UNIVERSITY OF CAMBRIDGE
Department of Engineering

Predicting the Risk of Induced Atrial Fibrillation During Electrophysiological Studies Through the Automated Detection of Fractionation

Author Name: Matthew Ashman
Supervisor: Dr. Elena Punskaya

Date: November 12, 2019

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

Signed: _____ Date: _____

Technical Abstract

Electrophysiology (EP) studies play an important role in the management of cardiac arrhythmias. The procedure involves the insertion of catheter electrodes into the heart, from which electrical signals, collectively referred to as an electrogram, are recorded and used to make a diagnosis. During an EP study, there is a risk of putting the patient into a state of atrial fibrillation (AF). If left untreated, the patient can experience life threatening complications such as cardiac arrest or stroke. In most cases, a large electric shock (3000-5000 Volts) is necessary to recover. Research has shown that electrogram recordings from patients at risk of entering AF exhibit conduction delays, prolonged responses and significant fragmentation. The term 'fractionation' has been widely adopted by the medical community to describe these changes in the shape of a patient's response. In this report, an exploration is made into the use of advanced signal processing and machine learning techniques to 'automatically' determine the degree to which a patient's response becomes fractionated, and in turn evaluate the risk a patient is at to entering AF.

An initial dataset, comprised of nine patients who entered AF during the EP study and 37 who did not, was available for the development of a model. A test dataset, comprised of 10 patients who entered AF and 30 who did not, was later made available to evaluate the performance of the model. The data for each patient consisted of a collection of text files corresponding to each step of the Antegrade Curve, a pacing manoeuvre in which the patient's heart is increasingly stressed by externally induced stimuli (pulses). In the data extraction process, the location of each pulse was identified and the subsequent 125ms interval capturing the patient's atrial response was extracted. A categorisation scheme was designed with consultation from several medical professionals, whereby each response was assigned a label indicating no, mild or severe fractionation. A selection of features identified as being indicative of fractionation, computationally efficient and interpretable was extracted from each response. These feature were then normalised with respect to each patient to account for natural patient variability. To overcome the limitations of having few severely fractionated responses, a data augmentation technique, whereby fractionated segments are scaled by a randomly generated cubic spline, was employed.

An investigation was made into the performance of the logistic regression and naïve Bayes classification algorithms. Whilst both models were able to distinguish between not at all and severely fractionated responses, the logistic regression classifier was shown to outperform the naïve Bayes classifier, achieving an average F_1-score of 0.884 on a validation dataset. This was attributed to the violation of assumptions made by the naïve Bayes classifier, specifically that feature values are mutually independent and distributed normally. By contrast, the assumption of a linear decision boundary made by the logistic regression classifier was appropriate, since each feature value can be expected to monotonically increase with degree of fractionation.

The performance of the logistic regression classifier trained on augmented responses of

varying degrees of distortion, σ_A, was investigated. It was found that the use of data augmentation improved the performance of the logistic regression classifier, increasing the average F_1-score to 0.892 when trained on augmented responses of degree $\sigma_A = 0.3$. An analysis of the weights of the logistic regression classifier indicated that measures of conduction delay, response duration and proportion of the response populated by peaks had the most significance in deciding whether or not a response is fractionated. The classifier achieved an average F_1-score of 0.853 on the test dataset, with no severely fractionated responses misclassified as having no fractionation, and only two out of 820 responses vice-versa.

The f_{score}, a continuous quantitative metric for the degree of fractionation comprised of the probabilistic outputs of the logistic regression classifier, was used to compare groups of patients who entered AF and those who did not. Patients who entered AF were shown to exhibit a sharp rise in f_{score} as the electrical stress on the heart increased, which was particularly apparent in the recordings from the electrode nearest the site of the pulse. The f_{score} was shown to more effectively differentiate between the two groups of patients than the number of peaks and duration of atrial response, which are two of the metrics most frequently used in previous studies.

A questionnaire was designed and completed by three medical professionals, in which each participant was asked to interpret 58 patients' responses and indicate whether or not the EP study should have been stopped early to prevent induced AF. In 10 of the 14 cases in which patients entered AF, medical professionals unanimously agreed that the patient was at no risk, suggesting that the currently accepted warning signs are flawed. By contrast, the model was able to identify, and thus potentially prevent, 12 out of the 14 cases of induced AF.

The results strongly support the use of the model in both determining the degree of fractionation and predicting the risk a patient is at to entering AF, even in the presence of a medical professional. Future efforts could investigate whether or not the success of the techniques used in this project can be repeated in the analysis of ventricular tachycardias, such as predicting the onset of ventricular fibrillation. Additionally, a review into how medical professionals decide when to stop an EP study could identify areas for improvement and lead to subsequent changes to the criteria that are currently used.

The main contributions of this report are: the application of machine learning techniques to evaluate the degree of fractionation of a patient's atrial response and predict the risk of atrial fibrillation during EP studies; the introduction of a continuous metric for the degree of fractionation; the use of features conditioned on typical responses; the use of data augmentation to synthesise fractionated responses; the development of novel measures for the conduction delay, and duration, of a patient's atrial response; and a robust peak detection algorithm.

Contents

1 Introduction

Over two million people in the UK are affected by abnormalities of the heart's rhythm, more commonly known as cardiac arrhythmias, which lead to complications varying from mild palpitations to strokes or even death in the most extreme cases [1]. The economic cost is significant, accounting for an estimated 0.9-2.4% of the total NHS budget in 2000 [2]. Cardiac electrophysiology (EP) studies are a relatively recent advancement in the management of arrhythmias, which have previously relied on clinical history and ECG recordings. Catheter electrodes are inserted into the heart and used to electrically stress cardiac tissue, enabling an in-depth analysis of the causes of the arrhythmia, and in many cases the ablation[1] of the cardiac tissue responsible for the abnormality. These studies are now the recommended diagnostic and treatment strategy for most arrhythmias [3].

A serious drawback of using EP studies to electrically stress the heart is that in many cases it is too aggressive and there is a danger of putting the patient into a state of atrial fibrillation (AF). Without immediate intervention and subsequent recovery, the patient is at significant risk of life threatening complications, including cardiac arrest and stroke [4]. The current methodology for assessing the risk a patient is at to entering AF relies upon monitoring simple features of the patient's response, such as conduction delay and number of peaks, with a crude visual interpretation being made by the cardiologist [5]. This motivates the overarching goal of this project - to use advanced signal processing, feature extraction and analysis to develop a more accurate and rigorous process for evaluating this risk.

1.1 Electrograms, Fractionation and Induced Atrial Fibrillation

Inserting catheter electrodes into the heart allows changes in the potential field to be recorded, from which cardiologists can observe how electrical activity spreads through the cardiac chambers, and localise the source of any abnormality [6]. Electrical pulses can also be delivered through the catheters to trigger a heart beat, permitting external control, or pacing, of a patient's heart rate [4]. Typically, multiple catheters are inserted at specific locations[2] from where recordings, collectively referred to as electrograms, are taken. Figure 1 provides a snapshot of part of an electrogram over a 750ms interval. The two green signals highlighted by the arrows are the patient's *atrial response*[3], and the preceding 'spikes', shown in red, occur at the instance of the delivery of an electrical pulse.

[1] The delivery of electrical energy to destroy the cardiac tissue.

[2] One of which being the coronary sinus.

[3] The fluctuations in voltage due to the electrical activity of the atria, as recorded by electrodes located in the coronary sinus.

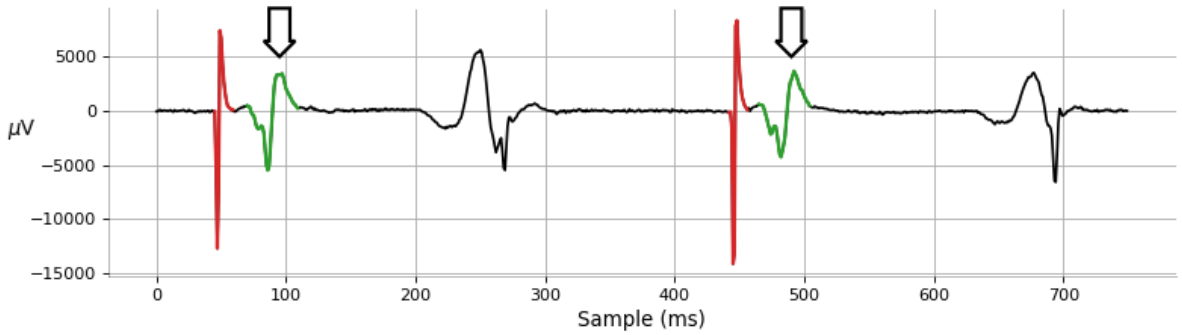
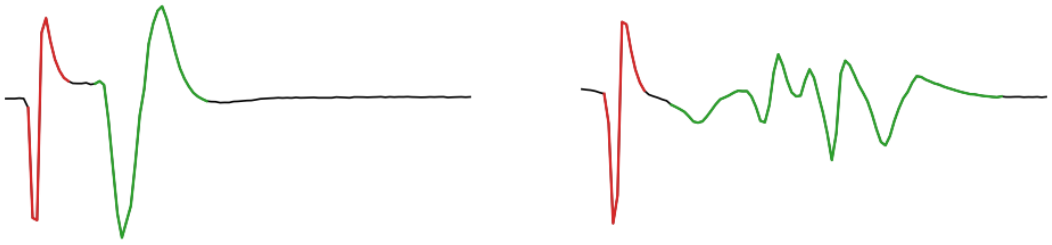


Figure 1: A snapshot of an electrogram recording. The arrows identify the two atrial responses, which are shown in green. The pacing 'spikes' are shown in red.

When pacing is performed from the coronary sinus, under normal circumstances the two responses highlighted in Figure 1 should have the same shape and conduction delay[4]. As the pacing interval[5] is decreased, the conduction of cardiac tissue will decrease slightly, resulting in an increase in conduction delay. If the atrial tissue is diseased, then as the pacing interval is decreased the conduction of cardiac tissue can become abnormal as the electrical activity spreads in a chaotic manner through the atria and coronary sinus. This is characterised by a significant change in the shape of the patient's atrial response, which becomes prolonged and fragmented, or 'fractionated' [7]. Figure 2 illustrates the changes from a) a normal response to b) a fractionated response.



(a) A normal, non-fractionated, response.

(b) Significant fractionation indicated by high-frequency oscillations and prolonged duration.

Figure 2: A comparison between a) a normal response and b) a fractionated response, following the pacing 'spike'. The colour scheme used in Figure 1 is adopted.

If the pacing interval continues to decrease beyond this point, then this chaotic conduction can become so severe that atrial tissue activation breaks down entirely, at which point

[4]The interval between the delivery of an electrical pulse and patient's atrial response.
[5]The interval between successive pulses.

the patient goes into a state of AF[6]. AF is an irregular rhythm, in which waves of activity circulate the atria continuously. If left untreated this abnormal activity will continue, putting the patient at significant risk of stroke [4]. For around 50-70% of patients, a large electric shock (3000-5000 Volts) is necessary to recover from the AF state [5].

1.2 The Need for a Solution

Precautions are taken to avoid inducing AF in patients during EP studies by rarely pacing at less than 230-240ms intervals and by monitoring for signs of fractionation, such as delayed, prolonged and fragmented atrial responses [5], as seen in Figure 2. However, the difficulty of evaluating the risk a patient is at to entering AF during EP studies cannot be understated. Cardiologists are expected not only to monitor 14 separate electrogram recordings moving across a screen in real-time, each showing electrical activity from a distinct location, but also to take into account the differing degrees of fractionation in each recording due to their location relative to the source of arrhythmia, large amounts of natural variability between the responses of patients and the presence of other, non-AF related irregularities such as atrial premature beats. This is all in addition to the need for a diagnosis. This is an extremely demanding task, even for an experienced cardiologist. Moreover, within the NHS the use of EP studies for procedures such as catheter ablation has increased significantly in recent years [8]. It can therefore be expected that many cardiologists conducting these studies are relatively inexperienced, and are even more likely to miss the signs indicative of AF. This prompts the question - could these warning signs be detected, and evaluated, automatically?

Existing papers [9–13] have shown a systematic link between patients at risk of entering AF and an increase in conduction delay and duration of atrial response as the pacing interval is reduced. These studies make no attempt to classify atrial responses using features beyond conduction delay and duration. Moreover, methods to quantify the degree of fractionation have been based upon seemingly arbitrary definitions of fractionation [14, 15]. No known attempts to automate the detection of fractionation with the aim of preventing induced AF have been made.

[6]This state of AF is referred to as *induced* AF, which is distinct from general AF. Anyone, with or without an arrhythmia, can be forced into a state of AF if the heart is put under sufficient electrical stress.

1.3 Project Goals and Specification

The goals of this project are as follows:

1. To develop a model that 'automatically' determines the degree of fractionation of a patient's atrial response.
2. To evaluate the ability of this model to predict the onset of induced AF during EP studies.

As discussed, signs of fractionation include delayed, prolonged and fragmented responses. These are all 'features' of the response that would be used collectively to determine the degree of fractionation. Clearly these are not the only features indicative of fractionation, just those that can be easily evaluated by visual inspection. It is expected that a judicious selection of these features would provide enough information to determine the degree of fractionation of a response. This motivates the following design choice and expansion of our first goal: to build an automated model based upon sophisticated feature extraction and supervised machine learning techniques.

Supervised machine learning implies the existence of target labels, however no such standard definition of fractionation exists [14, 15]. After a discussion with several cardiologists to identify a form of categorisation that would be most helpful in real-time, it was agreed that responses would be assigned one of three labels:

Green A green label indicates the response shows no signs of fractionation. The pacing interval can be safely decreased without risk of entering AF.

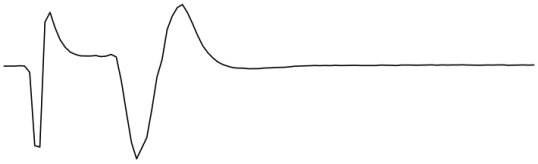
Amber An amber label indicates the response shows very mild fractionation, however retains most of its original shape. The pacing interval can be decreased, but the response may soon become more severely fractionated.

Red A red label indicates that the response has become significantly fractionated. At this point, decreasing the pacing interval could result in the patient entering AF.

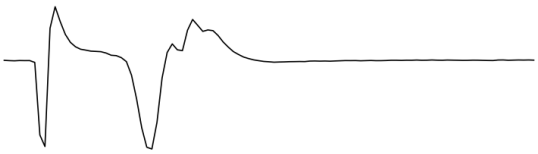
This categorisation scheme closely aligns with how cardiologists instinctively interpret a patient's response, ensuring that the model's output is intuitive to its users. Figure 3 provides examples of 'green', 'amber' and 'red' responses, respectively. All labels were confirmed individually by Dr. Andrew Grace, a leading cardiologist based at Papworth Hospital.

The performance of the model in determining the degree of fractionation of an electrogram signal will be measured by its ability to assign previously unseen signals into one of the three categories above with high accuracy. Whether or not the model is of use in predicting the onset of induced AF will be determined by its ability to correctly label patients as

being at risk of AF or not.



(a) Green response: no sign of fractionation.



(b) Amber response: very mild fractionation, as indicated by the additional deflections in the main peak.

(c) Red response: severe fractionation with little resemblance to the normal 'green' response.

Figure 3: A comparison between the different categories of fractionation.

2 Related Work

The research conducted in this area is limited. Both Qin et al. [9] and Plantanov et al. [10] identify a statistically significant relationship between patients with AF and an increase in conduction delay as the pacing interval is reduced. Tanigawa et al. [11] find a close relationship between the duration and number of fragmented deflections of atrial response and the vulnerability of the atrial muscle. Tai et al. [12] go further, concluding that prolonged electrogram activity may be related to the development of AF. These relationships are supported by Pytowski et al. [13], whose conclusions hinge upon 'intra-atrial conduction curves' (a plot of the number of peaks, and associated delays, of the electrogram signal).

Within the two most commonly used electroanatomical mapping systems, EnSite and CARTO, algorithms exist to identify 'Complex Fractionation Atrial Electrograms' (CFAEs) for the purpose of catheter ablation [14, 15]. However, in both cases, seemingly arbitrary

definitions of fractionation are used, based upon the intervals between successive deflections [14, 15]. An attempt to automate the detection of fractionation metrics (number of peaks, electrogram duration and conduction delay) in patients with ventricular tachycardias (VTs) was made by Gupta, Redfearn, Hashemi and Akl [16]. The analysis of their results is limited, providing only a comparison of feature values between groups of patients. They conclude that automated feature extraction is possible.

The lack of related work is reflective of the historical infrequency of EP studies relative to other diagnostic procedures in the treatment of cardiac arrhythmias. However, as discussed, the use of EP studies for procedures such as catheter ablation has risen significantly in recent years [8]. Thus, research involving the analysis of EP recordings is of increasing importance. Similar to the analysis of electrocardiogram recordings [17–19], the most significant developments in this area can be expected to come from within the machine learning community. This project is one of the first.

3 Background Theory

3.1 Supervised Machine Learning: Classification

Supervised machine learning describes the process of 'learning' an adaptive mathematical model, say $\mathbf{y}(\mathbf{x})$, that captures regularities within data through the availability of a training dataset, $\mathcal{D}$, comprised of a set of input vectors $\mathbf{x}$ along with their corresponding target vectors $\mathbf{t}$ [20]. Specifically, the training dataset $\mathcal{D}$ is used to tune the parameters, $\boldsymbol{\theta}$, of the model such that when given an unseen input vector $\mathbf{x}^*$, $\mathbf{y}(\mathbf{x}^*)$ accurately predicts its target vector $\mathbf{t}^*$. As we shall see, this is achieved by finding the parameters which minimise a loss function $E(\boldsymbol{\theta}, \mathcal{D})$,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, E(\boldsymbol{\theta}, \mathcal{D}). \tag{1}$$

Typically, $E(\boldsymbol{\theta}, \mathcal{D})$ is a proxy for the inverse-likelihood of the training data.

Classification problems are those in which the goal is to assign input vectors to one of a finite number of classes $\{\mathcal{C}_k\}_{k=1}^K$, e.g. predicting a person's gender given an image of their retina. The performance of a classification algorithm is determined by its ability to correctly classify unseen data, which forms the test dataset. In most cases, the training data comprises only a 'handful' of all possible input vectors - this makes *generalisation*, the ability of the model to adapt to unseen data, a principal goal of machine learning. Broadly speaking, we can divide classification algorithms into three distinct approaches:

Discriminant Functions whereby a function $f(\mathbf{x})$ is learnt that directly assigns each

input $\mathbf{x}$ into a specific class.

Discriminative Models whereby the posterior class probability $p(\mathcal{C}_k|\mathbf{x})$ is learnt, and the subsequent assignment of $\mathbf{x}$ is made.

Generative Models whereby the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ are learnt and combined with prior class probabilities $p(\mathcal{C}_k)$ to determine $p(\mathcal{C}_k|\mathbf{x})$ using Bayes' theorem.

There are strengths and weaknesses to each approach [20]. Discriminant functions are often the easiest to learn, however provide no measure of uncertainty in their predictions. In contrast, both discriminative and generative models learn $p(\mathcal{C}_k|\mathbf{x})$, which implicitly provides a measure of confidence the model has in its predictions. Whilst there are benefits to modelling $p(\mathbf{x}|\mathcal{C}_k)$, generative models are the most computationally demanding to learn and are often unnecessary when we are only interested in $p(\mathcal{C}_k|\mathbf{x})$ [20].

3.2 Logistic Regression Classifier

Consider the case in which we wish to classify an input vector $\mathbf{x}$ into one of K classes, $\{\mathcal{C}_k\}_{k=1}^K$. Our training data, $\mathcal{D} = \{\mathbf{x}_i, \mathbf{t}_i\}_{i=1}^N$, consists of N input/output pairs $(\mathbf{x}_i, \mathbf{t}_i)$, where $\mathbf{t}_i \in \{0,1\}^K$ represents a 1-of-K coding scheme of the class assignment of $\mathbf{x}$ such that $t_{ik} = 1$ if $\mathbf{x}_i \in \mathcal{C}_k$, and $t_{ij} = 0$ if $\mathbf{x}_i \notin \mathcal{C}_j$. The logistic regression classifier, as applied to this multiclass problem, models the posterior probability of class $\mathcal{C}_k$ as a softmax function acting on linear functions of feature values

$$p(\mathcal{C}_k|\mathbf{x}) = y_k(\mathbf{x}) = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)} \tag{2}$$

where $a_k = \mathbf{w}_k^T\mathbf{x}$ is termed the activation for class $\mathcal{C}_k$, and $\{\mathbf{w}_k\}_{k=1}^K$ are weight vectors and are the parameters of the model to be tuned. Introducing weight matrix $\mathbf{W}$ such that $\mathbf{W}_{kj} = w_{kj}$, we may now form an expression for the likelihood of our training dataset:

$$p(\{\mathbf{t}_i\}_{i=1}^N|\{\mathbf{x}_i\}_{i=1}^N, \mathbf{W}) = \prod_{i=1}^N \prod_{k=1}^K y_k(\mathbf{x}_i)^{t_{ik}}. \tag{3}$$

Taking the natural logarithm of the right hand-side in Equation 3 gives an expression for the log-likelihood, the negative of which shall be denoted $\mathcal{L}(\mathcal{D}, \mathbf{W})$:

$$\mathcal{L}(\mathcal{D}, \mathbf{W}) = -\sum_{i=1}^N \sum_{k=1}^K t_{ik} \ln y_k(\mathbf{x}_i). \tag{4}$$

To prevent overfitting of the model to the training data, a regularisation term is added to obtain the loss function,

$$E(\mathcal{D}, \mathbf{W}) = \mathcal{L}(\mathcal{D}, \mathbf{W}) + cR(\mathbf{W}) \tag{5}$$

where c is a constant. $R(\mathbf{W})$ is used to penalise large weights, and its form is a design choice. Two standard choices are either the L1-regularisation of weights, $R(\mathbf{W}) = \sum_{k=1}^{K}\sum_{j=1}^{d}|w_{kj}|$, which encourages the sum of absolute values of weights to be small, or the L2-regularisation of weights, $R(\mathbf{W}) = \sum_{k=1}^{K}\sum_{j=1}^{d}w_{kj}^2$, which encourages the sum of squared values of weights to be small. Applications of L1-regularisation, such as the Lasso algorithm [21], have been found to strongly encourage many of the weights to equal zero. In situations in which many features are believed to be 'useless', the use of L1-regularisation naturally identifies the most relevant features [22].

We wish to find the weights, $\mathbf{W}^*$, which minimise the loss-function,

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}}\, E(\mathcal{D}, \mathbf{W}). \tag{6}$$

Unfortunately there is no closed form solution for the above expression, however it is concave [22] and therefore has a unique minimum. An iterative procedure can be employed, the most basic of which is gradient descent. The gradient descent algorithm applies the update procedure at iteration $\tau + 1$

$$\mathbf{W}^{\tau+1} = \mathbf{W}^{\tau} - \eta\nabla E(\mathcal{D}, \mathbf{W})\Big|_{\mathbf{W}=\mathbf{W}^{\tau}} \tag{7}$$

where η controls the step-size at each iteration. It is straightforward to show that

$$\nabla_{\mathbf{w}_k} E(\mathcal{D}, \mathbf{W}) = \sum_{i=1}^{N} y_k(\mathbf{x}_i - t_{ik})\mathbf{x}_i + c\nabla_{\mathbf{w}_k} R(\mathbf{W}) \tag{8}$$

where $\nabla_{\mathbf{w}_k} R(\mathbf{W})$ depends on the type of regularisation used.

Since the activations $\{a_k\}_{k=1}^{K}$ are formed from a linear combination of feature values, the decision boundary will be linear in feature space. Classifiers with linear decision boundaries are termed *linear models*. An alternative to the logistic regression classification algorithm, in which decision boundaries are not linear in feature space, is the naïve Bayes classifier, which shall be described in the next Section.

3.3 Naïve Bayes Classifier

Similarly to the logistic regression classifier, the naïve Bayes classifier models the posterior probability of each class, $\mathcal{C}_k$, given an input vector $\mathbf{x} \in \mathbb{R}^d$,

$$p(\mathcal{C}_k|\mathbf{x}) = p(\mathcal{C}_k|x_1, ..., x_d) \tag{9}$$

where the dependence on each feature of $\mathbf{x}$ has been emphasised. Using Bayes' theorem, we can express Equation 9 as

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{p(x_1, ...x_d|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}. \tag{10}$$

The naïve Bayes assumption is that each of the features of $\mathbf{x}$ are mutually independent given class $\mathcal{C}_k$, such that $p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_d, \mathcal{C}_k) = p(x_i|\mathcal{C}_k)$. This reduces Equation 10 to

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}) &= \frac{p(x_1, ...x_d|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \\ &= \frac{p(x_1|\mathcal{C}_k)...p(x_d|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \\ &= \frac{p(\mathcal{C}_k)\prod_{j=1}^{d} p(x_j|\mathcal{C}_k)}{p(\mathbf{x})} \end{aligned} \tag{11}$$

where $p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathcal{C}_k)\prod_{j=1}^{d} p(x_j|\mathcal{C}_k)$ is the normalisation constant. An estimate for $p(\mathcal{C}_k)$ can be formed from the proportion of training data belonging to class $\mathcal{C}_k$. The form of $p(x_j|\mathcal{C}_k)$ is a design choice. When x_j is continuous, a common choice for $p(x_j|\mathcal{C}_k)$ is the Gaussian distribution,

$$p(x_j|C_k) = \mathcal{N}(x_j; \mu_{kj}, \sigma_{kj}^2) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp\left(-\frac{(x_j - \mu_{kj})^2}{2\sigma_{kj}^2}\right) \tag{12}$$

where μ_{kj} and σ_{kj} are the mean and standard deviation of the Gaussian distribution of feature x_j for class $\mathcal{C}_k$, respectively.

As with the logistic regression classifier, the optimum parameters of the model $\boldsymbol{\theta}^*$ can be found by maximising the log-likelihood. If all distributions $p(x_j|\mathcal{C}_k)$ are chosen to be Gaussian, the log-likelihood is given by

$$\ln p(\{\mathbf{t}_i\}_{i=1}^N|\{\mathbf{x}_i\}_{i=1}^N, \boldsymbol{\theta})) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d}\left\{-\frac{1}{2}\ln 2\pi\sigma_{kj}^2 - \frac{1}{2\sigma_{kj}^2}(x_{ij} - \mu_{kj})^2\right\} + C \tag{13}$$

where C has no dependence on parameters $\boldsymbol{\theta}$ and $\{\mathbf{t}_i\}_{i=1}^N$ are the same 1-of-K coded target

vectors as described in the previous Section. The learning objective is then

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathcal{L}(\mathcal{D}, \boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \left\{ -\frac{1}{2}\ln 2\pi\sigma_{kj}^2 - \frac{1}{2\sigma_{kj}^2}(x_{ij} - \mu_{kj})^2 \right\}. \tag{14}$$

Unlike the form of the log-likelihood for the logistic regression classifier, the expression above has a unique closed-form solution given by:

$$\mu_{kj}^* = \frac{\sum_{i=1}^{N} t_{ik} x_{ij}}{\sum_{i=1}^{N} t_{ik}} \tag{15}$$

$$\sigma_{kj}^{2\,*} = \frac{\sum_{i=1}^{N} t_{ik}(x_{ij} - \mu_{kj}^*)^2}{\sum_{i=1}^{N} t_{ik}}. \tag{16}$$

3.4 Feature Extraction and Selection

Consider the task of classifying a signal into one of two categories. It is expected that the signal will contain a great deal of redundant information, and that only a few characteristic features of the signal are necessary for classification. Moreover, if the signals are recorded over a significant time period with a high sampling rate then the signal will comprise a large number of data points. Treating the raw signal values as input vectors would be computationally unfeasible, and more importantly unnecessary. In circumstances such as this, we would seek to extract features from the raw data values and concatenate them to form a feature vector describing the signal. For example, features such as the standard deviation, mean, maximum and minimum value can be readily extracted and are often sufficient to perform signal classification. However, in practice we may have limited knowledge about which features are relevant to the classification task. Feature selection describes the process of selecting a subset of features which can distinguish between classes.

There exist three distinct categories of feature selection methods: filter methods, wrapper methods and ensemble methods [23]. Filter methods select a subset of features based upon a statistical measure of the relationship between the feature and the output[7], such as correlation or mutual information. Wrapper methods search through subsets of features, training a machine learning algorithm on each subset and selecting the subset that results in the best predictive performance. These methods often result in the subset of features with the best performance, however they are usually the most computationally expensive [23]. Unlike wrapper and filter methods, in ensemble methods feature selection is *implicit*, and is incorporated as part of the model building process - for example, by including

[7] In more sophisticated methods, the relationship between the feature and other features is also taken into account.

an L1-regularisation term in the loss function of the logistic regression classifier. As discussed, this form of regularisation encourages sparsity amongst the weights such that those corresponding to features with little 'relevance' are driven to zero.

3.5 Data Augmentation

Within classification problems, data augmentation is a technique used to generate more training data when the original training data is not sufficient to learn a model that can generalise [24]. For images, generation of new training data is typically done by cropping, or rotating, the original data [25]. The goal is to provide the model with a more diverse representation of each class, such that it generalises better to unseen data.

Data augmentation for time-series classification problems is not as commonly used as in image classification problems. However, techniques do exist. One such approach, introduced by Um et al. [26], is described as follows. Given a time-series, $\mathbf{s} \in \mathbb{R}^m$, augmentation is achieved by distorting the magnitude by a cubic spline $f(x)$. $f(x)$ consists of N polynomial pieces $\{f_i(x)\}_{i=0}^{N-1}$ of the form

$$f_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{17}$$

generated from a set of $N + 1$ evenly spaced points $\{(x_i, y_i)\}_{i=0}^{N}$, where $x_i = \frac{mi}{N}$ and $y_i \sim \mathcal{N}(y_i; 1, \sigma_A^2)$. For $1 \leq i \leq N - 1$, the following constraints are imposed:

$$\begin{aligned} &f_0(x_0) = y_0 \\ &f_{i-1}(x_i) = f_i(x_i) = y_i \\ &f'_{i-1}(x_i) = f'_i(x_i) \\ &f''_{i-1}(x_i) = f''_i(x_i) \\ &f''_0(x_0) = f''_{N-1}(x_{N-1}) = 0. \end{aligned} \tag{18}$$

Figure 4 illustrates three randomly generated cubic splines using $\sigma_A = 0.1$, 0.3 and 0.5. It can be inferred that as σ_A increases, the magnitude of the distortion of the cubic spline also increases. From here on, we shall refer to σ_A as the degree of distortion. The distorted time-series, $\tilde{\mathbf{s}}$, has elements

$$\tilde{s}_i = f(i)s_i. \tag{19}$$

Since each cubic spline is generated from random variables $y_i \sim \mathcal{N}(y_i; 1, \sigma_A^2)$, the cubic splines themselves are random. Thus, an arbitrary amount of new time-series data can be generated from a finite set of original data.

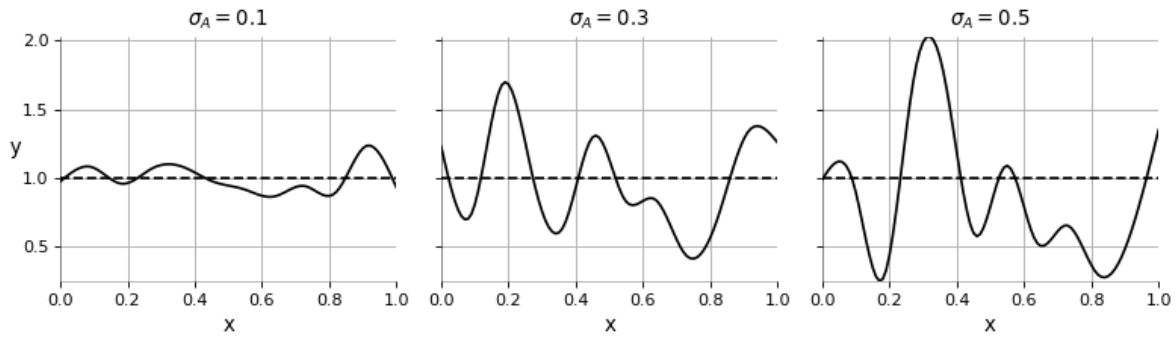


Figure 4: Three examples of randomly generated cubic splines for $\sigma_A = 0.1$, 0.3 and 0.5 with $N = 5$.

4 Clinical Methodology

The data used in the project was obtained from a retrospective study of patients undergoing EP studies at Papworth Hospital over the last six years. For each patient, the data collected using a pacing technique known as the Antegrade Curve was available. Catheter probes, similar to the illustration provided in Figure 5, are inserted in various cardiac chambers including the coronary sinus. Along each catheter are a collection of bi-polar electrodes, each with the capability to either record voltage fluctuations in the surrounding cardiac tissue or deliver electrical pulses.

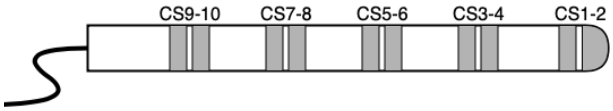


Figure 5: An illustration of the catheter probe inserted into the coronary sinus. Five bi-polar electrodes, CS1-2, CS3-4, CS5-6, CS7-8 and CS9-10, are situated along the probe.

Pacing is performed from electrodes positioned at the rear of the catheter in the coronary sinus (electrodes CS9-10 or CS7-8 in Figure 5) and consists of eight evenly spaced pulses (S1) at a pacing interval of 600ms, followed by an additional pulse (S2) at a shorter interval of 400ms, as illustrated in Figure 6. The purpose of the S1 pulses are to regularise the heart, prior to electrically stressing it with a premature S2 pulse. It is the patient's response to the S2 pulse which is of interest to the cardiologist[8]. The process is repeated, with the interval before the extra pulse reduced in 10ms intervals until either enough information to make a diagnosis is gathered, or an abnormal rhythm is induced.

[8]The S2 pulse probes the heart to reveal abnormalities (such as fractionation), which subsequently enables a diagnosis to be made.

Figure 6: S1 pulses at a fixed (600ms) interval, followed by a premature S2 pulse.

Only patients with recordings that met the following criteria were included in the study:

1. The EP study was performed with a catheter located in the coronary sinus, with pacing performed from electrodes CS9-10 or CS7-8 to allow signals from electrodes CS5-6, CS3-4 and CS1-2 to be recorded.
2. Pacing was performed with intervals covering at least a 100ms range.
3. Visual inspection indicated a stable catheter position and minimal noise.
4. Detailed records were available to determine if the patient had a history of arrhythmias.
5. Recordings over the entire study were available to determine whether or not AF was induced at any point.

Patients were divided into two groups, the first being those with no history of AF and who underwent an EP study without entering induced AF, and the second being those who developed AF in response to one of the pacing manoeuvres during the EP study. No patients developed AF in response to the standardised Antegrade Curve, but all had sustained (>30s) AF after pacing within 10 minutes of completion. The reason for this was that it was deemed unethical to induce AF in patients for the purpose of the study, and an experienced member of staff was on hand to halt the test if the risk was deemed unacceptable[9]. Furthermore, pacing was not performed at short intervals in patients if it was evident that this would provide no additional diagnostic value. All patients included in the study had some form of arrhythmia involving the atria, including atrial tachycardia (AT), atrio-ventricular nodal re-entry tachycardia (AVNRT) and atrio-ventricular reentrant tachycardias (AVRT).

A training dataset, consisting of nine patients who entered AF and 37 who did not, was available from the beginning of the project. A test dataset was later made available, consisting of 10 patients who entered AF and 30 who did not, upon which the performance of any model trained on the original dataset would be assessed.

[9]There is no quantitative threshold for this - it is based on visual analysis of the recordings by the cardiologist performing the EP study.

5 Methodology

To avoid confusion, Table 1 provides some of the nomenclature that will be used throughout the rest of the report.

Key Word	Meaning
Atrial response	Fluctuations in voltage recorded at an electrode due to the electrical activity of the atria.
Segment	A fixed length snapshot of time-series data encapsulating an individual patient's atrial response recorded at a single electrode.
Pulse	An externally induced electrical stimulus, typically delivered from an electrode towards the rear of a catheter inserted in the coronary sinus.
S1/S2 pulse	As illustrated in Figure 6, the S1 pulses are used to regularise the heart and are followed by a premature S2 pulse.
S1/S2 interval	The interval between the final S1 and S2 pulse.
AF Patients	Patients who experienced induced AF during the course of the EP study.
Non-AF Patients	Patients who did not experience induced AF during the course of the EP study.

Table 1: Nomenclature used throughout the report.

5.1 Data Pre-processing

For each patient, the data from the EP study is stored in a collection of text files, corresponding to each step of the Antegrade Curve. Each text file consisted of a unique title comprised of a patient I.D. and S1/S2 interval (e.g. 'AFPATIENT1-0230.txt', where 'AFPATIENT1' is the I.D. and '0230' indicates an S1/S2 interval of 230ms), and contents including header information detailing the properties of the electrodes (sampling rate[10], electrode name and voltage range) together with data collected from each electrode over a 2.5s interval. The 2.5s interval captures the patient's atrial response to the S2 pulse at the corresponding step of the Antegrade Curve, as well as the patient's atrial response to the preceding two S1 pulses. This is illustrated in Figure 7. The dashed lines identify the location of the S1 and S2 pulses, and the red segments highlight the patient's corresponding atrial response. It is important to note that although data recorded from 12-14 electrodes is present in each file, only the data recorded from the coronary sinus electrodes (CS1-2, CS3-4, CS5-6, CS7-8 and CS9-10) is of interest. This is because the remaining electrodes do not capture the patient's atrial response (rather they record the patient's ventricular response, or are electrocardiogram (ECG) signals).

[10]This was 1000Hz in all cases.

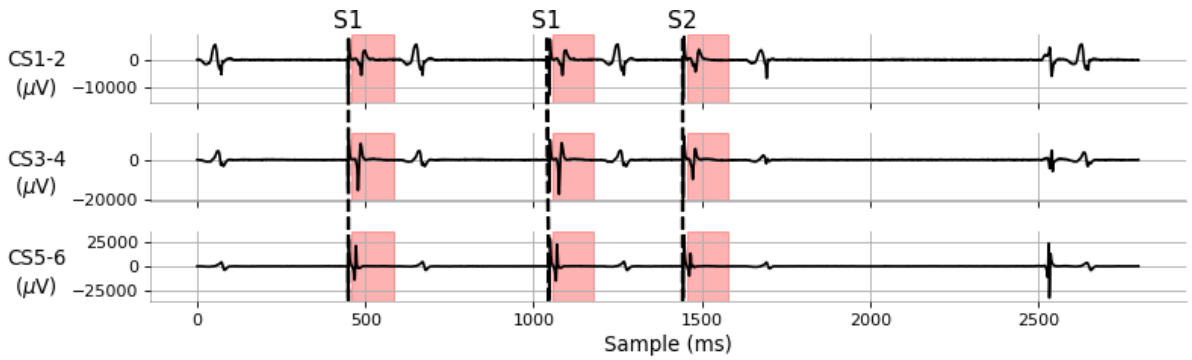


Figure 7: Identification of the S1/S2 pulses, and the corresponding atrial responses contained in an individual text file.

As pacing is performed from electrodes CS9-10 or CS7-8, only data from electrodes CS5-6, CS3-4 and CS1-2 was to be extracted. It was assumed that the atrial response of all patients falls within 125ms following the electrical pulses. In practice, this assumption was found to be accurate and simplified the data extraction process significantly.

5.1.1 Data Parsing

A Python module was built to parse the information contained in each text file into a Pandas DataFrame. This format kept the data well structured whilst permitting the necessary data analysis.

5.1.2 Identifying the S1/S2 Pulses

Given that pulses were delivered from electrodes CS9-10 and CS7-8, it was expected that sharp 'peaks' in recordings would mark the instance of a pulse. Robust detection of these 'peaks' could be achieved by first convolving the signal with a rectangular window, and then applying a peak detection algorithm to locate the maxima. However, after an initial investigation into the form of these recordings it was found that consistent 'clipping' occurred at the instance of the pulse, as seen in Figure 8.

To account for this, an alternative approach was developed to detect the location of the S1 and S2 pulses with very high accuracy. First, positions at which the amplitude of the signal exceeded 95% of the maximum value were initially identified. Moving forwards (in time) through these positions, removing those at a separation of less than 200ms following the previous position, ensured that the pulse detector 'fired' only the rising edge of the clipping, marking the start of the delivery of a pulse. The intervals between each neighbouring pair of remaining positions was then calculated. The interval closest to the

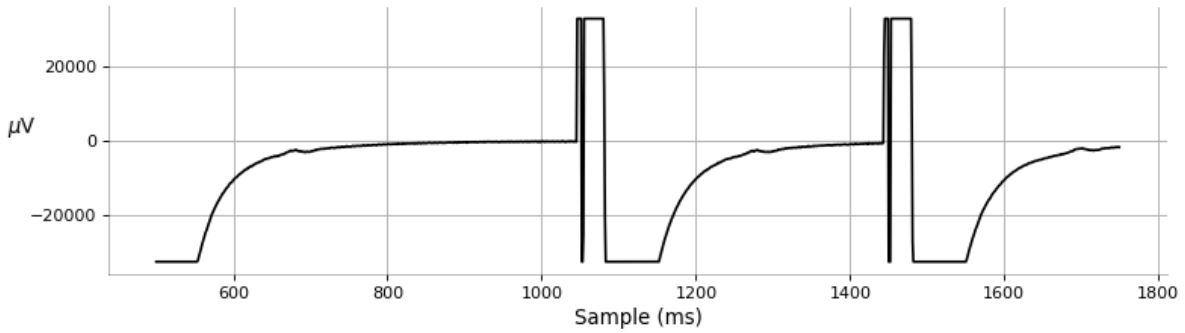


Figure 8: Clipping of the signal from the CS9-10 electrode at the point at which a pulse is delivered.

S1/S2 interval specified in the file name was identified, and the corresponding S2 pulse marked. The remaining positions were marked as S1 pulses.

This technique was found to be remarkably effective, not only enabling correct detection of all S1 and S2 pulses but additionally estimating the S1/S1 and S1/S2 intervals to an accuracy of 99.8% across the 785 files available.

5.1.3 Segment Extraction

Segments from each of the CS1-2, CS3-4 and CS5-6 electrode recordings were extracted according to the start and end positions, relative to the detected pulses, provided in Table 2. For example, suppose an S2 pulse is detected at position i. The segment extracted from the recording by electrode CS1-2 contains the values between positions $i + 27$ and $i + 152$.

	Electrode		
	CS1-2	CS3-4	CS5-6
Start (ms)	27	23	10
End (ms)	152	148	135

Table 2: Start and end positions, relative to the pulse location, of the segment extracted from each electrode recording.

Natural conduction delay of the pulse between the electrode locations meant that a lag existed between the atrial response as recorded in each electrode. This lag is visible in Figure 9 - the pulse reaches each electrode simultaneously at 5ms from the start of the recording, however there is a much greater delay between the atrial response recorded by electrode CS1-2 compared with CS5-6, with CS3-4 intermediate.

It is known that conduction delay (i.e. the interval between the start of the segment and

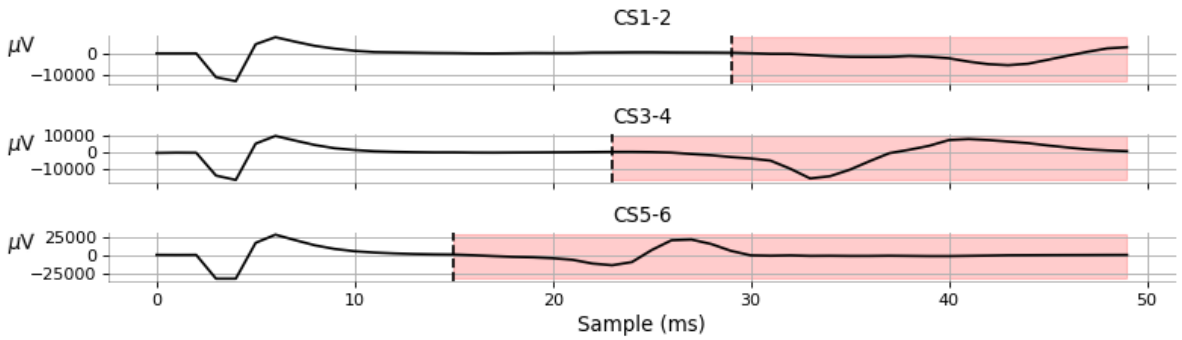


Figure 9: The conduction lag between electrodes CS1-2, CS3-4 and CS5-6. The part of the signal that is extracted from each electrode recording is highlighted in red.

atrial response) is associated with the degree of fractionation of the response. The values in Table 2 reflect an increase in delay for electrodes further from the site of the atrial response. It is desirable[11] that the interval between the start of each segment and the atrial response is roughly the same across all segments, and so accounting for this natural delay was deemed necessary.

To summarise, the input to the data extraction process for an individual patient consists of a collection of text files containing electrogram recordings at each step of the Antegrade Curve. Suppose the patient has recordings taken from an S1/S2 interval of 400ms to 220ms in steps of 10ms - the input then consists of 19 text files. As described earlier, each text file records 2.5s worth of data from 14 electrodes, capturing the patient's response to a single S2 pulse and the two preceding S1 pulses. Extracting these three responses from electrodes CS1-2, CS3-4 and CS5-6 gives nine segments in total. Since this is done across all 19 text files, the output is $9 \times 19 = 171$ segments with composition detailed in Table 3.

	Electrode		
S1/S2	CS1-2	CS3-4	CS5-6
S1	38	38	38
S2	19	19	19

Table 3: An example of the composition of segments outputted by the data extraction process for a single patient.

[11] Variations in conduction delay are known to contribute to the degree of fractionation; however, if we don't account for the natural variations in conduction delay between electrodes, a model trained on segments from all electrodes will learn that large variations in conduction delay are often not indicative of fractionation.

5.2 Data Labelling

As outlined in Section 1.3, individual segments were assigned into one of three categories ('green', 'amber' and 'red') with confirmation from Dr. Andrew Grace, a leading cardiologist at Papworth Hospital. This labelling scheme closely aligned with cardiologists' instinctive interpretation of a patient's response, ensuring that the output of the model is intuitive to its users. However, these labels alone were not sufficient to evaluate the full complexity of a cardiologist's considerations. It was necessary to collect information that summarised cardiologists' interpretation of the progression of a patient's response over the entire Antegrade Curve, not just the interpretation of each response in isolation. To achieve this, a questionnaire was designed and presented to a number of medical professionals individually. For each patient, the medical professional was asked to evaluate the following:

1. The degree to which the response becomes delayed[12].
2. The degree to which the response becomes fractionated.
3. The first sign of fractionation (coupling interval and channel).
4. Whether the EP study have been stopped early.

This information can be used to evaluate the performance of the model in determining the degree of fractionation. A 'useful' model is not only expected to be able to accurately determine the degree to which the response becomes fractionated, but also identify the first sign of fractionation and suggest whether or not the EP study should be halted.

A single page of the questionnaire is shown in Figure 10. The image on the page displays the progression of the atrial response of a single patient as recorded by electrodes CS1-2, CS3-4 and CS5-6 (displayed from left to right). At the top, the image shows a longer segment illustrating the location from which the segment is extracted from the electrode recording. This was included to aid the medical professionals in developing an intuitive understanding for this form of presentation of a patient's response. Moving downwards from the top, the responses for incrementally decreasing S1/S2 intervals are shown. Although most patients do not have responses for the entire range of intervals, it was important to keep the vertical (and horizontal) scale consistent across patients such that the atrial response to a specific S1/S2 interval is at the same vertical position on the page for all patients. AF and non-AF patients were mixed and presented in a random order, such that participants wouldn't develop a bias[13] in their assessments after seeing

[12]Medical professionals typically consider fractionation and conduction delay as two separate, distinct properties of a patient's response. For ease of use, a distinction between the two was explicitly made. To clarify: in this project only fractionation was considered, with conduction delay being a feature that is associated with fractionation.

[13]AF patients' responses are expected to show more fractionation than those of non-AF patients. Seeing

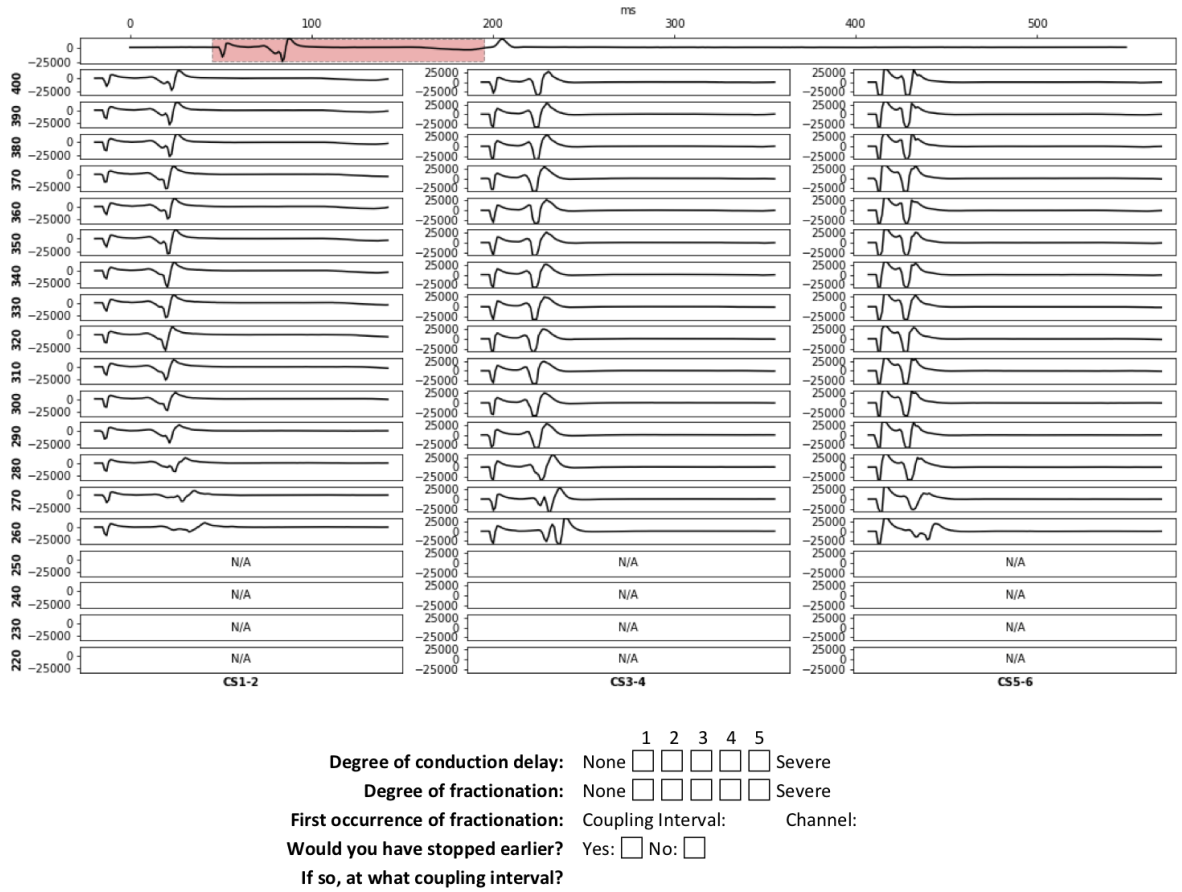


Figure 10: A single page of the questionnaire provided to medical professionals.

many AF (or non-AF) responses back-to-back.

5.3 Choice of Machine Learning Model

The aim was to develop a model that takes an input segment capturing a patient's atrial response to an S2 pulse, and determine the degree to which it is fractionated. The data available for the development of such a model consisted of input/output pairs, whereby the output is a label indicating the class the input segment belongs to. Clearly, this describes the role of a supervised machine learning algorithm. A brief justification for the choice of employing a 'traditional'[14] supervised machine learning model to achieve this was provided in Section 1. This Section begins with further justification for this choice.

The following choices for the development of the model were considered: a classification model based upon a measure of 'closeness' to responses representative of each label (e.g. Dynamic Time Warping), training a recurrent neural network (e.g. a Long-Short-Term-

lots of fractionation in a short space of time may cause participants to become 'insensitive' to it, biasing their assessment of fractionation downwards.

[14]Traditional refers to the use of feature extraction, followed by supervised training of a machine learning algorithm.

Memory network [27]), or building a traditional machine learning algorithm trained on extracted features. In order to identify which of these is most suitable, it is instructive to first consider what is required of it:

1. **Data Efficiency** As is often the case in health sciences [28], the availability of data in the project is limited. It is crucial that the model remains effective in spite of this. The lack of regularisation in RNN models makes them difficult for use in situations with small quantities of data, as the desire to avoid overfitting frequently results in the use of networks that are too small to model the complexity of the problem [29].

2. **Understandable Decisions** An extremely desirable property of the model is for its decisions to be understandable. An RNN would be the epitome of a 'black box' classification model. No medical professional could be expected to 'lift open the lid' of the weight matrices and understand the decisions made. By contrast, a model based on extracted features has the advantage that the features can be engineered to be interpretable. For example, the value associated with the feature 'number of peaks' can help the user to understand why a response was deemed fractionated.

3. **Generalisation** It is essential that the model has the ability to generalise the concept of fractionation, such that it performs highly on unseen data. Classification models based on a measure of closeness do not generalise. Instead, their performance hinges upon the availability of examples that are representative of a particular label. Although this technique has proven successful in many applications [30], it is known that fractionation is characterised by chaotic irregularities. No single example is representative of fractionation and thus a similarity based metric is expected to perform poorly.

Only a traditional machine learning algorithm trained on extracted features is able to meet the requirements we demand of it. Moreover, only those that model class posterior probabilities[15], $p(\mathcal{C}_k|\mathbf{x})$, are of interest. The labels 'green', 'amber' and 'red' are not an intrinsic property of each segment, like 'dog' and 'cat' would be to images of dogs and cats. Rather, they are used as an ordinal measure of an expert's assessment of the degree to which a patient's atrial response appears fractionated. It is anticipated that segments will lie between these labels, e.g. somewhere between mildly fractionated and heavily fractionated. A probabilistic class output permits mapping from feature space to a continuous scale of degree of fractionation by interpolating between class labels. Such an output provides a more useful assessment of a patient's response than a discrete label.

With these considerations in mind, it was decided to investigate and compare the performance of the logistic regression and naïve Bayes classification algorithms. In addition to satisfying each of the criteria outlined above, these algorithms have the advantages of

[15] i.e. discriminative and generative models.

being computationally efficient, easy to understand and relatively simple to implement[16], and have both proved extremely effective in a wide range of medical applications [31–34]. For each model, there remains a number of design decisions that need to be made and investigated. Specifically, the form of the regularisation term in the loss function of the logistic regression classifier and the assumed distribution for feature values in the naïve Bayes classifier. These decisions are described below.

5.3.1 Logistic Regression Classifier Implementation

Recall the form of the regularised loss function derived in Section 3.2,

$$E(\mathcal{D}, \mathbf{W}) = \mathcal{L}(\mathcal{D}, \mathbf{W}) + cR(\mathbf{W}).$$

The form of the regularisation term $R(\mathbf{W})$ is a high-level design choice that influences the values of the weights at the optimum. As discussed, L1-regularisation encourages sparsity amongst the weights, providing implicit feature selection within the model. This enables a much richer analysis of the model - not only can we access its performance on a test dataset, we can also identify the features that play an important role. Thus, it was decided to use L1-regularisation of the form $R(\mathbf{W}) = \sum_{k=1}^{K} \sum_{j=1}^{d} |w_{kj}|$. The parameter c influences the strength of regularisation applied to the weights w_{kj}. Since the optimal choice for c is not known a priori, the dependence of the performance of the model on c will be analysed by training separate models with $10^{-5} \leq c \leq 10^{5}$.

5.3.2 Naïve Bayes Classifier Implementation

The only design choice for the naïve Bayes classifier is the form of $p(x_j|\mathcal{C}_k)$. We shall assume that the feature values associated with each class are distributed according the Gaussian distribution, reproduced here for completeness:

$$p(x_j|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp\left(-\frac{(x_j - \mu_{kj})^2}{2\sigma_{kj}^2}\right).$$

5.4 Feature Extraction and Selection

As discussed in Section 3.4, taking raw time-series data as the input vectors to our model is inefficient. Instead, we seek to extract features from each segment that can be used to distinguish between class labels. Ideally, these features should have the following characteristics:

[16] This reduces the likelihood of 'bugs' present in the code.

Indicative of Fractionation This is more of a necessity than a desire. The model must use the input features to determine the degree of fractionation of the signal.

Interpretable One of the main advantages in using extracted features to determine fractionation is that the features can be chosen to be interpretable by the end user. This allows the user to more readily understand the model's decisions.

Computationally Efficient The model is of limited use if it is unable to analyse signals, and evaluate fractionation in real time. We wish to achieve this without interfering with the medical procedure, hence cannot use features that require inordinate amounts of computational effort as this may impede the natural pace of the EP study.

Table 4 details a selection of features that were found to satisfy each of the above criteria after an initial investigation. The features are divided into two distinct categories: standard features and hand-engineered features. Standard features are those that are regularly used in practice throughout time-series data analysis. Hand-engineered features are those that have been tailored for the purpose of determining the degree of fractionation of a segment. Interestingly, it was found that features involving frequency properties of the segments (e.g. power spectral entropy) did not distinguish between fractionated segments and non-fractionated segments.

Category	Feature	Description
Standard	*Average Magnitude*	The (normalised) mean absolute value of sample values.
	Ratio Above σ	The proportion of (absolute) sample values greater than $\mu + \sigma$.
	Sample Entropy	A measure of the complexity of a short time-series segment. See Section 5.4.2.
Hand-engineered	*Number of Peaks*	The number of significant peaks. See Section 5.4.3.
	Location of Max Energy	A measure of conduction delay of the atrial response. See Section 5.4.1.
	Width of Max Energy	A measure of the duration of the atrial response. See Section 5.4.1.
	Percentage Fractionation	The proportion of the atrial response that is 'fractionated'. See Section 5.4.3.

Table 4: The features that will be used to determine the degree of fractionation of a segment.

For each S2 response[17], these seven features were extracted to form an initial feature

[17]It is only the patient's response to the S2 pulse which we analyse for signs of fractionation, since the

vector, which will be denoted $\phi_1(\mathbf{x})$, where $\mathbf{x}$ are the raw time-series data values for the segment. Whilst these features provide a summary of each segment in isolation (i.e. not conditioned upon any other segment), they do not explicitly take into account *patient variability*. Patient variability refers to the natural variation in atrial responses between patients. For example, it may be typical of patient A to have five peaks in their atrial activity; however, if five peaks were detected for patient B, who usually only has two, then it is likely that patient B's response is fractionated. In this sense, the determination of the degree of fractionation present in a patient's atrial response should be conditioned on that patient's normal response. We can introduce this conditioning with an additional feature vector, $\phi_2(\mathbf{x})$ capturing the deviation from normal feature values:

$$\phi_2(\mathbf{x}) = \phi_1(\mathbf{x}) - \phi_1(\tilde{\mathbf{x}}) \tag{20}$$

where $\phi_1(\tilde{\mathbf{x}})$ is the initial feature vector of the segment capturing the patient's normal response, $\tilde{\mathbf{x}}$. An additional 'conditioned' feature was also added: the Dynamic Time Warping (DTW) distance between $\mathbf{x}$ and $\tilde{\mathbf{x}}$, $\phi_{DTW}(\mathbf{x}, \tilde{\mathbf{x}})$. The new feature vector is then

$$\phi(\mathbf{x}) = [\phi_1(\mathbf{x});\ \phi_2(\mathbf{x});\ \phi_{DTW}(\mathbf{x}, \tilde{\mathbf{x}})]. \tag{21}$$

For each patient, three normal responses (corresponding to electrodes CS1-2, CS3-4 and CS5-6) were identified. In most cases, the response to the first S1 pulses recorded by each electrode (i.e. at the largest S1/S2 interval) was taken to be typical. In some cases this response was actually deemed atypical, and typical responses were identified by eye[18].

Before proceeding, some of the less familiar features seen in Table 4 are described in more detail below.

5.4.1 Location and Width of Maximum Energy

The location and width of maximum energy are two hand-engineered features that are calculated by convolving the magnitude of each segment with a rectangular window of width M (a choice of $M = 14$ was found to produce convolved signals that had a clear maximum). This provides a much more robust estimate of the delay, and breadth, of the atrial response than the use of a peak detection algorithm, which is heavily dependent on implementation details and very sensitive to noise. Figure 11 illustrates the extraction of these two features. The *location of maximum energy* is defined as the position at which the convolved signal is maximum, and the *width of maximum energy* as the distance

S1 pulses do not electrically stress the heart.

[18]The typical response is the most frequently occurring response to S1 pulses. In most cases, all responses to S1 pulses were identical.

between the positions either side of this point at which the convolved signal drops below 20% of the maximum amplitude.

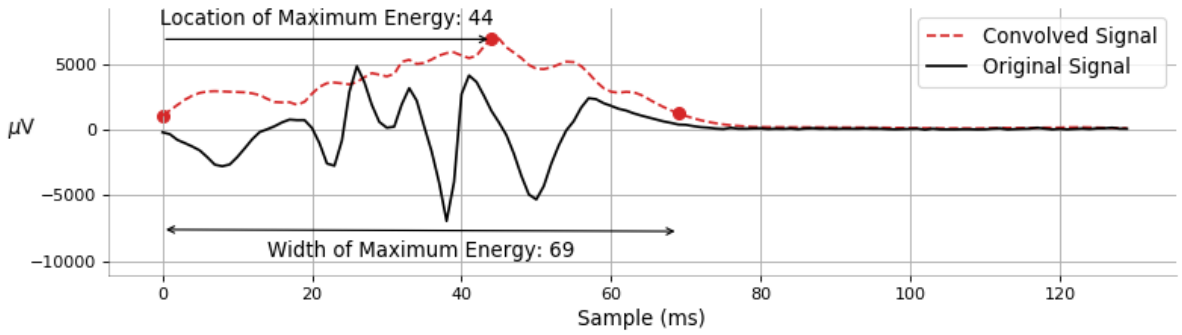


Figure 11: The location and width of maximum energy of a severely fractionated segment.

5.4.2 Sample Entropy

Sample entropy is an effective method of measuring the complexity of short segments of time-series data [35]. Given a time-series data $\mathbf{x} = [x_1, x_2, ..., x_N] \in \mathbb{R}^N$ and parameters m and r, the matrix $\mathbf{X}_m \in \mathbb{R}^{(N-m+1)\times m}$ is constructed, consisting of rows $\mathbf{x}_m(i)$ defined by

$$\mathbf{x}_m(i) = [x_i, x_{i+1}, \ldots, x_{i+m-1}]. \tag{22}$$

The Chebyshev distance between each row $\mathbf{x}_m(i)$ and $\mathbf{x}_m(j)$,

$$D_{\text{Chebyshev}}\left(\mathbf{x}_m(i), \mathbf{x}_m(j)\right) := \max_k \left(|\mathbf{x}_m(i)_k - \mathbf{x}_m(j)_k|\right) \tag{23}$$

where $i \neq j$, is then computed. The number of occurrences, A, of this distance being less than r is found. The matrix $\mathbf{X}_{m+1}$ is then constructed, and the number of times, B, the Chebyshev distance between each row $\mathbf{x}_{m+1}(i)$ and $\mathbf{x}_{m+1}(j)$, $i \neq j$, is less than r is also found. Sample entropy is computed as:

$$sample\ entropy = -\log \frac{A}{B}. \tag{24}$$

A choice of $m = 3$ and $r = 0.15x_{max}$, where x_{max} is the maximum absolute value of $\mathbf{x}$, was found to be effective in distinguishing between fractionated, and non-fractionated, responses.

Computing the sample entropy for a 125 sample long segment is computationally expensive. As discussed, it is of great importance that features can be efficiently extracted to avoid having any impact on the pace of the EP study. A cheap, yet effective, alternative is to compute the sample entropy for the 30 samples nearest the location of maximum

energy. Intuitively, we are only interested in the complexity of the atrial response, not the samples before or after its onset. It was found that 30 samples were enough to capture the entire response, whilst achieving a massive increase in computational efficiency.

5.4.3 Number of Peaks and Percentage Fractionation

The main concern in the design of a peak detection algorithm is providing robustness to noise. When denoising a signal prior to peak detection, care must be taken to avoid removing deflections due to fractionation. Since the frequency of these deflections vary across patients, the use of a common low-pass filter is unsuitable. Instead, wavelet denoising was performed. First, the segment is decomposed into its Daubauchy db6 wavelet coefficients. Soft thresholding is then applied using the universal threshold [36]. Coefficients with values less than T, defined by

$$T = \hat{\sigma}\sqrt{2\log N} \tag{25}$$

where $\hat{\sigma}$ is an estimate of the noise level σ and N is the length of the segment, are set to zero whilst others are decreased by T. The segment is then reconstructed using the modified wavelet coefficients - the result being a 'truer' version of the input segment.

For a time-series $\mathbf{x} = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^N$, peaks are identified at points x_i that satisfy the following conditions:

- x_i is a local maximum/minimum: $|x_i| > |x_{i-1}|$ and $|x_i| > |x_{i+1}|$.
- x_i has significant amplitude: $|x_i| > 0.1x_{max}$, where x_{max} is the maximum absolute value of $\mathbf{x}$.
- x_i is not at the boundary: $i \neq 0$ and $i \neq N$.
- The value at the midpoint, x_{mp}, between the previously detected peak, x_{prev}, and x_i is sufficiently different to either x_{prev} or x_i: $\max\{|x_{prev} - x_{mp}|, |x_i - x_{mp}|\} > 0.2x_{max}$.

The summation of all peak-to-peak intervals less than 10ms divided by the total segment length gives *percentage fractionation*. This is a slight modification to the metric introduced by Haley et al. [37], in which a threshold of 120ms is applied to the peak-to-peak intervals.

Figure 12 displays the location of peaks detected by the peak detection algorithm as applied to a fractionated segment, and the formation of *percentage fractionation*.

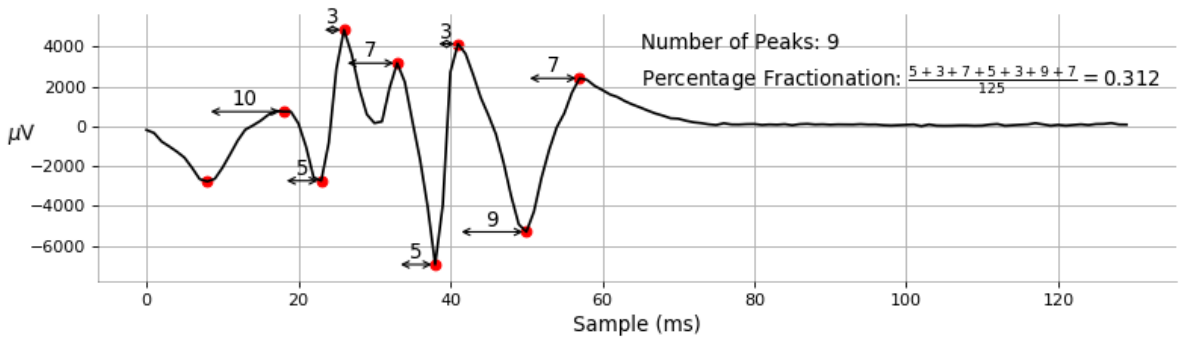


Figure 12: The peak detection algorithm as applied to a severely fractionated segments. Detected peaks are shown in red.

5.4.4 Dynamic Time Warping Distance

Although the use of DTW as a classification method was rejected, it remains an effective choice of feature for evaluating the 'distance' between a segment and a 'typical' segment. Given two time-series data $\mathbf{a} = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^n$ and $\mathbf{b} = [b_1, b_2, \ldots, b_m] \in \mathbb{R}^m$, an alignment, p, between $\mathbf{a}$ and $\mathbf{b}$ is a sequence of matched points,

$$p = \{p(1) = (i_1, j_1), \ldots, p(|p|) = (i_{|p|}, j_{|p|})\} \tag{26}$$

where (i_k, j_k) indicates that point a_{i_k} is matched with point b_{j_k}. The DTW algorithm searches for optimal alignment p^* that minimises the distance $D(\mathbf{a}, \mathbf{b}, p)$,

$$p^* = \underset{p}{\operatorname{argmin}}\, D(\mathbf{a}, \mathbf{b}, p) = \sum_{k=1}^{|p|} |a_{i_k} - b_{j_k}| \tag{27}$$

subject to the following constraints:

- p must satisfy the boundary conditions: $p(1) = (1, 1)$ and $p(|p|) = (n, m)$.
- p must move forward in time: if $p(k) = (a, b)$ and $p(k+1) = (c, d)$, then $c \geq a$ and $d \geq b$.
- p must include all of $\mathbf{a}$ and $\mathbf{b}$: $p(k) = (a, b)$ and $p(k+1) = (c, d)$, then $c - a \leq 1$ and $d - b \leq 1$.

The *DTW distance* between $\mathbf{a}$ and $\mathbf{b}$ is then $D(\mathbf{a}, \mathbf{b}, p^*)$ [38].

5.5 Train-Validation-Test Dataset Split

As stated previously: a dataset, consisting of nine AF patients and 37 non-AF patients, was available at the beginning of the project. A test dataset was later made available,

consisting of 10 AF patients and 30 non-AF patients. Since we seek to investigate the effect of model selection (i.e. model choice and feature subset), it is necessary to introduce an additional split to the initial dataset to create separate training and validation datasets. Model selection will be performed by evaluating the performance of each model on the validation data, after being trained on the training data. Once the best model is selected, it shall be trained on the training and validation datasets combined (i.e. the initial dataset), and its performance on the test dataset will be evaluated. Whilst this may sound convoluted, this arrangement prevents any knowledge of the test data 'leaking' into the model selection process resulting in artificially good performance.

The training/validation split is typically performed by assigning, say, 75% of data to the training dataset and the other 25% of data to the validation dataset. A naïve approach would be to apply this split directly to the segments. However, this does not take into account the similarity between each patient's atrial responses. Consider a patient with 27 segments. If 24 of these segments are assigned to the training dataset, and the other three to the validation dataset, then if these three segments resemble any of the 24 in the training dataset (which is extremely likely), then any model trained on the training dataset will not need to have generalised at all to correctly classify the three 'unseen' segments. The ability of the model to generalise to unseen data is equivalent to the ability to generalise to unseen patients. Thus, the dataset should be split on a patient-by-patient basis. This was achieved by randomly assigning 25% of AF, and 25% of non-AF, patients to the validation dataset, with the remainder forming the training dataset. Table 5 provides details of the composition of each dataset.

	Patient Group		Segment Label		
Dataset	AF	Non-AF	Green	Amber	Red
Training	7	28	810	147	28
Validation	2	9	266	43	13
Test	10	30	951	173	33

Table 5: Composition of the training, validation and test datasets.

5.6 Applying Data Augmentation

As seen in Table 5, there is a great imbalance between classes in all datasets. For example, in the training dataset, there are only 28 segments labelled as 'red' and 810 labelled as 'green'. We would not expect these 28 segments to offer a sufficiently diverse representation of fractionated responses to prevent the model from overfitting. It was decided to investigate the performance of the model when trained on non-augmented, and augmented datasets. It is important that each augmented segment falls within the same class

as the segment it originated from, as re-labelling each augmented segment is not practical for large quantities of augmented data. Segments labelled as 'green' are characterised by showing no signs of distortion - augmenting these segments may introduce distortion such that they become 'amber' or even 'red'. Similarly, applying augmentation to segments labelled as 'amber' may push them either side of the decision boundary (i.e. becoming 'green' or 'red'). Thus, augmentation was only applied to 'red' segments under the assumption that this form of distortion could not sufficiently de-fractionate a response such that it moves across the decision boundary.

For each 'red'-labelled segment present in the training dataset, four augmented segments were generated as described in Section 3.5, with $N = 50$. This was repeated four times for degree of augmentation $\sigma_A = 0.1$, 0.2, 0.3 and 0.4 - the result being four new datasets consisting of both the original training data, and augmented training data of differing degrees. Figure 13 compares the augmentation introduced by each value of σ_A on the same original fractionated segment. As expected, increasing σ_A corresponds to an increase in the severity of distortion.

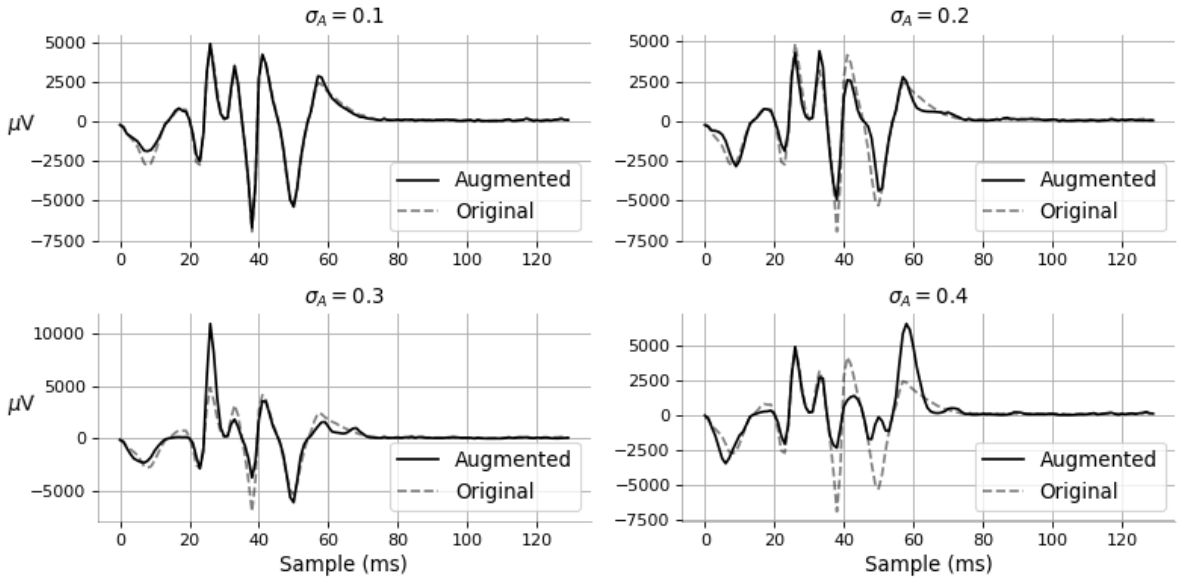


Figure 13: A comparison between augmented segments of degree $\sigma_A = 0.1$, 0.2, 0.3 and 0.4.

The composition of each augmented training dataset is detailed in Table 6.

	Patient Group		Segment Label		
Dataset	AF	Non-AF	Green	Amber	Red
Augmented	7	27	810	147	**140**

Table 6: Composition of each augmented dataset. The number of segments labelled as 'red' has increased from 28 in the non-augmented training dataset to 140.

5.7 Evaluation of Model Performance

In classification tasks, the three most widely used metrics to evaluate the performance of a model are *accuracy*, *precision* and *recall*. Consider the simple two class case, in which predictions made by the model can be summarised by the following confusion matrix:

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Accuracy is defined as the number of correct predictions, divided by the total number of predictions,

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}. \tag{28}$$

In cases in which classes are heavily unbalanced (i.e. negative labels are much more frequent), this metric is clearly unsuitable as a high accuracy is attainable by labelling all examples as negative.

Mathematically, we define precision and recall as

$$\text{precision} = \frac{TP}{TP + FP} \tag{29}$$

$$\text{recall} = \frac{TP}{TP + FN}. \tag{30}$$

If the model were to assign all examples as negative and thus achieve high accuracy, it would have a precision and recall of 0. However, achieving good precision and good recall are opposing forces. A recall of 1 is achieved by labelling all examples as positive - this results in a low precision if the number of negative examples is large. In contrast, a precision of 1 is achieved by labelling only almost certain positives as positive (i.e. $TP > 0$ and $FP = 0$) - this results in a low recall if the model is sufficiently confident in only a few positive examples (i.e. $FN \gg TP$).

To quantify the balance between recall and precision of our model, a modification to the F_1-score will be used, defined as

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1}. \tag{31}$$

The F_1-score is widely used by machine learning practitioners as a metric for the performance of classification models [39]. It can be adapted to three classes by computing the F_1-score for each class individually using a one-vs-all approach, whereby a single class is the positive class, and the other two classes are the negative. The modified F_1-score is an

average of the scores for each class:

$$\tilde{F}_1 = \frac{1}{3}(F_1^{green} + F_1^{amber} + F_1^{red}). \tag{32}$$

6 Results and Discussion

To get a sense of the separability of the classes in feature space, Figure 14 shows a plot of the first two principal components of the linear discriminant analysis (LDA) projection of feature vectors in the combined training and validation dataset. LDA is a dimensionality reduction technique similar to principal component ananlysis (PCA), except the principal components identified are those which maximise between class variability and minimise within class variability, as opposed to maximising the overall variability irrespective of class assignments [40]. It can be seen that a large cluster of green dots surround the origin,

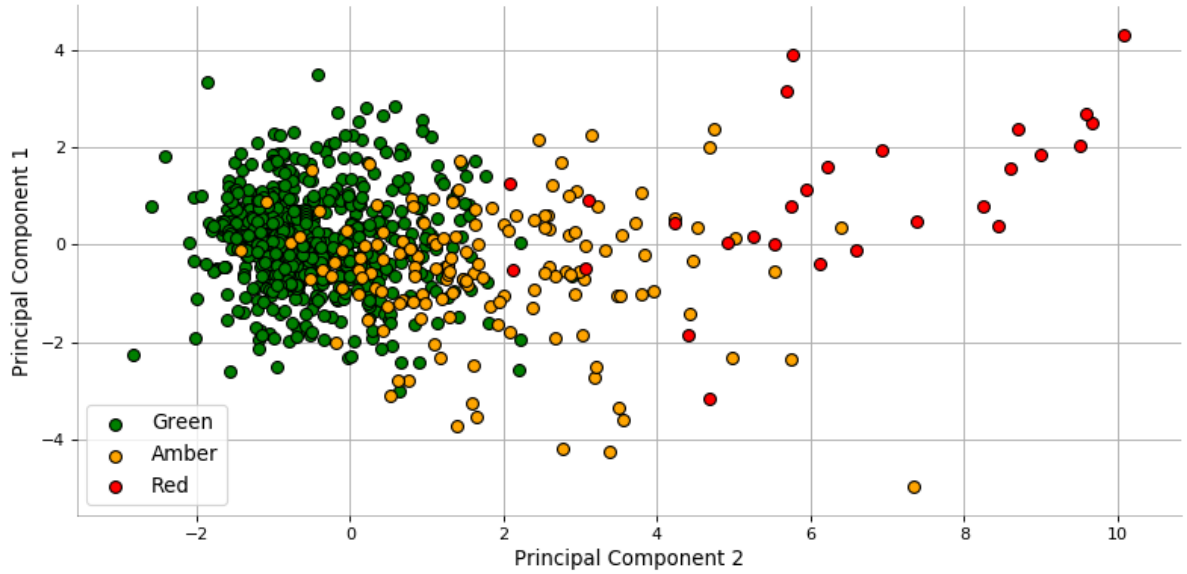


Figure 14: The first two principal components of the LDA transformation of the training dataset. The proportion of variance explained by 'Principle Component 1' and 'Principle Component 2' is 0.962 and 0.038, respectively.

with a band of amber dots separating this cluster from the collection of red dots situated some distance away from the origin. The result is promising, suggesting that the features are indeed sufficient to distinguish between responses that are heavily fractionated, and those that are not. However, it is clear that all three classes are not linearly separable. Specifically, the separation between green and amber is muddled with a significant proportion of amber dots existing within the cluster of green. We would expect that finding a decision boundary between these two classes that generalises to unseen data will be difficult.

'Raw' Feature	p_1	p_2	'Conditioned' Feature	p_1	p_2
Average Magnitude	**-0.409**	-0.230	*Average Magnitude*	0.130	**0.338**
Ratio Above σ	-0.258	0.257	*Ratio Above σ*	**0.505**	**-0.304**
Sample Entropy	-0.085	-0.039	*Sample Entropy*	**0.368**	**-0.336**
Number of Peaks	**0.421**	**-0.741**	*Number of Peaks*	**0.314**	-0.027
Location of Max Energy	-0.073	**-0.740**	*Location of Max Energy*	**0.324**	0.203
Width of Max Energy	**0.471**	**0.780**	*Width of Max Energy*	0.025	-0.016
Percentage Fractionation	-0.055	**0.495**	*Percentage Fractionation*	-0.217	**0.614**
			DTW Distance	**0.584**	**-0.343**

Table 7: The feature components for each axis shown in Figure 14. p_1 values form 'Principal Component 1' and p_2 values form 'Principal Component 2'. Values with magnitude exceeding 0.3 are shown in bold.

Table 7 provides a more detailed analysis of each of the principal components. Note that feature values were standardised prior to performing LDA, so the values shown in Table 7 can be compared directly. Interestingly, whilst raw *sample entropy* and conditioned *width of maximum energy* contribute very little to the separation seen in Figure 14, their conditioned and raw counterparts, respectively, are significant components of each axis. This supports the use of the conditioned feature vector in achieving greater class separation.

6.1 Logistic Regression versus Naïve Bayes Performance

Table 8 shows the confusion matrices, together with the $\tilde{F}_1$-score, for the logistic regression and naïve Bayes classifiers as trained on the training dataset with no augmented examples. The normalisation constant of the logistic regression classifier was set to $c = 1$. The $\tilde{F}_1$-score for the logistic regression and naïve Bayes classifiers were 0.884 and 0.860,

		Predicted		
		Green	Amber	Red
Actual	Green	**241**	20	5
	Amber	7	**30**	6
	Red	0	2	**11**

a) Logistic regression classifier.

		Predicted		
		Green	Amber	Red
Actual	Green	**248**	16	2
	Amber	17	**20**	6
	Red	0	4	**9**

b) Naïve Bayes classifier.

Table 8: A comparison between the predictions made by each model on the validation dataset.

respectively. The results indicate that the logistic regression classifier is able to identify a better decision boundary for both 'amber', and 'red', labelled feature vectors. Considering the differing assumptions made by each model, these results are not surprising. The key assumption in the naïve Bayes classifier is that feature values are mutually independent given the class $\mathcal{C}_k$. This is clearly violated using the features described in Section 5.4 -

as the degree of fractionation increases, each feature value can be expected to increase. Moreover, the Gaussian form for $p(x_i|\mathcal{C}_k)$ is unable to accurately model the distribution of some feature values. For example, Figure 15 shows a histogram of *sample entropy* feature values within the training dataset. The shape of the distribution has a heavy

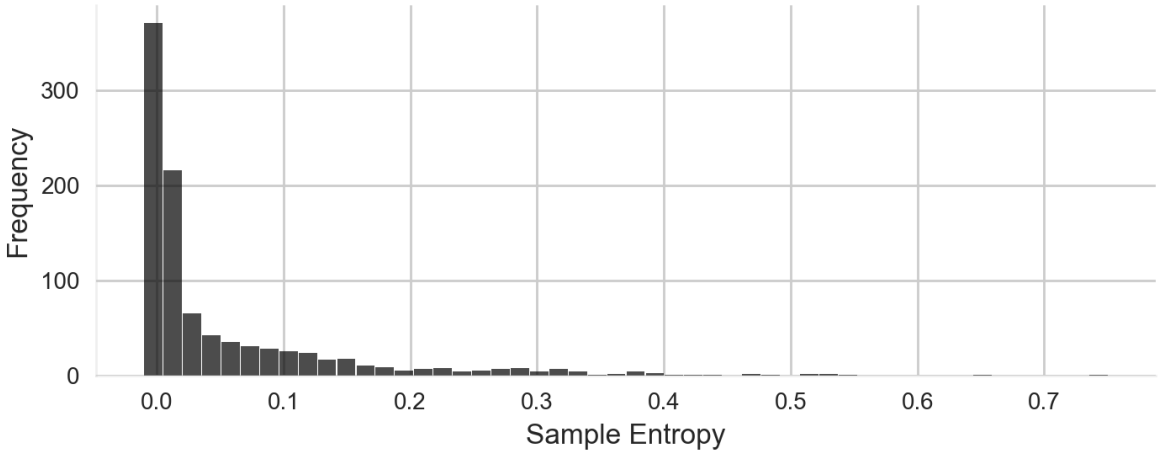


Figure 15: A histogram of *sample entropy* feature values.

positive skew - this property cannot be modelled by a Gaussian which is symmetric in nature. On the other hand, the only assumption made by the logistic regression classifier is that $p(\mathcal{C}_k|\phi(\mathbf{x}))$ is a transformation of a linear superposition of feature values. In other words, the decision boundary is linear in feature space. It is expected that the feature values monotonically increase with degree of fractionation. A linear decision boundary is suited to modelling such a relationship as we do not expect there to be regions of feature space in which class labels are isolated from others (in which case a more complex form, such as those formed using a support vector machine with a Gaussian kernel, would be required). Although the data does not appear to be linearly separable in Figure 14, we can see that reasonably good predictive performance could be achieved by dividing this two-dimensional projection using linear decision boundaries.

As expected, the difficulty experienced by each model was in its ability to distinguish 'amber' segments from 'green'. This is not too concerning - the labelling of a mildly fractionated response was often subjective when the differences were subtle (e.g. a slight increase in response duration). Moreover, whilst class predictions are helpful in accessing model performance, in practice we are more interested in the model's predictive probabilities $p(\mathcal{C}_k|\phi(\mathbf{x}))$. As we shall see, misclassifications made by the logistic regression model are on the 'right side' of misclassification. It is reassuring that neither model classified any responses labelled as heavily fractionated ('red') as having no fractionation ('green'). In a real-world scenario, failure to identify heavily fractionated responses could result in the patient entering AF and experiencing its potentially life threatening side-effects. However, we do note that the precision of heavily fractionated classifications is only 0.5

(i.e. there is a 50% probability that a response classified by the model as heavily fractionated is incorrectly done so). This is undesirable - sounding the alarm too frequently could halt the progression of an EP study unnecessarily and cause medical professionals to lose confidence in predictions made by the model.

6.2 The Effects of Data Augmentation and Normalisation

Figure 16 compares the effect that the normalisation constant c has on the $\tilde{F}_1$-score for logistic regression classifiers trained on the augmented datasets detailed in Section 5.6. For the model trained on the non-augmented dataset, we see rather noisy variations in

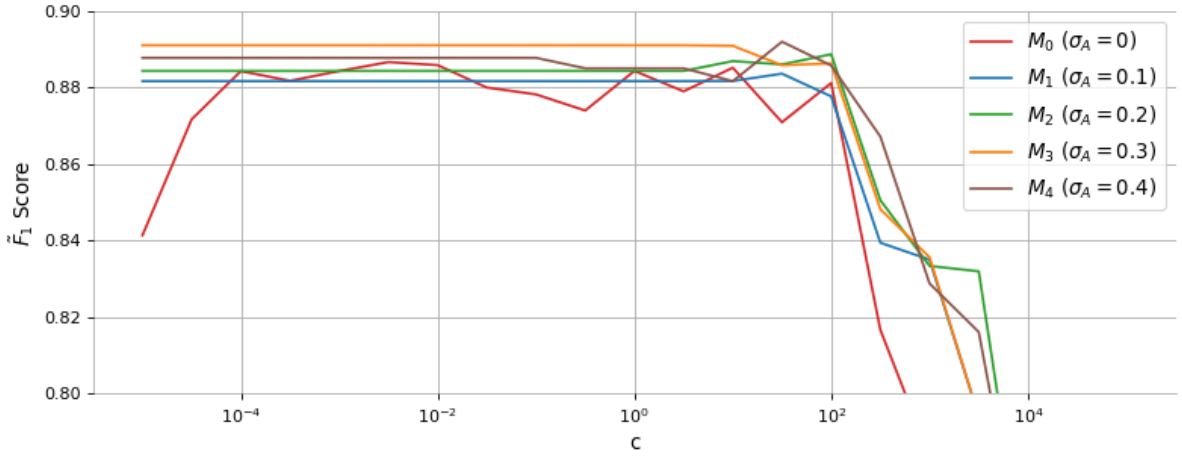


Figure 16: For each logistic regression classifier model trained on augmented datasets, a plot of $\tilde{F}_1$-score against c is shown. M_0 is the model trained on the non-augmented training dataset.

model performance with c, dropping off at very small values ($c < 10^{-4}$) and at large values ($c > 10^2$). For small values of c, this is a consequence of the model overfitting the training data. For large values, the strength of regularisation is too large for the model to fit the data at all - it simply labels all examples as 'green'.

Interestingly, in the interval $10^{-5} < c < 10^1$ the $\tilde{F}_1$-score remains constant for all models trained on augmented datasets. For small values of c, the model is prevented from overfitting to the training data by instead fitting to the augmented data. This implies that the large quantities of augmented data are indeed a more generalised representation of fractionation than the small quantity of real fractionated segments, validating the use of the augmentation technique. Figure 17 shows the location of augmented segments (with degree $\sigma_A = 0.3$) in feature space using the same LDA projection as in Figure 14. We see that the augmented feature vectors do not appear to be biased towards either end of the principal components. Instead, they extrapolate into the feature space near the actual heavily fractionated examples. σ_A controls the degree to which this extrapolation

is made. As σ_A increases, the augmented segments become increasingly dissimilar to their parent segment, moving deeper into the surrounding feature space.

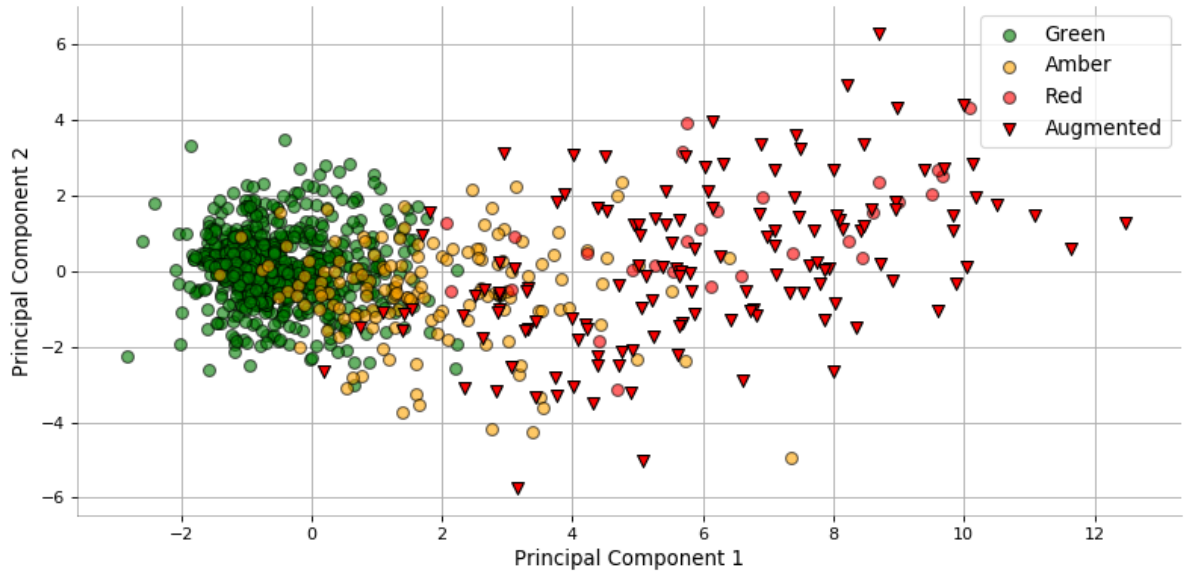


Figure 17: The same LDA transformation as shown in Figure 14 together with augmentated examples using $\sigma_A = 0.3$.

In Figure 16, it can be seen that model M_3 outperforms all other models, with a maximum $\tilde{F}_1$-score of 0.892 for all values of $c \leq 10$. The maximum $\tilde{F}_1$-score for model M_0 was 0.887 at $c = 10^{-2.5}$. For this value of c, the confusion matrices for models M_0 and M_3 are shown in Table 9.

		Predicted		
		Green	Amber	Red
Actual	Green	**238**	23	5
	Amber	4	**33**	6
	Red	0	2	**11**

a) Model M_0.

		Predicted		
		Green	Amber	Red
Actual	Green	**241**	24	1
	Amber	5	**33**	5
	Red	0	3	**10**

b) Model M_3.

Table 9: A comparison between the predictions made by each models M_0 and M_3 on the validation dataset with $c = 10^{-2.5}$.

It can be seen that the marginal improvement in performance of model M_3 is due to reducing the misclassification of 'green' segments as 'red' from five to one. However, model M_3 incorrectly labels one more 'red' segment as 'amber' than M_0. This presents a conflict of interests - whilst we wish to detect when a response is heavily fractionated with high accuracy, we don't wish to halt the EP study unnecessarily. The precision of model M_0 in identifying heavily fractionated responses is 0.5, which is low relative to the precision of M_3, 0.625. Moreover, given the 'noisy' plot of the performance of M_0 it is likely that this value of c is only optimal for this specific validation dataset. In other

words, there is a risk of overfitting c to the validation dataset. This is not the case for M_3, which shows consistent performance as c varies. Whether or not M_0 or M_3 is 'better' is largely subjective, and is dependent on the preferences of the user. For this report, however, it was decided that only the performance of M_3 on the test dataset would be evaluated.

The $\tilde{F}_1$-scores for the naïve Bayes classifier are 0.850, 0.844, 0.857 and 0.853 for models trained on augmented datasets of degree $\sigma_A = 0.1$, 0.2, 0.3 and 0.4, respectively. Again we see that the greatest performance is achieved using $\sigma_A = 0.3$, suggesting that $\sigma_A = 0.3$ generates responses that offer the best generalised representation of fractionated responses. However, for the naïve Bayes classifier all $\tilde{F}_1$-scores for models trained on augmented data are lower than for the model trained on non-augmented data. One possible explanation for this is that the naïve Bayes classifier is not sufficiently 'complex' to fit the data, let alone overfit to the training data. Thus, the benefits of data augmentation are not felt. We conclude that the logistic regression classifier outperforms the naïve Bayes classifier as applied to this problem.

6.3 Evaluating Feature Importance

As discussed, an analysis of the weights of the logistic regression classifier can be used to determine the relevance of each feature. To enable a fair comparison of the weights corresponding to each feature, each weight was divided by the standard deviation of the corresponding feature values in the training dataset. The resulting normalised weights can be compared directly. For our three class classification problem, there are three weight vectors, $\mathbf{w}^{green}$, $\mathbf{w}^{amber}$ and $\mathbf{w}^{red}$. The weight vectors of models trained on the non-augmented training dataset using $c = 0.01$, 1, and 10 are shown in Table 10.

As expected, for $c = 100$ the values of the weight vector are much smaller than for $c = 0.01$. Furthermore, the number of significant features in $\mathbf{w}^{green}$, $\mathbf{w}^{amber}$ and $\mathbf{w}^{red}$ decreases from 14, 9 and 14 to 6, 7 and 5, respectively. This is the sparsity inducing effect of the L1-regularisation term discussed in Section 3.2. For $c = 100$, it can be seen that the largest value of $\mathbf{w}^{red}$ corresponds to the conditioned *percentage fractionation*, implying that it is the most important feature for determining whether a segment is heavily fractionated. Interestingly, the second most important feature, the conditioned *width of maximum energy*, is also the second most significant in $\mathbf{w}^{green}$, the difference being the sign of the value. We can interpret this as follows: any increase in the *width of maximum energy* with respect to the patient's typical response is strongly indicative of fractionation, and any decrease indicates the response is not fractionated at all.

Another revealing feature of the weight vectors for $c = 100$ is the values corresponding

	$c = 0.01$			$c = 1$			$c = 100$		
Feature	Green	Amber	Red	Green	Amber	Red	Green	Amber	Red
Average Magnitude	0.251	-9.95e-3	-0.237	0	-5.74e-4	-3.18e-3	0	0	0
Average Magnitude 2	-0.199	-0.110	0.315	0	-4.77e-4	4.94e-4	0	0	0
Ratio Above σ	0.0529	0.0139	-0.0651	0	-3.82e-4	4.18e-4	0	0	0
Ratio Above σ 2	-0.324	0.0231	0.302	0	-1.35e-4	1.81e-4	0	0	0
Sample Entropy	0.352	0.687	-0.333	-0.210	0.547	-0.213	0	-8.62e-5	8.99e-5
Sample Entropy 2	-0.235	0.679	-0.443	-0.104	-1.60e-3	-0.307	0	-8.69e-5	8.99e-5
Number of Peaks	-0.312	0.146	0.166	-0.314	0.136	0.167	-9.04e-4	1.08e-4	8.09e-4
Number of Peaks 2	-0.382	-0.0916	0.474	-0.392	-0.0889	0.491	-7.91e-4	5.13e-4	1.29e-3
Location of Max Energy	2.52e-3	2.19e-3	-4.73e-3	2.54e-3	1.08e-3	-4.76e-3	-2.87e-4	4.96e-4	-4.81e-3
Location of Max Energy 2	-0.0109	-1.12e-3	0.0120	-9.73e-3	-1.17e-3	0.0115	-5.43e-3	7.94e-4	8.95e-3
Width of Max Energy	0.0200	0	-0.0201	0.0213	3.41e-3	-0.0247	7.89e-3	-2.22e-4	-5.60e-3
Width of Max Energy 2	-0.0352	4.97e-3	0.0302	-0.0406	-1.60e-3	0.0411	-0.0240	1.78e-3	0.0121
Percentage Fractionation	-0.0177	-5.10e-3	0.0229	-0.0176	-6.61e-3	0.0262	-8.29e-3	5.98e-4	1.81e-3
Percentage Fractionation 2	-0.0192	-0.0240	0.0432	-0.0214	-0.0250	0.0493	-2.77e-3	8.53e-4	0.0810
DTW Distance	-0.0942	0.0381	0.0560	-0.0957	0.0291	0.0618	-0.0527	0.0110	8.60e-4
Significant Features:	**14**	**9**	**14**	**9**	**6**	**9**	**6**	**7**	**5**

Table 10: The values for each of the three weight vectors, $\mathbf{w}^{green}$, $\mathbf{w}^{amber}$ and $\mathbf{w}^{red}$, for the logistic regression classifier trained on the non-augmented training dataset for c =0.01, 1 and 100. Within each weight vector, the values greater than 2.5% of the maximum value are highlighted and labelled as 'significant'. Feature names followed by a '2' indicate conditioned features.

to *DTW distance*. It was argued in Section 5.3 that classification based on a similarity metric, such as *DTW distance*, was inappropriate for this task as fractionated responses are inherently atypical. Indeed, it can be seen that the *DTW distance* of a segment is not a significant feature in $\mathbf{w}^{red}$. However, *DTW distance* is the most significant feature in $\mathbf{w}^{green}$, reflecting the fact that if the response is similar in shape to that typical of the patient then it is likely to show no signs of fractionation.

Finally, it is seen that the six largest weights in $\mathbf{w}^{red}$ when $c = 100$ correspond to three pairs of raw and conditioned feature values, *percentage fractionation*, *location of maximum energy* and *width of maximum energy*. We deem these three features to be the most important in identifying severe fractionation. This also supports the use of conditioned feature vectors - if they were redundant then it is expected that their weight values would be driven to zero.

6.4 Performance on Test Data

The results of the logistic regression classifier with $c = 1$, trained on the entire initial training dataset including augmented examples with degree $\sigma_A = 0.3$, is shown in Table 11. The corresponding $\tilde{F}_1$-score is 0.853. The $\tilde{F}_1$-score for the test dataset is significantly lower than that for the validation dataset, even after being trained on more data. However, there are no instances in which the model incorrectly identifies a severely fractionated segment as having no fractionation, and only two vice-versa. This is an extremely desirable property - as discussed, mistaking a severely fractionated segment as having no signs of fractionation puts patients at unnecessary risk of entering AF.

In order to understand why the model made the mistakes it did, it is instructive to identify

		Predicted		
		Green	Amber	Red
Actual	Green	**820**	129	2
	Amber	20	**117**	36
	Red	0	2	**31**

Table 11: Test predictions, as made by the logistic regression classifier trained upon the combined training and validation datasets, including augmented examples of degree $\sigma_A = 0.3$.

the locations of the misclassified segments in feature space. Figure 18 shows the LDA projection of test feature vectors onto the same principal components shown in Figure 14. The two segments misclassified as 'red' appear to lie in a region of feature space

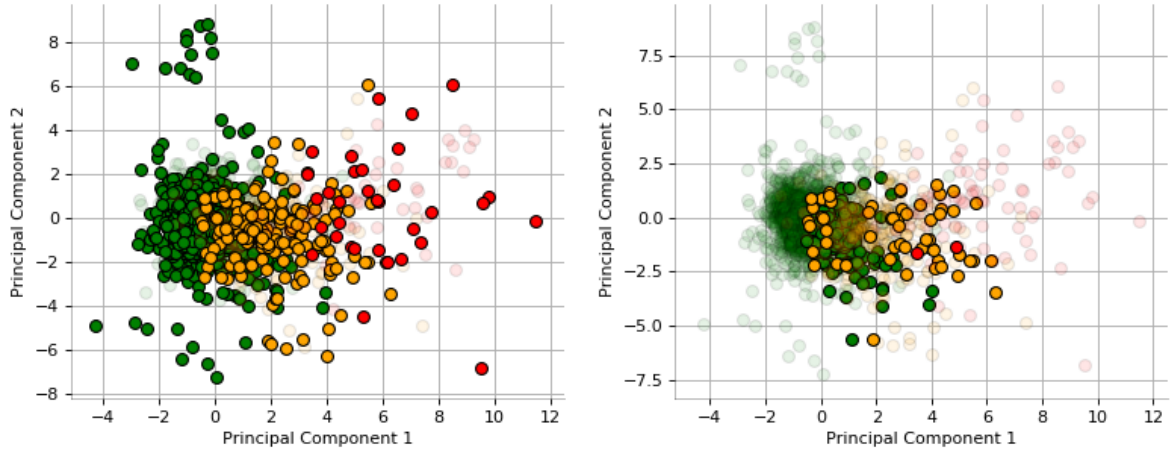


Figure 18: An LDA projection of test feature vectors onto two dimensions. The left plot shows the locations of all feature vectors and their corresponding true labels. The right plot shows only the locations of misclassified feature vectors, and their corresponding true labels.

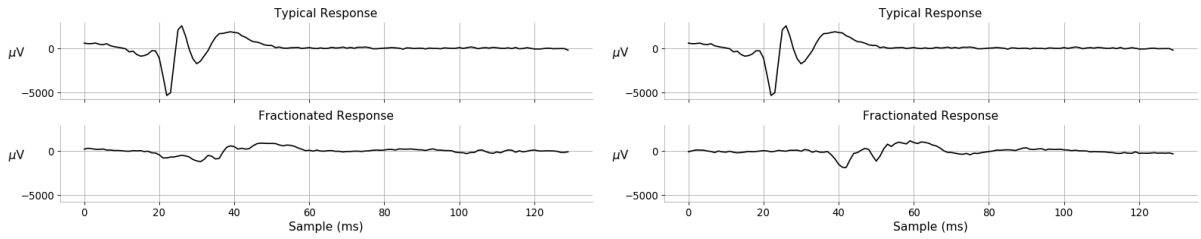
dominated by 'amber'. In fact, these two misclassifications are for responses recorded by the same electrode for the same patient, who experienced AF later on in the EP study. Figure 19 directly compares these segments with the 'typical' segments for that patient.

Given the absence of a quantitative measure of fractionation, the fractionation score, defined by

$$f_{score} = p(\mathcal{C}_{amber}|\phi(\mathbf{x})) + 2p(\mathcal{C}_{red}|\phi(\mathbf{x})) \tag{33}$$

can be used to get a sense of how fractionated the model believes these segments to be. Clearly $0 \leq f_{score} \leq 2$, with an f_{score} of 0, 1 or 2 implying the model believes the segment to be not at all, mildly or heavily fractionated, respectively. The fractionated segment shown in Figure 19a has an f_{score} of 1.490, whereas the segment shown in Figure 19b has an f_{score} of 1.163. Reassuringly, the model predicts both segments to be more than just mildly fractionated. It is also clear that segment in Figure 19b resembles the typical

response more so than the segment in Figure 19a, which is reflected in their corresponding f_{score}. Thus, although the model misclassified these heavily fractionated segments, there is some truth in the model's predictions.



(a) CS1-2: 290ms. **(b)** CS1-2: 280ms.

Figure 19: A comparison between the misclassified responses, and typical response, for patient AF14.

6.5 AF versus Non-AF Patients

For both groups of AF and non-AF patients, the progression of f_{score}[19] for each patient's atrial response, in each electrode, with S1/S2 interval is shown in Figure 20. The bold black lines show the median f_{score} in each group, and the individual red and green lines show individual patients' f_{score} in the AF and non-AF group, respectively.

These plots offer a great deal of insight. First, the difference in the progression of median f_{score} within each electrode between AF and non-AF patients is significant. We see a sharp rise in the median f_{score} for AF-patients at around 310ms in all electrodes, whereas for non-AF patients this rise is much more gradual. This difference is most striking in the f_{score} of electrode CS5-6 - all but four AF patients ends on an f_{score} greater than 1.5, whereas only seven of the 37 non-AF patients do so. Furthermore, the responses of AF patients in CS5-6 are seen to become fractionated much earlier than non-AF patients, at around 320ms compared to around 270ms for non-AF patients. The electrodes are identical in all but their position along the catheter, as shown in Figure 5, and therefore their location in the heart. A possible explanation for the more severe fractionation in the responses recorded by electrode CS5-6 could be due to local conduction delays in the cardiac tissue. When an S2 pulse is delivered prematurely, conduction in the nearby tissue can become delayed [9–13]. In such cases, since the conduction of the final S1 pulse is not as delayed as the conduction of the S2 pulse, the 'effective' S1/S2 interval increases further from the site of delivery. As described in Section 5.1.3, the CS5-6 electrode is located nearer the point of delivery of the pulse than electrodes CS3-4 and CS1-2. Thus,

[19]The f_{score} used in this Section comprises predictions made using the logistic regression model trained on both the original and test datasets. This is because we are interested in the mapping from input to f_{score}, not the predictive ability of the model.

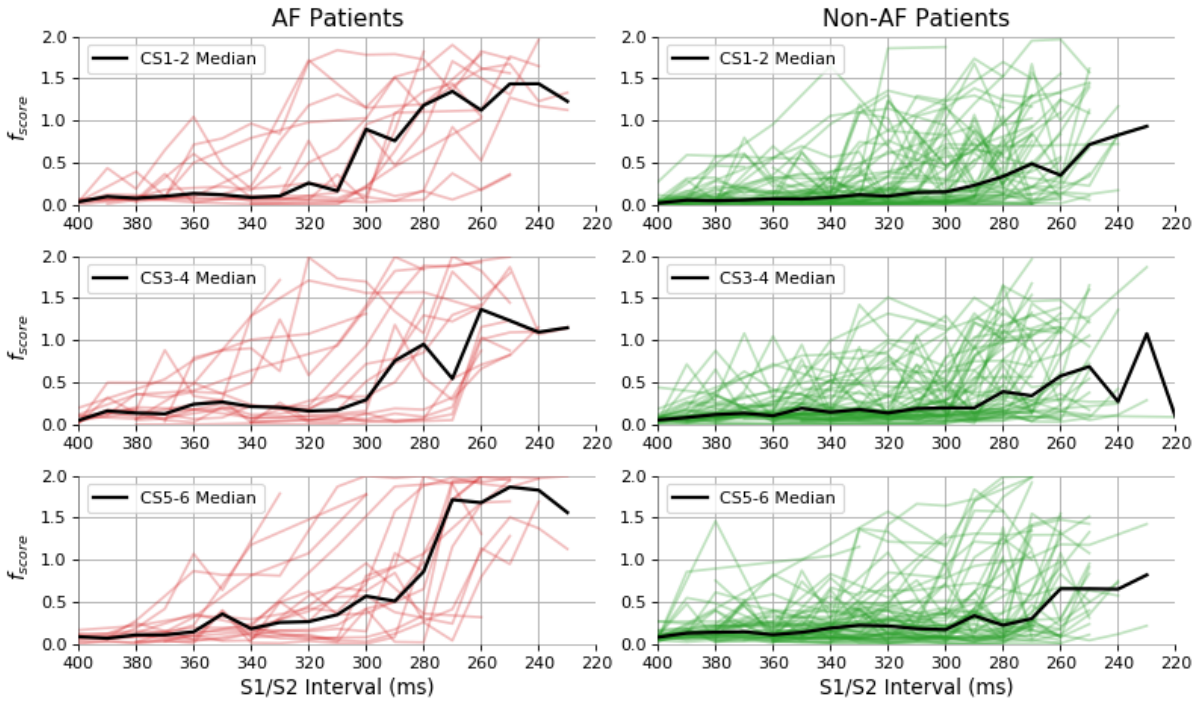


Figure 20: The progression of f_{score} for each patient, at each electrode, in the groups of AF and non-AF patients. The progression of the median f_{score} for all patients in each group is shown in black.

the 'effective' S1/S2 interval at electrode CS5-6 is shorter than that at electrodes CS3-4 and CS1-2, resulting in the largest degree of fractionation being observed in recordings by electrode CS5-6.

There is a large amount of variability between the degree of fractionation within each group, even within non-AF patients. This is expected - the cause of fractionation (e.g. diseased cardiac tissue) will vary between patients. Additionally, it is not surprising that many non-AF patients have a 'noisy' progression of f_{score}, especially at short S1/S2 intervals. All non-AF patients underwent the EP study because they experienced supra-ventricular tachycardias (SVTs), i.e. a rapid, regular heart rhythm arising from tissue above the ventricles which usually involves sections of atrial tissue. It is therefore likely that irregularities exist in their responses as recorded by each electrode, and that these irregularities will be more severe as the heart is put under increasing amounts of electrical stress when the S1/S2 interval is decreased.

Figures 21 and 22 show similar plots comparing the progression of the features *number of peaks* and *width of maximum energy*, respectively. As described in Section 2, most previous efforts to distinguish between AF and non-AF patients compared features equivalent to these. In both cases, the differences between AF and non-AF patients are much more subtle than those seen in Figure 20. In electrodes CS1-2 and CS3-4, the changes in both feature values as the S1/S2 interval is decreased are less significant than those in the

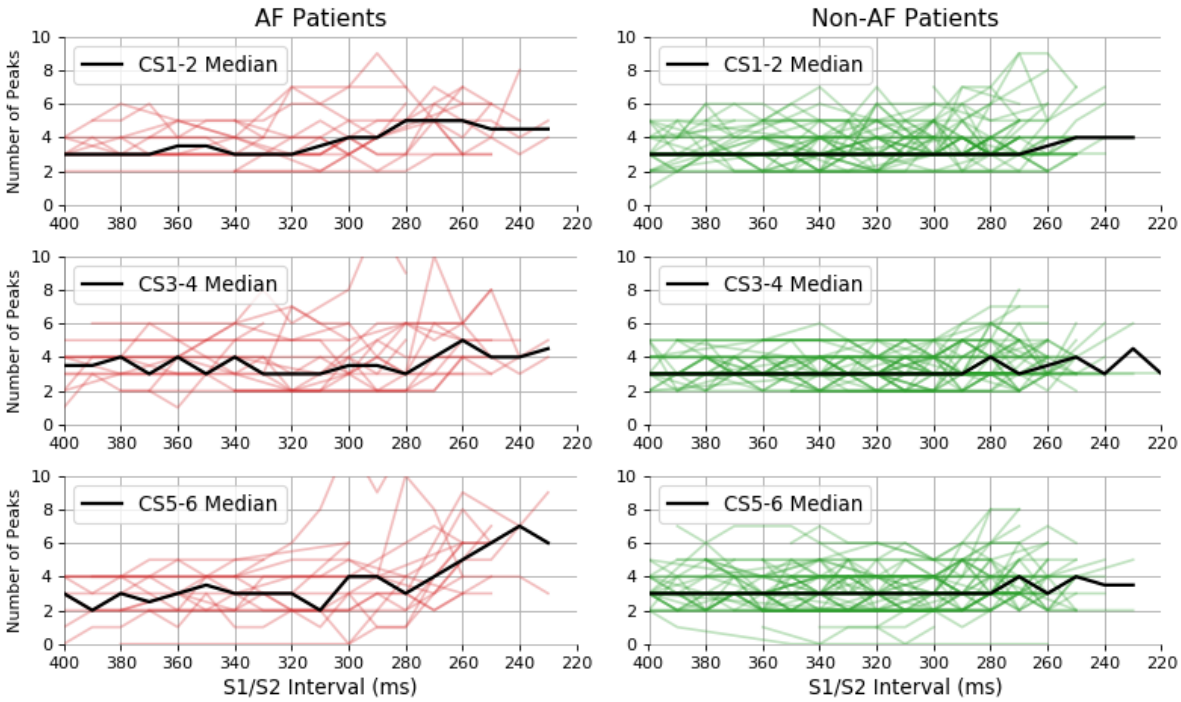


Figure 21: The progression of *number of peaks* for each patient, and electrode, in the groups of AF and non-AF patients. The progression of the median *number of peaks* for all patients in each group is shown in black.

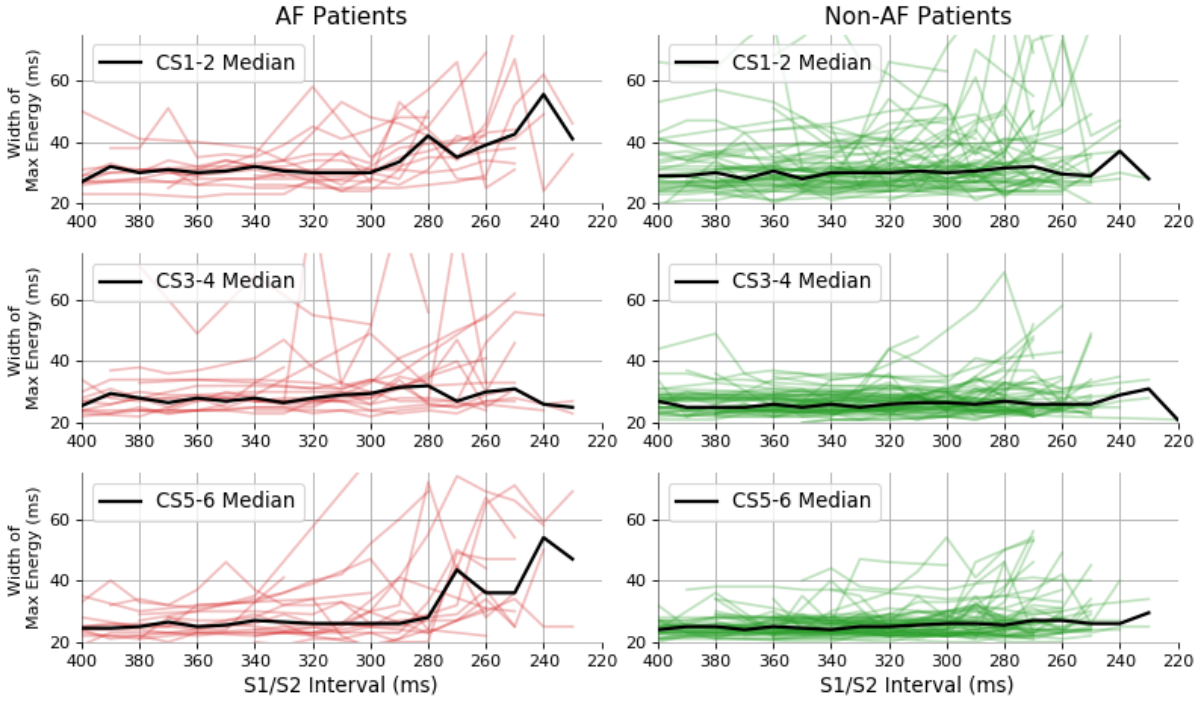


Figure 22: The progression of *width of maximum energy* for each patient, and electrode, in the groups of AF and non-AF patients. The progression of the median *width of maximum energy* for all patients in each group is shown in black.

f_{score}. It is only the responses in electrode CS5-6 for which the difference in progression of feature values between AF and non-AF patients is obvious. However, this difference is still less so than that seen in Figure 20. The use of a model to determine degree of fractionation is not only successful in achieving this, but is also shown to be more successful in distinguishing between the responses of AF and non-AF than the features used in previous studies [9–13].

The success of these results motivates the development of a model which explicitly predicts the risk a patient is at to entering atrial fibrillation. However, there are several reasons why this was decided against:

1. **Incomplete feature vectors** There is little consistency in the S1/S2 intervals recorded for each patient. For some patients, as few as three recordings were available, whereas for others there were as many as 18. This makes the formation of a fixed length feature vector representing the progression of the patient's atrial response an extremely challenging task.

2. **Unknown risk labels** The only known labels were whether or not patients entered AF during the EP study - this is not the same as a label identifying the risk the patient was at. Whilst those who entered AF were clearly at risk, it is probable that several non-AF patients were actually at high risk of entering AF. As a result, we cannot assign the label 'not at risk' to any patient.

3. **Lack of data** There were only 13 patients in the training dataset who entered AF. Since there are no patients who were known to be at no risk of induced AF, the total number of patients with labels is 13. This is clearly too small a dataset to construct a model that makes confident predictions.

Instead, the risk a patient is at can be inferred from a plot of the f_{score}, as shown in Figure 20, generated in real-time. This is possible due to the computational efficiency of the algorithm and feature extraction process. If the patient's response exceeds some threshold, say $f_{score} = 1$, for consecutive S1/S2 intervals, then the patient could be deemed at being at risk of entering AF. Alternatively, an assessment of risk could be made by comparing the patient's progression to the characteristic median AF curves seen in Figure 20.

6.6 Comparison with Expert Analysis

The questionnaire described in Section 5.2 was completed by three medical professionals. The results are summarised in Table 12 alongside the corresponding re-scaled f_{score}[20]

[20] i.e. $\frac{5 f_{score}}{2}$, such that it covers the range 0 to 5.

and *location of maximum energy* for each patient. The decisions made by the medical professionals as to whether to halt the EP study are compared to the decisions that would have been made by the model, using a threshold of 4.45 for the re-scaled f_{score}. Cells highlighted in green show when the decision to halt the EP study would have prevented induced AF, and cells highlighted in red show when the decision to halt the EP study was unnecessary.

		Questionnaire Results											
		Fractionation			Delay			Stop EP Study (S1/S2)				Model Analysis	
Type	Patient	Mean	Min	Max	Mean	Min	Max	Count	Mean	Min	Max	Max f_{score}	Max Delay
AF	1	1	1	1	1	1	1	0	-	-	-	4.98	60
AF	2	1	1	1	1	1	1	0	-	-	-	3.30	46
AF	3	3.33	3	4	1.33	1	2	3	287	280	300	4.97	71
AF	4	1	1	1	1	1	1	0	-	-	-	4.83	60
AF	5	1	1	1	1	1	1	0	-	-	-	4.91	93
AF	6	3.67	3	5	4.33	4	5	3	310	310	310	4.65	76
AF	7	1	1	1	1	1	1	0	-	-	-	1.65	102
AF	8	1.67	1	2	1	1	1	0	-	-	-	4.97	84
AF	9	2.33	2	3	2.33	1	3	3	320	320	320	4.97	68
AF	10	1	1	1	1	1	1	0	-	-	-	4.96	52
AF	11	1.67	1	2	1.33	1	2	0	-	-	-	4.55	87
AF	12	1.67	1	2	1	1	1	0	-	-	-	4.86	82
AF	13	1.67	1	3	1	1	1	0	-	-	-	4.46	43
AF	14	3	2	4	2	2	2	1	350	350	350	4.99	73
						Non-AF							
AT	1	1	1	1	1	1	1	0	-	-	-	4.83	132
AT	2	1.33	1	2	1	1	1	0	-	-	-	3.32	56
AT	3	1.33	1	2	1	1	1	0	-	-	-	4.88	87
AVNRT	1	1	1	1	1	1	1	0	-	-	-	3.32	69
AVNRT	2	3.33	3	4	2	2	2	3	257	240	290	2.51	60
AVNRT	3	1.67	1	2	1.33	1	2	0	-	-	-	3.06	32
AVNRT	4	1	1	1	1	1	1	0	-	-	-	1.02	38
AVNRT	7	4	4	4	2.33	2	3	3	300	300	300	1.80	40
AVNRT	8	4.33	4	5	3	3	3	3	260	260	260	0.51	42
AVNRT	9	1	1	1	1	1	1	0	-	-	-	1.13	38
AVNRT	10	1	1	1	1	1	1	0	-	-	-	3.75	46
AVNRT	11	1	1	1	1	1	1	0	-	-	-	4.28	43
AVNRT	12	3	2	4	3	3	3	3	313	310	320	2.53	36
AVNRT	13	4	4	4	3	3	3	3	343	340	350	3.36	46
AVNRT	14	1.67	1	2	1	1	1	0	-	-	-	2.69	39
AVNRT	15	3.33	3	4	1	1	1	3	303	280	330	4.96	71
AVNRT	16	2.67	1	4	2	1	3	2	310	300	320	4.49	67
AVNRT	17	1.33	1	2	1	1	1	0	-	-	-	4.06	40
AVNRT	18	1.67	1	2	2.67	2	4	3	333	320	360	3.39	57
AVNRT	19	5	5	5	3.67	3	5	3	340	330	360	4.98	82
AVNRT	20	3.67	3	4	1.67	1	2	3	293	290	300	2.20	33
AVNRT	21	3.67	3	4	2.33	2	3	3	280	280	280	4.20	55
AVNRT	22	1	1	1	1	1	1	0	-	-	-	1.75	38
AVNRT	23	1	1	1	1.67	1	2	0	-	-	-	4.95	63
AVRT	2	1	1	1	1	1	1	0	-	-	-	2.85	41
AVRT	3	1	1	1	1	1	1	0	-	-	-	3.32	56
AVRT	4	1	1	1	1	1	1	0	-	-	-	3.42	45
AVRT	7	4.67	4	5	4	3	5	3	310	300	330	1.46	27
AVRT	8	1.67	1	2	1.67	1	2	0	-	-	-	2.39	39
AVRT	9	1.33	1	2	1	1	1	0	-	-	-	3.39	45
AVRT	10	2.67	2	3	1.67	1	2	3	300	300	300	4.08	102
AVRT	11	1	1	1	1	1	1	0	-	-	-	4.95	48
AVRT	12	3.33	3	4	2.33	2	3	3	300	300	300	2.73	67
AVRT	13	3.33	3	4	1	1	1	3	280	280	280	3.03	53
EP	1	2.67	2	3	1	1	1	3	290	290	290	4.68	41
EP	2	1.67	1	2	3	3	3	3	283	280	290	4.07	39
EP	3	4.33	4	5	2.33	2	3	3	360	360	360	2.39	42
EP	4	3.67	3	4	1.67	1	2	3	280	270	300	3.57	43
EP	5	1.33	1	2	1	1	1	0	-	-	-	0.45	31
EP	6	4.33	4	5	3	3	3	3	313	300	340	4.04	52
EP	7	4.33	3	5	3.33	3	4	3	297	280	320	3.57	49
EP	8	1	1	1	1	1	1	0	-	-	-	2.07	49
EP	9	1	1	1	1	1	1	0	-	-	-	3.86	68
EP	10	2	1	3	1	1	1	2	280	280	280	1.71	44

Table 12: A comparison of the analyses of patient responses made by three medical professionals and the output of the model. The f_{score} was re-scaled such that it covered the range 0 to 5. The *location of maximum energy* is shown in the column labelled 'Max Delay'. The cells highlighted in green indicate decisions to stop the EP study that would have prevented induced AF (using a threshold of 4.45 for the re-scaled f_{score}). Those highlighted in red indicate decisions to stop the EP study unnecessarily. The questionnaire described in 5.2 was used to collect the data.

The results indicate that the model can more accurately evaluate the risk a patient is at to entering AF than medical professionals. Of the 26 cases in which a single medical professional declared that the EP study should have been stopped early, in only four instances did the patient enter AF, corresponding to a precision and recall of 0.15 and 0.286, respectively. By contrast, stopping the EP study if a re-scaled f_{score} greater than 4.45[21] was recorded would have prevented 12 out of 14 cases of induced AF and stopped only eight studies unnecessarily. This corresponds to a precision and recall of 0.6 and 0.857, respectively. The results overwhelmingly support the use of the model developed in this project in preventing induced AF, even in the case when a medical professional is at hand. Although the f_{score} decision boundary has been chosen to maximise the predictive performance on this dataset only, given more data it could be optimised across patients.

Interestingly, whilst the results indicate that the ability of medical professionals in predicting when to halt an EP study is poor, the assessments made are rarely divided - in only three cases was the decision to stop the EP study not unanimous. The same consistency is seen in the assessments of fractionation and conduction delay. Furthermore, Table 12 shows a strong disagreement between the severity of fractionation indicated by the medical professionals and by the model. For example, all three medical professionals indicate that no fractionation is seen in the responses of patient 'AF 1', whereas the model indicates the presence of severe fractionation. It is not simply the case that medical professionals systematically underestimate the degree to which a response is fractionated - for patient 'AVNRT 8', the unanimous opinion of the medical professionals is that responses become severely fractionated (an average score of 4.38), whereas the model indicates that very little fractionation is present. This suggests that the signs of fractionation, and thus warning signs of induced AF, taught to medical professionals are incorrect, resulting in systematic inaccuracies in predicting the risk a patient is at to entering AF.

7 Conclusion

7.1 Key Findings

The aim of the project was to address the need for a more rigorous evaluation of the risk a patient is at to entering atrial fibrillation (AF) during electrophysiology (EP) studies through the application of advanced signal processing and machine learning techniques. This was successfully achieved through the development of a supervised machine learning model with the capability to identify the degree of fractionation of a response with high accuracy, and thus predict the risk of entering AF.

[21] Considering the decisions to be predictions that the patient enters AF, a threshold of 4.45 maximises the familiar F_1-score.

Specifically, it was found that whilst both the logistic regression and naïve Bayes classifiers were able to distinguish between not at all and heavily fractionated responses, the performance of the logistic regression classifier was better. This can be attributed to the more appropriate assumptions made by the logistic regression classifier. The performance of the model was improved further through the implementation of sophisticated feature extraction and data augmentation techniques, overcoming the difficulties presented by patient variability and the limited availability of data, respectively. Features associated with conduction delay, response duration and proportion of the response populated by peaks were found to have the most significance in determining the degree of fractionation.

Using the probabilistic output of the logistic regression classifier as a quantitative measure of fractionation, it was found that as the electrical stress on the heart increased, patients who entered AF exhibited a greater degree of fractionation in their atrial response than patients who did not. This is consistent with previous studies, which show that the number of peaks and duration of a patient's atrial response, two features associated with fractionation, are also greater in patients who enter AF. However, the use of the model's output was found to be much more effective in differentiating between the two groups. Furthermore, it was shown that the model's output could be used to directly predict whether a patient entered AF to a much greater accuracy than predictions made by medical professionals.

Collectively, these results strongly support the use of the model in not only determining the degree of fractionation, but also in evaluating the risk a patient is at to entering AF during EP studies, even in the presence of a medical professional.

7.2 Limitations

In Section 6.5, a comparison was made between the progression of the degree of fractionation of patients who entered AF and patients who did not (Figure 20). Whilst the differences between the two groups of patients were significant, it was clear that the responses of some non-AF patients became extremely fractionated. This could be because such patients actually were at risk to entering AF, but it was not induced during the EP study. However, it is also true that all patients underwent the EP study because they experienced supra-ventricular tachycardias (SVTs). It would be expected that this population of patients is more likely to show fractionated responses than the general population. Thus, the difference between the development of fractionation in AF patients and non-AF patients taken from the general population is likely to be more significant than that observed in this study.

7.3 Suggestions for Future Work

A decision was made against the construction of a model to explicitly predict the risk of AF due to the lack of availability and incompleteness of the data. Moreover, the biases within the data (i.e. all patients having SVTs) prevented the responses of patients who entered AF being compared directly to the general population. Thus, future efforts could be made to collect a dataset without these limiting aspects and extend the analysis conducted in this report. In Section 5.4, features were normalised by subtracting values associated with a 'typical' response. This was done to improve the robustness of the model to patient variability. Although this was effective, an investigation into alternative methods of patient normalisation could lead to further improvement of the model's performance. It would also be interesting to investigate whether or not the success of the techniques used in this project can be repeated in the analysis of ventricular tachycardias (e.g. in predicting ventricular fibrillation, the chaotic contractions of the ventricles).

The consistent inability of medical professionals to predict when to halt an EP study suggests that the accepted warning signs of induced AF are incorrect. A review of how medical professionals make their decision, and how they differ to those made by the model developed in this project, could identify the mistakes being made and lead to subsequent changes to the currently accepted warning signs.

Acknowledgements

I would like to thank both Ian Turner and Dr. Andrew Grace for not only providing the data, but also for their patience in answering my many questions. I would also like to thank my supervisor, Dr. Elena Punskaya, for her support and guidance throughout the project.

References

[1] NHS. Arrhythmia, 2018.

[2] S. Stewart, N. Murphy, A. Walker, A. McGuire, and J. McMurray. Cost of an emerging epidemic: an economic analysis of atrial fibrillation in the UK. *Heart*, 90(3):286–292, 2004.

[3] R. L. Page, J. A. Joglar, M. A. Caldwell, H. Calkins, J. B. Conti, B. J. Deal, N. M. Estes, M. E. Field, Z. D. Goldberger, S. C. Hammill, and et al. 2015 ACC/AHA/HRS guideline for the management of adult patients with supraventricular tachycardia. *Circulation*, 133(14), 2016.

[4] M. A. B. Garcia, L. Macle, and P. Khairy. *Electrophysiology for Clinicians*. Cardiotext Publishing, 2012.

[5] I. Turner. Atrial pacing explained - notes. 2019.

[6] D. Bhakta and J. M. Miller. Principles of electroanatomic mapping. *Indian Pacing and Electrophysiology Journal*, 8(1), 2008.

[7] K. C. Roberts-Thomson, P. M. Kistler, P. Sanders, J. B. Morton, H. M. Haqqani, I. Stevenson, J. K. Vohra, P. B. Sparks, and J. M. Kalman. Fractionated atrial electrograms during sinus rhythm: Relationship to age, voltage, and conduction velocity. *Heart Rhythm*, 6(5), 2009.

[8] NHS. NHS standard contact for cardiology: Electrophysiology and ablation services (adult), 2013.

[9] Y. Qin, M. Kaibara, T. Hirata, O. Hano, Z. Liu, K. Tsukahara, T. Lshimatsu, C. Ueyama, M. Hayano, K. Yano, et al. Atrial conduction curves in patients with and without atrial fibrillation. *Japanese Circulation Journal*, 62(4), 1998.

[10] P. Platonov. Further evidence of localized posterior interatrial conduction delay in lone paroxysmal atrial fibrillation. *Europace*, 3(2), 2001.

[11] M. Tanigawa, M. Fukatani, A. Konoe, S. Isomoto, M. Kadena, and K. Hashiba. Prolonged and fractionated right atrial electrograms during sinus rhythm in patients with paroxysmal atrial fibrillation and sick sinus node syndrome. *Journal of the American College of Cardiology*, 17(2), 1991.

[12] C.-T. Tai, S.-A. Chen, J.-W. Tzeng, B. I. Kuo, Y.-A. Ding, M.-S. Chang, and L.-Y. Shyu. Prolonged fractionation of paced right atrial electrograms in patients with atrial flutter and fibrillation. *Journal of the American College of Cardiology*, 37(6), 2001.

[13] M. Pytkowski, A. Jankowska, A. Maciag, I. Kowalik, M. Sterlinski, H. Szwed, and R. C. Saumarez. Paroxysmal atrial fibrillation is associated with increased intra-atrial conduction delay. *Europace*, 10(12), 2008.

[14] J. Caldwell and D. Redfearn. Ablation of complex fractionated atrial electrograms in catheter ablation for AF; where have we been and where are we going? *Current Cardiology Reviews*, 8(4), 2012.

[15] T. P. Almeida, J. L. Salinet, G. S. Chu, G. A. Ng, and F. S. Schlindwein. Different definitions of complex fractionated atrial electrograms do not concur with the clinical perspective. In *Computing in Cardiology Conference 2013*. IEEE, 2013.

[16] D. Gupta, D. Redfearn, J. Hashemi, and S. Akl. A novel method for automated fractionation detection in ventricular tachycardia. *2016 Computing in Cardiology Conference (CinC)*, 2016.

[17] H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin. A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997*. IEEE, 1997.

[18] M. M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345, 2016.

[19] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati. Classification of ECG signals using machine learning techniques: A survey. In *2015 International Conference on Advances in Computer Engineering and Applications*. IEEE, 2015.

[20] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 1996.

[22] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004.

[23] J. Tang, S. Alelyani, and H. Liu. Feature selection for classification: A review. *Data classification: Algorithms and Applications*, 2014.

[24] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: when to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2016.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[26] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić. Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017.

[27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8), 1997.

[28] A. Pasini. Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease*, 7(5), 2015.

[29] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[30] E. J. Keogh and M. J. Pazzani. Derivative dynamic time warping. *Proceedings of the 2001 SIAM International Conference on Data Mining*, 2001.

[31] S. C. Bagley, H. White, and B. A. Golomb. Logistic regression in the medical lit-

erature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54(10), 2001.

[32] R. Bender and U. Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of Physicians of London*, 31(5), 1997.

[33] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[34] S. A. Pattekari and A. Parveen. Prediction system for heart disease using naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3), 2012.

[35] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6), 2000.

[36] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 1994.

[37] C. L. Haley, L. J. Gula, R. Miranda, K. A. Michael, A. M. Baranchuk, C. S. Simpson, H. Abdollah, A. J. West, S. G. Akl, and D. P. Redfearn. Validation of a novel algorithm for quantification of the percentage of signal fractionation in atrial fibrillation. *Europace*, 15(3), 2012.

[38] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1994.

[39] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 2009.

[40] S. Balakrishnama and A. Ganapathiraju. Linear discriminant analysis - a brief tutorial. *Institute for Signal and Information Processing*, 18, 1998.

A Appendix

All the code used in this report can be found at:
https://github.com/MattAshman/MasterProjectCode.

A.1 Risk Assessment

Prior to beginning work on the project, a risk assessment was completed. This detailed computer related hazards, such as repetitive strain injuries, and outlined precautions that would be taken to avoid them - all of which were implemented. No injuries were sustained over the duration of the project.