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: May 15, 2019

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

Signed: ______________________________ Date: _______________

Summary

Contents

1 Introduction

Over two million people in the UK are affected by an abormality of the hearts rhythm, more commonly known as cardiac arrhythmias, with symptoms varying from mild palpitations to the more extreme, such as strokes and even death[1]. The cost to the NHS is significant - around 1% of the total budget is spent on treating atrial fibrillation alone[2], a common type of arrhythmia. Cardiac Electrophysiology (EP) studies are a relatively recent advancement in the managing of arrhythmias, which have typically relied on clinical history and ECG recordings. Catheter electrodes are inserted into the heart to start and stop episodes of arrhythmia enabling an in-depth analysis of the causes of the arrhythmia, and in many cases the ablation[1] of the cardiac tissue behind the irregularity. These studies are now the recommended diagnostic and treatment strategy for most abnormal heart rhythms[3].

However, a drawback of using EP tests to electrically stress the heart is that in many cases it is too aggressive and can result in the onset of other abnormal heart rhythms, putting the patient at danger[4]. Current methodology for assessing the risk a patient is at is based upon crude visual interpretations of the recorded signal, whereby simple features such as conduction delay and number of peaks are monitored. 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 Atrial Fibrillation

Inserting catheter electrodes into the heart allows the changing potential field to be recorded, from which cardiologists can discover how electrical activity spreads through the cardiac chambers, and also localise the source of abnormality. 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. Typically, multiple catheters are inserted into specific locations from which recordings, collectively referred to as electrograms, are taken. Figure 1 provides a brief snapshot of part of an electrogram over a 750ms second interval.

When pacing is performed in the atria from the coronary sinus electrode, under normal circumstances the two signals highlighted in Figure 1 should have the same shape and conduction delay after the pacing pulse. As the duration before the final pulse 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 very abnormal as the electrical activity spreads in a chaotic

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

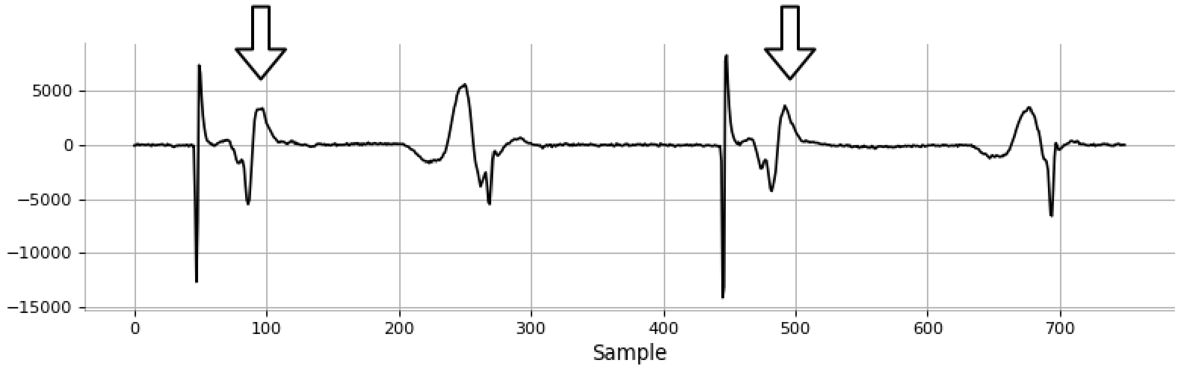
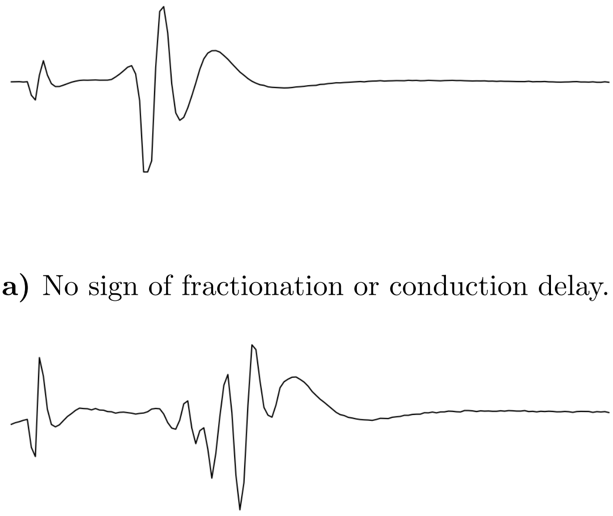


Figure 1: A short snippet of an electrogram recording, showing two atrial responses.

manner through the atria and coronary sinus. This is characterised by a significant change in the shape of the signals recorded from the coronary sinus, which become fractionated[5]. Figure 2 illustrates the changes from a) a normal response to b) a delayed and fractionated response.



(a) No sign of fractionation or conduction delay.

(b) Significant fractionation and some conduction delay.

Figure 2: A comparison between a) a normal response and b) a delayed and fractionated response.

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 and the atria go into atrial fibrillation (AF). 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]. 30-50% of patients are treated with anti-arrhythmic drugs, with the remainder of patients being sedated and given a large electric shock (3-5000 Volts) to re-set the rhythm. Understandably, these shocks are painful and carry their own risks.

1.2 The Need for a Solution

Consider again the electrogram shown in Figure 1. There is a great deal of information on display here. Now imagine 13 more signals, collectively moving across the screen in real-time, with the additional stress of monitoring the health of the patient. This is the environment in which cardiologists are expected to not only monitor the health of the patient, but also to make a diagnosis. Precautions are taken to avoid inducing AF in patients during EP studies is avoided by rarely pacing at less than 230-240ms intervals and by monitoring signs of fractionation, such as delayed, prolonged and fragmented signals from the coronary sinus electrodes. For experienced medical professionals, this is manageable. However, for less experience staff it is understandable that the signs indicative of fractionation may be missed. This begs the question - could these warning signs be detected automatically?

Existing papers[6–9] have shown a systematic link between an increase in conduction delay as the pacing interval is reduced and patients diagnosed with AF. These studies make no attempt to classify signals beyond simple conduction delay and signal width. Moreover, attempts to determine the degree of fractionation have been based upon either seemingly arbitrary definitions of fractionation[10] or are limited in their analysis[11]. No known attempts to automate the detection of atrial fractionation with the aim of preventing induced AF have been made.

1.3 Project Goals and Specification

The goals of this project are as follows:

1. Develop an novel tool to ‘automatically’ determine the degree of fractionation of an electrogram signal,
2. To evaluate the ability of this tool to predict the onset of induced AF during EP studies.

As discussed, signs of fractionation include delayed, prolonged and fragmented signals. These are all features of the signal that one would use collectively to determine the degree of fractionation. One can imagine that these are not the only features indicative of fractionation, just those that can be easily evaluation by visual inspection. It is expected that a good selection of these features provide enough information to determine the degree of fractionation, which motivates the following design choice and expansion of our first goal:

to build an automated tool based upon sophisticated feature extraction and supervised machine learning techniques.

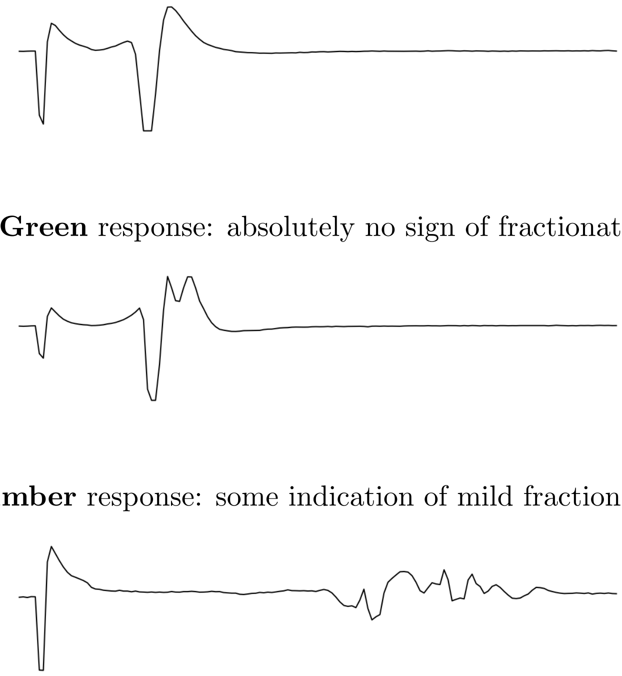
‘Supervised’ machine learning implies the existence of target labels, however no such standard measure of fractionation exists[10]. It was therefore decided to assign a signal into one of three categories:

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

Amber An amber label indicates the response has begun to shown some indication of fractionation, however retains most of it’s original shape. The pacing interval can be decreased, however may soon become more severely fractionated.

Red A red label indicates that the response has become significantly fractionated. At this point, the tissue activation could break down entirely if the pacing interval is decreased further and the patient enter AF.

Figure 3 provide examples of ‘green’, ‘amber’ and ‘red’ responses, respectively.



(a) Green response: absolutely no sign of fractionation.

(b) Amber response: some indication of mild fractionation.

(c) Red response: severe fractionation, little resemblance to a normal response.

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

This categorisation was validated by Dr. Andrew Grace, a leading cardiologist based at Papworth Hospital, who confirmed all labels individually.

The performance of the tool to determine the degree of fractionation of an electrogram signal will be measured by its ability to assign previously unseen signals into one of three three categories above with high accuracy. Whether or not the tool is of use in predicting the onset of induced AF will be determined by it's ability to correctly label patients at being at risk of AF or not.

2 Related Work

Both Qin et al.[7] and Plantanov et al.[9] identify a statistically significant relationship between conduction delay and patients with AF. Tai et al.[6] find a similar relationship between prolonged electrogram activity and atrial fibrillation. These relationships are confirmed by [8], in which conclusions hinge upon 'intra-atrial conduction curves' (i.e. 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. However, in both cases seemingly arbitrary definitions of fractionation, based upon the intervals between successive, are used [10]. An attempt to automate the detection of fractionation metrics (number of peaks, electrogram duration and conduction delay) in patients with ventricular tachycardias (VT) was made by Gupta, Hashemi, Akl and Redfearn[11]. The analysis of their results is limited, providing only a comparison between each raw feature values between groups of patients. They conclude that automated feature extraction is possible.

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, $\mathcal{D}$, comprising of input vectors $\mathbf{x}$ along with their corresponding target vectors $\mathbf{t}$. 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 it's target vector $\mathbf{t}^*$. As we shall see, this is done by finding the parameters which minimise a loss function $\mathcal{L}(\boldsymbol{\theta}, \mathcal{D})$,

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

Typically, $\mathcal{L}(\boldsymbol{\theta}, \mathcal{D})$ 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, e.g. predicting a persons gender given an image of their retina. The performance of a classification algorithm is determined by it's ability to correctly classify unseen data, which forms the test dataset. Typically, the training data comprises of only a (relatively speaking) 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 may divide classification algorithms into three distinct approaches [12]:

Descriminant Functions whereby a function $f(\mathbf{x})$ is learnt that directly assigns each input $\mathbf{x}$ into a specific class.

Descriminative 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 combined with prior class probabilities $p(\mathcal{C}_k$ to determine $p(\mathcal{C}_k|\mathbf{x})$ using Bayes' theorem.

There are pros and cons to each approach. Discriminant functions are the easiest to learn, however provides no measure of uncertainty in its 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 be had from modelling $p(\mathbf{x}|\mathcal{C}_k)$[12], generative models are the most computationally demanding to learn and are often unnecessary when we are only interested in $p(\mathcal{C}_k|\mathbf{x})$.

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, $\mathcal{D} = \{\mathbf{t}_i, \mathbf{x}_i\}_{i=1}^N$, 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

$\mathbf{W}_{kj} = \mathbf{w}_{kj}$, 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 right hand-side in Equation 3 gives us an expression for the negative log-likelihood, which we shall denote as $\mathcal{L}(\mathcal{D}, \mathbf{W})$:

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

To prevent overfitting of the model to the training data, a regularisation term is added to $\mathcal{L}(\mathcal{D}, \mathbf{W})$ 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. The form of $R(\mathbf{W})$ is a design choice, and in typically either L1-regularisation of weights, $R(\mathbf{W}) = \sum_{k=1}^K \sum_{j=1}^d |w_{kj}|$, or L2-regularisation of weights, $R(\mathbf{W}) = \sum_{k=1}^K \sum_{j=1}^d w_{kj}^2$. In most cases, the use of L2-regularisation is encouraged to achieve the best performance. However, L1-regularisation more strongly encourages sparsity amongst the weights, and can be used to identify the most relevant features.

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

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

Unfortunately there is no closed form solution for the above expression, however it is concave and therefore has a unique minimum. An iterative procedure can be employed, the most basic of which being 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 taken at each iteration. It simple to show that

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

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

Since activation functions a_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 classification algorithm to logistic regression in which decision boundaries are not linear is feature space is the naïve Bayes classifier, which shall be described in the following section.

3.3 Naïve Bayes Classifier

Similar to the logistic regression, the naïve Bayes classifier models the posterior probability of each class 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. By 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

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

where $p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathcal{C}_k)\prod_{i=1}^{d} p(x_i|\mathcal{C}_k)$ 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 $p(x_i|\mathcal{C}_k)$ is a design choice. When x_i is continuous, the most common choice for $p(x_i|\mathcal{C}_k)$ is the Gaussian distribution,

$$p(x_i|C_k) = \mathcal{N}(x_i; \mu_{ki}, \sigma_{ki}^2) = \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{-\frac{(x_i-\mu_{ki})^2}{2\sigma_{ki}^2}} \tag{12}$$

where μ_{ki} and σ_{ki} are the mean and standard deviation of the Gaussian distribution of feature x_i 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. When all $p(x_i|\mathcal{C}_k)$ are chosen to be Gaussian, the log-likelihood is given by

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

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

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

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

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

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

3.4 Feature Extraction and Selection

Consider the case in which we wish to classify 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 large sampling rate then the signal will comprise of a great deal 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 together 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. Since data processing can only reduce the separation between classes (by consideration of the data-processing inequality), it is of great importance to identify features that can be used to distinguish between classes - this is the process of feature selection. Typically, the feature selection problem is formulated as the following: given a set of candidate features, identify the subset of features that result in the best performance of the model on unseen data.

There exist three distinct categories of feature selection methods: filter methods, wrapper methods and ensemble methods. Filter methods select a subset of features based upon a statistical measure of the relationship between the feature and the output[2], such as

[2] In more sophisticated methods, the relationship between the feature and other features is also taken

correlation or mutual information. Wrapper methods search through subsets of features, training a machine learning model 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 are the most computationally expensive to perform[13]. Additionally, the chosen subset is the best for only the model used. Unlike wrapper and filter methods ensemble methods cannot be used to explicit select a subset of features. Instead, feature selection is *implicit*, and is incorporated as part of the model building process - for example, by including an L1-regularisation term in the loss function of the logistic regression classifier. As touched upon, this form of regularisation encourages sparsity amongst the weights such that those corresponding to features that have 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. For images, generation of new training data is typically done by cropping, or rotating, the original data. 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.[14], 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 $f(x)$, for $x \in \{0, m\}$. $f(x)$ 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 $y_i \sim \mathcal{N}(y_i; 1, \sigma^2)$. For $1 \leq i \leq N - 1$, the following constraints are imposed:

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

Figure 4 illustrates three randomly generated cubic splines using $\sigma = 0.1$, 0.3 and 0.5. It

into account.

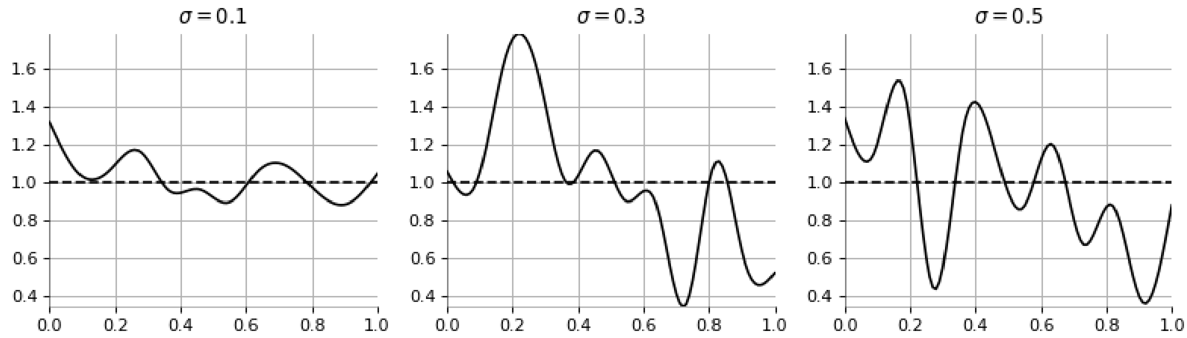


Figure 4: Three examples of randomly generated cubic splines for $\sigma = 0.1$, 0.3 and 0.5 with $N = 5$.

can be inferred that as σ increases, the magnitude of the distortion of the cubic spline also increases. 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 $y_i \sim \mathcal{N}(y_i; 1, \sigma^2)$, 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.

4 Clinical Methodology

The data used in the project was obtained from a retrospective study of patients undergoing EP studies at Papworth Hostiptal over the last 6 years. For each patient, the data collecting during a pacing technique known as the Antegrade Curve was available. Pacing is performed from electrodes positions at the rear of the catheter in the coronary sinus, electrodes CS9-10 or CS7-8, and consists of 8 evenly spaced pulses (S1) at an interval of 600ms, followed by an extra pulse (S2) at a shorter interval of 400ms, as illustrated in Figure 5. The process is repeated, with the interval before the extra pulse decreasing in 10ms intervals until either enough information to make a diagnosis is gathered, or an abnormal rhythm is induced.

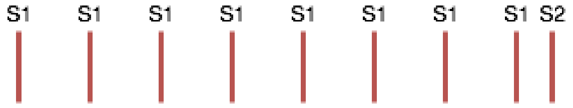


Figure 5: S1 pulses at a fixed (600ms) interval, followed by a premature S2 pulse.

Only patients who 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 as 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. The signals for 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 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. None of the patients developed AF in response to the standardised Antegrade Curve, but all had sustained (>30s) AF after pacing within 10 minutes after the end of the curve. 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 exceeded was was deemed acceptable. Furthermore, pacing was not performed at short intervals in many patients if it was evident that this would provide no additional diagnostic value. All patients used 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).

An training dataset, consisting of 9 AF patients and 37 non-AF patients, was available from the beginning of the project. A test dataset was later made available, consisting of 4 AF patients and 8 non-AF patients, upon which the performance of any model trained on the initial dataset would be assessed.

5 Methodology

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

Key Word	Meaning
Atrial response	Fluctuations in voltage as recorded at an electrode due to the electrical activity of the atria.
Segment	A fixed length snippet of time-series data encapsulating an individual patient's atrial response recorded at a single electrode.
Pulse	An externally induced electrical stimuli, typically delivered from an electrode towards the rear of a catheter inserted in the coronary sinus.
S1/S2 pulse	As illustrated in Figure 5.
S1/S2 interval	The interval between the final S1 and S2 pulse.

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 comprising of a patient I.D. and S1/S2 interval (e.g. 'AFPATIENT1-0230.txt', where 'AFPATIENT1' is the I.D. and '0230' indicates that the S1/S2 interval is 230ms) , and contents including header information detailing the properties of the electrodes (sampling rate, 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 the patient's atrial response to the preceeding two S1 pulses. This is illustrated in Figure 6. 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, we are only interested in data recorded from the coronary sinus channels (CS1-2, CS3-4, CS5-6, CS7-8 and CS9-10). This is because the remaining electrodes do not capture the patient's atrial response (rather they record the patient's ventricular response, or are ECG signals).

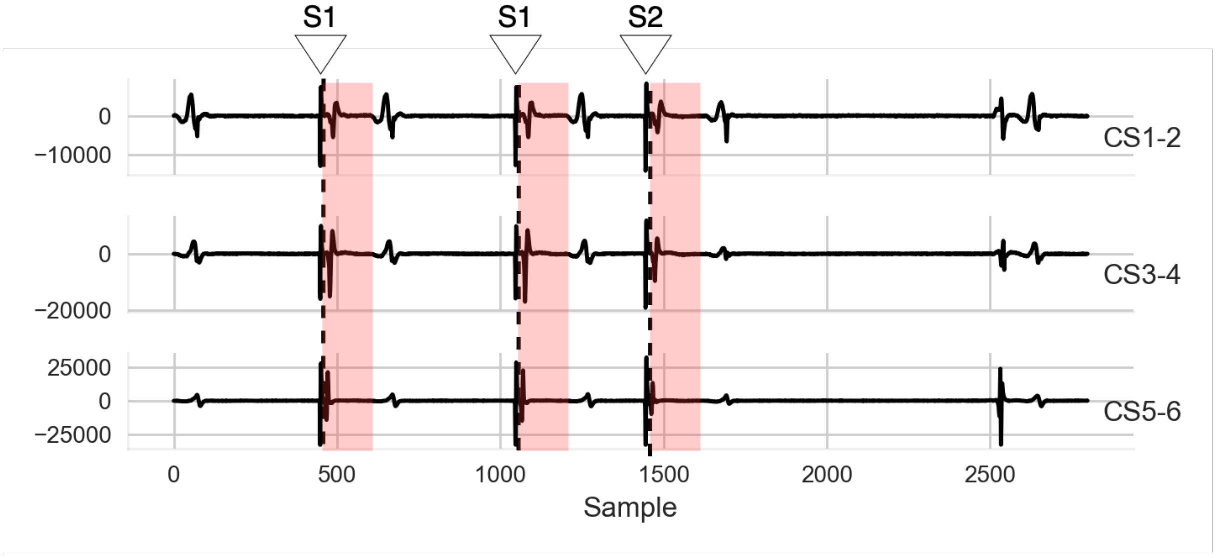


Figure 6: Identification of the S1/S2 pulses, and the corresponding atrial responses contained in an individual text file.

As pacing is performed from electrodes CS9-10 or CS7-8, only data from electrodes CS5-6, CS3-4 and CS1-2 will be extracted. It shall be assumed that all patients atrial response falls within 150ms proceeding 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

An initial investigation into the form of signal recorded from the electrode from which pacing was delivered (i.e. CS9-10 or CS7-8) revealed that consistent ‘clipping’ occurred at the instance of the pulse, as seen in Figure 7.

This consistency can be exploited to detect the location of the S1 and S2 pulses with very high accuracy. Positions at which the amplitude of the signal exceeded 95% of the maximum value were first identified. Moving forwards (in time) through these positions, removing those at a separation of less than 200ms proceeding 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 is then calculated. The interval closest to the S1/S2 interval specified in the file

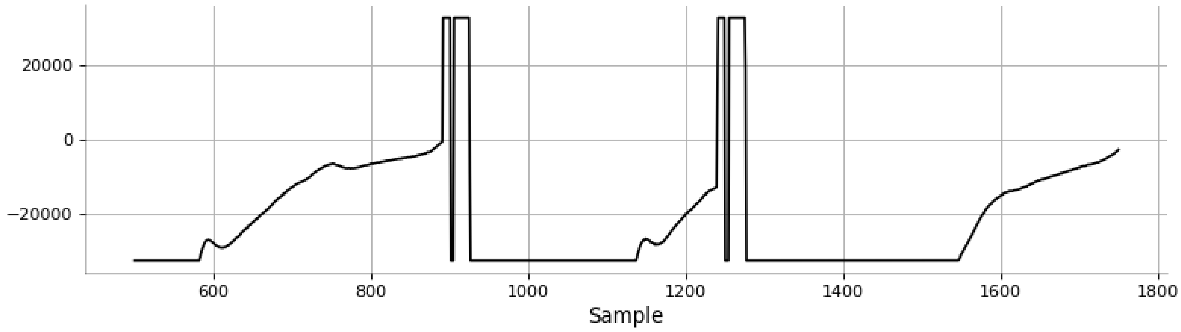


Figure 7: Clipping of the signal from the CS9-10 electrode at the point at which a pulse is delivered.

name is identified, and the latter position marked as the point of delivery of the S2 pulse. The remaining positions are marked as points of delivery of 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 electrodes are extracted according to the start and end positions, relative to the detected pulses, provided in Table 2.

	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 8 - the pulse reaches each electrode simultaneously at 5 samples in, however there is a much greater delay between the atrial response recorded by electrode CS1-2 than CS5-6.

As we shall see, conduction delay (i.e. the interval between the start of the segment and atrial response) will play a role in determining the degree of fractionation of the response. The values in Table 2 reflect an increase in delay for electrodes further from the site of

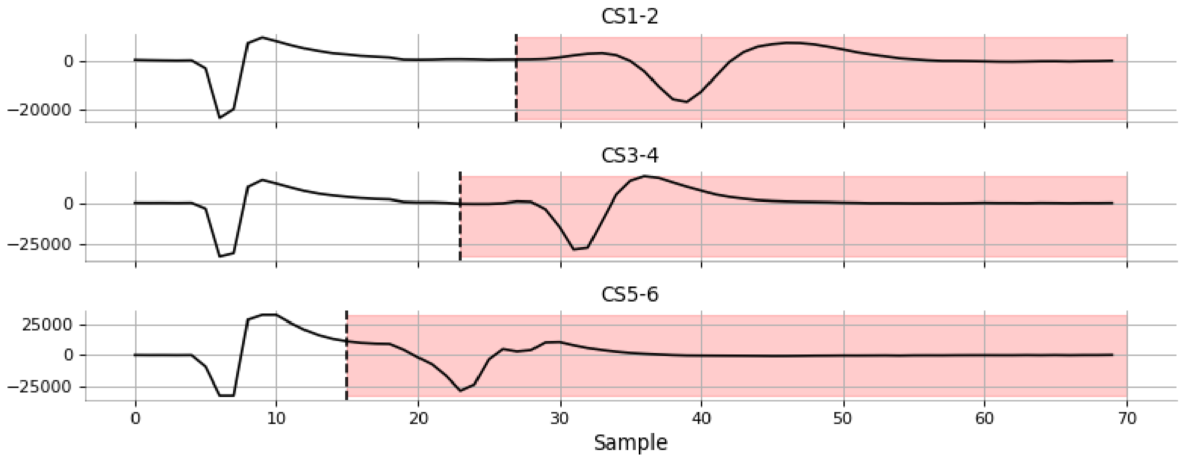


Figure 8: The conduction lag between electrodes CS1-2, CS3-4 and CS5-6. The part of the signal that is extracted from each electrode recording is highlighted in red.

the atrial response, and are included as to provide invariance of conduction delay to the electrode such that only deviations from what is typical of each channel is recorded.

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. From each text file the atrial response to two S1 pulses and a single S2 pulse is extracted from three electrode recordings (CS1-2, CS3-4 and CS5-6) - nine segments in total. Since this is done across all 19 text files, the output is 9 × 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

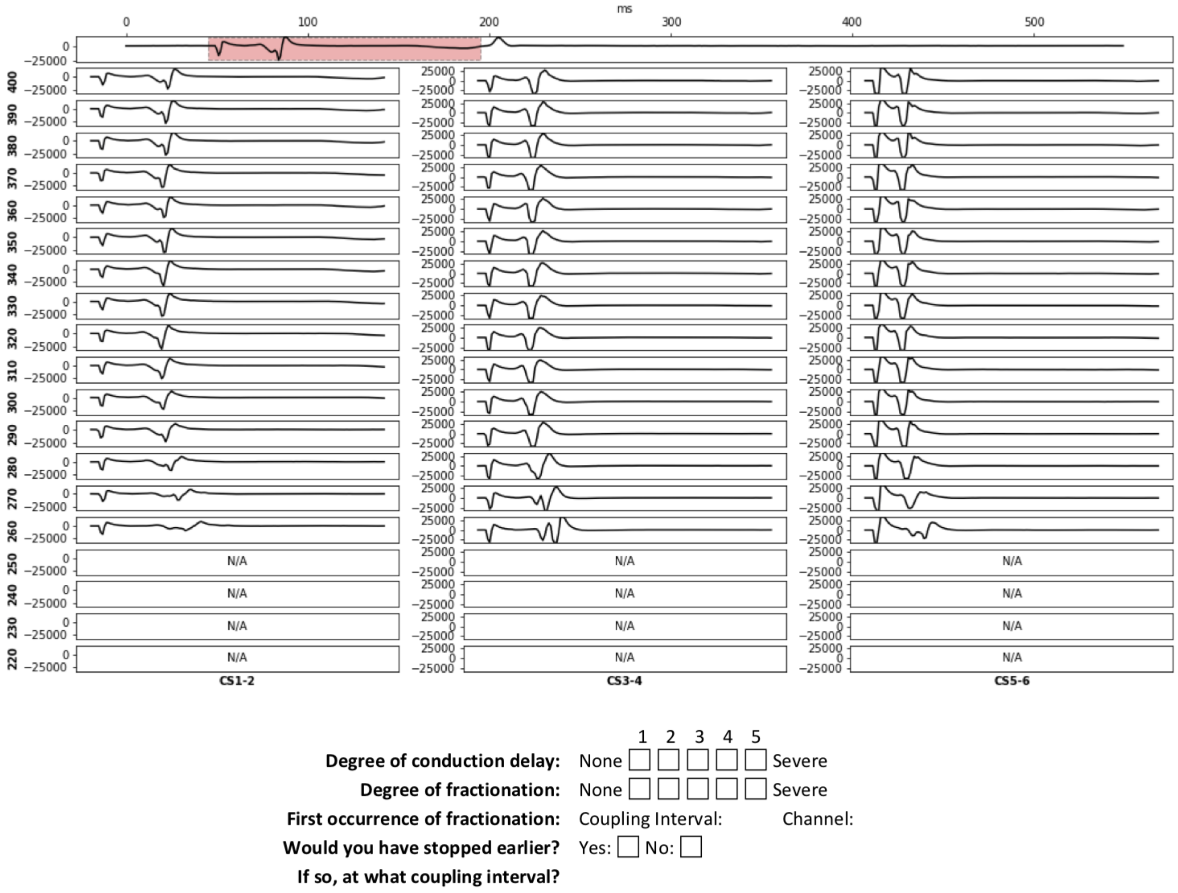
Table 3: An example of the composition of segments outputted by the data extraction process for a single patient.

5.2 Data Labelling

As outlined in section 1.3, individual segments were assigned to one of three categories (green,'amber'or 'red') with verification from Dr. Andrew Grace, a leading cardiologist at Papworth Hospital. It was also decided to collect additional information that captured an assessment of each patient's atrial activity during the EP study by a number medical professionals. For each patient, the medical professional was asked to evaluate the following:

1. To what degree does the response become delayed?
2. To what degree does the response become fractionated?
3. Where is the first sign of fractionation (coupling interval and channel)?
4. Should the EP study have been stopped earlier.

This information can be used to evaluate the performance, and use, of the model for 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 as to whether or not the EP study should be halted. These questions were presented in the form shown in Figure 9. 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 (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 may seem unnecessary, however it was included to aid cardiologists in developing an intuitive understanding for this form of presentation of the patient's response. Moving downwards from the top, the responses for incrementally



1 2 3 4 5

Degree of conduction delay: None □ □ □ □ □ Severe

Degree of fractionation: None □ □ □ □ □ Severe

First occurrence of fractionation: Coupling Interval: Channel:

Would you have stopped earlier? Yes: □ No: □

If so, at what coupling interval?

Figure 9: A single page of the questionnaire provided to medical professionals.

decreasing S1/S2 intervals are shown. Although most patients don't have responses for the entire range of intervals, it was important to keep the vertical (and horizontal) scale consistent across patients. Additionally, this enables the atrial response to a specific S1/S2 interval to be at the same vertical position on the page for all patients.

5.3 Choice of Machine Learning Model

Let's reiterate the first goal of the project. We wish to develop a model that takes an input segment capturing a patient's atrial response, and determine the degree of fractionation. The data available for the development of such a model consists 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 was provided in section 1 for the choice of employing a 'traditional'[3] supervised machine learning algorithm to achieve this. We will begin this section with further justification for this choice.

The choices for the development of the model are as follows: 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-Memory network[15]), or building a traditional machine learning model trained on extracted features. There are limitations associated with each of of these approaches. To begin, building a classification model based upon a measure of closeness hinges upon the availability of examples that are representative of a particular label. Although this technique has proven successful in many applications[16], we know 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.

Recurrent neural networks (RNNs) have the same limitation as many of its deep learning counterparts: the large quantity of training data required. The quantity of data available in this project is simply too small to expect a deep learning based model to be effective. Moreover, a 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. On the other hand, a model based upon extracted features has the benefit that the features can be engineered to be interpretable. For example, the value associated with the hypothetical feature number of peaks can assist the user of the model in understand why a certain prediction was made.

For these reasons, we limited our choice to supervised learning algorithms that permit multiclass classification. Moreover, we restrict our search to discriminative and generative

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

models, i.e. those that can model $p(\mathcal{C}_k|\mathbf{x})$. 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 a quantitative measure of an experts assessment of the degree to which a patients atrial response appears fractionated. It is anticipated that segments will lie inbetween 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 patients response than a discrete label.

With these considerations in mind, it was decided that to investigate and compare the performance of the logistic regression and naïve Bayes classification algorithms. Whilst these algorithms are relatively simple in comparison to some of their more advanced counterparts, they both meet each of the criteria of a 'desirable' machine learning algorithm outlined above. If the performance of one, or both, of the algorithms is high then we will have found the simplest, best solution. If not, then it will be necessary to explore more advanced alternatives.

For each model, there remains a number of design decisions that need to be made and/or investigated - these 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 the influences the values of the weights at the optimum. As we have discussed, whilst L2-regularisation often achieves the best results, 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 it's 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 influence the degree 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 for by training separate models with $c \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$.

5.3.2 naïve Bayes Classifier Implementation

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

$$p(x_i|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{-\frac{(x_i-\mu_{ki})^2}{2\sigma_{ki}^2}}.$$

As seen in section 3.3, the loss-function for this model has the unique closed-form solution

$$\begin{aligned}\mu_{kj}^* &= \frac{\sum_{i=1}^N t_{ik} x_{ij}}{\sum_{i=1}^N t_{ik}} \\ \sigma_{kj}^{2\;*} &= \frac{\sum_{i=1}^N t_{ik}(x_{ij} - \mu_{kj}^*)^2}{\sum_{i=1}^N t_{ik}}.\end{aligned}$$

5.4 Feature Extraction and Selection

As discussed in section 3.4, treating raw data time-series data as the input vectors to our model is inappropriate. 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:

Indicative of Fractionation This is more of a necessity than a desire. The model must be able to 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 the features can be chosen to be interpretable by the end user. This allows the user to more readily understand the model's output.

Computationally Efficient The tool is useless if it is unable to analyse signals, and evaluate fractionation in real time. We wish to do this whilst avoiding interfering with the medical procedure, hence cannot use features that require inordinate amounts of computational effort compute 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. These 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) were not distinguishable 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.
Hand-engineered	Number of Peaks	-
	Location of Max Energy	A measure of the conduction delay of the atrial activity.
	Width of Max Energy	A measure of the duration of atrial activity.
	Percentage Fractionation	A measure of the proportion of a signal which is fractionated.

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

For each segment, these eight features were extracted forming an initial feature vector, which we shall denote $\phi_1(\mathbf{x})$, 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 5 peaks in their atrial activity; however, if 5 peaks were detected for patient B, who usually only has 2, then alarms should be raised. In this sense, the determination of the degree of fractionation present in a patient's atrial response should be conditioned upon that patient's normal response. In terms of features, 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 DTW distance between $\mathbf{x}$ and $\tilde{\mathbf{x}}$, $\phi_{DTW}(\mathbf{x})$. The new feature vector is then

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

For each patient, three normal responses (corresponding to each electrode CS1-2, CS3-4 an CS5-6) were identified. In most cases, the response in each electrode to the first S1 pulse recorded (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.

Before proceeding, some of the more obscure features given in Table 4 are described in more detail below.

5.4.1 Location of Maximum Energy/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 effective). 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 both heavily on implementation details and very sensitive to noise. Figure 10 illustrates the extraction of the two features. We define the location of maximum energy as the position at which the convolved signal is maximum, and the width of maximum energy is the distance between the positions either side of this point at which the convolved signal drop below 20% of the maximum amplitude.

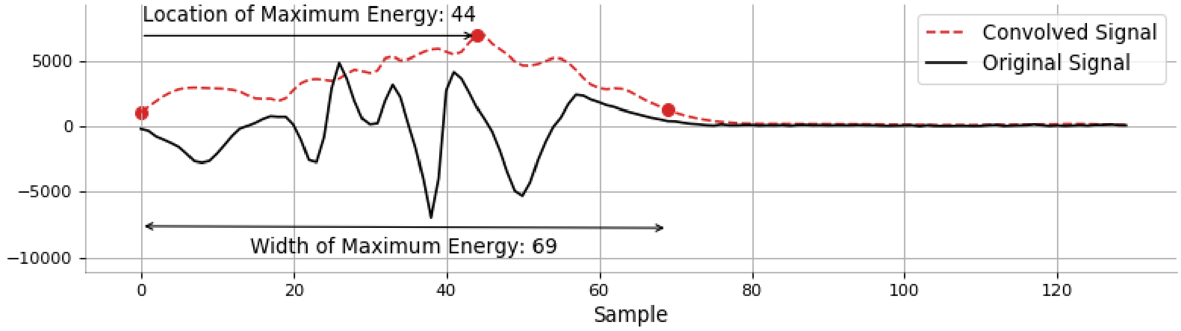


Figure 10: The location, and width, of maximum energy of a rather fractionated segment.

5.4.2 Sample Entropy Around Maximum

Sample entropy is an effective method of measuring the complexity of short segments of time series data [17]. Given N time series data points $\mathbf{x} = [x_1\ x_2\ \dots\ x_N]^T$ 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

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

The Chebyshev distance between each row $\mathbf{x}_m(i)$ and $\mathbf{x}_m(j)$, $i \neq j$, is then computed and 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:

$$SampEntr = -\log \frac{A}{B}. \tag{23}$$

Computing the sample entropy between two 150 sample long segments 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 was enough to between capturing the entire response, whilst achieving a massive increase in computational efficiency. A choice of $r = 3$ was found to be effective.

5.4.3 Number of Peaks/Percentage Fractionation

The main concern in the design of a peak detection algorithm is providing robustness to noise. Care must be taken to avoid removing fluctuations due to fractionation, which varies in frequency across patients implying 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 [18]. Coefficients with values less than $T = \hat{\sigma}\sqrt{2\log N}$, where $\hat{\sigma}$ is an estimate of the noise level σ, are set to zero whilst others are decreased by T. The segment is then re-constructed using the modified wavelet coefficients - the result being a 'truer' version of the input segment.

For a time series $\mathbf{x} = [x_1, x_2, \dots, x_m] \in \mathbb{R}^m$, 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 $i \neq m$.
- 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 10 samples divided by the total segment length gives percentage fractionation. Figure 11 displays location of peaks detected by the peak detection algorithm as applied to a fractionated segment, and the formation of percentage fractionation.

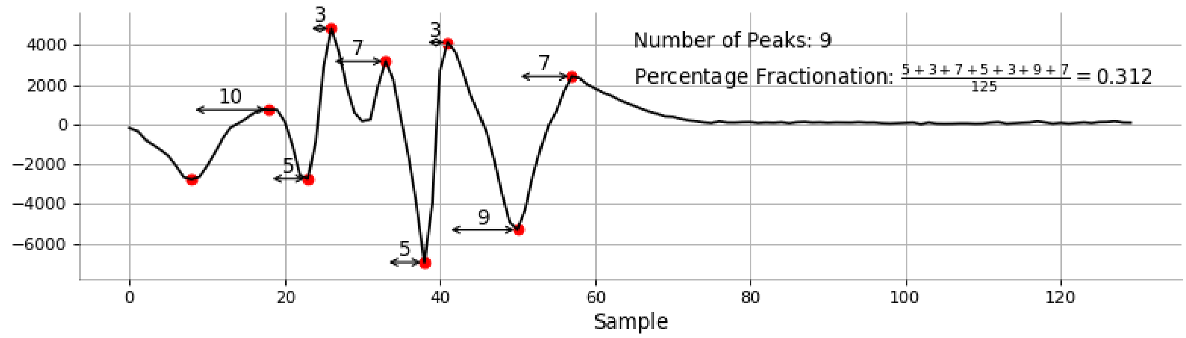


Figure 11: The peak detection algorithm as applied to a rather fractionated segments. Detected peaks are shown in red.

5.4.4 DTW Distance

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

$$p = \{p(1) = (i_1, j_1), \dots, p(|p|) = (i_{|p|}, j_{|p|})\} \tag{24}$$

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 $D(\mathbf{a}, \mathbf{b}, p)$,

$$p^* = \operatorname*{argmin}_p D(\mathbf{a}, \mathbf{b}, p) = \sum_{k=1}^{|p|} |a_{i_k} - b_{j_k}| \tag{25}$$

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 $D(\mathbf{a}, \mathbf{b}, p^*)$.

5.5 Train-Validation-Test Dataset Split

To reiterate: a dataset, consisting of 9 AF patients and 37 non-AF patients, was available from the beginning of the project. A test dataset was later made available, consisting of

4 AF patients and 8 non-AF patients. Since we seek to investigate the affect of model selection (i.e. model choice and feature subset), it is necessary to introduce an additional split to the initial dataset to create a validation dataset (and training dataset). Model selection shall 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 both the validation and training datasets (i.e. the initial dataset), and it's 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, 80% of data to the training dataset and the other 20% 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 S2 segments. If 24 of these segments are assigned to the training dataset, and the other 3 the validation dataset, then if these 3 segments resemble any of the 24 in the training dataset (which is extremelly probable), then a model trained on the training dataset will not need to have generalised at all to correctly classify the 3 '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 20% of AF, and 20% of non-AF, patients to the validation dataset, with the remaining forming the training dataset. Table 5 provides details of the composition of each dataset.

Dataset	AF	Non-AF	Green	Amber	Red
Training	7	27	810	147	28
Validation	2	10	266	43	13
Test	4	8	308	66	14

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 the training dataset, there are only 28 segments labelled as 'red' and 810 labelled as 'green' . We do not expect that these 28 segments offer a sufficiently diverse representation of fractionated responses to prevent a model from overfitting. It was decided to investigate the performance 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. Segments labelled as ‘green’ are characterised by showing no signs of distortion - augmenting these segments may introduce distortion such that they become ‘amber’ . 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, five augmented segments were generated as descried in section 3.5, with $N = 50$. This was repeated four times for $\sigma = 0.1$, 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 12 compares the augmentation introduced by each σ upon the same original fractionated segment. As expected, increasing σ corresponds to an increase in the severity of distortion.

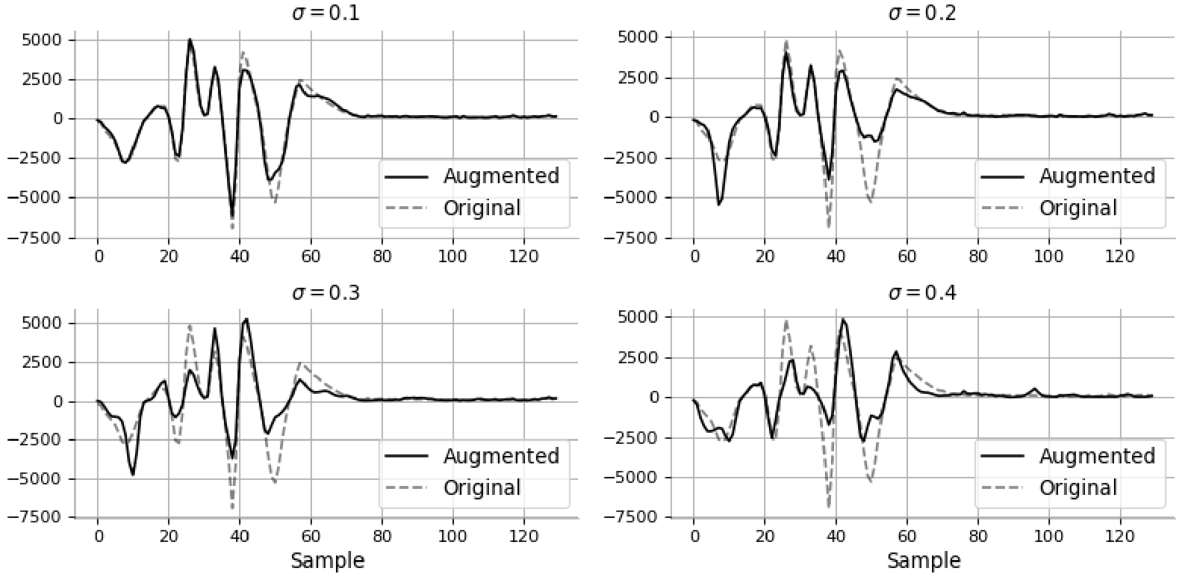


Figure 12: A comparison between augmented segments of degree $\sigma = 0.1$, 0.2, 0.3 and 0.4.

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

Dataset	AF	Non-AF	Green	Amber	Red
Augmented	7	27	810	147	**140**

Table 6: Composition of each augmented dataset.

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 the predictions made by the model can be summarised by the following confusion matrix:

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

Accuracy is defined as the number of correct predictions $(TP + TN)$, divided by the number of predictions $(TP + FP + FN + TN)$. 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

$$\text{precision} = \frac{TP}{TP + FP} \tag{26}$$

$$\text{recall} = \frac{TP}{TP + FN}. \tag{27}$$

If the model were to assign all examples as negative as to achieve a 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 very 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. $FN >> TP$).

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

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{ precision }^{-1}}{2} \right)^{-1}. \tag{28}$$

The F_1-score is widely used by machine learning practitioners as a metric for the performance of models on unbalanced datasets. 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 a sample average of the scores for each class:

$$\tilde{F}_1 = \frac{1}{3}(F_1^{green} + F_1^{amber} + F_1^{red}). \tag{29}$$

6 Results and Discussion

To get a sense of the separability of the classes in feature space, Figure 13 shows of a plot of the first two principle 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 principal components which maximise the between class variability and minimise the within class variability are identified[19]. It can be seen that a large cluster of green dots surround the origin, with a band of amber dots separating this cluster from the collection of red dots situated far 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, whilst this is only a two dimensional projection (whereas the true feature space exists in 15 dimensions) 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 can expect that finding a decision boundary between these two classes that generalised to unseen data will be difficult.

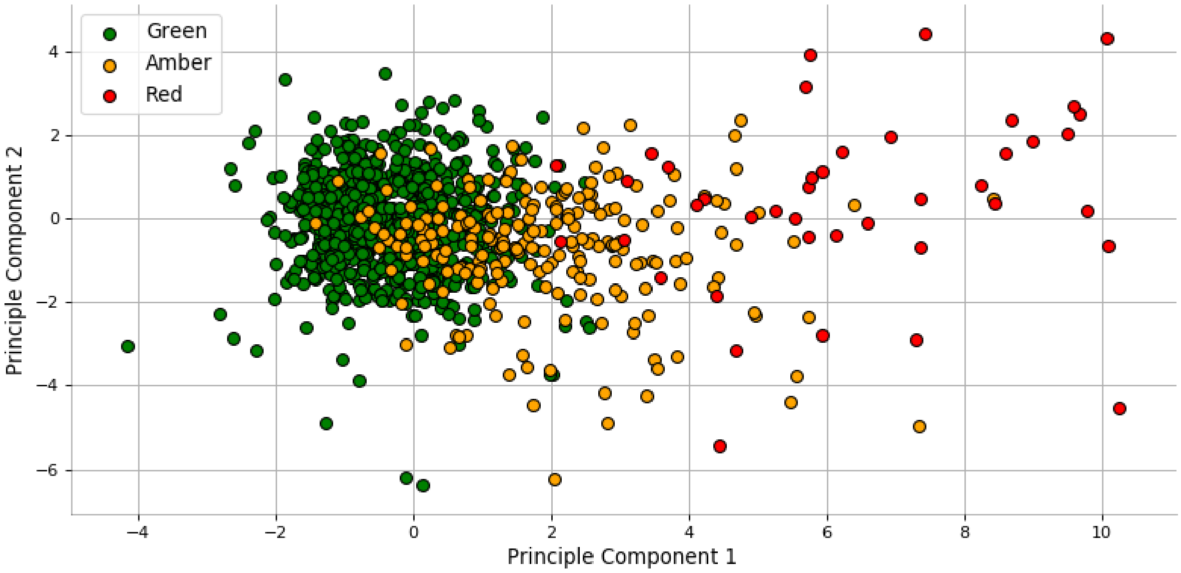


Figure 13: The first two principal components of the LDA transformation of the training dataset.

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 13 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.

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 13. 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.

6.1 Logistic Regression versus Naïve Bayes Performance

Table 8 shows the confusion matrices, together with the modified $\tilde{F}_1$-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 combined $\tilde{F}_1$-score for the logistic regression and naïve Bayes classifiers was 0.884 and 0.860, respectively. The results indicate that the logistic regression classifier can identify a better decision boundary for both 'amber', and 'red', labelled feature vectors. Consideration the differing assumptions made by each model, these results are expected. 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 of these feature values can be expected to increase. Moreover, the Gaussian form for $p(x_i|\mathcal{C}_k)$ is not able to accurately model the distribution of some feature values. For example, Figure 14 shows a histogram of sample entropy around maximum energy feature values within the training dataset. The shape of the distribution is heavily skewed, with a large tail towards one end - these properties 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. Again, it is expected that the feature

		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.

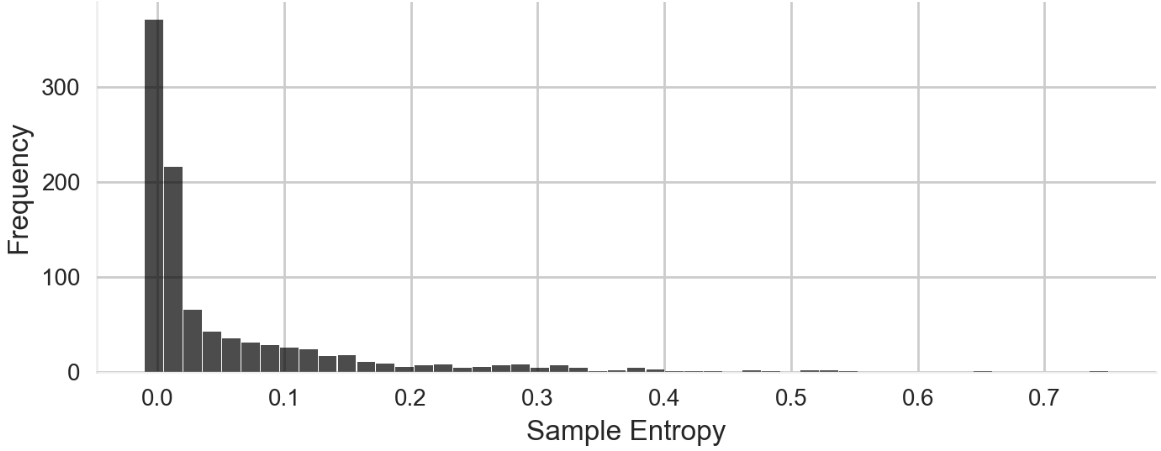


Figure 14: A histogram of sample entropy around maximum energy feature values.

values monotonically increase with degree of fractionation. A linear decision boundary is suited to model such a relationship as we don't expect their 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.) The suitability of a linear decision boundary can be seen in Figure 13, in which we can see that reasonably good predictive performance could be achieve by dividing this two-dimensional projection linearly into decision regions.

As expected, the difficulty experienced by each model was in the ability to distinguish'amber'feature vectors from 'green' feature vectors. This is not too much of a concern - the labelling of a mildly fractionated response was often subjective when the difference were subtle. Moreover, whilst class predictions are helpful in accessing model performance, in practice we are more interested in the models 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 of the models 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 induced AF and experiencing it's 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 not only halt the progression of an EP study unnecessarily, but could also cause medical professionals to lose confidence in predictions made by the model.

6.2 The Effects of Data Augmentation and Normalisation

Figure 15 compares the effect 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 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 over-fitting 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' .

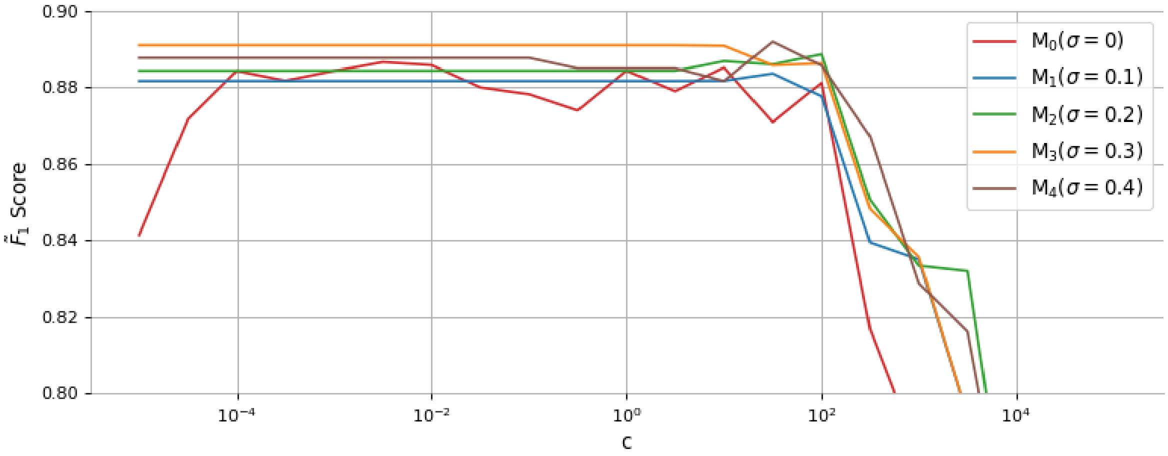


Figure 15: A comparison between the variation of the $\tilde{F}_1$-score with c between logistic regression models trained on datasets including augmented examples to varying degrees. $\sigma = 0$ denotes the training dataset with no augmented examples.

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 over-fitting 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 16 shows the location of augmented segments (with degree $\sigma = 0.3$) in feature space using the same LDA projection as in Figure 13. 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 real heavily fractionated examples. σ controls the degree to which this extrapolation is made. As σ increases, the augmented segments become more and more dissimilar to their parent segment moving deeper into the surrounding feature space.

It can be seen that model M_3 outperforms all other models, having a maximum $\tilde{F}_1$-score of 0.892 for all values of $c \leq 10^1$. 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

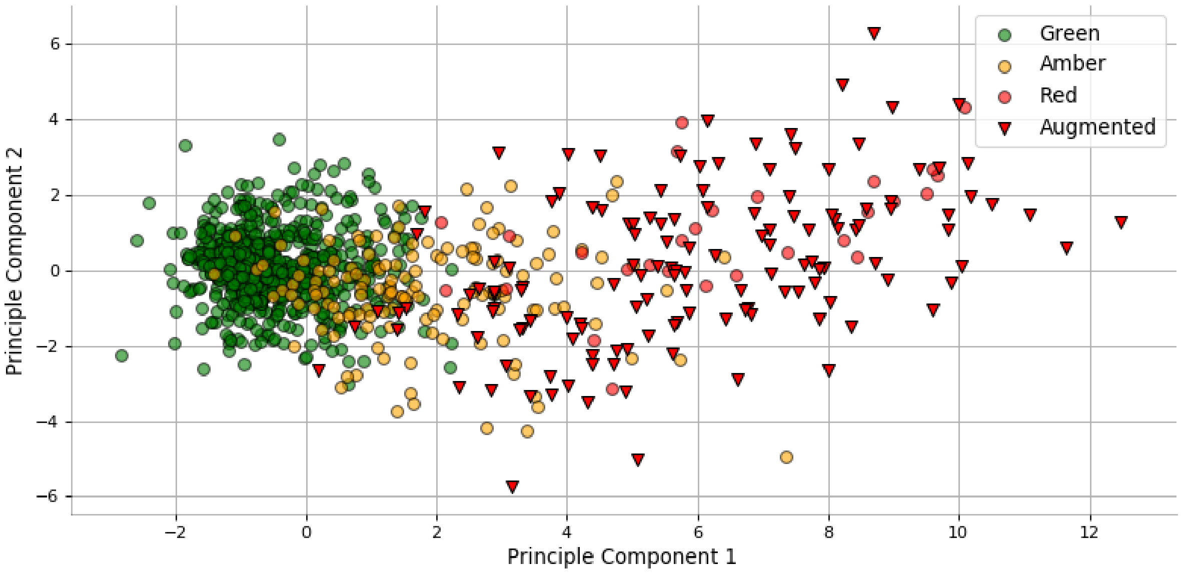


Figure 16: The same LDA transformation as shown in Figure 13 together with augmentated examples using $\sigma = 0.3$.

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 5 to 1. 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 sound the alarm bells when a response is heavily fractionated with high accuracy, we don’t wish to sound the alarm bells 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 otherwords, 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 M_3 would be carried forward to the test dataset.

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 $\sigma = 0.1$, 0.2, 0.3 and 0.4, respectively. Again we see that the greatest performance is achieved using $\sigma = 0.3$, suggesting that $\sigma = 0.3$ 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 we have 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 =0.1, 1, and 10 are shown in Table 10.

	$c = 0.01$			$c = 1$			$c = 100$		
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'.

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 in 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 whatsoever.

Another revealing feature of the weight vectors for $c = 100$ are the values corresponding to DTW Distance. We 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 a-typical. 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 the patient's typical, it is likely to show no signs of fractionation.

Finally, it is seen that the 6 largest weight in $\mathbf{w}^{red}$ when $c = 100$ correspond to three pairs of raw and conditioned feature values, percentage fractionation, width of maximum energy and location of maximum energy - we can deem these three features as the most important in identifying severe fractionation (it is somewhat re-assuring that the carefully hand-engineered features have been selected as opposed to some of the more simple values). 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 $\sigma = 0.3$ is shown in Table 11. The corresponding $\tilde{F}_1$-score is 0.835. 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 visa-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.

		Predicted		
		Green	Amber	Red
Actual	Green	**259**	49	0
	Amber	7	**47**	13
	Red	0	2	**12**

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

In order to understand why the model made the mistakes it did, it is instructive to identify the locations of the misclassified segments in feature space. Figure 17 shows the LDA projection of test feature vectors onto the same principal components shown in Figure 13. The two misclassified 'red' feature vectors appear to lie in a region of feature space dominated by amber . In fact, these two misclassifications are all for responses recorded by the same electrode for the same patient, who experienced AF later on in the EP study. Figure 18 directly compares these segments with the 'typical' segments for that patient.

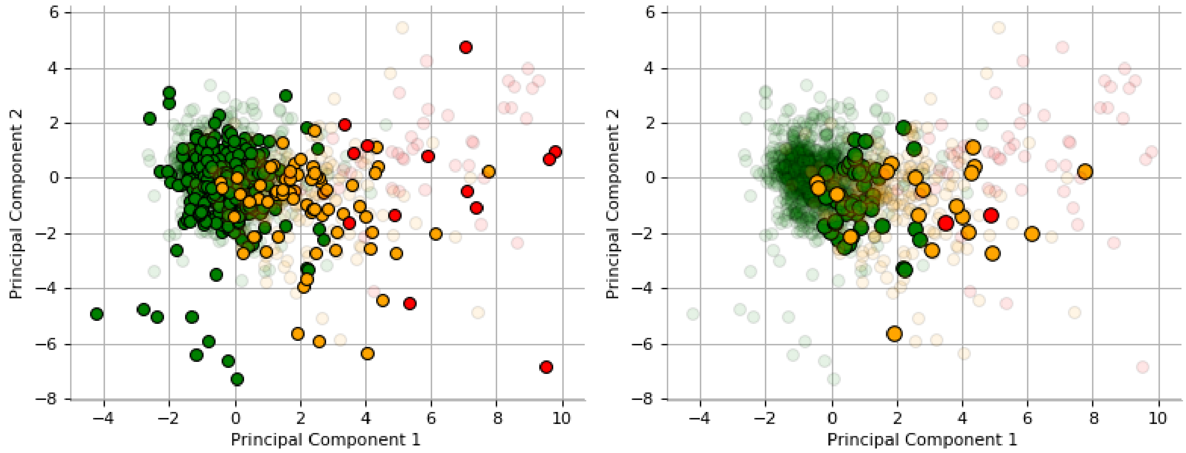
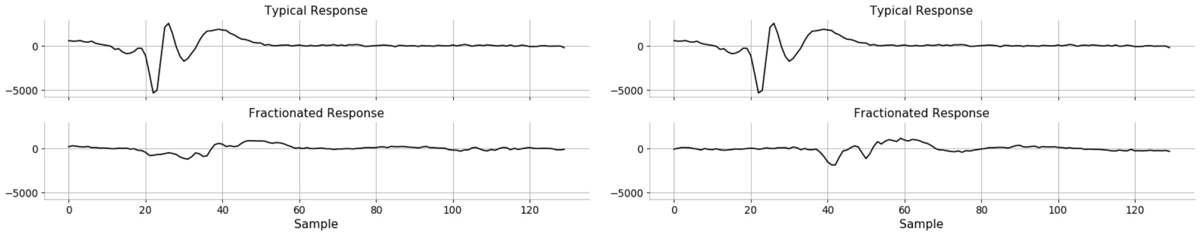


Figure 17: An LDA projection of test feature vectors onto two dimensions. The left plot shows the locations of all feature vectors and their corresponding true labels. The right plot shows only the locations of misclassified feature vectors, and their corresponding true labels.



(a) CS1-2: 290ms.

(b) CS1-2: 280ms.

Figure 18: A comparison between the misclassified responses, and typical response, for patient AF14.

We can get a sense for how fractionated the model believes these segments to be by computing the fractionation score, defined by

$$f_{score} = p(\mathcal{C}_{amber}|\phi(\mathbf{x})) + 2p(\mathcal{C}_{red}|\phi(\mathbf{x})). \tag{30}$$

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 18a has an f_{score} of 1.490, whereas the segment shown in Figure 18b has an f_{score} of 1.163. Reassuringly, the model predict both segments to be more than just mildly fractionated. It is also clear that segment in Figure 18b resembles the typical response more so than the segment in Figure 18a, which is reflected in their corresponding f_{score}. Thus, although the model misclassified these heavily fractionated segments, there is some truth in the models predictions.

6.5 AF versus Non-AF Patients

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

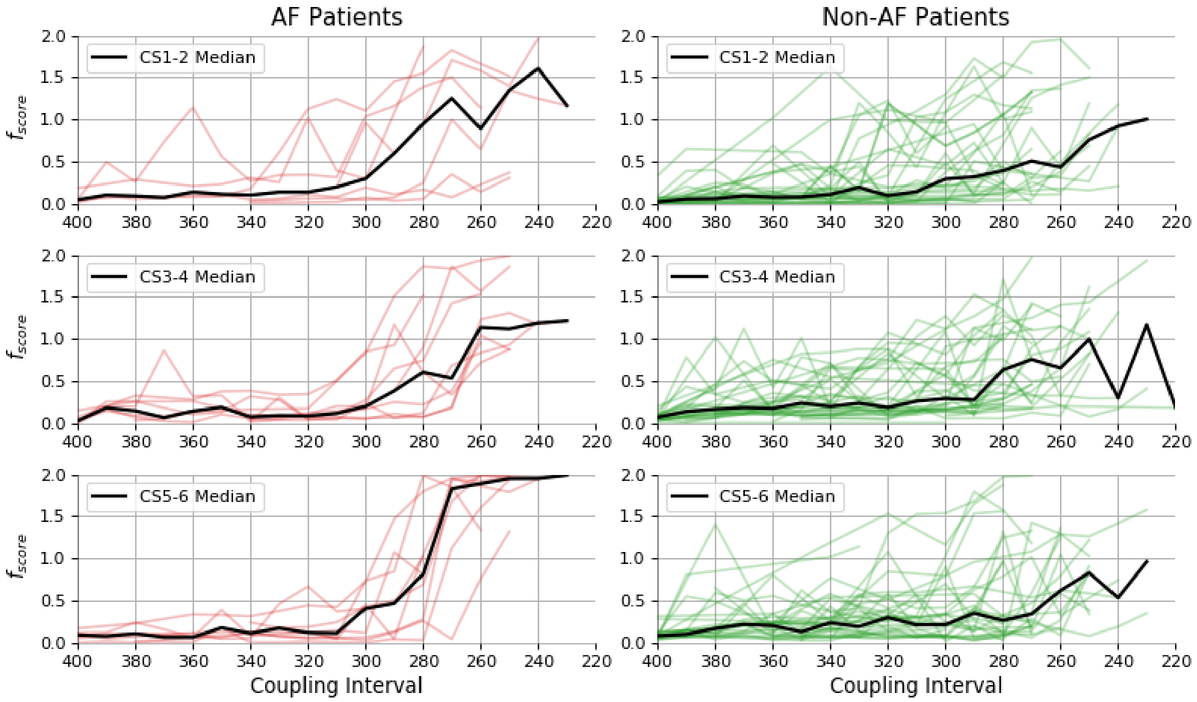


Figure 19: The progression of f_{score} for each patient, and electrode, in the groups of patients AF and non-AF. The progression of the median f_{score} for all patients in each group is shown in black.

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 300ms 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 one AF patient ends on an f_{score} greater than 1.5, whereas only 3 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 310ms compared to around 270ms for non-AF patients. It is also clear that a great deal of variability between the degree of fractionation within each group exists, even within non-AF patients. We expect this large amount of patient variability - 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 (SVT’s), i.e. irregular heart rhythms involving the ventricles. 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 as the S1/S2 interval is decreased.

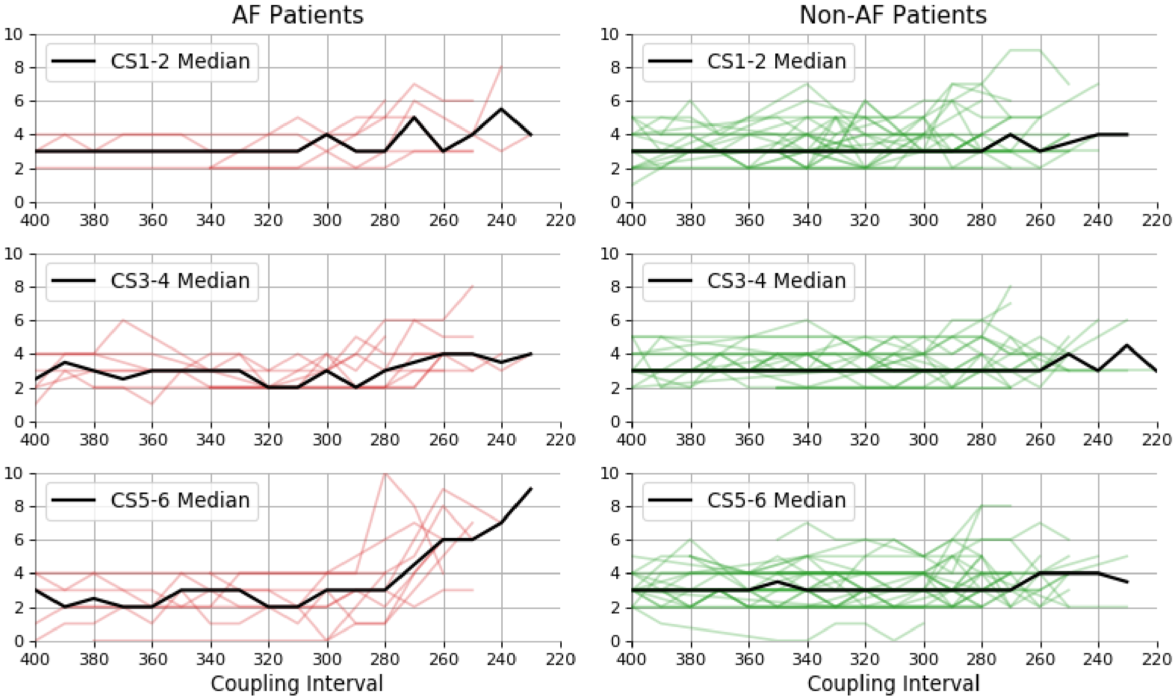


Figure 20: The progression of number of peaks for each patient, and electrode, in the groups of patients AF and non-AF.

Figures 20 and 21 shows similar plots, comparing the progression of the features number of peaks and width of maximum energy, respectively. As detailed in section 2, most previous efforts to distinguish between AF and non-AF patients compared features equivalent to these. In both cases, the difference between AF and non-AF patients are much more subtle than seen in Figure 19. 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 f_{score}. It only the responses in electrode CS5-6 for which the difference in progression of feature values between AF and non-AF is obvious; however, less so than the progression seen in Figure 19. The use of a model to determine degree of fractionation is not only successful in doing so, but is also shown to distinguish between the responses of AF and non-AF more so than the features used in previous studies [6–9].

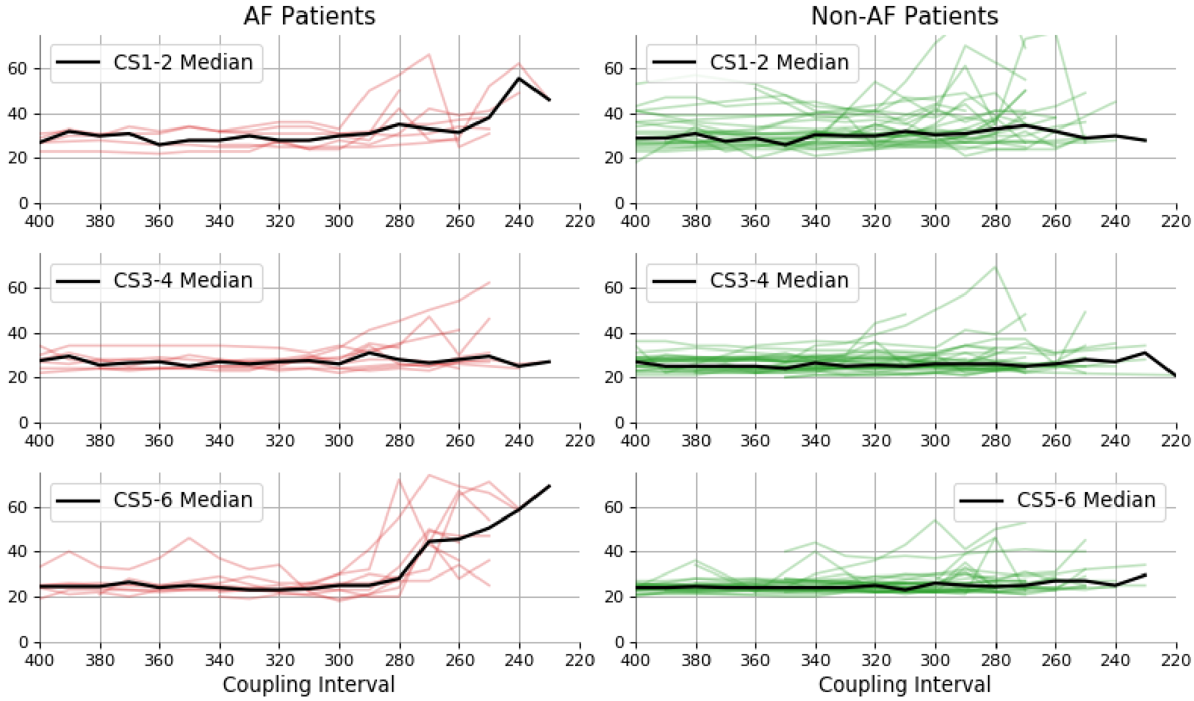


Figure 21: The progression of width of maximum energy for each patient, and electrode, in the groups of patients AF and non-AF.

These success of these results motivate 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 little 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 ground truth labels available 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 at 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 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 of a dataset to construct a model that makes confident predictions.

Instead, the degree of risk a patient is at can be inferred from a plot of the f_{score}, as in Figure 19, generated in real-time - this is possible due to the computationally efficiency of the algorithm and feature extraction process. If the patients response exceeds some threshold, say $f_{score} = 1$, for consecutive S1/S2 intervals then the patient could be deemed at being at risk to 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 19.

6.6 Comparison of with Expert Analysis

Table 12, shown in the Appendix, summarises the analysis of patient responses across three medical professionals. It is compared with the maximum f_{score} and maximum location of maximum energy across all patient responses. Patients highlighted in green and red indicate a strong agreement and disagreement, respectively, between the analysis made by the model and medical professionals. In total, there are only 7 cases in which the analyses strongly agree, compared to 16 strong disagreements, most of which occurring for AF patients. The results favour the analysis of the model over that of expert assessment in being able to predict the risk the patient is at to AF - almost all maximum f_{scores} indicate severe fractionation ($\frac{5f_{score}}{2} \geq 4$) in AF patient responses. In comparison, in only three AF patients does the average professional assessment of fractionation exceed 3. In fact, stopping the EP study if an f_{score} greater than 4.45 was recorded would have prevented 10 out of 13 cases of induced AF, and halted only 18 out of 55 studies in total. In contrast, in the 26 cases in which a single medical professional declared that the EP study should have been stopped early, in only four of which did the patient go on to experience induced AF. This overwhelmingly supports 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 decision boundary for f_{score} is largely arbitrary, given more data it could be optimised across patients.

7 Limitations

7.1 Limitations of the Model

One limitation in the construction of the logistic regression classifier is the dependence of half of the features on a patients 'typical' response. In this sense, the model does not determine the fractionation of truly unseen data. However, in practice it is entirely feasible that prior to the onset of the Antegrade Curve, the 'typical' atrial response is identified by a medical professional and fed into the model for use later on.

7.2 Limitations of the Analysis

In section 6.5 we compared the progression of the degree of fractionation of AF and non-AF patients (Figure 19). Whilst the differences between the AF group and non-AF group where significant, it is clear that the responses of some patients in the non-AF group become extremely fractionated. As noted - this is could be because such patients actually were at risk to AF, but avoided entry during the EP study. However, it is also true that all non-AF patients underwent the EP study because they experienced supra-ventricular tachycardias (SVT's), i.e. irregular heart rhythms involving the ventricles. It can be expected that this population of patients are more likely to show fractionated responses than the general population. Thus, the difference between the progression of f_{score} of AF patients and non-AF patients from the general population is likely to be more significant than the difference between the group of non-AF patients used in this study.

8 Suggestions for Future Work

Decisions were made against the use of a recurrent neural network and the construction of a model to explicitly predict the risk of AF due to the availability and incompleteness of the data. Moreover, the biases within the data (i.e. all patient's having SVT's) prevented the responses of AF patients being compared directly to a general population. Thus, future efforts could be made to collect a dataset without these limiting aspects and perform the subsequent analysis that was decided against in this report.

Additionally, it would be beneficial 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 (VF).

9 Conclusion

In this report, a model to determine the degree of fractionation of a patient's atrial response based upon extracted features was developed. It was found that the logistic regression classifier was able to distinguish between not at all, mildly and heavily fractionated responses more so than the naïve Bayes classifier. Furthermore, it was found that augmenting fractionated responses using scaling via a cubic spline provided a more generalised representation of fractionation than the few real fractionated responses. The logistic regression classifier, after training on an augmented dataset, achieved a sample average F_1-score of 0.835 over the three classes of fractionation on a test dataset. More importantly, the classifier did not misclassify any heavily fractionated responses as having no fractionation, and visa-versa. An analysis of the weights of the logistic regression classifier suggested that the location of maximum energy, width of maximum energy and percentage fractionation are the three most important features in determining whether or not a response is fractionated.

A continuous metric for the degree of fractionation, the f_{score}, comprising of a linear combination of the probabilistic outputs of the logistic regression classifier was introduced. Comparing the progression of f_{score} as the S1/S2 interval decreased between AF and non-AF groups of patients showed that AF patients exhibit a characteristic sharp increase in degree of fractionation at around 300ms, whereas for non-AF patients this increase was both smaller and much more gradual - agreeing with prior assumptions. This was found to be especially true in electrode CS5-6. A comparison between the progression of f_{score} and metrics used in previous studies suggested that f_{score} was more useful in tracking the progression of fractionation, and hence determining the risk a patient is at to AF. It was shown that the maximum f_{score} across all responses for each patient was **not** consistent with the assessment of fractionation by medical professionals. However, the use of f_{score} in determining when to halt the progression of an EP study to prevent induced atrial fibrillation was significantly more effective than professional assessment. This strongly supports the use of the model in assisting in the analysis of electrogram recordings, even when a medical professional is available.

References

[1] NHS. Arrhythmia, Jun 18AD. URL https://www.nhs.uk/conditions/arrhythmia/.

[2] Elise Dusseldorp, Thrse Van Elderen, Stan Maes, Jacqueline Meulman, and Vivian Kraaij. A meta-analysis of psychoeducational programs for coronary heart disease patients. *Health Psychology*, 18(5):506519, 1999. doi: 10.1037/0278-6133.18.5.506.

[3] Richard L. Page, Jos A. Joglar, Mary A. Caldwell, Hugh Calkins, Jamie B. Conti, Barbara J. Deal, N.a. Mark Estes, Michael E. Field, Zachary D. Goldberger, Stephen C. Hammill, and et al. 2015 acc/aha/hrs guideline for the management of adult patients with supraventricular tachycardia. *Circulation*, 133(14), 2016. doi: 10.1161/cir.0000000000000311.

[4] Miguel A. Barrero. Garcia, Laurent Macle, and Paul Khairy. *Electrophysiology for Clinicians*. Cardiotext Publishing, 2012.

[5] Kurt C. Roberts-Thomson, Peter M. Kistler, Prashanthan Sanders, Joseph B. Morton, Haris M. Haqqani, Irene Stevenson, Jitendra K. Vohra, Paul B. Sparks, and Jonathan M. Kalman. Fractionated atrial electrograms during sinus rhythm: Relationship to age, voltage, and conduction velocity. *Heart Rhythm*, 6(5):587591, 2009. doi: 10.1016/j.hrthm.2009.02.023.

[6] Ching-Tai Tai, Shih-Ann Chen, Jyh-Woei Tzeng, Benjamin I Kuo, Yu-An Ding, Mau-Song Chang, and Liang-Yu Shyu. Prolonged fractionation of paced right atrial electrograms in patients with atrial flutter and fibrillation. *Journal of the American College of Cardiology*, 37(6):16511657, 2001. doi: 10.1016/s0735-1097(01)01215-3.

[7] Yuewu Qin, Muneshige Kaibara, Tetsuya Hirata, Osamu Hano, Zhigan Liu, Kimio Tsukahara, Takashi Lshimatsu, Chiaki Ueyama, Motonobu Hayano, Katsusuke Yano, and et al. Atrial conduction curves in patients with and without atrial fibrillation. *Japanese Circulation Journal*, 62(4):289293, 1998. doi: 10.1253/jcj.62.289.

[8] 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):14151420, 2008. doi: 10.1093/europace/eun282.

[9] P Platonov. Further evidence of localized posterior interatrial conduction delay in lone paroxysmal atrial fibrillation. *Europace*, 3(2):100107, 2001. doi: 10.1053/eupc.2001.0150.

[10] Jane Caldwell and Damian 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):347353, Nov 2012. doi: 10.2174/157340312803760848.

[11] Divyanshu Gupta, Damian Redfearn, Javad Hashemi, and Selim Akl. A novel method for automated fractionation detection in ventricular tachycardia. *2016 Computing in Cardiology Conference (CinC)*, Sep 2016. doi: 10.22489/cinc.2016.269-519.

[12] Christopher M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[13] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: algorithms and applications*, page 37, 2014.

[14] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana 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*, ICMI 2017, pages 216–220, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5543-8. doi: 10.1145/3136755.3136817.

[15] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):17351780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[16] Eamonn J. Keogh and Michael J. Pazzani. Derivative dynamic time warping. *Proceedings of the 2001 SIAM International Conference on Data Mining*, 2001. doi: 10.1137/1.9781611972719.1.

[17] Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6), 2000. doi: 10.1152/ajpheart.2000.278.6.h2039.

[18] David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. doi: 10.1093/biomet/81.3.425.

[19] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial.

A Appendix

		Fractionation			Delay			Stop EP Study				Model Analysis	
Type	Patient	Mean	Min	Max	Mean	Min	Max	Count	Mean	Min	Max	Max F1	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
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	0	280	280	280	1.71	44

Table 12: A comparison between a group of three medical professionals collective analysis of each patient's response, and the output of the model. In green: experts and model agree on severe fractionation. In red: experts and model disagree on sever fractionation. Average professional assessment of fractionation greater than 3, and maximum f_{score} greater than 4, are highlighted in bold.