Adapatable Forecasting Framework in Real-tiMe (AFFIRM)

A thesis presented for an MRes degree in Artificial Intelligence-enabled Healthcare Systems

Candidate number: KFPR2

University College London

Institute of Health Informatics

1st September 2021

Abstract

This thesis presents the first attempt to predict atrial fibrillation (AF) in the Intensive Care Unit (ICU) within hours using machine learning. New-onset AF in the ICU is correlated with a longer length of stay and a greater chance of mortality. Being able to predict whether a patient will suffer an adverse event, such as AF, can allow early treatment of at-risk patients. In this project, a framework called Adaptable Forecasting Framework in Real-tiMe (AFFIRM) was developed for such adverse event predictions.

AFFIRM is an end-to-end pipeline that takes raw time series patient data from High time-Resolution ICU Dataset (HiRID), and trains a machine learning model to predict adverse events during a patient's ICU stay.

For AF prediction, AFFIRM was trained on the data from 31236 patients, 2341 of which developed AF at least two hours after admission. Classification of AF is a particularly challenging task due to several factors: (1) the lack of formal definition of new-onset AF; (2) AF time points cannot be labelled after data collection because AF must be diagnosed from an electrocardiogram (ECG); (3) AF can be transient and asymptomatic and easily missed by clinicians; (4) only around 5.47% of time points are labelled with AF (when data is grouped every two hours), creating a large data imbalance. Despite these difficulties, AFFIRM achieved an area under the precision-recall curve (AUPRC) of 0.59 ± 0.01, with a ratio of one false alert for every true alert and 2.6 missed alerts for every true alert. This value is comparable to state-of-the-art prediction models for circulatory failure (AUPRC = 0.63) and acute kidney injury (AUPRC = 0.297).[1,2] In agreement with previous literature, the most predictive variables for AF were increased age, increased heart rate and low mean arterial pressure.

AFFIRM was also used to predict two other events tachycardia and circulatory failure, achieving AUPRCs of 0.91 ± 0.00 and 0.60 ± 0.01 respectively. Circulatory failure prediction achieved better results than AF prediction despite having an even more imbalanced dataset verifying that AF is uniquely difficult to predict in advance.

Keywords: Atrial Fibrillation, Event Prediction, Machine Learning, ICU, Electronic Health Records

Glossary

Adaptable Forecasting Framework in Real-tiMe (AFFIRM) The machine learning end-to-end pipeline developed in this project. It makes adverse event predictions in the ICU.. 1, 18, 19, 24, 26, 39–42, 44–46

Amsterdam University Medical Centers Database (AmsterdamUMCdb) Critical care dataset of over 20,000 patients from hospitals in the Netherlands. 6

Area Under the Precision-Recall Curve (AUPRC) A machine learning metric that is appropriate for identifying rare events such as AF incidences.. 12, 29, 32, 34–36, 40, 41, 44, 45

Area Under the Receiver Operating Characteristic Curve (AUROC) A machine learning metric that is measures the proportion of correct predictions.. 32, 36

Atrial Fibrillation (AF) The most common form of cardiac arrythmia. The condition causes and irregular and/or fast heartbeat.. 1, 7–13, 15, 16, 19–26, 30, 32, 34–41, 44–46

Body Mass Index (BMI) A measure of height and weight which assesses whether a person is underweight, healthy, overweight or obese.. 19

Clinical Decision Support Systems (CDSS) Technology designed to assist clinicians and improve patient care.. 6, 7

Electrocardiogram (ECG) A medical test used to check the heart's rhythm and electrical activity. It is currently the only way to confirm an atrial fibrillation diagnosis.. 1, 8, 10, 13, 22, 46

Electronic Health Record (EHR) Any electronic document containing information on a patient's health. 6, 7, 12, 13, 46

Electronic Intensive Care Unit (eICU) Multi-center critical care database of over 100,000 patients made by Philips Healthcare in partnership with the MIT. 6

eXtreme Gradient Boosting (XGBoost) Machine learning framework that uses gradient boosting decision trees.. 26, 27, 29, 30, 32, 34–38, 40–45

High time-Resolution ICU Dataset (HiRID) A critical care dataset containing patient data from over 30,000 patients from Bern University Hospital ICU.. 1, 6, 13–19, 23, 30, 44–46

Intensive Care Unit (ICU) Hospital ward for treating critically ill patients. 1, 6, 8, 9, 12, 15, 16, 23, 25, 26, 30, 35, 44, 46

International Classification of Diseases (ICD) Originally used as medical billing codes. ICD codes are used globally to categorise patients' health conditions.. 6

Light Gradient Boosting Machine (LightGBM) Machine learning framework that uses gradient boosting decision trees.. 26, 27, 29, 30, 32, 35, 36, 44

Machine Learning (ML) A form of artificial intelligence. It comprises of algorithms that can learn patterns in data to perform tasks that it is not implicitly programmed for such as event prediction.. 6–10, 12, 15, 17, 25, 26, 28–30, 44, 45

Medical Information Mart for Intensive Care (MIMIC) A critical care dataset containing patient data from over 40,000 patients from ICU of the Beth Israel Deaconess Medical Center. 6, 14, 46

SHapley Additive exPlanations (SHAP) Values calculated that interpret the impact of different features for a machine learning prediction or outcome.. 13, 17, 26, 27, 29, 38, 39, 41–43, 45

Contents

Chapter 1

Introduction

1.1 Machine Learning and Electronic Health Records

The amount of Electronic Health Record (EHR) data has increased rapidly with technological innovation. This has resulted in better record keeping and easier sharing of patient data between healthcare systems. EHRs were originally used for medical billing, reliant on International Classification of Diseases (ICD) hospital codes. ICD codes fail to capture the status of the patient which is often too complex to be denoted by a single code, and patient status can change throughout their hospital stay. However, EHRs can also contain a variety of different data including vital signs and laboratory test results which can be leveraged for machine learning (ML) research. Particularly, there is a large amount of patient data that can be collected from intensive care units (ICU), due to intensive patient monitoring. This has prompted the creation of large openly available Intensive Care Unit (ICU) datasets including: Medical Information Mart for Intensive Care (MIMIC),[3] Amsterdam University Medical Centers Database (AmsterdamUMCdb)[4] , Electronic Intensive Care Unit (eICU)[5] and High time-Resolution ICU Dataset (HiRID).[1] These databases contain time-series data, which is produced by recording successive measurements of quantities over time such as heart rate or temperature.

Clinicians are exposed to a large amount of patient information, particularly in the intensive care unit. Humans have a limited ability to remember vast quantities of information and it is also impossible for them to closely monitor many patients at once.

Clinical Decision Support Systems (CDSS) were designed to assist clinicians in their day-to-day tasks. For example, CDSS can help minimise errors in data entry by specifying a range for a

variable e.g. patient temperature must be between $\sim$ 35-40 degrees Celsius. CDSS may also alert the clinician if a monitored value dips below or exceeds a suitable range e.g. if the patients' heart rate drops to less than 50 beats per minute.

There are two types of CDSS, knowledge-based and non-knowledge based. Knowledge-based CDSS were created based on prior knowledge and "if-then" rules e.g. if variable x is elevated then prescribe y drug. Non-knowledge-based CDSS is based on ML.

If adverse event prediction ML models can be sufficiently trained and rigorously tested, then they could be used to efficiently aid intensive care clinicians. ML would be able to make decisions personalised to the patient and, in theory, provide better patients outcomes. Currently, there is no such machine learning CDSS tools used in practice as, much like clinical drug trials, ML tools must be evaluated over many years before deployment. Moreover, ML models tend to use "black boxes" to make decisions on data which means the models cannot be explained. Ethically, it is also unclear who is responsible if a machine learning model causes a negative outcome. The difficulty in explaining ML decisions and the lack of accountability are major hurdles to employing ML models as CDSS.

Many ML models have been trained on EHR data to predict adverse events continuously over a patients' hospital stay, however, such methods have yet to be applied to atrial fibrillation (AF) prediction.[6–7]

1.2 Atrial Fibrillation

Atrial fibrillation (AF) is the most common type of cardiac arrhythmia, with a lifetime risk of 25%. AF occurs when the electrical signals that maintain heart rhythm start to fire more rapidly and chaotically, leading to an increased and abnormal heart rhythm.[8,9] Some people can live for years with AF without problems, however, without treatment, AF can cause the heart to work less efficiently leading to heart failure and blood clots. AF condition is independently correlated with a 3-5 fold increased risk of stroke[10] and double the risk of mortality.[9]

There is currently no known cause for AF however it is associated with several conditions. Risk factors for developing AF in the community include advanced age, heart disease and recent cardiac surgery. In 10% of cases, no underlying heart disease is found and, in these cases, there is an indication that factors such as alcohol, stress, electrolyte imbalances, infections and genetic

factors may also increase the risk of AF. In some cases, no cause can be found.

AF is diagnosed using an electrocardiogram which captures the electrical impulses that travel through the cardiac muscle. The aim for treating AF is regaining normal heart rhythm and controlling heart rate which is called cardioversion. Cardioversion can be achieved using medications called rate or rhythm control drugs. Rate control drugs reduce the heart rate while rhythm control drugs help to return the heart to a normal rhythm. Rhythm control drugs may have adverse side effects and patients require monitoring when they take such drugs. Anticoagulants may also be administered to reduce the risk of blood clots and stroke. If these drugs fail to correct or control AF, patients may be defibrillated, this is electrical cardioversion when an electrical signal is sent to the heart to restart it. Defibrillation is often successful, but the heart can revert to AF so it must still be controlled using drugs.[11]

It is also common for cardioversion to never succeed and the patient will have permanent AF and their condition must be managed using drugs. AF can also be transient lasting for only a few hours at a time.

1.3 Atrial Fibrillation in Critically Ill Patients

ICU patients frequently develop AF as a complication of critical illness; the incidence of new-onset AF ranges from 5% to 46%.[12−13] New-onset AF is, therefore, a marker of disease severity and has been shown to lead to a longer length of stay and higher risk of morbidity. While there are extensive guidelines in treating AF in a community setting, there is insufficient evidence to make recommendations about a standard treatment for patients in the ICU.[11] Not only is it more difficult to restore a critically ill patient's heart rate to normal rhythm[14] , but also medication options are more limited depending on the patient's co-morbidities.[15]

AF is a condition that can be easily misdiagnosed due to its similarity with other arrhythmias.[16] Furthermore, AF can only be diagnosed using an electrocardiogram (ECG), as such transient episodes of AF may be undiagnosed.[17] Undiagnosed and misdiagnosed AF can lead to late treatment which puts the patient at greater risk of further complication.

1.4 Rationale

Evidence shows that new-onset AF may be preventable.[18] Given that critically ill patients that develop new-onset AF have a longer length of stay in the ICU and higher mortality; an ML event

prediction model for AF would be valuable.[19−20] Furthermore, as AF is likely a predictor of disease severity, early prediction of AF will highlight which patients' conditions will deteriorate. While risk factors for AF in the community are known, factors that contribute to new-onset AF are yet to be confirmed. Creating an explainable ML model may also illuminate which patient variables are causing patients to develop AF.

1.5 Project Aims and Objectives

This project aims to create a pipeline that will preprocess raw ICU time-series quickly and make a prediction on whether a patient will develop AF within the next few hours. The prediction will be made using a binary classifier model that should satisfy three criteria to maximise its clinical usefulness for intensivists: (1) change over time dependent on new data; (2) rely on modifiable variables such as electrolyte levels that can be adjusted by a clinician, as well as static factors; (3) be transparent by ranking variable importance in prediction.

Additionally, the preprocessing pipeline and prediction model must be easy to use and generalise to predict other adverse events with varying hours of prediction into the future and with different binary classifiers. This will allow other health data scientists to make baseline predictions before optimising their models for a specific task.

Chapter 2

Literature Review

2.1 Machine Learning for AF Classification

The focus of applying machine learning (ML) to atrial fibrillation (AF) has been for the detection/classification of electrocardiogram (ECG). AF ECG signals can often be mistaken for other arrhythmia. Also transient and/or asymptomatic episodes of AF can go unnoticed.[21,22]

Hannun et al. developed a deep learning convolutional neural network classifier that detects and distinguishes between 12 different heart rhythm classes, including AF. The neural network was trained on expert annotated single-lead ECGs. The resulting model was able to exceed the positive predictive value of an average cardiologist, F1-score = 0.837 vs 0.780. While the results of this paper are highly impressive, it is noted that AF and other heart rhythm classes are diagnosed with a 12-lead ECG rather than a single-lead. Therefore, the annotations from cardiologists may be inaccurate and hence lead the average cardiologist, who would usually examine 12-lead ECGs, to perform worse than the model.[21]

Attia et al. used 650,000 12-lead ECGs to train a convolutional neural network that could identify AF. However, the performance of this network was much lower than that of Hannun et al. with an F1-score = 0.392.[22] This suggests that 12-lead ECGs are harder to diagnose than single-lead.

The Apple Heart Study is conducting an ongoing study monitoring $\sim$ 400,000 consenting Apple Watch customers, for cardiac arrhythmia. Apple watches can measure pulse rate from the wrist using photoplethysmography which can identify pulse irregularity or variability. The results of this study can inform AF screening in the population. Screening AF early can potentially inform management, increase the probability of successful cardioversion and in turn reduce future costs

and burdens to the health system. However, if the precision of this technology is insufficient, then the greater proportion of the population will be screened unnecessarily for AF, thereby increasing the burden on the health system.[23]

2.2 Rule-based AF Prediction

While there has been significant progress in AF detection/classification using ML, there is limited innovation in ML-based AF prediction. Predicting future AF may illuminate how it develops in the first place. By gaining such knowledge, there is a potential to prevent AF before it occurs. The majority of AF prediction models use rule-based risk scores calculated from static features, such as age, sex and pre-existing conditions, to predict AF within years.[24−25] Alonso et al. created the CHARGE-AF risk score, a simple risk score that predicts AF in the population using features such as age, race, height, weight, blood pressure, medication and pre-existing conditions. They validated their risk score on a diverse population of over 20,000 people and found discrimination between case and control of 60-70%.[26−27]

Similarly, Saliba et al. used the existing CHADS2 and CHA2DS2-VASc risk scores, originally designed to predict the risk of stroke in AF, to predict new-onset AF in the population. These scores are calculated using age, sex and pre-existing conditions such as congestive heart failure. They studied around one million adults for an average of three years each. They found that the incident rate of AF increased with increasing CHADS2 and CHA2DS2-VASc scores.[24]

Christophersen et al. compared the CHARGE–AF and CHA2DS2-VASc risk scores and found that CHARGE-AF performs better in both discrimination and calibration. CHARGE–AF is regarded as the most validated and most accurate rule-based risk score for AF.[26]

Hill et al. created a machine learning prediction model based on primary care data collected from 3 million people with no history of AF in the five years before the study. As well as static data, the authors examined temporal data such as read codes and blood pressure within a rolling 12-month window. They followed the patients until AF diagnosis, death, loss to follow-up or the end of the study. They compared several machine learning models: neural network, random forest, SVM and logistic regression. They chose to use the neural network based on the AUROC scores for each model. The authors conclude that their neural network had greater precision and accuracy than CHARGE-AF. In this paper, they only describe the neural network

as "time-varying" but do not describe the type of neural network they used or the architecture of the network. They also do not describe any data preprocessing steps before the use of their ML algorithm. Furthermore, AUROC is an inappropriate metric for comparing their ML models as there is a large imbalance between the number of patients who develop AF and those who do not. It would be more appropriate to measure the Area Under the Precision-Recall Curve (AUPRC) or use F1-scores.

Tiwari et al. used Electronic Health Record (EHR) data from 2 million individuals to classify whether they would develop AF within six months. Using a shallow neural network on 200 EHR features only yielded a result slightly better than using logistic regression on age, sex and known AF risk factors.[28]

Karnik et al. used free-text and coded EHR data to predict AF within 1, 3, 5 and all years. The text-based dataset achieved the best model with an F1-score of 0.601 where the coded data did not significantly increase performance.

While these studies are useful for preventative guidance, they are limited in clinical usefulness in the Intensive Care Unit (ICU) as clinicians require a short-term assessment of their patients within hours or days.[29] Furthermore, these predictions are based on static features which clinicians cannot treat or manage. Currently, several rule-based scoring systems are used in the ICU including APACHE, SOFA and Glasgow Coma Scale however, there is no such score for AF.

2.3 Atrial Fibrillation Risk Factors in the ICU

Several studies have identified risk factors associated with new-onset AF. New-onset AF is more common in patients with electrolyte derangement and greater disease severity.[30–31] Inflammation may also damage cardiac muscle tissue which leads to atrial contractile dysfunction, causing AF.[32] Elevated levels of adrenaline and noradrenaline, a consequence of greater illness severity, are also associated with AF development.[33] Comparisons, using echocardiography, showed that patients with new-onset AF have a larger left atrium than those who had no AF where enlarged left atrial size is thought to be associated with diastolic dysfunction.[34] The patient population sizes of these studies were limited so the exact causes of AF remain unknown.[35] ML can detect patterns in data that are not accessible to humans, which makes it suitable for defining personalised disease risk factors.[36]

2.4 Event Prediction Using Machine Learning

Using time-series databases, machine learning time point classification models have been developed to predict diseases and events including sepsis[37] , acute kidney injury[2] and circulatory failure.[1] For the prediction of events in the future, labels are brought forward in time.

Tomašev et al. developed an early prediction model for acute kidney injury within 48 hours. Out of over 700,000 patients, the model could predict 55.8% of acute kidney injury episodes. The rolling predictions occurred within 6 hour time windows. They transformed the high dimensional patient data into a lower-dimensional representation using embedding. The embeddings were fed into a recurrent neural network which in turn was fed into a final prediction layer.

Hyland et al. created an early warning system for circulatory failure using the High time-Resolution ICU Dataset (HiRID).[1] Their system used machine learning to provide a rolling prediction of cardiovascular failure within eight hours, every five minutes. They found that a LightGBM classifier resulted in the best performance, they predicted 90% of circulatory failure events and predicted 82% more than two hours in advance. The authors also used shapelets, time series subsequences, to improve their model's predictive power. Features used in the model were ranked by importance for each prediction using SHapley Additive exPlanations (SHAP) values, to increase the transparency of their model.[38]

This project aims to create a warning system similar to that of Tomašev et al. and Hyland et al., applied to AF prediction. AF event prediction poses additional challenges as it must be diagnosed using ECGs and cannot be diagnosed directly from EHR data. Therefore, AF labels must be created during the patient's stay and cannot be annotated AFterwards like most other conditions. Furthermore, AF can be transient and asymptomatic so some episodes of AF may go undetected. AF usually lasts for hours, however, a clinician who enters AF into the database may only label AF once rather than for every time point throughout the whole duration of AF.

Another aim of this project is to create a generalised event prediction pipeline. The van der Schaar lab has already achieved such a pipeline called Clairvoyance.[39] Clairvoyance takes static and temporal patient data to create a rolling prediction of an event. However, Clairvoyance requires data to be inputted with a specific format i.e. a table with columns: patient id, time, patient variable and value. The authors of Clairvoyance also offer preprocessing code for MIMIC. However, this project will differ from Clairvoyance by being tailored to take the raw HiRID data

as input as well as allow the user more choices concerning preparing the data. HiRID has a far greater time-resolution compared to Medical Information Mart for Intensive Care (MIMIC) which is more beneficial for time-dependent event prediction. HiRID has measurements up to every two minutes whereas MIMIC data is collected hourly.[1]

Chapter 3

Method

3.1 Setting

The research setting for this project is the Intensive Care Unit (ICU), where 5-46% of patients develop new-onset atrial fibrillation (AF).[12−13] New-onset AF leads to immediate haemodynamic effects, which requires immediate management of the patient's condition. A machine learning (ML) model can be trained on ICU data to evaluate which patients are at high risk of developing AF during their stay.

3.2 Dataset

This project will use the High time-Resolution ICU Dataset (HiRID), a freely accessible, deidentified, time-series ICU dataset. HiRID was collected between 2008 and 2016 from Bern University Hospital's interdisciplinary intensive care unit. The data contains 712 routinely collected physiological variables, test and treatment parameters stored at a high time resolution with at least one entry every two minutes throughout the patients' stay.

HiRID was chosen over the other three publicly available large ICU databases (eICU, MIMIC, AmsterdamUMCdb) because of its distinct advantages for training an ML model for AF prediction. HiRID contains time-stamped episodes of AF which is a necessity for training the ML model. HiRID also has measurements every 2 minutes which is the highest time resolution out of all of the public ICU databases. This high time resolution is suitable for time series classification and, by extension, event prediction.

$\sim$15% (4991/33905) patients in the HiRID database have recorded instances of AF which is sufficient for ML model development and it is in keeping with the literature case rate suggesting

that the AF annotations are reliable.

The raw HiRID database is $\sim$ 6Gb in size and is split into four tables:

- *General* - Patient information including admission time, age and sex
- *Observations_table* - Timestamped physiological patient measurements
- *Pharma_records* - Timestamped pharmacological information
- *Reference* - The variables in *Observations_table* and *Pharma_records* are recorded with IDs. The reference table provides the variable names and units for the variables according to the ID.

The data structure is illustrated in Fig. 3.1.

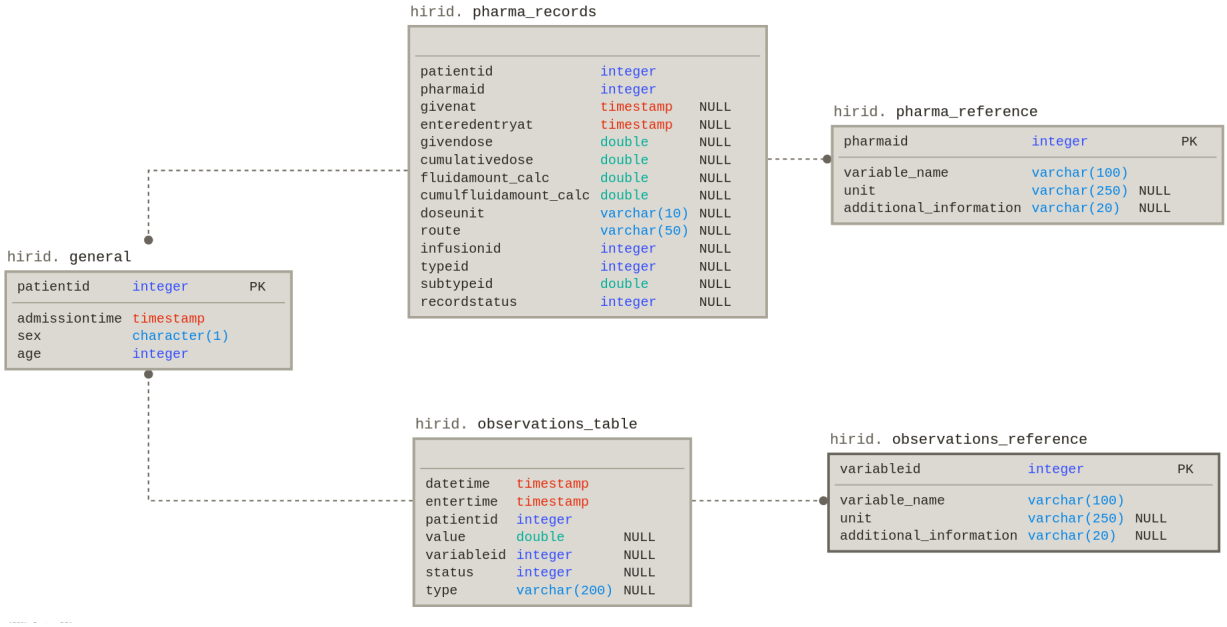

Figure 3.1: Database schematic for the high time resolution ICU data set (HiRID)

3.3 Ethics

HiRID is anonymised per HIPAA and GDPR, laws responsible for patient privacy in the US and individual privacy in the EU, respectively. The anonymisation process consisted of removing all 18 identifying data elements listed in HIPAA, shifting admission dates into the future and

binning the patients' age, height and weight into intervals of five. Furthermore, the dataset does not include any clinical notes that could potentially contain patient information.

While the previously discussed ML algorithms show promising results, there are still many boundaries for them to be used in real-world healthcare systems. Issues such as lack of transparency, bias and privacy are problems in all areas for ML research, however, these issues can have life-threatening consequences when ML is applied in healthcare settings.

Bias in ML can arise from the unbalanced data the models are trained on. For instance, Hyland et al's early prediction model was trained on the HiRID database which is collected from Bern University Hospital in Switzerland. The model is trained on the demographic of the city of Bern which is predominantly white and affluent. Hence, if the model was deployed in other areas of the world, the model would be biased against other ethnic groups and potentially produce harmful predictions. In order, to mitigate bias algorithms must be tested on varieties of groups. This may be mitigated in future work by evaluating the model on external data. Bias can also be a result of biased domain knowledge. ML models for healthcare are built with input from clinicians. However, if the clinicians are biased according to their experience, the models may also have the same bias.

ML models that do not explain how it comes up with its results are called "black box" models. Clinicians are unlikely to heed the advice given from black box models because their predictions may not arise from true patterns seen in the data but confounders. Therefore, ML models must be transparent and explainable. In EHR research, this involves highlighting which features are most important during diagnosis e.g. heart rate. If the most important variables seem irrelevant to the diagnosis, it could be confounding or it could be that the model has discovered a new correlation. SHapley Additive exPlanations (SHAP) values can rank feature importance during a binary classification and therefore improve the transparency of the model.

3.4 Design

As mentioned previously, this project aims to use raw HiRID data for event prediction. This process requires three main steps:

1. Data Preprocessing. A pipeline for preprocessing the raw HiRID data

2. Data Prepararation. A pipeline for preparing the data for binary classification

3. Event Prediction and Evaluation. A framework that takes the prepared data and evaluates several different binary classifiers using 10-fold stratified cross-validation. It can then perform hyperparameter optimisation and calculate feature importance.

These three steps represent parts of the whole pipeline which will be referred to as Adaptable Forecasting Framework in Real-tiMe (AFFIRM). The next sections will describe how AFFIRM works, any variables that are required as input into AFFIRM will be written in italics. A summary of AFFIRM is shown in Fig. 3.2.

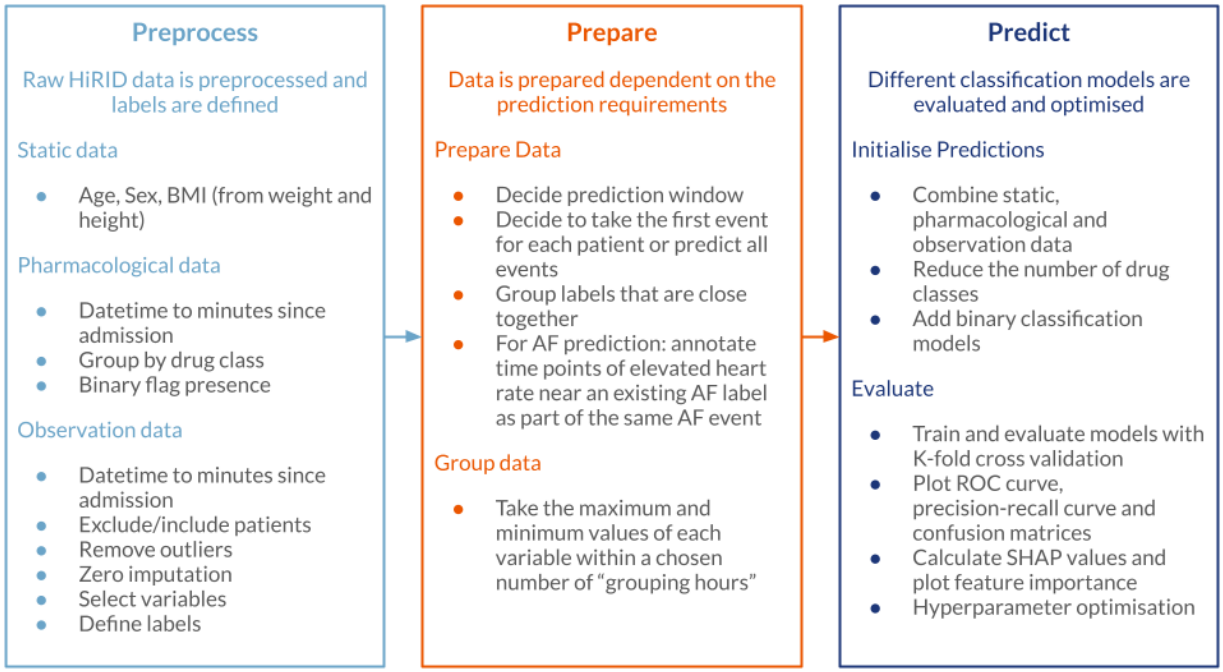

Figure 3.2: Overview of the steps involved in the AFFIRM (Adaptable Forecasting Framework In Real-tiMe) pipeline. AFFIRM consists of a preprocessing pipeline, a data preparation pipeline and a prediction/evaluation pipeline. The overall pipeline can be easily adapted to predict different adverse events.

3.4.1 Data Preprocessing

The data was preprocessed to reduce its size by removing unnecessary columns and changing the format to create a more efficient representation of the data.

The HiRID data is stored in 250 CSV or parquet files. This project used parquet files which are much faster to read than CSV files. To process the data, the 250 files are looped through the preprocessing pipeline. The pipeline then saves the 250 preprocessed files inside another directory.

The static data is extracted from the raw data for each patient. This includes sex, age, weight and height. To reduce the number of variables and to create a more meaningful representation of the patient, the weights and heights were converted to Body Mass Index (BMI) values.

The timestamps for the pharmacological data was changed to minutes since admission with intervals of five minutes. The pharmacological variables were then grouped by drug class using a manually created annotation. There are over 500 pharmacological variables however most of the drugs are not unique. Drugs with the same active ingredients have unique variable IDs because they are recorded with varying dosages, brand names and delivery systems. The amount of pharmacological data is less than a tenth of the observational data so the drugs were grouped by drug class to decrease the number of variables with relatively few values. Grouping the variables by drug class means equating the dosage and the delivery system, therefore the drug classes are represented by a binary flag describing whether the drug was given or not at a time point.

The timestamps for the observed physiological measurements were similarly converted to minutes since admission. Any values that were 0 were removed because they are likely to be errors, this is apart from variables where 0 is a valid categorical value such as the Glasgow coma scale. Any variables with different names that represent the same measurement are grouped and can be specified by defining *rename_dict*. For example, the temperature of the patient is recorded in the data as core body temperature, rectal temperature and axillary temperature. These variables are grouped together as temperature.

The variables *include_patients* or *exclude_patients* can be specified in order to include or exclude patients according to the type of patient i.e. trauma patients. As the number of AF patients is limited, it was decided to include all types of patients.

The *parameter_dict* allows the user to specify the event labels for example AF is described in HiRID as a circadian rhythm value of 10. *parameter_dict* could, for example, be changed to predict circulatory failure by specifying lactate below 2 and mean arterial pressure above 65. The pipeline also gives the option of adding previously labelled data, which would be necessary for events that require more complex labelling.

Next, the outliers have to be filtered out. There are two ways to combat outliers, either specifying a range for each variable or by taking a range of quantiles for the data. Both types of filtering were tried with similar results, therefore AFFIRM uses quantiles to save the user time as specifying

ranges for each variable can take a long time and would require domain knowledge. The filtering takes the values between the 0.01 and the 0.99 quantiles. Then, if the standard deviation is greater than the mean after the filtering, values between the 0.1 and 0.9 quantiles are taken. This is because if the standard deviation is greater than the mean, then there are a significant number of values that are much larger than the mean. These values would be highly unlikely for physiological measurements, as physiological measurements tend to have a small valid range. Outliers arise from human error when the clinicians enter the patient data into the database, for example, one patient's temperature was erroneously entered as 463 degrees, as this is impossible, it was most likely meant to represent 46.3 degrees.

The number of observational variables is then reduced. The variable selection must be a balance between trying to keep as many variables but also discarding any variables that have little data. To do this, the user can specify *percentage_patients_per_variable* which the percentage of patients that have at least one measurement for each value and the *avg_values_each* which is the average number of values that a patient has for each variable. The default values are 0.8 and 2 respectively which yields around 40 variables.

Finally, all three table types, static, pharmacological and observation are converted from long-form to wide form data, i.e. the data changes from a table with patient id, time, variable as columns to a table where patient id and minutes since admission form the index. Fig. 3.3 shows the difference between the structure of the raw data and the preprocessed data.

Fig. 3.4 shows an example of a patient stay after the data was preprocessed with a few of the features.

3.4.2 Data Preparation

The 250 preprocessed files of the observation data and pharmacological data are fed through a data preparation pipeline. The static data is not processed any further during data preparation.

The preparation pipeline first groups any event labels that are close to each other. The proximity of the labels required for grouping is specified by the user variable *group_within*. When clinicians label AF, they can only label one time point at a time rather than a range of time points as shown in Fig. 3.4. It would be inefficient for the clinician to label every single time point for the duration of the AF however it can be inferred that the patient has continuous AF between the AF labels. Several recorded AF times are in temporal proximity suggesting that they form the same

a)

	datetime	value	variableid	Label
0	2158-12-07 08:55:00.000	165.00	10000450	Body height measure
1	2158-12-07 08:55:00.000	70.00	10000400	Body weight
2	2158-12-07 08:59:38.620	0.50	212	ST elevation
3	2158-12-07 09:00:45.200	0.20	212	ST elevation
4	2158-12-07 09:01:51.720	0.30	212	ST elevation
...	...	...	...	...
267513	2158-12-24 06:00:00.000	5.00	24000427	Amylase [Enzymatic activity/volume] in Serum o...
267514	2158-12-24 06:00:00.000	5.00	24000427	Amylase [Enzymatic activity/volume] in Serum o...

b)

Minutes Since Admission	Base excess in Arterial blood by calculation	Bicarbonate [Moles/volume] in Arterial blood	Carbon dioxide [Partial pressure] in Arterial blood	Heart rate	Invasive mean arterial pressure	Lactate [Mass/volume] in Arterial blood	Oxygen [Partial pressure] in Arterial blood	Potassium [Moles/volume] in Blood	pH of Arterial blood
295	-2.575660	-0.719549	-0.320909	1.447431	0.874534	0.209345	-0.767897	0.051243	-0.520693
300	-2.573929	-0.728227	-0.289402	1.389235	0.834710	0.206834	-0.766378	0.049456	-0.555052
305	-2.572199	-0.736904	-0.257895	1.447431	0.927633	0.204323	-0.764860	0.047668	-0.589412
310	-2.570469	-0.745581	-0.226388	1.389235	0.861260	0.201812	-0.763341	0.045881	-0.623772
315	-2.568739	-0.754258	-0.194882	1.412513	0.847985	0.199300	-0.761822	0.044093	-0.658131
...	...	...	...	...	...	...	...	...	...
25465	-1.357592	-0.055463	-0.973073	0.679239	-0.134337	1.393301	-0.294209	0.197825	0.888060
25470	-1.357592	-0.055463	-0.973073	0.667600	-0.174160	1.393301	-0.294209	0.197825	0.888060

Figure 3.3: A diagram to show the difference between the raw and cleaned data. a) Example of raw data from one patient, b) Data after processing

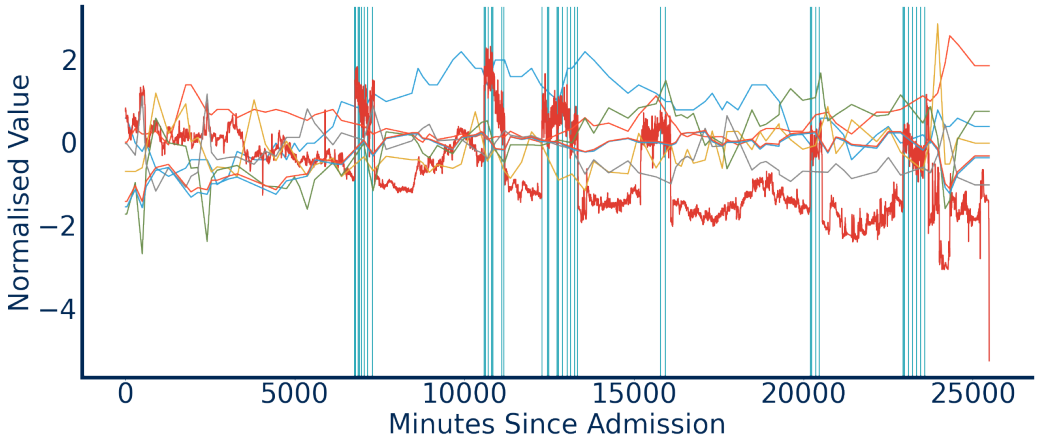

Figure 3.4: Example patient stay. Raw AF labels are shown by vertical blue lines. Only a select number of variables shown.

AF episode. As such, AF times are grouped if they occurred within *group_within* of each other. Fig. 3.5 illustrates the grouping of the AF labels.

Furthermore, the clinician may not mark the exact start or end of an AF episode. They may even label one instance of AF during a whole duration. AF often lasts for a few hours, but the time point labels have a resolution of seconds. To try to combat this, we try to annotate any missing

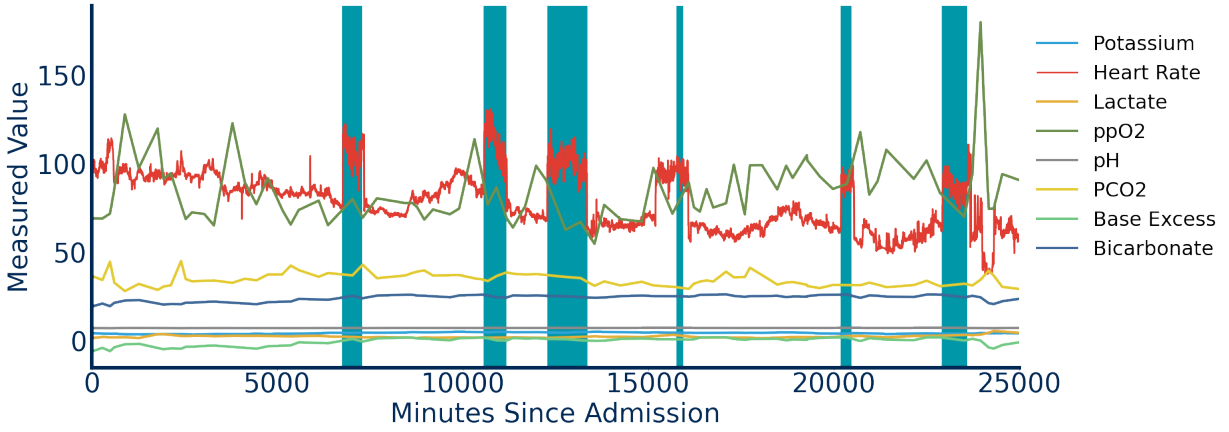

Figure 3.5: Example patient stay with the AF labels grouped if the AF labels are within two hours of each other. AF labels shown as blue vertical lines. Only a select number of variables shown.

AF labels. AF is known to coincide with a sudden increase in heart rate or a rapid heart rate (over 100 bpm), Fig. 3.6 also shows that regions of increased heart rate correspond to episodes of AF. Sometimes AF may be recorded late, this may be due to human error as clinicians diagnose AF from electrocardiogram (ECG)s. Therefore, if there is a sudden peak in heart rate or the patient has over 100 bpm within *group_within* number of minutes of an AF label, then those time points are also labelled with AF.

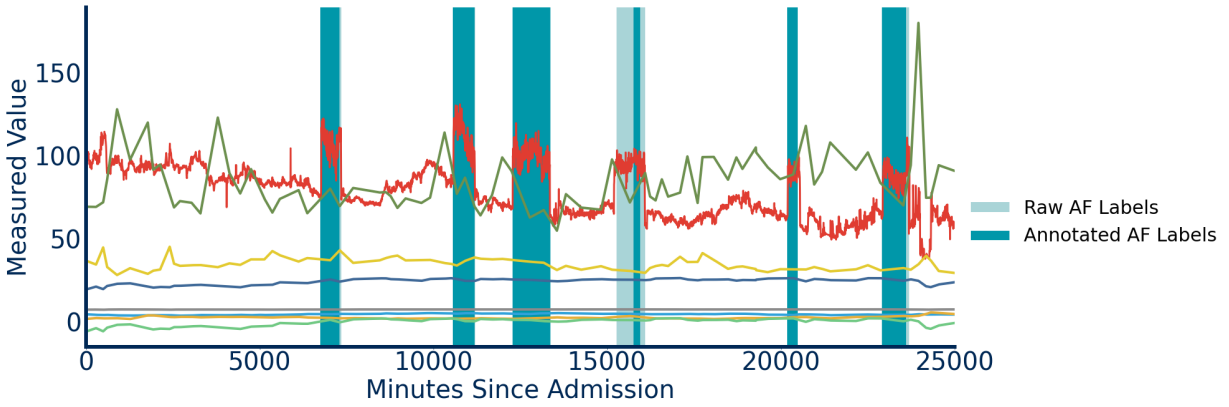

Figure 3.6: AF annotations including areas of elevated heart rate within two hours of an AF label.

As stated before, the preprocessed data has measurements every five minutes. However, the majority of physiological measurements were taken over much larger time intervals, resulting in many missing values in the data. Therefore, to reduce the number of missing values the data points are further grouped every two hours. Two hours was chosen as it was deemed long enough to reduce the amount of missing data but also short enough to make more predictions during a

patient's ICU stay. The number of hours in which to group the data can be varied by the user using the *grouping_hours* variable. Patients who developed AF within two hours of admission or any patients with a length of stay less than two hours were subsequently discounted. Fig. 3.7 shows a histogram describing the distribution of the first AF episode for patients that developed AF, measured by hours since admission.

The majority of patients develop AF within the first hour, particularly in the first twenty minutes. It would be difficult to predict which patients will develop AF within the first hour as there would be insufficient data to predict from. Furthermore, it may be that patients who develop AF so early in their ICU stay have a pre-existing AF condition. As this project is focused on new-onset AF, it is beneficial to filter these patients out, despite the loss in case numbers. The larger the *grouping_hours*, the more patients would have to be discounted. Choosing to group the data every two hours reduces the number of case-patients from 4991 to 2213. The data can be grouped using the maximum, minimum or mean of the two hours. For AF, the mean did not have much predictive power, so the maximum and the minimum values were taken e.g. the heart rate variable was split into the variables maximum heart rate and minimum heart rate. This resulted in doubling the number of observation variables, the number of static and pharmacological variables stayed the same. For some variables, taking the maximum and minimum of the variable over two hours resulted in the same value, due to infrequent measurements. Any duplicate columns were subsequently removed. The final number of variables was 83.

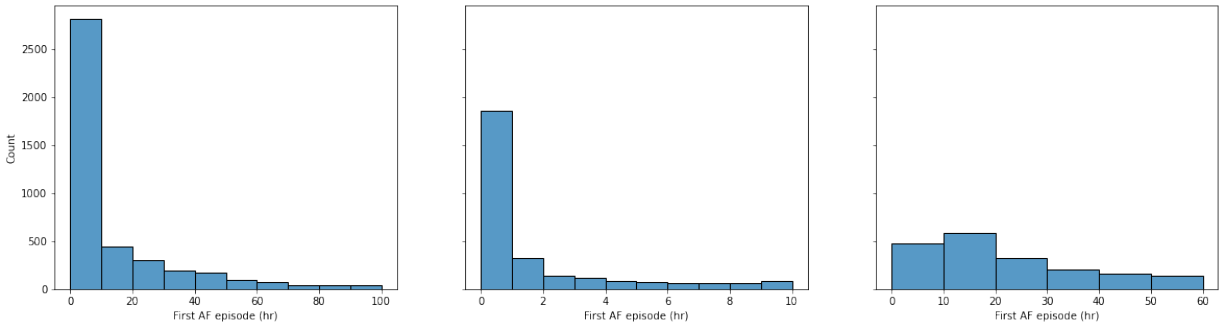

Figure 3.7: A figure to show the distribution of the first onset of AF of patients in HiRID, measured by hours since admission.

Not only does grouping the measurements reduce the amount of missing data, but it also serves to reduce the number of time points to classify which is necessary due to the large imbalance of data. There are 2×10^7 control time points compared to 6×10^4 (0.28%) case time points when

the measurements are every five minutes. When the measurements are grouped every two hours, the number of control time points is 8.3×10^5 and the case time points are 4.8×10^4 (5.47%).

The user can also optionally only take the first instance of an event. When patients develop AF they are far more likely to have it reoccur during their stay. It could therefore be beneficial to the clinician to only predict the first instance of AF rather than predict all occurrences. However, taking the first instance of AF will cause the already large data imbalance to increase drastically. Therefore, all occurrences of AF were kept.

Finally, the data labels must be shifted into the past dependent on the number of hours ahead the user wants to predict an adverse event (*predict_hours*). As a result of the label shifting, the labels do not represent the presence of AF but instead the onset of AF within the next *predict_hours*.

Hyland et al. chose to predict circulatory failure eight hours into the future and Tomašev et al. predict acute kidney injury 48 hours into the future. These time frames were chosen according to the amount of time before onset the clinicians have to make an effect such as administering a certain drug in order to prevent the adverse event. As it is currently unclear how to prevent AF, there is no established time window before onset for which it can be prevented. As a result, the best amount of time must be deduced based on the known characteristics of onset AF.

Choosing too few hours ahead of time may result in better predictions because the patients' physiological measurements would be closer to the AF case however, it may not be enough time for the clinician to act. Choosing too many hours ahead of time, however, decreases the certainty of when AF will occur. The gap between AF events can range from hours to days, and the average patient length of stay was $\sim$ 54 hours, therefore it is likely to be useful to predict AF within hours rather than days. Therefore a comparison between the number of hours in advance can be made including 4, 6, 8 and 10 hours. Comparison of results should reveal an optimum value for *predict_hours*.

3.4.3 Event Prediction and Evaluation

After data preparation, all the data is stored in three separate tables: static, pharmacological and observations. Grouping the variables every two hours, greatly reduces the size of the tables and allows for the data to be handled easily rather than looping through 250 files.

A Python class called AFFIRM was created for event prediction. The AFFIRM object takes as its input the three tables and concatenates all three of them into one dataframe. By keeping the

three tables separate the user has to option to use the tables separately. Also, the observation table has far more rows than the other two tables because it has data every five minutes from the beginning to the end of admission. Whereas the static dataframe only has one data point per patient and the pharmacological dataframe only has rows with at least one non-missing value. Therefore, it would be inefficient to store the three tables joined together as there would be many missing and repeated values.

Previously, all the pharmacological variables were kept in the data because the pharmacological data was not very large and all the variables could therefore be easily stored. However, many of the variables are not useful for prediction because of the number of missing values e.g. anaesthetic drugs are common in patients whereas only a few patients are prescribed antiparkinsonian drugs. In this stage, the user can choose which pharmacological variables they are interested in keeping. Previous, literature suggests that vasopressors and low potassium levels can increase the chances of AF in the ICU. Therefore, it was important to keep all the vasopressor and potassium information. The other drugs were kept according to the variable *pharma_quantile*. The number of times each drug class was prescribed was summed, the drugs that had greater than the *pharma_quantile* of the sum range were kept. For example, the maximum number of drug counts was the antithrombotic drugs with 475045 counts and the minimum number of drug counts was chemotherapy drugs with only 381 counts. The default *pharma_quantile* value is 0.75 which is 56586 counts. 12 variables have greater than 56586 counts: anaesthetic, antibiotic, antithrombotic drugs, bronchodilators, insulin, magnesium, non-opioid painkillers, nutrition supplements, opioid painkillers, potassium, vasodilators, vasopressors.

Even after the data preprocessing and data preparation pipelines, there remain missing data points in the dataframe. Most ML models require no missing data. There are several methods for filling in the missing data. Interpolation is usually used for chronological data such as this ICU data because it uses the other time points in the same sequence to plot a polynomial and therefore estimate the missing values. However, some patients didn't have measurements for certain features, as such those features could not be interpolated. In this use case, interpolation is also considerably slower than other imputation methods because interpolation must be carried out patient-wise. A much faster imputation method is simply filling the missing values with the mean or median for all the data. However, this method highly skews every value towards the

control time points because there is such a large data imbalance. Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost) ML classifiers can handle inputs with missing values. In initial tests, the performance of these classifiers exceeded other models that used interpolated or mean/median-imputed data. To handle missing values, LightGBM and XGBoost simply use zero-imputation, which involves replacing all the missing values with zeroes. Therefore, zero imputation was chosen to be used for all the data.

After initialising the data, the user can specify some binary classification models to train. AFFIRM contains some prebuilt models that can be added by just using some keywords for example *AFFIRM.add_model("lightgbm")* will add a LightGBM model to the framework. The user can choose LightGBM, XGBoost, logistic regression, random forest and a Keras model. These models will only have the default classifier parameters and the Keras model is a fully connected model with three layers of eight nodes. Keras was trained until the validation metric stopped improving. The user can also add in their own models so that they can explore different types of binary classifiers or with optimised parameters.

After adding all the models to the framework, they can all be trained and tested. The user can specify how many folds they want to cross-validate. In this project, 10-fold stratified cross-validation was used, which means splitting the data and training into nine portions of data with one portion as the test set, repeated ten times. This is illustrated in Fig. 3.8.

Cross-validation verifies whether the results produced are reliable and consistent. Stratified cross-validation ensures that there is the same proportion of control and case data points in each fold. This is especially important for imbalanced data where one fold may not contain any case data points and therefore not be representative of the whole dataset. After the cross-validation, receiver operating characteristic curves and precision-recall curves are plotted side by side containing the results of every model. The plots will include a shaded area denoting the standard deviation of values during each of the validation folds and the mean average values are plotted as a solid line.

SHAP values can then be calculated for the classifier predictions. SHAP values are a measurement of the contribution of each variable to the classification and therefore explains which features are most important in the classification. This can be especially significant for AF as there is no known cause for AF in the ICU. ML can spot patterns that humans cannot, so it may

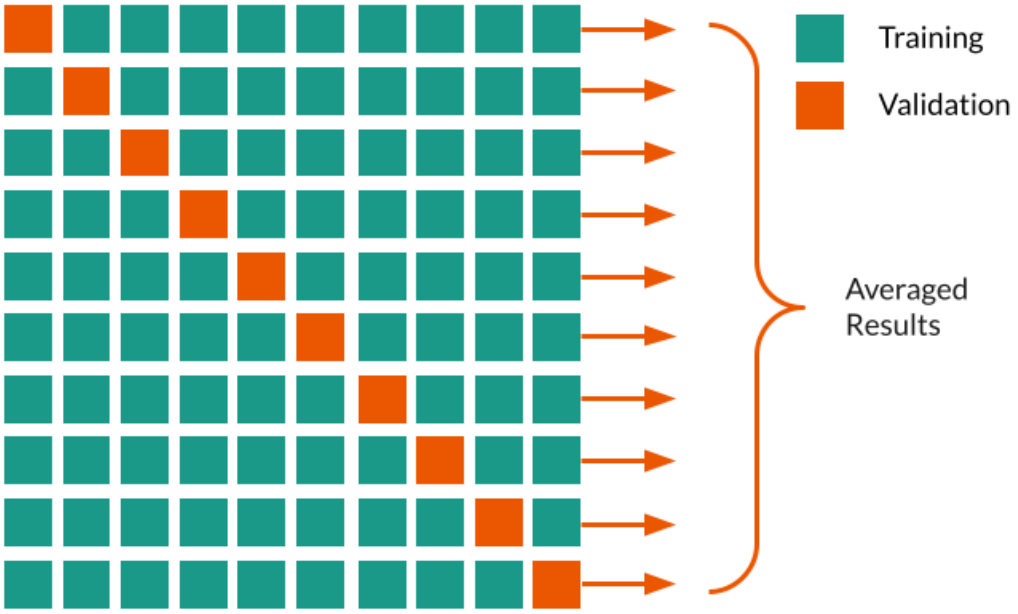

Figure 3.8: Illustration of 10-fold cross validation. The data is split into 10 "folds". With each iteration one of the folds becomes the validation set and the other folds are the training set. The model is trained on the training set and validated on the validation set. This happens until each fold has been validated. The results from each fold is then averaged.

find a variable that has been overlooked by clinicians.

3.5 Binary Classifier models

The performance of six binary classifier models are compared in this project: logistic regression, Keras, decision tree, random forest, LightGBM and XGBoost. The decision tree model will act as a baseline with fewer variables, in order to simulate rule-based based on information from previous literature. The six models are illustrated in Fig. 3.9.

In logistic regression, the inputs x are entered into the model. A weighted sum of the inputs, $h(x)$ shown in Eq. 3.2 are calculated then entered into a sigmoid function. The sigmoid function, shown in Eq. 3.1 takes in any input and outputs a value between zero and one which represents a probability where a probability of one indicates the positive class and the zero indicates the negative class. The SHAPe of a sigmoid curve is also found in Fig. 3.9.

As the logistic regression model is trained i.e. when the model is given an input, the output will be compared to the ground truth label then the weights of the equation are updated. The updated weights will mean that the model has a greater ability to predict the correct class with a

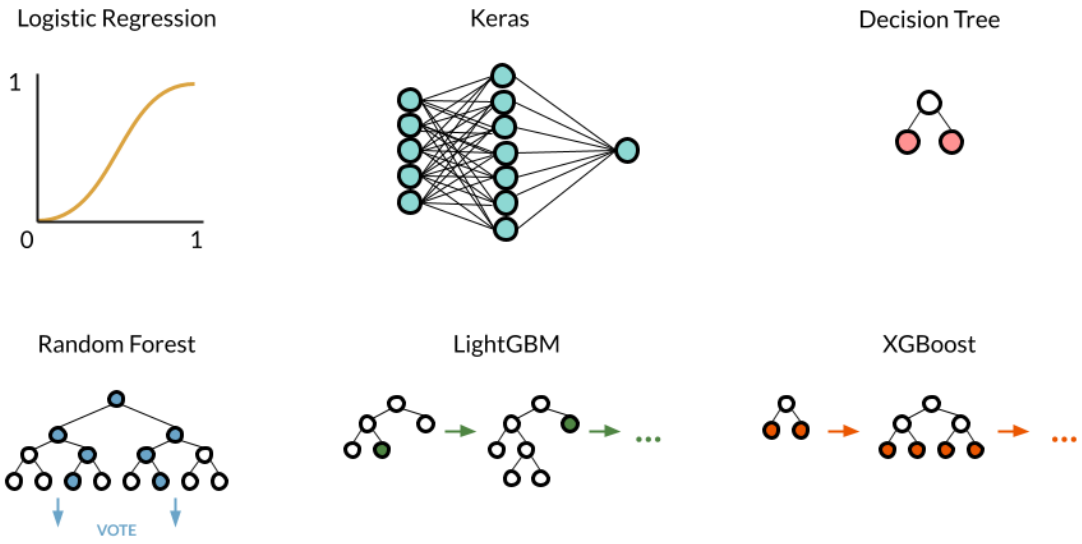

Figure 3.9: An illustration of the six binary classification models used in this project.

probability. Training is finished when the loss, which is the difference between the ground truth label and the output, is at a minimum. The process of training is known as gradient descent. When the logistic regression model is tested, it will output a probability for each test case. The probabilities above 0.5 will be determined as the positive prediction and below 0.5 will be a negative prediction.

Eq. 3.2 shows the equation for weighting the inputs.

$$\sigma(h) = \frac{1}{1+e^{-h}} \tag{3.1}$$

$$h(x) = \Theta^T X \tag{3.2}$$

Where h is the weighted inputs, Θ are the weights, X is is the input and σ is the sigmoid function.

Keras is a ML library that allows the user to easily create deep neural networks. In this project, Keras is used to create a fully connected multilayer perceptron neural network. The neural network receives input and outputs a prediction, in this case, a binary prediction, on the input. Between the input and output layers is at least one hidden layer. Each node has a linear activation function that maps the weighted inputs into outputs at each node. The activation function used

in this project will be ReLU which often achieves the best possible performance. The equation for ReLU is given in Eq. 3.3.

$$ReLU(x) = \begin{cases} 0, & \text{for x} < 0 \\ x, & \text{for x} \geq 0 \end{cases} \tag{3.3}$$

Where x is the input data.

The fully connected neural network means all the nodes are interconnected, so all the outputs of a previous layer become the inputs of the current layer. Similarly to logistic regression, the weights at each node are updated as the input is fed through the neural network and compared with a ground truth label. Deep learning models, such as those built with the Keras library, can be more powerful than other ML models. Deep learning models are especially useful for processing more complex data such as images. Due to the relatively few features and small dataset, a deep learning model may not be the best model because it may overfit the small dataset. This is because deep learning models work by abstracting data through neural network layers. These relatively simple EHR variables do not have significant "hidden" information, unlike images. Furthermore, the more complex a deep learning model becomes, the slower the computation time. Considering these factors, several different Keras models will be compared where the number of layers and the number of nodes will be varied to see which has the best performance. Also, the abstraction of information in deep learning means that it is difficult to explain deep learning predictions and ML requires transparency, especially in healthcare. On the other hand, all the other models are easily explained using SHAP values.[38] Six Keras model architectures will be compared, these will comprise of either two or three hidden layers of 8, 16 and 64 nodes. For simplicity, the same number of nodes will be used for each layer. The models will be trained until the Area Under the Precision-Recall Curve (AUPRC) fails to improve.

The LightGBM, XGBoost and random forest classifiers are based on decision trees. Decision trees are supervised learning algorithms. The input is given at the "root node" and different decisions are made about the data which splits the data into further decision nodes until a "leaf node" is reached at which point the data is classified. The main advantage of decision trees is that it is easy to interpret. As decision trees are meant to mimic the process of "decision making", a decision tree classifier will be used as a baseline model to compare the performance of the other models. This baseline model will also have a limited number of variables based

on literature, whereas the other models will rely on as many variables as possible from the data. From literature, new-onset AF in the ICU is more likely to occur due to advanced age, being male, having increased heart rate, an electrolyte imbalance, illness severity and the use of vasopressors. As such these variables will be used in the baseline classification. Lactate levels will be used as an indicator of illness severity.

Random forest ensembles are constructed of many decision trees where the output of the random forest is the class selected by the majority of the decision trees. Fig. 3.9 shows how a random forest can be constructed out of many decision trees, where there is a vote on the output of the trees.

Relatively recently, gradient boosting decision trees have been developed and typically outperform random forest ensembles. One such gradient boosting framework is XGBoost which has gained popularity for winning many ML competitions. In random forests, the decision trees are built at the beginning whereas, in gradient boosting frameworks, decision trees are added at each stage to compensate for the existing weak learners. LightGBM is a relatively new gradient boosting framework that works similarly to XGBoost but tends to be much faster and requires less memory to run. XGBoost was originally based on level-wise growth where the decision trees were grown level by level. On the other hand, LightGBM uses a leaf-wise growth where the leaf node with the highest loss is split into more nodes which is why it converges faster than XGBoost however it is more prone to overfitting on smaller datasets. The HiRID dataset is large enough for LightGBM to not overfit, hence why the LightGBM and XGBoost have similar performances. Fig. 3.9 illustrates the difference between the models growing leaf-wise (LightGBM) and level-wise (XGBoost).

3.6 Hyperparameter Optimisation

The best model from the evaluation will be chosen for hyperparameter optimisation. In this project, the results will show that the XGBoost model performed the best. The optimisation framework, Optuna, was used for hyperparameter optimisation of XGBoost.[40] Optuna is also compatible with all six of the binary classifiers used in this project. During hyperparameter optimisation, the framework shuffles through the parameter space to find the best values for each parameter, by using AUPRC as its evaluation metric. The parameters that achieve the

highest AUPRC value are the best parameters for the task. The more iterations the optimiser goes through, the more of the parameter space can be searched. More iterations require greater computation time and power. To find the best possible performing model, the UCL's Myriad high performance computing cluster is used to search through many more parameter combinations that would otherwise take a very long time on a local machine. The optimiser was run for 5000 trials.

Chapter 4

Results

4.1 Initial Results

There are many classes of binary classifiers to choose from, each of which works in different ways. Several binary classifiers can be evaluated to see which is the best for atrial fibrillation (AF) prediction. Furthermore, these classifiers are initially tested on four future prediction windows (4, 6, 8 and 10) to see which prediction window is the best for AF prediction. Hyperparameter optimisation of the best classifiers will produce even better AF prediction results.

In these initial results, five different binary classifiers are compared: Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), logistic regression, random forest and a fully connected deep learning model made in Keras, using only default parameters. The usual evaluation metric for binary classification is the Area Under the Receiver Operating Characteristic Curve (AUROC) which delineates the balance between the true positive rate and the false positive rate. However, the AF prediction data is highly imbalanced which means that the AUROC will give an optimistic measure of how well the classifier performs. The Area Under the Precision-Recall Curve (AUPRC) is more illustrative of the performance of a classifier with imbalanced data. Recall describes the number of correct positive predictions out of all positive predictions made whereas precision quantifies the number of correct positive predictions made. Both recall and precision are focused on the minority positive class as opposed to the majority negative class. The AUPRC is, therefore, the ideal metric as, in the context of event prediction, the positive class is the most important.

The complexity of a deep learning model architecture can greatly affect prediction performance.

Table 4.1: Table showing the patient characteristics of the control and case patients used to train the ML model for AF prediction. Patients who develop AF within two hours, or have a length of stay less than two hours were filtered out

		Case	Control
Total Patients		4598 (13.7%)	28920 (86.3%)
Male		2893 (67.1%)	17443 (64%)
Female		1420 (32.9%)	9792 (36%)
Age			
Mean		69.8 ± 11	59.3± 15.8
Median			
Range		20-90	20-90
Mortality			
Alive		4017 (88.8%)	27239 (94.7%)
Dead		507 (11.2%)	1517 (5.3%)
Patient Type			
Cardiovascular	*Non-surgical*	1063 (21.4%)	3539 (11.2%)
	Surgical	958 (19.2%)	7685 (24.2%)
Neurological	*Non-surgical*	821 (16.5%)	5211 (16.4%)
	Surgical	261 (5.2%)	4368 (13.8%)
Gastrointestinal	*Non-surgical*	367 (7.4%)	2282 (7.2%)
	Surgical	236 (4.7%)	1712 (5.4%)
Trauma	*Non-surgical*	149 (3%)	1507 (4.8%)
	Surgical	46 (0.9%)	407 (1.3%)
Pulmonary		538 (10.8%)	1997 (6.3%)
Sepsis		201 (4%)	541 (1.7%)
Other		90 (1.8%)	625 (2%)
Surgical Respiratory		71 (1.4%)	678 (2.1%)
Metabolic/Endocrinology		71 (1.4%)	592 (1.9%)
Surgical Gynecology		40 (0.8%)	304 (1%)
Surgical Orthopedics		32 (0.6%)	124 (0.4%)
Hematology		20 (0.4%)	84 (0.3%)
Surgical Urogenital		13 (0.3%)	70 (0.2%)

Before comparing the five binary classifies, several Keras models, of different complexities, were first compared.

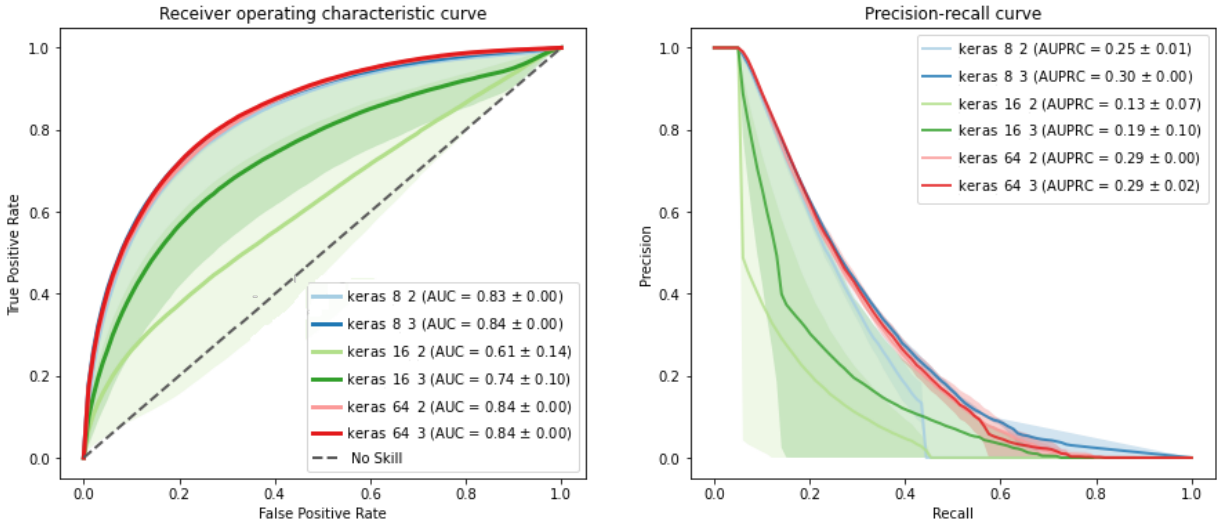

Figure 4.1: Comparison of fully connected neural networks using Keras for AF prediction, six hours in advance. The first number indicates number of nodes per layer and the second number is the number of layers e.g. keras 8 2 means two layers of eight nodes.

Fig. 4.1 is a comparison of Keras models that vary in complexity. In general, the eight nodes per layer and the 64 nodes per layer achieved better results than the 16 nodes per layer despite being intermediate. The 16 nodes per layer also had the highest standard deviation. The best performing model was the three layers with eight nodes in each layer, so this architecture was chosen for comparison with the other binary classifiers.

Table 4.2 shows the AUPRC for AF prediction for five different binary classifier models (not including the baseline) and four different prediction windows, where the models are initiated with default parameters and the Keras model has three layers of eight nodes. Table 4.2 shows that the models at each prediction window have similar AUPRC values which suggest that the models are almost equally skilled at predicting 4, 6, 8 and 10 hours ahead. The lack of variation suggests that the time range is not significant i.e. there would likely be a greater range in AUPRC if the prediction window was increased to 24 or 48 hours. The random forest classifier was the worst-performing whereas the XGBoost had the best performance. Out of every prediction window for XGBoost, the 4-hour and 6-hour had the greatest AUPRC values, when taking into account the standard deviation. The 6-hour prediction window was deemed the best prediction window. The 6-hour time window was chosen over the 4-hour time window as it is more helpful to clinicians

Table 4.2: Initial AF prediction results. Five binary classifier models and four prediction windows are compared using the AUPRC metric.

	Area under Precision-Recall Curve			
	Prediction Window			
Model	4	6	8	10
Logistic Regression	0.06 ± 0.00	0.07 ± 0.00	0.07 ± 0.00	0.07 ± 0.00
Keras	0.28 ± 0.02	0.26 ± 0.08	0.27 ± 0.04	0.29 ± 0.02
Random Forest	0.26 ± 0.01	0.26 ± 0.01	0.26 ± 0.01	0.26 ± 0.00
LightGBM	0.47 ± 0.01	0.46 ± 0.01	0.45 ± 0.01	0.44 ± 0.01
XGBoost	**0.53 ± 0.01**	**0.52 ± 0.00**	**0.51 ± 0.01**	**0.51 ± 0.01**

in the Intensive Care Unit (ICU) to have access to AF prediction information earlier.

Fig. 4.2 shows the results for each binary classifier after 10-fold stratified cross-validation, where the shaded area shows the standard deviation between folds. The receiver operating characteristic curve is not an ideal measure for classifier performance due to the high imbalance in data, however, it is included as a sanity check.

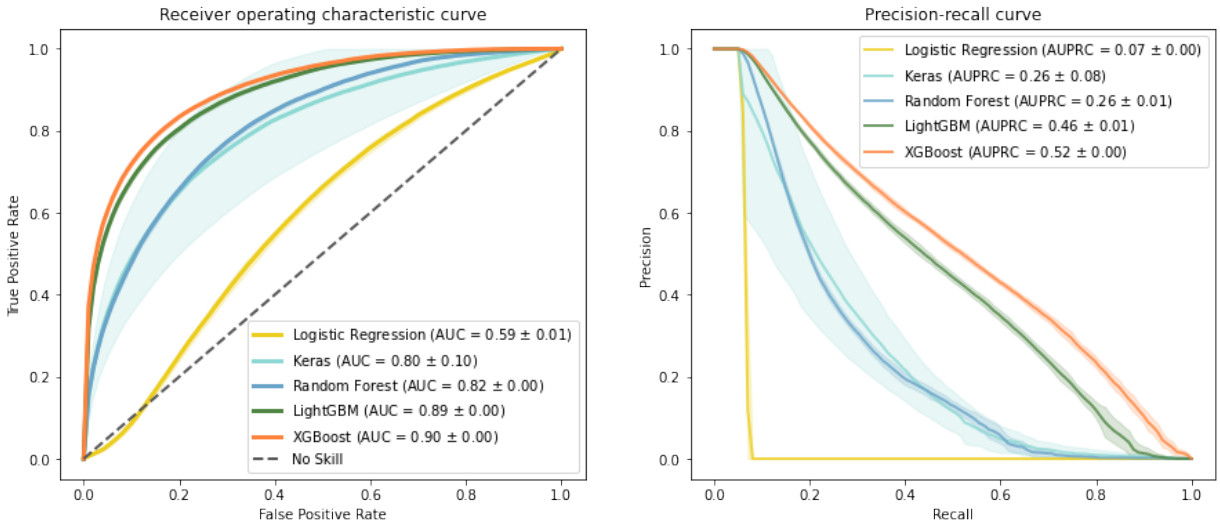

Figure 4.2: Five binary classification results using default parameters for predicting atrial fibrillation within six hours.

Notably, the Keras model has a high standard deviation compared to the other models. The same model in Fig. 4.1 yielded a mean AUPRC of 0.3 whereas in Fig. 4.2 shows a mean AUPRC of 0.26 with a standard deviation of 0.08. This is due to the stochastic nature of neural networks, the weights on all the nodes are initialised randomly therefore as they undergo gradient descent, the weights will reach different minima and therefore have different performances.

Table 4.3: Time taken to train each model 10 times. Training times will be different depending on what machine is used.

Model	Time taken to evaluate 10 fold (mins)
Baseline (decision tree, fewer variables)	*0.51*
Logistic Regression	2.09
Keras (batch size = 512, average 30 epochs)	34.66
Random Forest	8.52
LightGBM	**0.54**
XGBoost	7.81

Table 4.3 shows the amount of time it took to train all of the models for 10 fold cross-validation. On average LightGBM took the fastest time to train, taking only 0.54 minutes, whereas XGBoost took 7.81 minutes. A short computation time is highly desirable in a prediction model as, in a real-world setting, having more time to take action on these ML predictions is vital. The baseline model took only 0.02 minutes less time to train than the LightGBM. It is to be expected that the baseline model should train the fastest because it is a simpler binary classifier with fewer variables and therefore less data from which to train. As the variables are grouped every two hours, 7.81 minutes is a relatively short amount of time, therefore XGBoost would still be useful in a real-world setting. XGBoost is therefore still the best model, with a relatively fast training time, the highest AUPRC value, consistency and ease of explainability. Fig 4.4 shows the confusion matrix for the XGBoost predictions. To produce an even better model, hyperparameter optimisation was performed on XGBoost using the Optuna library over 5000 iterations, maximising the AUPRC. Optuna was able to find models that achieved a higher AUPRC than the default but this was at the expense of the area under the receiver-operating curve (AUROC). This would not be a problem as the main metric used for evaluation was AUPRC however the AUROC was close to the baseline, meaning that the results were inconsistent. AUROC should naturally be close to 0.9 because of the large data imbalance. However, a better XGBoost model was found after manual hyperparameter optimisation and the result is plotted in Fig. 4.3.

The optimised XGBoost model shows an improvement on the default XGBoost model. Due to the infinite possible combinations of hyperparameter values, it is uncertain whether it is the best possible model. For comparison, a baseline model was also plotted in Fig. 4.3. This baseline is meant to mimic “decisions” made by the clinician dependent on features that have known, but not yet validated, correlation to new-onset AF. Therefore, a decision tree classifier was trained

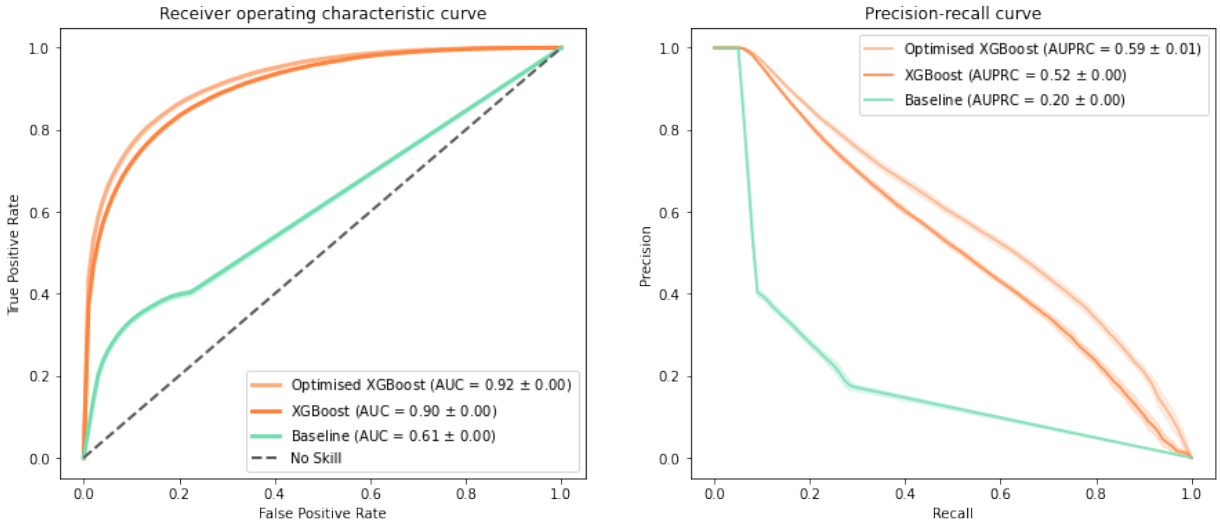

Figure 4.3: The results for predicting AF six hours in advance using a XGBoost model, an optimised XGBoost model and a baseline model which is a decision tree with fewer variables chosen based on previous literature.

on variables including static features, heart rate, electrolytes and the presence of vasopressors.

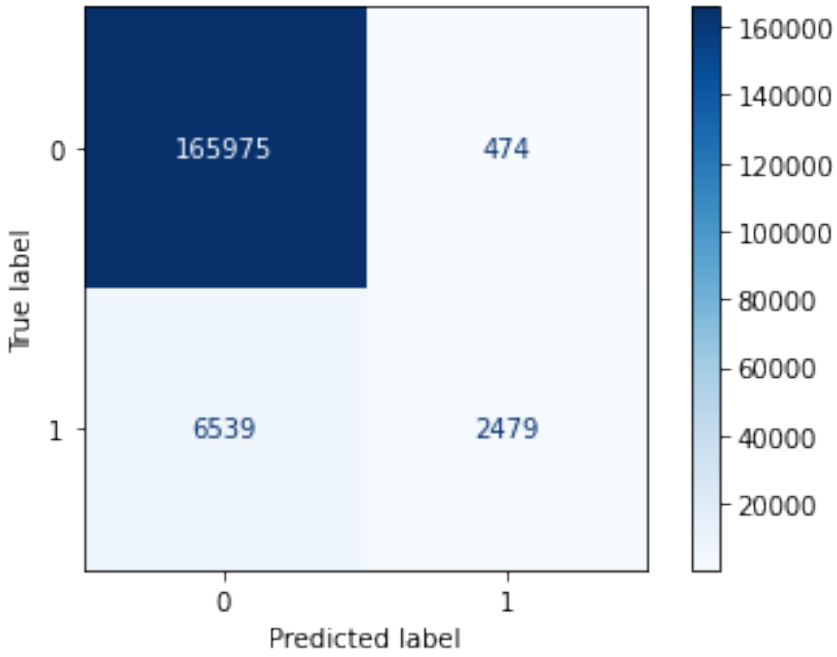

Figure 4.4: Confusion matrix of the optimised XGBoost model, predicting six hours in advance

The confusion matrix Fig. 4.4 shows that there is one false alert for every true alert and 2.6 missed alerts for every true alert.

4.2 SHAP values

SHapley Additive exPlanations (SHAP) comes from the game theory concept of the Shapely value. SHAP values are a way to explain the outputs of machine learning models. In this case, SHAP values will highlight which variables, and whether these variables are relatively high or low, contribute most to determining if a patient will develop AF.[38]

Explainability is highly important in machine learning, especially in the application to healthcare. Clinicians and patients are unlikely to trust black-box machine learning model predictions. Examining SHAP values can help determine if the machine learning model is working correctly by comparing it to previous literature and clinician experience.

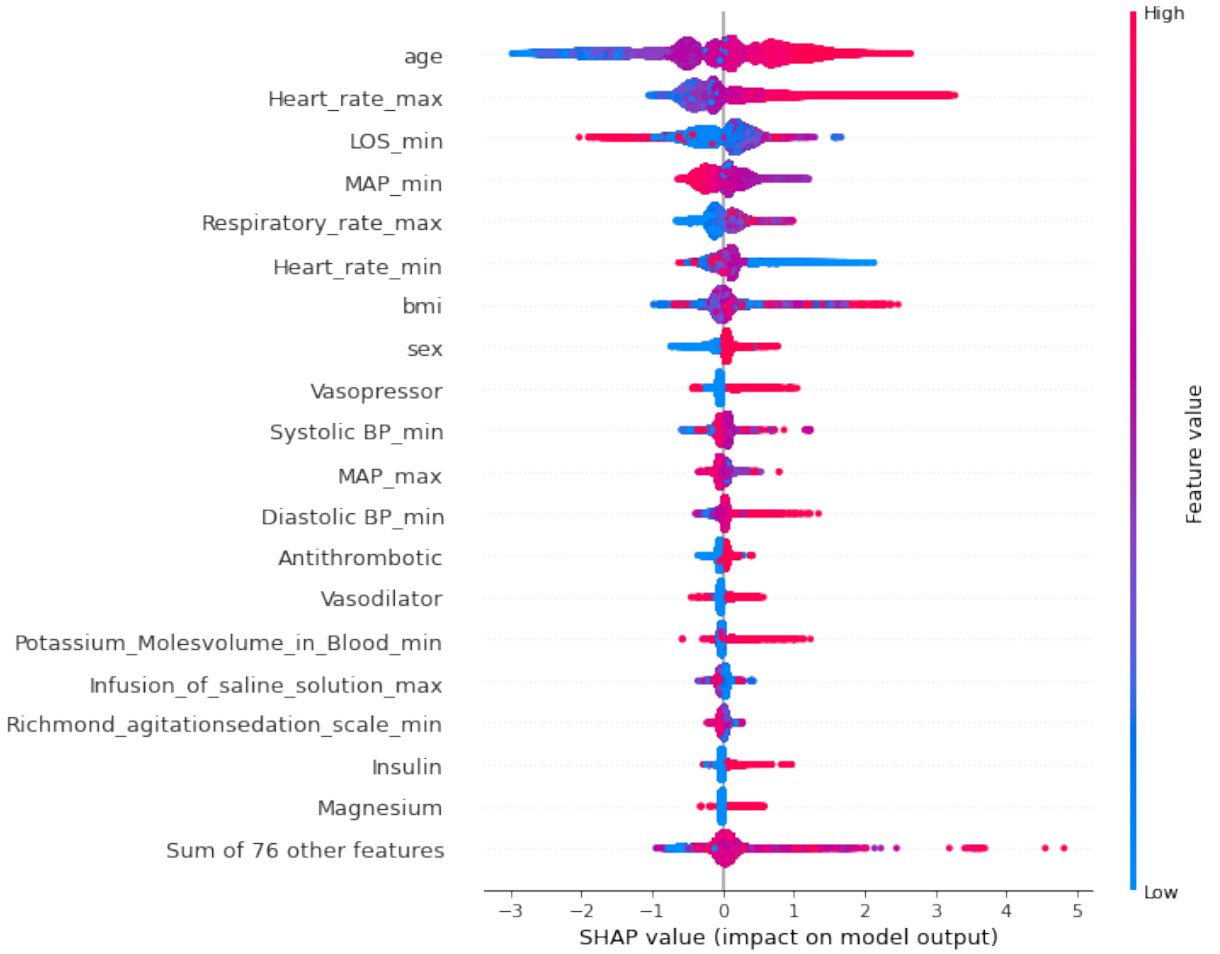

Figure 4.5: SHAP values of the XGBoost model. Variables are plotted in order of feature importance, red values mean that higher values are more important whereas blue values indicate that lower values are more important for a given variable. Only the top most important variables are shown, with the other variables summed together.

Fig. 4.5 is a beeswarm plot of the SHAP values corresponding to the XGBoost model. It shows

that advanced age and increased maximum heart rate are the highest contributors to the AF prediction which is consistent with previous literature. The third most significant variable is the length of stay at the measurement time point where lower values mean the patient is more likely to have AF. This is consistent with Fig. 3.7 where the majority of patients developed AF soon after admission. Consistent with the literature, the use of vasopressors also increases the likelihood of AF and males are more likely to develop AF than females (males are labelled with 1 and females are labelled with 0).

The SHAP values can also help the user identify possible confounding. Antithrombotics such as anticoagulants are often used on patients that develop AF to prevent blood clots that lead to stroke. As antithrombotics are one of the most indicative variables for AF, according to the SHAP values, the machine learning model may be learning from the clinician actions. However, this would be difficult to prove as antithrombotics could be used for other reasons.

Previous literature suggests that AF may also occur due to a low potassium level and one of the treatments for AF is to increase the patients' potassium level. Fig. 4.5 shows that high levels of potassium are indicative of AF. However, this could be due to the clinicians increasing the patients' potassium levels due to their AF diagnosis. Again, the machine learning model is learning from the clinicians. However, instances, where the clinician administers potassium to the patient, are recorded as a pharmacological variable named "potassium". This pharmacological variable is not listed as among the topmost important variables according to the SHAP values. If the machine learning model was truly learning from the clinicians' use of potassium, then the "potassium" variable should be one of the most important. This suggests there may not be confounding due to potassium levels.

4.3 Predicting Other Adverse Events

One of the main aims of this project was to make sure that Adaptable Forecasting Framework in Real-tiMe (AFFIRM)was easily generalised to other prediction tasks. AFFIRMwas used to predict tachycardia and circulatory failure.

AF is associated with tachycardia (increased heart rate). Predicting tachycardia may be a way to indicate the possibility of AF. Tachycardia is much more common and can be labelled after data collection with less human error compared to AF. The AFFIRM was therefore adjusted to

predict tachycardia defined as a heart rate above 100 beats per minute. The results are shown in Fig. 4.6 where an 8-hour prediction window was used.

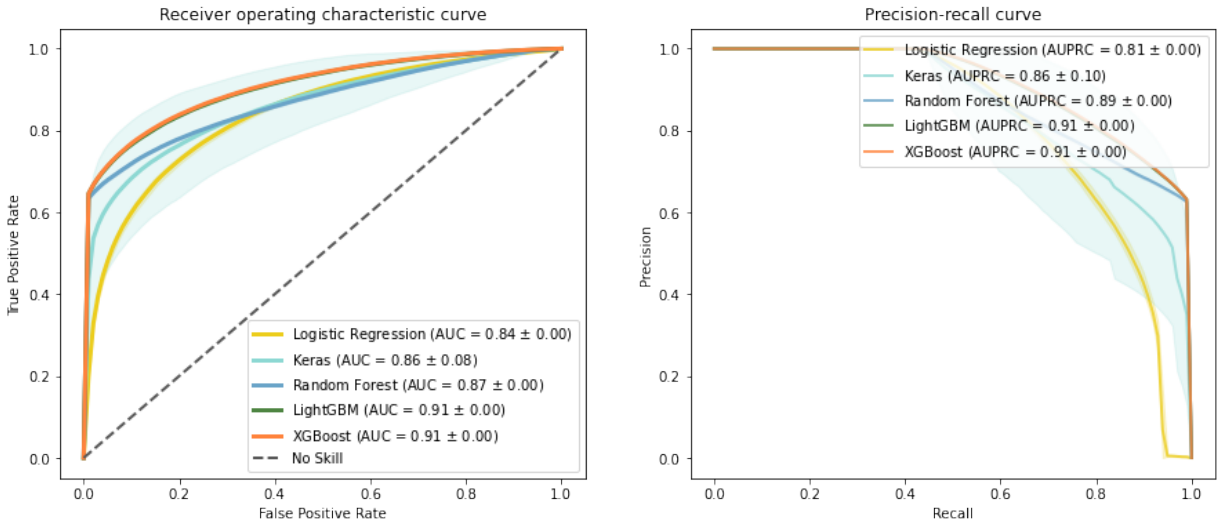

Figure 4.6: Five binary classification results using default parameters for predicting tachycardia (¿ 100 bpm) within six hours.

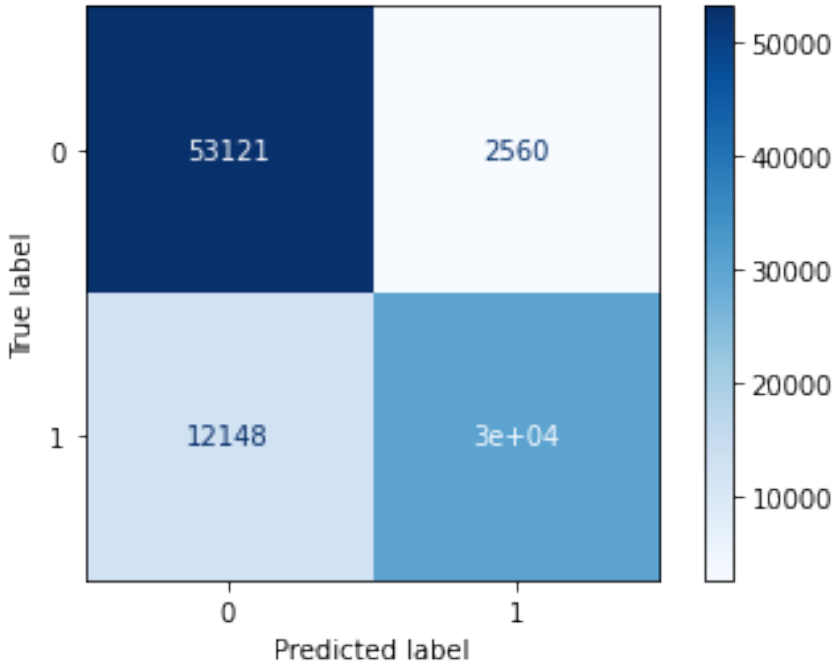

Figure 4.7: XGBoost confusion matrix for predicting tachycardia.

Fig. 4.6 shows that AFFIRM is far more effective at predicting tachycardia compared to AF, AUPRC = 0.91 compared to 0.59. This is likely due to the balanced dataset, where there is

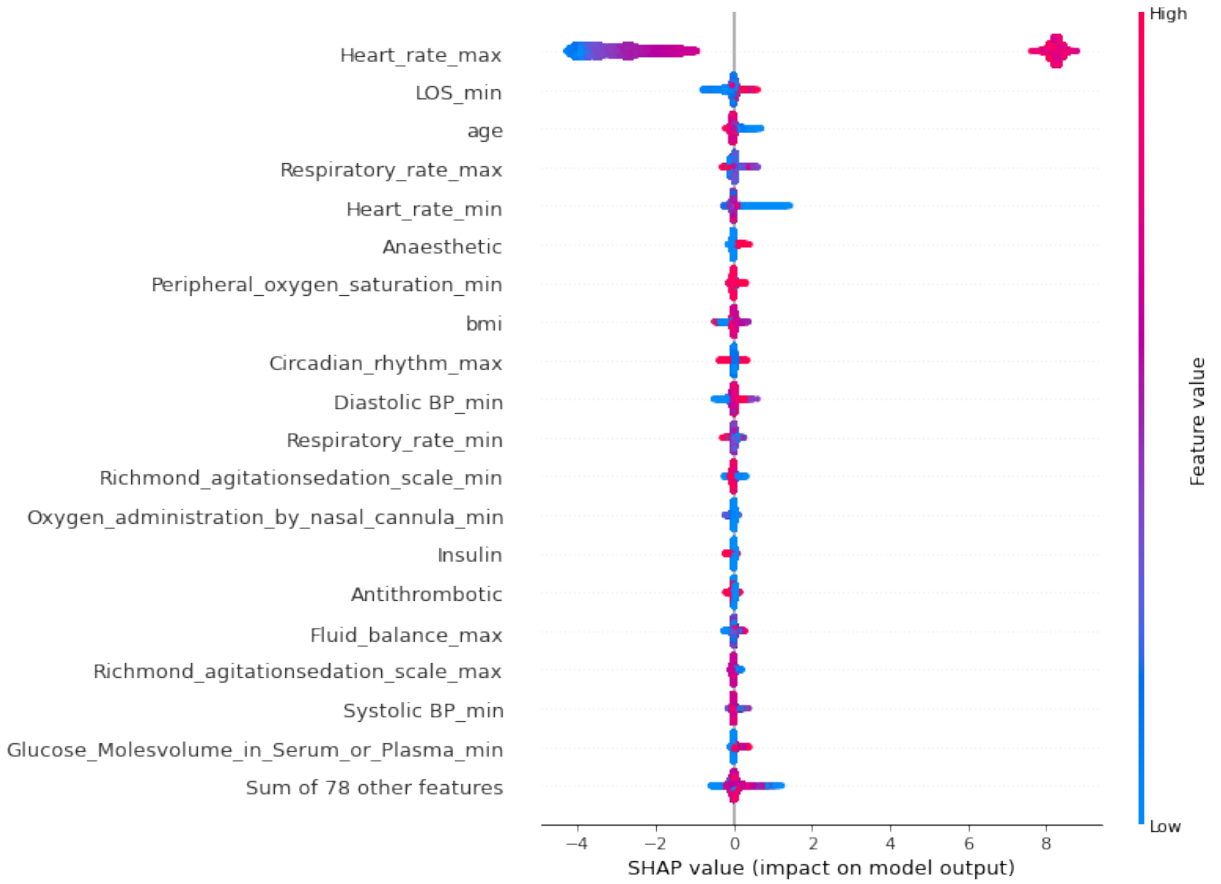

Figure 4.8: SHAP values from XGBoost, showing the most important features for predicting tachycardia.

approximately the same number of cases with tachycardia as without (46% case time points). The very high performance of AFFIRM on tachycardia is also likely since tachycardia is defined by one variable that already exists in the features. Fig. 4.8 shows that heart rate is by far the most dominant feature for prediction.

In Hyland et al.'s paper, circulatory failure was defined as a lactate level less than 2 and mean arterial pressure (MAP) greater than 65. They also predicted 8 hours into the future. These parameters can be entered into AFFIRM and the results are shown below.

Fig. 4.9 show that AFFIRM works better for predicting circulatory failure than AF with an AUPRC of around 0.60 (without hyperparameter optimisation). These results are similar to the Hyland et al. paper where they achieved an AUPRC of 0.63. Understandably, AFFIRM performs less well compared to Hyland et al.'s model because it has not been optimised for that purpose. Hyland et al also used a more sophisticated feature set. Furthermore, Hyland et al. made predictions every five minutes which is a more complex task.

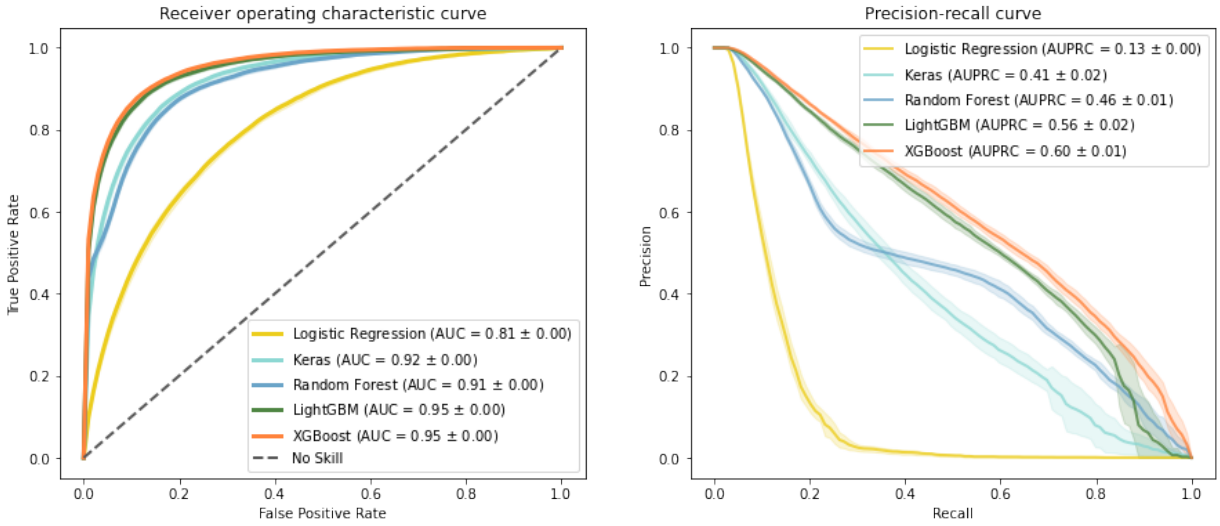

Figure 4.9: Five binary classification results using default parameters for predicting circulatory failure as defined by Hyland et al. within eight hours.

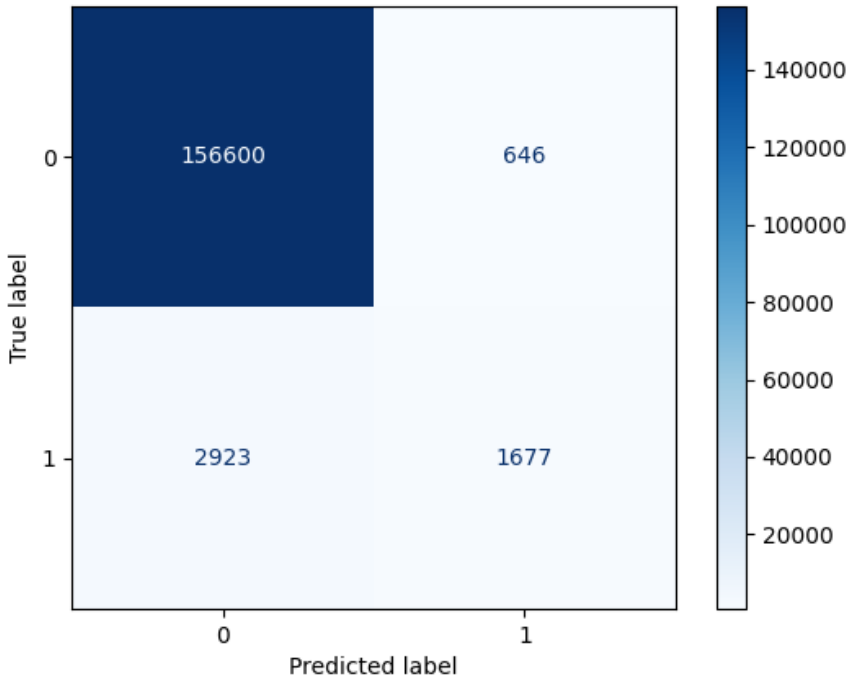

Figure 4.10: XGBoost confusion matrix for predicting circulatory failure.

Hyland et al. calculated SHAP values for their prediction of circulatory failure. They found that the top predictive features included maximal lactate, minimal MAP, age and time since admission. Fig. 4.11 shows that AFFIRM also confirms the importance of these features with the minimum MAP being the most important, the current length of stay (LOS) as the third most important and maximal lactate at fifth. The agreement in SHAP rankings suggests that AFFIRM

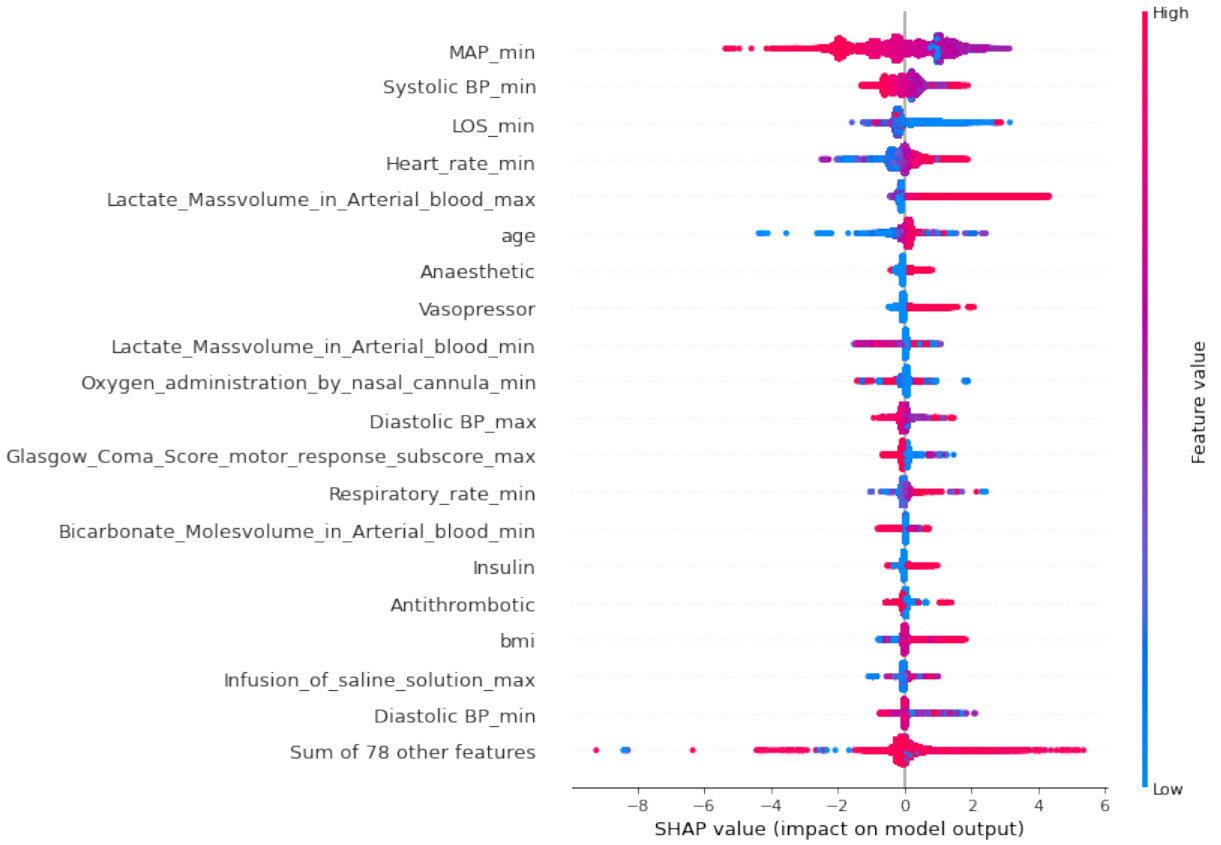

Figure 4.11: SHAP values from XGBoost, showing the most important features for predicting circulatory failure.

is working correctly.

Chapter 5

Discussion

This project represents the first time attempt at atrial fibrillation (AF) prediction in the Intensive Care Unit (ICU) within hours using only EHR data. New-onset AF in the ICU affects 5-46% of patients and is associated with a longer length of stay and a greater chance of mortality. Early prediction of AF using machine learning (ML) may be able to guide AF management in the ICU.

In this project, a framework called Adaptable Forecasting Framework in Real-tiMe (AFFIRM) was created. Initially, the framework was built for AF prediction, but it is now generalised to predict any adverse event. AFFIRM consists of three main stages preprocessing raw High time-Resolution ICU Dataset (HiRID) data, preparing the data for classification and evaluating ML models. Several machine learning algorithms are included in the AFFIRM framework including the Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), logistic regression, random forest and a deep learning model made using Keras. The user can also specify their machine learning models.

AF prediction was evaluated using the five binary classifiers in AFFIRM for several prediction windows. A prediction window is how far in advance the user wants to predict AF. Four prediction windows were tested 4, 6, 8 and 10. Out of all the baseline binary classifiers, the XGBoost model had the best performance for all prediction windows. The 4 and 6 hour prediction time windows produced the best results with an Area Under the Precision-Recall Curve (AUPRC) = 0.52 ± 0.00. The six-hour time window was chosen as it is beneficial to clinicians to have knowledge earlier in advance.

Hyperparameter optimisation was performed on the XGBoost model using a random search for the best value parameters. XGBoost has the advantage over other binary classifiers of having a

relatively fast computation time and being transparent. It is simple to calculate SHapley Additive exPlanations (SHAP) values from the XGBoost predictions and thereby determine which patient variables were the most important during classification. Following the literature, these were advanced age, increased heart rate and decreased mean arterial pressure.

Classification of AF is a particularly challenging task due to several factors. Unlike other adverse events that have been predicted in the past, there is no explanation for AF. Also, AF can only be diagnosed using an ECG. Therefore, unlike other events, the data cannot be annotated after data collection. Therefore, accurate AF data is reliant on the clinician, accurately inputting AF episodes into the database. However, AF can be transient and can look similar to other arrhythmias therefore it could go undiagnosed or be misdiagnosed. Furthermore, AF usually occurs over several hours, however, the HiRID system only allows for data collection at specified time points rather than a range of time. Likely, clinicians may only document one time point as AF within a whole time range. This was combated by grouping time point data into hours and also filling in time points that were likely to contain AF due to the increased heart rate. Furthermore, there is a very large data imbalance between the time points containing AF and the time points not containing AF, around 5.47%.

These difficulties make ML-enabled AF prediction difficult and which is likely why there has been no previous work done. Despite these difficulties, the final machine learning model achieved an AUPRC of 0.59 ± 0.01 after 10-fold cross-validation. This AUPRC is comparable to previous literature where Hyland et al. achieved AUPRC = 0.63 for circulatory failure and Tomašev et al. achieved AUPRC = 0.297 for acute kidney injury.

To test its adaptability, AFFIRM was tested on two other adverse events with the same variables. AFFIRM predicted AUPRC of 0.91 ± 0.00 and 0.60 ± 0.01 for tachycardia and circulatory failure respectively with no hyperparameter optimisation.

AFFIRM was thought to predict tachycardia better than AF because of the balanced dataset, however, circulatory failure is even more imbalanced than AF. Circulatory failure makes 2.8% of case time points whereas AF has 5.5% of case time points. This suggests either that circulatory failure is easier to predict with the variables used i.e. those variables are more predictive for circulatory failure or there are many time points unlabelled AF time points due to human error that cannot be labelled post data collection.

The pharmacological data used in this project was represented very simply. The pharmacological variables were grouped by drug class and its presence indicated with a binary flag. The drug classes used were very generic e.g. vasodilators described a whole range of drugs designed to decrease blood pressure. The vasodilators used in this project could have been subclassed into drugs such as beta-blockers, calcium channel blockers or alpha-blockers. Despite having the same main function, all of these drugs may have different effects on AF. Furthermore, continuous dosages and the size of dosage of a particular drug can result in many types of outcomes that are not captured by the use of simple binary indicators. More work is required to accurately describe the effect of the drugs.

AFFIRM only works with HiRID data which is a single centred database. To be clinically useful AFFIRM must be able to generalise to different patient populations and different hospitals where there are different protocols for measurement e.g. frequency of measurement. Therefore, AFFIRM needs to be evaluated on other databases such as the Medical Information Mart for Intensive Care (MIMIC) database.

AF prediction can also be improved by examining multi-modal data. There is significant research in AF prediction using electrocardiogram (ECG), so AF prediction could be improved by combining information from ECG and Electronic Health Record (EHR). However, collecting ECG and EHR at the same time for the same patient is challenging as they are collected using different computerised systems. Information leveraged from clinical notes using natural language processing could improve AF prediction as clinical notes can provide a more detailed account of a patients' stay as well as including patient history. The HiRID database had the clinical notes removed due to anonymisation issues, but there is a potential that the authors will anonymise the notes and release them for general use.

Further work could also include stratifying the patients by diagnostic group e.g. whether the patient had cardiac surgery. Post-cardiac surgery patients who develop new-onset AF are known to exhibit unique characteristics compared to other patients in the ICU. Stratification was not carried out in this project because stratifying the patients would reduce the already small number of AF patients. If more data is collected, it would be easier to stratify patients with enough data to train and test each cohort.

Bibliography

[1] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

[2] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[3] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[4] Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, Gilles Clermont, Mihaela van der Schaar, Ari Ercole, Armand R. J. Girbes, and Paul W. G. Elbers. Sharing ICU patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The amsterdam university medical centers database (AmsterdamUMCdb) example$*$. $CriticalCareMedicine, 49(6) : e563 - -e577, February 2021$.

[5] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1), September 2018.

[6] Rahul C Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[7] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[8] Stanley Nattel. New ideas about atrial fibrillation 50 years on. *Nature*, 415:219–226, 01 2002.

[9] Emelia J Benjamin, Philip A Wolf, Ralph B D'Agostino, Halit Silbershatz, William B Kannel, and Daniel Levy. Impact of atrial fibrillation on the risk of death: the framingham heart study. *Circulation*, 98(10):946–952, 1998.

[10] Huey-Juan Lin, Philip A. Wolf, Margaret Kelly-Hayes, Alexa S. Beiser, Carlos S. Kase, Emelia J. Benjamin, and Ralph B. D'Agostino. Stroke severity in atrial fibrillation. *Stroke*, 27(10):1760–1764, October 1996.

[11] Paulus Kirchhof, Stefano Benussi, Dipak Kotecha, Anders Ahlsson, Dan Atar, Barbara Casadei, Manuel Castella, Hans-Christoph Diener, Hein Heidbuchel, Jeroen Hendriks, et al. 2016 esc guidelines for the management of atrial fibrillation developed in collaboration with eacts. *Kardiologia Polska (Polish Heart Journal)*, 74(12):1359–1469, 2016.

[12] Salam Salman, Abubakr Bajwa, Ognjen Gajic, and Bekele Afessa. Review of a large clinical series: Paroxysmal atrial fibrillation in critically ill patients with sepsis. *Journal of intensive care medicine*, 23(3):178–183, 2008.

[13] Philippe Seguin and Yoann Launey. Atrial fibrillation is not just an artefact in the icu. *Critical Care*, 14(4):1–2, 2010.

[14] Mattia Arrigo, Natalie Jaeger, Burkhardt Seifert, Donat R Spahn, Dominique Bettex, and Alain Rudiger. Disappointing success of electrical cardioversion for new-onset atrial fibrillation in cardiosurgical icu patients. *Critical care medicine*, 43(11):2354–2359, 2015.

[15] Allan J Walkey, Emily K Quinn, Michael R Winter, David D McManus, and Emelia J Benjamin. Practice patterns and outcomes associated with use of anticoagulation among patients with atrial fibrillation during sepsis. *JAMA cardiology*, 1(6):682–690, 2016.

[16] Frank Bogun, Daejoon Anh, Gautham Kalahasty, Erik Wissner, Chadi Bou Serhal, Rabih Bazzi, W Douglas Weaver, and Claudio Schuger. Misdiagnosis of atrial fibrillation and its clinical consequences. *The American journal of medicine*, 117(9):636–642, 2004.

[17] T Rizos, A Wagner, E Jenetzky, PA Ringleb, R Becker, W Hacke, and R Veltkamp. Paroxysmal atrial fibrillation is more prevalent than persistent atrial fibrillation in acute stroke and transient ischemic attack patients. *Cerebrovascular Diseases*, 32(3):276–282, 2011.

[18] Gregory YH Lip and Hung-Fat Tse. Management of atrial fibrillation. *The Lancet*, 370(9587):604–618, 2007.

[19] Peter MC Klein Klouwenberg, Jos F Frencken, Sanne Kuipers, David SY Ong, Linda M Peelen, Lonneke A van Vught, Marcus J Schultz, Tom van der Poll, Marc J Bonten, and Olaf L Cremer. Incidence, predictors, and outcomes of new-onset atrial fibrillation in critically ill patients with sepsis. a cohort study. *American journal of respiratory and critical care medicine*, 195(2):205–211, 2017.

[20] Sangita-Ann Christian, Christa Schorr, Lynn Ferchau, Maria E Jarbrink, Joseph E Parrillo, and David R Gerber. Clinical characteristics and outcomes of septic patients with new-onset atrial fibrillation. *Journal of critical care*, 23(4):532–536, 2008.

[21] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 01 2019.

[22] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019.

[23] Mintu P Turakhia, Manisha Desai, Haley Hedlin, Amol Rajmane, Nisha Talati, Todd Ferris, Sumbul Desai, Divya Nag, Mithun Patel, Peter Kowey, et al. Rationale and design of a large-

scale, app-based study to identify cardiac arrhythmias using a smartwatch: The apple heart study. *American heart journal*, 207:66–75, 2019.

[24] Walid Saliba, Naomi Gronich, Ofra Barnett-Griness, and Gad Rennert. Usefulness of chads2 and cha2ds2-vasc scores in the prediction of new-onset atrial fibrillation: a population-based study. *The American journal of medicine*, 129(8):843–849, 2016.

[25] Ingrid E Christophersen, Xiaoyan Yin, Martin G Larson, Steven A Lubitz, Jared W Magnani, David D McManus, Patrick T Ellinor, and Emelia J Benjamin. A comparison of the charge–af and the cha2ds2-vasc risk scores for prediction of atrial fibrillation in the framingham heart study. *American heart journal*, 178:45–54, 2016.

[26] Alvaro Alonso, Bouwe P Krijthe, Thor Aspelund, Katherine A Stepas, Michael J Pencina, Carlee B Moser, Moritz F Sinner, Nona Sotoodehnia, João D Fontes, A Cecile JW Janssens, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the charge-af consortium. *Journal of the American Heart Association*, 2(2):e000102, 2013.

[27] Nathan R Hill, Daniel Ayoubkhani, Phil McEwan, Daniel M Sugrue, Usman Farooqui, Steven Lister, Matthew Lumley, Ameet Bakhai, Alexander T Cohen, Mark O'Neill, et al. Predicting atrial fibrillation in primary care using machine learning. *PloS one*, 14(11):e0224582, 2019.

[28] Premanand Tiwari, Kathryn L. Colborn, Derek E. Smith, Fuyong Xing, Debashis Ghosh, and Michael A. Rosenberg. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. *JAMA Network Open*, 3(1):e1919396, January 2020.

[29] Sean McMillan, Ilan Rubinfeld, and Zeeshan Syed. Predicting atrial fibrillation from intensive care unit numeric data. In *2012 Computing in Cardiology*, pages 213–216. IEEE, 2012.

[30] Philippe Seguin, Bruno Laviolle, Axelle Maurice, Christophe Leclercq, and Yannick Mallédant. Atrial fibrillation in trauma patients requiring intensive care. *Intensive care medicine*, 32(3):398–404, 2006.

[31] Allan J Walkey, Melissa A Greiner, Susan R Heckbert, Paul N Jensen, Jonathan P Piccini, Moritz F Sinner, Lesley H Curtis, and Emelia J Benjamin. Atrial fibrillation among medicare beneficiaries hospitalized with sepsis: incidence and risk factors. *American heart journal*, 165(6):949–955, 2013.

[32] Michael J Mihm, Fushun Yu, Cynthia A Carnes, Peter J Reiser, Patrick M McCarthy, David R Van Wagoner, and John Anthony Bauer. Impaired myofibrillar energetics and oxidative injury during human atrial fibrillation. *Circulation*, 104(2):174–180, 2001.

[33] Allan J Walkey, Renda Soylemez Wiener, Joanna M Ghobrial, Lesley H Curtis, and Emelia J Benjamin. Incident stroke and mortality associated with new-onset atrial fibrillation in patients hospitalized with severe sepsis. *Jama*, 306(20):2248–2254, 2011.

[34] Ciara M Shaver, Wei Chen, David R Janz, Addison K May, Dawood Darbar, Gordon R Bernard, Julie A Bastarache, and Lorraine B Ware. Atrial fibrillation is an independent predictor of mortality in critically ill patients. *Critical care medicine*, 43(10):2104, 2015.

[35] Nicholas A Bosch, Jonathan Cimini, and Allan J Walkey. Atrial fibrillation in the icu. *Chest*, 154(6):1424–1434, 2018.

[36] J Larry Jameson and Dan L Longo. Precision medicine—personalized, problematic, and promising. *Obstetrical & gynecological survey*, 70(10):612–614, 2015.

[37] Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1):1–10, 2021.

[38] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[39] Daniel Jarrett, Jinsung Yoon, Ioana Bica, Zhaozhi Qian, Ari Ercole, and Mihaela van der Schaar. Clairvoyance: A pipeline toolkit for medical time series. In *International Conference on Learning Representations*, 2020.

[40] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.