, as opposed to using extrapolated signals, is that end-to-end DL methods can be particularly effective at denoising images [15]. In real-world clinical practice, the digitisation of paper ECGs can introduce a lot of noise and artefact [16]. This may lead to poor quality extrapolated signals using the rigid, rule-based approach described in chapter 4. The direct-from-image DL training method employed in chapter 5, on the other hand, may be more robust under these circumstances.

6.3 Implications for future ECG AI research

6.3.1 DL for AMI detection

As previously noted, this has been an active research field over the last few years. Figure 6.1 shows an approximate quantification of this trend.



PUBMED SEARCH TERMS: ((((artificial intelligence[Title/Abstract]) OR (deep learning[Title/Abstract])) AND ((electrocardiogram[Title/Abstract]) OR (ECG[Title/Abstract])) AND ((ischaemia[Title/Abstract]) OR (ischemia[Title/Abstract]) OR (myocardial infarction[Title/Abstract]))) AND (("2018/01/01"[Date - Publication] : "3000"[Date - Publication])))

Figures 6.1a & 6.1b: approximate trend in the rate of publications on ECG AI for ischaemia detection over the last 3-4 years. Note this search was undertaken on the 14th May 2022. Assuming an even distribution of publications over a 12 month period, the strong upwards trend is set to continue beyond 2022.

In terms of a more rigorous and detailed update: in 2021, Al Hinai et al. published a systematic literature review, broadly similar to the one presented in chapter 2 [17]. Whereas the chapter 2 review considered DL methods exclusively for the detection of myocardial ischaemia, the scope of the 2021 review extended to myocardial dysfunction more broadly, which notably included myocardial hypertrophy in addition to ischaemia. Those authors also limited inclusion criteria to studies that used end-to-end DL, rather than DL only for feature extraction, and to studies that used resting 12 lead ECGs. Both reviews used the PRISMA guidelines.

Al Hinai et al. identified six recent original research publications that employed DL for AMI detection and met the inclusion criteria. The results reported from these studies showed high sensitivity and specificity compared with conventional methods, leading the authors to conclude that DL technology is starting to show significant promise for AMI detection. CNNs were the dominant DL model architecture, in keeping with the trend identified in the introductory chapter of this thesis. In terms of future research, the two main challenges identified by the more recent systematic review were (1) interpretability of DL in this context and (2) a paucity of high quality training data.

As noted earlier, the results of the experiment described in chapter 3 offered limited insight into the broader ability of DL to detect AMI from ECG signals. Under those specific

experimental conditions, the DL approach employed in that study was unable to predict accurate diagnostic labels. That is not to say that the results would be the same under different conditions, or with a different DL approach.

However, the experiment was not without implications for future research in this field, nor beyond. Indeed, the work in this PhD was the subject of an editorial article in the EHJ Digital Health, which highlighted its importance for the clinical readership [18]. The key finding cited by the editorial was that the DL training and evaluation workflow used in the study, widely used within other medical AI research, offered little or no effective mechanism for detecting confounding data features learned by the DL model. This illustration of the need for greater interpretability of DL-enabled medical applications corresponded well with the subsequent findings of Al Hinai et al. Nonetheless, the implications of that experiment for future research specifically into DL for AMI detection are not held as the major contribution of this thesis.

6.3.2 DL for ECG image analysis

The results of the experiments described in chapters 4 and 5 suggest that DL methods may allow for more accurate ECG image analysis than conventional, non-ML methods. The motivation for further work in this field is discussed in those chapters. Broadly, there are three strands to the argument for a greater focus on ECG image, as opposed to raw sample, analysis.

1. Advances in ECG image analysis techniques can be applied directly to historical ECG data, which generally exists in paper form. To date, applying state-of-the-art analytics to these data has meant first recovering the raw samples from ECG images. Non-ML methods have not proven sufficiently robust for widespread adoption of this approach.

2. By the same token, advances in ECG image analysis can be potentially leveraged at the point of care by any clinician with access to a paper ECG and a mobile phone equipped with a camera. Conversely, access to state-of-the-art raw sample analysis is generally contingent upon owning expensive state-of-the-art ECG hardware and/or having the specification and computer program to parse proprietary file formats. Applying DL methods to ECG image analysis in addition to raw sample analysis can thus be seen as ‘democratising’ access to AI.

3. As discussed during chapters 4 and 5, transposing ECG samples to images can provide additional mechanisms to promote interpretable AI. These include the ability to visualise learned features in a human readable format prior to downstream analysis. The act of transposing an ECG signal into an ECG image may also act as a form of feature filtering to reduce the risk of confounding data features.

6.3.3 ECG image analysis using sample recapture

The work presented in chapters 4 and 5 is proposed to have several implications for future research into ECG image analysis. Firstly, the experiment in chapter 4 used non-ML methods

to recapture raw sample equivalents from an ECG image and analysed these 'extrapolated samples' using a DL model. The results suggested that the ability of this application to differentiate between normal sinus rhythm and AF was equivalent to applications trained directly on raw samples. This is held to be the first peer-reviewed experiment investigating the plausibility of sample recapture from ECG images followed by DL analysis. Other authors have since pursued a similar approach, only with DL methods incorporated into the sample recapture method, in addition to a separate DL model for diagnostic classification [19]. Li et al. investigated an end-to-end DL approach to sample recapture and diagnostic classification. They noted that the multi-stage approach described in chapter 4 has distinct advantages by comparison, and cited the write-up of the chapter 4 experiment as an alternative approach in their discussion [20].

Further research in this direction is felt to be warranted, both to establish the extent to which DL analysis of recaptured samples is effective across different ECG abnormalities, and to investigate whether this approach is robust enough to handle real-world data in a prospective setting. The advantage of the sample recapture approach, as opposed to the direct-from-image DL training described in chapter 5, is that sample recapture results in a 1D signal vector that is approximately a scaled version of the original raw sample recordings. Most emerging DL-enabled ECG analysers expect data in this format, and it is plausible that these models could be fine-tuned to analyse extrapolated samples relatively easily. They could not, on the other hand, be fine-tuned to handle 2D image files. Therefore, ECG image analysis using the sample recapture method may be the most effective way of applying state-of-the-art ECG analysis to historical data.

6.3.4 Direct-from-image analysis

Chapter 5 highlights a number of themes that may be promising avenues for future research. Firstly, the results for both AMI and AF detection obtained by DL models trained with ECG images suggests that direct-from-image ECG analysis using DL methods is a viable approach. At the time the experiment described in chapter 5 was undertaken, there was no known precedent for training DL models directly using 12-lead ECG images and this was held to be a first-in-field result. While corresponding article was being prepared for publication, however, Anwar et al. published a study investigating direct-from-image detection of AMI and COVID-19 using a DL model (21). Less than a fortnight after the chapter 5 experiment was published, Bridge et al. published a study investigating arrhythmia detection using a DL model trained directly with 12-lead ECGs [22]. Although the authors would not have had time to read the report of the chapter 5 study prior to submitting their own article, they did cite the chapter 4 experiment as an alternative approach worth considering for future work.

The key advantages of the direct-from-image approach, as opposed to the sample recapture approach, are:

1. An increased robustness to noise [20].
2. The ability to leverage advances in DL for image processing, as opposed to relying on the more niche area of DL for 1D signal analysis.

As discussed in the introductory chapter, the computer vision research community are widely held to be responsible for the major breakthrough in modern DL: the 'ImageNet moment'. The medical field continues to benefit from a large and active community of computer vision researchers [23]. ECG image analysis can benefit from the outputs of this community in a way that raw sample analysis cannot.

6.3.5 Synthetic ECG data for DL model training

The two other themes from chapter 5 that are felt to warrant further research are RL based on wave segmentation and the use of synthetic ECGs for training DL models. The latter theme has already been investigated by two other groups. However, there was an important difference in approaches. These groups both used a DL model to simulate ECGs, in addition to training a second DL model to analyse the ECGs [24, 25]. These two 'opposing' DL models are known as generative adversarial networks, or GANs. The 'generator' of the GAN pair is analogous to a forger, whose primary goal is to fool the detective or 'discriminator'. If the training of two models (also known as 'networks', as they are both ANNs) is properly synchronised, they remain neck-and-neck throughout, learning from each other as they improve [26].

The downside to this approach is that it remains entirely data driven, with no scope for input from human domain experts beyond the extent to which those experts can curate the training data. Consequently, the models can only learn features that are represented within the distribution of the training data. This can introduce significant bias into the model, which is held to be a major limitation of DL in the medical domain [27]. Noseworthy et al. have also demonstrated the effects of race and ethnicity on bias displayed by DL models specifically trained for ECG analysis [28].

Synthetising ECG data using a hand-crafted algorithm by no means negates the risk of bias. Rather, it gives domain experts a very high level of control over the feature distribution within the dataset. By extension, this confers some control over the types of bias contained within the data. Such control is important in a domain where certain types of error can be much more impactful than others [29]. Thus, while research into ECG GANs is ongoing, it is proposed that further research into training DL models with non-ML-based simulation methods is also important. It is posited that the work described in chapter 5, along with the accompanying open source ECG simulation toolkit, could provide an important catalyst for such work.

6.3.6 Wave segmentation as a type of RL

Finally, RL through wave segmentation pretraining is felt to be a particularly promising avenue for future research. As noted in chapter 5, other groups have investigated RL in the domain of ECG analysis. However, the use of a task that forces DL models to learn human-interpretable features addresses both the problem of learning generalisable representations and the problem of explainable AI. The generalisability of the features learned through WaSP was, to an extent, formally evaluated during the chapter 5 experiment. Further study in this direction should focus on its utility for diagnostic challenges beyond AF and AMI detection. The hypothesis that wave segmentation creates more explainable ECG AI, on the other hand, was proposed but not formally evaluated. In future, it will be important to investigate this more fully and, ideally, through some form of quantitative evaluation.

6.4 Related work

6.4.1 Personal ECG devices

During the course of this PhD programme, two related articles were published as part of this project. One was a review of the implications of consumer ECG wearables for current cardiology services, seen through the lens of a single centre case study [30]. This article served to elucidate the need for a new generation of AI-enabled ECG analysers to cope with an anticipated deluge of patient-instigated ECG recordings over the coming years. It was published in the proceedings of the 2019 Computing in Cardiology conference (IEEE), and can be found in appendix 1.

6.4.2 Reinforcement learning for resuscitation training

The second article was a position paper describing the role that reinforcement learning may play in digital clinical simulation for advanced life support (ALS) training. Reinforcement learning is an AI framework whose basic premise is the same as the generic ML training pipeline described in chapter 2. That is to say, a set of inputs and target outputs are given, and the ML model attempts to learn mapping between the two using a loss function and gradient descent-based trial and error. However, in reinforcement learning problems, the input is a 'state vector' that generally described some external environment and the output is an 'action vector'. The actions taken by the ML model have some impact on the external environment, and a new state vector is returned, and so on. The logic process learned by the ML model is described as a 'behaviour policy'.

The argument put forward in the position paper can be summarised thus:

- There are a high number of preventable deaths among acutely unwell patients, even in advanced healthcare systems like the National Health Service (NHS) (31-34).
- High fidelity simulation-based training is the most effective intervention to improve patient outcomes [35, 36]. The optimal frequency for this ALS training may be as often

as six-weekly [37]. Unfortunately, a shortage of trainers means real-world frequency is a low as four-yearly [38].

- Digital simulation has proven highly effective for clinical training [39]. A digital resuscitation simulator was developer prior to the commencement of this PhD project, so it is known to be feasible [40].
- The challenge is that real-world clinical emergencies evolve unpredictably. Research from the aviation industry has shown that unpredictable simulation training is the best way to prepare for such situations [41)]. Automated assessment of trainee performance during unpredictable (or 'stochastic') resuscitation simulation is felt to be intractable using conventional techniques. The 'correctness' of any action is determined by too many interdependent variables to faithfully capture in a hardcoded evaluation algorithm.
- Reinforcement learning is proposed as a means to (i) learn a behaviour policy that ranks a given set of actions for any state of the simulation; (ii) rate the actions of trainees in order to provide real-time feedback, which is considered essential for an effective training experience [42].

This article may seem very distantly related to the work presented in previous chapters. However, there was a clear segue from the reinforcement learning research into the ECG research presented in this thesis. Rather than train the reinforcement learning agent using raw data outputs from the resuscitation simulator, it was decided that the agent should be trained using an audio-visual feed. The idea was that this approach would allow for the agent to be fine-tuned on real-world audio-visual feeds at a later date. As noted at the end of the position paper, the longer term ambition was to create an AI application that could not only provide feedback to trainees using a simulator, but could also provide real-time decision support to clinicians in real-world emergencies.

Therefore, the first experiment following the position paper involved training the agent to analyse ECG traces shown on the screen of the defibrillator within the training environment. During the preliminary literature review on ECG image analysis technology, it was noted that there were a number of open research questions in the domain of DL-enabled ECG analysis. The research questions described in chapter 2 arose from this process, and it was decided that this research would have more immediate impact than the somewhat 'blue sky' topic of reinforcement learning in the context of medical emergencies.

Nonetheless, as a final note on future research directions, it is still felt that this topic would be an interesting and worthwhile one to explore as the field of medical DL evolves.

6.5 Overall contributions of this work

Chapter 2 of this thesis provided relevant background on the field of computerised ECG analysis, DL, and DL for ECG analysis. Three key research questions were identified:

1. Can DL be used to improve the detection of AMI from ECG signals?
2. Can DL improve the quality of ECG image analysis?
3. How can DL methods be 'democratised' so that their application is not limited to data-rich problems?

Chapter 2 also presented systematic literature review of DL for AMI detection, and concluded that this topic had been relatively underexplored compared to DL for arrhythmia detection.

Chapter 3 presented an investigation of DL methods for hyperacute detection of AMI. The results showed that the DL algorithm's performance was equivalent to a random chance classifier. However, the results also highlighted the risk of confounding data features going undetected using conventional AI research techniques. The importance of this finding for future research was highlighted by an editorial piece in the EHJ Digital Health journal.

Chapter 4 presented an investigation of DL for ECG image analysis using rule-based sample recapture. The results showed that the DL algorithm's performance was equivalent to models trained directly on raw samples for the same task. This was a first-in-field finding and the results have been cited by several original papers by author authors since.

Chapter 5 presented an investigation wave segmentation pretraining as a form of ECG RL. The results showed that the models pretrained using WaSP performed better than non-pretrained models, and than models pretrained using data from other domains, across both AF detection and AMI detection, and across both raw samples and ECG images. This suggests that the representations learned during WaSP can generalise to multiple downstream tasks. It was also proposed that WaSP allows for more interpretable AI applications than other approaches, given that clinicians are able to interrogate the quality of the learned features. However, this hypothesis was not objectively assessed. This study was only published very recently and does not yet have any citations, but its implications for future research are discussed in this chapter.

This sixth chapter summarised the conclusions of previous chapters, highlighted implications for future research, and described two related articles that were written and published as part of the work towards this thesis, but were not addressed directly the research questions described in chapter 2.

In summary, while the results presented in chapter 3 provide a relevant illustration of the pitfalls of the "black box" effect in DL, the key novel contributions of this thesis are as follows:

1) At the time that the study presented in chapter 4 was published, it was among the first pieces of peer-reviewed evidence to show that DL may be transformative for the field of ECG image

analysis. As discussed in chapters 4 and 5, this has significant implications for the democratisation of automated ECG analysis and, therefore, its global clinical impact.

2) The wavelet segmentation pretraining described in chapter 5 is one of the earliest pretraining techniques shown to be effective for DL-based ECG analysis, and the first to be made available to the wider ECG research community via an open-source toolkit. At time of thesis submission, it remains the only pretraining method shown to be effective across both major ECG data modalities (raw samples and images), and the only method for pretraining ECG image DL models that improves upon pretraining with "ordinary" image data (e.g. ImageNet).

At present, minority population groups, residents of the developing world, and sufferers of rare cardiac conditions are under-represented among digital ECG research datasets. This means that applications developed using these datasets are likely to serve them less well than groups well-represented within that data.

The option for including ECG images in future increases the likelihood that less digitally mature healthcare systems will be able contribute data from their patient populations. Effective pretraining methods reduce the need for task-specific training data, increasingly the likelihood that emerging DL-based applications can be applied to digitally under-represented groups. Therefore, it is proposed that the two key novel contributions of this thesis comprise a significant step towards to the broadening access to state-of-the-art ECG diagnostics for underserved patient populations.

6.6 Summary of limitations

A major limitation of this thesis is that all conclusions are drawn from experiments conducted with either retrospective observational data or synthetic data. A key part of the motivation for this work, as stated in section 1.1, was to explore some of the barriers to implementing AI at the point of care. Although this has been a consistent theme running through the research chapters, the lack on any prospective real-world validation of the conclusions presented here increases the risk that they will not generalise to clinical practice.

The three conclusions that may be particularly useful to test prospectively are:

1. That AI-enabled ECG image analysis can produce clinically useful outputs.
2. That pretraining on synthetic ECG data reduces the volume of training data required to achieve diagnostic accuracy equivalent to a non-pretrained model.
3. That segmentation masks provide a useful means for confidence calibration in a real-world setting.

In addition, the fast-moving nature of this field meant that the literature review presented in chapter 2 does not necessarily reflect the state of the art at the time of completing this thesis.

Over the last two years there have been significant advances in AI technology, of which the rise of transformer neural networks is a particular feature. There has also been progress in terms of guidance and regulation, such as the US Food and Drug Administration's "AI/ML-based Software as a Medical Device (SaMD) Action Plan" [43]. However, the field is continuing to evolve quickly, and it seems likely that by the time the work here was significantly updated, it would once again be out of date.

6.6 Concluding remarks

This research provides an important contribution to the field of AI-enabled ECG analysis. In particular, it highlights the importance of (1) interpretability, insofar as the mechanisms for this allow clinicians to effectively calibrate the confidence they place in AI-based analysis; (2) ECG image analysis as a means for democratising access to emerging AI technology and facilitating its application to historical data; (3) techniques that allow AI models to be applied to rarer diseases and minority patient groups where large volumes of data may be harder to come by.

Through a combination of reviewing the existing knowledge base and original research, these three key themes are described, their limits explored, and proposals for future work to further advance the field are proposed.

6.7 References

[1] Keeley EC, Boura JA, Grines CL. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: A quantitative review of 23 randomised trials. The Lancet. 2003;361:13-20.

[2] Yiadom MYA, Baugh CW, McWade CM, et al. Performance of emergency department screening criteria for an early ECG to identify ST-segment elevation myocardial infarction. Journal of the American Heart Association. 2017;6:e003528.

[3] Di Ieva A. AI-augmented multidisciplinary teams: Hype or hope? The Lancet. 2019;394:1801.

[4] Makimoto H, Höckmann M, Lin T, et al. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. Scientific reports. 2020;10:1-9.

[5] Chakraborty A, Chatterjee S, Majumder K, et al. A comparative study of myocardial infarction detection from ECG data using machine learning. In: Advanced Computing and Intelligent Technologies. Springer; 2022. p. 257-67.

[6] Rai HM, Chatterjee K. Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data. Appl Intell. 2022;52:5366-84.

[7] Liu WC, Lin CS, Tsai CS, et al. A deep learning algorithm for detecting acute myocardial infarction. EuroIntervention. 2021;17:765-73.

[8] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[9] Kawintiranon K, Singh L. In: Knowledge enhanced masked language model for stance detection. Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies; ; 2021. p. 4725-35.

[10] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.

[12] P. Sarkar, A. Etemad. Self-supervised ECG representation learning for emotion recognition. IEEE Transactions on Affective Computing. 2020:1-.

[13] Misra I, Maaten Lvd. In: Self-supervised learning of pretext-invariant representations. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; ; 2020. p. 6707-17.

[14] Liu R, Gillies DF. Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognit. 2016;53:73-86.

[15] Tian C, Fei L, Zheng W, et al. Deep learning on image denoising: An overview. Neural Networks. 2020;131:251-75.

[16] Garg DK, Thakur D, Sharma S, et al. ECG paper records digitization through image processing techniques. International Journal of Computer Applications. 2012;48:35-8.

[17] Al Hinai G, Jammoul S, Vajihi Z, et al. Deep learning analysis of resting electrocardiograms for the detection of myocardial dysfunction, hypertrophy, and ischaemia: A systematic review. European Heart Journal-Digital Health. 2021;2:416-23.

[18] PT de Jaegere P. Artificial intelligence for automated ECG analysis. an experimental study revealing knowns and mysteries. still a long pathway ahead? European Heart Journal-Digital Health. 2021.

[19] Mishra S, Khatwani G, Patil R, et al. ECG paper record digitization and diagnosis using deep learning. Journal of medical and biological engineering. 2021;41:422-32.

[20] Li Y, Qu Q, Wang M, et al. Deep learning for digitizing highly noisy paper-based ECG records. Comput Biol Med. 2020;127:104077.

[21] Anwar T, Zakir S. In: Effect of image augmentation on ECG image classification using deep learning. 2021 international conference on artificial intelligence (ICAI); IEEE; 2021. p. 182-6.

[22] Bridge J, Fu L, Lin W, et al. Artificial intelligence to detect abnormal heart rhythm from scanned electrocardiogram tracings. Journal of Arrhythmia. 2022.

[23] Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. NPJ digital medicine. 2021;4:1-9.

[24] Wulan N, Wang W, Sun P, et al. Generating electrocardiogram signals by deep learning. Neurocomputing. 2020;404:122-36.

[25] Li W, Tang YM, Yu KM, et al. SLC-GAN: An automated myocardial infarction detection model based on generative adversarial networks and convolutional neural networks with single-lead electrocardiogram synthesis. Inf Sci. 2022.

[26] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview. IEEE Signal Process Mag. 2018;35:53-65.

[27] Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Communications medicine. 2021;1:1-3.

[28] Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ECG analysis. Circulation: Arrhythmia and Electrophysiology. 2020;13:e007988.

[29] Sharkey SW, Berger CR, Brunette DD, et al. Impact of the electrocardiogram on the delivery of thrombolytic therapy for acute myocardial infarction. Am J Cardiol. 1994;73:550-3.

[30] R. Brisk, R. Bond, D. Finlay, et al. In: Personal ECG devices: How will healthcare systems cope? A single centre case study. - 2019 computing in cardiology (CinC); ; 2019. p. Page 1,Page 4.

[31] Sarwar S, Shafi MI. National confidential enquiry into patient outcome and death. Obstetrics, Gynaecology & Reproductive Medicine. 2007;17:278-9.

[32] Centre for Clinical Practice at NICE (UK. Acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital [internet]. . 2007.

[33] Urquhart A, Yardley S, Thomas E, et al. Learning from patient safety incidents involving acutely sick adults in hospital assessment units in england and wales: A mixed methods analysis for quality improvement. J R Soc Med. 2021;114:563-74.

[34] Hogan H, Healey F, Neale G, et al. Preventable deaths due to problems in care in english acute hospitals: A retrospective case record review study. BMJ quality & safety. 2012;21:737-45.

[35] Callaway CW, Soar J, Aibiki M, et al. Part 4: Advanced life support: 2015 international consensus on cardiopulmonary resuscitation and emergency cardiovascular care science with treatment recommendations. Circulation. 2015;132:S84-S145.

[36] Bellomo R, Goldsmith D, Uchino S, et al. Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates. Crit Care Med. 2004;32:916-21.

[37] Sutton RM, Niles D, Meaney PA, et al. Low-dose, high-frequency CPR training improves skill retention of in-hospital pediatric providers. Pediatrics. 2011;128:e145-51.

[38] Perkins GD, Kimani PK, Bullock I, et al. Improving the efficiency of advanced life support training: A randomized, controlled trial. Ann Intern Med. 2012;157:19-28.

[39] Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. Ann Surg. 2002;236:458,63; discussion 463-4.

[40] Brisk R. Virtu-ALS. 2016. Available from: www.virtu-ALS.com.

[41] Landman A, van Oorschot P, van Paassen MM, et al. Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios. Hum Factors. 2018;60:793-805.

[42] Hewson MG, Little ML. Giving feedback in medical education: Verification of recommended techniques. Journal of general internal medicine. 1998;13:111-6.

[43] FDA. AI/ML-based software as a medical device (SaMD) action plan. 2021 [cited 22/09/2022]. Available from: https://www.fda.gov/media/145022/download.

Appendix 1:

Personal ECG Devices: How Will Healthcare Systems Cope? A Single Centre Case Study

7.1 Introduction

With the prevalence of personal ECG devices already on the rise and the Apple Watch Series 4 recently hitting the market, the number of daily ECG recordings in the developed world is poised to explode [1]. While some individuals will undoubtedly benefit from enhanced diagnosis of cardiac arrhythmias, the consequences of false alarms are likely to be detrimental to patients and clinicians alike. This article aims to describe the possible impact of this technology on a typical cardiology department of a UK NHS Trust.

7.1.1 The clinical setting

The NHS is the largest single-payer healthcare system, the fifth largest employer globally (after the US Department of Defense, the Chinese People's Liberation Army, Wallmart and MacDonald's) and is responsible for a population of approximately 66 million [2,3]. For this reason, it is often considered to be a particularly good test bed for emerging healthcare technologies [4].

The Southern Health and Social Care Trust (SHSCT) is an NHS trust in Northern Ireland. The population of the SHSCT's catchment area is approximately 360,000, which has grown substantially over the last decade [5]. The Northern Irish population is ageing more rapidly, scores lower on socioeconomic metrics and has a lower average life expectancy than the rest of the UK [6-9]. Waiting times for outpatient consultations are correspondingly longer than the national average [10].

The cardiology department of the SHSCT runs a coronary care unit at each of two teaching hospitals, with six and eight beds respectively. There are an additional 25 permanent inpatient cardiology beds at the larger of the two hospitals, along with two interventional cardiac catheterisation laboratories. The department records about 55,000 inpatient episodes yearly and has a high volume of outpatient encounters (internal statistics).

7.1.2 Current ambulatory ECG service

At present, Holter monitors and cardiac event recorders can be requested directly by physicians outside the cardiology department, including general practitioners (GPs). Implantable loop recorders (ILRs) are only requested by members of the cardiology team and must be approved by a consultant prior to implantation.

Recordings from wearable devices are automatically annotated by specialized software (Sentinel, Spacelabs Healthcare, Snoqualmie, WA, US) and reviewed by one of 13 full time equivalent cardiac physiologists (CPs). Approximately 2,500 studies are undertaken and reported annually. According to SHSCT CPs, reporting a single study takes between 15 minutes and several hours.

The responsibility for acting upon the report generated by a wearable ECG monitor lies with the requesting consultant. No study is undertaken without a responsible consultant physician designated on the request form.

7.2 Personal ECG devices

In their 2018 review, Banshal and Joshi identified 15 widely available personal ECG devices, but only six that were associated with Pubmed-listed studies [11]. Three of these are currently intended for prescription by medical professionals and three are available to individual consumers online. Table 7.1 shows representative prices for the commercially available devices, with the addition of the Apple Series 4 Watch (not included in the 2018 review as it had not been released).

This review will focus primarily on the impact of two devices: the AliveCor Mobile device and the Apple smartwatch.

7.2.1 Characteristics of selected devices

The AliveCor device is chosen as both the most affordable of the peer reviewed devices and the best supported by published evidence. It has been reviewed favourably by the National Institute for Clinical Excellence (NICE), who noted that the sensitivity and specificity for the automated detection of AF has been reported in multiple studies as above 85% and 90%, respectively [12].

Figure 7.1. An Apple Watch. The Series 4 model is ECG capable.

The ECG-enabled Apple Series 4 Watch is chosen for having, by far, the highest predicted sales figures. In 2018, prior to the introduction of ECG capabilities, Apple is estimated to have shipped 22.5 million watches globally. [13] It is not clear how many of the Series 4 watches have been bought in the UK since it went on sale in September 2018, nor how many owners use the ECG technology. However, the authors of this study consider these sales figures to be the most compelling reason to begin thinking about the logistics of the widespread use of self-prescribed ECG monitoring.

Device	Cost	Outlet
Omron Heart Scan	£699.99	Amazon.co.uk
AliveCor Mobile	£99	AliveCor.com
REKA Health	N/A	
Zenicor ECG	N/A	
Schiller MINISCOPE	£1134	EKGshop.com
ZioPatch	N/A	
Apple Series 4 Watch	£389	John Lewis

Table 7.1. Representative costs of personal ECG devices.

According to a press release by Apple, the sensitivity and specificity for the automated detection of AF is 98.3% and 99.6%, respectively [14]. The study from which these figures were obtained has not yet been published in a peer reviewed journal. It is felt that there is currently insufficient evidence to support a significant difference in device performance between the Apple Watch and the AliveCor device, and will assume parity henceforth.

At present, both AliveCor and Apple aim to automatically diagnose NSR and AF. Other classifications are tachycardia, bradycardia or inconclusive / unreadable. Diagnoses of such traces must be made manually. To this end, both devices can store ECG tracings in PDF format for

transmission to the patient's clinician.

AliveCor offers US customers a free manual analysis of their first ECG trace by a cardiologist, and analysis of future recordings for a fee.

ECG recordings are single-lead and must be user-initiated on both devices. On the Apple Watch, users touch a finger to the digital crown of the watch. On the AliveCor device, users place a finger from each hand on the electrodes. Recordings on the Apple device last 30 seconds. Recordings on the AliveCor device last 40 seconds [15,16].

7.3 Pathway for abnormal recordings

If either Apple or AliveCor applications detect rhythms other than NSR, both companies return responsibility to the user by suggesting they consult a physician. It is at the point where a UK user sends ECG data to a physician that problems may begin arise from a healthcare provider's perspective. Within the NHS, self-referral to a specialist is only possible under exceptional circumstances (for example, one may see an ophthalmologist directly in eye casualty) or via a private clinic for a fee. In general, however, a patient's first point of contact is their GP or an emergency department (ED) physician. For personal ECGs, it is likely be the former.

At present, there appears to be significant variation in how comfortable GPs are with ECG interpretation. In the experience of the SHSCT cardiology department, some referrals from the community arrive with accurate interpretation of even relatively rare ECG abnormalities (e.g. "?Brugada"), whereas others include fundamental mistakes such as confusing sinus arrhythmia for AF due to irregular R-R intervals but in the presence of clear P waves.

Regardless of individual competence, however, there can be little doubt that primary care is under unprecedented pressure and that GPs are unlikely to relish the prospect of an additional source of work [17]. It is anticipated, therefore, that most personal ECG recordings submitted to GPs will be referred to the cardiology department for review.

7.4 Impact on cardiology services

In the SHSCT, the current wait for a routine Holter monitor is around 52 weeks (internal statistic). However, all studies are currently ordered by a qualified clinician who, if they deem the test to be urgent, can stipulate a shorter time frame. In the absence of any clear way to triage personal recordings, it seems likely that ECGs mandated directly by patients will be considered a lower

priority than studies ordered by qualified medical professionals. It is therefore likely that they will be associated with a substantial delay in reporting. This would put some patients at risk, particularly if they have declined to seek expert attention via established channels as a consequence of having submitted potentially diagnostic information.

Furthermore, if the uptake of personal ECG monitoring among SHSCT patients is significant, the extra workload could cause delays across the CP service. This includes all ambulatory ECG monitoring, pacemaker checks, exercise stress testing, echocardiography services and all cath lab sessions.

Let it be assumed that, two years from now, the entire Apple smartwatch range is ECG capable and continues to sell at 22.5 million units per year. If one disregards a likely preponderance of sales towards industrialised nations like the UK and instead assumes an even global sales distribution among 7 billion people, approximately 1000 watches would be acquired by the SHSCT population.

Selder et al. (2019) found that patients using the AliveCor Kardia Mobile device submitted a median of 28 ECGs per patient per year, though this was among patients presenting to cardiology

Figure 7.2. The AliveCor Kardia Mobile device.

services with palpitations and is likely to be higher than a non-selected population [18]. Indeed, 19% of ECGs submitted showed AF, whereas AF prevalence among under 65s in the general population (the demographic into which the majority of Apple Watch owners fall) is around 2% [19, 20].

Nonetheless, 20% of all ECGs submitted were flagged as potentially abnormal by the device software and subsequently found to either show NSR or be unclassifiable. A press release from Stanford University regarding the Apple Heart Study noted that a little over half of users receiving an abnormal pulse warning sought medical attention [21]. If there was a similar false positive rate among Apple Watch owners, and if 50% decided to seek medical review of these recordings, this

could result in 2,800 additional ECGs being analysed by SHSCT staff per year: a 100% increase in outpatient studies analysed.

7.5 Benefit to patients

Halcox et al. (2017) report AF diagnosis rates among over 65s undergoing routine care (RC) vs twice-weekly ECG monitoring with the AliveCor device. 5 patients receiving RC were diagnosed with AF over the course of a year, compared with 19 in the AliveCor group (hazard ratio 3.9, p=0.007) [22]. They were unable to demonstrate a statistically significant difference in rates of cerebrovascular events over the 12-month study, but Boriani et al. (2014) previously concluded that silent AF is associated with a modifiable risk of embolic stroke if anticoagulants are appropriately prescribed [23]. Though the duration of AF warranting anticoagulation remains a matter of debate, it seems reasonable to conclude that higher rates of AF and appropriate anticoagulant prescription may be associated with lower rates of embolic stroke.

However, the authors note that the population prevalence of AF among Halcox's over 65 subjects is 9%, compared with 2% among most Apple Watch owners [20]. Furthermore, the rate of embolic stroke among otherwise well, young patients diagnosed with AF on routine screening is unknown (Boriani studied patients with pre-existing cardiac conditions). There is, therefore, insufficient data to estimate the cost per quality adjusted life year (QALY) of personal ECG monitoring, nor to quantify the impact of high false positive rates on the wider cardiology service and the psychological wellbeing of patients. It is felt that it is not clear that the widespread uptake of personal ECG devices will benefit patients in the SHSCT.

7.6 A technological solution to a technological problem?

As a final note, recent developments in deep learning-based arrhythmia detection may prove timely in light of the issues discussed in this review. Hannun et al. (2019) claim to have achieved "cardiologist-level" using a 34-layer convolutional neural network trained on large scale ambulatory ECG data [24]. This is a relatively nascent technology but if the results reported in this paper are reproducible by other groups, this may substantially reduce the number of false positive results and shift the risk-benefit balance in favour of personal ECG devices.

7.7 References

[1] Giebel, G.D., Gissel, C. Accuracy of mHealth Devices for Atrial Fibrillation Screening: Systematic Review. JMIR Mhealth Uhealth, 2019; 7(6):e13641

[2] "The NHS is the world's fifth largest employer," The Nuffield Trust. October 2017. Report available at https://www.nuffieldtrust.org.uk/chart/the-nhs-is-the-world-s-fifth-largest-employer, last accessed 28th August 2019.

[3] "Population estimates," The Office for National Statistics. June 2019. Report available at https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates, last accessed 28th August 2019.

[4] "NHS England Test Beds Programme: Information Governance learning from Wave," NHS England, September 2018. Report available at https://www.england.nhs.uk/publication/nhs-england-test-beds-programme-information-governance-learning-from-wave-1/, last accessed 28th August 2019.

[5] "2017/18 Annual Report and Accounts," Southern Health and Social Care Trust, March 2018. Report available at http://www.southerntrust.hscni.net/pdf/SHSCT%20ANNUAL%20REPORT%20AND%20ACCOUNTS%2031%20March%202018.pdf, last accessed 28th August 2019.

[6] "Households below average income (HBAI) statistics," Department for Work and Pensions, May 2019. Report available at https://www.gov.uk/government/collections/households-below-average-income-hbai--2, last accessed 28th August 2019.

[7] "Health state life expectancies by national deprivation deciles, England and Wales Statistical bulletins," Office for National Statistics, March 2017. Report available at https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/healthstatelifeexpectanciesbyindexofmultipledeprivationimd/previousReleases, last accessed 28th August 2019.

[8] "Healthy life expectancy: Scotland," The Scottish Public Health Observatory, March 2019. Report available at https://www.scotpho.org.uk/population-dynamics/healthy-life-expectancy/data/scotland, last accessed 28th August 2019.

[9] "Health Inequalities Regional Report 2016," Department of Health Northern Ireland, Oct 2016. Report available at https://www.health-ni.gov.uk/sites/default/files/publications/health/hscims-report-2016.pdf, last accessed 28th August 2019.

[10] “Northern Ireland Waiting Time Statistics: Outpatient Waiting Times Quarter Ending March 2019,” Department of Health Northern Ireland, March 2019. Report available at https://www.health-ni.gov.uk/sites/default/files/publications/health/hs-niwts-outpatient-waiting-times-q4-18-19.pdf, last accessed 28th August 2019.

[11] Bansal, A., Joshi, R. Portable out-of-hospital electrocardiography: A review of current technologies. J Arrhythm, 2018;34(2):129-138.

[12] “AliveCor Heart Monitor and AliveECG app (Kardia Mobile) for detecting atrial fibrillation,” NICE, Aug 2015. Report available at https://www.nice.org.uk/advice/mib35, last accessed 28th August 2019.

[13] “U.S. Smartwatch Sales See Strong Gains, According to New NPD Report,” Press release – the NPD Group, Feb 2019. Available at https://www.npd.com/wps/portal/npd/us/news/press-releases/2019/us-smartwatch-sales-see-strong-gains-according-to-new-npd-report/, last accessed 28th August 2019.

[14] “ECG app and irregular heart rhythm notification available today on Apple Watch,” Press release – Apple Inc, Mar 2019. Available at https://www.apple.com/uk/newsroom/2019/03/ecg-app-and-irregular-rhythm-notification-on-apple-watch-available-today-across-europe-and-hong-kong/, last accessed 28th August 2019.

[15] “User Manual for Kardia™ by AliveCor®”, Instruction manual – AliveCor. Nov 2016. Available at https://www.AliveCor.com/previous-labeling/kardia/08LB12.13.pdf, last accessed 28th August 2019.

[16] “Taking an ECG with the ECG app on Apple Watch Series 4”, Instruction manual – Apple Inc, Aug 2019. Available at https://support.apple.com/en-gb/HT208955, last accessed 28th August 2019.

[17] “Working in a system that is under pressure”, BMA, Mar 2018. Report available at https://www.bma.org.uk/-/media/files/pdfs/collective%20voice/influence/key%20negotiations/nhs%20pressures/working-system-under-pressure-bma-council-report-mar-2018.pdf?la=en, last accessed 28th August 2019.

[18] Selder, LJ., Breukel, L., Blok, S., Van rossum, AC., Tulevski, II., Allaart, CP. A mobile one-lead ECG device incorporated in a symptom-driven remote arrhythmia monitoring program. The first 5,982 Hartwacht ECGs. Neth. Heart J.,2019;27(1):38-45.

[19] Campbell, M. Apple Watch, other wearables increasingly used to manage chronic health conditions, study says, News report – appleinsider, Aug 2018. Available at https://appleinsider.com/articles/18/08/30/apple-watch-other-wearables-increasingly-used-to-manage-chronic-health-conditions-study-says, last accessed 28th August 2019.

[20] "Atrial Fibrillation Fact Sheet", CDC, Aug 2017. Report available at https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_atrial_fibrillation.htm, last accessed 28th August 2019.

[21] "Apple Heart Study demonstrates ability of wearable technology to detect atrial fibrillation", Press release – Stanford University School of Medicine, Mar 2019. Available at http://med.stanford.edu/news/all-news/2019/03/apple-heart-study-demonstrates-ability-of-wearable-technology.html, last accessed 28th August 2019.

[22] Halcox, JPJ., Wareham, K., Cardew, A. Assessment of Remote Heart Rhythm Sampling Using the AliveCor Heart Monitor to Screen for Atrial Fibrillation: The REHEARSE-AF Study. Circulation, 2017;136(19):1784-1794.

[23] Boriani, G., Glotzer, TV., Santini, M. Device-detected atrial fibrillation and risk for stroke: an analysis of >10,000 patients from the SOS AF project (Stroke preventiOn Strategies based on Atrial Fibrillation information from implanted devices), Eur. Heart J., 2014;35(8):508-16.

[24] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65-69.

Appendix 2:

AI to enhance interactive simulation-based training in resuscitation medicine

8.1 Introduction

Much is made of the potential of emerging AI technology to bring badly-needed innovation to the field of medicine. Yet we see little evidence of this on the wards each day. Part of the problem might be that the vanguard of this progress is comprised almost exclusively of data scientists and ML researchers. Most healthcare workers remain entirely ignorant of even the basic concepts underpinning ML and, by extension, the technology we commonly define as being "artificially intelligent". Practitioners of ML are likely to gravitate towards clinical problems that present favourable targets for their science, for example single-step classification tasks in data-rich areas. Hence, disciplines like radiology are enjoying the lion's share of the attention from the ML community [1].

Here, the application of ML to a sequential decision-making task in a simulated clinical environment is described

8.2 The clinical need

***Figure 8.6**: High-fidelity, face-to-face clinical simulation (Reproduced with permission. © University of Dundee)*

There are over 10, 000 in-hospital cardiac arrests annually in the UK [2]. Outcomes for these patients are poor. Only one in five will survive to hospital discharge. Over half of these survivors will have some degree of neurological (brain) damage [3]. Some cardiac arrests happen "out of the blue", due to sudden events such as myocardial infarction or pulmonary embolus. But a significant proportion will be preceded by a gradual deterioration in the patient's condition. It has been concluded that as many as 5% of hospital deaths may be averted, largely by the prompt identification and effective treatment of acute illness [4 – 7]. The key question is: how do we improve the recognition and treatment of the deteriorating patient?

There are a range of novel technology-based solutions on the market but their efficacy remains unproven [8]. High quality simulation training for clinical staff is still by far the best-evidenced intervention [9, 10]. However, the resource-intensive nature of existing, face-to-face simulation methods is a limiting factor. This is due largely to the requirement for a high ratio of expert instructors to trainees.

There is evidence to suggest that the optimal training frequency might be as often as six-weekly [11]. In ever-shorter-staffed healthcare systems, even the logistic challenge of ensuring practitioners have access to an ALS course once every four years has necessitated a push by the European Resuscitation Council to streamline training and cut courses from two days to

one [12]. It is in this context that a novel human computer interaction-based solution is explored. This solution takes the form of an AI-supported digital simulation system. By negating the need for an expert human presence, this solution could facilitate the delivery of low-cost, high-impact training at unprecedented frequency.

8.3 Digital resuscitation simulation

The rationale for digital simulation in clinical training is well established. In fact, for certain procedural skills, such as those required to perform laparoscopic surgery, it has proven more efficacious even than conventional training methods [13].

At the simplest level, gated progression through a surgical simulation can be achieved using single-condition "if-then-else" statements. For example:

If [*the trainee performs step* A *according to the optimal method*] then [*the trainee is deemed to have demonstrated proficiency and can progress to step* B] else [*they receive constructive feedback and retry step* A].

This is both educationally viable, as it allows for the integration of a proficiency-based progression model [14], and computationally favourable. By restricting the permissible action space, the number of resultant states for which the simulation must account is very limited. Naturally, modern simulators have built upon this basic framework to develop less obviously linear narratives and to account for a number of common procedural complications. However, by continuing to restrict permissible user actions, they can continue to limit the state space to manageable dimensions.

***Figure 8.7**: Digital simulation for laparoscopic surgery Reproduced with permission. © Marcus Rall*

Resuscitation simulation cannot take advantage of the same approach. The focus when training for ALS moves from procedural to conceptual knowledge [15]. ALS providers are not required to become expert at tackling a fixed problem like their surgical colleagues. Rather, then need to develop cognitive processes that are generalizable to a wide range of disparate clinical scenarios. Research from the aviation industry has shown that unpredictable, or 'stochastic', simulation is most effective in this context [16].

Anyone who is familiar with model-free reinforcement learning will understand that development of generalizable behaviour policies first requires exploration of the action-state space [17] The same is true for humans, though psychologists would more likely term this as

the "active experimentation" phase of Kolb's experiential learning cycle [18]. In lay terms, this is simply known as trial-and-error learning. Thus, a simulator designed to develop generalizable behaviour is likely to maximise its efficacy by allowing users access to action-state spaces that reflect the true diversity of real-world experience. Hence, the need for stochastic simulation.

To test the feasibility of stochastic simulation for resuscitation training, a prototype stochastic resuscitation simulator was developed [19]. Figure 8.3 presents several screenshots from this programme. Stochastic simulation was felt by the developers to be a reasonable approach, and feedback from users was very positive. However, the downside is that the 'correct' path through the simulation cannot be determined in advance. Thus, it becomes much harder to give automated real-time feedback to trainees, which is considered vital for an effective educational experience [20].

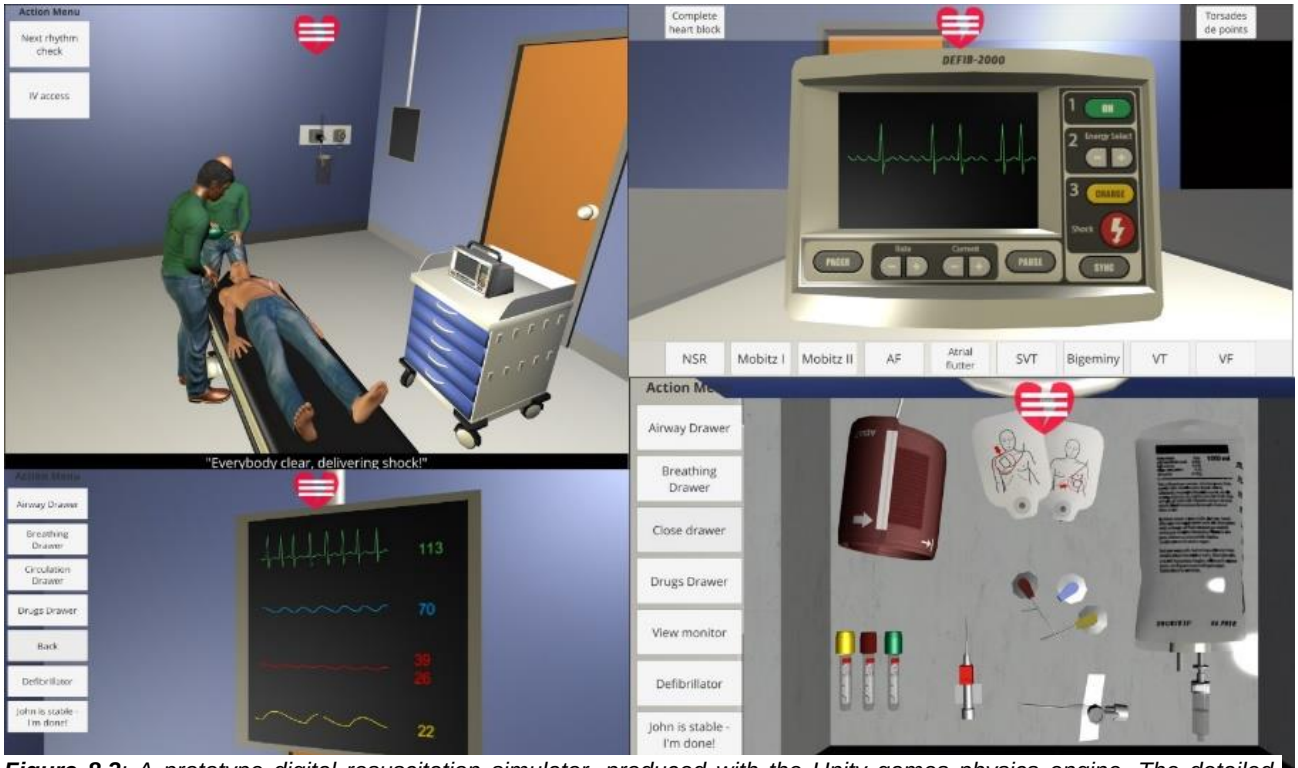


***Figure 8.3**: A prototype digital resuscitation simulator, produced with the Unity games physics engine. The detailed, stochastically-generated clinical environment makes for a particularly high-fidelity experience but necessitates a novel approach to automated trainee evaluation.*

8.4 The role of ML

Prespecifying a set of rules to evaluate the quality of a trainee's action for any given state is extremely challenging in a stochastic, high fidelity clinical simulation. The 'correctness' of actions depends on many interdependent variables. Hardcoding a set of rules to encapsulate

this knowledge was deemed to be infeasible. However, it was felt that the problem may be more tractable using ML methods.

The proposed approach takes its precedent largely from the seminal work by researchers at DeepMind in the field of deep reinforcement learning [21 – 23]. In the initial 2013 study, they employed a deep Q-learning strategy to attain human level performance in three of six complex reinforcement learning tasks involving Atari games. Their system, in short, consisted of a deep neural network tasked with predicting the action-value ("Q") function for a given behaviour policy (usually referred to as "π"). The network was updated during the training process using stochastic gradient descent, and the update process smoothed out using an experience replay mechanism (to avoid, say, a promising behaviour strategy being too heavily penalised for a single bad outcome). In 2015, further refinements to this process allowed them to surpass human performance in a large number of similar tasks.

This approach works well for environments like the game Space Invaders, where the action-state space is limited and there is minimal need for long-term planning. The ML model can develop generalizable skills quickly and rapidly transition to an "epsilon greedy" strategy, whereby it spends more time exploiting its new skillset and less time exploring its environment (or, as it was described earlier, engaging in trial-and-error learning).

The prototype simulator developed for resuscitation training, however, has a comparatively high-dimensional action space, a more diverse state space, and a greater need for long-term planning. A ML model targeted at this application would need both a much longer period of exploration (or "a slower epsilon decay") and less frequent policy updates to achieve a similar level of efficacy using Q-learning. This would result in exponentially increased computational cost. Furthermore, the stochastic nature of the simulator may confound attempts to learn an effective action-value policy (because the same action taken in the same state could potentially result in two different "rewards").

The aim, therefore, is to take a further lead from the DeepMind researchers. In 2016, they revisited some of the Atari problems, but this time using an "actor-critic" approach. Instead of trying to learn an optimal action-value function (from which the behaviour policy is then implied in a straight Q-learning approach), the actor-critic method employs two asynchronous models: an "actor" whose task is to directly learn an optimal behaviour policy and a "critic" whose task is to learn an action-value policy upon which the actor bases its updates. This has a few key advantages over the Q-learning framework. Namely, direct policy optimisation allows the model to more effectively deal with high-dimensional action spaces and to learn stochastic policies. The addition of a critic model as an action-value estimator offsets the increased variance and allows for more frequent policy updates than, say, a Monte Carlo approach to policy optimisation [24]. For this reason, it is proposed that the actor-critic framework represents the most promising solution to this particular problem.

8.5 Conclusions and future work

Based on the above review, it is proposed that an actor-critic model could be trained to resuscitate virtual patients within a stochastic clinical simulator. The intention would then be to employ the "critic" network from such a model as a means of evaluating the actions of our human trainees within any given state of the simulation, thus providing a basis upon which to provide constructive, real-time feedback within a complex, stochastic clinical simulation.

If this approach was successful, it could solve the problem of delivering high-frequency resuscitation simulation training within a resource-constrained healthcare system, and plausibly improve patient outcomes as a result. However, it could also open up a new paradigm of digital medical education. In this paradigm, medical students could hone their clinical skills on simulated wards, receiving constructive feedback as they assessed and treated virtual patients, before they ever made a management decision regarding a real-world patient.

Furthermore, a successful outcome from this research would lay the ground for an investigation of whether this framework – i.e. clinical simulation as a training environment for reinforcement learning models – could one day be used to train AI models to take decisions directly affecting patient care.

8.6 References

[1] Wang S, Summers RM. Machine learning and radiology. Medical image analysis. 2012; 16(5):933-51

[2] Nolan JP, Soar J, Smith GB, et al. Incidence and outcome of in-hospital cardiac arrest in the United Kingdom National Cardiac Arrest Audit. Resuscitation. 2014;85(8):987-92.

[3] Young GB. Clinical practice. Neurologic prognosis after cardiac arrest. N Engl J Med. 2009;361(6):605-11.

[4] National Confidential Enquiry into Patient Outcome and Death. An acute problem? London: NCEPOD; 2005.

[5] National Institute for Health and Clinical Excellence. Acutely Ill Patients in Hospital: Recognition and Response to Acute Illness in Hospital. National Institute for Health and Clinical Excellence, London. 2007.

[6] Thomson R.T., Leuttel D., Healey F. & Scobie S. National Patient Safety Agency, London. 2007.

[7] Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ quality & safety. 2012; 21(9):737-45.

[8] Amorim FF, Santana AN. Automated early warning system for septic shock: the new way to achieve intensive care unit quality improvement? Annals of translational medicine. 2017; 5(1):17.

[9] Callaway CW, Soar J, Aibiki M, et al. Part 4: Advanced Life Support: 2015 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation. 2015;132(16 Suppl 1):S84-145.

[10] Bellomo R, Goldsmith D, Uchino S, et al. Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates. Crit Care Med. 2004;32(4):916-21.

[11] Sutton RM, Niles D, Meaney PA, et al. Low-dose, high-frequency CPR training improves skill retention of in-hospital pediatric providers. Pediatrics. 2011; 128(1):e145-51.

[12] G.D. Perkins, P.K. Kimani, I. Bullock, et al. Improving the efficiency of advanced life support training: a randomized. Controlled Trial Ann Intern Med. 2012;157:19-28

[13] Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg. 2002; 236(4):458-63.

[14] Gallagher AG, Ritter EM, Champion H, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Ann Surg. 2005;241(2):364-72.

[15] Schneider M, Rittle-Johnson B, Star JR. Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. Developmental psychology. 2011; 47(6):1525-38.

[16] Landman A, van Oorschot P, van Paassen MM, et al. Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios. Hum Factors. 2018;60:793-805.

[17] R. Sutton and A. Barto. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA, 1998.

[18] Kolb, D. A. 1984. Experiential learning: Experience as the source of learning and development. New Jersey: Prentice-Hall.

[19] Brisk, R. Virtu-ALS. [COMPUTER / MOBILE APPLICATION]. Available at www.virtu-ALS.com. [Accessed 15 March 2018]

[20] Hewson MG, Little ML. Giving feedback in medical education: verification of recommended techniques. J Gen Intern Med. 1998;13(2):111-6.

[21] V Mnih, K Kavukcuoglu, D Silver, A Graves, I Antonoglou, D Wierstra, M Riedmiller. Playing Atari with Deep Reinforcement Learning. 2013. arXiv:1312.5602

[22] Mnih, V. Kavukcuoglu, K. Silver, D. Rusu, A. Veness, J. Bellemare, M. Graves, A. Riedmiller, M. Fidjeland, A. Ostrovski, G. Petersen, S. Beattie, C. Sadik, A. Antonoglou, I. King, H. Kumaran, D .Wierstra, D. Legg, S. Hassabis, D. Human level control through deep reinforcement learning. Nature. 2015; 518, p 529–533.

[23] Mnih, V. Badia, A. P. Mirza, M. Graves, A. Lillicrap, T. P. Harley, T. Silver, D. Kavukcuoglu, K. Asynchronous Methods for Deep Reinforcement Learning. 2016. arXiv:1602.01783

[24] Silver, D. Policy Gradient. [Lecture] University College London. 21 Dec 2015. Available at http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html [Accessed 01/02/2018]