# Development of Machine Learning Models with Applications in Cardiovascular Research

Ryan Andrew Alan Bellfield

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with the Liverpool Centre for Cardiovascular Science

February 2024

**Table of Contents**

# Abstract

Cardiovascular disease (CVD) is one of the leading burdens on modern healthcare globally in terms of mortality, loss of health and healthcare costs. CVD covers all conditions that affect the heart and circulatory systems. Artificial intelligence (AI) and machine learning (ML) are being increasingly leveraged to help improve diagnosis, prognosis, treatment, and management of CVDs. This thesis aims to develop ML approaches that can generate novel, meaningful insights into several aspects of cardiovascular research. First, in Chapter 3 we use convolutional neural networks (CNN) to quantify the effect ECG data format has on ML predictive performance, through the clinical task of detecting myocardial infarction, providing the first results in determining the optimal ECG data format for ML modelling.

The remaining analysis leverages the unsupervised, probabilistic ML technique generative topographic mapping (GTM). The analysis aims to generate 2-dimensional representations of data and propose different approaches that can identify large macro-clusters within the reduced dimension. Doing this gives an understanding of which patients/participants within a data set are clinically similar, along with interpretable visualisations that explain the rationale behind each cluster. Chapter 4 contains the first outline of this methodology, developed on a non-cardiovascular dataset, to demonstrate the generalisability of such a methodology. Through this approach, we propose a novel freedom of expression index that provides an understanding of the level of restrictions placed on the population of a country. This index is defined by macro clusters generated through aggregating the normalised information contained in the GTM reference vector outputs. Chapter 5 applies this methodology to generate clinically relevant AF phenotypes for specific patient cohorts, from the general and the critical care populations. We propose a new methodological approach to achieve this that implements hierarchical clustering, again on the GTM reference vector outputs, to generate the phenotypes.

Finally, Chapters 6 and 7 investigate the athlete's heart, defined as the physiological changes that the heart undergoes due to exercising for prolonged periods. Chapter 6 contains an in-depth scoping review, evaluating the current ML applications in athlete's heart and identifying the gaps for future research. Chapter 7 investigates features automatically extracted from ECG recordings from elite footballers, cyclists, rugby league players, and ultra runners to further the understanding of healthy athlete's hearts. The methodology in Chapter 5 was further developed here to define a novel approach that uses magnification factors to define neighbourhoods in the 2-dimensional data representation, to carry out constrained hierarchical clustering on the reference vectors.

# Declaration

I, Ryan A.A. Bellfield, confirm that the work presented in this thesis is my own. Furthermore, I confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Ryan A.A. Bellfield

Word count (excluding acknowledgement, appendices, and references): 39,717 words.

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors Dr Ivan Olier-Caparroso and Dr Sandra Ortega-Martorell. Your mentorship and tutelage have played an instrumental role in shaping my research and have provided me with the tools, experience, and confidence to continue with my academic career. You guys have gone above and beyond for me, from meeting with me every week (including the many, *many*, impromptu meetings) to including me in several external projects along the way, and for that, I am profoundly grateful. I would also like to thank the rest of my fantastic supervisory team, Prof David Oxborough and Prof Gregory Y. H. Lip for all their help and advice throughout my PhD journey. Thank you for all your time and effort in providing expert medical advice and support, which has been invaluable for my research. I would also like to thank LJMU for funding my research.

I would also like to say a huge thank you to the rest of the staff and my fellow researchers at LJMU. You lot have supported me all the way, provided many laughs, and made this whole journey that little bit sweeter. I would like to say a special thank you to Ian, who showed me early on in my academic journey that it's okay to love Maths and enjoy a bit of Rock N' Roll in the process!

I want to say another special thank you to my Mum and Dad who have been supportive during my PhD years. It can't be easy having a stressed-out 27-year-old PhD student moping around the house, so thank you for putting up with me! (At least you can now ask me questions other than "how much longer is it until you finish?").

Last, and by no means least, I want to say a massive thank you to my wonderful wife to be Hannah. For as long as we've been together, you have pushed me to be the best person I can be. You have shown nothing but unwavering support and encouragement throughout this whole PhD process. Even when working late and on weekends (and working late on weekends…), you have always been patient with me and for that, I can never say thank you enough, I couldn't have done this without you!

# Publications and Dissemination of Results

**Peer-Reviewed Journal Publications**

i. **Bellfield, R. A. A.**, Ortega-Martorell, S., Lip, G. Y., Oxborough, D., & Olier, I. (2022). The athlete's heart and Machine Learning: A review of current implementations and gaps for future research. *Journal of Cardiovascular Development and Disease*, *9*(11), 382. https://doi.org/10.3390/jcdd9110382 ***(Chapter 6)***

ii. Ortega-Martorell, S., **Bellfield, R. A. A.**, Harrison, S., Dyke, D., Williams, N., & Olier, I. (2023). Mapping the global free expression landscape using machine learning. *SN Applied Sciences*, *5*(12). https://doi.org/10.1007/s42452-023-05554-x ***(Chapter 4)***

iii. Ortega-Martorell, S., Olier, I., Hernandez, O., Restrepo-Galvis, P. D., **Bellfield, R. A. A.**, & Candiota, A. P. (2023). Tracking therapy response in glioblastoma using 1D convolutional neural networks. *Cancers*, *15*(15), 4002. https://doi.org/10.3390/cancers15154002

iv. **Bellfield, R. A. A.**, Ortega-Martorell, S., Lip, G. Y., Oxborough, D., & Olier, I. (2024) Impact of ECG data format on the performance of Machine Learning models for the prediction of Myocardial Infarction. *Journal of Electrocardiology*, https://doi.org/10.1016/j.jelectrocard.2024.03.005 ***(Chapter 3)***

**Manuscripts Submitted for Publication**

i. **Bellfield, R. A. A.**, Olier, I., Lotto, R., Jones, I., Dawson, E.A., Li, G., Tuladhar, A. M., Lip, G. Y, Ortega-Martorell, S. AI-based derivation of atrial fibrillation phenotypes in the general and critical care populations (Under Review as of February 2024) ***(Chapter 7)***

**Conferences and Presentations**

i. Presented at the 3-Minute Thesis Competition 2023 – Reached faculty finals - "Running into Issues": Development of Unsupervised Machine Learning Models with Applications in Athlete's Heart Research

ii. Presented at the LJMU Post Graduate Research Day 2023 - "Running into Issues": Development of Unsupervised Machine Learning Models with Applications in Athlete's Heart Research

iii. Liverpool Centre for Cardiovascular Science (LCCS) Research Group meetings. Dec 2021, Nov 2023, Oral Presentation.

# 1. Chapter 1: Background & Introduction

## 1.1. Introduction

Cardiovascular disease (CVD) is a hypernym for conditions that affect the heart and circulatory system, for example, congestive heart failure. CVD, in its many forms, is one of the largest burdens on modern healthcare globally [1] in terms of mortality, contributions to loss of health and healthcare costs [2]. Globally, as of 2021, it is estimated that 32% of all global deaths relate to CVD [3], putting the economic burden of CVD across the 27 European Union countries at €282 billion annually [4], with the projected 2035 figure for the USA being around $1.1 trillion [5]. CVD is not just a singular issue however, as people with CVD are more likely to suffer with one or more additional chronic diseases, referred to as comorbidities [6,7]. This again poses additional challenges as comorbidities can lead to lower quality of life and increased mortality [7]. Furthermore, comorbidities may also affect CVD treatment, as the presence of a certain disease may limit the effectiveness of treatment plans and clinical practical guidelines [8]. It is therefore imperative to diagnose CVD conditions early to maximise the possibility the disease can be properly managed with appropriate treatments [3]. Currently, the most effective methods for diagnosing CVD involve using invasive methods such as blood tests, and non-invasive methods such as electrocardiograms (ECGs), echocardiograms and cardiac magnetic resonance imaging (CMRI) [9] individually or as part of a series of tests. Diagnosing CVD is not straightforward however due to factors such as the heterogeneity of the diseases [10,11].

### 1.1.1. Machine Learning in Cardiovascular Research

In the pursuit of improving the speed and accuracy of CVD diagnoses and to further the understanding of CVD, the use of artificial intelligence (AI) and machine learning (ML) techniques has grown rapidly, as shown in Figure 1. With the constant developments and advancements of AI/ML models showing no signs of slowing down, the performance and capabilities of these approaches will continue to improve which has led to the sentiment that the use of AI in cardiology will not only be beneficial, but inevitable [12,13]. The popular belief for future AI usage within a clinical setting will be one that serves to aid medical experts through computer-aided decision-making systems, rather than serve as direct replacements [12].

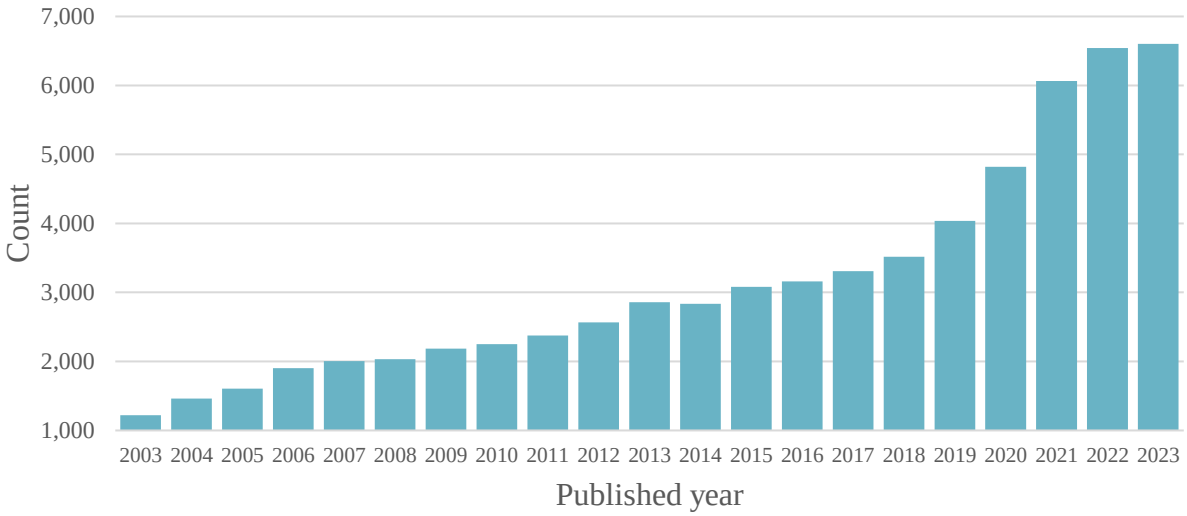



*Figure 1. Growth in the number of AI in CVD publications between 2003 and 2023*

The reach of AI models into the various cardiovascular research areas is widespread. Models have been developed using the full spectrum of data available, from routinely collected electronic health record data [14] to image and signal data generated from tests such as CMRI [15] and ECG recordings [16]. A plethora of techniques have also been employed to model this data from within both the supervised and unsupervised learning groups. Several studies also refer to using deep learning (DL) techniques. Techniques that fall under the umbrella of DL can be classified as either supervised or unsupervised, with the term deep referring to a model architecture that has many layers/parameters [17].

Supervised learning techniques are generally employed to tackle problems relating to prediction, which in the context of cardiovascular research, relates to tasks such as predicting the presence of a particular disease. These techniques have been successfully applied to predicting acute myocardial infarction [18], stroke risk prediction [19], and the prediction and identification of risk factors for congenital heart failure [20], to name a few examples. Several projects have also passed from theory to practical real-world applications, with a common example being the atrial fibrillation (AF) detection algorithms built into many of the current smart watches [21]. This is a key step as it is difficult to implement AI in real-world scenarios, especially within healthcare [22], so examples like this further emphasise the drive to utilise the power of AI applications.

Unsupervised learning techniques differ in that they are more generally applied to dimensionality reduction and clustering tasks. In the context of cardiovascular research, clustering techniques help further the understanding of diseases by providing information about key similarities and differences within the data. A common way this is implemented is to identify

clinically relevant subgroups within a population of patients with a condition of interest, and then highlight the key characteristics of each group, hence providing a deeper understanding of the different ways a condition may present itself. This was implemented successfully to characterise patients with known coronary artery disease undergoing myocardial perfusion imaging [23] and identify novel subgroups in heart failure patients [24].

For all the identified benefits that AI and ML bring to cardiovascular research, several challenges must be acknowledged and overcome to facilitate successful integration. First, as is the case with all ML modelling, is the concern about model overfitting. Overfitting refers to the situation whereby an ML model learns too much from the noise present within a data set rather than the key underlying relationships. This can result in the model not generalising well to new unseen data and lead to a higher-than-expected number of incorrect predictions [25]. This can arise in scenarios where a large, complex ML model is trained on a small dataset [26]. Even though a concern in every field, there is a particular sensitivity to it within healthcare applications as the cost of misclassifications is high, both monetarily and, more importantly, for the clinical outcome of the patient. Another key issue ML models face regarding widespread adoption lies in model interpretability. Being able to understand how and why a model made a particular decision is vital to not only ensuring optimal patient care, but also to building trust in such systems [27]. Linear models, such as logistic and linear regression, provide a fully interpretable result as to the rationale behind the final decision. As models become more complex, eventually reaching the high level of complexity of modern DL models, the sheer number of parameters makes interpreting the result extremely difficult. These models are commonly referred to as “black box models”. However, more complex models have been shown to perform better in many scenarios and be much more versatile as they can be applied to many more varied data formats. This leads to a situation whereby when developing ML there is a trade-off between model performance and interpretability and the correct decision can be difficult to determine [28]. Therefore, pursuing methodologies that provide interpretability is important and should be actively considered during the modelling phase of any analysis. Several methods have been proposed to achieve this, such as activation maps that highlight relevant areas of the input based on its effect on the output.

### 1.1.2. Thesis Scope and Motivation for Clinical Problems

The key goal of this thesis is to develop and implement ML approaches that can generate novel, meaningful insights into several aspects of cardiovascular research. Covering all areas of cardiovascular research within one thesis would neither be practical nor very informative, we therefore decided to address three specific areas: ECG ML modelling, AF, and athlete’s heart. Starting with the ECG ML modelling, as briefly mentioned above ECGs are commonly used in a

medical setting to aid in diagnosing CVD. They are also commonly used in ML analysis, with the ECG data format used for the analysis commonly being dictated by what was available. We consequently aimed to evaluate quantitatively which ECG data format was optimal for ML modelling performance. In keeping with wanting to address model interpretability, we also qualitatively assess each ECG format using activation maps to provide an understanding of how each model learns from the data, along with the pure overall performance. Addressing this will provide vital insight to help direct future ML development using ECG data.

The next clinical area addressed was AF, which has highly benefited from the use of AI and the development of AI applications [29]. The prediction of AF, and complications related to AF, is generally performed using clinical risk scores however their predictive accuracy is generally limited due to several factors such as the inherent complexity of the disease. There has been a growth in applications of unsupervised ML approaches that instead aim to understand the inherent key characteristics in the data to form clusters of similar patients. AF patients are conventionally classified based on the disease sub-types or arrhythmia patterns, which may not be adequate in certain situations. Unsupervised ML has therefore been applied to generate more in-depth sub-groups of AF, known as phenotypes, that can facilitate the development of more personalised treatments. The methodologies currently employed to achieve this though are not generally suited to modelling complex, non-linear relationships in the data along with other limitations such as being less interpretable. We therefore aimed to develop a more robust methodology, based upon generative topographic mapping (GTM) (discussed further in Chapter 2), to derive these AF phenotypes. This approach would provide advantages such as the ability to handle uncertainty and be less sensitive to noise within the data which would lead to more specific patient profiles.

The final clinical focus of the thesis is the athlete's heart. The athlete's heart differs slightly as it is not a disease, however the name given to the physiological changes a heart undergoes during extreme training regimes, like the ones followed by elite athletes. The problem lies in that the physiological changes can mask changes made due to cardiovascular disease, which if left undiagnosed can lead to increased risk of adverse cardiac outcomes. Therefore, having reliable ways to identify these is crucial to avoid scenes such as those seen in recent years in top sporting events, such as the Euro 2020 Football Tournament. The application of ML in the area continues to grow, with many approaches leveraging supervised ML. However, these current applications are limited in that access to high-quality labelled data is difficult as well as the low prevalence of adverse cardiac outcomes being difficult to model. Like with the AF phenotypes, we aimed to implement a similar unsupervised approach to instead generate specific groups within a healthy athlete's population to provide a deeper understanding of what different healthy hearts look like.

### 1.2. Research Novelty

This thesis builds upon existing ideas and provides several areas of novelty. This relates to both the ML methodologies as well as the application of said methodologies. These novelties are listed briefly below:

- Compared three common ECG data formats and provided the first quantitative answer as to what data format is optimal for ML modelling. Additionally, we proved the viability of digitising ECGs and using the extracted signals for ML modelling.
- Produced the "index index", a novel index that ranks countries based upon the degree of censorship their populus faces. Results from this work resulted in a journal publication.
- Outlined a methodological approach to identify and characterise clinically significant phenotypes within AF populations, which was successfully applied to two cohorts of patients that represented AF in the general and critical care populations.
- Developed and implemented an end-to-end methodological approach to model different clinically relevant groups within a healthy athlete's heart population.

### 1.3. Thesis Overview

The research conducted as part of this thesis serves to contribute to the overarching theme of developing ML models to aid in cardiovascular research. There are many areas and nuances within the cardiovascular research field, and as such is reflected by the range of work and specific clinical problems addressed as part of the overall analysis. Following on from the introduction, **Chapter 2** outlines the methodologies used or referenced throughout the thesis, ranging from statistical tests to complex ML techniques. **Chapter 3** then presents the first analysis of the thesis, where the optimal data format for ECG modelling is explored. This chapter aims to identify the optimal data format choice for predictive modelling, in this case in the context of myocardial infarction (MI) prediction, along with providing guidance as to situations where non-optimal data formats may be more appropriate. This chapter is the only one of the thesis to leverage supervised ML, with the remainder of the analysis using unsupervised ML techniques starting with **Chapter 4**. This chapter presents analysis that investigates the global free expression landscape to create a novel index that ranks the level of censorship and restrictions a country places on its populus. Even though this topic falls outside the umbrella of cardiovascular research and may appear out of place within this thesis, this analysis facilitated the development and validation of a clustering methodology and workflow capable of generating novel, impactful outputs. Having this validated approach proved crucial as it provided a solid methodological baseline that served as a blueprint upon which the methodology for the other analyses was developed from.

The research conducted in **Chapter 5** shifts its focus back to cardiovascular research, more specifically AF. Using the blueprint outlined in Chapter 4, this chapter aims to propose a new methodology for identifying clinically relevant AF phenotypes (think of phenotypes as clusters for the purposes of intuition). Two AF cohorts were used in this development that represented the general and critical care populations. Chapters 6 and 7 focus on a more niche area of cardiovascular research in the athlete's heart. **Chapter 6** contains a full and detailed scoping review of the area in the context of ML applications. The review aims to present the current state of the ML applications in the area and highlight the limitations of the research and the gaps for future research. One of the review outcomes was that implementing unsupervised ML could provide novel insights into the understanding of the athlete's heart. The research outlined in **Chapter 7** therefore aims to achieve this very thing. Again, further extending the methodology from chapters 4 and 5, we explored generated athlete clusters based on ECG measurements automatically extracted from image ECGs. Lastly, **Chapter 8** reflects on the methodologies, workflows and results presented throughout the thesis to review the work as a whole and identify any potential improvements or limitations.

# 2. Chapter 2: Methodological Approaches

## 2.1. Introduction

The terms AI and ML are used interchangeably, with both seemingly referring to the same thing. However, there is a difference: AI aims to create intelligent systems that can simulate human intelligence to solve complex problems, automate tasks, and assist in decision-making across various domains; whilst ML is a subgroup within AI focused on developing algorithms and techniques that can learn from data. ML algorithms consist of a blend of statistical and mathematical techniques, which when combined produce a toolbox of methodologies that can be applied to the simplest regression tasks or to solve complex non-linear problems. The different techniques within ML can be broadly categorised into 4 main categories: supervised, unsupervised, semi-supervised and reinforcement learning. These categories indicate how a particular methodology learns from data, with each category having its benefits and drawbacks.

This chapter will cover the different methodologies, ML and otherwise, used within the thesis. These techniques were used to develop and evaluate novel approaches for the analysis presented. Only supervised and unsupervised learning techniques will be discussed in this chapter as semi-supervised and reinforcement learning techniques were not used within the thesis. The chapter will be organised by first discussing the supervised learning techniques used in the thesis, followed by the unsupervised learning methods and finally, the performance and evaluation methods used, in addition to the traditional statistical tests implemented. Other methodologies that are mentioned in the thesis, even though they may not be used directly, may also be included here for the sake of parity, and will be identified accordingly.

## 2.2. Supervised Learning Approaches

### 2.2.1. Neural Networks

The modern neural networks (NN) are an ML methodology that can trace its beginnings to the development of the perceptron in 1957 [30] which was inspired by the function and structure of biological processes that occur in the human brain. There are many types of NNs with varying types of architectures such as recurrent neural networks (RNN), convolutional neural networks (CNN) and large language models (LLM). These architectures differ in their complexity and depth; however, this wide range of model structures means NNs can be used to solve an extensive assortment of problems.

The general form of NNs consists of an input layer, hidden layers, and an output layer, with each layer containing a pre-defined set of nodes (or neurons). These nodes take information in,

apply a mathematical operation to the input, and pass the data to an activation function ($\varphi(v)$) that dictates the nodes' output. These nodes are connected to each other by learnable weights that are updated as the model is trained using a process called backpropagation, allowing the NN to learn the key relationships within a given dataset. The input layer, usually fixed at the size of the input data, is used to pass data from a dataset to the NN and does not itself carry out any computations. The hidden layers of an NN are where the main bulk of the computation occurs. There can be as little as one hidden layer with there, in theory, being no limit to the number of hidden layers, with the term deep NN (DNN) being the term often used for NNs with many hidden layers [31]. However, if too many hidden layers are used this can cause problems such as overfitting and computational intractability [32]. The output layer converts the information that has been learned by the NN from the final layer to an output in the desired format using an appropriate activation function.

*2.2.1.1. Artificial Neural Network (ANN)*

Artificial neural networks (ANN), also referred to as fully connected feed-forward neural networks (FCNN) or multi-layer perceptron (MLP), are a simple NN architecture that consist of an input layer, that passes through one or more hidden layers and then is finally processed by an output layer, as shown in Figure 2. The information is transmitted through the network using the function in equation (1):

$$y_j = \varphi\left(b_j + \sum_{i=1}^{n} w_{ji} x_i\right) \quad (1)$$

Where $y_j$ will be the input to node $j$ in the following layer, $\varphi$ is the activation function (common functions used are sigmoid, tanh and ReLU), $b_j$ is the bias, $w_{ji}$ is the weight of the edge connecting nodes $i$ (in the previous layer) and $j$ (in the following layer), and $x_i$ is the output of node $i$ (from the previous layer).

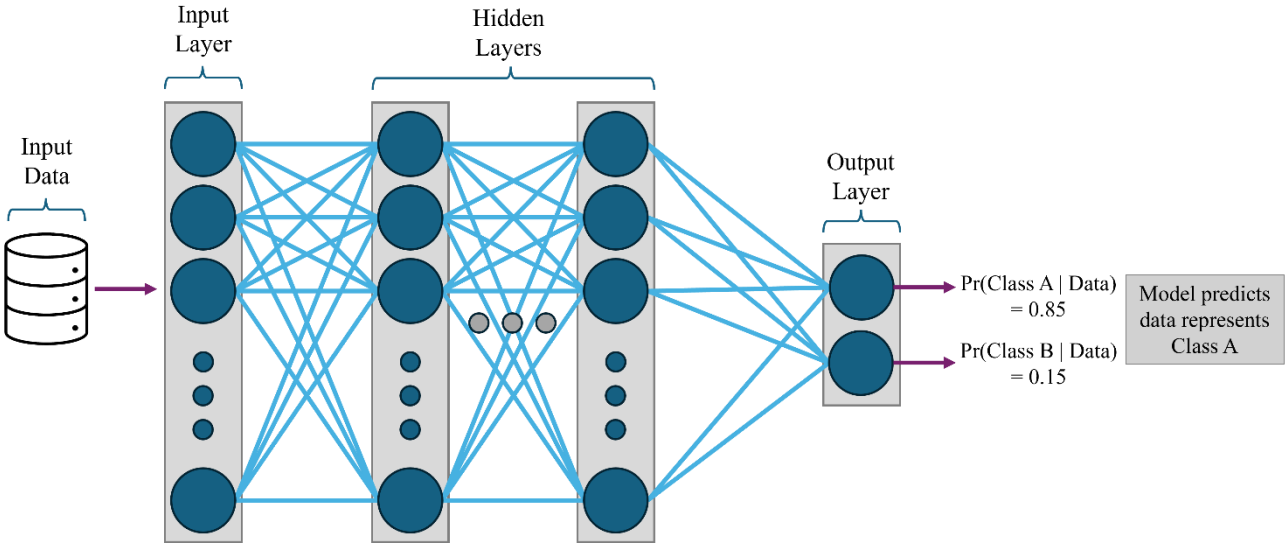



*Figure 2. Example of an ANN architecture*

#### *2.2.1.2. Convolutional Neural Network (CNN)*

CNNs are a NNs with a specific type of architecture that allow them to be applied to contextual datasets. 1-dimensional (1D) and 2-dimensional (2D) CNNs were used within the thesis that were applied to time series data (ECG signals) and image data (image ECGs) respectively. However, there are also 3-dimensional variants that can be applied to data such as videos. For the purposes of conciseness, CNNs will be explained through the lens of a 2D CNN, as the fundamentals apply to both 1D and 3D variants. The general CNN architecture consists of 3 main layer types [33] that can be grouped into two key parts: a feature learning stage and the prediction stage, as shown in Figure 3. The feature learning stage is used to capture the spatial and temporal dependencies within the input data. It achieves this by using two types of layer types called convolutional layers and the sub-sampling layers. In the convolutional layers, a mathematical operation called a convolution is used, which is a special type of linear function that explains the overlap between two functions as one function is shifted across another. This is applied in convolution layers by passing several learnable filters (also referred to as kernels) along the input to identify key elements, such as edges, with each filter output being passed through an activation function and combined to form the output of the convolutional layer in the form of feature maps. This is implemented using equation (2) [34,35].

$$y_j = \varphi\left(\sum_i K_{ij} \otimes x_i + b_j\right) \tag{2}$$

Where $y_j$ is the output of the j[th] convolutional layer, $\varphi$ is the activation function, $K_{ij} \otimes x_i$ represents the convolution of the filters with the i[th] input, with $b_j$ representing the bias. These convolution operations allow CNNs to achieve weight sharing [34], which is an inherent benefit of CNNs as it reduces the number of parameters in the network and can help improve the generalisability of the model by reducing the chance of overfitting [36].

The output from the convolutional layer is then passed to the sub-sampling layer, also known as the pooling layers, where a down sampling operation is performed. The down sampling operation does not affect the number of feature maps; however, it does reduce the dimensionality, with the size of the reduction dependant of the kernel that is passed over the feature maps. For example, if a 2x2 kernel is used, this will half the dimension of the feature map [33]. Common types of pooling operations are to take the max or average values of the section of the map covered by the kernel.

The prediction stage consists of an ANN that takes the output from the feature extraction stage and generates the desired output (for instance, scores of the likelihood that the image is part of a certain class [33]). The data must first be converted into a suitable form for the NN. This is achieved by flattening the output of the feature extraction layer into a 1-dimensional column vector.

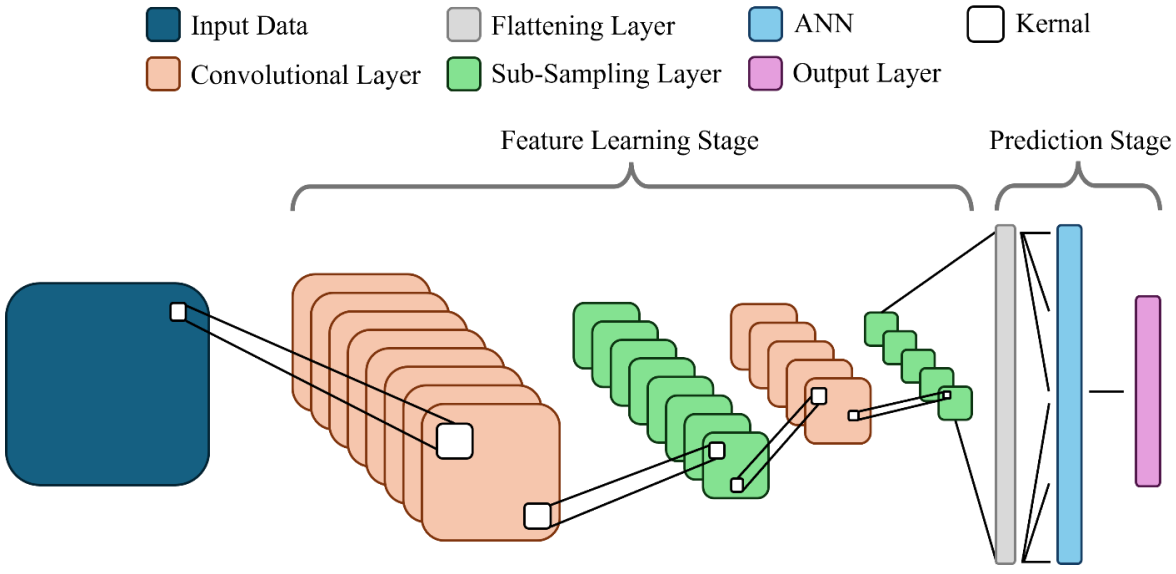



*Figure 3. Example of a CNN architecture*

*2.2.1.3. Regularisation methods*

Regularisation is key when training complex networks as it is an efficient way of reducing overfitting and improving model performance [37]. Overfitting refers to the situation whereby a model learns to represent the training data too closely, resulting in worse performance on new data as it can greatly affect the generalisability of the trained model. An example of this is shown

in Figure 4. The green line represents a model that has fitted the general relationship in the data, whereas the blue line represents a polynomial that has been trained to model the data too closely.

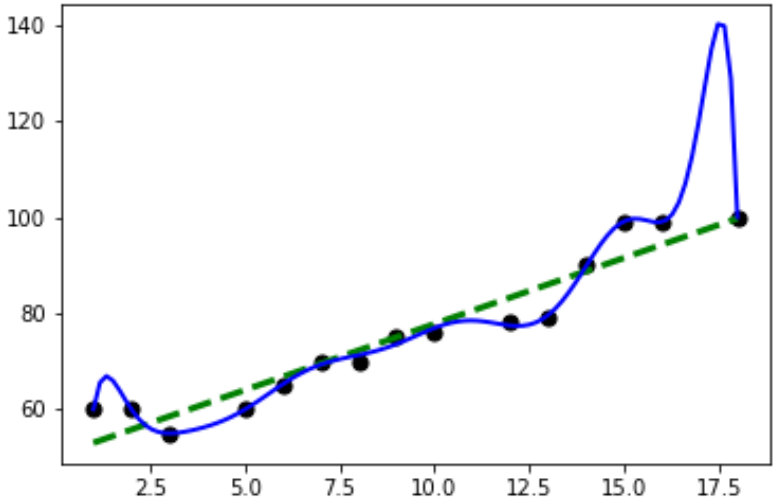



*Figure 4. An example of overfitting. The green line represents the line of best between the given data and the blue line represents a high order polynomial fitted to the same data.*

Whilst there are many methods for implementing regularisation within a NN (e.g. L1 and L2 regularisation and early stopping), this section will discuss only the two methods implemented within models developed as part of the thesis.

*Batch Normalisation*

Batch normalisation (BN) refers to the process of normalising the input to each layer for each batch during training. This approach has several benefits such as higher learning rates, that speed up the training time of the model, and reduces internal covariate shift, which is when there are changes within the distributions of internal nodes throughout the course of training [38]. BN works by applying the transformation in equation (3) to the input of a layer:

$$y_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2}}\gamma + \beta \tag{3}$$

Where $y_i$ is the i$^{th}$ output, $x_i$ is the i$^{th}$ input, $\mu_B$ and $\sigma_B^2$ are the batch mean and variance respectively, and $\gamma$ and $\beta$ represent learnable parameters that scale and shift the normalised value to ensure the original layer representation remains.

*Dropout*

Dropout is a stochastic regularisation technique that refers to the process of temporarily dropping nodes within a NN whilst training randomly with a probability $p$. This equates to sampling many "thinned networks" from the original architecture, During the testing stage, the full NN is used, with the weights of each node being scaled down by the probability $p$ a node was retained with during training. This is implemented within a NN by modifying slightly the equation (1) such that the term $x_i$ is replaced by equation (4) [39].

$$\overline{x_i} = r_i * x_i \tag{4}$$

Where $r_i$ is a vector of elements that have a probability $p$ of being 1. This vector is multiplied elementwise with $x_i$ to create the output $\overline{x_i}$, which defines which nodes will be included in the thinned network.

*2.2.1.4. High-Resolution Class Activation Maps (HiResCAM)*

It is not uncommon for ML models to depend on spurious correlations, relationships between two variables erroneously determined to be causal. Gradient based visual techniques are popular in helping provide understanding as to how a CNN model makes its predictions to aid in developing less biased models [40]. This is achieved by highlighting areas of the input data, via a heatmap, that are considered "important", providing context to the decision along with a visual method of evaluating if a model is learning from the correct area of the input. High resolution class activation mapping (HiResCAM) is an output level gradient-based method [40] that serves as a generalisation of the class activation mapping (CAM) technique [41]. HiResCAM addresses issues with other techniques, such as gradient-weighted class activation mapping (Grad-CAM) [42], whereby areas deemed as important to a model's decision-making do not reflect the actual locations used for prediction [40]. HiResCAM is calculated using equation (5) which takes the form:

$$\tilde{A}_m^{HiResCAM} = \sum_{f=1}^{F} \frac{\partial s_m}{\partial \boldsymbol{A}^f} * \boldsymbol{A}^f \tag{4}$$

Where $\tilde{A}_m^{HiResCAM}$ represents the attention map, $s_m$ represents the score of the model for class $m$ before it passes through the output activation function, $\boldsymbol{A}^f$ represents the feature maps produced by the final convolutional layer in a CNN, * represents the element-wise multiplication between, and $F$ represents the feature map dimension.

## 2.3. Unsupervised Learning Approaches

### 2.3.1. Clustering

Clustering is a subbranch of unsupervised learning and refers to methodologies grouping together data based on how similar they are, with these groups being known as clusters. These methods achieve this by identifying underlying relationships in the data without the need for the data to be labelled [43]. Clustering techniques can be generally categorised as falling into one of two categories: partitional clustering and hierarchical clustering [44]. Hierarchal clustering focuses on creating a hierarchal structure that clusters data together at different levels of granularity based on distance between individual or a subset of data. Partitional clustering on the other hand refers to the process of splitting data into distinct groups based on minimising some criterion function. This can include methods that assign each data point to one cluster only, or fuzzy methods that assign each data point a certain association to every clusters [44].

#### *2.3.1.1. Ward's Minimum Variance Method for Agglomerative Hierarchical Clustering*

Agglomerative hierarchical clustering refers to how the hierarchical structure of the data is generated. Using this approach, clusters are developed from the bottom up, starting with the individual data points essentially being their own cluster. An iterative process is then carried out that combines the closest clusters, based on a pre-defined similarity criterion, into a single cluster until all data are contained within a singular cluster [44].

Ward's minimum variance method [45] is one approach for implementing agglomerative hierarchical clustering that combines data into new clusters to minimise the total within-cluster variation. For this thesis, the metric used to determine the similarity between clusters is taken to be the squared Euclidean distance.

#### *2.3.1.2.Generative Topographic Mapping (GTM)*

Generative topographic mapping (GTM) [46,47] is an ML algorithm designed for clustering, data stratification and visualisation, which has sound foundations in probability theory and provides a principled alternative to another popular methodology, the Self-Organising Map (SOM) algorithm [48]. Rather than predicting whether two data points should be allocated to the same cluster, the GTM predicts the probability of belonging to the same cluster. The GTM performs a soft assignment of data to clusters. This is a robust approach that considerably reduces the risk of countries being assigned to the wrong clusters.

The GTM assumes that the observed data is generated through a nonlinear and topology-preserving mapping from a low-dimensional latent space in $\mathfrak{R}^{\mathfrak{L}}$ onto a manifold embedded in the

high-dimensional space, $\mathfrak{R}^{\mathfrak{D}}$, where the observed data resides. The function used to generate this embedding takes the form:

$$\mathbf{y} = \mathbf{W}\Phi(\mathbf{u}) \tag{1}$$

where $\mathbf{u}$ is a point in the L-dimensional latent space, $\mathbf{W}$ is a matrix containing parameters that govern the mapping, and $\Phi$ consists of S basis functions $\Phi_S$, which for the standard GTM are radially symmetric Gaussians. If a prior probability distribution of $p(u)$ is defined for the latent space, then the distribution of data $\mathbf{x}$, for a given $\mathbf{u}$ and $\mathbf{W}$, is chosen to be a radially symmetric Gaussian centred on $\mathbf{y} = \mathbf{W}\Phi(\mathbf{u})$ having a variance of $\beta^{-1}$ so that:

$$\boldsymbol{p}(\mathbf{x}|\mathbf{u}, \mathbf{W}, \boldsymbol{\beta}) = \left(\frac{\beta}{2\pi}\right)^{\frac{\boldsymbol{D}}{\boldsymbol{2}}} exp\left\{-\frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^{\mathbf{2}}\right\} \tag{2}$$

where $\mathbf{y}$ is as defined in (1). The GTM latent space is constrained to form a uniform discrete grid of M centres, analogous to the distribution of SOM units, in the form:

$$\boldsymbol{p}(u) = \frac{\mathbf{1}}{\boldsymbol{M}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{M}} \delta(u - u_i) \tag{3}$$

Each of these centres is responsible for generating a spherical Gaussian density function in the D-dimensional data space. In this sense, the GTM can be understood as a special case of a Gaussian mixture model in which each component in the mixture defines the probability of an observable data point given a latent centre. Therefore, assuming the observed data points $x_n$ are independent and identically distributed (i.i.d.), the parameter matrix $\mathbf{W}$ and the inverse variance $\beta$ can be determined by maximising the log-likelihood given by:

$$\mathbf{L}(\mathbf{W}, \beta|\mathbf{X}) = \sum_{\boldsymbol{n}=\mathbf{1}}^{\boldsymbol{N}} \ln p(x_n|\mathbf{W}, \boldsymbol{\beta}) = \sum_{\boldsymbol{n}=\mathbf{1}}^{\boldsymbol{N}} \ln\left\{\frac{\mathbf{1}}{\boldsymbol{M}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{M}} p(x_n|u_i, \mathbf{W}, \beta)\right\} \tag{4}$$

where

$$p(x_n|u_i, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{\boldsymbol{D}}{\boldsymbol{2}}} exp\left\{-\frac{\beta}{2}\|y_{\mathbf{i}} - x_i\|^{\mathbf{2}}\right\} \tag{5}$$

In equation (5), $y_i$ is defined using equation (1) and is a D-dimensional point the manifold embedded in the data space for the point $u_i$ in the latent space. The adaptive parameters of the model are optimised using the expectation-maximisation (EM) algorithm. Matrix $\mathbf{W}$ is updated as the solution to the following system of equations:

$$\Phi^T G_{old}\, \Phi W_{new}^T - \Phi^T R_{old} X = 0 \tag{6}$$

where Φ is a $M \times S$ matrix with elements $\phi_S(u_i)$; $X$ is the observed data matrix $N \times D$ matrix with elements $x_{nm}$; **R** is the matrix of responsibilities that define the probability of the data point $x_n$ being generated by the latent point $u_i$ defined as $R_{in} = p(u_i|x_n, W_{old}, \beta_{old})$; and $G$ is a diagonal matrix with elements $\sum_{n=1} NR_{in}$. Finally, the β parameter is updated according to the following:

$$(\beta^{\text{new}})^{-1} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{i=1}^{M} R_{in}\|y_i - x_n\|^2 \quad (7)$$

Note that the observed data $X$ requires to be normalised before training (e.g. by centring the data around zero and scaling the data so that the new standard deviation becomes 1). For further details on the calculations, please refer to the original publication [46].

The GTM can not only assign data points to clusters but also can visualise them in a cluster membership map by projecting the latent centres. The GTM latent space can serve for visualisation purposes if its number of dimensions is 1 or 2, to which the mode probability (i.e. the highest cluster probability) is used to decide a data's cluster membership.

For the trained GTM, each cluster centre $\boldsymbol{y}_i$, henceforth named as a reference vector, is a prototype of the data. Reference maps associated with each of the variables were generated based on the reference vector components. These reference maps can be visualised in the form of heatmaps, where the high and low values can be used to interpret the relationship between each variable and each data cluster. This can provide further information/interpretation about the role that each variable used in the model had in defining each cluster.

*2.3.1.3. GTM Magnification Factors*

As GTM maps nodes that lie on a uniform discrete grid in a lower dimensional latent space into the higher dimensional data space, regions in the latent space may experience distortions when the mapping is optimised to fit the data. Due to the pre-defined uniformity of the latent space, the visualisation of the membership map may not exhibit the natural separations present within the data space. This problem has been addressed by the creators of GTM by leveraging the concept of magnification factors [49]. These magnification factors are evaluated in terms of the mapping defined in equation (1) via differential geometry. The full proof and derivation of magnification factors can be found in the original publication [49], however, key equations will be outlined here to provide an appropriate description for the reader.

For a latent space with 2 dimensions, GTM maps an infinitesimal rectangle with area $dA = \prod_i dx^i$ from the latent space to another infinitesimal rectangle that resides in the data space, defined by equation (1), with area $dA'$ [50]. The magnification factors are therefore calculated as the determinant of the Jacobian of this transformation [51], and is expressed in matrix form as:

$$\frac{dA'}{dA} = det^{\frac{1}{2}}(\boldsymbol{\Psi^T W^T W \Psi}) \tag{8}$$

Where $\boldsymbol{\Psi}$ is a matrix, whose elements are comprised of the partial derivatives of the radial basis functions $\Phi$ with respect to the grid centres in the latent space, and $\boldsymbol{W}$ is the matrix of parameters that governs the GTM mapping. Once calculated, the magnification factors can then be superimposed onto the membership map visualisation to generate a magnification factor plot. By using a grey-scale representation for such plots, it provides a visual representation to the amount degree of distortion occurring during the GTM mapping at different areas of the latent space, with the extreme shades indicating large or small distortions [50,51].

### 2.4. Performance Evaluation (Metrics and Validation Methods)

#### 2.4.1. Area Under the Receiver Operating Characteristic Curve (AUC)

Being able to properly assess the performance of a ML model is crucial to determine the usefulness of its output. Model performance is measured using a metric, with different types of methods requiring different performance metrics. For example, regression style problems may use the mean squared error (MSE) to measure how accurate a prediction was to the actual value with the best model producing the lowest MSE. Classification style problems can use measures such as accuracy and F1 score. This thesis however employs the commonly used area under the receiver operating characteristic curve (AUC) as the metric to evaluate the performance of classification models. AUC values range from 0 to 1 where a model that can perfectly separate the classes in the data would have an AUC equal to 1 [52].

#### 2.4.2. K-fold Cross-validation

Cross-validation is a method of validating the performance of an ML model without an external validation dataset. This type of analysis is crucial as it provides an estimation of how the model will generalise to new data. For K-fold cross-validation, the training dataset is first split into K equal groups (usually performed whereby K = 10). A model is then trained on K-1 groups, with the remaining group left out to be used for testing. This process is then iterated K times such that each of the K groups is used as a testing set. During each iteration, the performance metric used to evaluate the model, which in the context of this thesis would be AUC, is recorded such

that after the K iterations, there is a set of K performance metrics. These metrics are then averaged to provide an overall model performance [53].

### 2.5. Statistical Tests

Statistical tests have been used in this thesis to provide an appropriate comparison between summary data for various variables. The idea behind using statistical tests is to determine whether two (or more) variables are independent from each other, or put another way, whether they are statistically different from one another. Several factors need to be accounted for when deciding which statistical tests to apply. Some of these factors are the type of data being considered, i.e. if the data is continuous or categorical, and the underlying distribution of the data [54].

#### 2.5.1. Chi-Squared Test

Pearson's Chi-Squared Test ($\chi^2$) is a nonparametric test used to determine the independence of two or more categorical variables. $\chi^2$ is calculated using equation (7) and (8):

$$\sum \chi^2 = \frac{(O-E)^2}{E} \tag{9}$$

$$E = \frac{M_r M_c}{n} \tag{10}$$

Where $\chi^2$ represents the Chi-Squared statistic, $O$ represents the observed counts in the data, $E$ represents the expected value calculated by multiplying the row and column marginals, $M_r$ and $M_c$ respectively, and dividing by the total data size $n$ [55].

#### 2.5.2. Kruskal-Wallis test

The Kruskal-Wallis test is also a non-parametric test that serves as an alternative to one-way analysis of variance, and is a more generalised form of the Mann-Whitney (Wilcoxon rank-sum test) that can be applied to two or more independent samples [56]. The Kruskal-Wallis test statistic is calculated using equation (9):

$$T = \left(\frac{12}{N(N+1)}\right) + \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{11}$$

Where $T$ is the test statistic, $N$ is the total dataset size, $n_i$ represents the size of sample $i$, and $R_i$ represents the sum of ranks assigned to sample $i$.

# 3. Chapter 3: ECG Data Format Comparison

## 3.1.Introduction

The ECG is a simple, non-invasive test used globally to detect numerous heart issues. ECGs measure the electrical activity of the heart using electrodes, known as leads, attached to different parts of the body. Older ECG machines directly record the electrical signals from each lead onto graph paper, which are then stored as physical copies and then manually scanned so they can be viewed electronically. With newer machines, signals can be directly recorded and stored electronically as portable document format (PDF) files [57]. In some instances, the electrical signals from each lead are recorded and stored digitally (as a signal, not as a PDF)[58], yet this is rare as the machines that provide the raw digital signals are more expensive and usually research-based.

There is an ever-increasing amount of ML research being completed whereby ECG data is used to develop models to address a variety of heart conditions [29,59–61]. Models have been developed using both digital ECG signals [62] and ECGs in an image format [63]. Image ECGs are records that were either physically recorded and scanned or were recorded electronically and stored in a PDF format. While it has been reported that analysing ECGs in a digital signal format is preferable [64], it is often the case that the choice of format is dictated by the data available. Several studies [64–68] focus on solving this problem, providing methods of digitising image ECGs by extracting the signals from the image and storing them as a multivariate time series. These extracted signals show promise for ML model development [64] but have not yet seen widespread adoption. Nevertheless, and to the best of our knowledge, there has been no research thus far that confirms a tangible benefit to developing ML models using one ECG data format over another.

To that end, we collated a large dataset of ECGs represented in three different data formats: original digital ECG signal recordings (Signal ECGs); the ECGs in an image format (Image ECGs); and ECG signals extracted by digitising the Image ECGs (Extracted Signal ECGs). The main objective of this chapter therefore is to quantify the effect ECG data format choice has on ML model performance in the context of MI prediction, thereby identifying the optimal format. In addressing the main objective, we also address an auxiliary objective by validating the feasibility of using Extracted Signal ECGs for ML cardiac outcome modelling.

## 3.2.Methods

### 3.2.1. Data source

We selected the PTB-XL database [69] for use in this analysis for several reasons. First, it is, to date, the largest open-source ECG dataset available, hosted by PhysioNet [70]. The dataset consists of the digital signals for 21,837 ECG records from 18,885 patients, with most records being assigned at least one of five main diagnoses (or "superclasses"): Normal, Myocardial Infarction, ST/T wave Change, Conduction Disturbance, and Hypertrophy. Each diagnostic superclass was assigned based on the written notes in the original ECG report. Each class received a likelihood score between 0 and 100, which represented the cardiologist's certainty of the diagnosis. The signal data provided within the PTB-XL database represent a 10 second ECG recording sampled at two frequencies, 100Hz and 500Hz. For this analysis, the data sampled at 100Hz was used as it falls within the common frequency range used by modern ECG machines [71].

### 3.2.2. Data extraction

Since 5 diagnostic superclasses are represented in these ECGs, this dataset has led to diverse study designs [72,73]. For this analysis, we designed a two-class classification task using only the two largest classes within the dataset: normal ECG (NORM) and MI. In this way, we could limit any source of variability that would interfere with evaluating the impact the data format has on model performance.

Another consideration was given to the diagnosis likelihoods. To ensure a thorough evaluation two subsets of the data were created: The 1st subset, referred to henceforth as the "conservative cohort", only included ECGs where the likelihood score equalled 100; the 2nd subset, referred to henceforth as the "speculative cohort", included all ECGs regardless of the likelihood score. Developing models on both data subsets allowed us to add a controlled amount of variability into our testing to provide a richer understanding of the optimal data format.

To provide transparency within our analysis, and allow for straightforward external validation and comparable inter-model results, we followed the suggested data splitting as defined in the original PTB-XL study [69], which recommends using ECGs assigned to folds 1-8 for model training, fold 9 for validation and fold 10 for testing.

### 3.2.3. Signal ECG data preparation

As previously mentioned, each Signal ECG recording within the PTB-XL database contains 12 signals, ten seconds in length, with each signal representing one of the 12 standard sets of leads used in ECG recordings (I, II, III, aVL, aVF, aVR, V1, V2, V3, V4, V5, V6). With the data being

sampled at 100Hz, each ten-second recording consists of 1,000 samples, giving a data dimension for each Signal ECG sample of 12x1000.

### 3.2.4. Image ECG data preparation

. The Image ECG data was generated manually using the Signal ECG recordings by leveraging the “wfdb” and “ecg-plot” Python packages. We formed the Image ECGs so they would resemble genuine, commonly found ECG recordings hence providing a realistic understanding of the performance ML models can achieve if deployed in a real-world application. To that end, we generated two sets of Image ECGs, with each set having different lead arrangements and displaying a different amount of the original signal. The first set of Image ECGs (arrangement A) arranged the 12 leads in a single column, with the full 10 seconds of data used for each lead (as shown in Figure 5b). One Image ECG was created for each set of the 12 lead ECG signals with dimensions 1200x1000. The second set (arrangement B) had the 12 leads arranged in a 3x4 grid, with 2.5 seconds of the full 10 seconds available used for each lead (see Figure 5c). The 2.5 seconds used for each lead was also staggered based on the column in which the lead was present such that:

- 0s - 2.5s used for leads I, II and III
- 2.5s - 5s used for leads aVR, aVL and aVF
- 5s - 7.5s used for leads V1, V2 and V3
- 7.5s - 10s used for leads V4, V5 and V6

One image ECG was created from each set of the 12 lead ECG signals with dimensions 300x1000. However, to ensure computational tractability for the proposed experiments, the images for both formats were reduced prior to model development. The Image ECGs were therefore analysed using a dimension of 165x500 and 330x275 for arrangement A and arrangement B respectively.

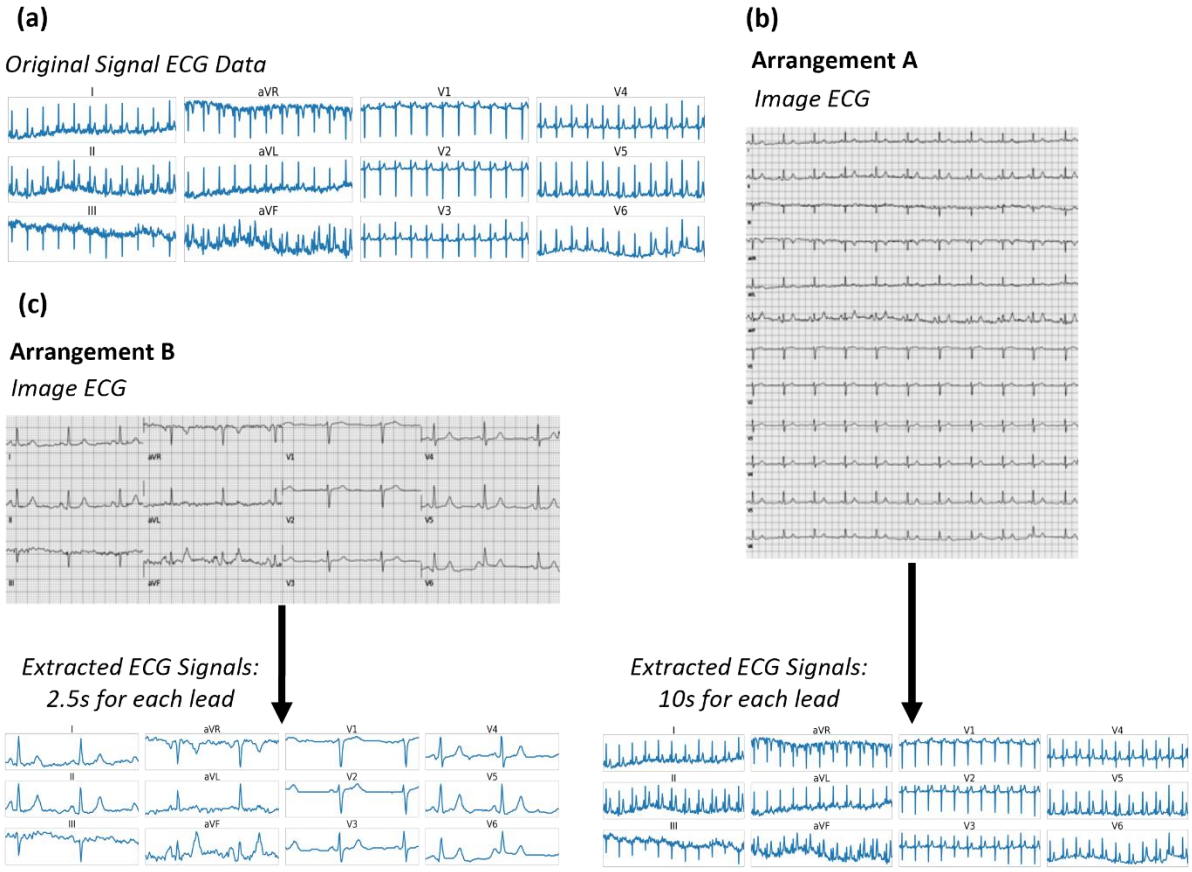



*Figure 5. Displays the same ECG in each of the three different data formats being evaluated, for both Image ECG arrangements. (a) Signal ECG data format; (b) Image ECG data and Extracted ECG Signal data format for arrangement A; (c) Image ECG data and Extracted ECG Signal data format for arrangement B.*

### 3.2.5. Extracted Signal ECG data preparation

We followed Fortune et al [66] ECG digitisation algorithm to extract the signals from the Image ECGs. They created an open-source application that allows a user to import an Image ECG, manually draw borders around each lead, then extract the ECG signal contained within each border and export the signals to a CSV file. The manual nature of this application meant it was not feasible for use in our analysis due to the volume of ECGs, as it would take too long and be prone to potential human errors. To overcome this, we extended their approach and implemented a semi-automatic signal extraction algorithm. Although our approach still requires the border positions to be manually set, this is performed only once as the Image ECGs are identical in layout and dimensionality. In addition, we added functionality that removed lead labels from the images, as they interfered with the extraction algorithm. The 10 second signal contained within the arrangement A Image ECGs were then extracted into a 12x1000 array, with the 2.5 second signal

contained within the arrangement B Image ECGs being extracted into a 12x250 array, with both then being exported as a CSV file. Figure 6 provides a flow chart of the extraction process.

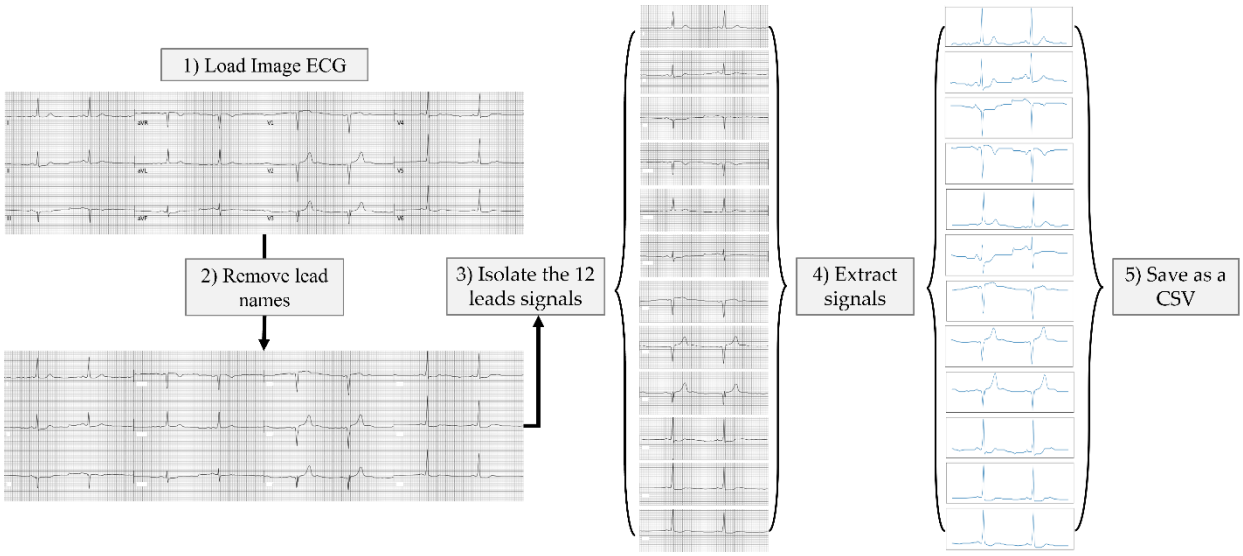



*Figure 6. Detailing the process used to extract the ECG signals from the image ECGs.*

### 3.2.6. ML modelling techniques

CNNs were used for this analysis due to their ability to be applied to contextual datasets of varying forms. Specifically, we utilised different structures to allow for both the 2-D image and 1-D signal inputs. Other ML methodologies (or even more complex versions of the selected methodology) could have been better suited for analysing the different formats being compared. However, the use of different methodologies would introduce a source of variation to the experiments, which would have detracted away from the direct comparison of the data formats, which was the aim of this chapter. Hyperparameter tuning was used to develop the models that will be applied to the different data formats. For completeness, each data format passed through three rounds of hyperparameter tuning, with each round using a different hyperparameter search space method: random search [74]; hyperband [75] and Bayesian optimisation [76]. Random search [74] works by selecting random combinations of hyperparameters from a pre-defined parameter space to train the model. This process is repeated for a set number of iterations, with each model being evaluated to find the combination that generates the best model. The hyperband [75] algorithm combines the ideas of random search and another algorithm known as Successive Halving [77]. It involves building and training multiple models with random hyperparameters and through principled early-stopping, poor-performing models are quickly identified and discarded. The remaining models are then trained longer, repeating the process of early stopping and discarding, until the single best-performing model remains. In contrast to random search and hyperband, Bayesian optimisation [76] works by using Bayesian inference to build a probabilistic

model of the objective function that is used to guide the hyperparameter search. After each model iteration, the Bayesian model is then updated based on the model performance and then subsequently used to choose the next set of hyperparameters.

Two hyperparameter search spaces were defined: one to develop 2-D CNN models to be applied to the Image ECGs, and one to develop 1-D CNN models to be applied to both the Signal ECG and Extracted Signal ECG data. For the 2-D CNNs applied to the Image ECGs, a "convolutional block" was defined that consisted of a 2-D convolutional layer, a ReLU activation layer, a batch normalisation layer, and a max pooling layer with a 2x2 pool size. Hyperparameter tuning was used here to select the number of filters used within each convolutional layer as well as the total number of convolutional blocks contained within the architecture up to a maximum of 6 (note: the filters selected for each convolutional layer in each convolutional block was done so individually). Then a flatten later was applied and hyperparameter tuning was used to select the number of hidden dense layers and their associated nodes before being passed to a dropout layer, where the dropout rate was also a tuned parameter, and finally to a sigmoid output classification layer.

Like with the 2-D CNN, for the 1-D CNNs used on the Signal ECG and Extracted Signal ECG data, a "convolutional block" was defined with changes to make it appropriate for the data type. The 1-D convolutional block consisted of a 1-D convolutional layer, ReLU activation layer and a batch normalisation layer. Hyperparameter tuning was used here to the number of filters and the kernel size within each convolutional layer as well as the total number of convolutional blocks up to a maximum of 3. Then a global average pooling layer was used before being passed to hidden dense layers, the number and size of which again selected through hyperparameter tuning like in the 2-D case, before finally being passed to a dropout layer, where again the dropout rate was selected through tuning, and a final sigmoid classification layer. In both the 2-D and 1-D cases, the learning rate was also a parameter that was tuned using the hyperparameter tuning method.

After we identified the best model for each data format, to help interpret the models generated we utilised the HiResCAM to visualise what areas of the inputs were considered most important by the CNN models when making their predictions. All models were evaluated using the results from the testing data split, with the performance metric used throughout being AUC.

### 3.3. Results

#### 3.3.1. Dataset generation

Applying the criterion set out in the *Data Extraction* section to the 21,837 ECGs, we were left with a total of 11,621 eligible ECGs: 9,083 of which were NORM and 2,538 were MI (21.7%

prevalence). The conservative and speculative cohort subsets were then created from the eligible ECGs. The conservative cohort contained a total of 8,358 ECGs: 7,017 of which were NORM and 1,341 are MI (16% prevalence), whilst the speculative cohort contained the full 11,621 ECGs. The ECGs were then grouped according to the fold they were assigned in the original PTB-XL dataset to generate the training, validation, and testing data splits. The data pre-processing steps outlined for the Signal ECGs, Image ECG and Extracted Signal ECGs data formats were then applied to generate the final datasets that would be used for the ML model developments. The full process is outlined in Figure 7.

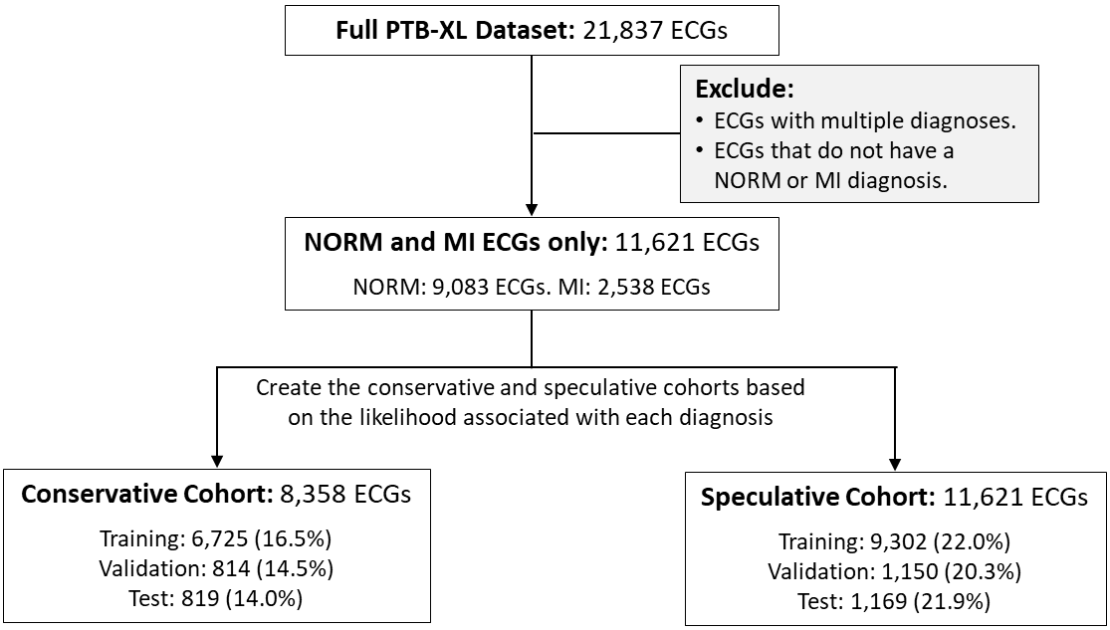



*Figure 7. Flowchart showing how the criteria was applied to the full PTB-XL dataset to generate both data cohorts. Values in brackets indicate the prevalence.*

### 3.3.2. Model Comparisons

Following the aforementioned framework for model development, we trained and tested models for both arrangement A and B data. The results of the best-performing models for each data format within each cohort are listed in Table 1 and Table 2 for arrangement A and B data respectively. Each of the best-performing model architectures are displayed in Figures S1 to S12 in the supplementary material. Starting with arrangement A, the Signal ECG and Extracted Signal ECG formats performed the best, with both also significantly outperforming the Image ECG format for both the conservative and speculative cohort tests. Additionally, the Signal ECG and Extracted ECG signal formats did not perform significantly differently from each other. Moving to the arrangement B data, here the Image ECG format performed the best, significantly outperforming the Signal and Extracted Signal ECG formats, in both the conservative and speculative cohort tests. Like with the arrangement A data, however, the Signal ECG and

Extracted ECG signals did not perform significantly differently from one another. Across both tests with the arrangement A and B data, we see a drop in performance across all the data formats when comparing the conservative cohort and speculative cohort test results.

*Table 1. Displays the modelling results using the arrangement A data. The AUCs of the best models trained using each ECG data format for both the conservative and speculative cohort are presented*

| **Arrangement A ECGs** | | | | | | |
|---|---|---|---|---|---|---|
| | **Conservative Cohort (AUC [95% CI])** | | | **Speculative Cohort (AUC [95% CI])** | | |
| Data Format | Signal ECG Data | Image ECG Data | Extracted Signal ECG Data | Signal ECG Data | Image ECG Data | Extracted Signal ECG Data |
| Training | 0.999 [0.998, 0.999] | 0.998 [0.998, 0.999] | 0.995 [0.993, 0.996] | 0.949 [0.945, 0.953] | 0.918 [0.913, 0.924] | 0.954 [0.95, 0.958] |
| Validation | 0.962 [0.951, 0.974] | 0.944 [0.929, 0.959] | 0.97 [0.96, 0.981] | 0.921 [0.906, 0.937] | 0.893 [0.874, 0.911] | 0.911 [0.895, 0.928] |
| Testing | 0.971 [0.961, 0.981] | 0.952 [0.938, 0.966] | 0.974 [0.965, 0.984] | 0.931 [0.918, 0.945] | 0.89 [0.871, 0.908] | 0.919 [0.903, 0.934] |

*Table 2. Displays the modelling results using the arrangement B data. The AUCs of the best models trained using each ECG data format for both the conservative and speculative cohorts are presented. ** The Signal ECG data used matches the same 2.5 seconds of signal used for the Extracted Signal ECG data to ensure a relevant comparison*

| **Arrangement B ECGs** | | | | | | |
|---|---|---|---|---|---|---|
| | **Conservative Cohort (AUC [95% CI])** | | | **Speculative Cohort (AUC [95% CI])** | | |
| Data Format | Signal ECG Data** | Image ECG Data | Extracted Signal ECG Data | Signal ECG Data | Image ECG Data | Extracted Signal ECG Data |
| Training | 0.985 [0.983, 0.988] | 0.978 [0.975, 0.981] | 0.979 [0.976, 0.982] | 0.966 [0.963, 0.969] | 0.963 [0.96, 0.967] | 0.951 [0.947, 0.955] |
| Validation | 0.946 [0.931, 0.961] | 0.933 [0.916, 0.951] | 0.949 [0.933, 0.963] | 0.907 [0.89, 0.924] | 0.900 [0.882, 0.918] | 0.903 [0.886, 0.910] |
| Testing | 0.938 [0.921, 0.954] | 0.960 [0.948, 0.973] | 0.937 [0.921, 0.953] | 0.886 [0.867, 0.905] | 0.903 [0.886, 0.92] | 0.864 0.843, 0.884] |

### 3.3.3. Class activation maps

We applied HiResCAM to the outputs of the best-performing models for each data format. This provided a visual heat map that we overlaid onto the inputted data to analyse the areas important to the decision-making of the model. Figures 8 and 9 display the same ECG of a patient with MI, represented in the three different formats for both data arrangements, with their respective activation maps superimposed on top, for the conservative cohort data. Figures 10 and 11 display the same ECG of a patient deemed normal, represented in the three different for both data arrangements, with the activation maps superimposed on top, this time for the speculative cohort data. For Figures 8, 9, 10, and 11, the red sections of the activation map represent regions of the input the model deemed most important, with the blue sections representing areas deemed less relevant to determining the outcome.

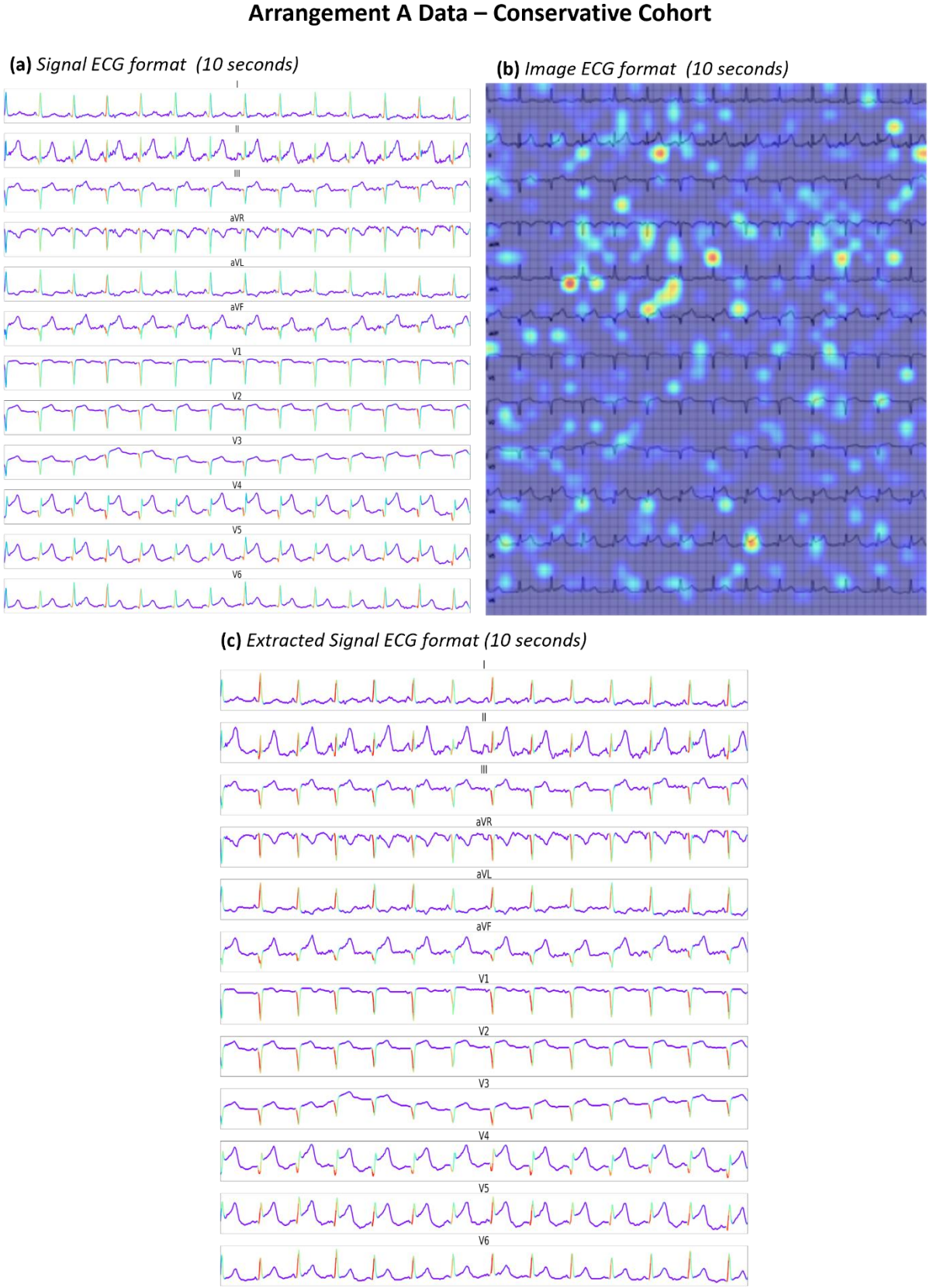



*Figure 8. Displays the HiResCAM activation maps generated using the three best models for each of the data formats in the conservative cohort for arrangement A data. (a) Signal ECG data format; (b) Image ECG data format; (c) Extracted ECG data format.*

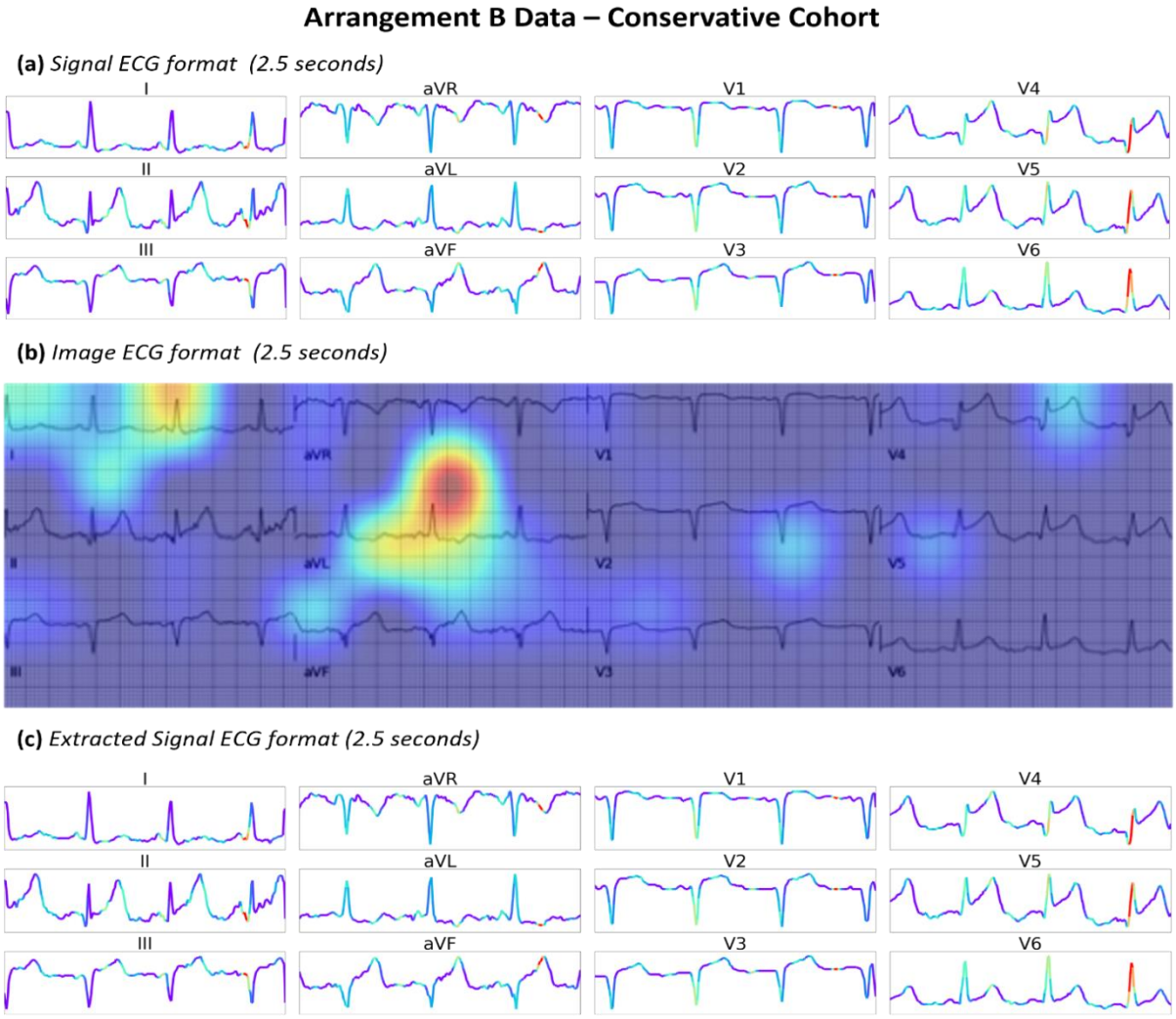



*Figure 9. Displays the HiResCAM activation maps generated using the three best models for each of the data formats in the conservative cohort for arrangement B data. (a) Signal ECG data format; (b) Image ECG data format; (c) Extracted ECG data format.*

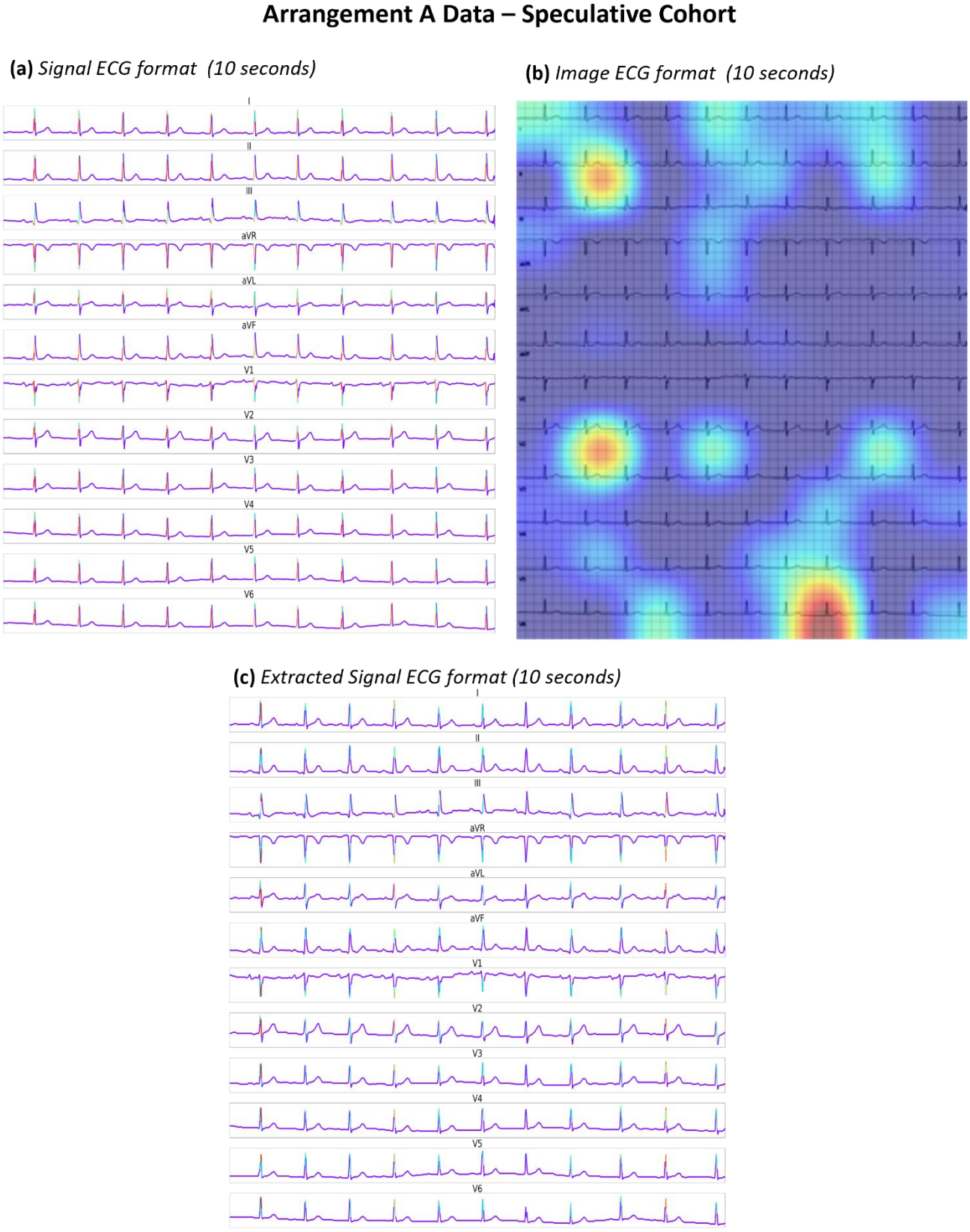



*Figure 10. Displays the HiResCAM activation maps generated using the three best models for each of the data formats in the speculative cohort for arrangement A data. (a) Signal ECG data format; (b) Image ECG data format; (c) Extracted ECG data format.*


*Figure 11. Displays the HiResCAM activation maps generated using the three best models for each of the data formats in the speculative cohort for arrangement B data. (a) Signal ECG data format; (b) Image ECG data format; (c) Extracted ECG data format.*

### 3.4. Discussion

The results highlight the very real presence of ML performance differences between the three different data formats, as well as how the data is represented within each format. As expected, Signal ECG should be the preferred choice for ML modelling, provided that such a format is available. If this is not the case, the decision would depend on the particular needs. Starting with arrangement A (10-second ECGs), the Extracted Signal ECG format seems to offer better performance results when using either the conservative or speculative cohort tests. Remarkably, the performance results of the Extracted Signal ECG were comparable to those of the Signal ECG format. This provides key quantifiable evidence that Extracted Signal ECGs are not only feasible

for ML modelling, but in some situations (such as with arrangement A data), it could be the preferred choice.

However, a drop in performance was observed when arrangement B (2.5-second ECGs) was used either with conservative or speculative cohort tests. The performance drop was particularly significant when the Extracted Signal ECG format was used, which was outperformed by Image ECG models when this arrangement was used. Interestingly, a similar drop in performance was observed when the original Signal ECG data was modelled using the same 2.5 seconds. This suggests that the drop in performance seen by the Extracted Signal ECG format results from the shorter ECG duration, and not because of an inherent issue with the ECG digitisation.

Overall, models developed using the conservative cohort subset performed better than the models developed using the speculative cohort subset. This is an expected result; the added uncertainty brought about by using noisier data in the speculative cohort was naturally harder to model than in the conservative cohort. Therefore, based on model performance alone, the Extracted Signal ECG format would be the preferred choice, should the Image ECG data contain 10 seconds of data per lead. Should the Image ECG contain less data per lead, then this becomes the preferred format. However, there may arise conditions brought on by external factors whereby choosing a format with slightly lower performance could yield more meaningful results.

One such example would be the interpretability of model output using techniques such as HiResCAM activation maps. Using the activation maps described in Figures 8 to 11 initially, the Image ECG maps highlight general regions of the signal that the model found important, making it difficult to precisely ascertain the key information. The maps for both the Signal ECG and Extracted Signal ECG data are much clearer, providing specific time points of interest on each digital signal that their respective models deemed important. For example, in Figure 8 the regions that have been deemed important relate primarily to the onset/upslope of the QRS signal alongside the known impact on the ST segment. The early part of the QRS signal is not routinely evaluated using conventional interpretation algorithms for MI. This also demonstrates the unique ability of the technique to identify new and novel patterns. Using a further example, Figures 12 and 13 contain the activation maps for a participant where the correct prediction was made for every data format to demonstrate the difference in interpretability. All three maps show that similar areas of the signal are considered by both the Signal ECG, Extracted Signal ECG, and Image ECG data formats. However, the maps associated with the Image ECGs show that the model has considered areas in between two signals as very important, implying that the model learning is far less intuitive. In contrast, the maps for the Signal ECGs and Extracted Signal ECGs show the models focus primarily on the peaks, and more specifically, what sections of the peaks were more

important than others. This allows the user to understand clearly what led to the prediction and if the correct point of the signal is being considered.

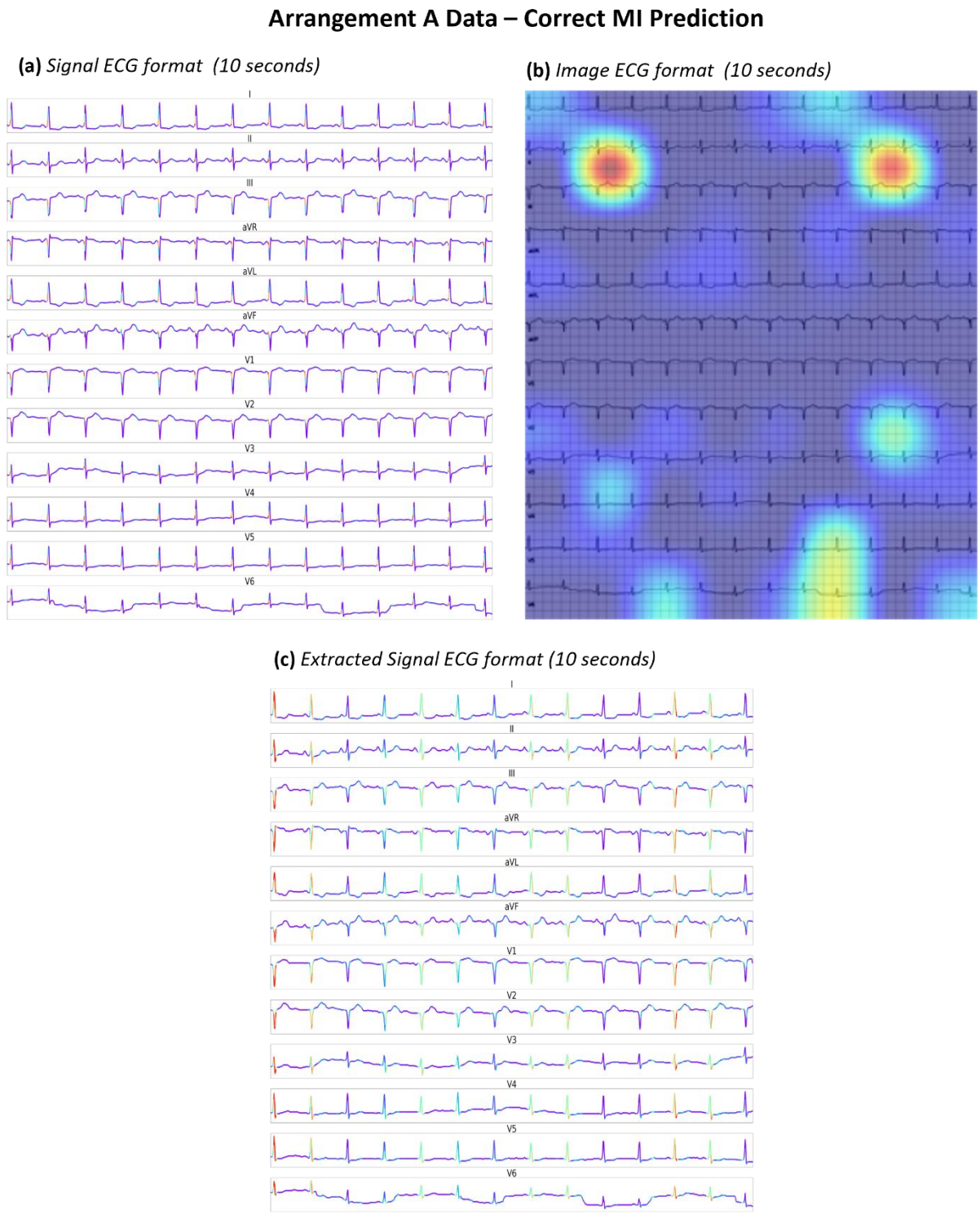



*Figure 12. Displays the HiResCAM activation maps generated for an ECG that represents MI, whereby the best models for each format all correctly predicted MI. (a) Signal ECG data format; (b) Image ECG data format; (c) Extracted ECG data format.*


*Figure 13. Displays the HiResCAM activation maps generated for an ECG that represents MI, whereby the best models for each format all correctly predicted MI. (a) Signal ECG data format; (b) Image ECG data format; (c) Extracted ECG data format.*

Further examples of situations that would favour the Extracted ECG Signals would be if key features from the ECG, such as QRS duration and P wave duration, need to be extracted. Extracting these features from the images is difficult, however, previous studies have seen success by first digitising the image and then extracting the features [78]. There are also open-source Python packages such as "neurokit2" [79] that can automatically detect key points on a digital ECG signal. Also, as briefly mentioned in the results section, the dimension of the Image ECGs had to be manually reduced to allow for ML modelling to be carried out. This highlights an inherent benefit of the Extracted Signal ECGs again over the Image ECGs in that they are computationally more efficient to process and model when compared to the images when working with large datasets, as was the case in this chapter.

One instance which would favour the Image ECG format over the Extracted Signal ECG format would be in the event whereby the recorded ECG was noisy. The ECG digitisation algorithm used in this chapter works by removing background noise within the image, isolating the ECG recording within a desired window which will be converted to a digital signal. For this analysis, as the Image ECGs were generated manually using the Signal ECG data, we intentionally cultivated a perfect scenario whereby we had fully clean Image ECGs. An Image ECG could be considered noisy if there is a significant overlap between the recordings of different leads, or a on the image such as a coffee stain (should the Image ECG be a scanned version of a physical copy). The presence of these could lead to the digitisation algorithm failing to extract the signal and therefore excluding that ECG from any further analysis. This favours the Image ECG format as it allows for the ECG to be analysed regardless of the state of the original image, reducing the chance data is removed.

The application of AI and ML in the utilisation of the 12-lead ECG has evolved in tandem with technological developments. Recent studies have demonstrated the role of AI on the ECG being able to predict disease that is not achievable through routine individual scrutiny [80] and hence there is a significant immediate and long-term clinical impact. The 12-lead ECG is the fundamental and primary cardiac investigation for patients presenting with symptoms and hence the ability of AI / ML technology to provide insight into structural and functional cardiac adaptation will improve patient diagnosis, management and reduce downstream costs secondary to a reduction in unnecessary investigations. It is apparent from our analysis that to build up large datasets with sufficient accuracy the signal format is important and should be considered when developing ML studies going forward. That aside, as hospital environments continue a transition to a full digital set-up the likelihood of securing widespread Signal ECGs is unlikely. Our data highlights the importance of digitally storing PDFs and refining methodology to better handle these image files and subsequently allowing more robust predictive models.

### 3.5. Chapter conclusion

The analysis conducted in this chapter provides an evaluation of three different data formats that can feasibly be used to analyse ECGs. Signal ECGs, Image ECGs and Extracted Signal ECGs were all compared using two different ECG arrangements and two data subsets: the first contained best-case scenario data with a clear separation between the classes; the second had more noise and less confident diagnoses. The results of the analysis showed that should the Signal ECG data be available, then this should always be used for any ML modelling. In the absence of data in this format, we showed that the optimal data regarding model performance is dependent on the way the data is arranged within the ECG: If the Image ECG contains 10 seconds of data for each lead,

digitising the signal and using the Extracted Signal ECGs is optimal; if the Image ECG contains 2.5 seconds of data per lead, then using the Image ECG data is optimal for ML performance. As highlighted in the discussion, the decision may become situational with certain criteria, such as noisy Image ECGs, meaning one is more effective than the other. What these results also speak to is the viability of extracting digital ECG signals from image ECGs and using those for ML model development. However, further analyses will be needed to investigate how factors such as changes in image resolution and in extraction algorithms influence model performance.

# 4. Chapter 4: GTM Methodology and Workflow Development – Mapping the global free expression landscape using machine learning

The methodology outlined in this chapter plays a key role in the story of the thesis from a methodological standpoint, as it served as the blueprint upon which the analysis in later chapters was based. The reader may find it unusual that the methodology was developed on a non-clinical application, however its inclusion is justified in so far as the primary focus of this thesis is methodological development, with the described methodology being applied to address clinical problems within cardiovascular research. The research presented in this chapter was instrumental as it served as the main influencing factor in deciding the methodological development path taken for all subsequent unsupervised learning carried out as part of this PhD project. This means its inclusion is vital to ensure that this thesis demonstrates appropriately that the methodology is properly validated. The work presented in this chapter resulted in a publication [81], which serves as a demonstration of the versatility and generalisability of the proposed methodology and further speaks to the robustness of the approach.

## 4.1. Introduction

In an increasingly atomised, polarised world, the free expression of ideas is more important than ever. But while the need for free expression has increased, so have the forces which seek to suppress it and the technologies which enable its suppression. Freedom of expression is among the core human rights set out in the United Nations (UN) Universal Declaration of Human Rights [82], the International Covenant on Civil and Political Rights (1966; entered into force in 1976) and subsequent treaties, including those in Europe, the Americas and Africa, for example, the European Convention on Human Rights [83], entry into force in 1953; the American Convention on Human Rights [84], entry into force in 1978; and the African Charter on Human and Peoples' Rights [85]; entry into force in 1986.

These global mechanisms uphold the principle that "everyone has the right to freedom of expression." Unfortunately, in today's world, this right is facing numerous challenges. Rapid advancements in technology have provided new avenues for those who wish to suppress freedom of expression. Censorship and surveillance tools are becoming more sophisticated and readily available, enabling governments and other entities to monitor and control the flow of information.

Censorship continues to operate across the globe, using several diverse tactics and drivers, including state laws or practices that restrict expression beyond what is included in international

instruments [86]. Examples of this include the mixture of technological and legislative mechanisms deployed by the Chinese state to block access to online resources (colloquially called the Great Firewall of China - see for example [87]), the reduction of civil space for protests and other acts of civic participation, and the use of strategic lawsuits against public participation (SLAPPs)[1] [88] to prevent journalists and other public watchdogs from being able to report in the public interest.

With the entry into force of such standards, the UN and regional inter-governmental organisations established bodies or mechanisms to assess state adherence to the standards. This required techniques of assessment and measurement which had been developed by scholars starting in the 1930s with Greer's study into the Reign of Terror in revolutionary France [89] and which have become increasingly sophisticated in terms of data sources and statistical techniques – see, for example, [90–92]. The purpose of measuring human rights is to assess the extent to which these rights are upheld in theory, manifested in reality, and advanced through effective policies [93]. By conducting such measurements, we aim to identify areas where human rights are being violated or neglected so that appropriate solutions to address these challenges can be developed.

This research introduces the Index Index, an innovative analysis of global censorship practices, and proposes a novel methodological approach to calculate it. Specifically, the Index Index focuses on academic, digital, and media/press freedom. It uses Generative Topographic Mapping (GTM, [46,47]), an unsupervised Machine Learning algorithm, to cluster and visualise countries in terms of their levels of freedom of expression. By utilizing established and robust indices and metrics, this research offers a comprehensive and nuanced assessment of the international landscape of free expression. It sheds light on the various threats that impede, curtail, suppress, or manipulate the public's right to access information, express themselves, and engage with others [2]. Unlike recent studies that solely rely on data related to internet accessibility, such as [94–96], the Index Index integrates a wide range of existing analyses and expertise to provide a comprehensive ranking of the free expression environment in all countries or nations where sufficient data is available.

[1] SLAPPs are vexatious lawsuits targeting journalists and other whistleblowers whereby powerful individuals and institutions use civil lawsuits to intimidate and financially threaten critics [7].

[2] For the purpose of measurement, the term 'country' refers to a state or political entity, including Kosovo, Palestine, and Taiwan, which are not recognized as states by the UN.

### 4.2. Methods

#### 4.2.1. Data and resources that informed the development of the Index Index

As this is an index of indices, the raw data comprises existing indices and metrics developed by a range of different national and international bodies such as research institutes, as well as international non-governmental organisations. Each pre-existing index has been selected based on several criteria, including its usage and reference by the wider community of practitioners, the robustness of its methodology, and its geographic scope. Individually, they are the product of internal testing and iterative development and as a result are used in a range of public advocacy and campaigning initiatives, including being referenced by international bodies, such as European institutions and UN bodies. For instance, Varieties of Democracy (V-Dem) is funded by, among others, the European Commission, the Swedish Ministry of Foreign Affairs and the World Bank [97]; the World Press Freedom Index is cited by the European Parliament in its Normandy Index 2023 [98]; and the Committee to Protect Journalists has submitted evidence to the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression [99].

We selected these indices on account of their robustness and completeness. The datasets were collated after in-depth conversations between the project team. Several other sources were explored and ultimately discounted. Overall, the Index Index models 178 variables, which are broken down into the following freedom categories:

- Academic - 14
- Digital - 50
- Media/Press - 114

Further details about the selected and discounted data sources are provided below, with the full list of variables used for modelling detailed in Table S1 in the Supplementary Materials.

*4.2.1.1. V-Dem (Varieties of Democracy)*

The Varieties of Democracy (V-Dem) Research Project [100] offers a nuanced and extensive analysis of democratisation, examining various dimensions and subcomponents. The data forming the foundation of V-Dem's component variables are collected through surveys administered to a network of over 3,500 Country Experts. The project aims to ensure a minimum of five experts for each indicator per country, facilitating a robust and diverse perspective. By employing a wide range of indicators and involving a substantial number of experts, V-Dem strives to provide a comprehensive understanding of democracy's complexities and variations across countries.

The V-Dem database offers a comprehensive range of democratic measures, surpassing the scope of the Index Index. Recognising this, the research team carefully extracted and isolated 171 variables from the extensive dataset that held significance for the model. These variables encompassed not only the three freedoms emphasised in the Index Index (academic, digital, and media/press freedom) but also encompassed broader contextual concerns, such as corruption and accountability measures, alongside various civil liberties.

*4.2.1.2. World Press Freedom Index*

The World Press Freedom Index, compiled by Reporters Without Borders (RSF), serves the purpose of comparing the level of press freedom across 180 countries and territories [101]. It provides a snapshot of the press freedom situation in these locations during the preceding calendar year prior to its publication. The Index utilises a scoring system ranging from 0 to 100 to rank each country or territory. This score is derived from two key components: a quantitative assessment of abuses against journalists and media outlets, and a qualitative analysis of the overall situation within each country or territory.

To obtain the qualitative analysis, RSF distributes a questionnaire in 23 languages to press freedom specialists, including journalists, researchers, academics, and human rights defenders. Following the calculation of scores, the countries and territories are arranged in an ordinal list from 1 to 180, with 1 indicating the highest level of press freedom. It is this raw score calculated for each country that we have utilised as a variable in our model's development.

*4.2.1.3. Committee to Protect Journalists (CPJ)*

The Committee to Protect Journalists (CPJ) collects comprehensive data [102] on the imprisonment, killing, and disappearance of journalists. The CPJ's annual imprisonment census provides a snapshot of incarcerated journalists each year. However, this census does not account for the numerous journalists who are imprisoned and released throughout the year. Additionally, journalists who go missing or are abducted by non-state entities such as criminal gangs or militant groups are not included in the prison census.

Since 1992, the CPJ has maintained detailed records of journalist fatalities. Their researchers independently investigate and verify the circumstances surrounding each death. The CPJ's database encompasses both "confirmed" cases, where it is evident that a journalist was murdered as a direct reprisal for their work, during combat or crossfire, or while undertaking a hazardous assignment, as well as "unconfirmed" cases that involve unclear motives but may have a potential link to journalism. Ongoing research allows for the reclassification of cases. It is important to note

that while both "confirmed" and "unconfirmed" cases are included in the CPJ's database, targeted statistical analyses only include the "confirmed" cases.

For the development of our model, we extracted the following information from the CPJ database for each country: the number of journalists and media workers killed, the number of journalists imprisoned, and the number of missing journalists. These variables serve as valuable inputs in our model development process.

*4.2.1.4. UNESCO Observatory of Killed Journalists*

The Observatory of Killed Journalists, managed by UNESCO [103], serves as a visual representation of the institution's strategic commitment to combating impunity and addressing crimes against journalists. This initiative aligns with the General Conference 36 C/Resolution 53 (2011), which urges UNESCO to collaborate with other United Nations bodies in monitoring the state of press freedom and the safety of journalists. In order to provide comprehensive insights, the chapter analyses information supplied by the UN Member States, which is then categorised as either Resolved or Ongoing/Unresolved, shedding light on the progress of investigations into journalist deaths. To conduct this analysis, we extracted data from the Observatory, specifically the number of journalists killed in each country, which was used as a variable in our model.

*4.2.1.5. Cost of Shutdown (COST)*

COST [104], developed by NetBlocks, is an invaluable data-driven online service that empowers a wide range of users, including journalists, researchers, advocates, policymakers, businesses, and others, to swiftly and effortlessly generate approximate assessments of the economic impact caused by Internet disruptions. By leveraging established methodologies pioneered by esteemed institutions such as the Brookings Institution and the Collaboration on International ICT Policy for East and Southern Africa (CIPESA), COST accurately gauges the potential economic consequences of internet shutdowns, mobile data blackouts, and social media restrictions. This powerful tool utilises publicly available economic indicators that pertain to the global digital economy. We utilised the COST platform to construct an additional variable for model development, specifically capturing the hourly cost of shutdown in each country, expressed in USD.

*4.2.1.6. Global Cybersecurity Index*

The Global Cybersecurity Index (GCI) [105] is a reputable source that evaluates countries' dedication to cybersecurity on a global scale, with the aim of raising awareness about the significance and diverse aspects of the issue. Given that cybersecurity encompasses a wide range of applications spanning multiple industries and sectors, each country's level of development and

engagement is assessed across five pillars: Legal Measures, Technical Measures, Organisational Measures, Capacity Development, and Cooperation. These pillars are then combined to form an overall score.

The GCI adopts a multi-stakeholder approach and relies on the expertise and capabilities of various organisations. Its objectives include enhancing the survey's quality, fostering international cooperation, and promoting knowledge exchange in the field of cybersecurity. The initiative is built upon the foundation and framework provided by the ITU Global Cybersecurity Agenda (GCA). To develop the model, the GCI score for each country were utilised as a variable.

#### 4.2.2. Data not included in the development of the Index Index

The model does not include metrics which have no immediate bearing on, or a proxy indication of, issues relating to free expression. We nevertheless provide socio-economic data and broader contextual information that can be viewed when viewing data from a specific country on the online map that accompanies this project, in a hover-over box that appears while viewing specific country data. The interactive map is included in the Supplementary Materials.

We included this information to provide broadly corollary metrics that immediately show texture and depth to the metrics featured. This first revived iteration of the Index Index is provided alongside contextual data on the UN Human Development Index (HDI), the Gross Domestic Product (GDP) per capita as compiled by the UN, and the Population data as compiled by the United Nations Population Fund (UNFPA), enabling the reader to explore links - if any - between this data and the core metrics.

#### 4.2.3. A note on the political entities included

Our modelling and visualisation are influenced by the indices comprising the dataset. This influence becomes evident through the inclusion and exclusion of various countries and political entities in the Index Index. The Index Index incorporates both UN and non-UN member states, countries with observer status, and other nations or regions that may be autonomous parts of other states. For example, Kosovo and Taiwan are included in the Index Index despite not being recognized as UN member states, while Greenland, an autonomous part of Denmark, lacks available data.

Moreover, the rankings of the British Overseas Territories, which are autonomous parts of the UK, and the overseas parts of France and the Netherlands, are attributed to their respective states. However, it is important to note that the nature of these overseas territories varies significantly.

Unfortunately, due to gaps in the datasets, the Index Index lacks data for several countries, including (but not limited to) Liberia, Papua New Guinea, Federated States of Micronesia,

Kiribati, Palau, Tonga, Tuvalu, Samoa, Dominica, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Grenada, Andorra, Liechtenstein, San Marino, and the Holy See.

#### 4.2.4. Index Index Ranking

The data was modelled using GTM to generate data clusters whereby each cluster represented at least one country that shared similar characteristics. A ranking was then generated by leveraging aggregated, normalised information from the reference maps that represent the relevant extracted variables. In this sense, a country will be given a score, which is calculated using equation (10):

$$Score_n = \sum_{i=1}^{M} R_{in}\tilde{\boldsymbol{y}}_i \qquad (10)$$

Where $\tilde{\boldsymbol{y}}_i$ is the normalised reference vector or centre $\boldsymbol{y}_i$. Countries are ranked according to their calculated score. This ranking is not a direct ranking of countries, but instead, it is a ranking of the different country clusters that were automatically identified from the data using GTM. This means that in a single position of the ranking, we could have more than one country sharing such a position. It should be noted that this score will be directly affected by the value chosen for the $\boldsymbol{M}$ latent nodes. Selecting a value for $\boldsymbol{M}$ that is too small may may lead to non-similar counties being cluster together, whereas if $\boldsymbol{M}$ is too large then GTM may separate similar countries into separate clusters. In either case, this emphasises the need to choose an appropriate value for $\boldsymbol{M}$ to optimise the ability of the GTM model to capture the underlying relationships in the data and generate meaningful country clusters. The developed ranking was then divided into 10 groups according to its distribution of scores to form the 10 deciles of the scale of free expression, where lower deciles represent higher levels of free expression and higher deciles represent lower levels.

### 4.3. Results

#### 4.3.1. Country cluster visualisation

The visualisation in Figure 14 (representing the cluster membership map in the GTM latent space of the developed model) shows a representation of a different kind of world map, where every circle represents a cluster, and each cluster represents one or more countries. Following the original GTM publication [46], we set the number of clusters to 100 (arranged in a grid of $10 \times 10$) and the number of basis functions to 16 (arranged in a grid of $4 \times 4$). The GTM regularisation term was optimised, and the one resulting in the lowest error (negative log-likelihood) was selected (Table S2, Supplementary Material). As discussed earlier, the GTM predicted the probability of countries belonging to the same clusters in the below visualisation. The top-left-hand side of the visualisation represents the highest deciles of free expression, while the top right represents the lowest. This visualisation of the data is intended to help identify

commonalities or differences and related factors to better understand the changing free expression landscape. Figure 14 shows the countries allocated to a selection of clusters. The full allocation of countries per cluster can be found in the Figures S14 and S15 in the supplementary material.

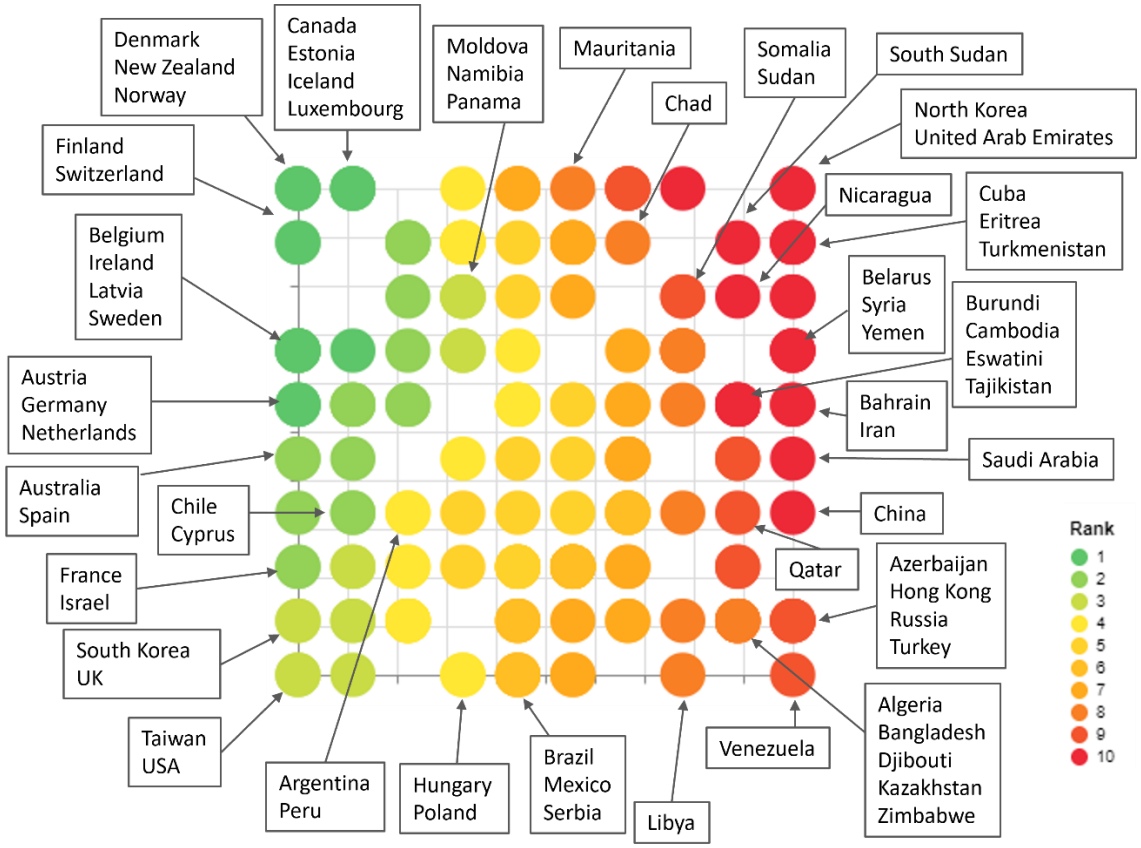



*Figure 14. Country clustering visualisation (cluster membership map) colour-coded by the cluster ranking. The countries allocated to a selection of clusters are displayed. Cluster separation indicates similarity (i.e. closer clusters are more similar than further clusters).*

### 4.3.2. Visualisation of the reference maps

A selection of reference maps is presented in Figure 15, showing the distribution of the clusters (and therefore countries) against the selected variables. They are organised by IoC freedom index areas: academic, media and digital freedom. The reference maps corresponding to all the variables used can be found in the Supplementary Materials, Figures S15 to S20.

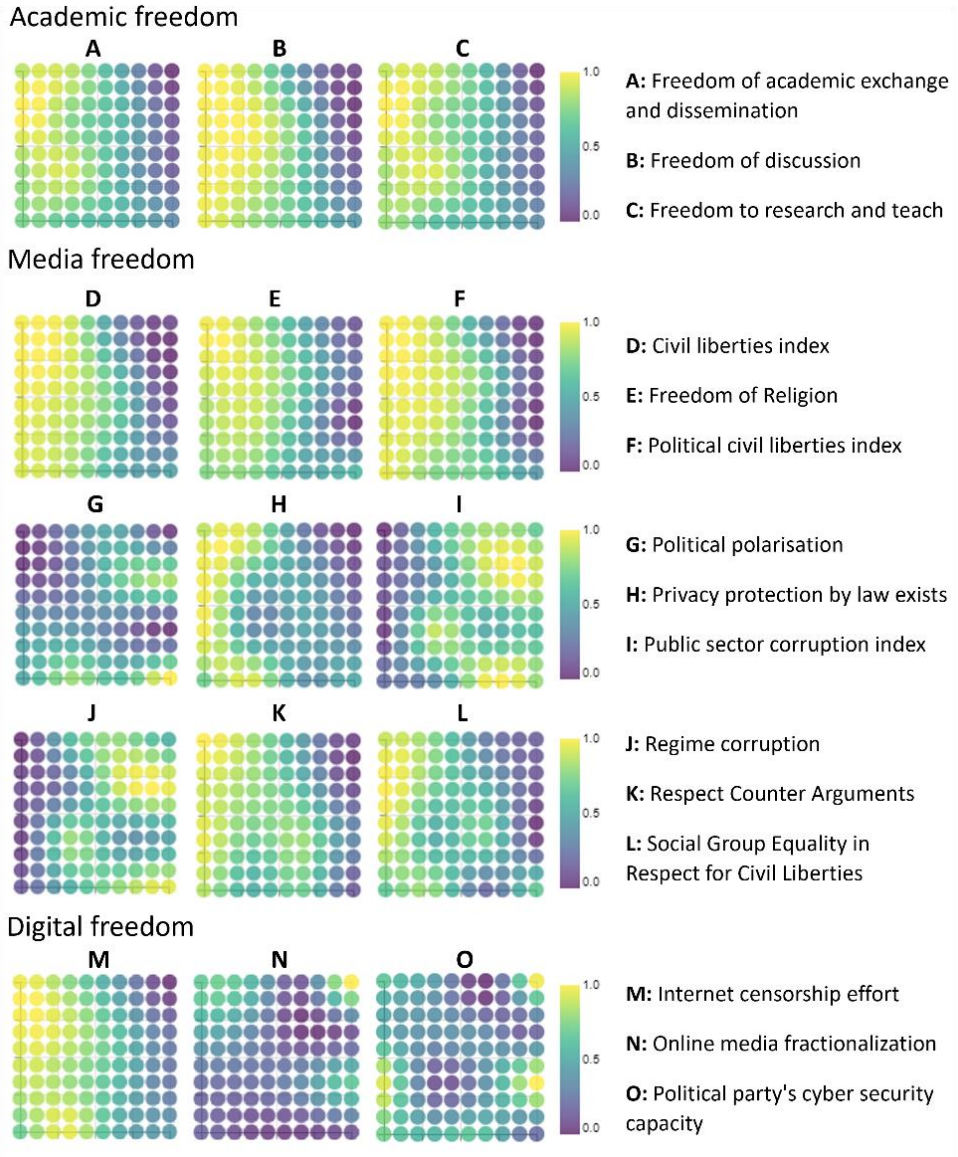



*Figure 15. Selected reference maps for 15 of the variables used to produce the GTM model.*

### 4.3.3. Global ranking of countries/nations – deciles

The Index Index groups states' free expression ranking into ten categories - deciles - intended to convey the complexity and nuance of the global practice of censorship, see Table 3. The deciles ensure the eventual ranking does not erase distinctions between countries/nations, but also presents a clear picture of the global free expression environment. A world map representation showing the global ranking of censorship by deciles is shown in Figure 16, with the highest deciles of free expression represented in green (lowest values), and the lowest levels in red (highest values). The rankings per area of freedom (academic, digital and media) can be found in Figures S21 to S23 and Table S3 to S5 in the Supplementary Materials.

*Table 3. Global ranking of free expression by deciles. Lower ranks represent higher levels of free expression while higher ranks represent lower levels of freedom. Note: The countries within each grouping are ranked alphabetically and do not present a ranking within the groupings.*

| Countries and nations | Global rank |
|---|---|
| Austria, Belgium, Canada, Denmark, Estonia, Finland, Germany, Iceland, Ireland, Latvia, Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Sweden, Switzerland | 1 |
| Australia, Barbados, Cape Verde, Chile, Costa Rica, Cyprus, Dominican Republic, France, Israel, Italy, Jamaica, Japan, Malta, Portugal, Slovakia, Spain, Trinidad and Tobago, Uruguay | 2 |
| Czechia, Greece, Moldova, Namibia, Panama, Romania, South Africa, South Korea, Suriname, Taiwan, Tunisia, United Kingdom, United States of America, Vanuatu | 3 |
| Argentina, Armenia, Benin, Botswana, Bulgaria, Croatia, Georgia, Ghana, Guyana, Hungary, Kosovo, Mongolia, Montenegro, Peru, Poland, Sao Tome and Principe, Senegal, Seychelles, Slovenia, Solomon Islands, Timor-Leste | 4 |
| Albania, Ecuador, Guatemala, Guinea-Bissau, Honduras, Madagascar, Malawi, Maldives, Mauritius, Mozambique, Niger, Nigeria, Paraguay, Sierra Leone, The Gambia | 5 |
| Angola, Bhutan, Bolivia, Bosnia and Herzegovina, Brazil, Indonesia, Ivory Coast, Jordan, Kenya, Kyrgyzstan, Lesotho, Mexico, Nepal, North Macedonia, Philippines, Serbia, Singapore | 6 |
| Burkina Faso, Central African Republic, Colombia, Comoros, Democratic Republic of the Congo, El Salvador, Fiji, Gabon, Haiti, India, Kuwait, Lebanon, Malaysia, Mali, Morocco, Pakistan, Sri Lanka, Tanzania, Togo, Ukraine, Zambia | 7 |
| Algeria, Bangladesh, Cameroon, Chad, Djibouti, Ethiopia, Guinea, Iraq, Kazakhstan, Libya, Mauritania, Rwanda, Thailand, Uganda, Zimbabwe | 8 |
| Afghanistan, Azerbaijan, Egypt, Hong Kong, Oman, Palestine, Qatar, Republic of the Congo, Russia, Somalia, Sudan, Türkiye, Uzbekistan, Venezuela, Vietnam | 9 |
| Bahrain, Belarus, Burma/Myanmar, Burundi, Cambodia, China, Cuba, Equatorial Guinea, Eritrea, Eswatini, Iran, Laos, Nicaragua, North Korea, Saudi Arabia, South Sudan, Syria, Tajikistan, Turkmenistan, United Arab Emirates, Yemen | 10 |

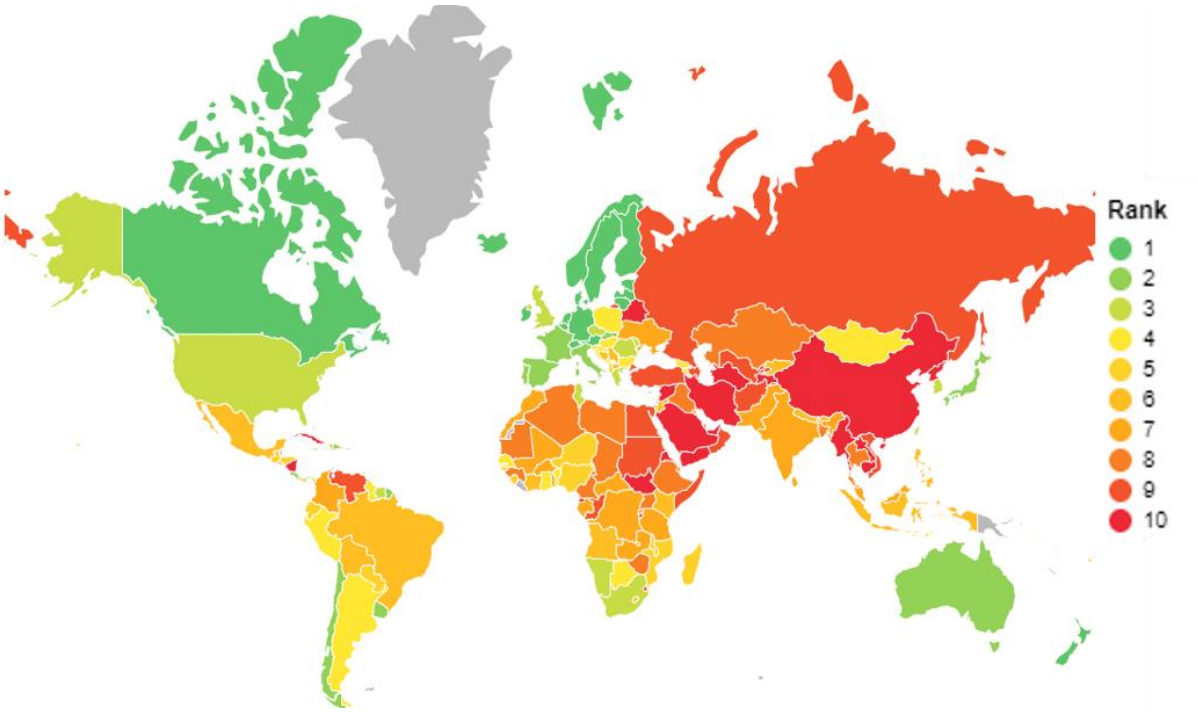



*Figure 16. World map showing the global free expression ranking*

### 4.4. Discussion

#### 4.4.1. Creating meaningful representations using GTM

Due to the challenges of data collection and data representation, there exist high levels of uncertainty that could potentially have a negative impact on the modelling process. GTM, being a robust probabilistic algorithm, calculates the probability of a cluster being responsible for a country while accounting for this uncertainty. In this analysis, the GTM cluster centres or prototypes serve as representations of freedom of expression, effectively stratifying the landscape of freedom of expression. A crucial property of GTM is the preservation of data topology, signifying that similar clusters will be positioned closer together in the latent space. Consequently, even if the most probable cluster assigned to a country does not precisely correspond to the actual one, it is expected to be closer to the correct one. In contrast, popular clustering techniques such as k-means, lacking probabilistic foundations, are not specifically designed to handle such levels of uncertainty.

In addition, GTM is particularly useful for crafting meaningful data representations by transforming high-dimensional information into a lower-dimensional space while retaining the intrinsic structure of the data. Alternative visualisation algorithms such as t-SNE [106] and UMAP [107] have gained popularity for data visualisation through dimensionality reduction. However, these techniques do not possess the capability to extract data prototypes in the manner that GTM does, which poses a challenge when it comes to stratifying countries based on freedom of

expression. In contrast, GTM creates a visualisation (the membership map) that captures the underlying patterns, relationships, and clusters within the data by mapping data points to these prototypes. This process allows for a more comprehensible and interpretable depiction of complex data, aiding in knowledge extraction and facilitating insights that might otherwise remain hidden in the original high-dimensional space. GTM has found applications in various real-world scenarios across different domains. In bioinformatics, it has been used to model protein structures and understand their conformational spaces, providing insights into protein folding and function, which is crucial for drug design [108], disease understanding [109], and other biomedical applications [110–112]. It has also been used to model species distributions and understand ecological patterns, e.g., to understand the species composition of a forest to assess biodiversity [113], and to study the ecological status of streams [114]. It has also been used in the financial sector, e.g., for early identification of business opportunities [115]. These examples highlight the versatility of the GTM in addressing real-world challenges across diverse fields. However, to the best of our knowledge, GTM has not been used before to study censorship or freedom of expression, hence making this a positional article in the application of GTM within this field.

#### 4.4.2. Interpreting the visualisations (membership and reference maps)

The country clustering visualisation (membership map) presented in Figure 14 provides another way to examine the data. It can then be used to show: i) the details of the individual countries within each cluster, indicating that they share very similar characteristics; ii) the location of the countries across all clusters, allowing for the representation of a certain degree of similarity if they are allocated to neighbouring clusters; and iii) the assigned colour-coded ranking to each of the clusters, and therefore to the countries that these clusters represent.

The reference maps provide further information/interpretation about the role played by each variable in the development of the GTM model, with high values representing areas of the maps where the variables had a higher influence, and low values representing otherwise. When exploring the reference maps of the academic freedom variables from Figure 15, which include freedom of academic exchange and dissemination (Figure 15.A), freedom of discussion (Figure 15.B), and freedom to research and teach (Figure 15.C); we can see that the higher values for those variables are on the left-hand side of the reference maps, which coincide with the areas with better rankings of freedom (see Figure 14).

Regarding the media freedom variables, we can also see high values on the left-hand side of Figure 15.D and Figure 15.F which represent civil liberties and political civil liberties, respectively. In the case of the public sector corruption index (Figure 15.I), we see high values in the top right quadrant where the clusters represent countries such as Nicaragua, Yemen, Somalia,

and Eswatini, among others. Also in this area, we see high values in the reference map of regime corruption (Figure 15.J).

In the case of the digital freedom variables, we can see high levels of online media fractionalisation (Figure 15.N) in the cluster of Eritrea, North Korea, and the United Arab Emirates, and high levels of internet censorship effort (Figure 15.M, which higher values meaning that the governments allow generally unrestricted Internet access) in countries represented by a higher level of freedom (left-hand side of Figure 14). These examples illustrate how the role of each of the variables used to produce the GTM model can be studied by visualising their respective reference maps.

### 4.4.3. Insights from the global ranking of countries/nations

A closer inspection of the global ranking in Table 2 and the rankings in the different areas of freedom (academic, digital, and media/press) show that Europe dominates the list of countries that were in the 1st decile (least censorship/greatest freedom) for all three freedoms. These include Austria, Belgium, Denmark, Estonia, Finland, Germany, Iceland, Ireland, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Sweden and Switzerland. The G20 Member States are spread across the full Index Index. Using the global ranking, Australia, Canada and Germany are the highest place members (1st decile), with Saudi Arabia and China being the lowest (10th decile).

For the global ranking, G7 Member States are placed: Canada = 1st, France = 2nd, Germany = 1st, Italy = 2nd, Japan = 2nd, United Kingdom = 3rd and USA = 3rd decile.

Much like G20 Members, UN Security Council members, including both permanent and non-permanent members, are spread across the full Index. Using the global ranking, Ireland and Norway are the highest place members (1st decile) and China and the United Arab Emirates are the lowest (10th decile). Out of the Permanent members, France (2nd decile) is the highest-ranking member, with Russia (9th decile) the lowest. Across the three freedoms, the United Kingdom is consistently found in the 3rd decile. This is similar to the United States of America. However, the latter is in the 4th decile for academic freedom.

The countries that were in the 10th decile for all three freedoms are Bahrain, Belarus, Burma/Myanmar, Cuba, Equatorial Guinea, Eritrea, Iran, Laos, North Korea, Syria, Turkmenistan, United Arab Emirates and Yemen.

### 4.4.4. Use and potential impact of the Index Index

By making available indices that provide objectively verifiable, clearly ranked data about rates of freedom of expression, in contrast to or perhaps as linked to academic freedom, the Index Index seeks to provide legislators and other policymakers, activists and governments, and non-

governmental and intergovernmental organisations, with tools to better inform policy or action decisions. Developing a wide range of campaigning and advocacy tools that can benefit from emergent and innovative technologies and research approaches to synthesize and present compelling and data-rich information is vital to ensure rights advocacy is underpinned by all available expertise that can be accessed easily and clearly. As seen in previous metrics, including those that are incorporated into the dataset for the Index Index, empirical data generated by this pilot project can be highly effective when communicated with policymakers to encourage more affirmative action when it relates to free expression, including more robust protection for journalists [116,117], the formulation of rights policies for educational institutions and ensuring all surveillance policies deployed for policing or national security purposes are rights-respecting. These are a few examples of how the Index Index can be used but should not be assumed to limit how it can be used by a wide range of stakeholders.

While the Index Index abstracts from the particular experiences of writers, journalists and academics facing daily repression across the globe, the overall ranking hints at what is at stake. It constitutes a call, directing the attention of those with a voice to denounce it, to where free expression is at greatest risk and providing insights into the granular policy areas needing attention. The global nature of the proposed index also means it can become a vital resource and tool for engagement with international and supranational bodies such as the United Nations, as well as other regional mechanisms such as the European Union, Council of Europe, African Union and the Inter-American Commission on Human Rights, whose work requires country-by-country, regional and global data sources.

As the Index Index is an index of existing respected and trusted indices and metrics it depends on robust and accurate data produced by the wider community of experts. The process of compiling and producing the Index has demonstrated its own use-case as it has identified the need for increased monitoring, verification and sharing of granular country-by-country level data on a wide range of markers against free expression more broadly, as well as academic, artistic, digital and media/press freedom. While also strengthening further iterations of this pilot project, this will also strengthen the global movement to protect free expression.

In this, too, the analysis provides the basis for developing insights into the political economy of censorship and freedom which shine a light - not always flattering - on human conduct towards others in our midst. Objective data and analysis provided by the Index Index encourage us to ask, simply, what will it take for us to live less censored lives and what must we do to achieve greater respect towards human dignity.

### 4.5. Chapter Conclusion

This project collected and collated pre-existing, robust data on the status of the free expression landscape on a global scale. We modelled the data using the GTM, an unsupervised, probabilistic machine learning method, to explore whether the model produced new insights into state conduct, human rights, and governance. The use of such a model removes an element of subjective interpretation from the modelling process and provides the resulting Index Index with a greater degree of rigour than previous rankings.

On close examination, the reader can be expected to find unexpected outcomes that call into question, correlation, or causality. The Index Index provides a powerful policy tool for all those seeking a clear picture of the health of the free expression environment, as well as what needs to happen to change the rankings.

# 5. Chapter 5: Phenotypes of atrial fibrillation in the UK population

Moving back towards cardiovascular research, this chapter outlines a novel methodological approach used to determine clinically relevant subgroups (referred to as phenotypes) within an AF population. The proposed methodology builds on the basic blueprint outlined in the previous chapter by implementing a more sophisticated approach to generating macro-clusters within the latent space generated by the GTM model. This is achieved by way of applying hierarchical clustering to the reference vectors.

## 5.1. Introduction

AF is the most common heart arrhythmia worldwide [118], affecting 2% of the European population (15 million patients). AF risk increases with age, with ~18 million AF patients estimated by 2060 [119]. AF is linked to a higher risk of mortality and morbidity from stroke, heart failure, dementia, and hospitalisations. Patients with AF are often associated with various cardiovascular and non-cardiovascular risk factors [119], and these often do not occur in isolation, co-existing in clusters of comorbidities, leading to multimorbidity, polypharmacy and frailty [120]. Such clinical complexity associated with AF patients has major implications for treatments and outcomes [121]. To predict AF and AF-related complications, clinical risk scores are commonly employed, but their predictive accuracy is generally limited, given the inherent complexity and heterogeneity of AF patients.

AI, and more specifically ML, is increasingly used in clinical practice for disease prediction and detection, as well as events and treatment optimisation [122]. Most ML applications in AF leverage supervised ML learning (requiring labelled data), however in recent years, there has been a rise in the application of unsupervised ML approaches as they can be used for exploring and understanding the inherent structure and characteristics of the data without requiring labelled outcomes or targets.

Conventional classification of patients with AF based solely on disease subtypes or arrhythmia patterns (e.g. paroxysmal, persistent, or permanent) may fall short of adequately characterising this diverse population [118]. The task of categorising patients into meaningful subgroups/phenotypes is inherently challenging and susceptible to misclassification. These phenotypes, in the context of medical research, are constructs based on clinical and physiological measurements that enable the characterisation of patient subgroups within a specific disease [123]. They comprise either individual disease attributes or combinations thereof, offering a comprehensive description of distinctions among affected individuals, including clinically

significant outcomes such as symptoms, exacerbations, treatment responses, disease progression rate, or mortality. By classifying different presentations of AF into coherent and manageable clinical phenotypes, the development of tailored prevention and treatment strategies can be facilitated. This is aligned with the current holistic approach to AF management [124], as recommended in guidelines [125].

Different approaches have been followed previously to identify AF phenotypes such as hierarchal clustering (namely Ward's minimum variance method [126–128] and complete linkage using Gowers distance [129]) and k-prototype [118]. These methods are not particularly suited to model complex relationships in the data, they assume clusters are generally homogeneous, they tend to be less interpretable, they may be sensitive to initialisation, they may not handle cluster membership uncertainty, and they lack robustness across datasets. However, these studies all demonstrate the potential value of phenotyping, with each identifying between three and six clinically distinct AF phenogroups. The population groups studied also vary, including Japanese [118,127,130], European [126,128,131], and North American [126] populations.

Our analysis proposes a novel methodological approach for generating clinically relevant AF phenotypes for specific patient cohorts, from the general and the critical care populations. To test the proposed approach, we generated phenotypes using two different AF cohorts: one derived from general population data from the UK-Biobank, and the other derived from critically ill patients admitted to the intensive care unit (ICU) from the MIMIC-IV database. These databases were chosen as they are both large and offer a rich pool of variables.

Our novel approach employs GTM, a probabilistic ML method chosen for its ability to elucidate meaningful data representations from large datasets. AF phenotypes were derived from the GTM model, and the inherent clinical characteristics associated with each of them were explored for both cohorts.

## 5.2. Methods

### 5.2.1. Proposed AI-based methodology to generate reliable phenotypes

#### *5.2.1.1. Micro-cluster segmentation using GTM*

Our novel approach, which is a developed version of the process seen in Chapter 4, first uses GTM that calculates the probability of an observed data point, represented here by a patient/participant, belonging to each cluster, as shown in Figure 17. The cluster with the highest probability determines the final cluster assignment, resulting in a fine-grained, micro-segmentation of the original data space. Since we have chosen to use a 2-dimensional latent space

(as was also used in Chapter 4), these data clusters can be visually represented on a 2-dimensional membership map.

Alternative algorithms such as t-SNE [106] and UMAP [107] have become popular for reducing dimensionality and visualising data. Whilst they have different mathematical underpinnings, both methods aim to reflect the underlying structure of the data. However, as opposed to GTM, they are not probabilistic methods; t-SNE and UMAP are deterministic techniques that focus on preserving local and global structures without explicitly modelling probability distributions. This is a limitation of the latter two methods since we are interested in generating probabilistic representations and explicit cluster modelling for the AF phenotypes. A probabilistic approach would offer advantages such as uncertainty quantification, robustness to noise, more specific patient profiles, and the ability to uncover hidden subgroups, ultimately contributing to a more robust stratification of patients.

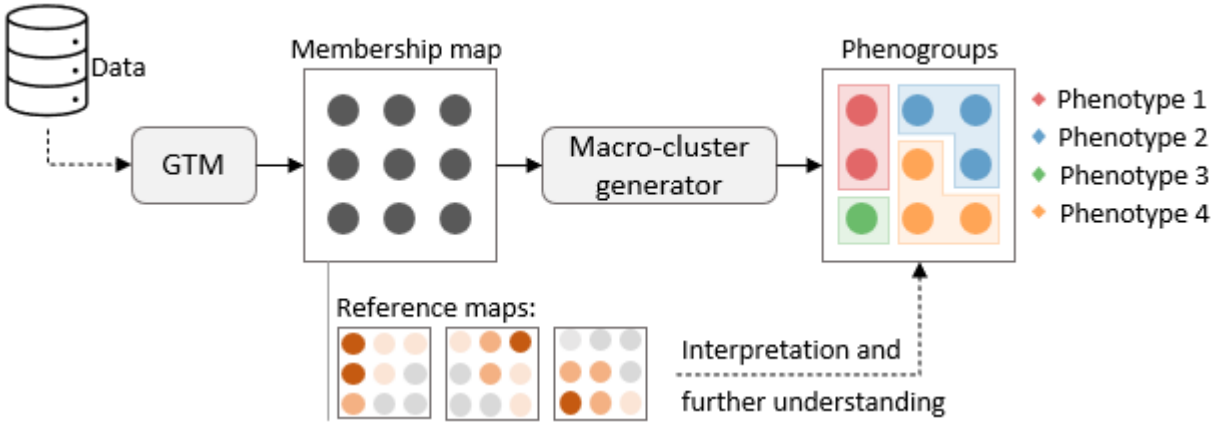



*Figure 17. Proposed AI-based methodology to generate reliable phenotypes. Data is modelled by the GTM algorithm, which projects the data into a 2-dimensional latent space, visualised in the membership map. The GTM also produces reference maps, which are used to indicate the influence of a variable over a micro-cluster. Hierarchical clustering is then applied to the reference vectors to group similar micro-clusters together into larger macro-clusters, which in turn are used to derive the phenotypes.*

As with any ML modelling, a crucial step in the development of ML models is the careful selection of appropriate hyperparameters. This is to ensure the model can learn the key relationships within the data whilst minimising the risk of overfitting and ensuring the model can generalise to unseen data. Although there are scenarios where hyperparameter tuning may be less critical with the GTM method, in this context, where the intended use of phenotypes is not purely prescriptive, the paramount objective was to ensure that the model could generalise effectively, and accurately project new, unseen patients into the most fitting phenotype.

Consequently, we conducted a comprehensive search of a predefined parameter space to identify the most suitable hyperparameters for our model. The specific hyperparameters subjected

to tuning included the number of radial basis functions (RBFs) employed for projecting data from the latent space to the data space, and the penalisation term used to regulate the mapping process.

Each combination of hyperparameters underwent rigorous evaluation through 10-fold cross-validation. The primary performance metric for each test involved assessing the log-likelihood projections of the test data folds. The optimal hyperparameters were selected based on their ability to perform exceptionally well on the test data while also exhibiting minimal standard deviation across all results from each cross-validation fold.

After obtaining a trained GTM model, the reference vectors were extracted and used to generate reference maps for each variable. As already mentioned, reference maps help to show each variables influence on each patient cluster through heatmap visualisations, i.e., the intensity of high and low values represents the extent to which each variable influences different areas of the membership map. An additional approach to interpreting the clusters involves superimposing other variables not seen by the model during the training, presented in the form of a heatmap onto the membership map visualisations. This provides users with an alternative method for comprehending the clusters through post-hoc analysis.

A crucial property of GTM is the preservation of data topology, meaning that similar clusters will be positioned closer together in the latent space. Even if the most probable cluster assigned to a participant does not precisely correspond to the actual one, it is expected to be closer to the correct one. This makes GTM representations valuable for visualising complex high-dimensional data in a more interpretable lower-dimensional space. In contrast, common clustering techniques such as k-means, lacking probabilistic foundations, are not specifically designed to handle such levels of uncertainty.

*5.2.1.2. Macro-cluster analysis to generate AF phenotypes.*

Defining macro-clusters within the array of micro-clusters generated by GTM is crucial for the identification of AF phenotypes. The outcome of such analysis would shed light on regions in the latent space where micro-clusters with similar characteristics are concentrated, representing natural groupings and inherent common patterns in the data space. As defined in Chapter 2, equation (1), the centres in the latent space are projected into the data space to create a non-linear manifold using GTM.

The approach outlined in this chapter (Figure 17) was inspired by an algorithm introduced by Vellido *et al* [50]. Instead of identifying macro-cluster regions in the latent space, we used agglomerative hierarchal clustering using Ward's minimum variance method on the reference vectors, and the distances between the vectors were computed using the Euclidean metric. The

reference vectors corresponded to the Gaussian centres projected from the centres in the latent space, each residing in the data space. Subsequently, the cluster assignment of each reference vector was mapped to their respective centres in the latent space, effectively generating the desired macro-clusters comprising the latent space's micro-cluster centres.

### 5.2.2. Data used for deriving AF phenotypes

*5.2.2.1. Modelling variables extracted from the UK-Biobank database*

The first data used for this analysis was a subset extracted from the UK-Biobank, a large, population-based database [132] encompassing over 500,000 participants aged 40-69 from across the UK. To identify eligible AF participants, we searched ICD-10 codes related to AF diagnosis recorded in the participants' conditions and causes of death variables. Eligible participants would have at least one of these codes recorded, with the full criteria described in Table 4.

*Table 4. Criteria used to identify eligible AF participants from the UK-Biobank*

| Field ID | Variable Value | Field ID - Description | Variable Value - Description |
|---|---|---|---|
| 20002 | 1471 | Non-cancer illness code, self-reported | atrial fibrillation |
| 20002 | 1483 | Non-cancer illness code, self-reported | atrial flutter |
| 41270 | I48 | Diagnoses - (main/secondary) ICD10 | Atrial fibrillation and flutter |
| 41270 | I480 | Diagnoses - (main/secondary) ICD10 | Paroxysmal atrial fibrillation |
| 41270 | I481 | Diagnoses - (main/secondary) ICD10 | Persistent atrial fibrillation |
| 41270 | I482 | Diagnoses - (main/secondary) ICD10 | Chronic atrial fibrillation |
| 41270 | I483 | Diagnoses - (main/secondary) ICD10 | Typical atrial fibrillation |
| 41270 | I484 | Diagnoses - (main/secondary) ICD10 | Atypical atrial flutter |
| 41270 | I489 | Diagnoses - (main/secondary) ICD10 | Atrial fibrillation and atrial flutter, unspecified |
| 40001 | I48 | Underlying (primary) cause of death: ICD10 | Atrial fibrillation and flutter |
| 40001 | I480 | Underlying (primary) cause of death: ICD10 | Paroxysmal atrial fibrillation |
| 40001 | I489 | Underlying (primary) cause of death: ICD10 | Atrial fibrillation and atrial flutter, unspecified |
| 40002 | I48 | Contributory (secondary) causes of death: ICD10 | Atrial fibrillation and flutter |
| 40002 | I480 | Contributory (secondary) causes of death: ICD10 | Paroxysmal atrial fibrillation |
| 40002 | I482 | Contributory (secondary) causes of death: ICD10 | Chronic atrial fibrillation |
| 40002 | I489 | Contributory (secondary) causes of death: ICD10 | Atrial fibrillation and atrial flutter, unspecified |
| 131350 | * Any Date | Date I48 First reported | Present |

.

In total, 67 variables from the UK-Biobank were used for modelling, 40 genomic variables and 27 biological sample variables. We only included these variables to ensure that participants were clustered based on the similarity of their biological and genetic profiles, rather than being influenced by external demographic factors. The genomic variables are a set of 40 principal components generated using >100,000 single nucleotide polymorphisms (SNPs) [133]. The 27 biological sample variables selected aim to represent key risk markers associated with AF: clotting, inflammation, renal function, liver function, cholesterol, diabetes, and sex-related markers [134].

*5.2.2.2. Modelling variables extracted from the MIMIC-IV database*

Data was extracted from the Medical Information Mart for Intensive Care IV (MIMIC-IV [135]), a freely available database of de-identified electronic health records linked to patients admitted to the Beth Israel Deaconess Medical Centre in Boston, Massachusetts. We used version 2.2 (January/2023), which includes 73,181 ICU stays.

Patients were included in this analysis if they had at least one episode of AF during the ICU admission. The latter was extracted from the *chartevent* table, using the code for heart rhythm: 220048, and identifying from those the ones that have value "AF (Atrial Fibrillation)". Therefore, this would include patients with pre-existing AF, and those with new-onset AF, although the first AF episode recorded occurred after the first 24 hours of the ICU admission. Patients <18 years old, patient admissions with short ICU stays (<24 hours), and patients with multiple ICU stays were excluded from the analysis.

In total, 21 variables from the MIMIC-IV database were used for modelling. These variables were extracted from sequences of vitals (e.g., temperature, and heart rate) and lab test results (e.g., glucose and haemoglobin) used to monitor the condition of the patient in the ICU. The variables used for modelling were selected as they represent key risk markers associated with AF in ICU [136,137].

### 5.2.3. Selection of variables associated with AF

*5.2.3.1. AF in the general population: UK-Biobank data*

AF is associated with ageing and comorbidities, as reflected in our phenotypic data. Indeed, multiple studies have shown how comorbid risk factors do not occur in isolation, but cluster together contributing to clinical complexity phenotypes [120,121]. There are well-recognised associations of common comorbidities such as hypertension, heart failure and diabetes, as well as renal and liver dysfunction [138]. The choice of biological sample variables selected for our modelling aims to represent key risk markers associated with AF since they are essential for a

comprehensive understanding of the factors contributing to AF. For example, inflammatory processes play a role in the development and progression of AF [139]. Certain genetic variants have also shown significant association with silent AF [140].

Various risk prediction tools have been proposed for the prediction of incident AF [141], e.g. CHARGE-AF (The Cohorts for Heart and Ageing Research in Genomic Epidemiology AF) score, developed for the general population, which uses variables such as age, ethnicity, height, weight, blood pressure, medication use, and comorbidities [142]. Simpler clinical risk factor scores such as $C_2$HEST have also been investigated to predict incident AF in population and post-stroke cohorts [143].

*5.2.3.1. AF in the critical care population: MIMIC-IV data*

AF stands as the most prevalent arrhythmia among critically ill patients, occurring at an incidence rate of 10–15% [144] within the critical care population. The risk factors for AF can significantly differ between the general and the critical care populations. Common risk factors for AF in the community involve structural and valvular heart disease, but these factors may not be distinctly associated with AF in critical illness [145]. In addition, acute factors are thought to be associated with increased risk for newly diagnosed AF during critical illness [142]. For example, invasive ventilation is associated with AF episodes in critically ill patients [145]. Monitoring oxygenation is crucial in these patients to assess respiratory function and optimise oxygen delivery, as compromised oxygenation can exacerbate cardiovascular stress and contribute to complications [146]. Electrolyte imbalances, such as phosphate abnormalities, observed in medical conditions like kidney dysfunction, may indirectly contribute to AF development.

### 5.2.4. Additional investigative variables

*5.2.4.1. Additional investigative variables extracted from the UK-Biobank database*

We used a set of 18 UK-Biobank variables for visualisation purposes. This selection consisted of 15 assessment centre variables, and two population characteristic variables, with the remaining variable belonging to the health-related outcomes category. Several of these variables were previously identified in prior AF studies [134].

We consider that incorporating comorbidity data is fundamental for understanding how various medical conditions can be differentiated among clusters of AF participants in the general population. To effectively convey information on thousands of diverse comorbidities in a clear, meaningful manner, we integrated the use of *phecode* [147]. Each phecode is composed of several individual diagnoses, defined using ICD-10 codes, which are subsequently grouped into various phecode categories.

In our analysis of AF participants from the general population using UK-Biobank data, we included several phecode categories that encompassed diagnoses from a predefined set of comorbidities commonly associated with individuals suffering from AF. To assign a phecode, and subsequently associate it with a phecode category, a patient's record was examined for a match with the ICD-10 code of either primary or secondary diagnoses to one within a phecode. The list of all phecodes, and their respective phecode categories, that were considered in this analysis can be found in Table S2. For the full details regarding which ICD10 codes make up each phecode, please refer to the original publication [147].

*5.2.4.2. Additional investigative variables extracted from the MIMIC-IV database*

A selection of 27 variables from the MIMIC-IV database were extracted for further investigation. They include demographic data and ethnicity. They also include the Glasgow Coma Scale (GCS), a neurological assessment tool commonly employed in critical care settings, which is used to evaluate a patient's level of consciousness based on their eye, verbal, and motor responses. Ventilation status (invasive and non-invasive), acute kidney injury (AKI) and acute respiratory distress syndrome (ARDS) are also investigated as variables of interest, as well as a series of variables related to length of stay and mortality.

### 5.2.5. Data pre-processing

To ensure the development of a robust and representative dataset for modelling, we undertook several pre-processing steps. First, we implemented a set of missingness criteria (defining appropriate levels/thresholds of data completion) to determine which variables and participants to include. The thresholds were set at 25% and 30% for data that could be missing for a variable or a participant, respectively. We also identified certain variables that exhibited positive skewness in their value distributions. To address this, we applied a log transformation to these variables, rendering their distributions more Gaussian in nature.

Subsequently, any remaining missing data were addressed through imputation, employing a multivariate imputer. This imputer estimated missing values by considering known values from other variables. To accomplish this, we utilised the "IterativeImputer" function, which is part of the Scikit-Learn Python package and draws inspiration from the R MICE package [148]. Invalid values of the variables (e.g., heart rate < 0) were marked as not available. Variables recorded with different units were harmonised, e.g., in MIMIC-IV, height was present in inches and centimetres (cm), and they were all converted to cm.

### 5.2.6. Statistical analysis

Medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (percentages) were used for categorical variables. There were several ordinal variables used for the exploratory analysis of the GTM output. These were one-hot encoded and then treated as a categorical variable and represented in the data as such.

To study the characteristics of the generated phenotype groups, differences between continuous variables were analysed using the Kruskal-Wallis test and differences between categorical variables were analysed using the Chi-squared test. In both cases, a p-value $<0.05$ was the threshold for statistical significance.

## 5.3. Results

### 5.3.1. Characteristics of the participants/patient cohorts

From the UK-Biobank we extracted 36,680 participants with AF from this general population cohort (median age 63 years (IQR 59-67), range 40 to 72 years; 63.5% male). Table 5 contains the summary of the biological variables used for modelling, and the investigative variables used in the post-hoc analysis. A second dataset of 2,695 critically ill patients with AF (median age 73 years (IQR 65-81), range 21 to 89 years; 60.3% male) was extracted from the MIMIC-IV, with the full summary presented in Table 6.

*Table 5. Characteristics of the participant subset extracted from the UK-Biobank database. Medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (as percentages) were calculated for the categorical variables. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.*

| Variable name | Value |
|---|---|
| **Modelling variables:** | |
| Inflammation markers: | |
| Neutrophil count [x$10^9$ cells/L] | 4.3 (3.49, 5.24) |
| Lymphocyte percentage [%] | 27.03 (22.3, 31.93) |
| Monocyte percentage [%] | 7.24 (5.91, 8.68) |
| C-reactive protein [mg/L] | 1.77 (0.86, 3.57) |
| Clotting markers: | |
| Haematocrit percentage [%] | 41.78 (39.31, 44.1) |
| Mean corpuscular volume [Femtolitres] | 91.7 (88.95, 94.5) |
| Red blood cell (erythrocyte) distribution width [%] | 13.5 (13.07, 14.09) |
| Platelet count [x$10^9$ cells/L] | 235 (201, 274) |
| Mean platelet (thrombocyte) volume [Femtolitres] | 9.3 (8.64, 10.06) |
| Platelet distribution width [%] | 16.5 (16.2, 16.86) |
| Mean reticulocyte volume [Femtolitres] | 106.99 (102.5, 111.83) |
| Mean sphered cell volume [Femtolitres] | 83.1 (79.8, 86.66) |
| Diabetes risk markers: | |

| | |
|---|---|
| Glucose [mmol/L] | 5.04 (4.68, 5.49) |
| Glycated haemoglobin (HbA1c) [mmol/mol] | 36.4 (33.8, 39.5) |
| Liver function: | |
| Albumin [g/L] | 44.65 (43.13, 46.1) |
| Alanine aminotransferase [U/L] | 21.56 (16.68, 28.19) |
| Direct bilirubin [umol/L] | 1.74 (1.39, 2.24) |
| Gamma glutamyltransferase [U/L] | 32.4 (22.2, 50.3) |
| Renal function: | |
| Creatinine [umol/L] | 75.6 (65.6, 86.1) |
| Sodium in urine [millimole/L] | 69.3 (44.0, 100.5) |
| Urea [mmol/L] | 5.69 (4.85, 6.63) |
| Urate [umol/L] | 338.01 (284, 393.7) |
| Cholesterol markers: | |
| Cholesterol [mmol/L] | 5.31 (4.53, 6.09) |
| HDL cholesterol [mmol/L] | 1.32 (1.11, 1.57) |
| Triglycerides [mmol/L] | 1.6 (1.14, 2.23) |
| Sex-related markers: | |
| SHBG [nmol/L] | 44.98 (33.62, 58.9) |
| Testosterone [nmol/L] | 8.73 (1.62, 12.2) |
| **Additional investigative variables:** | |
| Demographics: | |
| Age at recruitment [years] | 63 (59, 67) |
| Sex [Male] | 23,284 (63.5%) |
| Waist circumference [cm] | 96 (87, 106) |
| Hip circumference [cm] | 105 (99, 111) |
| Standing height [cm] | 172 (164, 178) |
| Weight [kg] | 83.3 (72.9, 95) |
| BMI [kg/m$^2$] | 28.16 (27.1, 29.98) |
| Activity level: | |
| Summed minutes activity [mins] | 95 (50, 180) |
| MET minutes per week for vigorous activity [mins/week] | 120 (0, 720) |
| Blood pressure: | |
| Diastolic blood pressure, automated reading [mmHg] | 82 (75, 90) |
| Systolic blood pressure, automated reading [mmHg] | 143 (130, 157) |
| Pulse rate, automated reading [bpm] | 68 (60, 77) |
| Respiratory measures: | |
| Forced expiratory volume in 1 second (FEV1) [L] | 2.68 (2.15, 3.27) |
| Peak expiratory flow (PEF) [L/min] | 383 (295, 484) |
| Forced expiratory volume in 1 second (FEV1) Z-score | 0.62 (-0.12, 1.37) |
| FEV1/ FVC ratio Z-score | 0.4 (-0.13, 1) |
| Alcohol intake frequency: | |
| Daily or almost daily [yes] | 7,170 (19.6%) |
| Three or four times a week [yes] | 6,417 (17.5%) |
| Once or twice a week [yes] | 6,869 (18.7%) |
| One to three times a month [yes] | 2,734 (7.5%) |
| Special occasions only [yes] | 3,354 (9.1%) |

| | |
|---|---|
| Never [yes] | 2,734 (7.5%) |
| Ethnic background: | |
| White [yes] | 35,536 (96.9%) |
| Asian or Asian British [yes] | 406 (1.1%) |
| Black or Black British [yes] | 247 (0.7%) |
| Mixed [yes] | 111 (0.3%) |
| Other ethnic group [yes] | 160 (0.4%) |
| Chinese [yes] | 36 (0.1%) |
| AF and flutter diagnosis (main/secondary): | |
| ICD10 - AF and flutter [yes] | 20,966 (57.2%) |
| ICD10 - Paroxysmal AF [yes] | 6,558 (17.9%) |
| ICD10 - Persistent AF [yes] | 1,274 (3.5%) |
| ICD10 - Chronic AF [yes] | 570 (1.6%) |
| ICD10 - Typical AF [yes] | 216 (0.6%) |
| ICD10 - Atypical atrial flutter [yes] | 86 (0.2%) |
| ICD10 - AF and atrial flutter, unspecified [yes] | 21,767 (59.3%) |
| Systems (phecode categories): | |
| Endocrine/metabolic [yes] | 10,119 (27.6%) |
| Circulatory system [yes] | 26,628 (72.6%) |
| Respiratory [yes] | 6,097 (16.6%) |
| Diabetes: | |
| Type 1 diabetes [yes] | 839 (2.3%) |
| Type 1 diabetes with ketoacidosis [yes] | 81 (0.2%) |
| Type 1 diabetes with renal manifestations [yes] | 60 (0.2%) |
| Type 1 diabetes with ophthalmic manifestations [yes] | 175 (0.5%) |
| Type 1 diabetes with neurological manifestations [yes] | 96 (0.3%) |
| Diabetes type 1 with peripheral circulatory disorders [yes] | 52 (0.1%) |
| Type 2 diabetes [yes] | 7,130 (19.4%) |
| Type 2 diabetes with ketoacidosis [yes] | 96 (0.3%) |
| Type 2 diabetes with renal manifestations [yes] | 233 (0.6%) |
| Type 2 diabetes with ophthalmic manifestations [yes] | 852 (2.3%) |
| Type 2 diabetes with neurological manifestations [yes] | 427 (1.2%) |
| Diabetes type 2 with peripheral circulatory disorders [yes] | 351 (1%) |
| Hypertension: | |
| Essential hypertension [yes] | 24,442 (66.6%) |
| Other hypertensive complications [yes] | 86 (0.2%) |
| Cardiovascular disease: | |
| Myocardial infarction [yes] | 6,544 (17.8%) |
| Other forms of chronic heart disease [yes] | 2 (0%) |
| Congestive heart failure (CHF) NOS [yes] | 3,760 (10.3%) |
| Chronic pulmonary heart disease [yes] | 1,105 (3%) |
| Heart failure NOS [yes] | 4,680 (12.8%) |
| Coronary atherosclerosis [yes] | 163 (0.4%) |
| Peripheral vascular disease: | |
| Peripheral vascular disease, unspecified [yes] | 1,911 (5.2%) |
| Other specified peripheral vascular diseases [yes] | 23 (0.1%) |

| | |
|---|---|
| Pulmonary hypertension: | |
| Primary pulmonary hypertension [yes] | 403 (1.1%) |
| Stroke: | |
| Hemiplegia [yes] | 1,214 (3.3%) |
| Liver disease: | |
| Liver abscess and sequelae of chronic liver disease [yes] | 373 (1%) |
| Alcoholic liver damage [yes] | 379 (1%) |
| Other chronic non-alcoholic liver disease [yes] | 1,441 (3.9%) |
| Other disorders of the liver [yes] | 808 (2.2%) |
| Kidney disease: | |
| End-stage renal disease [yes] | 484 (1.3%) |

*Table 6.* Characteristics of the ICU patient subset extracted from the MIMIC-IV database. *Summary statistics and colours as in Table 5*

| Variable name | Value |
|---|---|
| **Modelling variables:** | |
| Diabetes risk marker: | |
| Glucose [mg/dL] | 131.88 (118.17, 155.5) |
| Bone profile: | |
| Phosphate [mg/dL] | 3.58 (3.05, 4.22) |
| Oxygenation: | |
| Oxygen saturation [%] | 96.33 (94.38, 97.83) |
| Respiratory rate [breaths per min] | 18.51 (16.5, 21.27) |
| Fraction inspired oxygen, FiO2 [%] | 56.47 (50, 63.24) |
| Positive end-expiratory pressure (PEEP) [cmH2O] | 5.6 (5, 7.11) |
| Partial pressure of oxygen [mmHg] | 135.08 (99.15, 168.5) |
| Haemoglobin [g/dL] | 10.16 (9.11, 11.48) |
| Respiratory/metabolic markers: | |
| pH | 7.35 (7.21, 7.39) |
| Anion Gap [mEq/L] | 13.42 (11.33, 16.0) |
| Lactate [mmol/L] | 2.0 (1.49, 2.75) |
| Cardiac markers: | |
| Heart rate [beats per min] | 81.33 (74.24, 90.42) |
| Capillary refill rate | 0.0 (0.0, 0.02) |
| Diastolic BP [mmHg] | 57.25 (51.5, 63.38) |
| Systolic BP [mmHg] | 111.93 (104.73, 121.34) |
| Clotting markers: | |
| Prothrombin time [sec] | 14.47 (13.07, 16.45) |
| Platelet count [K/uL] | 165.0 (125.12, 223.0) |
| Renal function: | |
| Creatinine [mg/dL] | 1.03 (0.8, 1.56) |
| Electrolytes: | |
| Magnesium [mg/dL] | 2.15 (1.91, 2.44) |
| Potassium [mEq/L] | 4.29 (3.95, 4.61) |
| Other: | |

| | |
|---|---|
| Temperature [°C] | 36.74 (36.55, 37.0) |
| **Additional investigative variables:** | |
| Demographics: | |
| Age [years] | 73 (65, 81) |
| Sex [Male] | 1627 (60.4%) |
| Height [cm] | 170.09 (162.78, 177.9) |
| Weight [kg] | 82.43 (68.39, 97.37) |
| Ethnicity: | |
| White [yes] | 1971 (73.1%) |
| Other ethnic group [yes] | 453 (16.8%) |
| Black [yes] | 138 (5.1%) |
| Hispanic [yes] | 68 (2.5%) |
| Asian [yes] | 65 (2.4%) |
| Glasgow Coma Scale (GCS): | |
| GCS eye-opening | 2.88 (1.92, 3.75) |
| GCS motor response | 4.83 (3.5, 6) |
| GCS verbal response | 2.54 (1, 4.33) |
| Ventilation: | |
| Non-Invasive ventilation [yes] | 209 (7.8%) |
| Invasive ventilation [yes] | 2116 (78.5%) |
| Outcomes: | |
| Time to AF diagnosis [hours] | 53 (38, 83) |
| In-hospital length of stay [hours] | 256.78 (166.48, 407.12) |
| In-ICU length of stay [hours] | 109.18 (72.9, 200.43) |
| Death after ICU [hours] | 167.07 (17.64, 2700.04) |
| Death after hospital discharge [hours] | 20.44 (10.55, 2551.06) |
| Death after hospital discharge [days] | 0.85 (0.44, 106.29) |
| In-hospital mortality [yes] | 567 (21.0%) |
| In-ICU length of stay of 3+ days [yes] | 2040 (75.7%) |
| In-ICU length of stay of 7+ days [yes] | 840 (31.2%) |
| Mortality after hospital discharge within 30 days [yes] | 711 (26.4%) |
| Mortality after hospital discharge within 365 days [yes] | 936 (34.7%) |
| Mortality after hospital discharge after 365 days [yes] | 152 (5.6%) |
| Acute Kidney Injury (AKI) [yes] | 545 (20.2%) |
| Acute Respiratory Distress Syndrome (ARDS) [yes] | 174 (6.5%) |

### 5.3.2. Visualisation of the membership maps

The results of the GTM hyperparameter tuning showed that for a latent space of dimension 15x15 to provide an appropriate level of granularity, using 196 RBFs arranged in a 14x14 grid with a regularisation term of 1 was optimal and was therefore used when training the GTM models for both the UK Biobank and MIMIC-IV cohorts. Figure 18(A) and 18(B) show the membership map generated by the GTM models trained on the UK Biobank and MIMIC-IV cohorts, respectively. These maps display the latent space containing a compressed representation of the

entire original data space. Each point on the map represents a micro-cluster containing at least one participant, with the size of the point indicating the number of participants in the cluster: the larger the point, the more participants in the cluster and vice versa. Each participant has a probability of being assigned to every cluster, but the assignments below are the result of the participant being placed in the cluster with the highest probability.

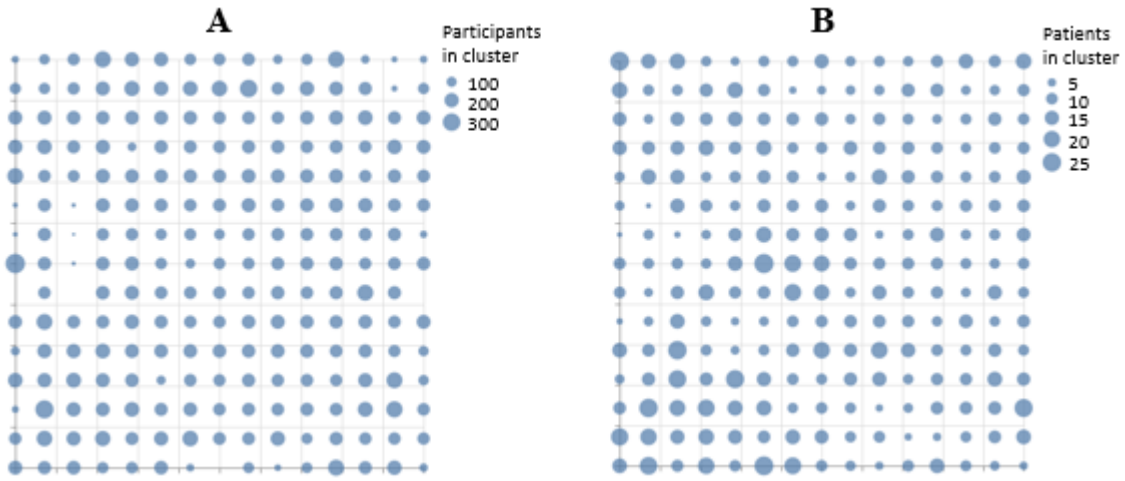



*Figure 18.* Membership maps showing how participants/patients are distributed in the latent space. *The size of each cluster reflects the number of participants/patients allocated to it. A) General population data from the UK Biobank. B) Critical care population data from the MIMIM-IV.*

### 5.3.3. Visualisation of reference vectors for the modelling variables

Figure 19 contains the reference vectors extracted from the trained GTM models for the UK-Biobank and MIMIC-IV AF cohorts. For the UK-Biobank data, it contains the reference vectors for the biological sample variables, with plots grouped by the different risk factors they relate to, whilst for the MIMIC-IV, it displays all modelling variables used for modelling. Each point in every plot within Figure 19 corresponds exactly to the same point in their respective membership maps in Figure 18. A light grey–red colour scheme was used for the reference vectors plot such that areas of the plots that are redder indicate that participants in that cluster had a higher value of that variable. Likewise, if the point in the reference vector is greyer, the lower the value is for participants in this cluster. All plots using the light grey–red colour scheme indicate variables used in the GTM model development, whereas plots using a light grey–teal represent variables that were not used in the modelling and have no direct impact on the clusters themselves.

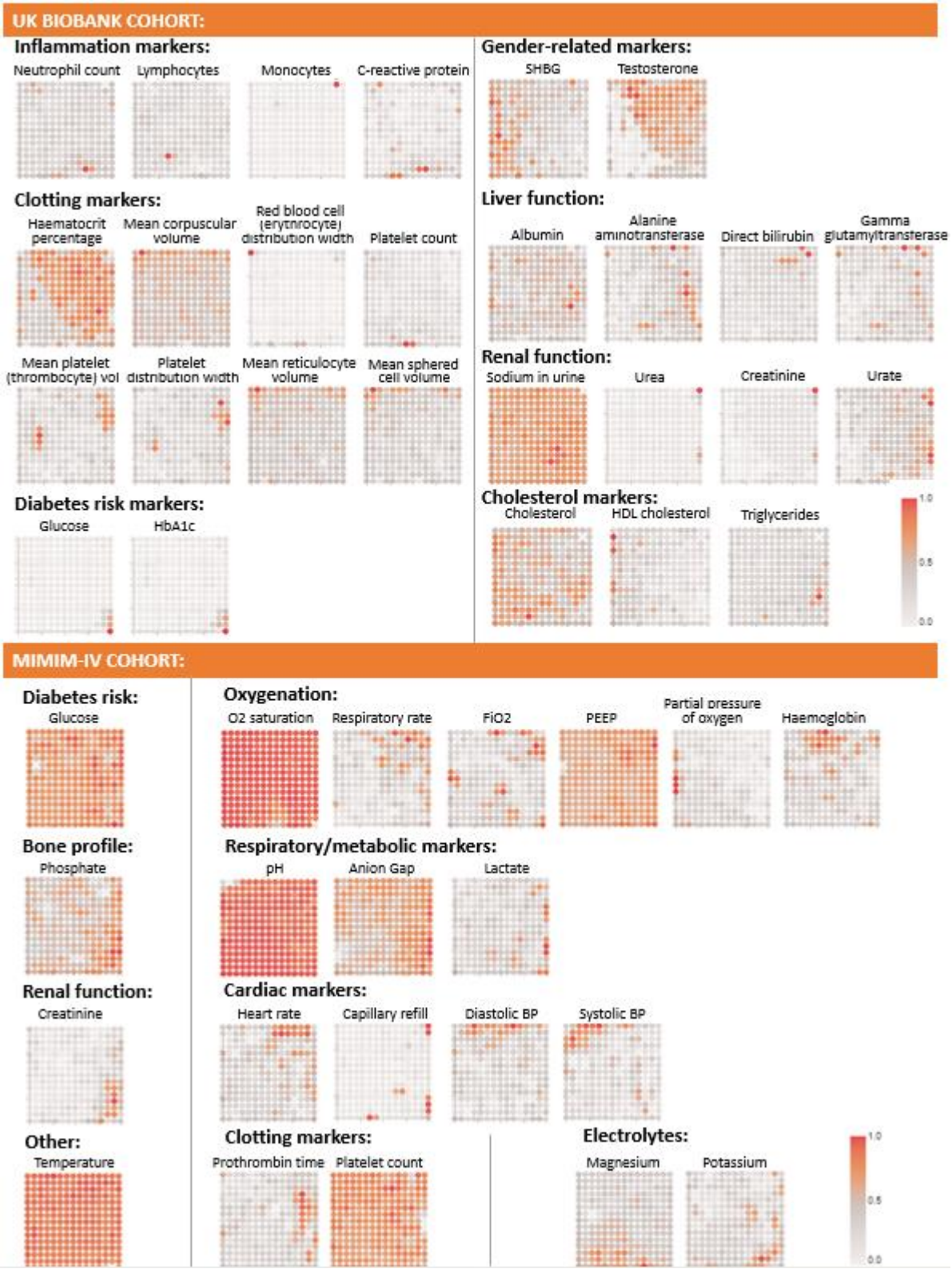



*Figure 19. Reference vector visualisations demonstrating how each biological sample variable affects the cluster distribution in the latent space for both, the UK-Biobank and the MIMIC-IV AF cohorts.*

### 5.3.4. Visualisation of additional investigative variables

Figure 20 contains a selection of visualisations showing how data from different investigative variables are distributed within the membership maps for the UK-Biobank and MIMIC-IV

cohorts. The visualisations representing the investigative variables all use a light grey-teal colour scheme as they were not used in model development. The value assigned to each micro-cluster is the average of the variable for all participants assigned to each cluster, the more teal a micro-cluster is, the higher the value. For the visualisations for all investigated variables described in this chapter, please refer to the "*Visualisation of all the additional investigative variables"* in the supplementary material section.

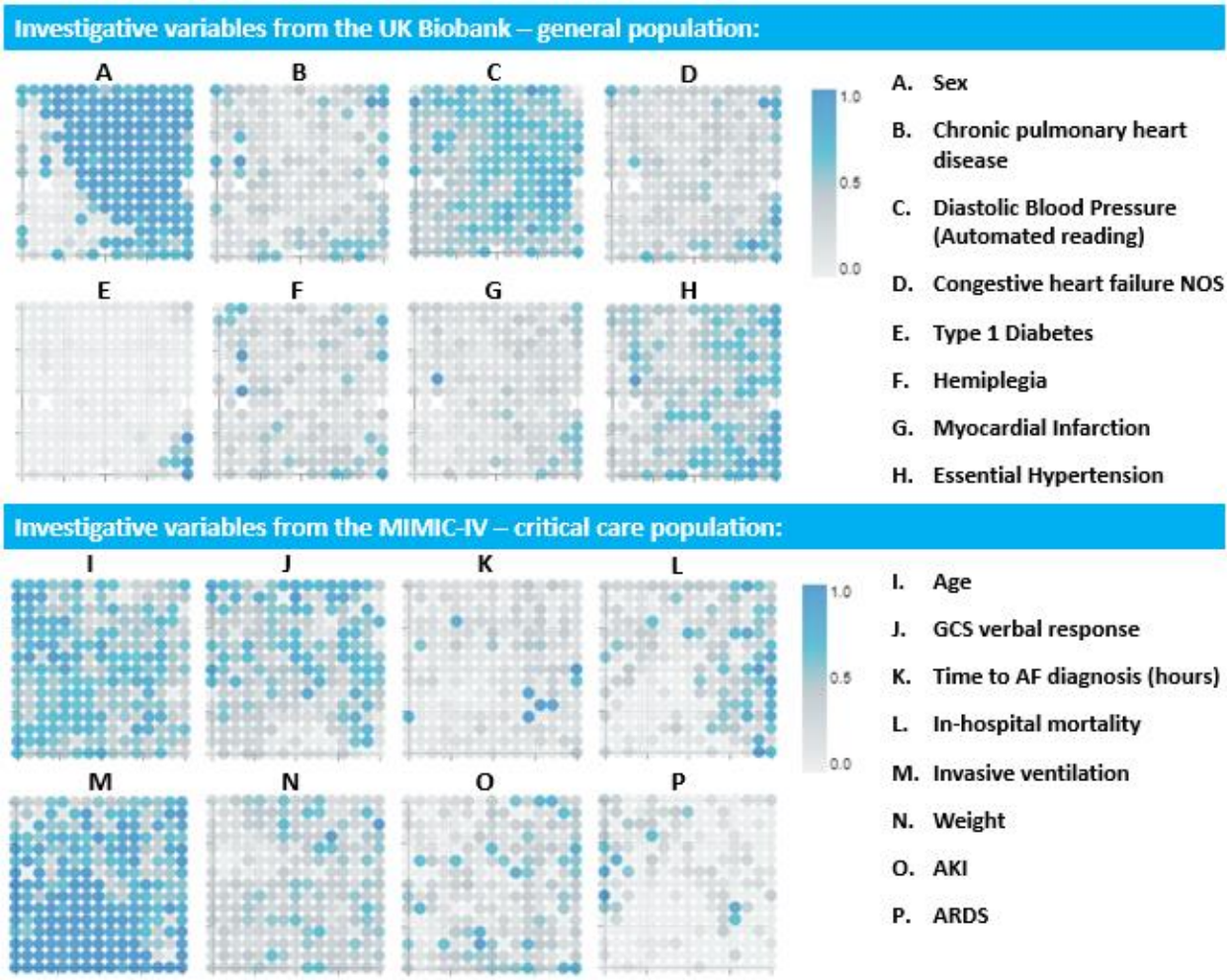



*Figure 20. Membership maps showing how a selection of investigative variables data are distributed within the latent space for the UK-Biobank and the MIMIC-IV cohorts. AF: Atrial Fibrillation. AKI: Acute Kidney Injury. ARDS: Acute Respiratory Distress Syndrome. GCS: Glasgow Coma Scale.*

### 5.3.5. Description of AF phenotypes

For the UK-Biobank cohort, we identified five clusters within the reference vectors residing in the data space, as demonstrated by the dendrogram in Figure 21(A). Transferring these reference vector cluster assignments to their corresponding latent centres gave five macro-cluster regions, which in turn were used to define the five AF phenotypes. These macro-cluster regions are visualised in Figures 21(B) and (C). Likewise, when applied to the MIMIC-IV cohort, the analysis identified four clusters within the reference vectors, as presented in Figure 22(A). The

macro-cluster regions generated by transferring these clusters to their respective latent centres are presented in Figures 22(B) and (C). The baseline data for each of the two databases were split according to the number of phenotypes and compared, in Tables 7 and 8 for the UK-Biobank and MIMIC-IV data, respectively. A description of the headline features that characterise both sets of phenotypes can be found in Figures 21(D) and 22(D).

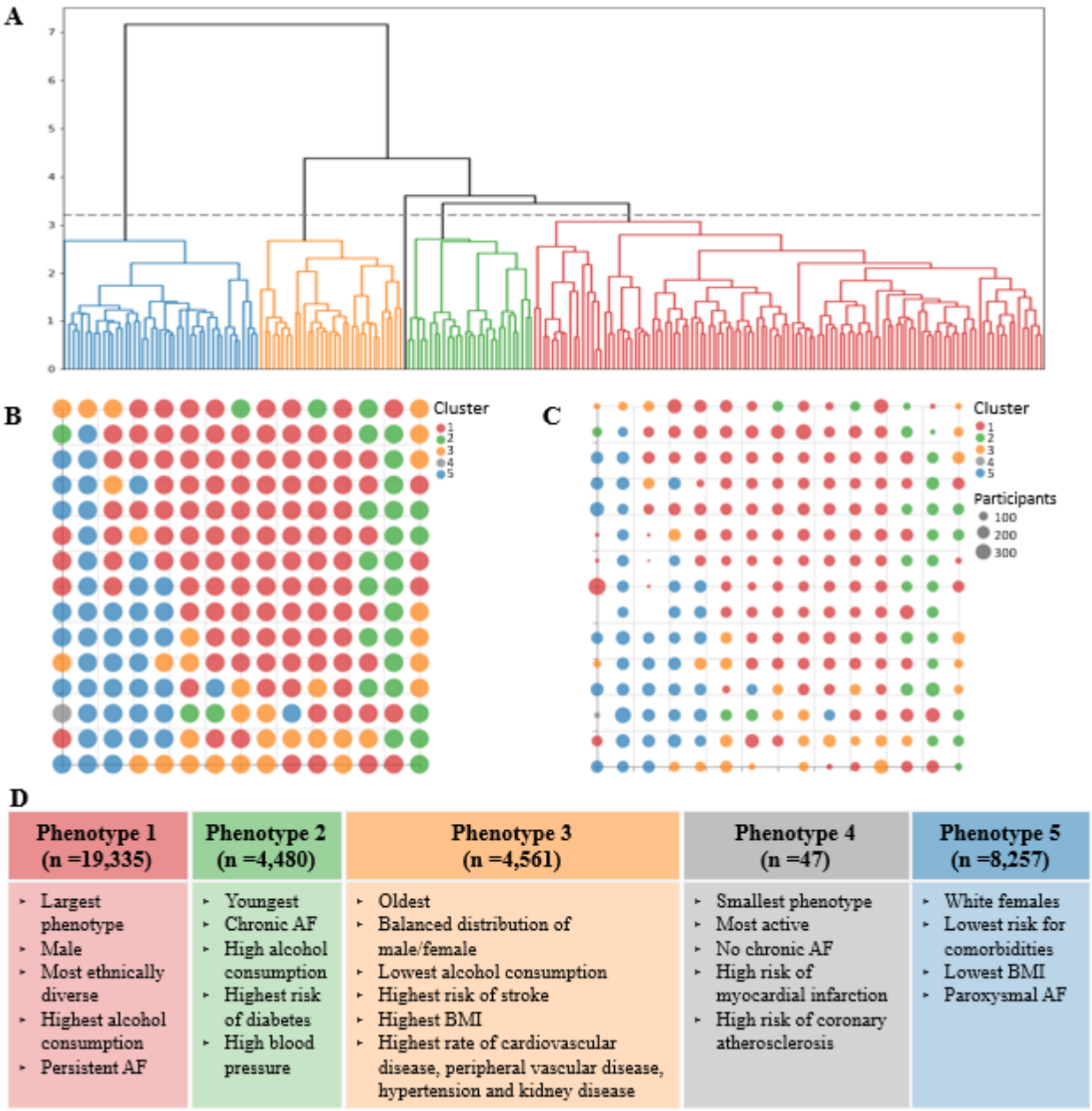



*Figure 21. Derived phenotypes of AF in the general population using UK-Biobank data. A) Dendrogram produced using Ward's minimum variance method. The graph shows the 5 clusters that are used to define the 5 AF phenotypes for the general population. B) Membership map with a uniform size for the micro-clusters to show the distribution of the macro-cluster regions. C) The size of the micro-clusters in the membership map dictated by the number of participants assigned to it. D) Main characterising features for each of the phenotypes*

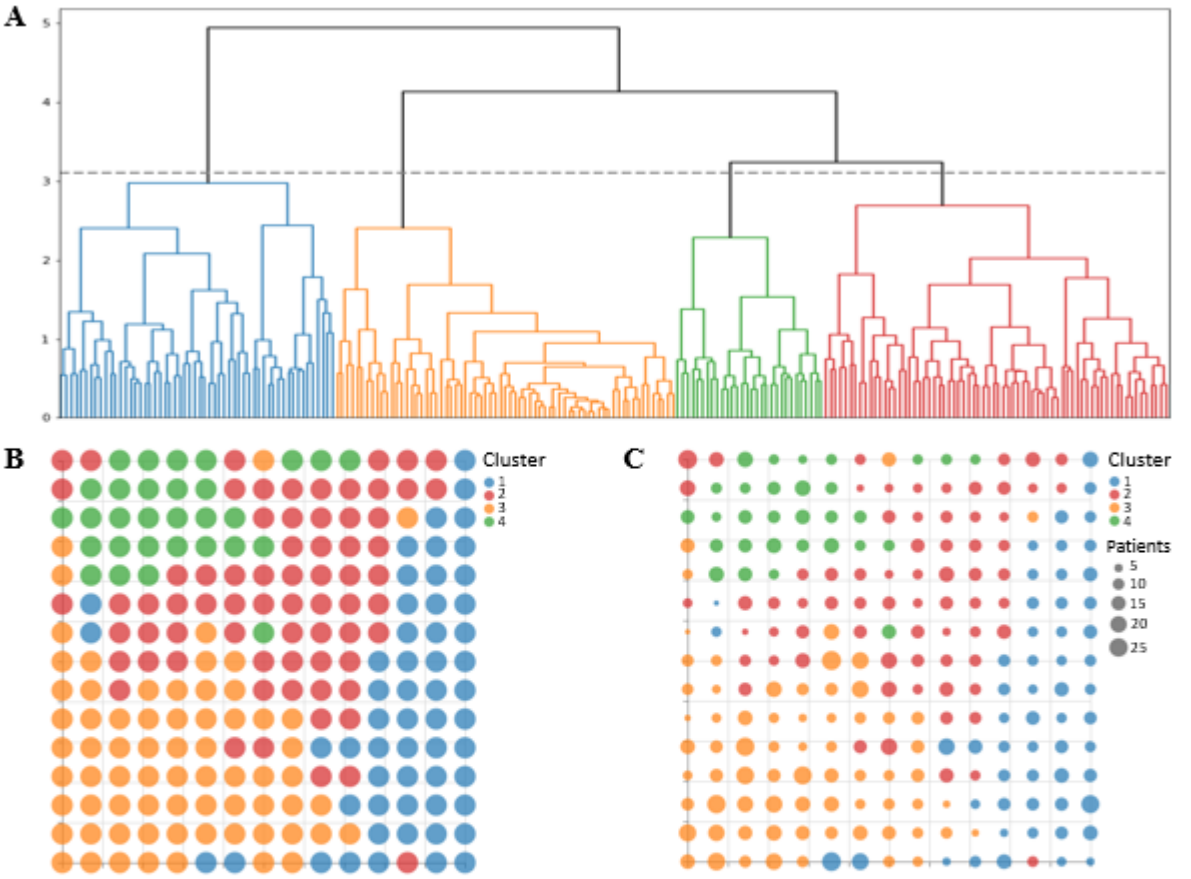



*Figure 22. Derived phenotypes of AF in the general population using MIMIC-IV data. A) Dendrogram produced using Ward's minimum variance method. The graph shows the 4 clusters that are used to define the 4 AF phenotypes for ICU patients. B) Membership map with a uniform size for the micro-clusters to show the distribution of the macro-cluster regions. C) The size of the micro-clusters in the membership map dictated by the number of participants assigned to it. D) Main characterising features for each of the phenotypes.*

*Table 7. Characteristics of the participants per phenotype of AF in the general population using UK-Biobank data. As in Table 5, medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (as percentages) were calculated for the categorical variables. Shades of red/blue were used per variable to illustrate differences between lower and higher values. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.*

| Variable name | Phenotype 1 | Phenotype 2 | Phenotype 3 | Phenotype 4 | Phenotype 5 | p-value |
|---|---|---|---|---|---|---|
| **MODELLING VARIABLES:** | | | | | | |
| **Inflammation markers:** | | | | | | |
| Neutrophil count | 4.27 (3.46, 5.18) | 4.37 (3.57, 5.3) | 5.1 (4.1, 6.35) | 4.46 (3.95, 5.22) | 3.99 (3.25, 4.77) | <0.05 |
| Lymphocyte percentage | 26.53 (22.1, 31.4) | 27.27 (22.63, 32.1) | 24.5 (19.01, 29.74) | 26.15 (23.53, 29.4) | 29.2 (24.67, 33.9) | <0.05 |
| Monocyte percentage | 7.6 (6.24, 9.04) | 7.4 (6.11, 8.8) | 6.7 (5.4, 8.15) | 7.45 (6.16, 8.76) | 6.7 (5.53, 7.91) | <0.05 |
| C-reactive protein | 1.54 (0.79, 2.94) | 2.15 (1.07, 4.11) | 4.75 (2.08, 10.82) | 2.05 (0.9, 3.33) | 1.44 (0.72, 2.86) | <0.05 |
| **Clotting markers:** | | | | | | |
| Haematocrit percentage | 43 (40.98, 44.93) | 42.92 (40.6, 45.13) | 39.82 (37.39, 42.18) | 42.3 (38.6, 45.16) | 39.3 (37.55, 41.07) | <0.05 |
| Mean corpuscular volume | 92.06 (89.46, 94.73) | 91.82 (89.03, 94.9) | 90.1 (86.8, 93.28) | 91.6 (89.03, 93.55) | 91.53 (88.9, 94.12) | <0.05 |
| Red blood cell distribution width | 13.5 (13.06, 14) | 13.43 (13, 13.99) | 13.95 (13.34, 14.89) | 13.6 (13.1, 13.94) | 13.47 (13, 14) | <0.05 |
| Platelet count | 228 (198, 261.45) | 209 (174, 248.53) | 262 (223.6, 308) | 242 (197.45, 275.5) | 253.4 (218.6, 292.8) | <0.05 |
| Mean platelet volume | 9.27 (8.6, 9.91) | 9.9 (9, 10.95) | 9.19 (8.53, 9.8) | 9.17 (8.65, 10.01) | 9.3 (8.61, 10.04) | <0.05 |
| Platelet distribution width | 16.5 (16.2, 16.8) | 16.9 (16.5, 17.36) | 16.49 (16.2, 16.8) | 16.5 (16.17, 16.9) | 16.37 (16.08, 16.7) | <0.05 |
| Mean reticulocyte volume | 107.37 (102.93, 112.11) | 106.47 (101.9, 111.62) | 106.39 (101.8, 111.82) | 105.6 (101.82, 108.46) | 106.6 (102.28, 111.3) | <0.05 |
| Mean sphered cell volume | 83.27 (80, 86.7) | 82.55 (79.36, 86.5) | 81.9 (78.5, 85.56) | 81.71 (79.19, 85.15) | 83.7 (80.4, 87.13) | <0.05 |
| **Diabetes risk markers:** | | | | | | |
| Glucose | 5.02 (4.66, 5.44) | 5.28 (4.8, 6.36) | 5.13 (4.72, 5.76) | 5.09 (4.73, 5.42) | 4.97 (4.67, 5.31) | <0.05 |
| HbA1c | 36.2 (33.6, 39.1) | 37.6 (34.2, 44.63) | 38.5 (35.6, 42.6) | 37.2 (33.35, 40.95) | 35.6 (33.4, 37.9) | <0.05 |
| **Liver function:** | | | | | | |
| Albumin | 44.81 (43.38, 46.2) | 45.09 (43.42, 46.9) | 43.72 (41.96, 45.22) | 44.42 (43.41, 46.41) | 44.5 (43.08, 45.86) | <0.05 |
| Alanine aminotransferase | 22.72 (17.88, 28.67) | 30.56 (22.69, 42.64) | 20.25 (15.68, 26.3) | 21.54 (16.14, 28.11) | 17.33 (14.13, 21.39) | <0.05 |
| Direct bilirubin | 1.91 (1.52, 2.41) | 1.88 (1.47, 2.46) | 1.57 (1.25, 1.99) | 1.66 (1.31, 2.11) | 1.48 (1.22, 1.81) | <0.05 |
| Gamma glutamyltransferase | 34.1 (24.3, 50.6) | 53.9 (34.5, 96.3) | 34.1 (24.2, 52.3) | 34.9 (22.1, 51.55) | 22 (16.9, 31.3) | <0.05 |
| **Renal function:** | | | | | | |
| Creatinine | 79.8 (71.7, 88.8) | 77.1 (67.2, 87.4) | 76.1 (64, 95) | 81.8 (63.3, 90.25) | 63.8 (57.3, 71.5) | <0.05 |
| Sodium in urine | 76.4 (49.5, 108.6) | 74.9 (48.9, 106) | 69 (43.5, 96.3) | 57.4 (35.65, 86.15) | 53.2 (34.3, 77.7) | <0.05 |
| Urea | 5.73 (4.94, 6.63) | 5.69 (4.83, 6.66) | 6.08 (5, 7.79) | 5.94 (5.05, 6.52) | 5.41 (4.61, 6.23) | <0.05 |
| Urate | 354.8 (310.5, 402.6) | 370.1 (312.3, 428.42) | 354.44 (297.4, 429) | 358.9 (312.05, 402.7) | 269.3 (230.3, 311.2) | <0.05 |
| **Cholesterol markers:** | | | | | | |
| Cholesterol | 5.16 (4.4, 5.88) | 5.3 (4.44, 6.18) | 5.09 (4.29, 5.95) | 5.13 (4.41, 6.07) | 5.8 (5.1, 6.53) | <0.05 |
| HDL cholesterol | 1.26 (1.08, 1.46) | 1.17 (0.99, 1.43) | 1.24 (1.04, 1.45) | 1.27 (1.16, 1.56) | 1.6 (1.4, 1.84) | <0.05 |
| Triglycerides | 1.6 | 2.24 | 1.82 | 1.78 | 1.32 | <0.05 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | (1.14, 2.18) | (1.44, 3.4) | (1.32, 2.5) | (1.27, 2.44) | (0.99, 1.76) | |
| **Sex-related markers:** | | | | | | |
| SHBG | 42.75 (33.16, 53.73) | 37.27 (26.76, 50.01) | 38.3 (28.58, 50.66) | 46.57 (37.98, 59.25) | 61.46 (48.69, 78.23) | <0.05 |
| Testosterone | 11.03 (8.4, 13.72) | 9.28 (6.05, 12.16) | 5.03 (1.09, 9.64) | 9.8 (1.34, 13.18) | 1.17 (0.76, 2.48) | <0.05 |
| **ADDITIONAL INVESTIGATIVE VARIABLES:** | | | | | | |
| **Demographics:** | | | | | | |
| Age at recruitment | 63 (59,67) | 62 (58,66) | 64 (60,67) | 63 (60.5,67) | 63 (60,67) | <0.05 |
| Sex [Male] | 16,842 (87.1%) | 3,535 (78.9%) | 2,216 (48.6%) | 30 (63.8%) | 661 (8%) | <0.05 |
| Waist circumference | 98 (91,106) | 102 (94,111) | 100 (91,110) | 100 (91.75,105.5) | 85 (77,93) | <0.05 |
| Hip circumference | 104 (100,110) | 107 (101,113) | 107 (101,116) | 106 (101.75,113) | 103 (97,109) | <0.05 |
| Standing height | 175 (169,180) | 174 (168,180) | 168 (161,175) | 173 (163.25,180) | 164 (159,169) | <0.05 |
| Weight | 86.2 (77.2,96.7) | 90.2 (80.3,102.5) | 85.4 (74.4,98.8) | 86.5 (73.2,95.3) | 70.9 (63.4,80.33) | <0.05 |
| BMI | 28.15 (27.03,29.85) | 29.79 (28.45,31.64) | 30.26 (28.7,32.26) | 28.9 (27.47,29.41) | 26.36 (25.08,28.12) | <0.05 |
| **Activity level:** | | | | | | |
| Summed minutes activity | 100 (50,180) | 90 (40,160) | 80 (30,150) | 120 (62.5,180) | 105 (55,180) | <0.05 |
| MET minutes/week for vigorous activity | 160 (0,960) | 0 (0,720) | 0 (0,480) | 320 (0,960) | 120 (0,720) | <0.05 |
| **Blood pressure:** | | | | | | |
| Diastolic BP | 83 (76,91) | 84 (77,92) | 81 (74,89) | 81.5 (73,87) | 80 (73,88) | <0.05 |
| Systolic BP | 144 (131,157) | 145 (133,160) | 143 (130,157) | 145 (124.75,151.75) | 142 (128,156) | <0.05 |
| Pulse rate | 67 (59,76) | 70 (61,80.25) | 71 (63,81) | 69 (63.75,76.25) | 68 (61,76) | <0.05 |
| **Respiratory measures:** | | | | | | |
| (FEV1) | 2.99 (2.42,3.49) | 2.85 (2.26,3.39) | 2.28 (1.84,2.77) | 2.71 (2.19,3.18) | 2.27 (1.93,2.64) | <0.05 |
| PEF | 433 (334,520) | 414 (313,507.75) | 332 (258,415) | 366 (304.5,469.5) | 318 (260,375) | <0.05 |
| FEV1 Z-score | 0.57 (-0.18,1.33) | 0.77 (0.07,1.53) | 0.97 (0.22,1.73) | 0.72 (0.08,1.08) | 0.5 (-0.22,1.2) | <0.05 |
| FEV1/FVC ratio Z-score | 0.36 (-0.17,0.98) | 0.29 (-0.22,0.95) | 0.43 (-0.12,1.08) | 0.45 (-0.28,0.95) | 0.51 (0.01,1.02) | <0.05 |
| **Alcohol intake frequency:** | | | | | | |
| Daily or almost daily | 4,196 (21.7%) | 1,071 (23.9%) | 624 (13.7%) | 15 (31.9%) | 1,264 (15.3%) | <0.05 |
| 3 or 4 times a week | 3,761 (19.5%) | 794 (17.7%) | 580 (12.7%) | 5 (10.6%) | 1,277 (15.5%) | <0.05 |
| Once or twice a week | 3,665 (19%) | 801 (17.9%) | 822 (18%) | 7 (14.9%) | 1,574 (19.1%) | 0.363 |
| 1 to 3 times a month | 1,241 (6.4%) | 299 (6.7%) | 409 (9%) | 5 (10.6%) | 780 (9.5%) | <0.05 |
| Special occasions only | 1,404 (7.3%) | 329 (7.3%) | 640 (14%) | 4 (8.5%) | 977 (11.8%) | <0.05 |
| Never | 1,172 (6.1%) | 311 (6.9%) | 532 (11.7%) | 4 (8.5%) | 715 (8.7%) | <0.05 |
| **Ethnic background:** | | | | | | |
| White | 18,578 (96.1%) | 4,445 (99.2%) | 4,264 (93.5%) | 46 (97.9%) | 8,203 (99.4%) | <0.05 |
| Asian or Asian British | 157 (0.8%) | 2 (0%) | 244 (5.4%) | 1 (2.1%) | 2 (0%) | <0.05 |
| Black or Black British | 243 (1.3%) | 1 (0%) | 2 (0%) | 0 (0%) | 1 (0%) | <0.05 |
| Mixed | 72 (0.4%) | 9 (0.2%) | 15 (0.3%) | 0 (0%) | 15 (0.2%) | 0.0641 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Other ethnic group | 135 (0.7%) | 5 (0.1%) | 12 (0.3%) | 0 (0%) | 8 (0.1%) | <0.05 |
| Chinese | 36 (0.2%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | <0.05 |
| **AF and flutter diagnosis (main/secondary):** | | | | | | |
| ICD10 - AF and flutter | 11,302 (58.5%) | 2,684 (59.9%) | 2,740 (60.1%) | 27 (57.5%) | 4,213 (51%) | <0.05 |
| ICD10 - Paroxysmal AF | 3,209 (16.6%) | 696 (15.5%) | 811 (17.8%) | 7 (14.9%) | 1,835 (22.2%) | <0.05 |
| ICD10 - Persistent AF | 740 (3.8%) | 149 (3.3%) | 106 (2.3%) | 1 (2.1%) | 278 (3.4%) | <0.05 |
| ICD10 - Chronic AF | 327 (1.7%) | 80 (1.8%) | 63 (1.4%) | 0 (0%) | 100 (1.2%) | <0.05 |
| ICD10 - Typical AF | 128 (0.7%) | 31 (0.7%) | 18 (0.4%) | 0 (0%) | 39 (0.5%) | 0.1045 |
| ICD10 - Atypical atrial flutter | 43 (0.2%) | 13 (0.3%) | 13 (0.3%) | 0 (0%) | 17 (0.2%) | 0.8072 |
| ICD10 - AF and atrial flutter, unspecified | 11,455 (59.2%) | 2,723 (60.8%) | 2,678 (58.7%) | 24 (51.1%) | 4,887 (59.2%) | 0.6494 |
| **Systems (phecode categories):** | | | | | | |
| Endocrine/metabolic | 4,467 (23.1%) | 1,865 (41.6%) | 1,947 (42.7%) | 12 (25.5%) | 1,828 (22.1%) | <0.05 |
| Circulatory system | 14,062 (72.7%) | 3,559 (79.4%) | 3,783 (82.9%) | 35 (74.5%) | 5,189 (62.8%) | <0.05 |
| Respiratory | 2,991 (15.5%) | 804 (18%) | 1,200 (26.3%) | 9 (19.2%) | 1,093 (13.2%) | <0.05 |
| **Diabetes:** | | | | | | |
| Type 1 diabetes | 300 (1.6%) | 258 (5.8%) | 225 (4.9%) | 0 (0%) | 56 (0.7%) | <0.05 |
| Type 1 diabetes with ketoacidosis | 18 (0.1%) | 40 (0.9%) | 14 (0.3%) | 0 (0%) | 9 (0.1%) | <0.05 |
| Type 1 diabetes with renal manifestations | 16 (0.1%) | 13 (0.3%) | 29 (0.6%) | 0 (0%) | 2 (0%) | <0.05 |
| Type 1 diabetes with ophthalmic manifestations | 58 (0.3%) | 61 (1.4%) | 41 (0.9%) | 0 (0%) | 15 (0.2%) | <0.05 |
| Type 1 diabetes with neurological manifestations | 26 (0.1%) | 36 (0.8%) | 29 (0.6%) | 0 (0%) | 5 (0.1%) | <0.05 |
| Diabetes type 1 with peripheral circulatory disorders | 13 (0.1%) | 13 (0.3%) | 23 (0.5%) | 0 (0%) | 3 (0%) | <0.05 |
| Type 2 diabetes | 3,400 (17.6%) | 1,620 (36.2%) | 1,462 (32.1%) | 9 (19.2%) | 639 (7.7%) | <0.05 |
| Type 2 diabetes with ketoacidosis | 35 (0.2%) | 41 (0.9%) | 14 (0.3%) | 0 (0%) | 6 (0.1%) | <0.05 |
| Type 2 diabetes with renal manifestations | 66 (0.3%) | 55 (1.2%) | 103 (2.3%) | 1 (2.1%) | 8 (0.1%) | <0.05 |
| Type 2 diabetes with ophthalmic manifestations | 326 (1.7%) | 244 (5.5%) | 226 (5%) | 3 (6.4%) | 53 (0.6%) | <0.05 |
| Type 2 diabetes with neurological manifestations | 132 (0.7%) | 139 (3.1%) | 137 (3%) | 2 (4.3%) | 17 (0.2%) | <0.05 |
| Diabetes type 2 with peripheral circulatory disorders | 122 (0.6%) | 109 (2.4%) | 110 (2.4%) | 0 (0%) | 10 (0.1%) | <0.05 |
| **Hypertension:** | | | | | | |
| Essential hypertension | 12,827 (66.3%) | 3,334 (74.4%) | 3,571 (78.3%) | 31 (66%) | 4,679 (56.7%) | <0.05 |
| Other hypertensive complications | 34 (0.2%) | 5 (0.1%) | 42 (0.9%) | 0 (0%) | 5 (0.1%) | <0.05 |
| **Cardiovascular disease:** | | | | | | |
| Myocardial infarction | 3,684 (19.1%) | 972 (21.7%) | 1,027 (22.5%) | 11 (23.4%) | 850 (10.3%) | <0.05 |
| Other forms of chronic heart disease | 2 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0.7735 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Congestive heart failure (CHF) NOS | 1,891 (9.8%) | 539 (12%) | 727 (15.9%) | 2 (4.3%) | 601 (7.3%) | <0.05 |
| Chronic pulmonary heart disease | 500 (2.6%) | 168 (3.8%) | 209 (4.6%) | 0 (0%) | 228 (2.8%) | <0.05 |
| Heart failure NOS | 2,452 (12.7%) | 662 (14.8%) | 834 (18.3%) | 8 (17%) | 724 (8.8%) | <0.05 |
| Coronary atherosclerosis | 92 (0.5%) | 28 (0.6%) | 30 (0.7%) | 1 (2.1%) | 12 (0.2%) | <0.05 |
| **Peripheral vascular disease:** | | | | | | |
| Peripheral vascular disease, unspecified | 934 (4.8%) | 316 (7.1%) | 408 (9%) | 2 (4.3%) | 251 (3%) | <0.05 |
| Other specified peripheral vascular diseases | 8 (0%) | 4 (0.1%) | 6 (0.1%) | 0 (0%) | 5 (0.1%) | 0.2495 |
| **Pulmonary hypertension:** | | | | | | |
| Primary pulmonary hypertension | 193 (1%) | 53 (1.2%) | 84 (1.8%) | 0 (0%) | 73 (0.9%) | <0.05 |
| **Stroke:** | | | | | | |
| Hemiplegia | 598 (3.1%) | 167 (3.7%) | 208 (4.6%) | 1 (2.1%) | 240 (2.9%) | <0.05 |
| **Liver disease:** | | | | | | |
| Liver abscess and sequelae of chronic liver disease | 165 (0.9%) | 117 (2.6%) | 60 (1.3%) | 0 (0%) | 31 (0.4%) | <0.05 |
| Alcoholic liver damage | 155 (0.8%) | 147 (3.3%) | 65 (1.4%) | 0 (0%) | 12 (0.2%) | <0.05 |
| Other chronic non-alcoholic liver disease | 654 (3.4%) | 300 (6.7%) | 268 (5.9%) | 1 (2.1%) | 218 (2.6%) | <0.05 |
| Other disorders of the liver | 387 (2%) | 135 (3%) | 125 (2.7%) | 0 (0%) | 161 (2%) | <0.05 |
| **Kidney disease:** | | | | | | |
| End-stage renal disease | 155 (0.8%) | 54 (1.2%) | 247 (5.4%) | 0 (0%) | 28 (0.3%) | <0.05 |

*Table 8. Characteristics of the participants per phenotype of AF in an ICU population using the MIMIC-IV database. As in Table 6, medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (as percentages) were calculated for the categorical variables. As in Table 7, shades of red/blue were used per variable to illustrate differences between lower and higher values. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.*

| Variable name | Phenotype 1 | Phenotype 2 | Phenotype 3 | Phenotype 4 | p-value |
|---|---|---|---|---|---|
| **MODELLING VARIABLES:** | | | | | |
| **Diabetes:** | | | | | |
| Glucose | 139 (115.22, 184.71) | 136.15 (118.3, 160.6) | 127.94 (119.24, 137.89) | 134.04 (114.84, 159.54) | < 0.05 |
| **Bone profile:** | | | | | |
| Phosphate | 4.57 (3.65, 5.65) | 3.5 (3, 4.1) | 3.38 (2.99, 3.79) | 3.36 (2.8, 3.86) | < 0.05 |
| **Oxygenation:** | | | | | |
| Oxygen saturation | 96.08 (93.88, 97.75) | 96.22 (94.67, 97.65) | 96.36 (93.66, 97.85) | 97.03 (95.37, 98.4) | < 0.05 |
| Respiratory rate | 19.25 (16.9, 22.32) | 20.5 (17.97, 23.09) | 16.98 (15.7, 18.62) | 18.46 (16.5, 20.63) | < 0.05 |
| FiO2 | 57.5 (50, 66.27) | 56.07 (50, 62.16) | 58.33 (52.08, 64.58) | 53.57 (46.15, 57.54) | < 0.05 |
| PEEP | 6.45 (5.08, 8.11) | 6.37 (5.1, 7.68) | 5.05 (5, 5.94) | 5.38 (5, 6.24) | < 0.05 |
| Partial pressure of oxygen | 109 (72.0, 150.97) | 114.79 (85.64, 139.45) | 168.96 (143.3, 205.26) | 133.93 (111.1, 152.19) | < 0.05 |

| | | | | | |
|---|---|---|---|---|---|
| Haemoglobin | 9.62 (8.58, 10.79) | 10.5 (9.14, 11.91) | 9.92 (9.23, 10.79) | 11.81 (10.4, 13.2) | < 0.05 |
| **Respiratory/metabolic markers:** | | | | | |
| pH | 7.29 (7.17, 7.36) | 7.32 (7.15, 7.38) | 7.37 (7.35, 7.4) | 7.22 (7.08, 7.38) | < 0.05 |
| Anion Gap | 17 (14.0, 20.21) | 13.83 (12, 15.97) | 11.67 (10, 13.08) | 14 (12.16, 15.94) | < 0.05 |
| Lactate | 2.33 (1.6, 3.39) | 1.9 (1.4, 2.62) | 2.14 (1.62, 2.78) | 1.6 (1.16, 2.12) | < 0.05 |
| **Cardiac markers:** | | | | | |
| Heart rate | 83.39 (73.36, 93.75) | 85.03 (76.83, 96.06) | 80.46 (75.33, 85.81) | 75.86 (68.26, 86.2) | < 0.05 |
| Capillary refill | 0.03 (0, 0.42) | 0 (0, 0.02) | 0 (0, 0) | 0 (0, 0) | < 0.05 |
| Diastolic BP | 56 (50.34, 61.62) | 58.21 (52.49, 63.98) | 55.19 (50.21, 59.87) | 65.62 (59, 72.69) | < 0.05 |
| Systolic BP | 109.24 (101.74, 119.08) | 110.19 (103.3, 118.4) | 111.01 (105.38, 117.38) | 131.09 (121.45, 143.2) | < 0.05 |
| **Clotting markers:** | | | | | |
| Prothrombin time | 16.53 (13.95, 22.29) | 14.65 (13.02, 16.7) | 14.2 (13.2, 15.37) | 13.1 (12.17, 14.3) | < 0.05 |
| Platelet count | 148.42 (102.73, 223.19) | 187.79 (139.22, 254.14) | 146.29 (120.05, 185.56) | 197 (151.08, 245.71) | < 0.05 |
| **Renal function:** | | | | | |
| Creatinine | 2.12 (1.3, 3.7) | 1 (0.75, 1.33) | 0.9 (0.73, 1.16) | 0.9 (0.7, 1.2) | < 0.05 |
| **Electrolytes:** | | | | | |
| Magnesium | 2.11 (1.91, 2.4) | 2 (1.8, 2.25) | 2.4 (2.19, 2.7) | 2 (1.8, 2.13) | < 0.05 |
| Potassium | 4.49 (4.05, 4.92) | 4 (3.83, 4.55) | 4.33 (4.11, 4.57) | 4.05 (3.74, 4.33) | < 0.05 |
| **Other:** | | | | | |
| Temperature | 57.5 (36.45, 36.97) | 56.07 (36.62, 37.11) | 58.33 (36.52, 36.85) | 53.57 (36.67, 37.24) | < 0.05 |
| **ADDITIONAL INVESTIGATIVE VARIABLES:** | | | | | |
| **Demographics:** | | | | | |
| Age | 71.0 (63.0, 81.0) | 73.0 (64.0, 82.0) | 74.0 (67.0, 80.0) | 75.0 (65.75, 84.0) | < 0.05 |
| Sex | 405 (63.4%) | 453 (57.2%) | 563 (62.3%) | 206 (57.2%) | 0.3317 |
| Height | 172.86 (162.78, 177.9) | 170.09 (162.72, 177.9) | 170.09 (162.78, 177.9) | 172.86 (162.78, 180.17) | 0.2896 |
| Weight | 83.93 (69.84, 98.29) | 81.42 (65.9, 99.36) | 83.05 (70.33, 95.92) | 79.79 (65.85, 95.97) | 0.0768 |
| **Ethnicity:** | | | | | |
| White | 434.0 (67.9%) | 589.0 (74.4%) | 700.0 (77.4%) | 248.0 (68.9%) | 0.1263 |
| Other ethnic group | 117.0 (18.3%) | 128.0 (16.2%) | 136.0 (15.0%) | 72.0 (20.0%) | 0.1785 |
| Black | 51.0 (8.0%) | 43.0 (5.4%) | 23.0 (2.5%) | 21.0 (5.8%) | < 0.05 |
| Hispanic | 16.0 (2.5%) | 16.0 (2.0%) | 28.0 (3.1%) | 8.0 (2.2%) | 0.5508 |
| Asian | 21.0 (3.3%) | 16.0 (2.0%) | 17.0 (1.9%) | 11.0 (3.1%) | 0.2400 |
| **Glasgow Coma Scale (GCS):** | | | | | |
| GCS eye-opening | 2.83 (1.75, 3.83) | 3.29 (2.29, 4.0) | 2.5 (1.67, 3.08) | 3.29 (2.34, 4.0) | < 0.05 |
| GCS motor response | 5.0 (3.06, 6.0) | 5.67 (4.28, 6.0) | 4.12 (2.79, 4.75) | 5.79 (4.67, 6.0) | < 0.05 |
| GCS verbal response | 2.04 (1.0, 4.62) | 3.33 (1.0, 5.0) | 2.25 (1.0, 3.5) | 3.25 (1.0, 5.0) | < 0.05 |
| **Ventilation:** | | | | | |
| Non-Invasive ventilation | 56.0 (8.8%) | 75.0 (9.5%) | 54.0 (6.0%) | 24.0 (6.7%) | < 0.05 |

| | | | | | |
|---|---|---|---|---|---|
| Invasive ventilation | 485.0 (75.9%) | 557.0 (70.3%) | 852.0 (94.2%) | 222.0 (61.7%) | < 0.05 |
| **Outcomes:** | | | | | |
| Time to AF diagnosis (hours) | 59.0 (41.0, 94.0) | 52.0 (36.0, 91.0) | 49.0 (37.0, 70.0) | 55.0 (36.75, 89.0) | < 0.05 |
| In-hospital length of stay (hours) | 296.32 (180.18, 498.3) | 262.41 (169.22, 427.41) | 228.08 (159.62, 340.88) | 246.62 (161.07, 413.97) | < 0.05 |
| In-ICU length of stay (hours) | 143.89 (82.93, 264.94) | 112.97 (70.99, 211.78) | 98.33 (69.63, 148.69) | 110.16 (69.28, 212.26) | < 0.05 |
| Death after ICU (hours) | 26.57 (16.46, 1021.61) | 183.27 (17.17, 2350.5) | 1558.35 (21.98, 10015.99) | 394.49 (18.8, 3513.18) | < 0.05 |
| Death after hospital discharge (hours) | 17.5 (8.5, 849.0) | 20.25 (10.1, 2106.35) | 1330.07 (16.3, 9930.92) | 27.81 (12.62, 3271.6) | < 0.05 |
| Death after hospital discharge (days) | 0.73 (0.35, 35.38) | 0.84 (0.42, 87.76) | 55.42 (0.68, 413.79) | 1.16 (0.53, 136.32) | < 0.05 |
| In-hospital mortality | 245.0 (38.3%) | 191.0 (24.1%) | 60.0 (6.6%) | 71.0 (19.7%) | < 0.05 |
| In-ICU length of stay of 3+ days | 526.0 (82.3%) | 587.0 (74.1%) | 665.0 (73.6%) | 262.0 (72.8%) | 0.1785 |
| In-ICU length of stay of 7+ days | 274.0 (42.9%) | 257.0 (32.4%) | 186.0 (20.6%) | 123.0 (34.2%) | < 0.05 |
| Mortality after hospital discharge within 30 days | 301.0 (47.1%) | 245.0 (30.9%) | 77.0 (8.5%) | 88.0 (24.4%) | < 0.05 |
| Mortality after hospital discharge Within 365 days | 368.0 (57.6%) | 325.0 (41.0%) | 121.0 (13.4%) | 122.0 (33.9%) | < 0.05 |
| Mortality after hospital discharge after 365 days | 36.0 (5.6%) | 49.0 (6.2%) | 43.0 (4.8%) | 24.0 (6.7%) | 0.5042 |
| AKI | 161.0 (25.2%) | 159.0 (20.1%) | 184.0 (20.4%) | 41.0 (11.4%) | < 0.05 |
| ARDS | 33.0 (5.2%) | 58.0 (7.3%) | 37.0 (4.1%) | 46.0 (12.8%) | < 0.05 |

A more detailed breakdown of the key features for each phenotype derived from the AF participants in the UK Biobank database is as follows:

Phenotype 1 (n =19,335)

The largest phenotype identified as part of the analysis, containing approximately 53% of the participants, shows the highest haematocrit percentage and mean corpuscular volume, as well as the highest levels of sodium in urine and direct bilirubin. This phenotype also contained the highest proportion of male participants (87.1%) which would be expected as the highest testosterone levels are seen in this phenotype. Additionally, this phenotype is categorised by participants that have the lowest pulse rate whilst having the highest peak expiratory flow rate and FEV1. This is also the most ethnically diverse phenotype, with it containing the height percentage of participants categorised as Black or Black British, Chinese, Mixed and Other ethnic group. It also features the highest alcohol consumption of the 5 phenotypes. A final distinguishing feature is that this phenotype has the highest levels of persistent AF across all phenotypes.

Phenotype 2 (n =4,480)

Characterised by having the youngest participant age group, participants in this phenotype were also the most likely to have chronic AF. Participants also showed the highest mean platelet distribution width, platelet distribution width, urate levels, albumin, alanine aminotransferase and

gamma glutamyl transferase. Participants in this phenotype also showed a high, albeit not the highest, drinking levels with 2nd highest saying they drink daily/almost daily and the least amount of people saying they drink infrequently at special occasions only.

Other key characteristics of this phenotype are the highest levels of triglycerides, lowest HDL cholesterol and SHBG levels. These features appear to be captured in the investigative variables with participants in this phenotype having the highest diastolic and systolic blood pressure, highest rates of type 1 and 2 diabetes and the highest rate of liver conditions.

Phenotype 3 (n =4,561)

Phenotype consists of the oldest participants whilst also being the most balanced concerning sex 51.4% being female. Participants in this phenotype are characterised by the highest values for neutrophil count and c-reactive protein and the lowest values for lymphocyte and monocyte percentages. They also have the highest levels of glycated haemoglobin and the highest urea levels. However, a higher level of urea is common in older people which may explain this reading. A key distinguishing factor is the participants have the highest hemiplegia levels, indicating that they are the most at risk of stroke.

The participants clustered in this phenotype also have the highest BMI and the lowest amount of weekly activity. The lowest alcohol intake levels are seen within the phenotype shown by having the highest percentage of participants that never drink or only drink on special occasions. The lowest number of white participants if seen in the phenotype, with it also having the highest percentage of Asian or Asian British participants. Regarding additional diseases, these participants are most likely to be diagnosed with additional comorbidities, with the phenotype showing the highest levels of circulatory system, endocrine/metabolic and respirator conditions. More specifically, the phenotype shows the highest rates of cardiovascular disease, peripheral vascular disease, hypertension, and kidney disease.

Phenotype 4 (n =47)

By far in the way the smallest phenotype, consisting of only a singular micro cluster. The participants here showed the lowest values for mean platelet volume, mean reticulocyte, and mean sphered cell volume whilst having the highest creatine value. Although unmentioned up until this point, this phecode has the most diverse genetic makeup, with 36 out of the 40 values being either a maximum or minimum value. This indicates that there may be some genetic difference in this between this phenotype when compared to the other four. These participants had the highest amount of weekly activity but the joint highest systolic blood pressure, matching that seen in phenotype 2. This is the only phenotype with no participants diagnosed with chronic AF

participants, however they are the most likely to be diagnosed with a neoplasm condition, myocardial infarction, and coronary atherosclerosis.

Phenotype 5 (n =8,257)

This phenotype is defined as consisting almost entirely of white female participants (92% female participants with 99.4% being White or White British). Across the board, the participants in this phenotype show the lowest risk factors in every category in comparison to the other 5 phecodes. In addition to this, the participants also have the lowest BMI and lowest chance of having an additional comorbidity across all categories considered here. One feature that does stand out however is that participants in this phenotype are the most likely to be diagnosed with paroxysmal AF.

As with the UK Biobank phenotypes, a detailed breakdown of the key features for each phenotype derived from the AF participants in the MIMIC-IV database is as follows:

Phenotype 1 (n = 1,705)

This is the largest phenotype identified out of the 4, consisting of 63% of patients. Patients in this phenotype are the youngest of the four, with a median age of 71. These patients also showed the highest anion gap, capillary refill, glucose, lactate, PEEP, phosphate, potassium, prothrombin time, and creatine, whilst also having the lowest haemoglobin, partial pressure of oxygen, oxygen saturation, systolic BP and joint lowest temperature with phenotype 3. The patients in this phenotype have the lowest GCS verbal response score, with highest rates of AKI.

Patients in this phenotype had the longest time to AF diagnosis once admitted to the ICU, as well as longest time spent in hospital and in the ICU. They also presented the highest rate mortality both in-hospital and after discharge, with the lowest time between mortality and ICU and hospital discharge.

Phenotype 2 (n = 631)

With regards to the variables used for modelling, patients in this phenotype had the least amount of standout characteristics, with the exceptions being that they presented the highest heart rate and respiratory rate and the joint lowest magnesium levels (shared with phenotype 4). Patients in this did however have the highest rate of non-invasive ventilation being used as well as the highest GCS scores for eye-opening and verbal response.

Phenotype 3 (n = 250)

This phenotype is defined by patients having the lowest values for anion gap, diastolic BP, glucose, PEEP, phosphate, platelet count, respiratory rate, creatine, and temperature, whilst also

having the highest partial pressure of oxygen, pH, magnesium and fraction inspired oxygen. Outside of the modelling variables, patients were also the lowest on the GCS eye-opening and motor response scales and had the highest rates of invasive ventilation.

As opposed to phenotype 1, patients in this phenotype where diagnosed the quickest, and had the shortest length of stay both in-hospital and in the ICU. Furthermore, they had the lowest mortality rate both in-hospital and after discharge, with the longest time between discharge and mortality. Finally, patients here showed the lowest rates or ARDS.

Phenotype 4 (n = 109)

The final phenotype identified was the smallest of the 4, defined by contained the oldest patient's cohort. The patients were also characterised with having the highest oxygen saturation, systolic and diastolic BP, temperature, haemoglobin, and platelet count. They also displayed the lowest fraction inspired oxygen, heart rate, lactate, magnesium, pH, phosphate, potassium, prothrombin time and creatine. Other defining features of this phenotype are the highest scores on GCS eye-opening and motor response scales, highest rates of ARDS and lowest rates of AKI and invasive ventilation.

### 5.3.6. Interpreting the visualisations

The membership maps show us which participants share the same cluster indicating that they share similar features. To unlock deeper insights, superimposing modelling data onto the membership maps provides a better understanding of why patients/participants were clustered in such a way (Figures 19 and 20). Extra insights can be learnt by superimposing post-hoc data, unseen during modelling. One example from the UK-Biobank cohort relates to sex-related markers, specifically testosterone and SHBG levels. By assessing their respective reference vectors, individuals with higher testosterone and lower SHBG tended to be in the middle and top-

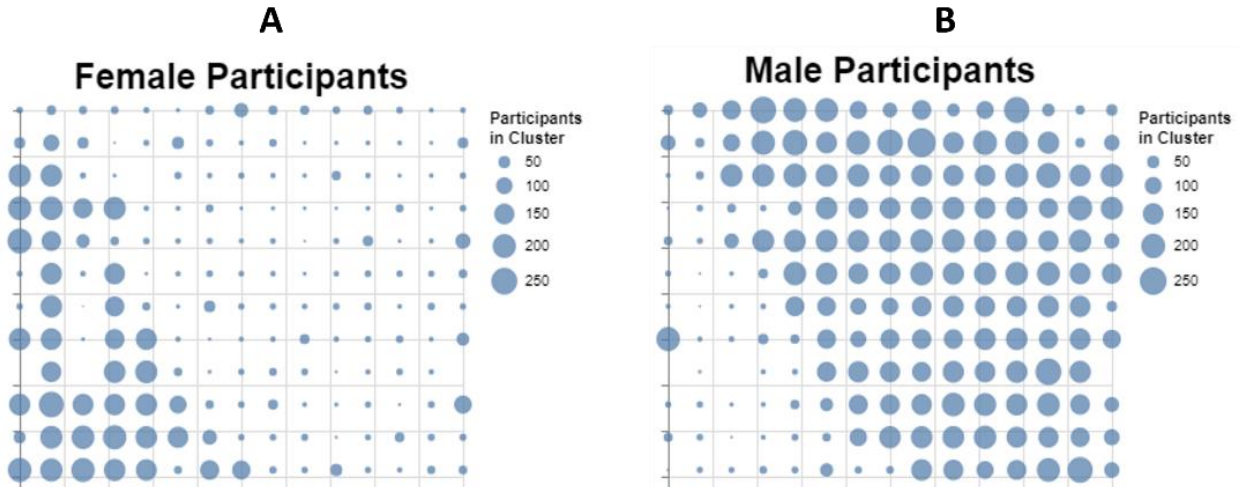



*Figure 23. Membership map generated by GTM stratified by the sex of the participant. A) Only female participants; B) Only male participants.*

right sections of the membership map. In contrast, those with heightened SHBG and lower testosterone were clustered towards the bottom left. Given that testosterone levels are generally higher in males [149], and SHBG levels are typically elevated in females [150], we can deduce that the membership map effectively delineated male and female participants during clustering. This can be seen in Figure 20(A), where we visually represent the participants' sex (the bluer area in Figure 20(A) predominantly corresponds to males), and in Figure 23, which shows the membership map stratified by sex.

### 5.4. Discussion

Using our novel AI methodology, we have identified and characterised clinical phenotypes of AF across diverse patient populations, which could facilitate the tailoring of prevention and treatment programs specific to each phenotype.

The principal findings of this chapter are: (i) The proposed AI-based methodology showed its ability to derive meaningful clinical phenotypes of AF in the general and critical care populations. (ii) Our approach is probabilistic, offering advantages such as the ability to handle uncertainty, robustness to noise, more specific patient profiles, and the ability to uncover hidden subgroups, contributing to more robust patient stratification and visualising complex high-dimensional data in a more interpretable lower-dimensional space, enhancing understanding.

#### 5.4.1. Meaningful data representation using GTM

Identifying clinical phenotypes of diseases using methods like hierarchical clustering (specifically Ward's minimum variance method and complete linkage with Gowers distance) and k-prototype used in previous phenotyping studies [126–129], may not always be the best option for several reasons: 1) Clinical data often contains diverse information, and these methods may not effectively capture the complexity of relationships within the data, and they may also be influenced by outliers or noise. 2) In clinical phenotyping, diseases may exhibit considerable heterogeneity [126–129], however hierarchical clustering assumes that data points within a cluster are homogeneous. 3) High-dimensional clinical data may pose challenges for hierarchical clustering and k-prototype methods for interpreting results, which in the context of clinical phenotypes may render unintuitive. 4) In the case of k-prototype, it can be sensitive to the choice of initial cluster centroids and may converge to local minima. 5) Clinical data often includes a mix of continuous and categorical variables. Some clustering methods, like k-prototype, handle both types, but the integration of different variable types can be challenging and may not fully capture the information. 6) Results obtained from these methods may not generalise well across different datasets or populations due to variations in data characteristics [128]. 7) They lack

probabilistic foundations and hence are not specifically designed to handle such levels of uncertainty [46,47].

Alternative approaches, such as probabilistic or ensemble methods, may provide more robust and interpretable clinical phenotypes. Our approach involves deriving micro-clusters using a probabilistic method (i.e. GTM), followed by hierarchical clustering to identify macro-clusters, i.e. the phenotypes. The latter differs from previous studies as the hierarchical methods were applied to the reference vectors from a probabilistic model rather than the original data space, which makes the clusters more stable and resilient to data uncertainty. Our use of GTM often provides highly interpretable representations as it explicitly models clusters and prototypes, offering insights into the underlying structure of the data. The membership map produced by GTM captures the underlying relationships and clusters within the data by mapping data points to these prototypes. This enables comprehensible and interpretable representations of complex data, aiding in knowledge extraction and facilitating insights that might otherwise remain hidden in the original high-dimensional space. Indeed, GTM has been applied in diverse real-world situations spanning various domains such as bioinformatics [111,112]; the financial sector [115]; and more recently also in modelling freedom of expression (Chapter 4)[81]. To the best of our knowledge, GTM has not been used before to study AF or to generate clinical phenotypes.

### 5.4.2. Clinical significance of the identified phenotypes

The identification and characterisation of clinical phenotypes of AF across diverse patient populations show potential for personalised risk assessment and prognosis. Leveraging these phenotypes could facilitate the tailoring of prevention and treatment programs specific to each phenotype.

The proposed methodology provides several advantages to extract meaningful phenotypes. First, as opposed to previous approaches [118,126,127,129,131], we define phenotypes based on a non-linear clustering approach which can capture more complex relationships. Furthermore, we can visualise the clusters, and by extension the phenotypes, and how each variable affects each cluster, which provides interpretability, crucial for validation and understanding. It also allows for a convenient method of looking at phenotype differences. For example, phenotype 2 in Figure 21(b) occupies predominantly the right side of the membership map. The reference vector for glucose in Figure 19 (top) highlights that participants in the bottom right micro-clusters have the highest glucose values when compared to the other micro-clusters. This information can be translated back to phenotype 2 to provide more context about its participants, and how risk factors may not be uniformly distributed within a given phenotype.

Another difference is in the selection of modelling variables. The phenotypes for both data cohorts were generated using only vitals and laboratory test data, as opposed to previous studies that also included demographics and medical history/comorbidity information in the modelling. This leads to each of the phenotypes having significant differences for such variables as they were used to initially stratify the data. The phenotypes generated in our analysis show significant differences with these key risk factors, but without including explicit information on these variables during modelling. Additionally, as the between-phenotype differences for variables such as demographics and comorbidities are performed post-hoc, should new data become available from variables not yet examined, their distribution between and within each phenotype can be swiftly identified.

### 5.4.3. Analysis limitations

One of the limitations of this analysis relates to the genomic principal components used for the UK-Biobank cohort, as their loadings were not available, limiting the ability to interpret them. Another limitation is related to the transferability of the derived phenotypic clusters to other cohorts of data, as they could vary across diverse populations due to genetic, environmental, and cultural differences. Additionally, differences in clinical settings, such as healthcare access, diagnostic criteria, and treatment approaches, may contribute to distinct phenotypic patterns among various patient groups. Since this chapter's main objective is to present a robust AI methodology for the derivation of AF phenotypes, this limitation can be mitigated by the derivation of specific phenotypes for different patient cohorts, as and when required. The dynamic nature of risk is also another possible limitation, as the current approach does not address how phenotypes change over time.

## 5.5. Conclusion

This chapter proposed a novel, AI-based approach for the derivation of clinically meaningful AF phenotypes. We applied it to two large cohort databases representing general and critical care populations. Our approach is probabilistic, contributing to robust patient stratification. It produces interpretable visualisation of complex high-dimensional data, enhancing understanding. It showed its ability to identify clinical phenotypes of AF, which could enable prevention and treatment programs specific to each phenotype. Our methodology can be applied to other datasets to derive clinically meaningful phenotypes of other conditions.

# 6. The Athlete's Heart and Machine Learning: A Review of Current Implementations and Gaps for Future Research

This chapter and Chapter 7 both focus on the same area of cardiovascular research known as the athlete's heart. Unlike the other clinical applications in other chapters, the application of AI techniques within the area of athlete's heart is much smaller in comparison. We therefore conducted an in-depth scoping review to gain a better understanding of the current applications and ascertain what approaches had already been taken; what clinical questions were researchers trying to address within the area; and what the gaps were for future novel research. This work in this chapter has also been published, and cited, demonstrating the value this research has added to the field [61].

## 6.1. Introduction

Heart disease is the leading cause of death worldwide, accounting for 16% of the total world's deaths in 2019 [151]. In the UK alone, around 7.6 million people are living with heart disease which causes, on average, one death every three minutes [152]. Exercise is one of the best methods for improving health and reducing cardiovascular risk factors [153]. However, extreme exercise regimes, such as those followed by athletes, cause physiological changes in the heart to help it cope with the increased demands placed upon it [154]. These physiological changes, also known as the "athlete's heart", can cause issues as they are difficult to distinguish from pathological changes, exposing athletes to sudden cardiac death [154].

Sudden cardiac death is the most common cause of death in young athletes, with current estimates placing its incidence rate between 1 in 40,000 and 1 in 80,000 athletes per year [155]. To prevent this, pre-participation screening, using techniques such as electrocardiography (ECG) and echocardiography, is used to identify the cardiovascular conditions associated with sudden cardiac death, allowing for appropriate treatments, and avoiding adverse outcomes. Although shown to be generally effective, there are still approximately 1% false positives, resulting in some athletes going undiagnosed, e.g., the cardiac arrest of Christian Eriksen at the Euro 2020 tournament and Fabrice Muamba in the FA Cup quarterfinals in 2012).

AI has rapidly grown over the last decade, with ML accounting for the majority of this growth [29]. ML techniques, powered by advances in computational performance and very large datasets, have shown great success and they frequently outperform human performance [156]. ML is commonly used in supervised and unsupervised learning tasks. Supervised ML techniques work in two parts: first, the ML algorithm is trained using input variables and labelled output variables to learn the associates between the two, then, the trained model is used to make predictions on a

test set, again where the labels of the outputs are known, to assess the performance [157]. Some examples of these methods include ANN, Random Forest, etc. Unlike supervised ML, unsupervised ML uses unlabelled data and automatically finds the key relationships and structures within the data. Two examples of such methodologies are t-distributed stochastic neighbourhood embedding (t-SNE) and principal component analysis (PCA).

The use of ML techniques applied to diagnostic investigations may prove valuable to help detect cardiac conditions in athletes, establish the risk levels, and develop an understanding of the physiological changes more accurately. ML models trained using different data modalities and data formats have been applied successfully in detecting many cardiovascular issues [158–162], showing how ML can solve a range of tasks, such as predicting mortality following a cardiac intervention [162], in specific populations of individuals [161], predicting coronary heart disease [159] and estimating the prognosis of patients with congenital heart failure [160].

The aim of this chapter, therefore, is to review the current state of ML applied to the athlete's heart by evaluating the current trends regarding the ML methodologies and approaches used within the area and determining the relevant questions and problems ML currently faces. To this end, we plan to focus the review on the following: (1) ML applications in the assessment of the athlete's heart, and (2) understanding the desire to implement ML approaches within this area of research.

## 6.2. Methods

### 6.2.1. Search strategy and selection process

To obtain the data needed to carry out the review, the Scopus and PubMed online electronic databases were searched to return the relevant literature. Table 9 outlines the criteria used to define the search term and where, within the manuscript, each term focuses. The literature returned from the searches was then reviewed and filtered by two authors, RAAB and DLO, by the titles and abstracts, and then through full-text readings, which were carried out by RAAB, so that only the studies relevant to the review were included.

*Table 9. Criteria used to build the literature search.*

| Criteria | | Term Location |
|---|---|---|
| **A** | "deep learning" OR "machine learning" OR "artificial intelligence" | Anywhere within the manuscript |
| **B** | electrocardio* OR echocardio* | Anywhere within the manuscript |
| **C** | "athletes heart" OR "athlete*" | Title, Abstract, or Keywords |

### 6.2.2. Search results

The search process is detailed in Figure 24. Based on the search criteria, 132 total studies were returned from the searches performed on the Scopus and PubMed online databases. The unique studies from these searches were subsequently extracted, which left a total of 128 studies. The titles and abstracts of these 128 studies were reviewed, resulting in 79 studies being excluded as they were deemed to be not relevant due to having a different focus area than the one specified for this review.

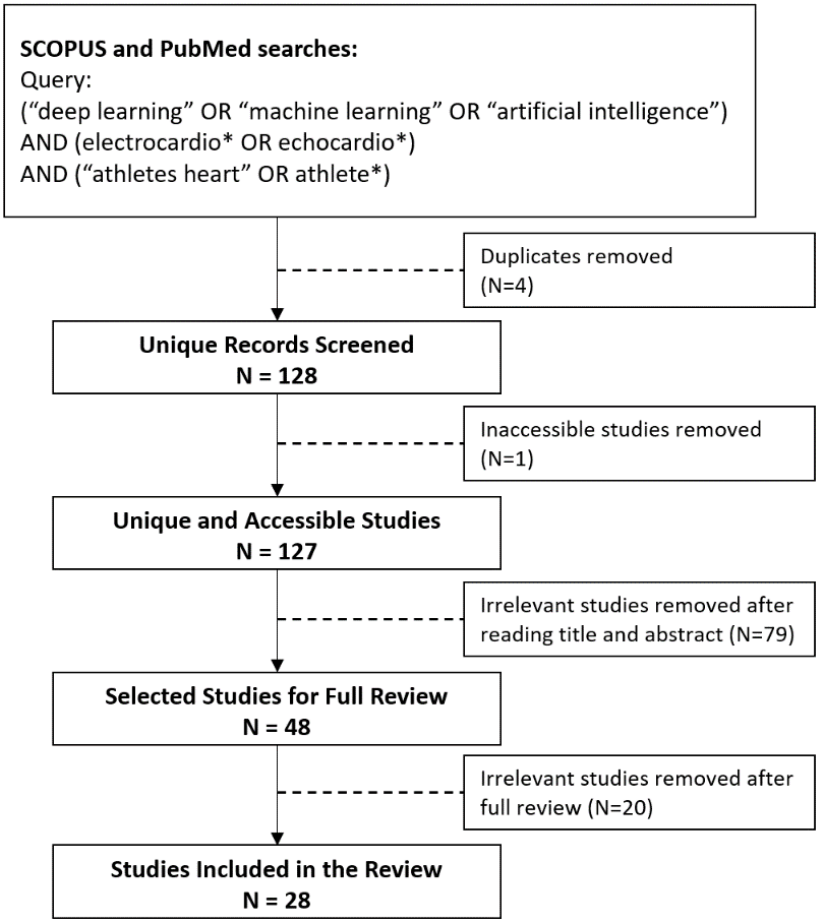



*Figure 24. Study selection flow chart.*

Of the 49 studies that remained, 1 study was excluded from the review due to issues with accessing the full manuscript, leaving 48 studies to be included for full-text readings and to form the dataset for this review. However, during the full-text readings, a further 20 studies were excluded: 16 were excluded as they were deemed to be not relevant to the review, and the other 4 were excluded due to concerns about their quality, i.e., being vague and having an unclear description of either their methodology or approach used to develop their models, how the evaluation criteria were presented, and why certain metrics were used over others. After all the

exclusions had been applied, this left a final total of 28 studies that were considered for this review [154,163,172–181,164,182–189,165–171].

### 6.3. Results

#### 6.3.1. Study subgroups

Of the 28 studies, several different approaches were taken. We clustered the studies into four subgroups: predictive modelling, reviews, wearables and others. Each study was then assigned to one of these four groups using the criteria outlined in Table 10. "Predictive Modelling" made up most of the studies with 10 (36%) [163,165,171,173,178,179,182,183,187,188] being assigned to this group. "Reviews" was the next single largest group with eight (29%) studies [166,169,172,174,175,180,185,186]. "Wearables" was the smallest single group with four (14%) studies [164,168,177,184]. The final six (21%) unassigned studies [154,167,170,176,181,189] were placed in the "Others" group as they did not meet the inclusion criteria for the previous groups.

*Table 10. Criteria for classification.*

| Group | | Criteria |
|---|---|---|
| **1** | Predictive Modelling | The main aim is to use some methodology to create a model or framework that can be used to classify data |
| **2** | Review | Consolidate existing literature in some way to construct practical guidelines or conduct a systematic review, etc. |
| **3** | Wearables | The main aim is the discussion or development of wearable technology for use as either a solely data collection enterprise or to conduct automatic analysis |
| **4** | Others | Does not fit the above criteria |

##### *6.3.1.1. Predictive modelling*

The studies within this group are focused on using methods that can be applied to a dataset to attribute one of two or more classes to each patient or participant. This has been approached in two main ways. The first and most popular type of approach implemented was to use ML to learn from the data and make predictions on what class each patient/participant should be classified as automatically. Eight of the studies [163,165,171,173,178,179,182,183] use this approach, applying ML algorithms in varying levels of complexity, from linear discriminant analysis (LDA) to ANN. A more in-depth discussion of the individual methods which were used and their respective applications can be found in the section "Machine learning approaches used".

The second approach used in the other two studies forgoes the use of ML and instead focuses on defining an algorithm tree that can be manually followed by a human user to help improve the

accuracy of their diagnoses. Vergani et al. [187] proposed a diagnostic algorithm that can be used by healthcare professionals to distinguish between a hypertrabeculation phenotype, noncompaction phenotype, and left ventricular noncompaction cardiomyopathy. Viviers et al. [188] focused on comparing the predictions made by a sports physician using a history questionnaire and a physical examination, to a technician using computer-assisted auscultation on the nature of cardiac murmurs in collegiate athletes. These two approaches are focused on classification, as with the ML-focused studies, but they have done so in a way that only utilizes human expertise.

*6.3.1.2. Reviews*

Within the data, there were eight studies which were classified as reviews. Georgijević and Andrić [174] and Lucas et al. [180] had relatively similar aims: they both reviewed the current use of different modalities in the pre-participation screening of athletes, with Georgijević and Andrić [174] looking specifically at ECG and Lucas et al. [180] concentrating on echocardiography. These studies also review the guidelines for how their respective modality should be used in the pre-participation environment and the benefits that they provide. Higgins et al. [175] had a different focus and instead reviewed the different defects that can cause sudden cardiac death in young athletes and recommended which modalities are best suited to diagnose each. Chang [169] also focused on the ECG, but their approach was to consider the positives and negatives of applying it to screening young adults, as well as a brief discussion on how AI is likely to shape the future of the heart screening of athletes. Conversely to the studies already mentioned, Beavers and Chung [166] and Seshadri et al. [185] both centred their reviews on wearables. More specifically, Beavers and Chung [166] highlighted the emerging wearable technologies and how they can be used to aid in heart assessments, with specific examples focused on minimising cardiovascular risk in athletes. Seshadri et al. [185] reviewed how the data collected from wearables had been analysed with ML to evaluate athletes' heart health, with several successful implementations reported to have achieved accuracies as high as 98% in the prediction of obstructive hypertrophic cardiomyopathy.

The remaining two studies are systematic reviews: Claudino et al. [172] focused on the sports performance and injury risk of athletes within team sports and highlighted which AI techniques have been applied within each sport, while Van Eetvelde et al. [186] looked more specifically at the ML methods which have been used in the prediction and prevention of general sports injuries. Our review differs from both Claudino et al. [172] and Van Eetvelde et al. [186] in two key areas: (1) we focus on highlighting ML applications towards the athlete's heart exclusively, instead of the wider research area of injury prevention and risk, and (2) we aim for a more comprehensive

overview of the ML approaches, and emphasise the relevant challenges that are present and how to address them through future research.

*6.3.1.3. Wearables*

The four studies in this category share the same goal: they describe the development or implementation of wearable hardware that can be used by athletes to help collect physiological data automatically. However, they differ in their individual implementations of the wearable technology, and in how the data are collected, stored, and analysed. Adetiba et al. [164] developed a smart jersey to be worn by athletes to automatically record an ECG signal. These data are then automatically passed through an ANN that has been pre-trained to identify heart defects and returns whether the result is normal or not to a smartphone application. Hussain et al. [177] proposed a fog-centric, wireless, and real-time framework for health and fitness analysis, which consists of collecting data such as ECG recordings, body movement, and posture from multiple wearables simultaneously, which is then fed into two ML models: one to predict the exercise being performed by the athlete; the other to predict the athlete's health state. Similar to the aforementioned studies, Castillo-Atoche et al. [168] described the development of a new wearable ECG with a dynamic power management strategy that then automatically passes the collected data to an ML model to detect arrhythmias in real time. Unlike Adetiba et al. [164] and Hussain et al. [177], the final study in this group by Rymarczyk et al. [184] concentrated exclusively on the development of a new type of electrode for physiological signal sensing as an alternative to a conventional gelled electrode.

*6.3.1.4. Others*

The remaining six studies do not match any of the criteria for the three main groups. Instead, these are individual pieces of research that provide a different overview of the athlete's heart. Chatzakis et al. [170] focused on developing an electronic health record, with a built-in decision support system, to support paediatric cardiovascular disease screening. Dockerill et al. [189] utilised a case series approach to assess the hearts of 27 runners before and after an extreme running event whilst documenting the changes in the cardiac structure caused by an acute bout of exercise. Similarly, Kerkhof et al. [154] investigated the changes in the heart of a select group: three division one undergraduate crew athletes explored the use of 'focused' echocardiography in screening athletes to assess their heart health and function.

The studies by Bernardino et al. [167], Huang et al. [176], and Mlynczak and Krysztofiak [181] bring unique approaches. Bernardino et al. [167] used cardiac magnetic resonance imaging data for athletes and non-athletes and applied several techniques, such as statistical shape analysis and dimensionality reduction, to highlight the areas of the heart that underwent a remodelling due

to endurance exercise (more details on the methods used are discussed in the section on the "Machine learning approaches used"). Huang et al. [176] is the only study to leverage unsupervised clustering to investigate the validity of sport-specific adaptions in athletes' hearts (the methods are further discussed in the section on the "Machine learning approaches used"). Mlynczak and Krysztofiak [181] focused on discovering causal relationships between cardiovascular and respiratory variables in elite athletes whilst they were supine and standing, aimed at developing appropriate training plans.

### 6.3.2. Data modalities used for athlete's heart assessment

Within this review, a data modality refers to the type of data collected. There are various modalities mentioned within the studies being reviewed, from images to signal data. There are examples of these being used as a sole modality as well as examples where information from multiple modalities have been used to evaluate the heart, with the splits for all the modalities mentioned displayed in Figure 25.

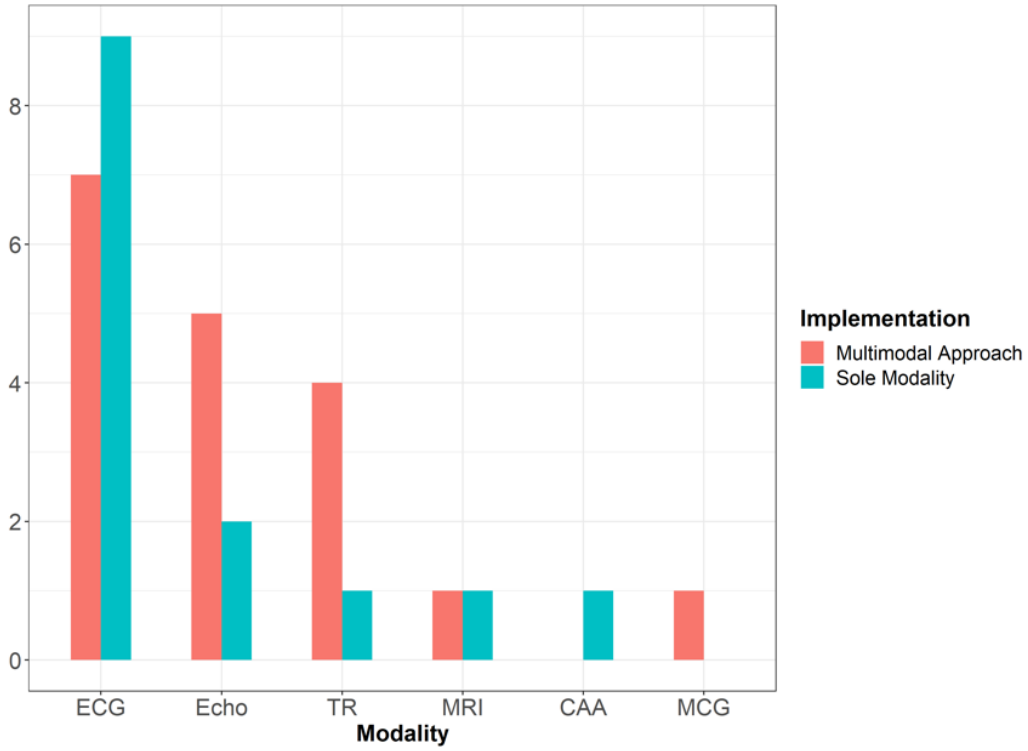



*Figure 25. Displays the number of times each modality was mentioned within the studies. It also displays how often the modality was implemented on its own, or in conjunction with another modality. Abbreviations: ECG = electrocardiogram, Echo = echocardiogram, TR = tabular records, MRI = cardiac magnetic resonance imaging, CAA = computer-assisted auscultation, MCG = magnetocardiogram.*

Our review highlighted that only 23 of the 28 studies mentioned which modality, or a combination of modalities, were used to either review or generate their dataset. The most commonly used was an ECG, with it listed in 16 of the studies, and it is the sole modality used in 9 of the studies. This is expected due to it being able to detect several conditions associated with

sudden cardiac death in athletes, such as hypertrophic cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy, myocarditis, dilated cardiomyopathy, brigade syndrome, long QT syndrome, and Wolff-Parkinson-White syndrome [175]. The use of the ECG as part of athletes' screening is recommended by associations worldwide, including the European Society of Cardiology (ESC) and the International Olympic Committee, highlighting its widespread application within the literature [174]. ECGs are also very commonly used among healthcare practitioners due to them being a cost-effective, non-invasive technique with a relatively high sensitivity for detecting underlying cardiac disease [180].

Echocardiography is the next most commonly used modality, with it being used in conjunction with other techniques in seven studies, with it being the sole modality used in two. Like with an ECG, echocardiography is widely used for many of the same reasons. It is non-invasive and, compared with other imaging modalities such as CT imaging and MRI, it is cost-effective [175]. It also plays a crucial role in diagnosing some conditions where the ECG is less sensitive such as coronary anomalies, and dilated cardiomyopathy [175]. Echocardiography also yields positive results when used in conjunction with the information generated from other sources, such as ECGs, in a multimodal approach [180].

A widely adopted modality is tabular records, which consolidate diverse sources of information concerning the patient/participant such as their age, sex, and race. This modality was referred to in five studies, with it appearing as the sole modality once. Rahman et al. [183] gave a compelling reason as to why tabular records should not be used as a sole modality in regard to the evaluation of athletes' hearts. Their use of the tabular information taken from the American Heart Association questionnaire for classification was not able to perform as well as a cardiologist that had both ECG and echocardiographic data available. However, it does serve an important purpose, as certain demographics such as age, race, and sex have already been shown to affect the heart differently, so ignoring this information may lead to overlooking a key insight. This point is further supported by Narula et al. [182], whereby using information derived from both tabular records and the echocardiogram, they built an accurate predictive model (the specific model performances, with metrics, can be found in the section on the "Machine learning approaches used").

Other modalities are referenced; however, they are used less frequently than the three most popular modalities: electrocardiography, echocardiography, and tabular records, discussed above. CMRI is referenced twice [167,187], and computer-assisted auscultation [188] and magnetocardiography [179] are both mentioned once. The reasons for this trend likely lie in the

already highlighted cost-effectiveness and non-invasive nature of the three popular modalities when compared to their alternatives.

A common theme throughout the studies is that in the majority of cases, the key features pre-extracted from the modality are analysed instead of the raw data itself. The features can either be extracted manually by a healthcare professional, such as the physical measurements [178,179,182], or by using a technique to generate statistical features instead [163,171]. The only study that bucks this trend was by Castillo-Atoche et al. [168], where authors developed their model on ECGs in an image format instead.

### 6.3.3. Machine learning approaches used

The application of ML has been used in 13 of the studies considered for this review. Eight of them were assigned to the "predictive modelling" group, three were assigned to the "wearables" group, and two were assigned to the "other" group. The most commonly used method was the ANN, with it being used in 5 out of the 13 studies [163,164,171,178,182]. This was then closely followed by support vector machines [171,178,182,183], used in 4 out of the 13 studies, and then random forest [171,182,183] and logistic regression [165,167,173], tested in 3 out of the 13 studies. Other techniques that were also mentioned within the literature but were less commonly used were decision trees [165,173], naïve Bayes classifiers [178,183], multiple linear regression [176], k nearest neighbours [178], LDA [179] and long-term short memory neural networks (LSTM) [177], CNN [168], and hierarchical clustering [176]. A summary of all 13 studies can be found in Table 11 which details the aims of each study along with other key information.

The main application of ML within these 13 studies is towards classifying whether a patient/participant has a particular heart disease or defect, with 8 out of the 13 having this focus. Adetiba et al. [163] used an artificial neural network to classify whether an athlete's heart is normal, or whether one of the following defects was present: tachyarrhythmia, bradyarrhythmia, or hypertrophic cardiomyopathy. This was done by extracting the ECG signals, applying a first-order statistical signal processing technique, and passing these features as inputs to train the model. The final model reported an accuracy of 90%. A subsequent study [164] from the same authors, published two years later, performed the same classification task, included feature extraction methods, and used only ANN. However, this time the data were generated by a wearable jersey they designed, reporting an accuracy of 100%.

Lombardi et al. [179] used LDA to determine whether patients with idiopathic ventricular arrhythmias with a left bundle branch block and inferior axis morphology arrhythmia originated from the aortic sinus cusps or the right ventricular outflow tract. Manually extracted features from multiple modalities were used to create the linear separation between the two classes, achieving a

final accuracy of 94.7%. The aim of Narula et al. [182] was to discriminate between hypertrophic cardiomyopathy and physiological hypertrophy in athletes. The manually extracted features from the echocardiographic scans as well as tabular records were used as the inputs to train a support vector machine, random forests, and an ANN model. The predictions from each model were taken and a voting system was used to determine the overall class of the patient. The reported performance of this ensemble method was an AUC of 0.795.

Długosz et al. [173] used different ML techniques in an attempt to address the two aims of the study, which were to use ECGs to estimate the level of cardiac troponin (cTnI) in amateur athletes as well as detect coronary artery disease (CAD) in the same cohort of patients. The cTnI levels of the athletes were recorded at several times before and after a sporting event, and CAD was confirmed in six athletes. The study attempted (unsuccessfully) to train a logistic regression model to estimate the cTnI levels. However, they were able to detect CAD successfully by training a grid search optimised decision tree using the pre-extracted features from ECGs performed on the athletes and tabular records such as their BMI and age and the blood levels of the cTnI. The best performing model achieved an AUC of 0.91.

The work by Rahman et al. [183] differs from the above three studies as it forwent any formal screening test data such as ECGs or echocardiograms and used the tabular record information collected from the American Heart Association questionnaire. It aimed to predict whether an athlete's heart was normal or not and it did this by training three models: a support vector machine, a random forest, and a naïve Bayes classifier. They performed two experiments, the first was on the whole dataset, which contained a large positive class (representing healthy hearts) bias, and another on a dataset where the positive class had been subsampled to create a biased dataset. The best results reported for these experiments were an accuracy of 0.742 using the support vector machine for the first experiment, and 0.553 using the random forest for the second experiment.

Regardless of their stated results and methodology, many of the studies referred to previously share a similar drawback: they all used a small dataset for their analyses. The size of the dataset used by Adetiba et al. [163,164] is n = 40, Lombardi et al. [179] is n = 26, with Narula et al. [182], Długosz et al. [173], and Rahman et al. [183] using larger datasets of n = 139, n = 160, and n = 470 participants, respectively. The use of small datasets can lead to problems when trying to leverage ML methods such as ANN, whereby the model will not learn the underlying relationship between the input variables and the output, potentially resulting in the model overfitting the data and reducing its ability to generalise to new, unseen data. Barbieri et al. [165] and Castillo-Atoche et al. [168] both addressed this issue by using much larger datasets for their analysis. Barbieri et al. [165] used 26,002 participants for their analysis, to classify whether an athlete is at a

cardiovascular risk or not. For this, the authors use tabular record information as well as the features manually extracted from ECGs as inputs to train and test two models, one built using a decision tree, and the other using logistic regression. The logistic regression model provided the best performance, generating an AUC of 0.78. Castillo-Atoche et al. [168] used a much larger dataset consisting of 56,542 ECG samples taken from 487 patients to automatically predict arrhythmias in athletes in real time. The ECG samples were analysed in an image format, with 55,222 samples taken from 480 subjects used for training and 1320 samples taken from 7 athletes used for the test. The training dataset was pieced together using several open-access online datasets, with the test set comprised of a manual reading taken from their discussed wearable. The model used to make the predictions was developed using a CNN and achieved an accuracy of 94.3% on the training set and an average accuracy across the seven athletes in the test set of 93.9%.

The remaining five studies that applied ML techniques have a different focus other than disease classification. Christ and Rückert [171] aimed to use ML to predict whether a participant was an athlete or not based on their ECG criteria. The authors used statistical measurements for time-domain features and discrete Fourier transforms to extract the frequency-domain features that were then used as model inputs. An ANN, a support vector machine, and a random forest model were trained and tested on the data, with the best performance coming from the random forest model which generated an accuracy of 98.1%.

Laurino et al. [178] focused on classifying the heart states in athletes, distinguishing between heart rates that were at rest and those during stressful conditions. Like with many of the approaches stated thus far, the features from the ECGs were manually extracted to be used as the dataset for this analysis. K nearest neighbours, support vector machines, naïve Bayes, and artificial neural networks were all tested, and the best result came from the artificial neural network, which successfully managed to separate the two classes with an accuracy of 0.87 and 0.86 on the training and test set, respectively.

Hussain et al. also used a similar application of ML [177] whereby they used an LSTM neural network on the waveforms generated from the heart rate, breathing rate, and heart rate variability, to predict the athletes' health state. The health state considered for the analysis were aerobic, anaerobic, V02 max, hazardous, and moderate, and their model was able to classify the athletes with an accuracy of over 97%. Hussain et al. [177] also described a second ML application, where they again used an LSTM network to predict what activity the athlete was performing. They trained four models for four different experiments, all using breathing waveform data and the ECG data as the inputs, with the best-stated predictive performance being an accuracy of over 83%.

*Table 11. Summary of studies that applied ML methods.*

| Study | Sample Size (N) | Type of Method | Problem Addressed | Performance Metrics Stated |
|---|---|---|---|---|
| Adetiba et al. [163] | 40 | ANN | Automatic heart defect detection for athletes | Accuracy = 0.9 |
| Adetiba et al. [164] | 40 | ANN | Develop a wearable ECG that can be worn by athletes to help automatically detect defects | Accuracy = 1 |
| Barbieri et al. [165] | 26,002 | Decision trees<br>Logistic regression | Classify whether an athlete is at cardiovascular risk or not | AUC = 0.78 |
| Bernardino et al. [167] | - | Logistic regression<br>Principal component analysis<br>Statistical shape analysis | Highlight areas of the heart that undergo cardiac remodelling due to endurance exercise | - |
| Castillo-Atroche et al. [168] | 56,542 samples from 487 patients | CNN | Automatically predict arrhythmias in athletes in real time | Accuracy = 0.939 |
| Christ and Rückert [171] | 22 and 9 | ANN<br>Random forest<br>Support vector machine | Predict whether a patient was an athlete or not based on ECG readings | Accuracy = 0.981 |
| Długosz et al. [173] | 160 | Decision tree<br>Logistic regression | (1) Use ECGs to estimate the level of cardiac troponin (cTnI) in amateur athletes<br>(2) Detect coronary artery disease (CAD) in athletes | AUC = 0.91 |
| Huang et al. [176] | 598 | Agglomerative hierarchical Clustering<br>Multiple regression analysis | (1) Identify athlete groups with similar characteristics<br>(2) Investigate the validity of sport-specific adaption for evaluating athlete's hearts | - |
| Hussain at al [177] | 7200 data points from 4 athletes | LSTM | (1) Predict an athlete's health state<br>(2) Predict the activity being performed by an athlete | 1) Accuracy = 0.97<br>2) Accuracy = 0.83 |
| Laurino et al. [178] | 14 and 12 | ANN<br>K nearest neighbours<br>Naïve Bayes<br>Support vector machines | Classifying heart states in athletes between those at rest and those in stressful conditions | Accuracy = 0.86 |
| Lombardi et al. [179] | 26 | Linear discriminant analysis | Determine whether patients with idiopathic ventricular arrhythmias with left bundle branch block and inferior axis morphology arrhythmia originated from the aortic sinus | Accuracy = 0.947 |

| | | | | |
|---|---|---|---|---|
| | | | cusps or the right ventricular outflow tract | |
| Narula et al. [182] | 139 | ANN<br>Random forest<br>Support vector machine | Discriminate between hypertrophic cardiomyopathy from physiological hypertrophy in athletes | AUC = 0.795 |
| Rahmen et al. [183] | 470 | Naïve Bayes<br>Random forest<br>Support vector machines | Predict whether an athlete's heart is normal or not | Accuracy 0.742 and 0.553 for experiments 1 and 2, respectively |

Huang et al. [176] are different from the former as they leveraged unsupervised learning in an attempt to find hidden clusters within the dataset. The study had two aims: (1) to explore the natural clustering of echocardiographic variables to identify athlete groups with similar characteristics; and (2) to investigate the validity of sport-specific adaption through a data-driven approach for evaluating the athlete's heart. To address the first aim, through utilising standard statistical tests such as an ANOVA and t-tests as well as multiple regression analysis, they were able to show clear training-related adaptations between the groups which were defined by using Mitchell's classification. For the second aim, the agglomerative hierarchical clustering managed to find two distinct clusters for both male and female athletes, confirming sport-specific adaptions.

The final study by Bernardino et al. [167] used a different approach and ML implementation to the other twelve studies. They presented a linear statistical shape analysis framework that looked for shape differences between the athletes and a set of control participants. This framework works by using a combination of dimensionality reduction techniques, principal component analysis, and partial least squares to reduce the high dimensional shape vectors to a latent space that contains the most relevant shape patterns. Logistic regression was then used to classify what shape patterns were the most discriminating between the two populations, and then they used this information to provide a visual representation of the changes. This framework was applied to cardiac magnetic resonance imaging for the study population which was able to highlight areas of the heart that undergo a cardiac remodelling due to endurance exercise.

There is a total of 11 years between the earliest study published by Laurino et al. [17] in 2011 and the most recent study published by Castillo-Atoche et al. [168] in 2022. Over most of this time, the implementation of machine learning was fairly straightforward: selecting a classification task, testing several techniques to find which performed the best, and reporting the results. However, more recently, the types of ML techniques which have been used have become more complex and intricate, as seen in Hussain et al. [177] being the first to leverage deep learning

methodologies in the form of an LSTM, and Castillo-Atoche et al. [168] leveraging the power of CNNs for image analysis. Additionally, the problems ML are being applied to are becoming more focused and novel, as seen in Bernardino et al. [34] and Huang et al. [176]. This indicates the beginning of a trend towards a more in-depth ML analysis being implemented within the research area.

### 6.4. Discussion

The studies evaluated as part of this review indicate that there is a clear drive within the research area of the athlete's heart to leverage ML. This is shown by 57% of the 28 studies either using ML to create a model to answer a question or solve a particular problem [14,16,19–22,25–27,34,36,37,40], or to evaluate how ML is being implemented in similar areas through review studies [172,185,186]. The most popular application of ML is in its use to generate models for classifying patients/participants to aid in diagnosing heart defects at an early stage.

The results stated in the research are very positive, showing the real benefit ML could have should it see a widespread adoption. What the studies also show is that alongside the traditional disease and heart health predictive modelling, there is also a desire to use ML to help further develop the knowledge surrounding the athlete's heart itself. This has been done by studies aimed at quantifying the magnitude of exercise volume on cardiac adaptations within athletes' hearts when compared to that of the general population.

#### 6.4.1. Limitations of current research

The use of ML is desirable in many tasks, including health care, as properly trained models can help reduce errors in diagnosis by either matching human performance [190] or even being superior in some cases [156]. Even though the ML applications in this area have shown promise, several issues could potentially slow the adoption of such techniques and limit their application in the real world. First, the vast majority of the data used in the studies that reference ML, or any of the 28 studies in the full literature reviewed, do not use an open-source dataset. This is problematic, making it difficult for external groups to assess the data to determine potential biases that were missed in the study or to validate the stated models being presented. This will likely be a difficult problem to overcome, due to the nature of the data being analysed. For the teams elite athletes compete for, having information about their players' health, or obtaining it for elite athletes from other teams, can give an unfair advantage in situations such as transfer markets [191]. Therefore, it is not in the best interests of the main collectors of athletes' data to make it publicly available and performing the data collection at the scale needed by third parties would become very costly. The only study that uses open-source data is by Castillo-Atoche et al. [168], where they fuse several open-source databases to form a training dataset. The datasets used are

the MIT-BIH Arrhythmia Database [192], ECG-ID Database [193], MIT-BIH Supraventricular Arrhythmia Database[194], MIT-BIH Atrial Fibrillation Database [195], QT Database [196], and Long Term ST Database [197] which are all hosted on PhysioNet [70]. Even though the use of open-source data is positive, this approach does not solve the issues discussed. First, there is not a clear description of how the data has been fused and pre-processed, hampering validation efforts. Additionally, the open-source data used does not contain the athlete's data used in the study, again further hindering the ability for external validation.

Another issue relating to the data is that many of the studies use small sample sizes for their analyses. This poses a problem, especially with ML applications, as it is well known that having sufficiently large data available is required not only to increase the model's performance but also to increase the generalisability of the model. Adetiba et al. [164] is an example where a classic symptom of overfitting is present, as the stated accuracy is unusually high at 100%. This, paired with the very small data size and the inability to reproduce the work due to non-public data, further supports the idea that overfitting may be present in this model. Furthermore, heart defects in athletes are generally at a low prevalence in a given population, hence a small sample size is unlikely to be fully representative of the disease which the authors are attempting to analyse. Hussain et al. [177] further demonstrated the effects imbalanced data can have on the results in the accuracy stated for the health state predictions of 97%. Even though the dataset is large, the prevalence for the class of concern, whether the health state is currently hazardous, represents only 0.085% of the dataset, with the classes aerobic and moderate accounting for 92.7% of the dataset. This causes an issue as it becomes very easy for a model to overfit and generate good performance metrics by mainly predicting the majority classes. This increases the difficulty for an ML model to fully understand trends that distinguish what separates the class of interest and reduces the likelihood of it generalising well to an unseen dataset.

As briefly mentioned previously, the analyses are mainly performed using the features extracted from the different modalities, such as electrocardiograms, as inputs for their analysis instead of the raw input itself. With the successes seen by using the raw data as inputs to develop ML models for the prediction of different heart conditions [62,63,198], it is surprising that none of the studies has attempted to implement this approach toward the athlete's heart. Additionally, restricting the input data to the pre-extracted features only means working under the assumption that the features themselves explain enough variance between the different output classes to enable accurate predictions, which may not hold true. Another potential issue arises due to either the time and associated cost of a healthcare practitioner extracting these features manually, which can

further exacerbate the small dataset issue, or using a feature extraction technique which may not fully capture all the relevant features of the original input, harming the model's performance.

A further point here is that most models which have been built have used supervised ML as the basis of the analysis. The difficulty here again is that the data are required to be labelled for supervised ML to be carried out, meaning an expert practitioner will need to analyse the data to provide an appropriate diagnosis or status to each sample, which can be costly and time-consuming. There may also be a situation where assigning labels to the data is not appropriate or even possible to do accurately, for example where a cross-sectional study was performed with no specific outcome in mind, or if the equipment needed for a gold standard diagnosis is unavailable. This problem will only be worsened by the ever-increasing volume of the data generated and could result in large numbers of datasets being underutilized, again exacerbating the issues surrounding the lack of open-source data and small sample sizes. Another problem supervised ML has in this context is, as described in the previous paragraph, the low prevalence of adverse outcomes in the athlete's heart. Having limited information on non-healthy hearts will likely impact the ability of any supervised ML to properly model the underlying structures that distinguish a healthy and non-healthy heart. Some techniques can be applied to help improve the performance on imbalanced datasets, however, again, these come with their own challenges, such as potentially introducing an additional bias to the results. Considering the numerous challenges associated with supervised ML in this area, it hints that a different approach may be appropriate to generate an optimal output.

### 6.4.2. Future research and impact

A great first step would be an organised effort to generate large, open-source datasets consisting of athletes' hearts data so that ML models can be built, tested, and validated by external researchers to confirm the performances of different models. This should also help in building up the trust between those developing the models and those that will be using them, which in turn may help speed up the adoption. This approach is not novel, with the creation of public databases playing a pivotal role in pushing key areas of research in closely related disciplines, such as atrial fibrillation detection [29].

A further area for future research will be to focus on applying ML models to the raw data instead of using pre-extracted features. This will have obvious benefits which have already been stated of saving time and money, in theory allowing the scope of future projects to be more ambitious. The main reason for this approach, however, is the potential for the discovery of novel biomarkers by the ML model, finding associations between the features in the raw data and the previously unknown outcome. These discoveries would help push this research field forward,

helping to strengthen the understanding of the athlete's heart. Castillo-Atoche et al. [168] is the only study within the review that embraces the raw data in the form of ECG images. Their study clearly shows the benefits of this approach with their model being able to carry out analysis automatically and to a high degree of accuracy.

In addition to this, there should also be a focus on developing frameworks that can use ML models that can analyse the raw data from several modalities simultaneously to make its decisions. Rahman et al. [183] suggested that the data from electrocardiography and echocardiography should be considered by healthcare professionals when performing athletes' heart screenings to yield the best results. Therefore, it seems a logical next step to evaluate whether this hypothesis transfers to ML models and if it yields tangible improvements to the model's performance.

Another potential avenue that could be pursued is to look at developing models to determine disease progression alongside the physiological adaptations of the athlete's heart. All the predictive modelling conducted in the above literature centres on determining the presence of the disease, not necessarily the severity of the disease or how it will develop within the subject. Expanding the research in this area will provide healthcare professionals with the tools and information needed to help properly manage the disease and provide the appropriate treatments earlier.

Finally, future research should start to focus on expanding the implementation of unsupervised ML due to its advantages in certain situations over supervised ML. As unsupervised learning does not require labelled data, instead finding key relationships within the data automatically, it provides a solution to the issues with datasets which were mentioned above, relating to the time and cost of labelling, as well as the data where labels are simply not appropriate. A more significant benefit of unsupervised ML in this context is that it allows for a rephrasing of the problem and provides an alternative look into the data. For example, instead of taking the classical approach and phrasing the problem as a binary classification problem, such as is the athlete's heart healthy or not healthy, the problem can instead be constructed as an anomaly detection task and answer "What does a healthy athlete's heart look like?". This approach provides compelling solutions to the issues discussed in the limitations section surrounding the low prevalence of adverse outcomes in athletes' hearts, as only healthy data would be required to develop such a model, providing solutions which give a deeper understanding of the raw data itself, as well as looking at to what degree the data are similar.

### 6.5. Chapter Conclusion

This review shows that there is a clear interest in the use of ML to study the athlete's heart. The most commonly used ML methodologies within this research area were ANNs, support vector

machines, and random forests, where the most common implementation was to perform predictive modelling in the form of disease classification. With continued development and sustained advancements, the future potential of ML applications is promising, not only in improving model prediction accuracies, but in aiding in the understanding of the underlying physiological changes within an athlete's heart.

# 7. Modelling Healthy Athlete's Hearts: Applying GTM to ECG Rhythm Strip Data to Identify Clinically Relevant Sub-Groups

## 7.1. Introduction

As thoroughly discussed in Chapter 6, the athlete's heart describes the physiological adaptations the heart undergoes during extreme training regimes to cope with the increased stresses placed upon it. A heart that has undergone these adaptations differs from the general population in several ways, such as increased cavity sizes (both left and right) and left ventricular wall thickness. There are also differences between athletes as well, caused by factors such as age, sex, and sport played [199–201]. These adaptations can result in cardiac measurements frequently exceeding normal limits [199] that can overlap significantly with identifiers of cardiac disease [200], which if left undiagnosed, can lead to adverse cardiac outcomes. They also make it difficult to develop a complete picture of what a healthy athlete's heart looks like.

Pre-participation screening is used to identify abnormalities in athletes before partaking in sports to reduce any potential risks [61]. Again, as discussed in Chapter 6, ECG tests are commonly used in these screenings as they can help improve the sensitivity of cardiovascular disease detection [200]. In prior years there was some debate around the effectiveness of ECGs for mass pre-participation screenings, with the European model recommending the use of 12-lead ECGs which contrasted with the American Heart Association's opposition to its inclusion based on potentially high false positive rates [202]. However, more recently the inclusion of the ECG is more favourable, with it now being supported by the European Society of Cardiology, the International Olympic Committee, and many National Collegiate Athletic Associations [203]. Regardless, the need for minimising false positives within screening an athlete's heart remains crucial as they can cause a large financial burden and, more importantly, a psychological impact on the athlete in question [204]. Therefore, ensuring a principled approach to interpreting the athlete's heart is vital to minimise this risk.

There have been continued developments for guidelines that provide recommendations to clinicians and medical professionals on how to best interpret ECGs for athletes [204–207]. Alongside these developments, there has also been a growth in the application of AI and ML to help further understand the athlete's heart [61]. An extensive review of these applications is presented in Chapter 6, which highlights unsupervised ML as a promising future avenue for athlete's heart research. Since the publication of the work contained in Chapter 6, a subsequent review looking into the same area has also been published [208], further strengthening the overall message previously presented. The authors identify that the area is still in its infancy, with there

being several promising areas for future AI integration into sports cardiology. They suggest that a hybrid approach that allows AI to be leveraged by clinicians as a decision-making aid will be an optimal format, enhancing the medical expertise they already have. Naturally, however, there are several challenges in developing and implementing AI applications, especially in a healthcare setting. There are general problems that apply, such as the lack of interpretability and transparency of certain ML models, alongside more specific challenges, such as a lack of available data [61,204]. In any case, the inclusion of AI will likely lead to a better understanding of the athlete's heart and in turn, lead to more positive outcomes by proper management and treatment of disease [61,208], should it address the challenges and concerns surrounding its implementation.

This chapter proposes a novel methodological workflow to identify different sub-groups within a population of healthy athletes based on their ECG recordings. The proposed methodology will consist of three core sections that combine and build upon many novel elements described in previous chapters. These core sections are data extraction, feature generation, and data clustering. First, the data extraction stage applied the ECG digitisation algorithm defined in Chapter 3 to extract the signals presented within the ECGs stored as PDFs. The feature extraction stage then takes an ECG signal and extracts human readable features, such as P wave and QRS duration, automatically. Finally, the data clustering section groups together athletes based on the similarity of their ECG features. To carry out this clustering, we applied a modified version of the methodology described in Chapter 5 that now performs constrained hierarchical clustering, based on neighbourhoods derived using the magnification factors calculated from the GTM output. This proposed methodology aims to address several of the areas highlighted for future research, providing a robust and trustworthy output [61,208].

## 7.2. Materials and Methods

### 7.2.1. Data Source

The data used for this analysis was derived from a set of 854 ECG recordings from 611 healthy athletes from four different sporting disciplines: footballers, cyclists, rugby league players, and ultra runners. The ECG recordings were carried out by cardiologists at Liverpool John Moores University over several years for the purposes of pre-participation screening. The ECG recordings contained the 12 standard leads arranged in a 3x4 grid (identical to the arrangement B ECGs outlined in Figure 5c in Chapter 3) along with a full 10-second recording of the signal generated by lead II. These prolonged single lead recordings are known as rhythm strips and are used to accurately assess the cardiac rhythm, with lead II commonly being used for this purpose as it generally provides a good view of the P-wave [209]. The analysis carried out within this chapter

used these rhythm strip signals. The ECG recordings were stored in a PDF format and recorded at a sample rate of 150Hz.

### 7.2.2. Methodological Workflow

#### *7.2.2.1. Data Extraction*

The first section of the approach focuses on converting the ECGs stored in the PDF format into digital signals. The ECG signals were embedded within a page surrounded by other demographical data such as age and sex. An important note here is that the data had already been pseudo-anonymised, meaning any identifiable information had already been omitted from the PDF, and replaced with appropriate IDs. We then created an algorithm that could extract the rhythm strip ECG signal portion of the PDF and store it as a PNG file. This step is crucial as converting the data to an image format allowed for the signal extraction algorithm defined in Chapter 3 to be employed to convert the rhythm strip image into a digital signal. As the signal was sampled at 150Hz, and the rhythm strip represents a 10-second recording of lead II, then each generated signal had a dimension of 1x1500, with each data point representing 6.67ms.

Within the algorithm used to locate the rhythm strip and convert it to an image, we also leveraged the Python package "Pytesseract" to perform optical character recognition. This allowed for metadata to be extracted for each athlete simultaneously from the information surrounding each ECG signal. From the PDFs, we were able to extract the following demographical information for each athlete: sporting discipline, ethnicity, age (at ECG recording), sex, height, and weight (and by extension, BMI). These variables were not used for modelling and instead used as additional investigative variables to provide a deeper understanding and context to the makeup of the identified clusters.

#### *7.2.2.2. Feature Generation*

The next part of the methodology takes the extracted rhythm strip signals from the previous section and generates human interpretable features from these signals to be used during the modelling phases of the analysis. To achieve this, an algorithm was developed using functions provided in the Python packages "heartpy" and "neurokit2" [79]. The heartpy package is a Python-based heart rate analysis toolkit that provides several functions to both pre-process an ECG signal and extract heart rate variability features. The neurokit2 package is an open-source Python-based neurophysiological signal processing toolkit that provides several functions that can be applied to a range of bodily signals, however, for our purposes, only ECG-related functions were required. These functions were used primarily to identify key event locations within the ECG signal, such as the onset and offset points of P and T waves.

Before any feature extraction is performed, each signal is min-max scaled to normalise its values between 0 and 1. Then, the first step is to identify the locations for the peaks of the R waves for each signal. This is an important step as the subsequent functions rely on these locations to accurately extract their relevant features. This is achieved by employing the *enhanced_peaks* heartpy function that enhances the signal-to-noise ratio by emphasizing the highest peaks, which generally within an ECG is the R-wave [210]. This processed signal is then passed through the *ecg_peaks* neurokit2 function that identifies the peak of the R-wave by detecting the local maxima of the absolute gradient within the QRS complex.

From this point the feature generation is split into two parts: the first part extracted features from the ECG relating to heart rate variability, with the second part focusing on extracting information regarding wave and interval durations. Starting with the heart rate variability, all features extracted are derived from the time differences between subsequent R-wave peaks. The heart rate variability features calculated from each signal, along with a description of each variable, are outlined in Table 12.

*Table 12. Defines each of the heart rate variability features extracted from each ECG rhythm strip, along with a description of each variable.*

| **Heart Rate Variability Extracted Feature** | **Description** |
|---|---|
| Beats per minute (BPM) | Describes the resting heart rate of the athlete, calculated by the average number of large squares in between consecutive R waves per minute |
| Interbeat Interval (IBI) | Another term for the RR Interval, this describes the average time between consecutive R peaks |
| RR Interval (sd) | Describes the standard deviation of all the calculated IBIs |
| Standard deviation of successive differences | Describes the standard deviation of the differences between consecutive IBIs |
| Root mean square of successive differences | Describes the root mean square of the differences between consecutive IBIs |
| Proportion of successive differences above 20ms | A value between 0 and 1 that indicates what proportion of the differences between the consecutive IBIs were over 20ms |
| Proportion of successive differences above 50ms | A value between 0 and 1 that indicates what proportion of the differences between the consecutive IBIs were over 50ms |

The second part of the feature extraction generates wave and interval durations from the ECG signal, commonly used when interpreting an ECG recording. This was achieved by first using the *ecg_delineate* neurokit2 function that identified key points along the ECG signal using discrete

wavelet transformations [211]. The identified points relate to the onset, peak, and offset locations of the P wave, QRS complex and T Wave. A set of points is generated for each R wave peak identified earlier in the process, i.e. for every beat in the ECG signal. Figure 26 displays one of the ECG rhythm strips with the identified points marked on the signal. Before any durations are calculated, however, a check is performed to ensure that the points that have been extracted are correct and accurate. The check looks to see if the order of the locations identified is in the correct sequence, with the correct order demonstrated in the legend of Figure 26. If all the locations are in the correct order, the beat is included for the calculations of the interval and wave durations. Again, using Figure 26 to demonstrate this, we see that 7 out of the 8 beats passed the check and will be used in the duration calculations, with the first beat of the signal excluded. This will be due to an error in the order of the points identified, for example, the R wave onset being identified as occurring earlier than the p wave offset for that beat. From these identified positions, we were able to calculate the following features:

- P wave duration (mean and standard deviation)
- PR interval (mean and standard deviation)
- QRS complex (mean and standard deviation)
- ST segment (mean and standard deviation)
- T wave (mean and standard deviation)

By applying all the feature generation steps to each ECG rhythm strip, we were able to generate a total of 17 features which serve as the data used during the modelling stage of the analysis.

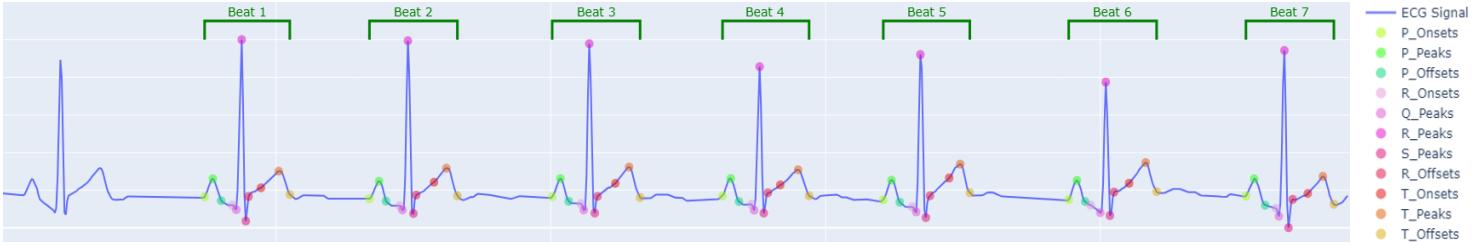



*Figure 26. An athlete's ECG rhythm strip with the key points relating to the P wave, QRS complex and T wave marked on the signal.*

*7.2.2.3. Data Clustering*

The final stage of the proposed methodology takes the features extracted from the ECG rhythm strips and applies a clustering approach to identify similar sub-groups within the data. The approach outlined in this chapter builds upon the methodology outlined in Chapter 5, with GTM again being the method of choice to model the data and generate the latent clusters. Like with

Chapters 4 and 5, GTM was selected due to its ability to identify complex non-linear relationships within the data, as well as provide interpretability of the cluster compositions via the reference vectors and reference maps.

This chapter provides new developments in the identification of the macro clusters in the latent space. Here, we introduce a novel methodology for identifying clinically relevant macro clusters by using a constrained hierarchical clustering of the reference vectors. When considering the latent nodes generated via GTM, nodes that are closer together will be mapped to points that are close together in the data space. The idea behind applying a constraint to the hierarchical clustering is to help preserve these latent neighbourhoods within the macro-clusters to generate more accurate, easier-to-interpret macro-clusters [50]. This work is essential due to the GTM micro-cluster visualisation being too granular for practical use [212]. Like in Chapter 5, macro clusters solve this issue by providing aggregated cluster partitions that can be more easily interpreted. Using constrained hierarchical clustering for this purpose is not novel in and of itself, with a version of the approach being used by Vellido *et al* [50]. The study implemented a simple neighbourhood constraint to the reference vectors such that all immediate surrounding nodes were considered neighbours, with non-neighbouring nodes restricted from merging.

The novel aspect of our approach therefore is that we first define the neighbouring condition in the latent space using magnification factors. The neighbours for each latent node were defined using unsupervised k nearest neighbours' (KNN) algorithm. Given a set of points $\boldsymbol{U} = \{u_1, u_2, \cdots, u_l,\}$ in $R^n$, for every point $u_i$, unsupervised KNN defines $k$ neighbours for that point that have the smallest distance away [43,213]. For our purposes, each point $u_i$ resides in $R^3$, with it being defined by the 2-dimensional coordinates of each latent centre along with the magnification factor corresponding to that centre. Euclidean distance was used to evaluate the distance between pairs of points. We set $k = 4$ as we wished to define highly nuanced neighbourhoods that captured the information being provided through the magnification factors [43]. Once the neighbouring conditions had been defined for each latent centre, these were transferred to their corresponding reference vector to be used as the constraint in the hierarchical clustering.

The magnification factors generated from the trained GTM model provides information for how the lower dimensional latent manifold distorts when being mapped to the higher dimensional data space. Higher magnification factors show areas of high distortion during the projection, which corresponds to areas where data is sparse, with the reverse being true for lower magnification values. These can be assessed visually by superimposing the magnification factors onto the membership map generated from GTM and using a grey-colour scale to represent the

values of the factors. Magnification factors were used in the study by Vellido *et al* [50], however, they were used indirectly to visually assess the macro-clusters to see if they estimated the magnification factor distribution. Our approach instead looks to directly include the magnification factor information to influence the macro-clusters to create a simpler, equally informative, output. This process is visually outlined in Figure 27.

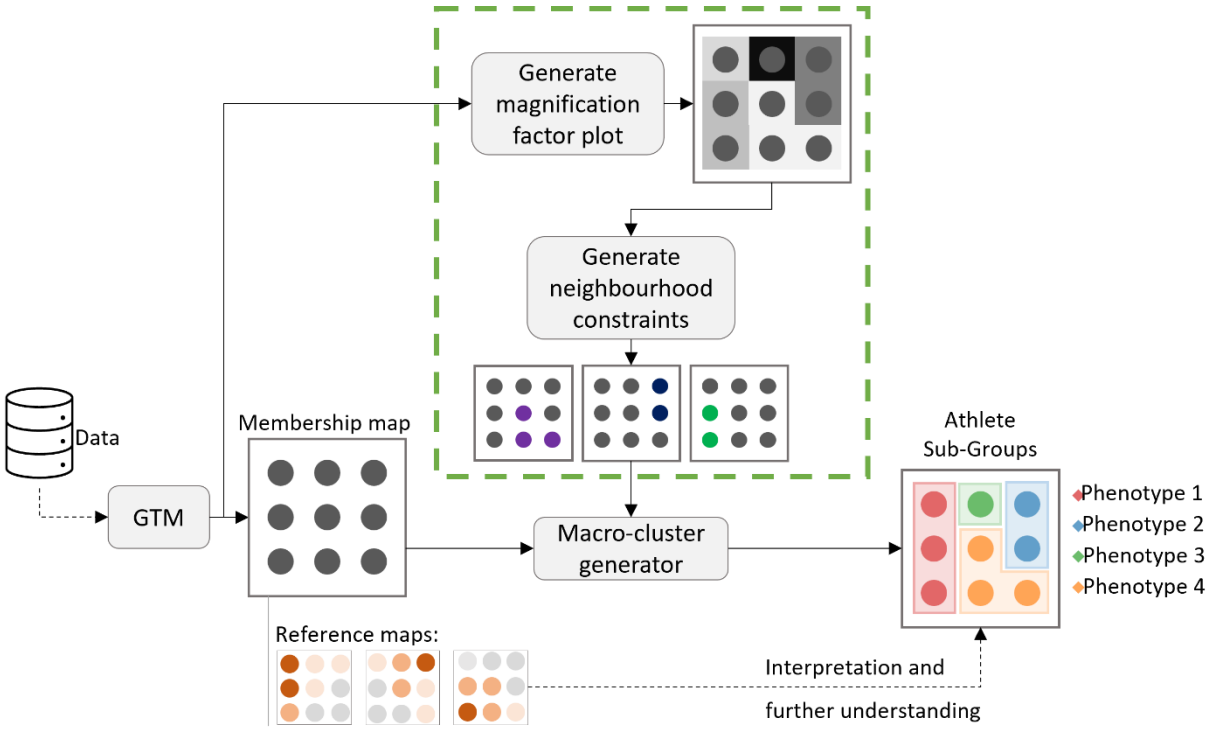



*Figure 27. Proposed AI-based methodology that builds on the approach outlined in Chapter 5. The modification to the approach is contained with the green dashed box and demonstrates how the magnification factors are generated from the GTM model, used to identify neighbourhood constraints, which are then used to influence the hierarchical clustering applied to the reference vectors.*

### 7.2.3. Statistical analysis

Medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (percentages) were used for categorical variables. There were several ordinal variables used for the exploratory analysis of the GTM output. These were one-hot encoded and then treated as a categorical variable and represented in the data as such.

To study the characteristics of the generated phenotype groups, differences between continuous variables were analysed using the Kruskal-Wallis test and differences between categorical variables were analysed using the Chi-squared test. In both cases, a p-value $<0.05$ was the threshold for statistical significance.

### 7.3. Results

#### 7.3.1. Data Summary

The data extraction algorithm extracted all available demographical information contained within the PDFs. From this, we learned that the median age of the athletes is 20 (IQR 16-28) with a range between 13 and 62, with ~81% being male. Table 13 contains the summary of these additional investigative variables, along with the modelling variables extracted during the feature generation stage of the methodology. A note here is that the colour scheme used to differentiate between the modelling and additional investigative variables, both in the tables and in the visualisations, is the same as that used within Chapter 5.

*Table 13. Summary characteristics of the athlete's data*

| Variable name | Value |
|---|---|
| **Modelling variables:** | |
| Heart Rate Variability: | |
| BPM | 58 (52, 65) |
| IBI [ms] | 1034 (924, 1163) |
| RR Intervals (sd) [ms] | 47 (30, 73) |
| Standard deviation of successive differences [ms] | 57 (36, 94) |
| Root mean square of successive differences [ms] | 59 (37, 96) |
| Proportion of successive differences above 20ms | 0.83 (0.67, 1) |
| Proportion of successive differences above 50ms | 0.43 (0.14, 0.67) |
| Waves and Intervals: | |
| P-wave duration (mean) [ms] | 75 (55, 93) |
| P-wave duration (sd) [ms] | 16 (8, 23) |
| PR-interval duration (mean) [ms] | 151 (134, 173) |
| PR-interval duration (sd) [ms] | 16 (10, 26) |
| QRS-complex (mean) [ms] | 107 (91, 120) |
| QRS-complex (sd) [ms] | 9 (4, 17) |
| ST-segment duration (mean) [ms] | 154 (135, 175) |
| ST-segment duration (sd) [ms] | 16 (8, 28) |
| T-wave duration (mean) [ms] | 140 (123, 153) |
| T-wave duration (sd) [ms] | 14 (8, 27) |
| **Additional investigative variables:** | |
| Demographics: | |
| Age at ECG [years] | 20 (16, 28) |
| Height [cm] | 179 (171, 184) |
| Weight [kg] | 76 (65, 87) |
| BMI [$kg/m^2$] | 24 (22, 26) |
| Sex [Male] | 691 (81%) |

| | |
|---|---|
| Ethnicity (70% populated): | |
| White | 469 (55%) |
| Black | 124 (15%) |
| Hispanic | 3 (0%) |
| Asian | 1 (0%) |
| Sporting Discipline (98% populated): | |
| Football | 375 (44%) |
| Cyclists | 213 (25%) |
| Rugby League | 130 (15%) |
| Ultra runner | 115 (13%) |

### 7.3.2. Data Clustering Results

#### *7.3.2.1. Membership map visualisation*

Like Chapter 5, we also performed hyperparameter tuning using 10-fold cross validation to determine the optimal parameters for the GTM modelling. For a latent space of dimension 10x10 to provide an appropriate level of granularity for this smaller dataset size, using 16 RBFs arranged in a 4x4 grid with a regularisation term of 0.1 was optimal. Figure 28 shows the membership map generated by the GTM model trained on the features extracted from the athlete's ECG rhythm strips. The maps display the latent space containing a compressed representation of the entire original data space. Each point on the map represents a micro-cluster containing at least one ECG, with the size of the point indicating the number of ECGs in the cluster: the larger the point, the more ECGs in the cluster and vice versa.

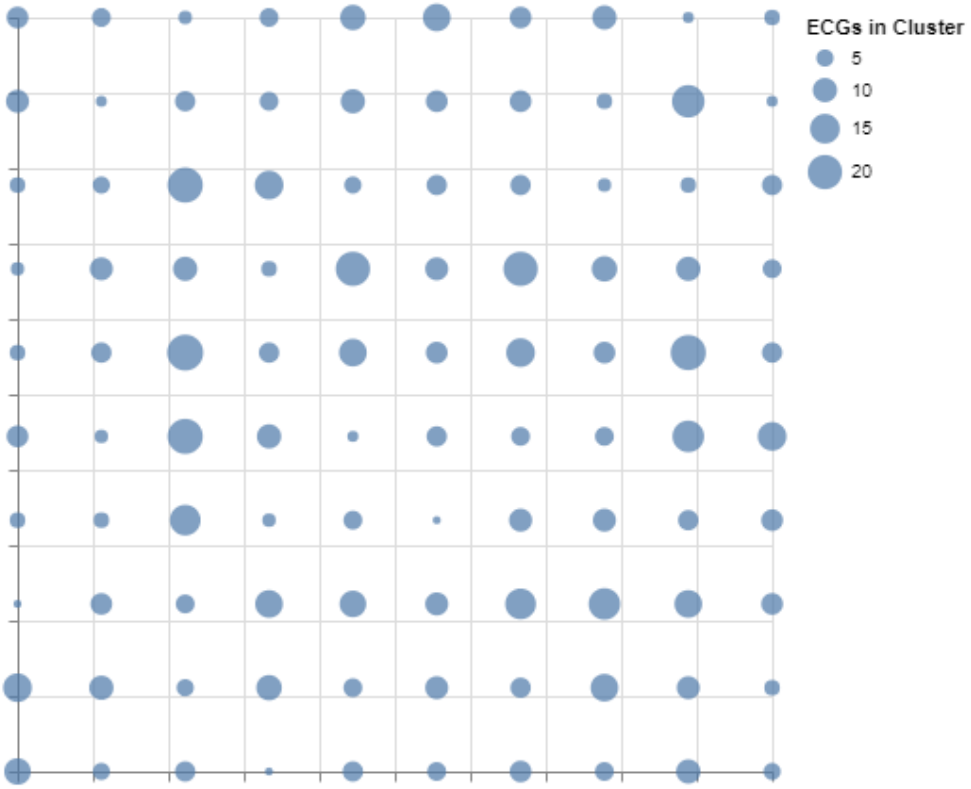



*Figure 28. Membership map representing the latent space generated from the GTM model trained of the features extracted from the athlete's ECG rhythm strips.*

*7.3.2.2. Reference map visualisation*

Figure 29 contains the reference vectors extracted from the trained GTM model trained on the features extracted from the athlete's ECG rhythm strips. The reference maps displayed how each of the modelling variables affected the latent clusters and have been split based on whether the variable relates to heart variability, or the wave and intervals of the rhythm strip for easier interpretation. Each point in every plot within Figure 29 corresponds exactly to the same point in their respective membership maps in Figure 28. A light grey–red colour scheme was used for the reference vectors plot such that areas of the plots that are redder indicate that participants in that cluster had a higher value of that variable. Likewise, if the point in the reference vector is greyer, the lower the value is for participants in this cluster. Again, like in Chapter 5, all plots using the light grey–red colour scheme indicate variables used in the GTM model development, whereas plots using a light grey–teal represent variables that were not used in the modelling and have no direct impact on the clusters themselves.

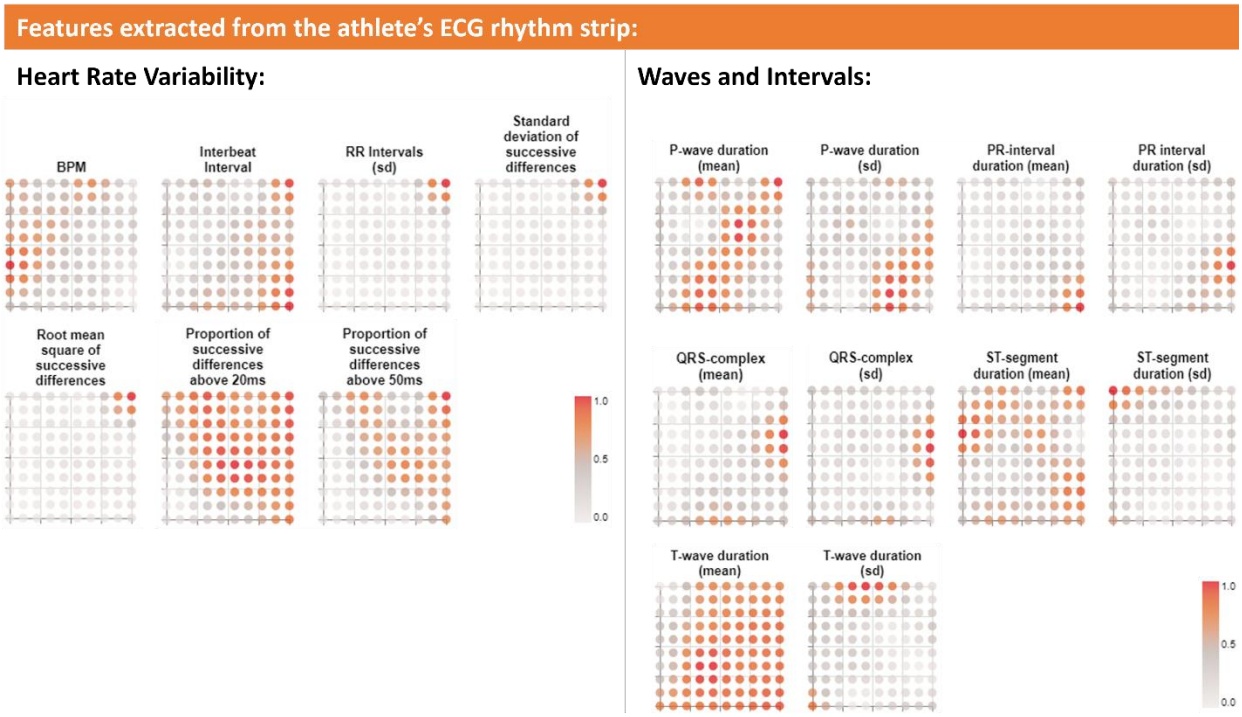



*Figure 29. Reference vector visualisations demonstrating how each variable extracted from the ECG rhythm strip affects the cluster distribution in the latent space for both, with variables split by their respective categories.*

*7.3.2.3. Additional investigative variable visualisations*

Figure 30 contains visualisations for all the additional investigative variables that show how data from different investigative variables are distributed within the membership maps. The visualisations representing the investigative variables all use a light grey-teal colour scheme as they were not used in model development. The value assigned to each micro-cluster is the average of the variable for all participants assigned to each cluster, the more teal a micro-cluster is, the higher the value.

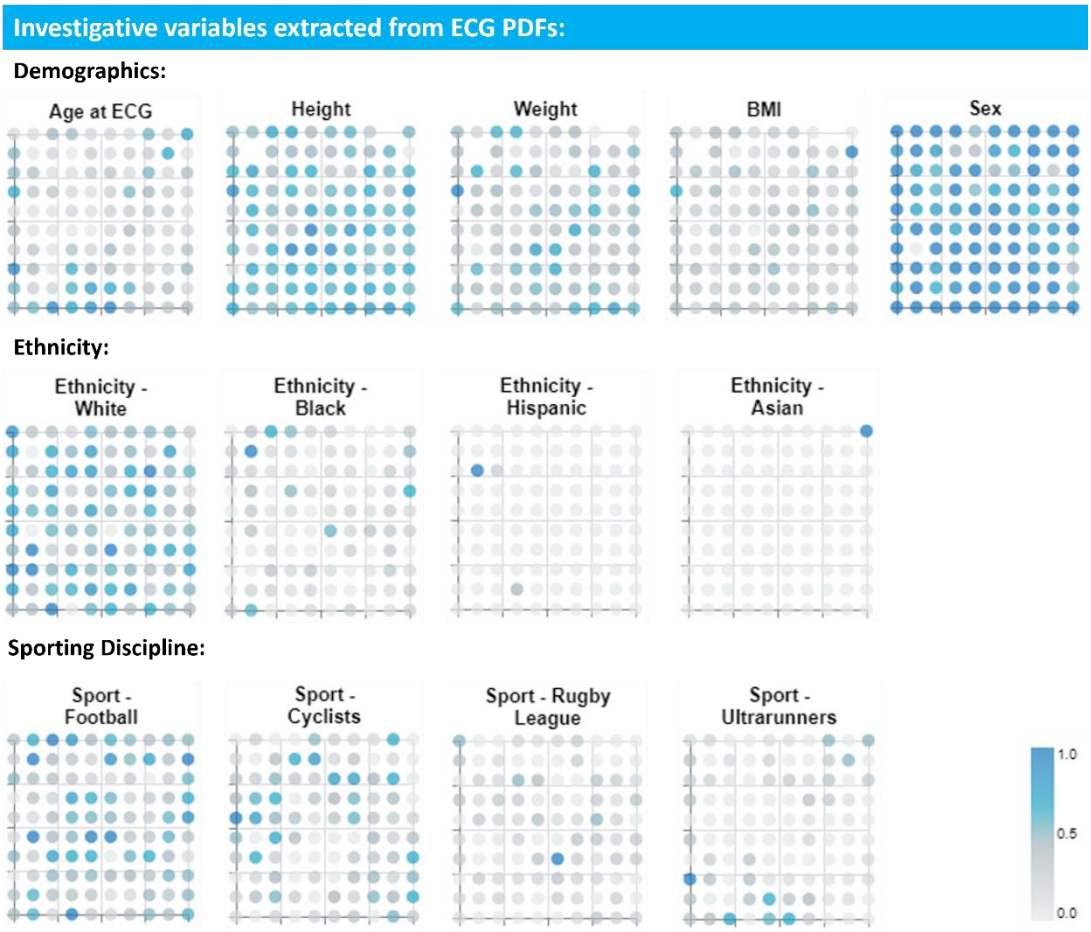



*Figure 30. Membership maps showing how the additional investigative variables data are distributed within the latent space.*

### *7.3.2.4. Magnification factors visualisation*

Figure 31 displays the magnification factor plot generated from the trained GTM, as described in section 7.2.2.3, using a grey-scale representation. This plot provides a visual representation of how the latent manifold distorts when projected and fitted to the data. Lighter areas of this visualisation demonstrate the regions whereby there was low distortion in the mapping, with the darker the grey corresponding to regions that experienced high distortion in the same mapping.

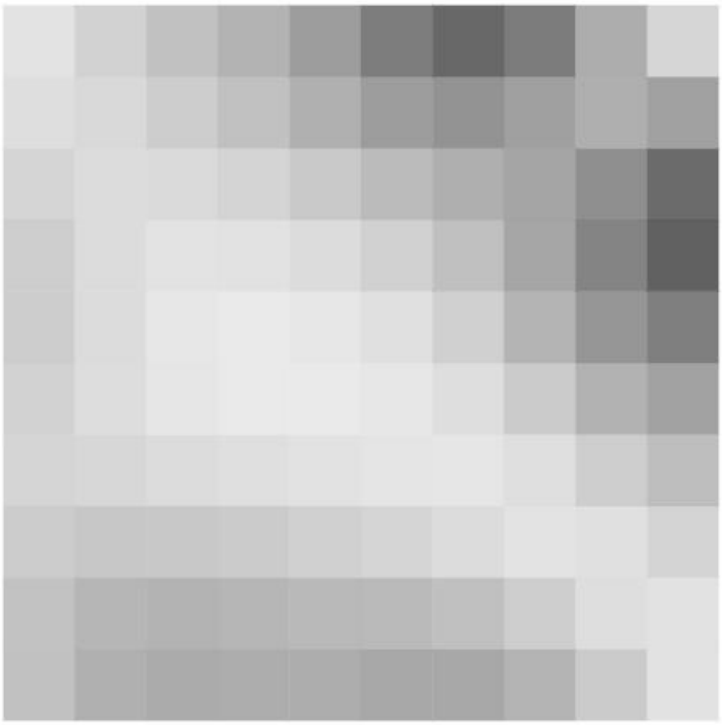


*Figure 31. Magnification map calculated from the trained GTM. Light areas of this map correspond to areas of low distortion during mapping, with the darker areas relating to area of high distortion during the mapping.*

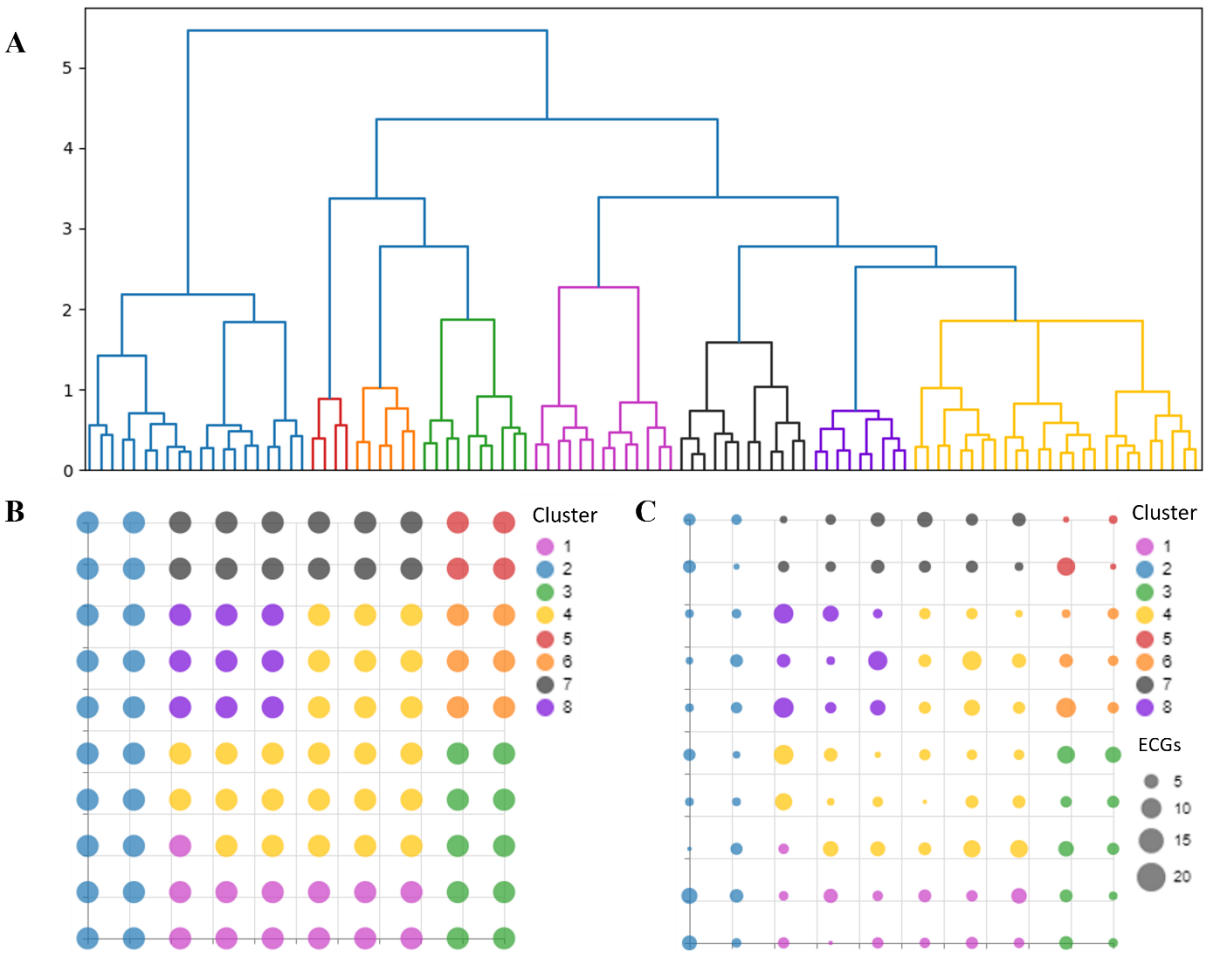
*7.3.2.5.Macro-cluster analysis and description of identified athlete sub-groups*

The data described in Table 13 was then split according to the number of sub-groups identified through the macro-cluster analysis and compared in Table 14. By applying the clustering approach outlined in section 7.2.2.3, we identified 8 clusters within the reference vectors, as demonstrated in the dendrogram in Figure 32(a). By transferring these cluster assignments to their corresponding latent centre, we were able to generate 8 macro cluster regions that also define the 8 different athlete sub-groups, as shown in Figures 32 (b) and (c). A full breakdown of the headline features of each sub-group derived using the information within Table 14 is presented in Figure 32 (d). Unlike Chapter 5, a full separate text breakdown is not required due to the smaller, more manageable, number of variables.

*Table 14. Characteristics of the participants split per athlete sub-group. Shades of red/blue were used per variable to illustrate differences between lower and higher values. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.*

| Variable name | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 | Group 7 | Group 8 | P-value |
|---|---|---|---|---|---|---|---|---|---|
| **Modelling variables:** | | | | | | | | | |
| Heart Rate Variability: | | | | | | | | | |
| BPM | 54 (51, 56) | 67 (63, 74) | 47 (43, 50) | 56 (52, 61) | 53 (48, 54) | 54 (51, 58) | 64 (61, 70) | 65 (62, 69) | <0.05 |
| IBI | 1119 (1064, 1187) | 897 (809, 953) | 1297 (1211, 1407) | 1068 (987, 1148) | 1230 (1179, 1325) | 1128 (1058, 1193) | 953 (865, 1005) | 925 (874, 973) | <0.05 |
| RR Intervals (sd) | 24 (17, 33) | 23 (15, 37) | 61 (42, 91) | 51 (37, 68) | 403 (344, 469) | 68 (46, 106) | 60 (41, 97) | 52 (38, 77) | <0.05 |
| Standard deviation of successive differences | 28 (20, 38) | 22 (16, 34) | 82 (60, 130) | 62 (44, 89) | 631 (540, 811) | 94 (68, 151) | 73 (46, 123) | 61 (48, 91) | <0.05 |
| Root mean square of successive differences | 30 (21, 38) | 23 (16, 34) | 83 (62, 130) | 63 (45, 90) | 631 (540, 812) | 96 (70, 153) | 75 (46, 123) | 63 (49, 93) | <0.05 |
| Proportion of successive differences above 20ms | 0.57 (0.39, 0.71) | 0.5 (0.31, 0.63) | 0.83 (0.8, 1) | 0.86 (0.72, 1) | 1 (0.83, 1) | 1 (0.86, 1) | 0.88 (0.75, 1) | 0.88 (0.75, 1) | <0.05 |
| Proportion of successive differences above 50ms | 0.11 (0, 0.17) | 0 (0, 0.13) | 0.6 (0.5, 0.75) | 0.5 (0.29, 0.71) | 0.78 (0.58, 0.83) | 0.6 (0.5, 0.82) | 0.5 (0.25, 0.71) | 0.5 (0.38, 0.67) | <0.05 |
| Waves and Intervals: | | | | | | | | | |
| P-wave duration (mean) | 88 (75, 107) | 59 (45, 74) | 59 (50, 76) | 92 (76, 107) | 97 (78, 112) | 77 (61, 91) | 76 (67, 89) | 49 (39, 63) | <0.05 |
| P-wave duration (sd) | 17 (7, 30) | 15 (10, 20) | 20 (13, 31) | 12 (6, 22) | 15 (6, 23) | 22 (12, 31) | 15 (11, 21) | 16 (12, 21) | <0.05 |
| PR-interval duration (mean) | 158 (143, 178) | 135 (118, 150) | 204 (170, 239) | 158 (139, 176) | 163 (141, 184) | 144 (132, 168) | 140 (128, 164) | 139 (125, 153) | <0.05 |
| PR-interval duration (sd) | 18 (11, 30) | 15 (10, 22) | 39 (27, 48) | 12 (7, 19) | 15 (9, 26) | 20 (14, 27) | 15 (11, 18) | 17 (13, 22) | <0.05 |
| QRS-complex (mean) | 119 (110, 136) | 89 (83, 99) | 121 (113, 139) | 110 (100, 117) | 111 (107, 116) | 153 (143, 168) | 92 (84, 101) | 92 (85, 102) | <0.05 |
| QRS-complex (sd) | 11 (3, 20) | 9 (5, 15) | 19 (5, 31) | 7 (3, 11) | 10 (6, 15) | 27 (22, 35) | 8 (4, 15) | 9 (4, 15) | <0.05 |
| ST-segment duration (mean) | 159 (145, 170) | 159 (129, 182) | 172 (148, 195) | 148 (132, 166) | 173 (149, 193) | 123 (104, 143) | 160 (143, 175) | 155 (143, 175) | <0.05 |
| ST-segment duration (sd) | 10 (6, 16) | 30 (20, 50) | 8 (5, 14) | 11 (7, 18) | 13 (8, 20) | 21 (14, 29) | 37 (31, 46) | 19 (14, 26) | <0.05 |
| T-wave duration (mean) | 146 (136, 160) | 98 (65, 125) | 152 (143, 161) | 148 (137, 158) | 138 (128, 153) | 150 (137, 162) | 128 (106, 142) | 127 (113, 141) | <0.05 |
| T-wave duration (sd) | 10 (6, 14) | 24 (18, 34) | 8 (5, 12) | 10 (7, 17) | 12 (6, 16) | 9 (6, 20) | 37 (30, 47) | 20 (13, 27) | <0.05 |

| Additional investigative variables: | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Demographics: | | | | | | | | | |
| Age at ECG | 30 (19, 43) | 21 (16, 31) | 22 (17, 28) | 19 (17, 27) | 32 (19, 46) | 19 (16, 26) | 20 (15, 26) | 18 (15, 21) | <0.05 |
| Height | 180 (177, 188) | 179 (172, 184) | 181 (171, 185) | 180 (172, 185) | 178 (168, 188) | 178 (170, 184) | 176 (170, 182) | 177 (170, 182) | <0.05 |
| Weight | 79 (73, 87) | 75 (66, 84) | 76 (67, 90) | 76 (69, 87) | 75 (62, 82) | 75 (64, 85) | 71 (61, 84) | 71 (60, 85) | 0.0802 |
| BMI | 24 (23, 26) | 24 (22, 26) | 24 (22, 26) | 24 (22, 27) | 23 (22, 25) | 24 (22, 26) | 24 (22, 25) | 23 (21, 26) | 0.5270 |
| Sex | 80 (87%) | 104 (83%) | 69 (73%) | 207 (83%) | 25 (96%) | 42 (76%) | 72 (77%) | 92 (79%) | 0.9262 |
| Ethnicity | | | | | | | | | |
| White | 56 (61%) | 74 (59%) | 50 (53%) | 130 (52%) | 16 (62%) | 24 (44%) | 47 (50%) | 72 (62%) | 0.735 |
| Black | 9 (10%) | 23 (18%) | 19 (20%) | 31 (12%) | 3 (12%) | 14 (25%) | 14 (15%) | 11 (9%) | 0.0975 |
| Hispanic | 1 (1%) | 1 (1%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (1%) | 0.6834 |
| Asian | 0.0 (0.0%) | 0.0 (0.0%) | 0.0 (0.0%) | 0.0 (0.0%) | 1.0 (4.0%) | 0.0 (0.0%) | 0.0 (0.0%) | 0.0 (0.0%) | <0.05 |
| Sporting Discipline: | | | | | | | | | |
| Football | 30 (33%) | 48 (38%) | 42 (44%) | 111 (44%) | 12 (46%) | 31 (56%) | 50 (53%) | 51 (44%) | 0.3808 |
| Cyclists | 20 (22%) | 29 (23%) | 28 (29%) | 68 (27%) | 3 (12%) | 10 (18%) | 22 (23%) | 33 (28%) | 0.6112 |
| Rugby League | 9 (10%) | 20 (16%) | 17 (18%) | 39 (16%) | 0 (0%) | 9 (16%) | 9 (10%) | 27 (23%) | 0.0672 |
| Ultra runner | 30 (33%) | 27 (21%) | 4 (4%) | 26 (10%) | 11 (42%) | 4 (7%) | 11 (12%) | 2 (2%) | <0.05 |




**D**

| Group 1 (n = 92) | Group 2 (n = 126) | Group 3 (n = 95) | Group 4 (n = 250) |
|---|---|---|---|
| ‣ 2nd lowest overall heart rate variability<br>‣ 2nd highest Age<br>‣ High proportion of ultra runners with lowest proportion of footballers | ‣ Highest BPM<br>‣ Lowest overall heart rate variability<br>‣ Lowest PR interval and QRS complex and T wave duration<br>‣ High variability in wave and interval durations<br>‣ High proportion of ultra runners | ‣ Lowest BPM<br>‣ Longest and most variable PR interval durations<br>‣ Longest and least variable T waves<br>‣ Lowest ST segment variability<br>‣ Tallest athletes<br>‣ Highest proportion of cyclists | ‣ Largest Group<br>‣ Long P wave durations<br>‣ Lowest overall wave and interval variability<br>‣ High proportion of cyclists |
| **Group 5 (n = 26)** | **Group 6 (n = 55)** | **Group 7 (n = 94)** | **Group 8 (n = 116)** |
| ‣ Smallest group<br>‣ Highest overall heart rate variability<br>‣ Longest P waves and ST segment durations<br>‣ Oldest athletes<br>‣ Highest proportion of Male athletes<br>‣ Highest proportion of ultrarunners | ‣ Longest and most variable QRS complex durations<br>‣ Most variable P wave durations<br>‣ Lowest ST segment duration<br>‣ Highest proportion of black athletes<br>‣ Highest proportion of footballs | ‣ High BPM<br>‣ High variability in ST segment and T wave durations<br>‣ High proportion of footballers | ‣ High BPM<br>‣ Lowest P wave durations<br>‣ Low PR interval and QRS complex durations<br>‣ Youngest athletes<br>‣ Highest proportion of rugby league players<br>‣ Lowest proportion of ultra runners |

*Figure 32. Derived sub-groups of athletes using data extracted from their ECG rhythm strips A) Dendrogram produced using a constrained Ward's minimum variance method. The graph shows the 8 clusters that are used to define the 8 athlete's sub-groups B) Membership map with a uniform size for the micro-clusters to show the distribution of the macro-cluster regions. C) The size of the micro-clusters in the membership map dictated by the number of ECGs assigned to it. D) Characterising features for each of the sub-groups.*

### 7.4. Discussion

By using the methodology outlined in this chapter, we were able to take ECGs stored in a PDF format, digitise the signal, and automatically extract features relating to heart rate variability as well as wave and interval durations, and generate clinically relevant athlete sub-groups. These sub-groups provide an alternative view to visualising the different adaptions athlete's hearts undergo, which in turn could enable a deeper understanding of these differences. Identifying sub-groups such as these could also a useful addition to a pre-participation screening toolbox. By applying the trained model to a new athletes' data, it would provide a user the ability to see what athletes are similar and aid in making an informed decision based on common outcomes for the group it is assigned to.

Along with the benefits of the approach already outlined in Chapter 5, a key finding from this chapter is that we were successfully able to develop upon the methodology. We achieved this through the inclusion of the magnification factor to directly influence the macro-cluster identification process. This approach resulted in the generation of macro-clusters with boundaries that better reflected the latent micro-cluster neighbours. To the best of our knowledge, GTM has not been used before to study the athlete's heart or to generate clinically relevant athlete sub-groups.

#### 7.4.1. Interpretation of athlete sub-groups

As in Chapter 5, through the combination of the membership map in Figure 28, the reference maps in Figure 29, and the additional investigative variable plots in Figure 30 we can gain a greater understanding of the reasons behind cluster assignments. The inclusion of the magnification factor plot in Figure 31, further improves the ability to interpret the results.

By comparing the magnification factor plot in Figure 31 and the athlete sub-groups presented in Figure 32 (b) and (c), we clearly see how the former successfully influenced the boundaries in the latter. To demonstrate this, we will consider 3 of the identified groups: group 5, group 6, and group 7. Figure 33 highlights the areas of interest within the magnification factor plot for this example, where the blue-shaded region encapsulates groups 6 and 7, which corresponds to an area that experienced high distortion during the mapping process. This region completely isolates a denser area of points contained within the green-shaded region, which corresponds to the location of group 5. This indicates that the ECGs assigned to group 5 are much further away from the others within the data space. By referring to Figure 29, we see the reference maps show that ECGs in cluster 5 are distinct in that they have the highest values for RR interval (sd), standard deviation of successive differences and root mean square successive differences. This is also confirmed

within Table 14, with the value differences for these variables being an order of magnitude larger. Upon further investigation, we uncovered that cluster 5 ECGs took one of two forms:

1. ECGs that genuinely displayed high levels of heart rate variability within the rhythm strip (Figure 34 (a))
2. ECGs that have high heart rate variability due to erroneous data extraction (Figure 34 (b)).

To further expand on point 2, the erroneous data extraction refers to the R wave between beats 7 and 8 in Figure 34 (b) that was not identified properly. This resulted in the R peak detection part of the feature generation section of the methodology failing to identify it. Therefore, the algorithm considers the two beats as consecutive, when in fact there should be another beat in between that would result in the heart rate variability being calculated as normal. This highlights a limitation in the data extraction and feature generation as there are certain instances whereby data is not obtained accurately. However, this result also highlights the benefit of using GTM for this type of analysis, as it has successfully managed to identify and isolate similar, erroneous data into one area of the latent space.

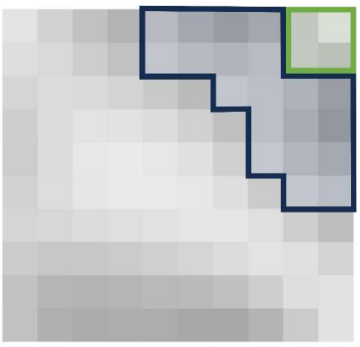


*Figure 33. Magnification factor plot from figure 31, with areas of interest highlighted. The blue shaded region highlights an area where there was high distortion when mapping to fit the data, with the green shaded region highlighting a denser region of points isolated in the top right of the plot.*

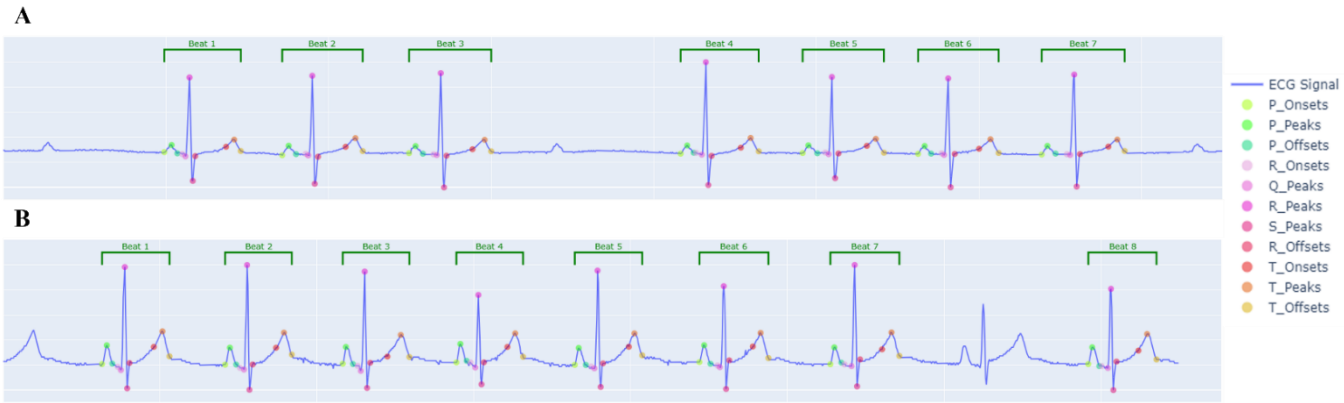



*Figure 34. Examples of the two types of ECG rhythm strips that were assigned to group 5. (a) Contains an ECG that genuinely exhibits high heart rate variability; (b) An ECG whereby an erroneous extraction has led to a false high heart rate variability calculation.*

### 7.4.2. Analysis limitations

As discussed in the previous sub-section, a limitation of this analysis lies in the data extraction and feature generation stage with regard to failed R wave peak detections, which in turn led to some heart rate variability calculations being skewed. An additional potential limitation lies within the GTM algorithm's use of Gaussian distributions, as these distributions can struggle to deal with outliers in the data, in this case, caused both by the erroneously calculated and genuinely high heart rate variability ECGs. This can result in the rest of the data being condensed down into a smaller area of the latent space, and in turn, not fully capturing the underlying relationships in the data. This could potentially be an explanation as to why in this analysis, sex and race were not significantly different among the identified groups, even when these are key factors used in real-world assessments to evaluate an athlete's heart [199–201]. This could be addressed in future analysis by employing *t-GTM,* an alternative to the regular GTM that redefines the methodology to instead use t-distributions, which are less sensitive to outliers [214].

## 7.5. Chapter Conclusion

The analysis in this chapter proposed a novel approach to analysing the athlete's heart and generating clinically relevant sub-groups from within a population of healthy athletes. We developed and applied data extraction and feature generation algorithms to convert PDFs containing ECGs, into human-readable features commonly used to evaluate an ECG recording. We then applied a modified version of the GTM macro clustering outlined in Chapter 5 to cluster these features to identify the sub-groups. This approach is the first of our knowledge to integrate the magnification factors to directly influence the macro-cluster boundaries, providing simpler more interpretable output.

# 8. Chapter 8: Conclusions and Future Research

## 8.1. Conclusion

The overall goal of this thesis was to develop and implement ML models to generate novel insights into several areas of cardiovascular research. The research outlined in the above chapters details this journey, outlining the various developed methodologies to help further the understanding of ECG ML modelling, atrial fibrillation, and the athlete's heart. Along the way, the methods developed as part of the research were also able to contribute novel and meaningful results to areas outside of cardiovascular research, further solidifying the robustness of the developed approaches.

Chapter 3 describes the comparative analysis of different ECG data formats to ascertain which is the best for ML modelling. We also aimed to address whether extracting signals from image ECGs and analysing the digitised signals was feasible within the context of ML. The data used for this analysis was taken from the PTB-XL dataset, an open-source database containing over 20,000 ECGs. From this data we defined the binary classification task of predicting whether an ECG was normal or showed MI, with three ECG data formats being tested: Signal ECGs, Image ECGs, and Extracted Signal ECGs. Several models were trained for each data format, using different data representations and different model tuning methods to provide as thorough a comparison as possible. The result of this analysis was that should the original signals be available, they should always be used for any analysis. Should these be absent, then Image ECGs should be used if they contain 2.5s of data for each lead and Extracted Signal ECGs should be used if the Image ECGs contain 10s of data for each lead. These results provided to first quantitative answer as to what the best ECG data format is for ML modelling, along with proving the viability of using extracted ECG signals for ML modelling.

Chapter 4 moved away from supervised ML to focus on developing a framework for applying GTM, a probabilistic unsupervised ML clustering technique. GTM was selected as the methodology of choice for this thesis due to its ability to identify complex non-linear relationships and generate robust data stratifications through its ability to handle high levels of uncertainty, along with the enhanced understanding of the output gained through the interpretable visualisations it produces.

The aim of this chapter from an analysis perspective was to generate the "Index Index": an index that ranks countries based upon the level of censorship they place on their populus. To achieve this, we collated data from several open-source data repositories, such as V-DEM, the World Press Freedom Index and the Committee to Protect Journalists. From these sources, we

created a dataset containing 178 different variables representing the academic, media and digital freedoms of a country. GTM was then applied to this data, which we then generated the rankings from using the normalised reference vectors extracted from the trained model. This therefore applied the rankings to each latent cluster, with each country inheriting the rank assigned to the cluster it was placed in. The use of this approach removed the subjective interpretation from the modelling process and provided the resulting Index Index with a greater degree of rigour than previous rankings. Not only did this work result in novel findings that led to a peer-reviewed published work [81], but it also served the purpose within the context of the thesis by facilitating the development and validation of a clustering methodology and workflow to serve as a blueprint for Chapters 5 and 7.

In Chapter 5, we aimed to develop a novel methodology for the identification of phenotypes within AF populations. The outlined methodology is built upon the developments made within Chapter 4, with the inclusion of hierarchical clustering being applied to the reference vectors instead, to generate the macro-clusters in the latent space. To develop and validate this methodology, we used two datasets: the first was the UK Biobank as it represented the general population; the second was the MIMIC-IV database as it represented a critically ill population. Our proposed methodology was able to identify 5 and 4 clinically relevant phenotypes respectively when applied to the two datasets. It demonstrated the ability of such an approach to identify clinical phenotypes of AF, which could enable prevention and treatment programs specific to each phenotype.

Both Chapters 6 and 7 aimed to address the same clinical area of the athlete's heart. Chapter 6 provides an in-depth review of the current state of AI applications within this area, as well as determining key avenues for future research. This was an important step as the area itself is still in its infancy, so understanding the full research landscape allows for the identification of the most effective research directions moving forward. The results highlighted that there was a clear desire for AI applications within the area, with most of the research implementing AI focusing on predictive modelling. The review highlighted several limitations such as small, unlabelled datasets that restrict the current approaches. However, the final prognosis for the future of the research area is very positive, with key areas for potential development being identified. The results of this review provided proven novelty to the area, with the work leading to a peer-reviewed publication [61].

One such avenue was explored within Chapter 7. The aim of this chapter was to implement unsupervised learning, an area identified in the review that could add key findings to the area. More specifically, we aimed to identify clinically relevant sub-groups with a healthy athlete

population consisting of 854 ECGs from 611 athletes. This was achieved by further developing the methodology outlined in Chapter 5. The hierarchical clustering element was replaced by a constrained variant, with the constraint being derived from the magnification factors derived from the trained GTM model. By using the magnification factors this way, we were able to generate a simpler, more interpretable output. It is also the first approach to directly use magnification factors to influence the identification of macro-clusters within a GTM latent space.

### 8.2. Future Work

While the work within this thesis is promising and provides various novel contributions, with further time and resources the work could be further developed and improved. Starting with Chapter 3, analysis was carried out successfully and thoroughly to identify the optimal ECG data format for ML modelling. However, there is a potential limitation in these results in that the prediction task carried out was a binary classification between normal ECGs and MI ECGs. To address this, future work should focus on defining a multi-class problem to re-test the same data formats to further validate the results. This could again be carried out using the same PTB-XL data, but remove the restriction applied to only include normal and MI data.

There are several interesting ways in which the analysis in Chapter 5 could be further progressed. First, the genomic data used within the analysis proved vital in stratifying participants into phenotypes, however, the interpretation of this is hampered due to the absence of the PCA loadings. By obtaining these PCA loadings, this analysis would unlock a new level of understanding of the results as it would improve the interpretability of the GTM output, providing insight into how specific genetic profiles can influence each phenotype. This could allow for research into causal relationships behind the different types of AF. Furthermore, further work could be done to validate the AF phenotypes generated as part of this paper by applying the methodology to different AF populations.

Another area of future work that would generate novel results would be to use GTM through time (GTM-TT) [47] to analyse how a subject changes within the latent space over time. This could be applied to both the AF analysis (on both the UK Biobank and MIMIC-IV data) as well as the athletes' ECG data. By developing such an approach, it would provide the user with the means to monitor an athlete or patient's clinical trajectory or disease progression, providing crucial updates sooner and improving outcomes.

An additional area for future work lies in the methodological workflow developed within Chapter 7. As already highlighted, there are known issues with certain parts of the data extraction relating to R wave peak detection. By addressing this, it would improve the accuracy of the features extracted, in turn leading to a more informative latent space. To further assist in this goal,

implementing t-GTM [214] would also likely result in the latent space better representing the relationships in the data space due to its better ability to handle outliers. In addition to this, work could be done to the feature generation section of the methodology to also extract amplitude and distance measurements, such as P wave amplitude or ST segment elevation, to be used as modelling variables. Being able to identify events such as T wave inversion, a potential sign of heart muscle disease [215], could lead to more accurate sub-groups and add functionality to the process to increase the capability of identifying potential issues.

# References


1 Roth GA, Mensah GA, Johnson CO, *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol*. 2020;76:2982–3021.

2 Vaduganathan M, Mensah GA, Turco JV, *et al.* The Global Burden of Cardiovascular Diseases and Risk. *J Am Coll Cardiol*. 2022;80:2361–71.

3 WHO. Cardiovascular diseases (CVDs). 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

4 Luengo-Fernandez R, Walli-Attaei M, Gray A, *et al.* Economic burden of cardiovascular diseases in the European Union: a population-based cost study. *Eur Heart J*. 2023;44:4752–67.

5 Dunbar SB, Khavjou OA, Bakas T, *et al.* Projected Costs of Informal Caregiving for Cardiovascular Disease: 2015 to 2035: A Policy Statement From the American Heart Association. *Circulation*. 2018;137:e558–77.

6 Kendir C, van den Akker M, Vos R, *et al.* Cardiovascular disease patients have increased risk for comorbidity: A cross-sectional study in the Netherlands. *Eur J Gen Pract*. 2018;24:45–50.

7 Buddeke J, Bots ML, Van Dis I, *et al.* Comorbidity in patients with cardiovascular disease in primary care: A cohort study with routine healthcare data. *Br J Gen Pract*. 2019;69:E398–406.

8 Cruz-Ávila HA, Vallejo M, Martínez-García M, *et al.* Comorbidity Networks in Cardiovascular Diseases. *Front Physiol*. 2020;11:1–19.

9 Jafari M, Shoeibi A, Khodatars M, *et al.* Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Comput Biol Med*. 2023;160:106998.

10 Baptiste DL, Turkson-Ocran RA, Ogungbe O, *et al.* Heterogeneity in Cardiovascular Disease Risk Factor Prevalence Among White, African American, African Immigrant, and Afro-Caribbean Adults: Insights From the 2010–2018 National Health Interview Survey. *J Am Heart Assoc*. 2022;11. doi: 10.1161/JAHA.122.025235

11 Simonetto C, Rospleszcz S, Kaiser JC, *et al.* Heterogeneity in coronary heart disease risk. *Sci Rep*. 2022;12:1–9.

12 Johnson KW, Torres Soto J, Glicksberg BS, *et al.* Artificial Intelligence in Cardiology. *J Am Coll Cardiol*. 2018;71:2668–79.

13 Krittanawong C, Zhang HJ, Wang Z, *et al.* Artificial Intelligence in Precision Cardiovascular Medicine. *J Am Coll Cardiol*. 2017;69:2657–64.

14 Weng SF, Reps J, Kai J, *et al.* Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12:e0174944.

15 Leiner T, Rueckert D, Suinesiaputra A, *et al.* Machine learning in cardiovascular magnetic resonance: Basic concepts and applications. *J Cardiovasc Magn Reson*. 2019;21:1–14.

16 Pham H, Egorov K, Kazakov A, *et al.* Machine learning-based detection of cardiovascular disease using ECG signals: performance vs. complexity. *Front Cardiovasc Med*. 2023;10:1–11.

17 Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Mark*. 2021;31:685–95.

18 Than MP, Pickering JW, Sandoval Y, *et al.* Machine Learning to Predict the Likelihood of Acute Myocardial Infarction. *Circulation*. 2019;140:899–909.

19 Dritsas E, Trigka M. Stroke Risk Prediction with Machine Learning Techniques. *Sensors*. 2022;22. doi: 10.3390/s22134670

20 Pachiyannan P, Alsulami M, Alsadie D, *et al.* A Cardiac Deep Learning Model (CDLM) to Predict and Identify the Risk Factor of Congenital Heart Disease. *Diagnostics*. 2023;13:1–21.

21 Marcus GM. The Apple Watch can detect atrial fibrillation: so what now? *Nat Rev Cardiol*. 2020;17:135–6.

22 Petersson L, Larsson I, Nygren JM, *et al.* Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. *BMC Health Serv Res*. 2022;22:1–16.

23 Williams MC, Bednarski BP, Pieszko K, *et al.* Unsupervised learning to characterize patients with known coronary artery disease undergoing myocardial perfusion imaging. *Eur J Nucl Med Mol Imaging*. 2023;50:2656–68.

24 Sun J, Guo H, Wang W, *et al.* Identifying novel subgroups in heart failure patients with unsupervised machine learning: A scoping review. *Front Cardiovasc Med*. 2022;9:1–12.

25 Gygi JP, Kleinstein SH, Guan L. Predictive overfitting in immunological applications: Pitfalls and solutions. *Hum Vaccines Immunother*. 2023;19. doi: 10.1080/21645515.2023.2251830

26 Quer G, Arnaout R, Henne M, *et al.* Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. *J Am Coll Cardiol*. 2021;77:300–13.

27 khan B, Fatima H, Qureshi A, *et al.* Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. *Biomed Mater Devices*. 2023;1:731–8.

28 Doshi-Velez F, Kortz M, Budish R, *et al.* Accountability of AI Under the Law: The Role of Explanation. *SSRN Electron J*. Published Online First: 2017. doi: 10.2139/ssrn.3064761

29 Olier I, Ortega-Martorell S, Pieroni M, *et al.* How machine learning is impacting research in atrial fibrillation: Implications for risk prediction and future management. *Cardiovasc Res*. 2021;117:1700–17.

30 Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65:386–408.

31 Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput Sci*. 2021;2:420.

32 Uzair M, Jamil N. Effects of Hidden Layers on the Efficiency of Neural networks. *2020 IEEE 23rd International Multitopic Conference (INMIC)*. IEEE 2020:1–6. https://doi.org/10.1109/INMIC50486.2020.9318195

33 Alom MZ, Taha TM, Yakopcic C, *et al.* The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. Published Online First: 2018.

34 Zhang Q, Yang LT, Chen Z, *et al.* A survey on deep learning for big data. *Inf Fusion*. 2018;42:146–57.

35 Yamashita R, Nishio M, Do RKG, *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9:611–29.

36 Indolia S, Goswami AK, Mishra SP, *et al.* Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. *Procedia Comput Sci*. 2018;132:679–88.

37 Mehdi CA, Nour-Eddine J, Mohamed E. Regularization in CNN: A Mathematical Study for L1, L2 and Dropout Regularizers. 2023:442–50. https://doi.org/10.1007/978-3-031-26384-2_38

38 Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Journal Pract*. 2015;10:730–43.

39 Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–58.

40 Draelos RL, Carin L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv Prepr*. 2020;1–20.

41 Zhou B, Khosla A, Lapedriza A, *et al.* Learning Deep Features for Discriminative Localization. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016;2016-Decem:2921–9.

42 Selvaraju RR, Cogswell M, Das A, *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis*. 2020;128:336–59.

43 Patel AA. *Hands-On Unsupervised Learning Using Python*. 2019. https://www.oreilly.com/library/view/hands-on-unsupervised-learning/9781492035633/

44 Ezugwu AE, Ikotun AM, Oyelade OO, *et al.* A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Eng Appl Artif Intell*. 2022;110:104743.

45 Ward Jr JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58:236–44.

46 Bishop CM, Svensén M, Williams CKI. GTM: The Generative Topographic Mapping. *Neural Comput*. 1998;10:215–34.

47 Olier I, Vellido A. Advances in clustering and visualization of time series using GTM through time. *Neural Networks*. 2008;21:904–13.

48 Kohonen T. *Self-Organizing Maps*. Berlin, Heidelberg: Springer Berlin Heidelberg 2001. https://doi.org/10.1007/978-3-642-56927-2

49 Bishop CM, Svensen M, Williams CKI. Magnification factors for the GTM algorithm. *IEE Conf Publ*. 1997;64–9.

50 Vellido A, Lisboa PJG, Meehan K. The generative topographic mapping as a principal model for data visualization and market segmentation: an electronic commerce case study. *International-Journal-of-Computers,-Systems-and-Signals*. 2000;1:119–38.

51 Maniyar DM, Nabney IT. Visual data mining using principled projection algorithms and information visualization techniques. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM 2006:643–8. https://doi.org/10.1145/1150402.1150481

52 Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27:861–74.

53 Yadav S, Shukla S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on

Colossal Datasets for Quality Classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. IEEE 2016:78–83. https://doi.org/10.1109/IACC.2016.25

54 Bellolio MF, Serrano LA, Stead LG. Understanding statistical tests in the medical literature: which test should I use? *Int J Emerg Med*. 2008;1:197–9.

55 McHugh ML. The Chi-square test of independence. *Biochem Medica*. 2013;143–9.

56 Bewick V, Cheek L, Ball J. Statistics review 10: further nonparametric methods. *Crit Care*. 2004;8:196–9.

57 Randazzo V, Puleo E, Paviglianiti A, *et al.* Development and Validation of an Algorithm for the Digitization of ECG Paper Images. *Sensors*. 2022;22:7138.

58 Bond RR, Finlay DD, Nugent CD, *et al.* A review of ECG storage formats. *Int J Med Inform*. 2011;80:681–97.

59 Somani S, Russak AJ, Richter F, *et al.* Deep learning and the electrocardiogram: Review of the current state-of-the-art. *Europace*. 2021;23:1179–91.

60 Ebrahimi Z, Loni M, Daneshtalab M, *et al.* A review on deep learning methods for ECG arrhythmia classification. *Expert Syst with Appl X*. 2020;7:100033.

61 Bellfield RAA, Ortega-Martorell S, Lip GYH, *et al.* The Athlete's Heart and Machine Learning: A Review of Current Implementations and Gaps for Future Research. *J Cardiovasc Dev Dis*. 2022;9:382.

62 Kwon J myoung, Kim KH, Medina-Inojosa J, *et al.* Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *J Hear Lung Transplant*. 2020;39:805–14.

63 Makimoto H, Höckmann M, Lin T, *et al.* Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep*. 2020;10:8445.

64 Mishra S, Khatwani G, Patil R, *et al.* ECG Paper Record Digitization and Diagnosis Using Deep Learning. *J Med Biol Eng*. 2021;41:422–32.

65 Swamy P, Jayaraman S, Girish Chandra M. An improved method for digital time series signal generation from scanned ECG records. *ICBBT 2010 - 2010 Int Conf Bioinforma Biomed Technol*. 2010;400–3.

66 Fortune JD, Coppa NE, Haq KT, *et al.* Digitizing ECG image: A new method and open-source software code. *Comput Methods Programs Biomed*. 2022;221. doi: 10.1016/j.cmpb.2022.106890

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

68 Baydoun M, Safatly L, Hassan OKA, *et al.* High Precision Digitization of Paper-Based ECG Records: A Step Toward Machine Learning. *IEEE J Transl Eng Heal Med*. 2019;7:1–9.

69 Wagner P, Strodthoff N, Bousseljot R-D, *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Sci Data*. 2020;7:154.

70 Goldberger AL, Amaral LA, Glass L, *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101. doi: 10.1161/01.cir.101.23.e215

71 Tereshchenko LG, Josephson ME. Frequency content and characteristics of ventricular conduction. *J Electrocardiol*. 2015;48:933–7.

72 Strodthoff N, Wagner P, Schaeffter T, *et al.* Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE J Biomed Heal Informatics*. 2021;25:1519–28.

73 Śmigiel S, Pałczyński K, Ledziński D. ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset. *Entropy*. 2021;23:1121.

74 Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13:281–305.

75 Li L, Jamieson K, DeSalvo G, *et al.* Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *J Mach Learn Res*. 2016;18:1–52.

76 Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst*. 2012;4:2951–9.

77 Karnin Z, Koren T, Somekh O. Almost Optimal Exploration in Multi-Armed Bandits. In: Dasgupta S, McAllester D, eds. *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, Georgia, USA: PMLR 2013:1238–46. https://proceedings.mlr.press/v28/karnin13.html

78 Wang S, Zhang S, Li Z, *et al.* Automatic digital ECG signal extraction and normal QRS recognition from real scene ECG images. *Comput Methods Programs Biomed*. 2020;187. doi: 10.1016/j.cmpb.2019.105254

79 Makowski D, Pham T, Lau ZJ, *et al.* NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav Res Methods*. 2021;53:1689–96.

80 Ito S, Cohen-Shelly M, Attia ZI, *et al.* Correlation between artificial intelligence-enabled electrocardiogram and echocardiographic features in aortic stenosis. *Eur Hear J - Digit Heal*. 2023;1–6.

81 Ortega-Martorell S, Bellfield RAA, Harrison S, *et al.* Mapping the global free expression landscape using machine learning. *SN Appl Sci*. 2023;5. doi: 10.1007/s42452-023-05554-x

82 United Nations. Universal Declaration of Human Rights. 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed 2 December 2022)

83 European Court of Human Rights. *European Convention on Human Rights*. 1950. www.conventions.coe.int

84 Organization of American States. *American Convention on Human Rights 'Pact of San Jose, Costa Rica'*. 1969.

85 African States members of the Organisation of African Unity. *African Commission on Human and Peoples' Rights Legal instruments*. 1979.

86 Bollinger LC, Callamard A. *Regardless of frontiers: global freedom of expression in a troubled world*. 2021.

87 Stevenson C. Breaching the {Great Firewall}: {China}'s Internet Censorship and the Quest for Freedom of Expression in a Connected World. *Bost Coll Int \& Comp Law Rev*. 2007;30:531–58.

88 Pring GW, Canan P. *{SLAPPs}: Getting Sued for Speaking Out*. Temple University Press

1996. https://tupress.temple.edu/books/slapps

89 Greer D. The incidence of the terror during the {French Revolution}: a statistical interpretation. *Harvard Hist Monogr*. 1935;196.

90 Landman T, Schwarz K. *Human rights indicators and implementation*. Edward Elgar Publishing 2022.

91 Restrepo JA, Spagat M, Vargas JF. Special Data Feature; The Severity of the {Colombian} Conflict: Cross-Country Datasets Versus New Micro-Data. *J Peace Res*. 2016;43:99–115.

92 Ball P, Asher J, Sulmont D, *et al.* How many {Peruvians} have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000. *Am Assoc Adv Sci*. Published Online First: 2003.

93 Landman T, Carvalho E. *Measuring human rights*. Routledge 2010.

94 Aryan S, Halderman JA. Internet Censorship in Iran: A First Look. *3rd USENIX Workshop on Free and Open Communications on the Internet*. 2013.

95 Dainotti A, Squarcella C, Aben E, *et al.* Analysis of country-wide internet outages caused by censorship. *IEEE/ACM Trans Netw*. 2014;22:1964–77.

96 Niaki AA, Cho S, Weinberg Z, *et al.* {ICLab}: A global, longitudinal internet censorship measurement platform. *Proc - IEEE Symp Secur Priv*. 2020;2020-May:135–51.

97 Coppedge M, Gerring J, Knutsen CH, *et al.* '{V-Dem Codebook} v12' Varieties of Democracy ({V-Dem}) Project. 2022. https://www.v-dem.net/data/the-v-dem-dataset/

98 Lazarou E, Stanicek B. *Mapping threats to peace and democracy worldwide*. 2023. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/751422/EPRS_STU(2023)751422_EN.pdf

99 Radsch C, Paterson K. Submission from the Committee to Protect Journalists to the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. 2021. https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/disinformation/2-Civil-society-organisations/Committee-to-Protect-Journalists.pdf

100 Coppedge M, Gerring J, Knutsen CH, *et al.* '{V-Dem Codebook} v12' Varieties of Democracy ({V-Dem}) Project. 2022.

101 Reporters Without Borders. World Press Freedom Index. 2022. https://rsf.org/en/index?year=2022

102 Committee to Protect Journalists. {CPJ}'s database of attacks on the press. 2022. https://cpj.org/data/

103 UNESCO. {UNESCO} observatory of killed journalists. 2022. https://en.unesco.org/themes/safety-journalists/observatory

104 NetBlocks. COST: The NetBlocks Cost of Shutdown Tool. 2022. https://netblocks.org/projects/cost

105 International Telecommunication Union ITU. Global Cybersecurity Index. 2022. https://www.itu.int/en/ITU-D/Cybersecurity/Pages/global-cybersecurity-index.aspx

106 Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–625.

107 McInnes L, Healy J, Saul N, *et al.* UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw*. 2018;3:861.

108 Horvath D, Marcou G, Varnek A. Generative topographic mapping in drug design. *Drug Discov Today Technol*. 2019;32–33:99–107.

109 Orlov AA, Khvatov E V., Koruchekov AA, *et al.* Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds. *Mol Inform*. 2019;38:1–12.

110 Bathen TF, Engan T, Krane J, *et al.* Analysis and classification of proton NMR spectra of lipoprotein fractions from healthy volunteers and patients with cancer or CHD. *Anticancer Res*. 2000;20:2393–408.

111 Gaspar HA, Hübel C, Breen G. Biological Pathways and Drug Gene-Sets: Analysis and Visualization. *Eur Neuropsychopharmacol*. 2019;29:S834.

112 Olier I, Amengual J, Vellido A. A variational Bayesian approach for the robust analysis of the cortical silent period from EMG recordings of brain stroke patients. *Neurocomputing*. 2011;74:1301–14.

113 Polyakova A, Mukharamova S, Yermolaev O, *et al.* Automated Recognition of Tree Species Composition of Forest Communities Using Sentinel-2 Satellite Data. *Remote Sens*. 2023;15. doi: 10.3390/rs15020329

114 Vellido A, Martí E, Comas J, *et al.* Exploring the ecological status of human altered streams through Generative Topographic Mapping. *Environ Model Softw*. 2007;22:1053–65.

115 Feng J, Liu Z, Feng L. Identifying opportunities for sustainable business models in manufacturing: Application of patent analysis and generative topographic mapping. *Sustain Prod Consum*. 2021;27:509–22.

116 Irum SA, Laila AS. Media censorship: Freedom versus responsibility. *J Law Confl Resolut*. 2015;7:21–4.

117 Clark M, Grech A. *JOURNALISTS UNDER PRESSURE Unwarranted interference, fear and self-censorship in Europe*. 2017.

118 Saito Y, Omae Y, Nagashima K, *et al.* Phenotyping of atrial fibrillation with cluster analysis and external validation. *Heart*. 2023;heartjnl-2023-322447.

119 Zhang J, Johnsen SP, Guo Y, *et al.* Epidemiology of Atrial Fibrillation. *Card Electrophysiol Clin*. 2021;13:1–23.

120 Lip GYH, Genaidy A, Tran G, *et al.* Improving Stroke Risk Prediction in the General Population: A Comparative Assessment of Common Clinical Rules, a New Multimorbid Index, and Machine-Learning-Based Algorithms. *Thromb Haemost*. 2022;122:142–50.

121 Romiti GF, Proietti M, Bonini N, *et al.* Clinical Complexity Domains, Anticoagulation, and Outcomes in Patients with Atrial Fibrillation: A Report from the GLORIA-AF Registry Phase II and III. *Thromb Haemost*. 2022;122:2030–41.

122 Olier I, Ortega-Martorell S, Pieroni M, *et al.* How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management. *Cardiovasc Res*. 2021;117:1700–17.

123 Chung KF, Adcock IM. How variability in clinical phenotypes should guide research into disease mechanisms in asthma. *Ann Am Thorac Soc*. 2013;10. doi: 10.1513/AnnalsATS.201304-087AW

124 Romiti GF, Pastori D, Rivera-Caravaca JM, *et al.* Adherence to the 'Atrial Fibrillation Better Care' Pathway in Patients with Atrial Fibrillation: Impact on Clinical Outcomes—A Systematic Review and Meta-Analysis of 285,000 Patients. *Thromb Haemost*. 2022;122:406–14.

125 Chao T-F, Joung B, Takahashi Y, *et al.* 2021 Focused Update Consensus Guidelines of the Asia Pacific Heart Rhythm Society on Stroke Prevention in Atrial Fibrillation: Executive Summary. *Thromb Haemost*. 2022;122:020–47.

126 Vitolo M, Proietti M, Shantsila A, *et al.* Clinical phenotype classification of atrial fibrillation patients using cluster analysis and associations with trial-adjudicated outcomes. *Biomedicines*. 2021;9:1–11.

127 Watanabe E, Inoue H, Atarashi H, *et al.* Clinical phenotypes of patients with non-valvular atrial fibrillation as defined by a cluster analysis: A report from the J-RHYTHM registry. *IJC Hear Vasc*. 2021;37. doi: 10.1016/j.ijcha.2021.100885

128 Proietti M, Vitolo M, Harrison SL, *et al.* Impact of clinical phenotypes on management and outcomes in European atrial fibrillation patients: a report from the ESC-EHRA EURObservational Research Programme in AF (EORP-AF) General Long-Term Registry. *BMC Med*. 2021;19:1–17.

129 Inohara T, Piccini JP, Mahaffey KW, *et al.* A Cluster Analysis of the Japanese Multicenter Outpatient Registry of Patients With Atrial Fibrillation. *Am J Cardiol*. 2019;124:871–8.

130 Ogawa H, An Y, Nishi H, *et al.* Characteristics and clinical outcomes in atrial fibrillation patients classified using cluster analysis: The Fushimi AF Registry. *Europace*. 2021;23:1369–79.

131 Bisson A, M. Fawzy A, Romiti GF, *et al.* Phenotypes and outcomes in non-anticoagulated patients with atrial fibrillation: An unsupervised cluster analysis. *Arch Cardiovasc Dis*. 2023;116:342–51.

132 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 2015;12:1–10.

133 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562:203–9.

134 Papadopoulou A, Harding D, Slabaugh G, *et al.* Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank. *medRxiv*. 2022;2022.10.28.22281669.

135 Johnson AEW, Bulgarelli L, Shen L, *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:1.

136 Ortega-Martorell S, Olier I, Johnston BW, *et al.* Sepsis-induced coagulopathy is associated with new episodes of atrial fibrillation in patients admitted to critical care in sinus rhythm. *Front Med*. 2023;10. doi: 10.3389/fmed.2023.1230854

137 Ortega-Martorell S, Pieroni M, Johnston BW, *et al.* Development of a Risk Prediction Model for New Episodes of Atrial Fibrillation in Medical-Surgical Critically Ill Patients Using the AmsterdamUMCdb. *Front Cardiovasc Med*. 2022;9. doi: 10.3389/fcvm.2022.897709

138 Allan V, Honarbakhsh S, Casas J-P, *et al.* Are cardiovascular risk factors also associated with the incidence of atrial fibrillation? *Thromb Haemost*. 2017;117:837–50.

139 Nso N, Bookani KR, Metzl M, *et al.* Role of inflammation in atrial fibrillation: A

comprehensive review of current knowledge. *J Arrhythmia*. 2021;37:1–10.

140 Mohanty S, Hall A, Mohanty P, *et al.* Being Asymptomatic With Atrial Fibrillation: Is It a Genetic Trait? *J Am Coll Cardiol*. 2016;67:677.

141 Kalarus Z, Mairesse GH, Sokal A, *et al.* Searching for atrial fibrillation: looking harder, looking longer, and in increasingly sophisticated ways. An EHRA position paper. *Europace*. 2023;25:185–98.

142 Alonso A, Krijthe BP, Aspelund T, *et al.* Simple Risk Model Predicts Incidence of Atrial Fibrillation in a Racially and Geographically Diverse Population: the CHARGE-AF Consortium. *J Am Heart Assoc*. 2013;2. doi: 10.1161/JAHA.112.000102

143 Lip GYH, Skjøth F, Nielsen PB, *et al.* Evaluation of the C2HEST Risk Score as a Possible Opportunistic Screening Tool for Incident Atrial Fibrillation in a Healthy Population (From a Nationwide Danish Cohort Study). *Am J Cardiol*. 2020;125:48–54.

144 Johnston BW, Chean CS, Duarte R, *et al.* Management of new onset atrial fibrillation in critically unwell adult patients: a systematic review and narrative synthesis. *Br J Anaesth*. 2022;128:759–71.

145 Bosch NA, Cimini J, Walkey AJ. Atrial Fibrillation in the ICU. *Chest*. 2018;154:1424–34.

146 O'Driscoll BR, Smith R. Oxygen Use in Critical Illness. *Respir Care*. 2019;64:1293–307.

147 Wu P, Gifford A, Meng X, *et al.* Mapping ICD-10 and ICD-10-CM Codes to phecodes: Workflow development and initial evaluation. *JMIR Med Informatics*. 2019;7:1–13.

148 van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45:1–67.

149 Handelsman DJ, Hirschberg AL, Bermon S. Circulating Testosterone as the Hormonal Basis of Sex Differences in Athletic Performance. *Endocr Rev*. 2018;39:803–29.

150 Qu X, Donnelly R. Sex hormone-binding globulin (Shbg) as an early biomarker and therapeutic target in polycystic ovary syndrome. *Int J Mol Sci*. 2020;21:1–17.

151 WHO. The top 10 causes of death. 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

152 British Heart Foundation. Facts and figures. 2021. https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures#:~:text=There are around 7.6 million,the single biggest killer worldwide

153 Wasfy MM, Hutter AM, Weiner RB. Sudden Cardiac Death in Athletes. *Methodist Debakey Cardiovasc J*. 2016;12:76.

154 Kerkhof DL, Lucas C, Corrado GD. Monitoring morphologic changes in male rowers using limited portable echocardiography performed by a frontline physician. *J Ultrasound Med*. 2018;37:2451–5.

155 Harmon KG, Drezner JA, Wilson MG, *et al.* Incidence of sudden cardiac death in athletes: A state-of-the-art review. Br. J. Sports Med. 2014;48:1185–92. https://doi.org/10.1136/bjsports-2014-093872

156 McKinney SM, Sieniek M, Godbole V, *et al.* International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.

157 Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. *Proc 10th INDIACom; 2016 3rd Int Conf Comput Sustain Glob Dev INDIACom 2016*.

2016;1310–5.

158 Katz DH, Deo RC, Aguilar FG, *et al.* Phenomapping for the Identification of Hypertensive Patients with the Myocardial Substrate for Heart Failure with Preserved Ejection Fraction. *J Cardiovasc Transl Res*. 2017;10:275–84.

159 Alizadehsani R, Hosseini MJ, Khosravi A, *et al.* Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. *Comput Methods Programs Biomed*. 2018;162:119–27.

160 Asselbergs FW, Meijboom FJ. Big data analytics in adult congenital heart disease: Why coding matters. Eur. Heart J. 2019;40:1078–80. https://doi.org/10.1093/eurheartj/ehz089

161 Jing L, Ulloa Cerna AE, Good CW, *et al.* A Machine Learning Approach to Management of Heart Failure Populations. *JACC Hear Fail*. 2020;8:578–87.

162 Agasthi P, Ashraf H, Pujari SH, *et al.* Artificial Intelligence Trumps TAVI2-SCORE and CoreValve Score in Predicting 1-Year Mortality Post-Transcatheter Aortic Valve Replacement. *Cardiovasc Revascularization Med*. 2021;24:33–41.

163 Adetiba E, Iweanya VC, Popoola SI, *et al.* Automated detection of heart defects in athletes based on electrocardiography and artificial neural network. *Cogent Eng*. 2017;4. doi: 10.1080/23311916.2017.1411220

164 Adetiba E, Onosenema EN, Akande V, *et al. Development of an ECG Smart Jersey Based on Next Generation Computing for Automated Detection of Heart Defects Among Athletes*. Springer International Publishing 2019. https://doi.org/10.1007/978-3-030-17935-9_47

165 Barbieri D, Chawla N, Zaccagni L, *et al.* Predicting cardiovascular risk in athletes: Resampling improves classification performance. *Int J Environ Res Public Health*. 2020;17:1–9.

166 Beavers DL, Chung EH. Wearables in Sports Cardiology. Clin. Sports Med. 2022;41:405–23. https://doi.org/10.1016/j.csm.2022.02.004

167 Bernardino G, Benkarim O, Sanz-de la Garza M, *et al.* Handling confounding variables in statistical shape analysis - application to cardiac remodelling. *Med Image Anal*. 2020;65:101792.

168 Castillo-Atoche A, Caamal-Herrera K, Atoche-Enseñat R, *et al.* Energy Efficient Framework for a AIoT Cardiac Arrhythmia Detection System Wearable during Sport. *Appl Sci*. 2022;12. doi: 10.3390/app12052716

169 Chang AC. Primary prevention of sudden cardiac death of the young athlete: The controversy about the screening electrocardiogram and its innovative artificial intelligence solution. *Pediatr Cardiol*. 2012;33:428–33.

170 Chatzakis I, Vassilakis K, Lionis C, *et al.* Electronic health record with computerized decision support tools for the purposes of a pediatric cardiovascular heart disease screening program in Crete. *Comput Methods Programs Biomed*. 2018;159:159–66.

171 Christ P, Rückert U. Identification of athletes during walking and jogging based on gait and electrocardiographic patterns. *Commun Comput Inf Sci*. 2014;452:240–57.

172 Claudino JG, Capanema D de O, de Souza TV, *et al.* Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. *Sport Med - Open*. 2019;5. doi: 10.1186/s40798-019-0202-3

173 Długosz D, Królak A, Eftestøl T, *et al.* ECG signal analysis for troponin level assessment

and coronary artery disease detection: The NEEDED study 2014. *Proc 2018 Fed Conf Comput Sci Inf Syst FedCSIS 2018*. 2018;15:1065–8.

174 Georgijević L, Andrić L. Electrocardiography in pre-participation screening and current guidelines for participation in competitive sports. *Srp Arh Celok Lek*. 2016;144:104–10.

175 Higgins JP, Ananaba IE, Higgins CL. Sudden cardiac death in young athletes: Preparticipation screening for underlying cardiovascular abnormalities and approaches to prevention. *Phys Sportsmed*. 2013;41:81–93.

176 Huang KC, Lin CE, Lin LY, *et al.* Data-driven clustering supports adaptive remodeling of athlete's hearts: An echocardiographic study from the Taipei Summer Universiade. *J Formos Med Assoc*. Published Online First: 2021. doi: 10.1016/j.jfma.2021.10.017

177 Hussain A, Zafar K, Baig AR. Fog-Centric IoT Based Framework for Healthcare Monitoring, Management and Early Warning System. *IEEE Access*. 2021;9:74168–79.

178 Laurino M, Piarulli A, Bedini R, *et al.* Comparative study of morphological ECG features classificators: An application on athletes undergone to acute physical stress. *Int Conf Intell Syst Des Appl ISDA*. 2011;242–6.

179 Lombardi G, Sorbo AR, Guida G, *et al.* Magnetocardiographic classification and non-invasive electro-anatomical imaging of outflow tract ventricular arrhythmias in recreational sport activity practitioners. *J Electrocardiol*. 2018;51:433–9.

180 Lucas C, Kerkhof DL, Briggs JE, *et al.* The use of echocardiograms in preparticipation examinations. *Curr Sports Med Rep*. 2017;16:77–83.

181 Mlynczak M, Krysztofiak H. Discovery of causal paths in cardiorespiratory parameters: A time-independent approach in elite athletes. *Front Physiol*. 2018;9:1455.

182 Narula S, Shameer K, Salem Omar AM, *et al.* Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. *J Am Coll Cardiol*. 2016;68:2287–95.

183 Rahman QA, Kanagalingam S, Pinheiro A, *et al.* What we found on our way to building a classifier: A critical analysis of the AHA screening questionnaire. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2013:225–36. https://doi.org/10.1007/978-3-319-02753-1_23

184 Rymarczyk T, Stanikowski A, Nita P. Wearable sensor array for biopotential measurements. *2019 Appl Electromagn Mod Eng Med PTZE 2019*. 2019;184–7.

185 Seshadri DR, Thom ML, Harlow ER, *et al.* Wearable Technology and Analytics as a Complementary Toolkit to Optimize Workload and to Reduce Injury Burden. *Front Sport Act Living*. 2021;2:1–17.

186 Van Eetvelde H, Mendonça LD, Ley C, *et al.* Machine learning methods in sport injury prediction and prevention: a systematic review. *J Exp Orthop*. 2021;8. doi: 10.1186/s40634-021-00346-x

187 Vergani V, Lazzeroni D, Peretto G. Bridging the gap between hypertrabeculation phenotype, noncompaction phenotype and left ventricular noncompaction cardiomyopathy. *J Cardiovasc Med*. 2020;21:192–9.

188 Viviers PL, Kirby JAH, Viljoen JT, *et al.* The Diagnostic Utility of Computer-Assisted Auscultation for the Early Detection of Cardiac Murmurs of Structural Origin in the Periodic Health Evaluation. *Sports Health*. 2017;9:341–5.

189 Dockerill C, Lapidaire W, Lewandowski AJ, *et al.* Cardiac remodelling and exercise: What happens with ultra-endurance exercise? *Eur J Prev Cardiol*. 2020;27:1464–6.

190 Liu X, Faes L, Kale AU, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1:e271–97.

191 Marr B. The Big Risks Of Big Data In Sports. Forbes. 2017. https://www.forbes.com/sites/bernardmarr/2017/04/28/the-big-risks-of-big-data-in-sports/%3E

192 Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng Med Biol Mag*. 2001;20:45–50.

193 Lugovaya T. *Biometric human identification based on electrocardiogram*. 2005. https://doi.org/10.13026/C2J01F

194 GreenWald SD. *Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information.* 1990.

195 Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. *Comput Cardiol*. 1983.

196 Laguna P, Mark RG, Goldberg A, *et al.* Database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. *Comput Cardiol*. 1997;24:673–6.

197 Jager F, Taddei A, Moody GB, *et al.* Long-term ST database: A reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia. *Med Biol Eng Comput*. 2003;41:172–82.

198 Erdenebayar U, Kim H, Park J-U, *et al.* Automatic Prediction of Atrial Fibrillation Based on Convolutional Neural Network Using a Short-term Normal Electrocardiogram Signal. *J Korean Med Sci*. 2019;34. doi: 10.3346/jkms.2019.34.e64

199 Flanagan H, Cooper R, George KP, *et al.* The athlete's heart: insights from echocardiography. *Echo Res Pract*. 2023;10:1–17.

200 Alasti M, Omidvar B, Jadbabaei MH. Heart and athlete. *J Tehran Univ Hear Cent*. 2010;5:1–8.

201 Parry-Williams G, Gati S, Sharma S. The heart of the ageing endurance athlete: The role of chronic coronary stress. *Eur Heart J*. 2021;42:2737–44.

202 Alattar A, Maffulli N. The Validity of Adding ECG to the Preparticipation Screening of Athletes An Evidence Based Literature Review. *Transl Med @ UniSa*. 2015;11:2–13.

203 Krivenko GS, Ribeiro ER, Walker S, *et al.* Feasibility of electrocardiogram screening in the USA prior to high school sport participation. *Prog Pediatr Cardiol*. 2022;65:101522.

204 Uberoi A, Stein R, Perez M V., *et al.* Interpretation of the electrocardiogram of young athletes. *Circulation*. 2011;124:746–57.

205 Pelliccia A, Maron BJ, Culasso F, *et al.* Clinical significance of abnormal electrocardiographic patterns in trained athletes. *Circulation*. 2000;102:278–84.

206 Prakash K, Sharma S. Interpretation of the Electrocardiogram in Athletes. *Can J Cardiol*. 2016;32:438–51.

207 Sharma S, Drezner JA, Baggish A, *et al.* International recommendations for

electrocardiographic interpretation in athletes. *Eur Heart J*. 2018;39:1466–80.

208 Palermi S, Vecchiato M, Saglietto A, *et al.* Unlocking the potential of artificial intelligence in sports cardiology: does it have a role in evaluating athlete's heart? *Eur J Prev Cardiol*. Published Online First: 10 January 2024. doi: 10.1093/eurjpc/zwae008

209 Meek S, Morris F. ABC of clinical electrocardiography: Introduction. I---Leads, rate, rhythm, and cardiac axis. *BMJ*. 2002;324:415–8.

210 Ashley EA, Niebauer J. *Cardiology Explained*. London: Remedica 2004.

211 Martinez JP, Almeida R, Olmos S, *et al.* A Wavelet-Based ECG Delineator: Evaluation on Standard Databases. *IEEE Trans Biomed Eng*. 2004;51:570–81.

212 Cipriano X, Vellido A, Cipriano J, *et al.* Influencing factors in energy use of housing blocks: a new methodology, based on clustering and energy simulations, for decision making in energy refurbishment projects. *Energy Effic*. 2017;10:359–82.

213 Eppstein D, Paterson MS, Yao FF. On Nearest-Neighbor Graphs. *Discrete Comput Geom*. 1997;17:263–82.

214 Vellido A. Missing data imputation through GTM as a mixture of t-distributions. *Neural Networks*. 2006;19:1624–35.

215 Stein R, Malhotra A. T wave inversions in athletes: A variety of scenarios. *J Electrocardiol*. 2015;48:415–9.

# Glossary of Aggregated Terms

| Abbreviation | Definition |
|---|---|
| **1D** | 1-dimensional |
| **2D** | 2-dimensional |
| **AF** | Atrial fibrillation |
| **AI** | Artificial intelligence |
| **AKI** | Acute kidney injury |
| **ANN** | Artificial neural network |
| **ANOVA** | Analysis of variance |
| **ARDS** | acute respiratory distress syndrome |
| **AUC** | Area under the receiver operating characteristics curve |
| **BMI** | Body mass index |
| **BN** | Batch normalisation |
| **BPM** | Beats per minute |
| **CAD** | Coronary artery disease |
| **CAM** | Class activation mapping |
| **CHARGE-AF** | The Cohorts for Heart and Ageing Research in Genomic Epidemiology AF |
| **CMRI** | Cardiac magnetic resonance imaging |
| **CNN** | Convolutional neural network |
| **COST** | Cost of shutdown |
| **CPJ** | Committee to Protect Journalists |
| **CT** | Computed tomography |
| **cTnI** | Cardiac troponin |
| **CVD** | Cardiovascular disease |
| **DL** | Deep learning |
| **DNN** | Deep neural network |
| **ECG** | Electrocardiogram |
| **EM** | Expectation-maximisation algorithm |
| **ESC** | European Society of Cardiology |

| | |
|---|---|
| **FCNN** | Fully connected neural network |
| **GCI** | Global cybersecurity index |
| **GCS** | Glasgow coma scale |
| **GDP** | Gross domestic product |
| **Grad-CAM** | Gradient-weighted class activation mapping |
| **GTM** | Generative topographic mapping |
| **HDI** | Human development index |
| **HiResCAM** | High resolution class activation mapping |
| **IBI** | Interbeat interval |
| **ICU** | Intensive care unit |
| **IQR** | Interquartile range |
| **KNN** | K nearest neighbours |
| **LDA** | Linear discriminant analysis |
| **LLM** | Large language models |
| **LSTM** | Long-teerm short memory neural networks |
| **MI** | Myocardial infarction |
| **MIMIC-IV** | Medical Information Mart for Intensive Care IV |
| **ML** | Machine learning |
| **MLP** | Multi-layer perceptron |
| **MRI** | Magnetic resonance imaging |
| **MSE** | Mean squared error |
| **NN** | Neural Networks |
| **NORM** | Normal ECGs |
| **PCA** | Principal component analysis |
| **PDF** | Portable document format |
| **PNG** | Portable network graphics |
| **RBF** | Radial basis function |
| **ReLU** | Rectified linear unit |
| **RNN** | Recurrent neural network |
| **RSF** | Reporters without borders |

| | |
|---|---|
| **sd** | Standard deviation |
| **SNP** | Single nucleotide polymorphisms |
| **SOM** | Self-organising map algorithm |
| **t-SNE** | t-distibuted stochastic neighbourhood embedding |
| **UMAP** | Uniform manifold approximation and projection |
| **UN** | United Nations |
| **UNFPA** | United Nations population fund |
| **V-Dem** | Varieties of democracy |

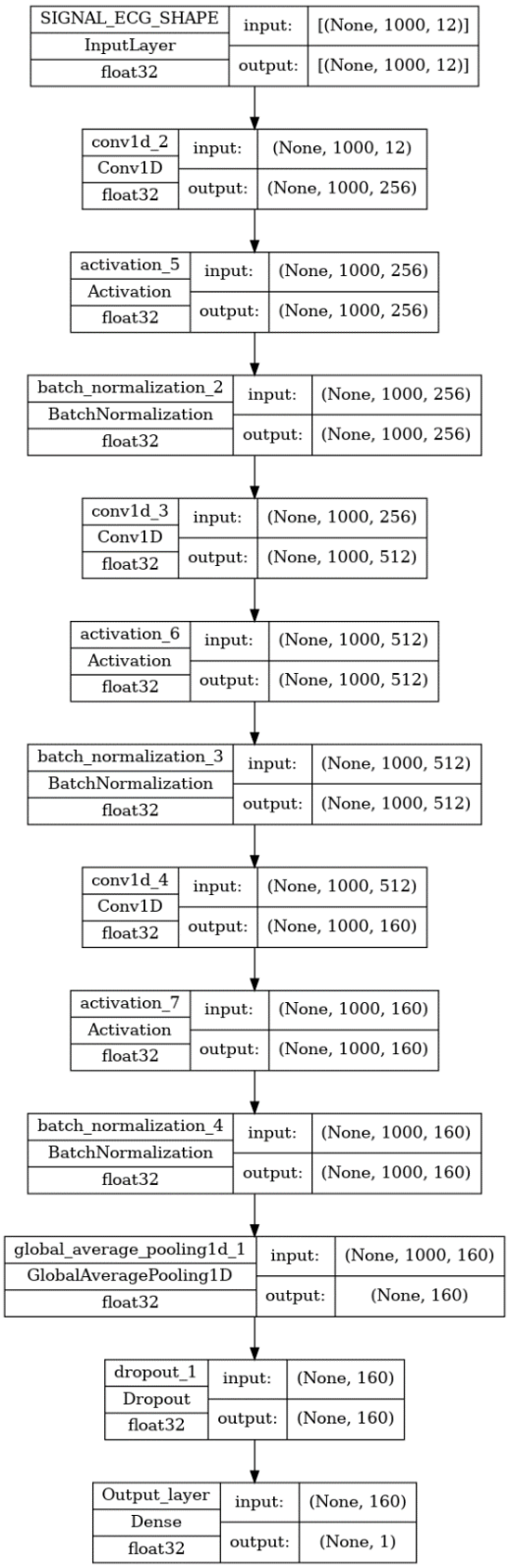



*Figure S1. Signal ECG model structure trained on arrangement A data from the conservative cohort.*

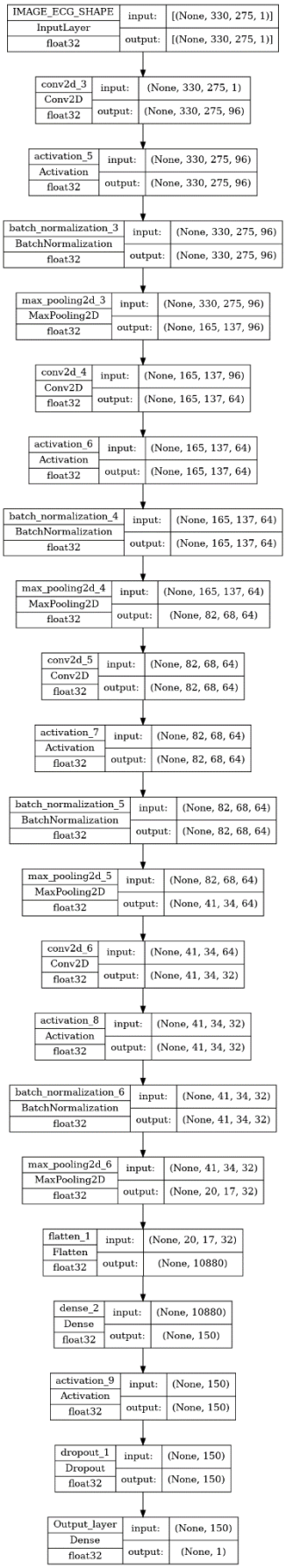



*Figure S2. Image ECG model structure trained on arrangement A data from the conservative cohort.*

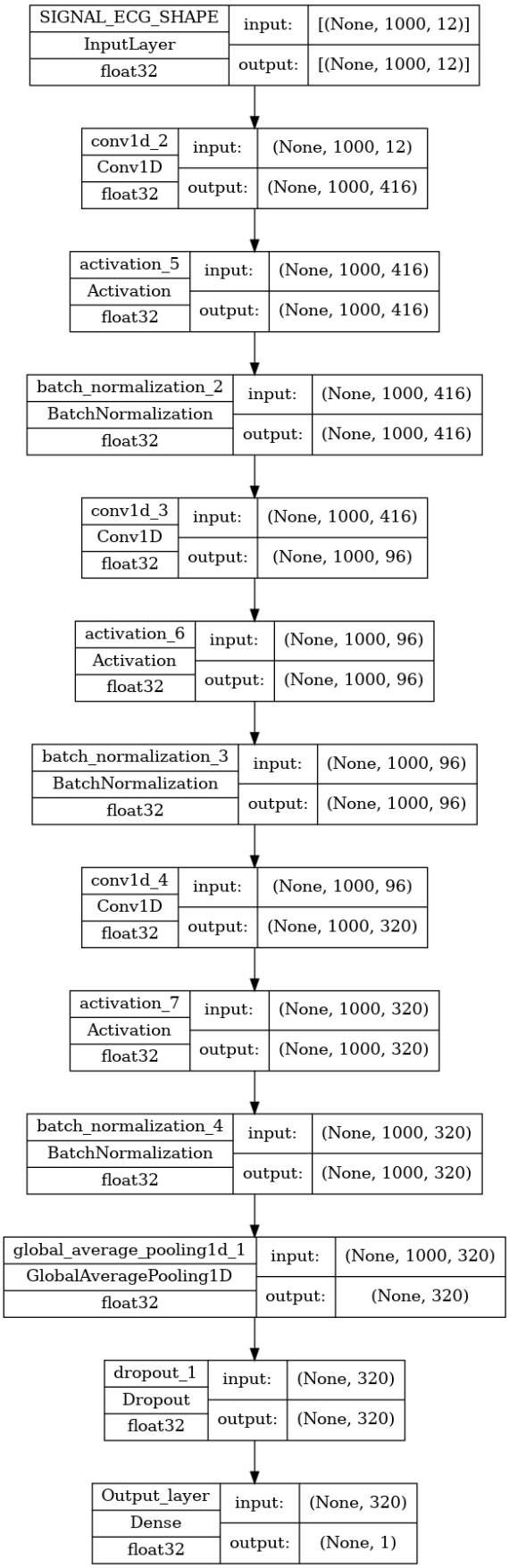



*Figure S3. Extracted Signal ECG model structure trained on arrangement A data from the conservative cohort.*

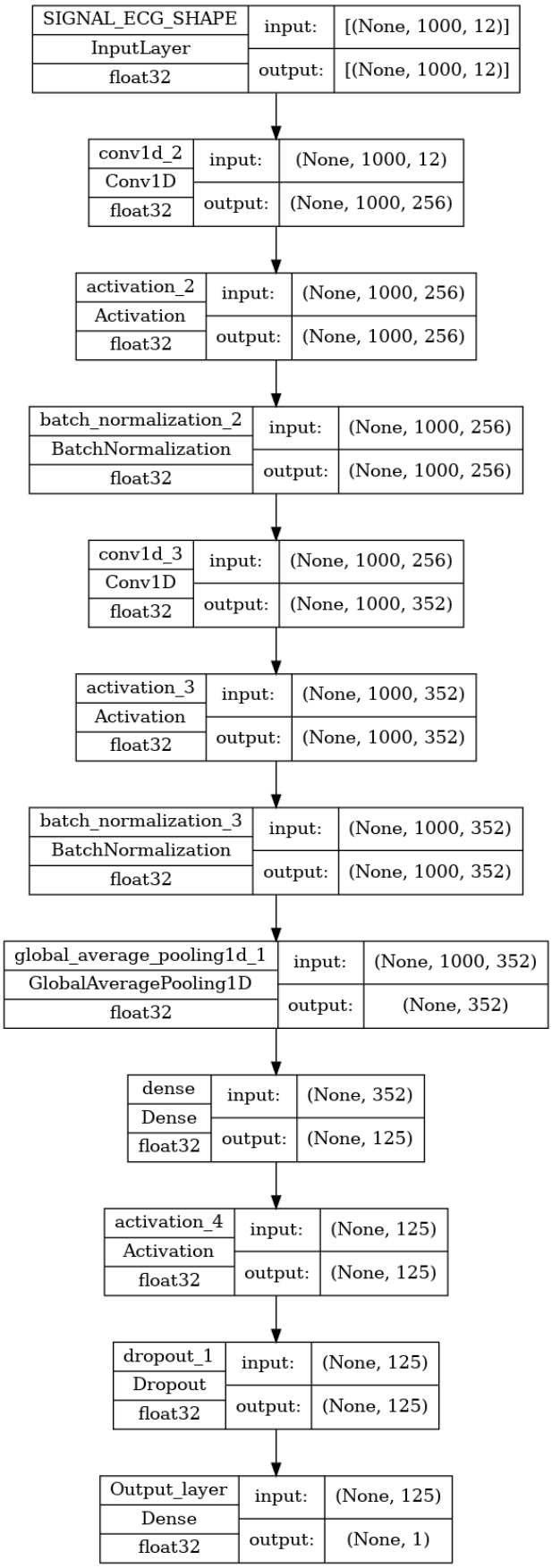



*Figure S4. Signal ECG model structure trained on arrangement A data from the speculative cohort.*

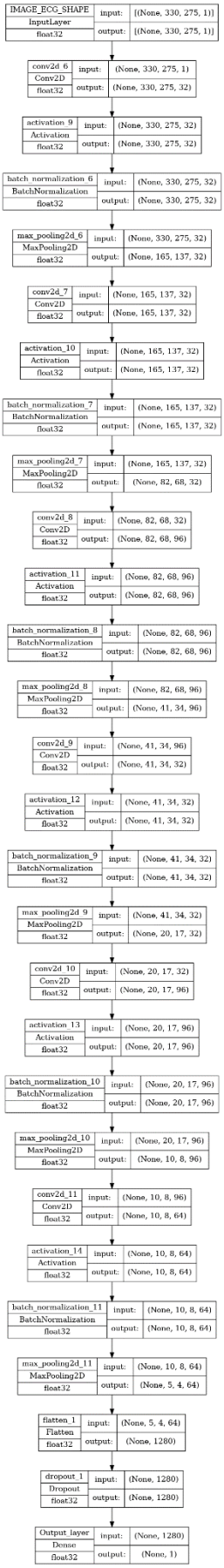



*Figure S5. Image ECG model structure trained on arrangement A data from the speculative cohort.*

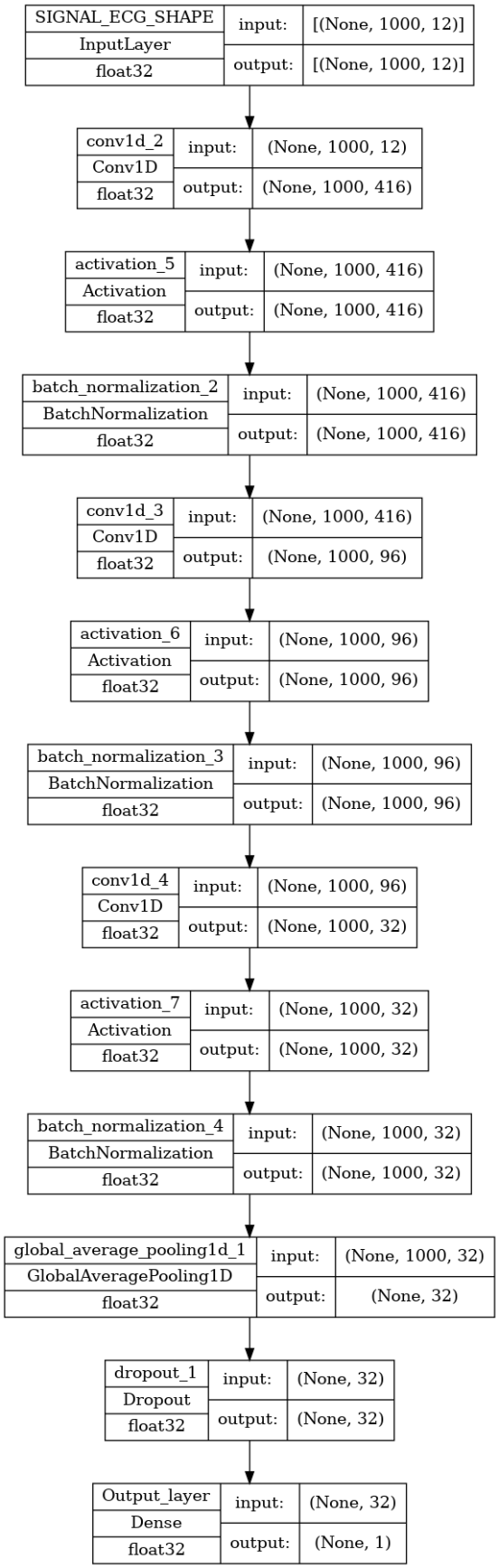



*Figure S6. Extracted Signal ECG model structure trained on arrangement A data from the speculative cohort.*

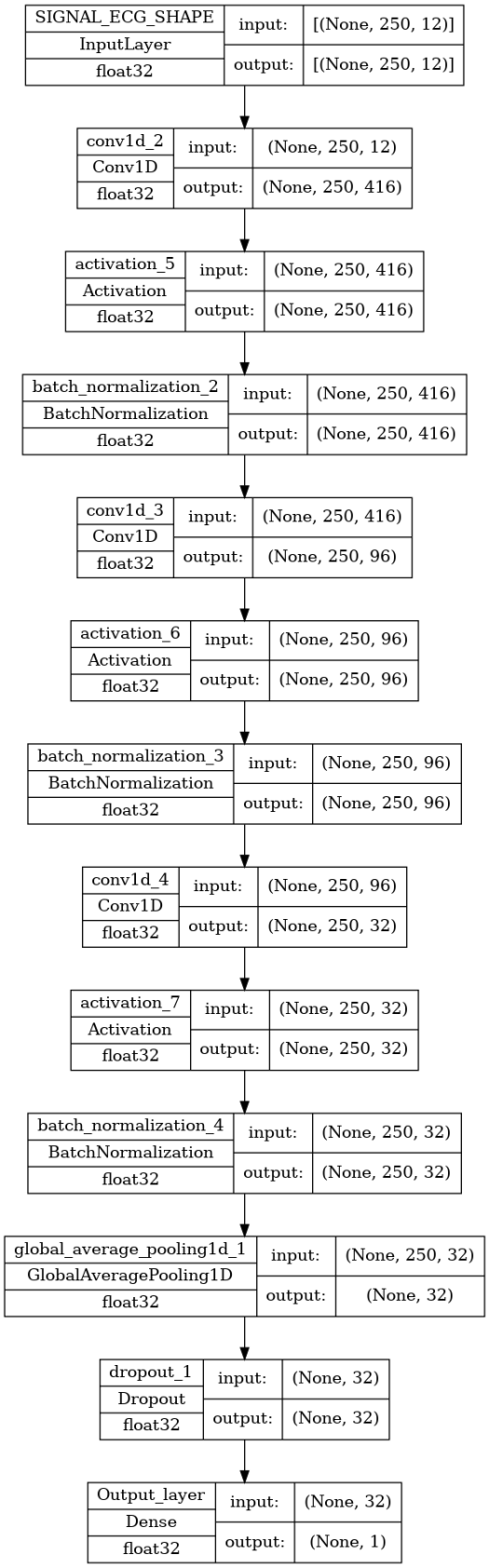



*Figure S7. Signal ECG model structure trained on arrangement B data from the conservative cohort.*

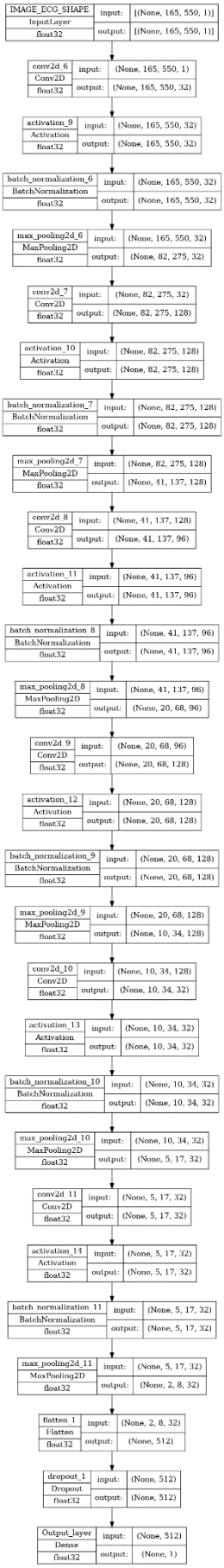



*Figure S8. Image ECG model structure trained on arrangement B data from the conservative cohort.*

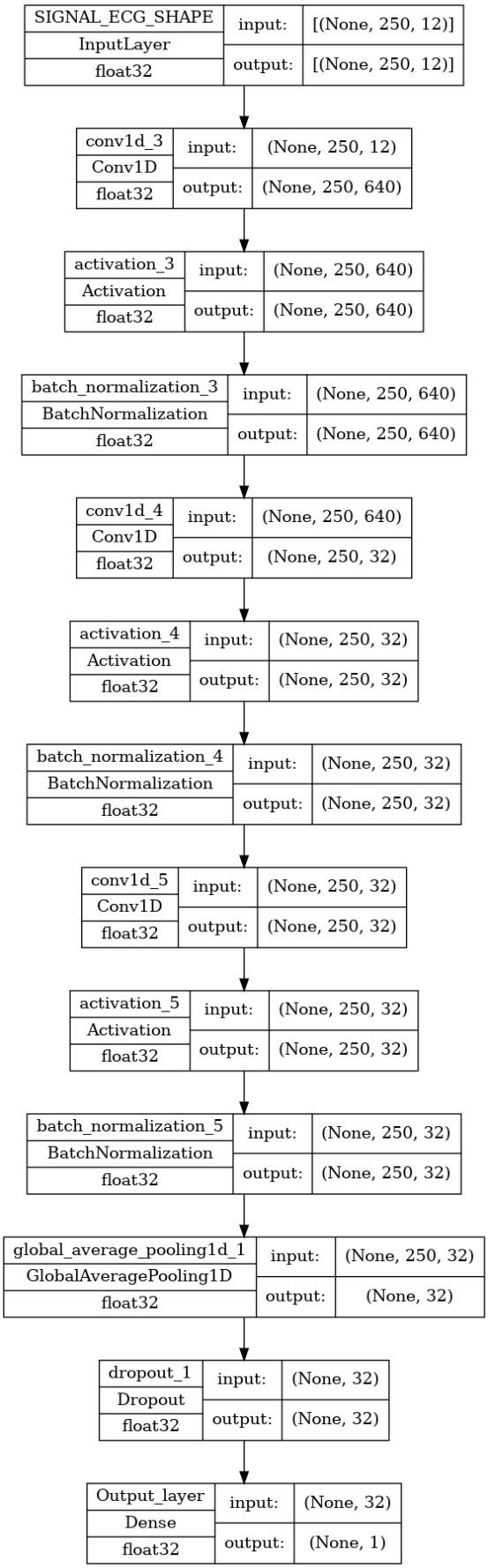



*Figure S9. Extracted Signal ECG model structure trained on arrangement B data from the conservative cohort.*

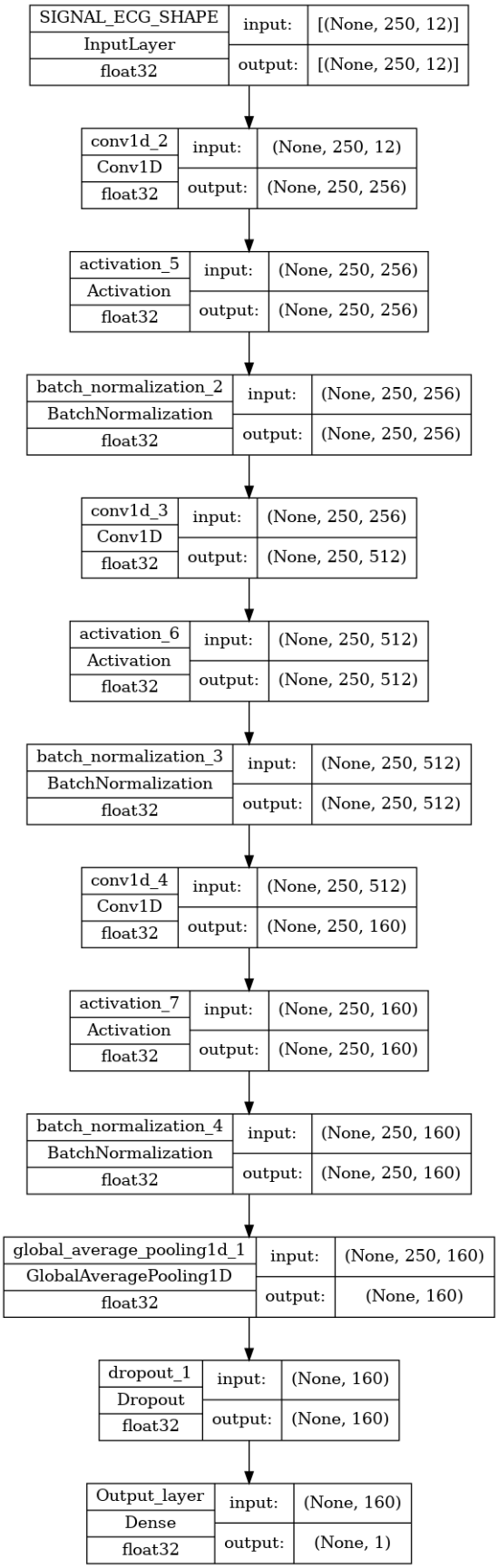



*Figure S10. Signal ECG model structure trained on arrangement B data from the speculative cohort.*

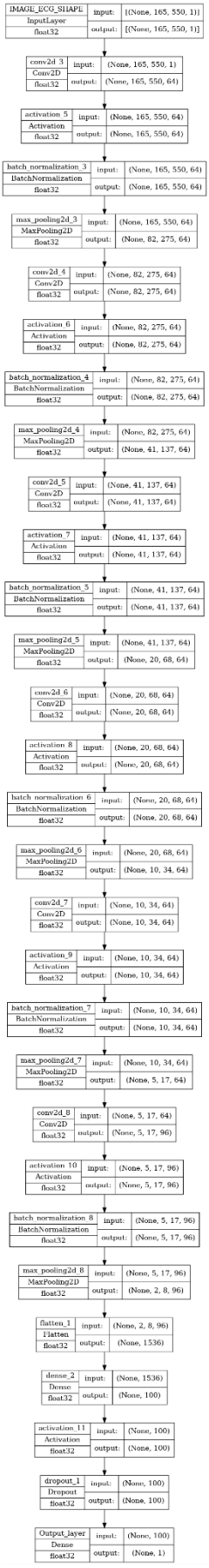



*Figure S11. Image ECG model structure trained on arrangement B data from the speculative cohort.*

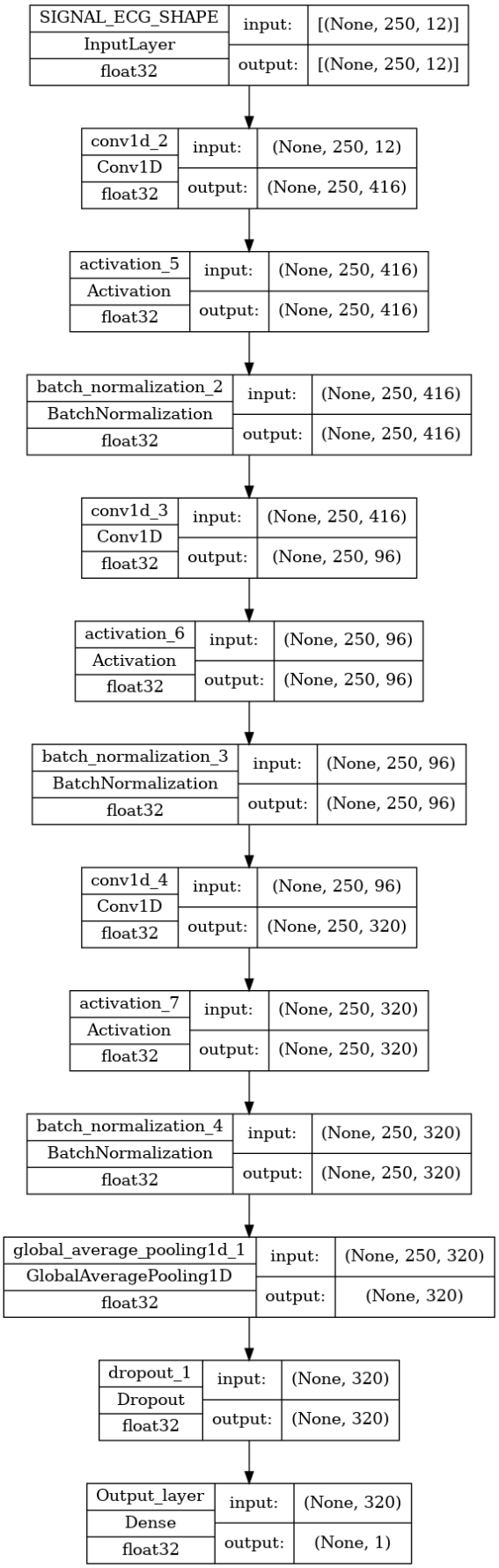



*Figure S12. Extracted Signal ECG model structure trained on arrangement B data from the speculative cohort.*

*Table S1. List of variables collected and used for analysis is compiled below. It includes the origin of the data, the V-Dem indicator code wherever applicable, the area of freedom (academic, digital or media) to which the variable was assigned during this study, and a variable reference used in this study, so they can be used further studied in the context of their reference maps*

| Name | Origin of data | V-Dem indicator code | Area of Freedom | Variable Reference |
|---|---|---|---|---|
| Abuse of defamation and copyright law by elites | V-Dem | v2smdefabu | Digital | DI1 |
| Abuse of defamation and copyright law by elites - Standard Deviation | V-Dem | v2smdefabu_sd | Digital | DI1SD |
| Academic Freedom Index | V-Dem | v2xca_academ | Academic | AC1 |
| Academic Freedom Index - Standard Deviation | V-Dem | v2xca_academ_sd | Academic | AC1SD |
| Academics as critics | V-Dem | v2cacritic | Academic | AC2 |
| Academics as critics - Standard Deviation | V-Dem | v2cacritic_sd | Academic | AC2SD |
| Access to Justice for Men | V-Dem | v2clacjstm | Media | ME1 |
| Access to Justice for Men - Standard Deviation | V-Dem | v2clacjstm_sd | Media | ME1SD |
| Access to Justice for Women | V-Dem | v2clacjstw | Media | ME2 |
| Access to Justice for Women - Standard Deviation | V-Dem | v2clacjstw_sd | Media | ME2SD |
| Accountability index | V-Dem | v2x_accountability | Media | ME3 |
| Alternative sources of information index | V-Dem | v2xme_altinf | Digital | DI2 |
| Alternative sources of information index - Standard Deviation | V-Dem | v2xme_altinf_sd | Digital | DI2SD |
| Arrests for political content | V-Dem | v2smarrest | Media | ME4 |
| Arrests for political content – Standard Deviation | V-Dem | v2smarrest_sd | Media | ME4SD |
| Average people’s use of social media to organize offline action | V-Dem | v2smorgavgact | Digital | DI3 |
| Average people’s use of social media to organize offline action - Standard Deviation | V-Dem | v2smorgavgact_sd | Digital | DI3SD |
| Campus integrity | V-Dem | v2casurv | Academic | AC3 |
| Campus integrity - Standard Deviation | V-Dem | v2casurv_sd | Academic | AC3SD |
| Civil liberties index | V-Dem | v2x_civlib | Media | ME5 |

| | | | | |
|---|---|---|---|---|
| Civil liberties index - Standard Deviation | V-Dem | v2x_civlib_sd | Media | ME5SD |
| Common Good | V-Dem | v2dlcommon | Media | ME6 |
| Common Good - Standard Deviation | V-Dem | v2dlcommon_sd | Media | ME6SD |
| Core civil society index | V-Dem | v2xcs_ccsi | Media | ME7 |
| Core civil society index - Standard Deviation | V-Dem | v2xcs_ccsi_sd | Media | ME7SD |
| Defamation protection | V-Dem | v2smlawpr | Media | ME8 |
| Defamation protection - Standard Deviation | V-Dem | v2smlawpr_sd | Media | ME8SD |
| Elites use of social media to organize offline action | V-Dem | v2smorgelitact | Digital | DI4 |
| Elites use of social media to organize offline action - Standard Deviation | V-Dem | v2smorgelitact_sd | Digital | DI4SD |
| Engaged Society | V-Dem | v2dlengage | Media | ME9 |
| Engaged Society - Standard Deviation | V-Dem | v2dlengage_sd | Media | ME9SD |
| Executive Bribery Corrupt Exchanges | V-Dem | v2exbribe | Media | ME10 |
| Executive Bribery Corrupt Exchanges - Standard Deviation | V-Dem | v2exbribe_sd | Media | ME10SD |
| Executive corruption index | V-Dem | v2x_execorr | Media | ME11 |
| Executive corruption index - Standard Deviation | V-Dem | v2x_execorr_sd | Media | ME11SD |
| Executive Respects Constitution | V-Dem | v2exrescon | Media | ME12 |
| Executive Respects Constitution - Standard Deviation | V-Dem | v2exrescon_sd | Media | ME12SD |
| Female journalists | V-Dem | v2mefemjrn | Media | ME13 |
| Female journalists - Standard Deviation | V-Dem | v2mefemjrn_sd | Media | ME13SD |
| Foreign governments ads | V-Dem | v2smforads | Media | ME14 |
| Foreign governments ads - Standard Deviation | V-Dem | v2smforads_sd | Media | ME14SD |
| Foreign government's dissemination of false information | V-Dem | v2smfordom | Media | ME15 |
| Foreign government's dissemination of false information - Standard Deviation | V-Dem | v2smfordom_sd | Media | ME15SD |
| Freedom of Academic and Cultural Expression | V-Dem | v2clacfree | Media | ME16 |

| | | | | |
|---|---|---|---|---|
| Freedom of Academic and Cultural Expression - Standard Deviation | V-Dem | v2clacfree_sd | Media | ME16SD |
| Freedom of academic exchange and dissemination | V-Dem | v2cafexch | Academic | AC4 |
| Freedom of academic exchange and dissemination - Standard Deviation | V-Dem | v2cafexch_sd | Academic | AC4SD |
| Freedom of discussion | V-Dem | v2xcl_disc | Academic | AC5 |
| Freedom of discussion - Standard Deviation | V-Dem | v2xcl_disc_sd | Academic | AC5SD |
| Freedom of Discussion for Men | V-Dem | v2cldiscm | Media | ME17 |
| Freedom of Discussion for Men - Standard Deviation | V-Dem | v2cldiscm_sd | Media | ME17SD |
| Freedom of Discussion for Women | V-Dem | v2cldiscw | Media | ME18 |
| Freedom of Discussion for Women - Standard Deviation | V-Dem | v2cldiscw_sd | Media | ME18SD |
| Freedom of domestic movement | V-Dem | v2xcl_dmove | Media | ME19 |
| Freedom of domestic movement - Standard Deviation | V-Dem | v2xcl_dmove_sd | Media | ME19SD |
| Freedom of Domestic Movement for Men | V-Dem | v2cldmovem | Media | ME20 |
| Freedom of Domestic Movement for Men - Standard Deviation | V-Dem | v2cldmovem_sd | Media | ME20SD |
| Freedom of Domestic Movement for Women | V-Dem | v2cldmovew | Media | ME21 |
| Freedom of Domestic Movement for Women - Standard Deviation | V-Dem | v2cldmovew_sd | Media | ME21SD |
| Freedom of Foreign Movement | V-Dem | v2clfmove | Media | ME22 |
| Freedom of Foreign Movement - Standard Deviation | V-Dem | v2clfmove_sd | Media | ME22SD |
| Freedom of peaceful assembly | V-Dem | v2caassemb | Media | ME23 |
| Freedom of peaceful assembly - Standard Deviation | V-Dem | v2caassemb_sd | Media | ME23SD |
| Freedom of Religion | V-Dem | v2clrelig | Media | ME24 |
| Freedom of Religion - Standard Deviation | V-Dem | v2clrelig_sd | Media | ME24SD |
| Freedom to research and teach | V-Dem | v2cafres | Academic | AC6 |
| Freedom to research and teach - Standard Deviation | V-Dem | v2cafres_sd | Academic | AC6SD |

| | | | | |
|---|---|---|---|---|
| Government capacity to regulate online content | V-Dem | v2smregcap | Digital | DI5 |
| Government capacity to regulate online content - Standard Deviation | V-Dem | v2smregcap_sd | Digital | DI5SD |
| Government censorship effort - Media | V-Dem | v2mecenefm | Media | ME25 |
| Government censorship effort - Media - Standard Deviation | V-Dem | v2mecenefm_sd | Media | ME25SD |
| Government cyber security capacity | V-Dem | v2smgovcapsec | Digital | DI6 |
| Government cyber security capacity - Standard Deviation | V-Dem | v2smgovcapsec_sd | Digital | DI6SD |
| Government dissemination of false information abroad | V-Dem | v2smgovab | Digital | DI7 |
| Government dissemination of false information abroad - Standard Deviation | V-Dem | v2smgovab_sd | Digital | DI7SD |
| Government dissemination of false information domestic | V-Dem | v2smgovdom | Digital | DI8 |
| Government dissemination of false information domestic - Standard Deviation | V-Dem | v2smgovdom_sd | Digital | DI8SD |
| Government Internet filtering capacity | V-Dem | v2smgovfilcap | Digital | DI9 |
| Government Internet filtering capacity - Standard Deviation | V-Dem | v2smgovfilcap_sd | Digital | DI9SD |
| Government Internet filtering in practice | V-Dem | v2smgovfilprc | Digital | DI10 |
| Government Internet filtering in practice - Standard Deviation | V-Dem | v2smgovfilprc_sd | Digital | DI10SD |
| Government Internet shut down capacity | V-Dem | v2smgovshutcap | Digital | DI11 |
| Government Internet shutdown capacity - Standard Deviation | V-Dem | v2smgovshutcap_sd | Digital | DI11SD |
| Government Internet shutdown in practice | V-Dem | v2smgovshut | Digital | DI12 |
| Government Internet shutdown in practice - Standard Deviation | V-Dem | v2smgovshut_sd | Digital | DI12SD |
| Government online content regulation approach | V-Dem | v2smregapp | Digital | DI13 |
| Government online content regulation approach - Standard Deviation | V-Dem | v2smregapp_sd | Digital | DI13SD |

| | | | | |
|---|---|---|---|---|
| Government social media alternatives | V-Dem | v2smgovsmalt | Digital | DI14 |
| Government social media alternatives - Standard Deviation | V-Dem | v2smgovsmalt_sd | Digital | DI14SD |
| Government social media censorship in practice | V-Dem | v2smgovsmcenprc | Digital | DI15 |
| Government social media censorship in practice - Standard Deviation | V-Dem | v2smgovsmcenprc_sd | Digital | DI15SD |
| Government social media monitoring | V-Dem | v2smgovsmmon | Digital | DI16 |
| Government social media monitoring - Standard Deviation | V-Dem | v2smgovsmmon_sd | Digital | DI16SD |
| Government social media shut down in practice | V-Dem | v2smgovsm | Digital | DI17 |
| Government social media shut down in practice - Standard Deviation | V-Dem | v2smgovsm_sd | Digital | DI17SD |
| Harassment of journalists | V-Dem | v2meharjrn | Media | ME26 |
| Harassment of journalists - Standard Deviation | V-Dem | v2meharjrn_sd | Media | ME26SD |
| Institutional autonomy | V-Dem | v2cainsaut | Academic | AC7 |
| Institutional autonomy - Standard Deviation | V-Dem | v2cainsaut_sd | Academic | AC7SD |
| Internet binary | V-Dem | v2mecenefibin | Digital | DI18 |
| Internet binary - Standard Deviation | V-Dem | v2mecenefibin_sd | Digital | DI18SD |
| Internet censorship effort | V-Dem | v2mecenefi | Digital | DI19 |
| Internet censorship effort - Standard Deviation | V-Dem | v2mecenefi_sd | Digital | DI19SD |
| Internet legal regulation content | V-Dem | v2smregcon | Digital | DI20 |
| Internet legal regulation content - Standard Deviation | V-Dem | v2smregcon_sd | Digital | DI20SD |
| Judicial corruption decision | V-Dem | v2jucorrdc | Media | ME27 |
| Judicial corruption decision - Standard Deviation | V-Dem | v2jucorrdc_sd | Media | ME27SD |
| Mass mobilization | V-Dem | v2cagenmob | Media | ME28 |
| Mass mobilization - Standard Deviation | V-Dem | v2cagenmob_sd | Media | ME28SD |
| Mass mobilization concentration | V-Dem | v2caconmob | Media | ME29 |

| | | | | |
|---|---|---|---|---|
| Mass mobilization concentration - Standard Deviation | V-Dem | v2caconmob_sd | Media | ME29SD |
| Media bias | V-Dem | v2mebias | Media | ME30 |
| Media bias - Standard Deviation | V-Dem | v2mebias_sd | Media | ME30SD |
| Media corrupt | V-Dem | v2mecorrpt | Media | ME31 |
| Media corrupt - Standard Deviation | V-Dem | v2mecorrpt_sd | Media | ME31SD |
| Media self-censorship | V-Dem | v2meslfcen | Media | ME32 |
| Media self-censorship - Standard Deviation | V-Dem | v2meslfcen_sd | Media | ME32SD |
| Mobilization for autocracy | V-Dem | v2caautmob | Media | ME33 |
| Mobilization for autocracy - Standard Deviation | V-Dem | v2caautmob_sd | Media | ME33SD |
| Mobilization for democracy | V-Dem | v2cademmob | Media | ME34 |
| Mobilization for democracy - Standard Deviation | V-Dem | v2cademmob_sd | Media | ME34SD |
| Online media existence | V-Dem | v2smonex | Digital | DI21 |
| Online media existence - Standard Deviation | V-Dem | v2smonex_sd | Digital | DI21SD |
| Online media fractionalization | V-Dem | v2smmefra | Digital | DI22 |
| Online media fractionalization – Standard Deviation | V-Dem | v2smmefra_sd | Digital | DI22SD |
| Online media perspectives | V-Dem | v2smonper | Media | ME35 |
| Online media perspectives - Standard Deviation | V-Dem | v2smonper_sd | Media | ME35SD |
| Party dissemination of false information abroad | V-Dem | v2smparab | Media | ME36 |
| Party dissemination of false information abroad - Standard Deviation | V-Dem | v2smparab_sd | Media | ME36SD |
| Party dissemination of false information domestic | V-Dem | v2smpardom | Media | ME37 |
| Party dissemination of false information domestic - Standard Deviation | V-Dem | v2smpardom_sd | Media | ME37SD |
| Party-candidate use of social media in campaigns | V-Dem | v2smcamp | Media | ME38 |
| Party-candidate use of social media in campaigns - Standard Deviation | V-Dem | v2smcamp_sd | Media | ME38SD |

| | | | | |
|---|---|---|---|---|
| Polarization of society | V-Dem | v2smpolsoc | Media | ME39 |
| Polarization of society - Standard Deviation | V-Dem | v2smpolsoc_sd | Media | ME39SD |
| Political civil liberties index | V-Dem | v2x_clpol | Media | ME40 |
| Political civil liberties index - Standard Deviation | V-Dem | v2x_clpol_sd | Media | ME40SD |
| Political corruption index | V-Dem | v2x_corr | Media | ME41 |
| Political corruption index - Standard Deviation | V-Dem | v2x_corr_sd | Media | ME41SD |
| Political parties' cyber security capacity | V-Dem | v2smpolcap | Digital | DI23 |
| Political parties’ cyber security capacity - Standard Deviation | V-Dem | v2smpolcap_sd | Digital | DI23SD |
| Political parties hate speech | V-Dem | v2smpolhate | Media | ME42 |
| Political parties hate speech - Standard Deviation | V-Dem | v2smpolhate_sd | Media | ME42SD |
| Political polarization | V-Dem | v2cacamps | Media | ME43 |
| Political polarization - Standard Deviation | V-Dem | v2cacamps_sd | Media | ME43SD |
| Political violence | V-Dem | v2caviol | Media | ME44 |
| Political violence - Standard Deviation | V-Dem | v2caviol_sd | Media | ME44SD |
| Print/broadcast media critical | V-Dem | v2mecrit | Media | ME45 |
| Print/broadcast media critical - Standard Deviation | V-Dem | v2mecrit_sd | Media | ME45SD |
| Print/broadcast media perspectives | V-Dem | v2merange | Media | ME46 |
| Print/broadcast media perspectives - Standard Deviation | V-Dem | v2merange_sd | Media | ME46SD |
| Privacy protection by law exists | V-Dem | v2smprivex | Media | ME47 |
| Privacy protection by law exists - Standard Deviation | V-Dem | v2smprivex_sd | Media | ME47SD |
| Private civil liberties index | V-Dem | v2x_clpriv | Media | ME48 |
| Private civil liberties index - Standard Deviation | V-Dem | v2x_clpriv_sd | Media | ME48SD |
| Public Sector Corrupt Exchanges | V-Dem | v2excrptps | Media | ME49 |
| Public Sector Corrupt Exchanges - Standard Deviation | V-Dem | v2excrptps_sd | Media | ME49SD |

| | | | | |
|---|---|---|---|---|
| Public sector corruption index | V-Dem | v2x_pubcorr | Media | ME50 |
| Public sector corruption index - Standard Deviation | V-Dem | v2x_pubcorr_sd | Media | ME50SD |
| Range of Consultation | V-Dem | v2dlconslt | Media | ME51 |
| Range of Consultation - Standard Deviation | V-Dem | v2dlconslt_sd | Media | ME51SD |
| Regime corruption | V-Dem | v2xnp_regcorr | Media | ME52 |
| Regime corruption - Standard Deviation | V-Dem | v2xnp_regcorr_sd | Media | ME52SD |
| Respect Counter Arguments | V-Dem | v2dlcountr | Media | ME53 |
| Respect Counter Arguments - Standard Deviation | V-Dem | v2dlcountr_sd | Media | ME53SD |
| Rule of law index | V-Dem | v2x_rule | Media | ME54 |
| Rule of law index - Standard Deviation | V-Dem | v2x_rule_sd | Media | ME54SD |
| Social Group Equality in Respect for Civil Liberties | V-Dem | v2clsocgrp | Media | ME55 |
| Social Group Equality in Respect for Civil Liberties - Standard Deviation | V-Dem | v2clsocgrp_sd | Media | ME55SD |
| Use of social media to organize offline violence | V-Dem | v2smorgviol | Digital | DI24 |
| Use of social media to organize offline violence - Standard Deviation | V-Dem | v2smorgviol_sd | Digital | DI24SD |
| Global Cybersecurity Index 2020 - Score | Global Cybersecurity Index | - | Digital | DI25 |
| Missing journalists | CPJ | - | Media | ME56 |
| Number of journalists and media workers killed - up to 2022 | CPJ | - | Media | ME57 |
| Number of journalists imprisoned as of 1 December 2021 | CPJ | - | Media | ME58 |
| RSF Global ranking - Score | World Press Freedom Index | - | Media | ME59 |
| UNESCO - observatory of killed journalists | UNESCO | - | Media | ME60 |
| Cost of Total Shutdown Per Hour (USD) | NetBlocks | - | Digital | DI26 |

*Table S2. Results of hyperparameter tuning carried out to select the optimal regularisation term*

| REGULARISATION TERM | ERROR |
|---|---|
| **100** | 83.5 |
| **10** | 64.82 |
| **1** | 59.89 |
| **0.1** | 59.14 |
| **0.01** | 59.09 |
| **0.001** | 58.88 |
| **0.0001** | 59.08 |

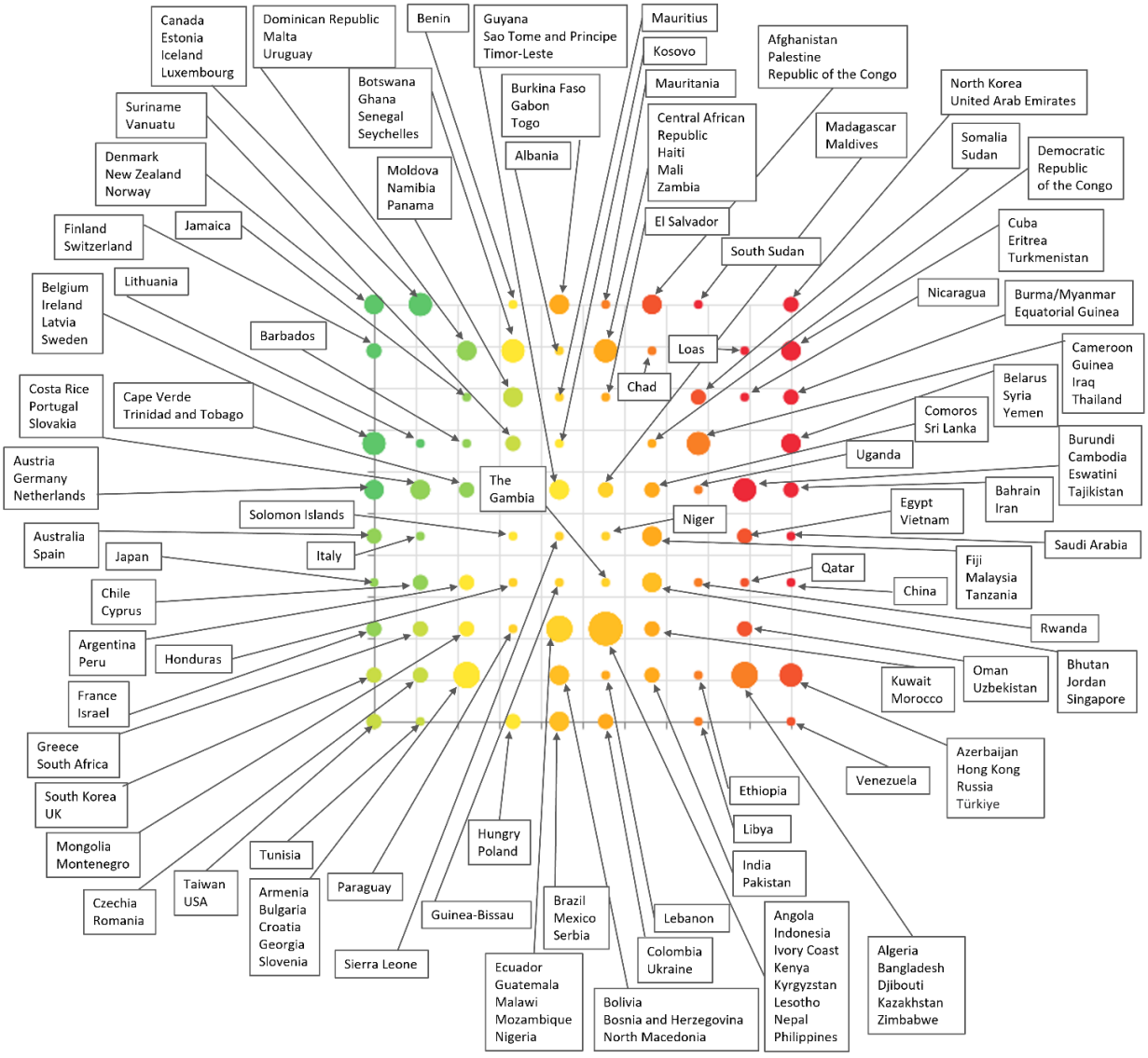



*Figure S13. Detailed country clustering visualisation (cluster membership map) colour-coded by the cluster ranking. The countries allocated to a selection of clusters are displayed. Cluster separation indicates similarity (i.e. closer clusters are more similar than further clusters).*

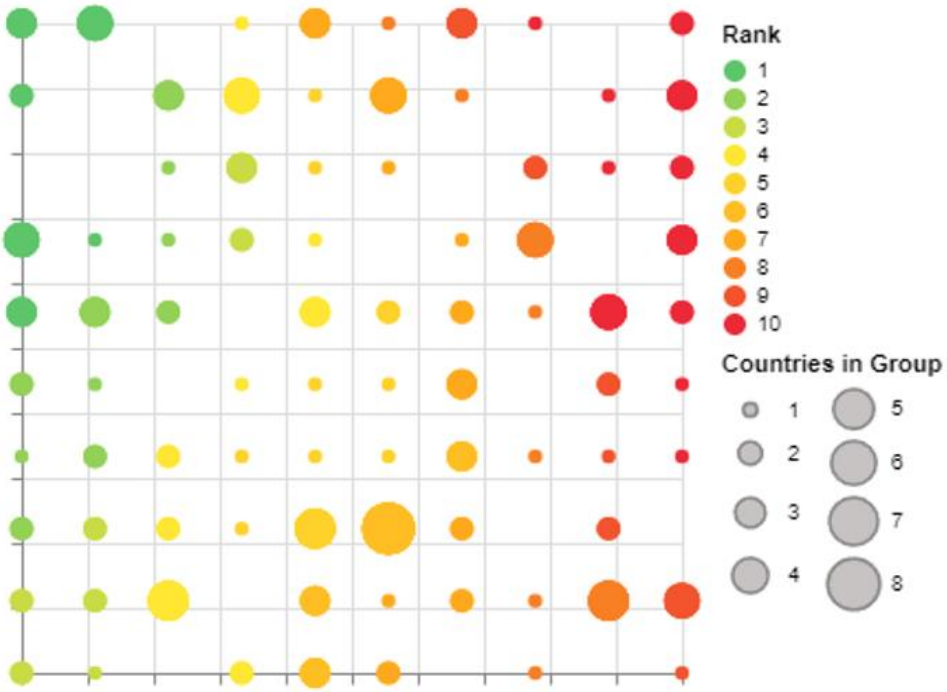
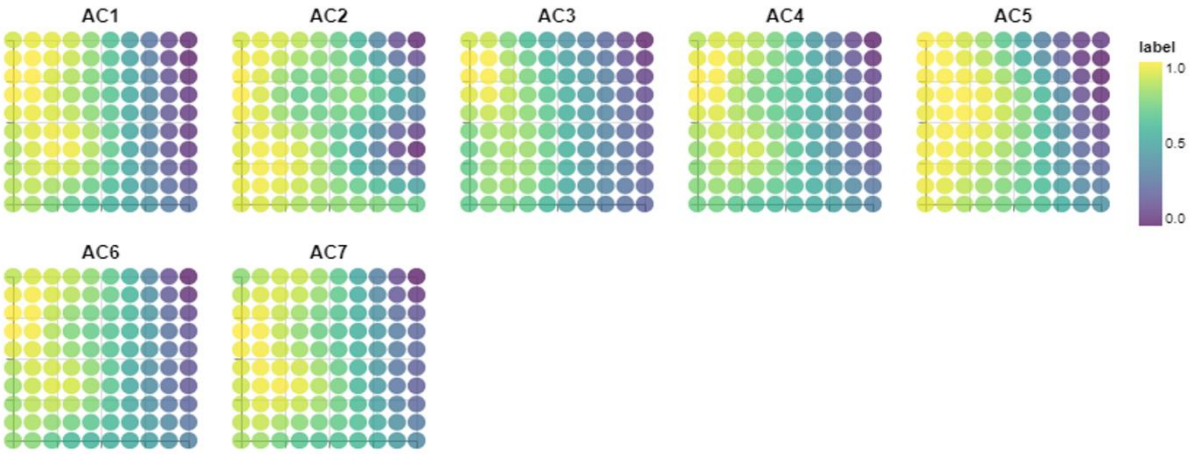
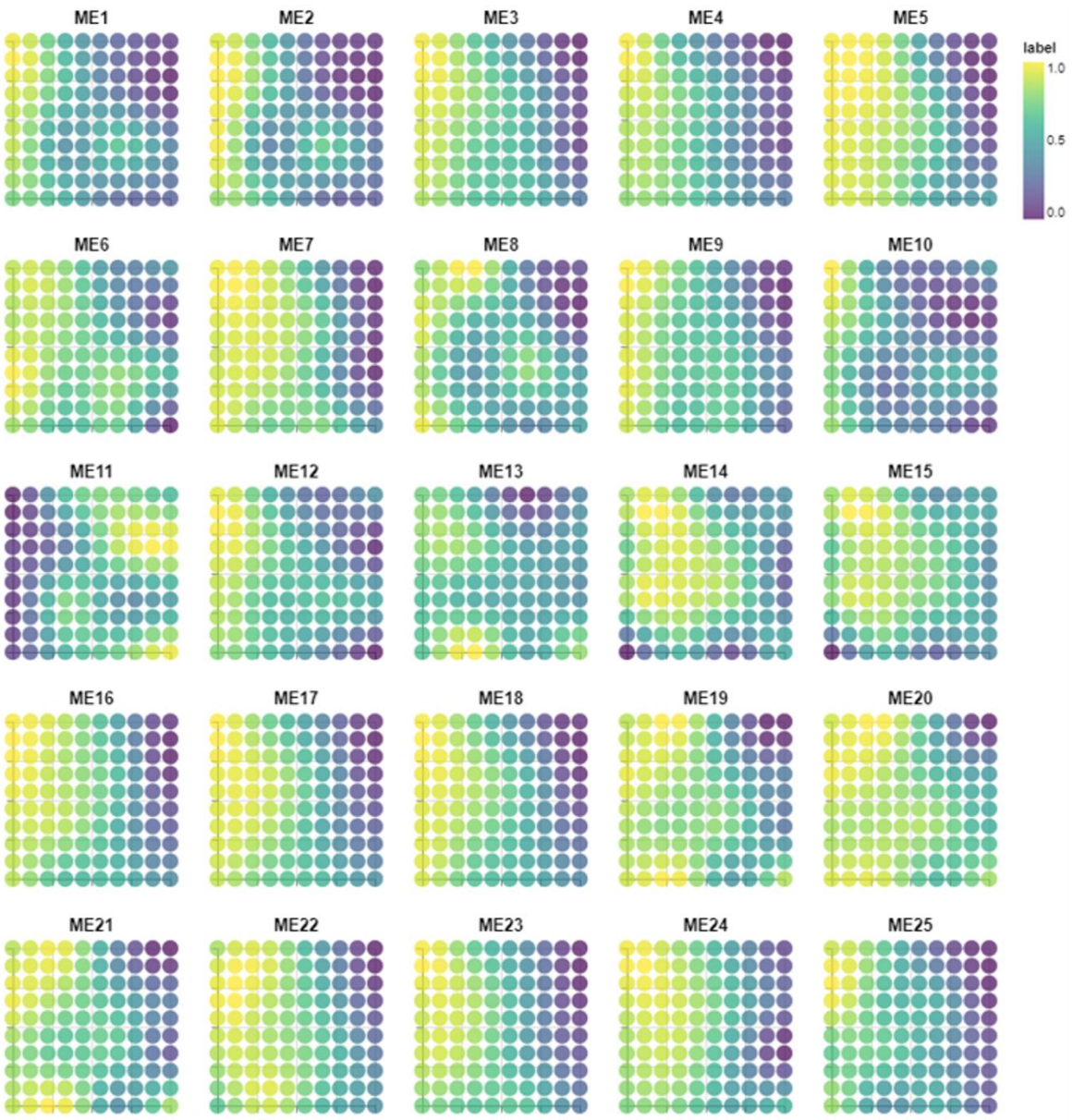



*Figure S14. This figure mirrors Figure S7 but omits the list of countries allocated to the clusters to offer a clearer view of the size of the clusters and the assigned ranking.*



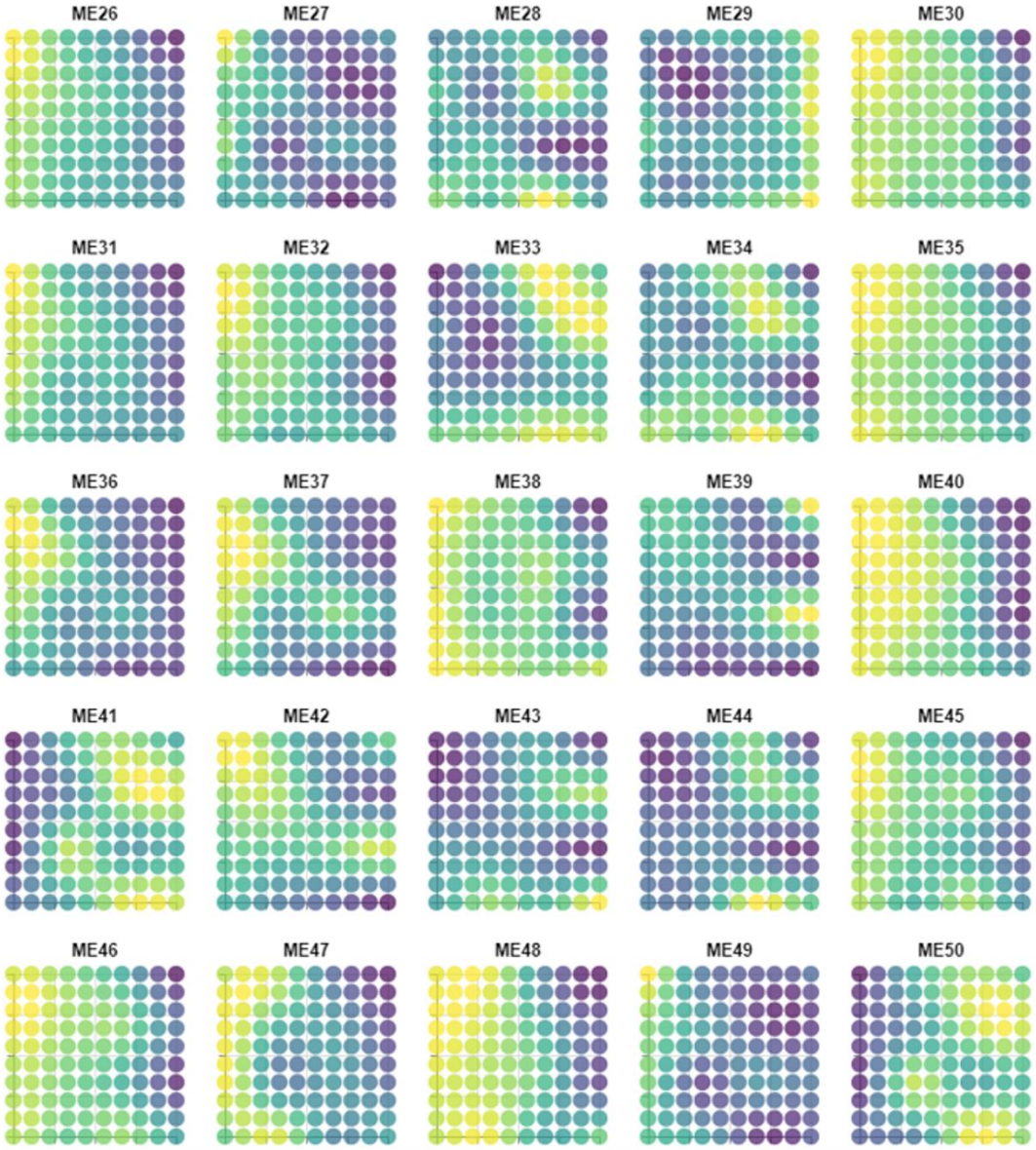

AC1: Academic Freedom Index
AC2: Academics as critics
AC3: Campus integrity
AC4: Freedom of academic exchange and dissemination
AC5: Freedom of discussion
AC6: Freedom to research and teach
AC7: Institutional autonomy

*Figure S15. Reference maps for the academic freedom-related variables used to produce the GTM model.*



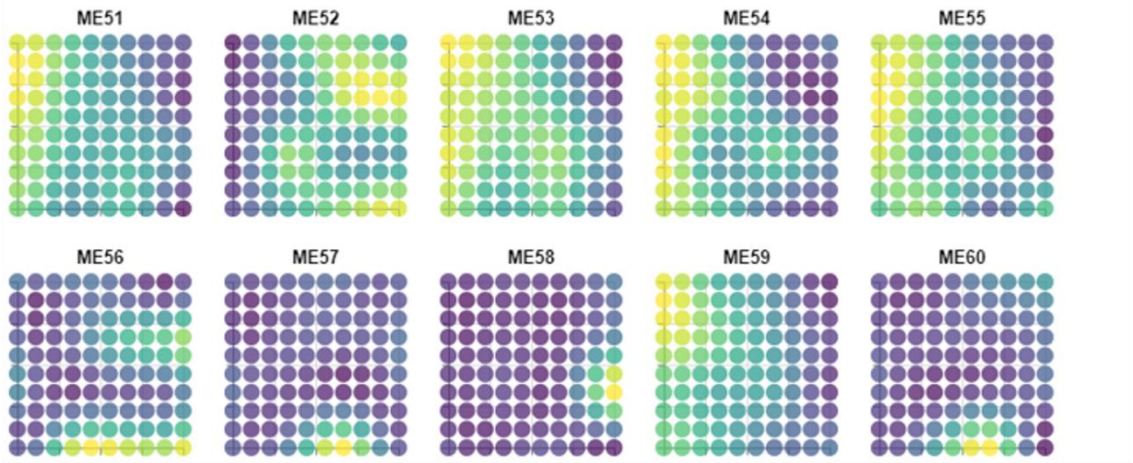

ME1: Access to Justice for Men
ME2: Access to Justice for Women
ME3: Accountability index
ME4: Arrests for political content
ME5: Civil liberties index
ME6: Common Good
ME7: Core civil society index
ME8: Defamation protection
ME9: Engaged Society
ME10: Executive Bribery Corrupt Exchanges
ME11: Executive corruption index
ME12: Executive Respects Constitution
ME13: Female journalists
ME14: Foreign governments ads
ME15: Foreign government's dissemination of false information
ME16: Freedom of Academic and Cultural Expression
ME17: Freedom of Discussion for Men
ME18: Freedom of Discussion for Women
ME19: Freedom of domestic movement
ME20: Freedom of Domestic Movement for Men
ME21: Freedom of Domestic Movement for Women
ME22: Freedom of Foreign Movement
ME23: Freedom of peaceful assembly
ME24: Freedom of Religion
ME25: Government censorship effort - Media

*Figure S16. Reference maps for 25 of the media freedom-related variables (ME1-ME25) used to produce the GTM model.*



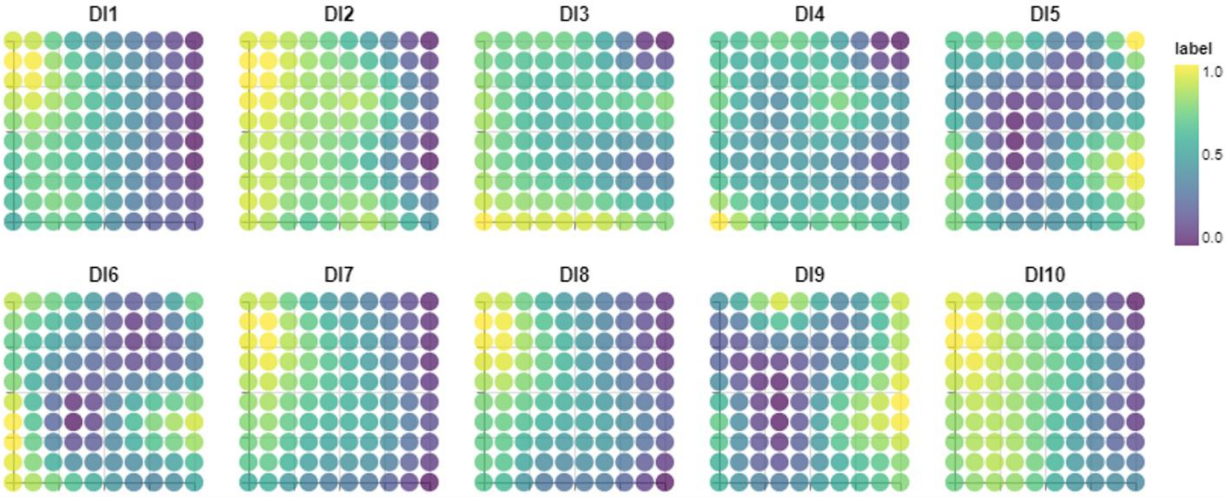

ME26: Harassment of journalists
ME27: Judicial Corruption Decision
ME28: Mass mobilisation
ME29: Mass mobilisation concentration
ME30: Media bias
ME31: Media corrupt
ME32: Media self-censorship
ME33: Mobilisation for autocracy
ME34: Mobilisation for democracy
ME35: Online media perspectives
ME36: Party dissemination of false information abroad
ME37: Party dissemination of false information domestic
ME38: Party-candidate use of social media in campaigns
ME39: Polarisation of society
ME40: Political civil liberties index
ME41: Political corruption index
ME42: Political parties hate speech
ME43: Political polarisation
ME44: Political violence
ME45: Print/broadcast media critical
ME46: Print/broadcast media perspectives
ME47: Privacy protection by law exists
ME48: Private civil liberties index
ME49: Public Sector Corrupt Exchanges
ME50: Public sector corruption index

*Figure S17. Continuation of Figure S16, including the reference maps for 25 of the media freedom-related variables (ME26-ME50) used to produce the GTM model.*



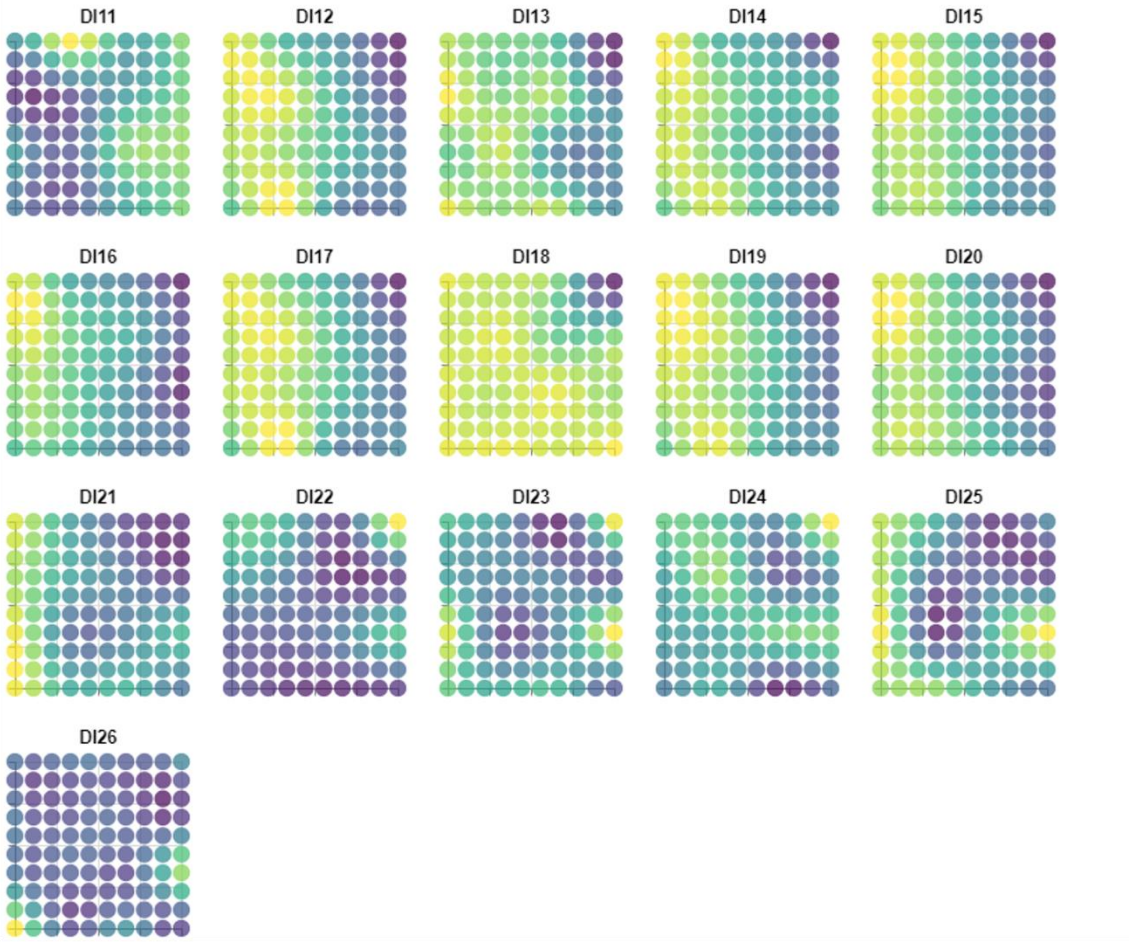

ME51: Range of Consultation
ME52: Regime corruption
ME53: Respect Counter Arguments
ME54: Rule of law index
ME55: Social Group Equality in Respect for Civil Liberties
ME56: Missing journalists
ME57: Number of journalists and media workers killed - up to 2022
ME58: Number of journalists imprisoned as of 1 December 2021
ME59: RSF Global ranking
ME60: UNESCO - observatory of killed journalists

*Figure S18. Continuation of Figure S17, including the reference maps for 10 of the media freedom-related variables (ME51-ME60) used to produce the GTM model.*




DI1: Abuse of defamation and copyright law by elites
DI2: Alternative sources of information index
DI3: Average people's use of social media to organize offline action
DI4: Elite's use of social media to organize offline action
DI5: Government capacity to regulate online content
DI6: Government cyber security capacity
DI7: Government dissemination of false information abroad
DI8: Government dissemination of false information domestic
DI9: Government Internet filtering capacity
DI10: Government Internet filtering in practice

*Figure S19. Reference maps for 10 of the digital freedom-related variables (DI1-DI10) used to produce the GTM model.*




DI11: Government Internet shut down capacity
DI12: Government Internet shut down in practice
DI13: Government online content regulation approach
DI14: Government social media alternatives
DI15: Government social media censorship in practice
DI16: Government social media monitoring
DI17: Government social media shut down in practice
DI18: Internet binary
DI19: Internet censorship effort
DI20: Internet legal regulation content
DI21: Online media existence
DI22: Online media fractionalization
DI23: Political party's cyber security capacity
DI24: Use of social media to organize offline violence
DI25: Global Cybersecurity Index 2020
DI26: Cost of Total Shutdown Per Hour – Dollar

*Figure S20. Continuation of Figure S19, including the reference maps for 16 of the digital freedom-related variables (DI11-DI26) used to produce the GTM model.*

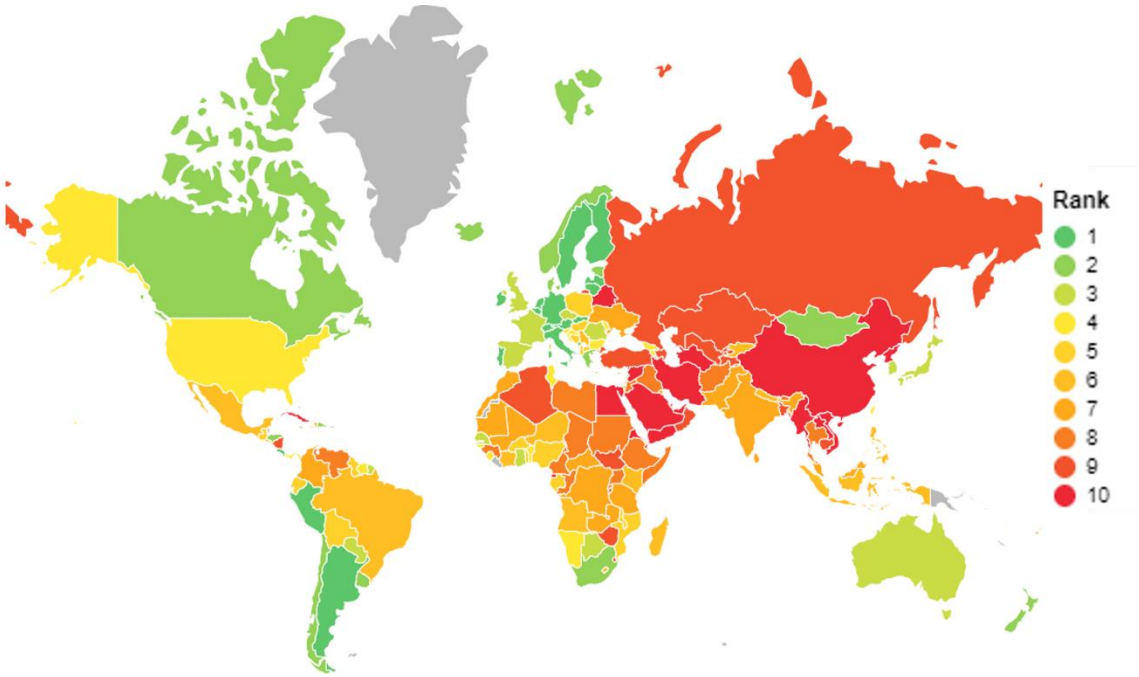



*Figure S21. World map showing the academic free expression ranking.*

*Table S3. Global ranking of academic free expression by deciles. Lower ranks represent higher levels of free expression while higher ranks represent lower levels of freedom.*

| COUNTRIES AND NATIONS | ACADEMIC FREEDOM RANKING |
|---|---|
| Argentina, Austria, Belgium, Costa Rica, Finland, Germany, Ireland, Italy, Latvia, Lithuania, Netherlands, Peru, Portugal, Slovakia, Sweden, Switzerland | **1** |
| Canada, Chile, Cyprus, Denmark, Dominican Republic, Estonia, Greece, Honduras, Iceland, Jamaica, Luxembourg, Malta, Mongolia, Montenegro, New Zealand, Norway, South Africa, Uruguay | **2** |
| Australia, Barbados, Botswana, Cape Verde, Czechia, France, Ghana, Israel, Japan, Paraguay, Romania, Senegal, Seychelles, Solomon Islands, South Korea, Spain, Trinidad and Tobago, United Kingdom | **3** |
| Armenia, Benin, Bulgaria, Croatia, Georgia, Guinea-Bissau, Moldova, Namibia, Panama, Sierra Leone, Slovenia, Suriname, Taiwan, Tunisia, United States of America, Vanuatu | **4** |

| | |
|---|---|
| Albania, Bolivia, Bosnia and Herzegovina, Burkina Faso, Ecuador, Gabon, Guatemala, Guyana, Hungary, Kosovo, Malawi, Mauritius, Mozambique, Nigeria, North Macedonia, Poland, Sao Tome and Principe, Timor-Leste, Togo | **5** |
| Angola, Brazil, El Salvador, Indonesia, Ivory Coast, Kenya, Kyrgyzstan, Lebanon, Lesotho, Madagascar, Maldives, Mexico, Nepal, Niger, Philippines, Serbia, The Gambia | **6** |
| Central African Republic, Colombia, Comoros, Democratic Republic of the Congo, Fiji, Haiti, India, Kuwait, Malaysia, Mali, Mauritania, Morocco, Pakistan, Sri Lanka, Tanzania, Ukraine, Zambia | **7** |
| Afghanistan, Bhutan, Cameroon, Chad, Ethiopia, Guinea, Iraq, Jordan, Libya, Palestine, Republic of the Congo, Rwanda, Singapore, Somalia, Sudan, Thailand, Uganda, Venezuela | **8** |
| Algeria, Azerbaijan, Bangladesh, Burundi, Cambodia, Djibouti, Eswatini, Hong Kong, Kazakhstan, Nicaragua, Oman, Russia, South Sudan, Tajikistan, Türkiye, Uzbekistan, Zimbabwe | **9** |
| Bahrain, Belarus, Burma/Myanmar, China, Cuba, Egypt, Equatorial Guinea, Eritrea, Iran, Laos, North Korea, Qatar, Saudi Arabia, Syria, Turkmenistan, United Arab Emirates, Vietnam, Yemen | **10** |

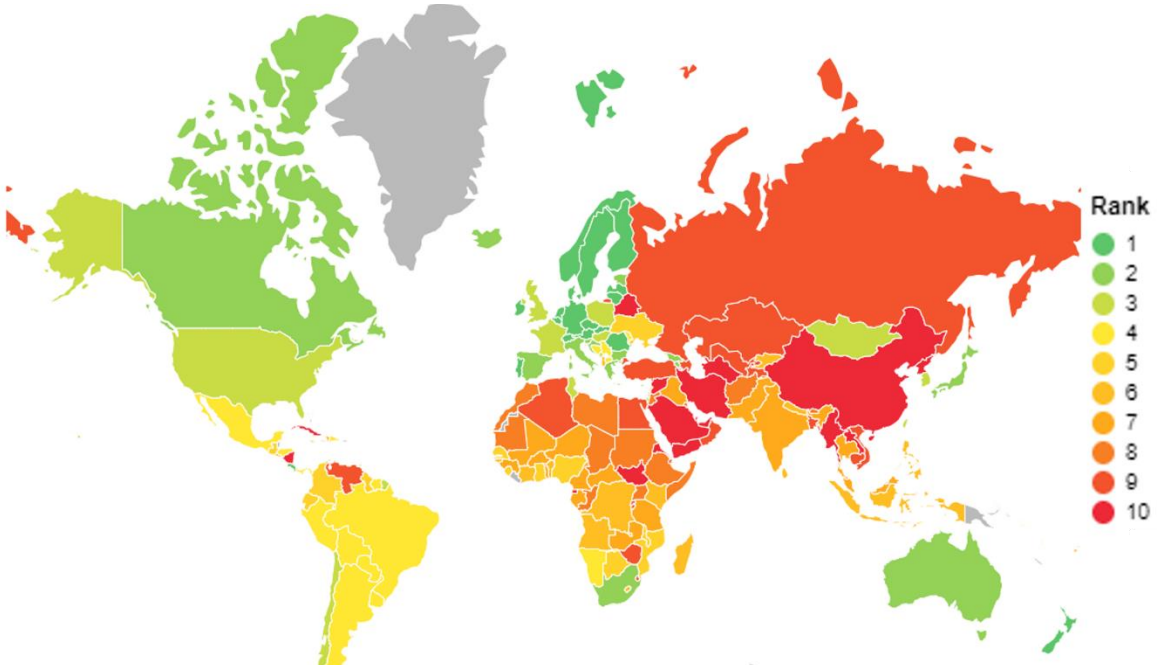



*Figure S22. World map showing the digital free expression ranking.*

*Table S4. Global ranking of digital free expression by deciles. Lower ranks represent higher levels of free expression while higher ranks represent lower levels of freedom.*

| COUNTRIES AND NATIONS | DIGITAL FREEDOM RANKING |
|---|---|
| Austria, Belgium, Costa Rica, Czechia, Denmark, Finland, Germany, Ireland, Latvia, Lithuania, Netherlands, New Zealand, Norway, Portugal, Romania, Slovakia, Sweden, Switzerland | **1** |
| Armenia, Australia, Barbados, Bulgaria, Canada, Croatia, Estonia, Georgia, Greece, Iceland, Italy, Japan, Luxembourg, Slovenia, South Africa, Spain | **2** |
| Cape Verde, Chile, Cyprus, France, Hungary, Israel, Jamaica, Mongolia, Montenegro, Poland, South Korea, Taiwan, Trinidad and Tobago, Tunisia, United Kingdom, United States of America | **3** |
| Argentina, Bolivia, Bosnia and Herzegovina, Brazil, Dominican Republic, Honduras, Kosovo, Malta, Mexico, Moldova, Namibia, North Macedonia, Panama, Paraguay, Peru, Serbia, Solomon Islands, Suriname, Uruguay, Vanuatu | **4** |
| Botswana, Colombia, Ecuador, Ghana, Guatemala, Guyana, Lebanon, Malawi, Mauritius, Mozambique, Nigeria, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Timor-Leste, Ukraine | **5** |
| Albania, Angola, Benin, Comoros, Democratic Republic of the Congo, El Salvador, Guinea-Bissau, Indonesia, Ivory Coast, Kenya, Kyrgyzstan, Lesotho, Madagascar, Maldives, Nepal, Philippines, Sri Lanka | **6** |
| Burkina Faso, Cameroon, Central African Republic, Fiji, Gabon, Guinea, Haiti, India, Iraq, Malaysia, Mali, Niger, Pakistan, Tanzania, Thailand, The Gambia, Togo, Zambia | **7** |
| Afghanistan, Bhutan, Chad, Ethiopia, Jordan, Kuwait, Libya, Mauritania, Morocco, Palestine, Republic of the Congo, Singapore, Somalia, Sudan, Uganda | **8** |
| Algeria, Azerbaijan, Bangladesh, Burundi, Cambodia, Djibouti, Egypt, Eswatini, Hong Kong, Kazakhstan, Oman, Russia, Rwanda, Tajikistan, Türkiye, Uzbekistan, Venezuela, Vietnam, Zimbabwe | **9** |

| | |
|---|---|
| Bahrain, Belarus, Burma/Myanmar, China, Cuba, Equatorial Guinea, Eritrea, Iran, Laos, Nicaragua, North Korea, Qatar, Saudi Arabia, South Sudan, Syria, Turkmenistan, United Arab Emirates, Yemen | **10** |

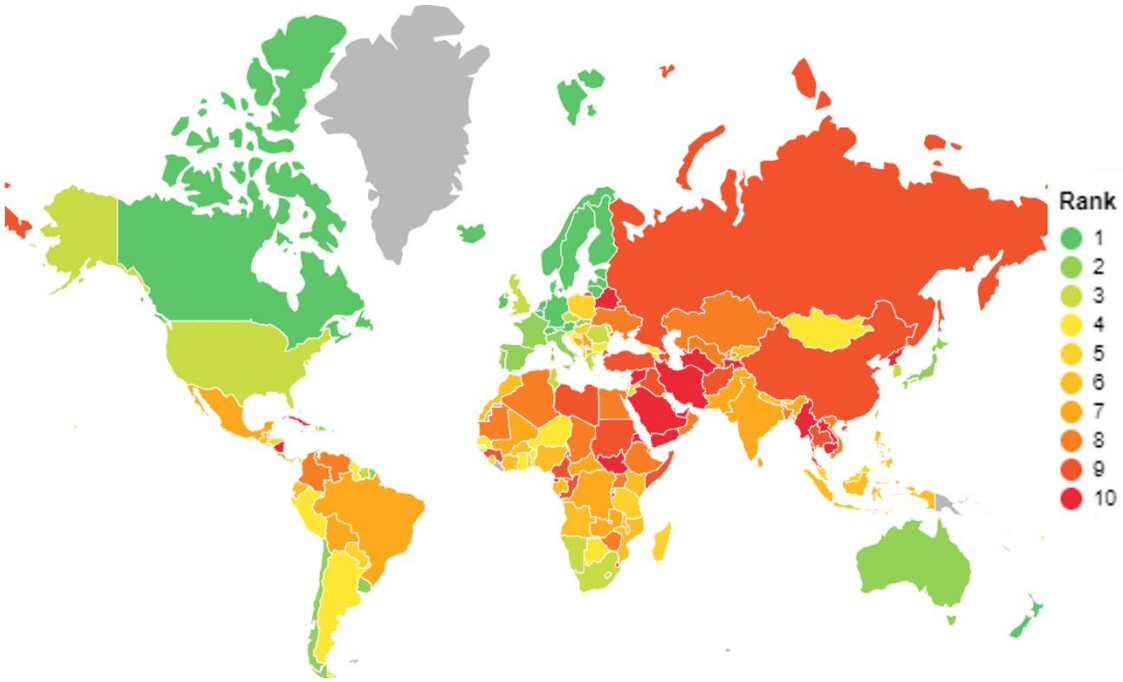



*Figure S23. World map showing the media free expression ranking.*

*Table S5. Global ranking of media free expression by deciles. Lower ranks represent higher levels of free expression while higher ranks represent lower levels of freedom.*

| **COUNTRIES AND NATIONS** | **MEDIA FREEDOM RANKING** |
|---|---|
| Austria, Belgium, Canada, Denmark, Estonia, Finland, Germany, Iceland, Ireland, Latvia, Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Sweden, Switzerland | **1** |
| Australia, Barbados, Cape Verde, Chile, Costa Rica, Cyprus, Dominican Republic, France, Israel, Italy, Jamaica, Japan, Malta, Portugal, Slovakia, Spain, Trinidad and Tobago, Uruguay | **2** |
| Czechia, Greece, Moldova, Namibia, Panama, Romania, South Africa, South Korea, Suriname, Taiwan, Tunisia, United Kingdom, United States of America, Vanuatu | **3** |
| Argentina, Armenia, Benin, Botswana, Bulgaria, Croatia, Georgia, Ghana, Guyana, Kosovo, Mauritius, Mongolia, Montenegro, Niger, | **4** |

| | |
|---|---|
| Peru, Sao Tome and Principe, Senegal, Seychelles, Slovenia, Solomon Islands, Timor-Leste | |
| Albania, Bhutan, Fiji, Guinea-Bissau, Honduras, Hungary, Jordan, Madagascar, Malaysia, Maldives, Paraguay, Poland, Sierra Leone, Singapore, Tanzania, The Gambia | **5** |
| Angola, Burkina Faso, Ecuador, Gabon, Guatemala, Indonesia, Ivory Coast, Kenya, Kuwait, Kyrgyzstan, Lesotho, Malawi, Morocco, Mozambique, Nepal, Nigeria, Philippines, Togo | **6** |
| Bolivia, Bosnia and Herzegovina, Brazil, Central African Republic, Comoros, Democratic Republic of the Congo, El Salvador, Haiti, India, Lebanon, Mali, Mexico, North Macedonia, Pakistan, Rwanda, Serbia, Sri Lanka, Zambia | **7** |
| Algeria, Bangladesh, Chad, Colombia, Djibouti, Egypt, Ethiopia, Kazakhstan, Mauritania, Oman, Qatar, Uganda, Ukraine, Uzbekistan, Venezuela, Vietnam, Zimbabwe | **8** |
| Afghanistan, Azerbaijan, Cameroon, China, Guinea, Hong Kong, Iraq, Libya, Palestine, Republic of the Congo, Russia, Somalia, Sudan, Thailand, Türkiye | **9** |
| Bahrain, Belarus, Burma/Myanmar, Burundi, Cambodia, Cuba, Equatorial Guinea, Eritrea, Eswatini, Iran, Laos, Nicaragua, North Korea, Saudi Arabia, South Sudan, Syria, Tajikistan, Turkmenistan, United Arab Emirates, Yemen | **10** |

*Table S6. List of phecodes, and their respective phecode categories, used in the analysis carried out in Chapter 4*

| **Phecode Category** | **Phecode** |
|---|---|
| Circulatory System | |
| | Cardiac complications, not elsewhere classified |
| | Chronic pulmonary heart disease |
| | Chronic venous hypertension |
| | Congestive heart failure (CHF) NOS |
| | Coronary atherosclerosis |
| | Essential hypertension |
| | Heart failure NOS |
| | Heart failure with preserved EF [Diastolic heart failure] |

| | |
|---|---|
| | Heart failure with reduced EF [Systolic or combined heart failure] |
| | Hypertension |
| | Late effects of cerebrovascular disease |
| | Myocardial infarction |
| | Other forms of chronic heart disease |
| | Other hypertensive complications |
| | Other specified peripheral vascular diseases |
| | Peripheral vascular disease, unspecified |
| | Primary pulmonary hypertension |
| **Dermatologic** | |
| | Unspecified diffuse connective tissue disease |
| **Digestive** | |
| | Liver abscess and sequelae of chronic liver disease |
| | Other chronic non-alcoholic liver disease |
| | Other disorders of liver |
| **Endocrine/Metabolic** | |
| | Acquired hypothyroidism |
| | Congenital hypothyroidism |
| | Diabetes insipidus |
| | Diabetes mellitus |
| | Diabetes type 1 with peripheral circulatory disorders |
| | Diabetes type 2 with peripheral circulatory disorders |
| | Diabetic retinopathy |
| | Hypothyroidism NOS |
| | Polyneuropathy in diabetes |
| | Secondary diabetes mellitus |
| | Secondary hypothyroidism |
| | Type 1 diabetes |
| | Type 1 diabetes with ketoacidosis |
| | Type 1 diabetes with neurological manifestations |
| | Type 1 diabetes with ophthalmic manifestations |
| | Type 1 diabetes with renal manifestations |
| | Type 2 diabetes |
| | Type 2 diabetes with ketoacidosis |
| | Type 2 diabetes with neurological manifestations |
| | Type 2 diabetes with ophthalmic manifestations |
| | Type 2 diabetes with renal manifestations |
| **Genitourinary** | |
| | End-stage renal disease |
| **Mental Disorders** | |
| | Alcoholic liver damage |
| | Alcoholism |
| | Delirium dementia and amnestic and other cognitive disorders |
| | Dementia with cerebral degenerations |

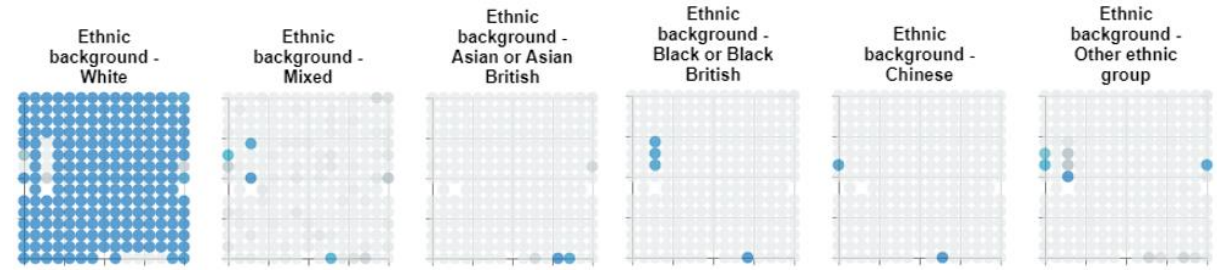
| | |
|---|---|
| | Dementias |
| | Senile dementia |
| | Vascular dementia |
| **Neoplasms** | |
| | Basal cell carcinoma |
| | Bone cancer |
| | Bone marrow or stem cell transplant |
| | Breast cancer |
| | Breast cancer [female] |
| | Breast cancer [male] |
| | Cancer of bladder |
| | Cancer of bone and connective tissue |
| | Cancer of brain |
| | Cancer of brain and nervous system |
| | Cancer of bronchus; lung |
| | Cancer of connective tissue |
| | Cancer of oesophagus |
| | Cancer of eye |
| | Cancer of hypopharynx |
| | Cancer of intrathoracic organs |
| | Cancer of kidney and renal pelvis |
| | Cancer of larynx |
| | Cancer of larynx, pharynx, nasal cavities |
| | Cancer of lip |
| | Cancer of liver and intrahepatic bile duct |
| | Cancer of major salivary glands |
| | Cancer of mouth |
| | Cancer of nasopharynx |
| | Cancer of nasal cavities |
| | Cancer of oropharynx |
| | Cancer of other endocrine glands |
| | Cancer of other female genital organs |
| | Cancer of other female genital organs (excluding uterus and ovary) |
| | Cancer of other lymphoid, histiocytic tissue |
| | Cancer of other male genital organs |
| | Cancer of prostate |
| | Cancer of stomach |
| | Cancer of the gums |
| | Cancer of the mouth floor |
| | Cancer of tongue |
| | Cancer of urinary organs (incl. kidney and bladder) |
| | Cancer within the respiratory system |
| | Cancer, suspected or other |
| | Carcinoma in situ of skin |

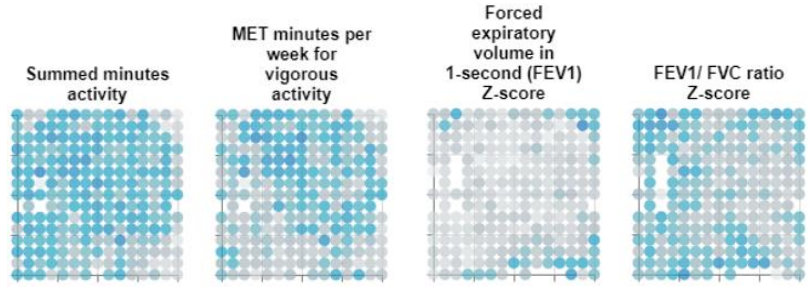
| | |
|---|---|
| | Cervical cancer |
| | Cervical intraepithelial neoplasia [CIN] [Cervical dysplasia] |
| | Chemotherapy |
| | Colon cancer |
| | Colorectal cancer |
| | Hemangioma and lymphangioma, any site |
| | Hemangioma of skin and subcutaneous tissue |
| | Hodgkin's disease |
| | Kaposi's sarcoma |
| | Large cell lymphoma |
| | Leukemia |
| | Lymphoid leukemia |
| | Lymphoid leukemia, acute |
| | Lymphoid leukemia, chronic |
| | Lymphosarcoma |
| | Malignant neoplasm of bladder |
| | Malignant neoplasm of female breast |
| | Malignant neoplasm of gallbladder and extrahepatic bile ducts |
| | Malignant neoplasm of head, face, and neck |
| | Malignant neoplasm of kidney, except pelvis |
| | Malignant neoplasm of liver, primary |
| | Malignant neoplasm of other and ill-defined sites within the digestive organs and peritoneum |
| | Malignant neoplasm of other urinary organs |
| | Malignant neoplasm of ovary |
| | Malignant neoplasm of ovary and other uterine adnexa |
| | Malignant neoplasm of rectum, rectosigmoid junction, and anus |
| | Malignant neoplasm of renal pelvis |
| | Malignant neoplasm of retroperitoneum and peritoneum |
| | Malignant neoplasm of small intestine, including duodenum |
| | Malignant neoplasm of testis |
| | Malignant neoplasm of unspecified male genital organ |
| | Malignant neoplasm of uterus |
| | Malignant neoplasm, other |
| | Manlignant and unknown neoplasms of brain and nervous system |
| | Melanomas of skin |
| | Melanomas of skin, dx or hx |
| | Monocytic leukemia |
| | Multiple myeloma |
| | Myeloid leukemia |
| | Myeloid leukemia, acute |
| | Myeloid leukemia, chronic |
| | Myeloproliferative disease |
| | Neoplasm of uncertain behavior |

| | |
|---|---|
| | Neoplasm of uncertain behavior of breast |
| | Neoplasm of uncertain behavior of male genital organs |
| | Neoplasm of uncertain behavior of skin |
| | Neoplasm of unspecified nature of digestive system |
| | Neurofibromatosis |
| | Nevus, non-neoplastic |
| | Nodular lymphoma |
| | Non-Hodgkins lymphoma |
| | Other non-epithelial cancer of skin |
| | Pancreatic cancer |
| | Polycythemia vera |
| | Reticulosarcoma |
| | Secondary malignancy of bone |
| | Secondary malignancy of brain/spine |
| | Secondary malignancy of lymph nodes |
| | Secondary malignancy of respiratory organs |
| | Secondary malignant neoplasm |
| | Secondary malignant neoplasm of digestive systems |
| | Secondary malignant neoplasm of liver |
| | Secondary malignant neoplasm of skin |
| | Squamous cell carcinoma |
| | Thyroid cancer |
| **Neurological** | |
| | Hemiplegia |
| **Pregnancy Complications** | |
| | Diabetes or abnormal glucose tolerance complicating pregnancy |
| | Endocrine and metabolic disturbances of fetus and newborn |
| | Hypertension complicating pregnancy, childbirth, and the puerperium |
| | Other complications of pregnancy NEC |
| | Preeclampsia and eclampsia |
| **Respiratory** | |
| | Asthma |
| | Asthma with exacerbation |
| | Chronic airway obstruction |
| | Obstructive chronic bronchitis |
| | Respiratory failure |
| | Respiratory insufficiency |
| | Wheezing |

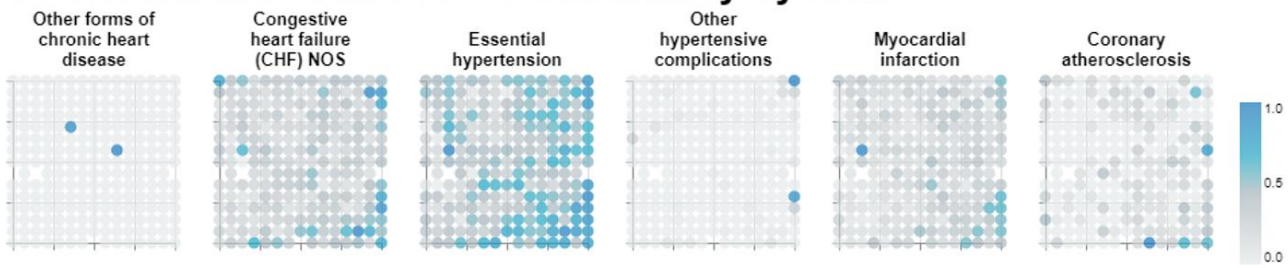
## Visualisation of all the additional investigative variables

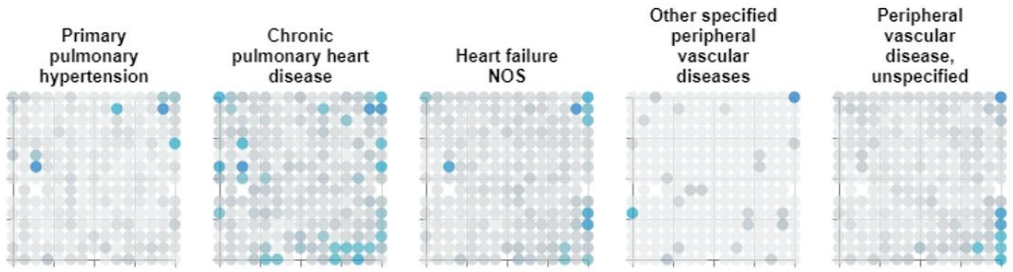
## UK Biobank Investigative variables

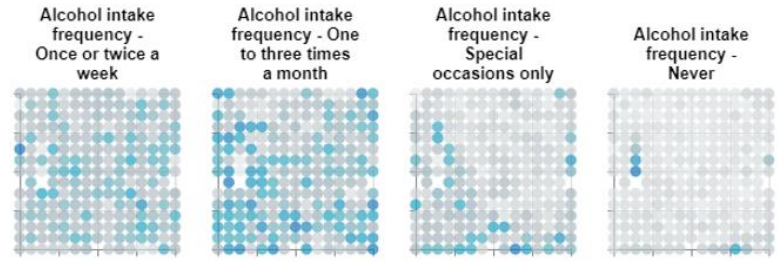


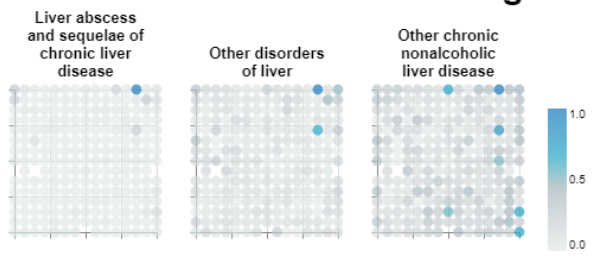

## Comorbidities - Phecodes: Circulatory system



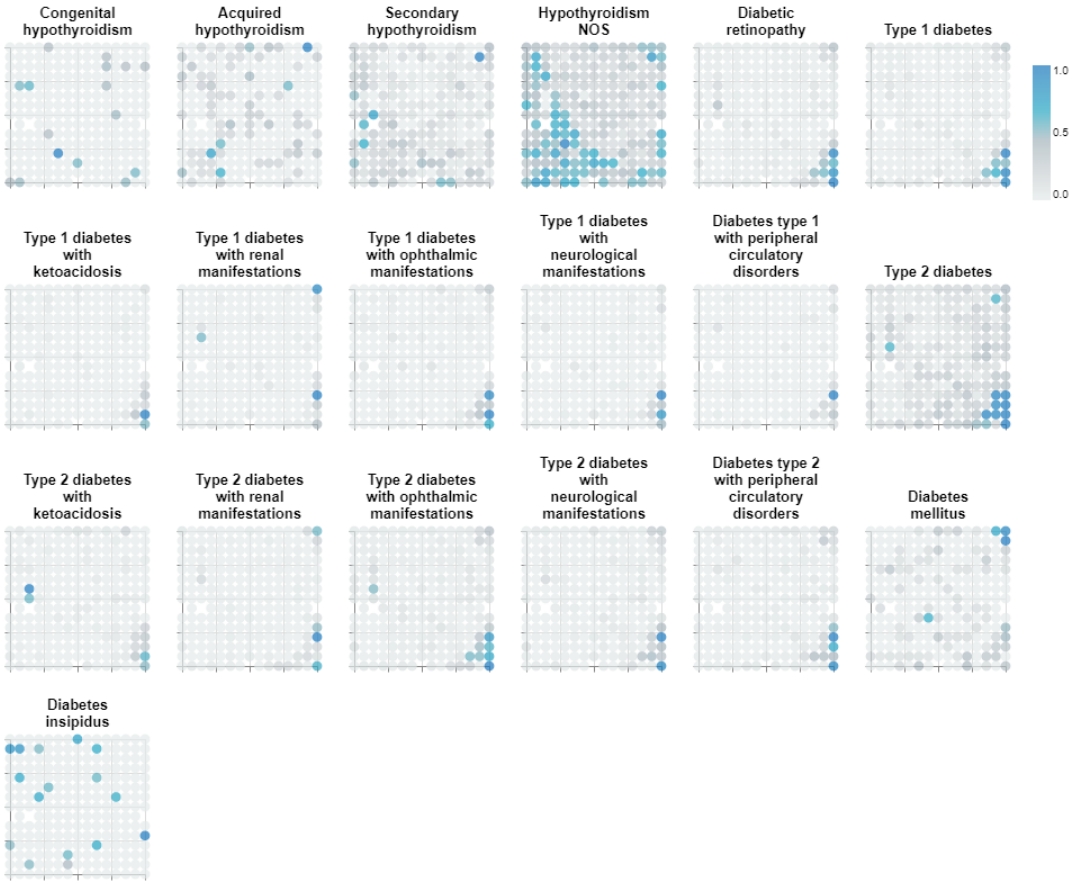



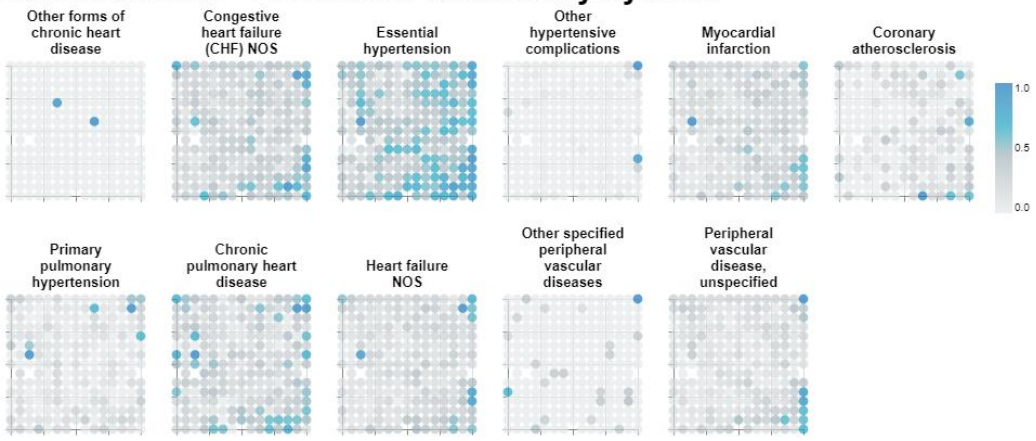

## Comorbidities - Phecodes: Digestive



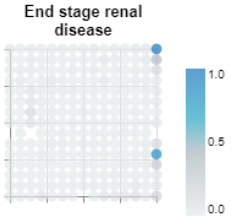
## Comorbidities - Phecodes: Endocrine/metabolic

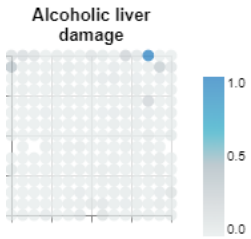



## Comorbidities - Phecodes: Circulatory system

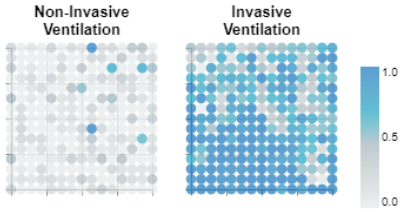



## Comorbidities - Phecodes: Genitourinary

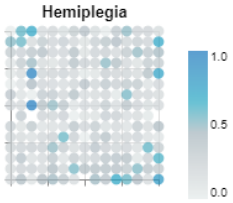




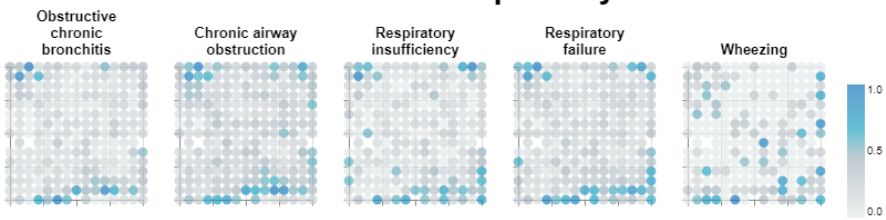

## Category: Procedures




## Comorbidities - Phecodes: Neurological

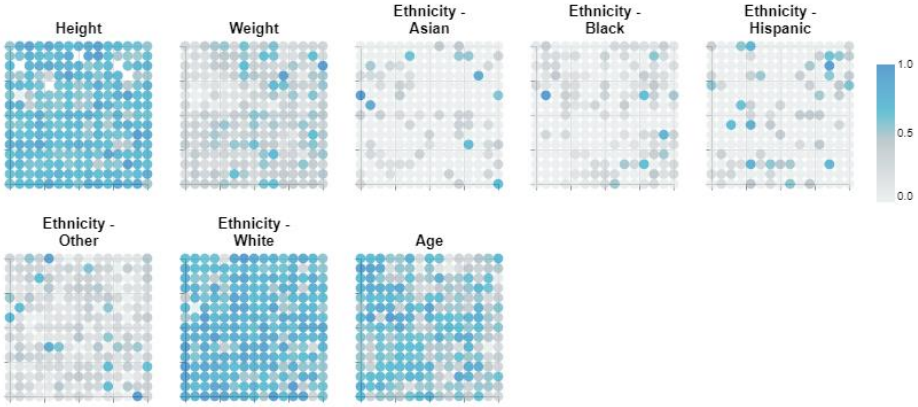


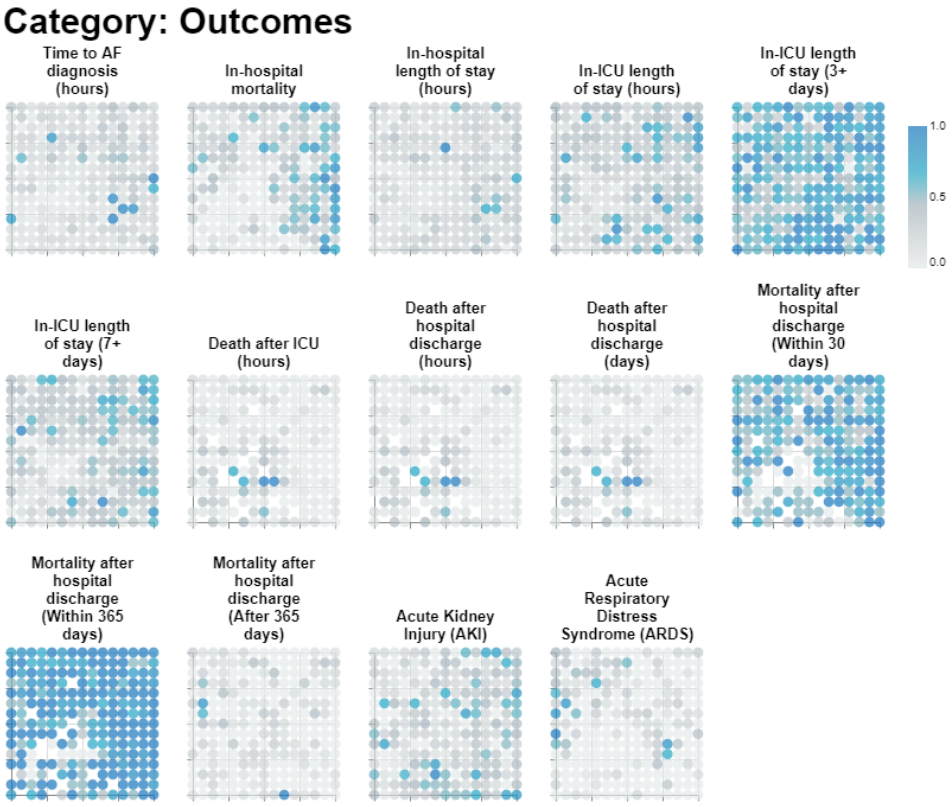

## Comorbidities - Phecodes: Respiratory




**MIMIC-IV Investigative Variables**

## Category: Characteristics




Category: Outcomes
Time to AF diagnosis (hours)
In-hospital mortality
In-hospital length of stay (hours)
In-ICU length of stay (hours)
In-ICU length of stay (3+ days)
1.0
0.5
0.0
In-ICU length of stay (7+ days)
Death after ICU (hours)
Death after hospital discharge (hours)
Death after hospital discharge (days)
Mortality after hospital discharge (Within 30 days)
Mortality after hospital discharge (Within 365 days)
Mortality after hospital discharge (After 365 days)
Acute Kidney Injury (AKI)
Acute Respiratory Distress Syndrome (ARDS)

SS