Explainable Intensive Care Discharge Predictions

Ryan Porter
Supervised by Raul Santos-Rodriguez, Jeff Clark and Alex Hepburn

April 30, 2024

Project Report submitted in support of the degree of Master of Engineering

Department of Engineering Mathematics, University of Bristol

Abstract

Predictions of intensive care discharge outcomes using machine learning methods is an increasingly popular area of research. Methods in the literature surrounding this topic leverage information from the most recent physiological measurements to predict readiness for discharge at the current point in time for intensive care patients. This study explores the use of features based on temporal gradients between the initial and current measurement values to improve readiness for discharge predictions. This was demonstrated on a cohort of intensive care patients from the Medical Information Mart for Intensive Care database, using explainable boosting machines and random forests for binary classification. The addition of these gradient features did not improve overall model performance. The methodology was expanded to predict readiness for discharge for discrete time steps into the future. This proved achievable at a granularity of 24 hours for up to 4 days into the future. The potential for the explainable boosting machine models to be trustworthy for deployment in clinical context was demonstrated with global and local prediction examples. However, testing and validation with expertise is required to evaluate the clinical meaning of model interpretations. This study used only a single data source, so generalisability of the methods is presented as a key area for further work.

Ethics Statement: This project fits within the scope of ethics pre-approval process, as reviewed by my supervisor Raul Santos-Rodriguez and approved by the faculty ethics committee as application 16028.

Contents

1 Introduction

Intensive Care Units (ICUs) provide care to hospital patients who depend on it for survival in the short term. However, over the course of a prolonged ICU admission the risk of deterioration increases, as ICU admissions entail high prevalence of invasive procedures and devices which leave patients susceptible to infections [3]. Prolonged periods in intensive care can also cause detrimental psychological effects [8]. Alongside this, hospitals are typically subject to resource constraints and high demands which can limit the flow of patients from admission to discharge. Therefore, it is important to prevent patients from staying in intensive care longer than necessary, whilst also allowing enough time in the ICU as to give the best chance of recovery before discharge.

One of the many roles of doctors in ICU wards is to assess whether a patient is ready for discharge throughout their stay. This decision is based on the opinions of the doctors, which may be influenced by trusted scoring systems and situational context [39]. Increasingly, there is a wealth of data collected from patients in the form of electronic health records (EHR). This data is particularly abundant for intensive care admissions, since patients in intensive care are routinely monitored by nurses as well as with automatic equipment. Doctors usually have a limited daily time window to make discharge decisions. During this time window, it is not humanly possible to take into account the vast majority of the data that is recorded from each patient. A tool that could digest this data and use it to offer decision support to ICU doctors offers potential to overcome this issue [21]. Such a tool could enable doctors to prioritise assessing the patients that are most likely to be ready for discharge during the decision time window, likely increasing the frequency of successful discharges. However, any modelling on EHR data must be explainable, such that the processes and outputs can be interpreted, validated and ultimately trusted in clinical context.

The increasing prevalence of EHR from ICUs has opened opportunities for model-led standardisation of the discharge decision process, as well as risk prediction systems. Many studies have explored using machine learning (ML) models with ICU data to predict discharge outcomes using a variety of datasets, such as the Medical Information Mart for Intensive Care (MIMIC) datasets [17]. Some of these studies are reviewed in Section 2. To predict discharge outcomes, the methods in the literature typically use current information, such as measurements from the last few hours of ICU admission [10,21,29]. However, measurements from earlier portions of ICU admission and their relationship to the current measurements may contain predictive information that can be exploited to improve discharge outcome predictions. Furthermore, methods in the literature are largely confined to predicting outcomes after discharge at the current point in time. It may be possible to predict outcomes for discharge at points in the near future, giving an indication of when patients will be ready for discharge. This could further benefit the scheduling of assessment from ICU doctors as well as potentially enabling some degree of ward planning.

The goal of this work is to build on the findings of the relevant literature to further explore the potential for using standard ICU data to support doctors with discharge decisions. This is achieved through the following objectives:

- Implement a highly explainable and editable binary classification model that predicts whether a patient in intensive care is ready for discharge at the current point in time.
- Design novel model features which exploit the predictive power of data from earlier times in admission by capturing the trajectories of recovery or deterioration over the course of admission.
- Assess the impact of the novel features on model performance by comparing with an an established

readiness-for-discharge prediction method.

- Explore the ability to predict when patients will be ready for discharge by extending the modelling method to predict whether patients will be ready for discharge at a set of discrete time points in the future.
- Demonstrate the suitability of the modelling methods used to the clinical setting. The deliverable is a model tool that, with further development and testing, could be implemented into bedside systems on ICUs to aid doctors with the discharge decision making process.

2 Literature Review

In this section, the key concepts and challenges pertaining to intensive care discharge and machine learning in the clinical context are discussed through a review of the related literature. This starts with some of the more simplistic and established methods before moving on to reviewing some notable advancements in data-driven tools that utilise machine learning. Two salient considerations that arise with the progression towards data-driven methods are the nature of the data itself and the issue of trust between clinicians and machine learning models.

2.1 Background

The Glasgow Coma Scale (GCS) is a scoring system for 'bedside assessment of the depth and duration of impaired consciousness and coma' [33]. It is comprised of three components; eye opening, motor response and verbal response. Since its introduction in 1974, GCS has stood the test of time for its use in clinical practice as well as research [34]. GCS scores are particularly relevant to the assessment of patients in intensive care. While GCS is not a system designed to directly recommend discharge decisions in itself, it frequently features as a parameter or consideration in methods in the literature.

In 2003, a set of nurse-led discharge (NLD) criteria on which nurses could decide to discharge a patient from an acute ward was proposed [18]. This entailed a set of equally-weighted boolean test conditions on measurements from the last 4 hours of the ICU admission. The measurements used were a set of standardly recorded vital signs, routinely collected laboratory results and GCS scores. [18]. This criteria offered the potential to improve patient flow while circumventing the need to have a doctor approve every discharge decision on the ward. However, when tested on an ICU cohort from the MIMIC dataset these criteria were shown to be very conservative, achieving a true positive rate of only 1.1% [21]. Discharging patients without a doctor's opinion does necessitate a conservative criteria, since it bypasses validation by a doctor's heuristics and expertise. However, such a conservative model is unlikely to allow satisfactory patient flow to cope with resource constraints on most hospitals.

A more popular approach has been to use a patient's measurements to calculate scores that may aid a doctor's decisions about their discharge. The Stability and Workload Index for Transfer score (SWIFT) is a validated numerical tool with the ability to measure a patient's safe and suitable discharge from the ICU [9]. Similarly to the nurse-led criteria, SWIFT uses test conditions on the most recently recorded measurements including GCS score, but with the improvement of weighted criteria. The Early Warning Score (EWS), and several variations of it, can also be used to this effect by predicting risk of deterioration [28]. While the performance of these methods may be limited by their simplicity, both SWIFT and EWS have widespread use due to their ease of implementation and trustworthiness [39]. Better performance may be achievable by making use of the increasingly abundant data that is collected from ICUs.

Datasets derived from EHR are typically heterogeneous and very large. ML models offer great performance on data with these attributes and are a popular choice for predictive tasks on EHR data as a result. However, trust between the medical community and ML methods must be established if they are to be adopted at a large scale. This brings about the topic of explainability and interpretability of ML methods.

2.2 Explainability and Interpretability

In the context of machine learning, explainability refers to the ability to explain the decision making process behind the output of a model in such a way that it is understandable to its human end user. Interpretability pertains to the ability of the inner mechanics of a model to be understood, such as the relationship between input, features, associated weights and output.

Many studies have explored using ML models for tasks such as regression, classification and clustering to make predictions on health data [32]. These models have varying degrees of explainability and interpretability. For some tasks, the highest accuracy may be achieved using models that come under the 'black box' category, meaning they are not interpretable or explainable. However, clinical understanding of model predictions is more important than accuracy alone, as model misclassification may lead to fatal consequences in the intensive care context. Model predictions that are explainable can be acted upon with more confidence after being approved in conjunction with clinical expertise. In one particular effort to increase trust between clinicians and ML models, explainable ML methods have been used to cluster hospital patients on their vital signs data and the characteristics of these clusters were mirrored by clinicians who independently evaluated samples from each cluster [38].

ML algorithms previously considered as black box methods are gaining trust due to recent advancements in explainability methods [1]. One such explainability method is Shapley Additive Explanations (SHAP), which utilises concepts from game theory to analyse the importance of features in a prediction task [20]. SHAP is demonstrable on EHR for tasks such as mortality prediction [31]. However, it is highlighted that while SHAP values give insight into how models consider input, they do not imply causality [31]. In other words, it is erroneous to equate high model performance to learned medical understanding. This is an inescapable limitation to modelling with data and further accentuates the importance of interpretability as well as explainability.

Another advancement into explainable AI is the explainable boosting machine (EBM) method [24]. Rather than retroactively explaining a black box model with something like SHAP values, EBM are constructed in such a way that they are inherently explainable and interpretable. For this reason it is known as a glass box model. In health data contexts, they have been demonstrated to have comparable performance to that of state of the art black box models, such as RF and XGboost [14,30]. An important advantage of EBM other the other methods is that it is highly editable. This means that its decision making process can be altered directly for each feature that it learns on. The editing process for EBM is visual and intuitive, making it ideal for review and improvement via clinical expertise.

With an understanding of the importance of explainability and interpretability for model trust, ML decision support tools can be evaluated for use in the intensive care context.

2.3 Machine Learning Decision Support Tools

This discussion of ML discharge decision support tools begins with [21], a popular paper in the literature with a comparatively high number of citations. Using the same standardly collected physiological measurements as the NLD criteria, a decision support model was created using a Random Forest (RF) classifier [21]. This was achieved by retroactively labelling patients as ready for discharge (RFD) or not (NRFD) at the time of their callout - the time at which a doctor declares a patient is RFD. This labelled data, which consisted of features based on physiological measurements over the final 4 hours of each ICU admission, was then used to train a binary classifier via supervised learning. The trained classifier was used to classify each patient as RFD of NRFD at the point in time of the end of their 4 hour data window.

In [21], readiness for discharge was operationalised using the assumption that if a patient left the hospital after discharge from the ICU without readmission to the ICU or in-hospital death they are considered truly RFD at the time that a doctor declared they were. Conversely, it was assumed that if a patient was readmitted to the ICU or died within the same hospital admission they are considered NRFD at the time that a doctor declared they were ready. This binary mapping of outcomes for ICU patients does not fit all possible scenarios. For example, patients who died during their ICU admission may not necessarily have had have a point in time at which they could have been successfully discharged, so it does not make sense to predict when they would be RFD. These patients were omitted from the study cohort. It is also possible for patients to have multiple hospital admissions, with the possibility for ICU admissions within each. Only the first ICU admission of each patient was included in the cohort, so that readmission could be used to define a negative outcome [21].

On a combined cohort from MIMIC and another source, this model performed better than the NLD criteria at classifying patients as RFD or not, according to several performance metrics [21]. The addition of some admission-level features such as age and length of stay further improved performance. This model was not explained using SHAP values. Instead, global feature importances were calculated using permutation feature importance [4]. One drawback of this model is that it had a relatively high rate of false positives [21]. In the clinical setting, this translates to premature flagging of RFD status for patients in the ICU, which may cause a doctor to prioritise the wrong patients when making assessments and decisions on discharge. Over time, consistent premature flagging may lead to model distrust. This model considered only the last 4 hours of recorded values, meaning that it did not have the ability to capture a relative trajectory of recovery or deterioration for each patient over their ICU admission. Another limitation of this model is that the scope of predictions is restricted to the current point in time. In this sense it answers the question of whether a patient is RFD now, rather than when will they be RFD.

Predicting risk of ICU readmission or mortality is a slightly different problem to classifying patients as ready for discharge or not, but the motivation, end use and results are comparable. The combination of XGBoost with SHAP for predicting in-ICU mortality has been demonstrated on ICU data from the MIMIC dataset [10]. This was achieved by classifying patients as having predicted in-ICU mortality using the last 24 hours of data as input. The SHAP explanations of the model were used to suggest which features are most important, and where to set threshold values for these features at which healthcare personnel should be alerted. Unlike [21], this study benefited from subtyping patients by age group. More features were modelled on each variable than in [21], including mean and standard deviation in addition to minimum and maximum and for vital measurements. The performance of this model was high, achieving an average accuracy of 0.916 across the age groups. However, predictions from this model were limited in that their outcomes were not time-bound. Furthermore, data from earlier on in the ICU admissions was not exploited for its predictive power.

In [29], a gradient-boosting machine model was used to predict risk of ICU readmission. The model was more accurate than the SWIFT and the Modified Early Warning Score. The feature selection was more expansive than in [21], including more demographic variables, medications administered during the ICU admission, nursing documentation, and International Classification of Diseases (ICD) codes from prior admissions at the same hospital. Features which captured temporal trends for vital measurements, such as 24-hour slope (gradient) of systolic blood pressure readings, are included. Some of these temporal features had greater feature importance than their individual, non-temporal equivalents. This suggests that features which represent time-based trajectories of measurements can lead to better performing models. Similarly to [10], these gradient features are limited in that they only utilise data from the last 24 hours.

More extensive patient subtyping was explored with the use of clustering in [2]. This involved clustering patients by similarity of trajectories of vital measurements over their ICU admissions. Dynamic time warping (DTW) was used to give similarities as pairwise distances between vital measurement trajectories of each patient in the cohort. DTW is suitable for quantifying similarities between trends in ICU time-series data because it is robust to irregular sampling frequencies and variable durations of admissions. However, DTW is not ideal in deployment as these pairwise distances must be recalculated for each new patient. This effectively makes the clustering method transductive, requiring frequent computation that is likely to be prohibitively expensive. The clusters were demonstrated to have captured some clinical meaning, as certain ICD codes were more prominent in some clusters than others. Following a similar approach, [38] demonstrates that patients admitted with the same condition may have different care needs, as indicated by similar frequencies of the same ICD codes between different clusters.

Continuing with [2], EBM were used for ICU mortality prediction on the MIMIC dataset, along with patient subtyping. When predicting in-hospital mortality using EBM, training models for each individual cluster yielded higher performance over training a single model on all of the unclustered data. This suggests that subtyping is one potential avenue that should be further explored in this field.

Another notable result from [2] was that classifiers trained on data from only the first 4 hours of ICU admissions outperformed classifiers that used data from up to the first week, according to several performance metrics. This suggests that data from the initial period of an ICU admission is useful for predicting eventual outcome. Classifiers that were trained on data including the end of stay period had a similar performance to those trained on the initial 4 hours. It is to be expected that data towards the end of the stay strongly indicates the outcome. However, these findings suggests an unexpected phenomenon; the predictive power of ICU data may decrease after the initial period after admission. Perhaps the impact of interventions and treatments that patients undergo while in the ICU is obscuring their condition by 'artificially' improving their vital readings. Further research is needed in order to determine this.

In another study into predicting ICU readmission risk with various ML methods, features were engineered by calculating the squares for variables that show a clear parabolic relationship to the outcome [35]. However, the value of the additional squared features was found to be nonsignificant in this model. This study also used features indicating whether or not a specific variable was measured and how often this variable was measured. This type of feature could be considered as a meta-feature, as it describes the level of missingness of the other features. These features were shown to have predictive power. This example brings up another important consideration of model choices in clinical context. That is, whether or not to allow for informative missingness when training an ML model.

2.4 Informative Missingness

Missing data can be informative in EHRs as the presence or lack of measurement readings may be indicative of a patient's wellbeing. This is because the patterns of missingness may reflect a nurse/doctor's assessment of whether a measurement is needed, which in itself could act as an informative feature. Due to this, missing indicator variables may be useful for improving predictions from modelling methods relating to EHRs. However, it is paramount to consider the deployment of the model in practice and whether missing data mechanisms are transportable to the setting of future application [11]. For example, where a nurse/doctor is currently required to decide to take a patient's measurement, the same measurement may be completely automated in the near-future. Missingness in this measurement would no longer capture the same predictive patterns, as the only missing values would be due to technical faults, rather than clinical decisions. Furthermore, knowledge of the predictive value of missingness may effect a doctor or nurse's decision to take a measurement, subconsciously or otherwise. This concept is sometimes referred to as 'the curse of knowing' [11].

2.5 Summary

The problem of when to discharge patients from intensive care can be approached from several angles, each with varying operationalisations and underlying assumptions. The work in this study is primarily predicated upon the problem definition and solution framework from [21] as it provides a reproducible framework using a readily available dataset, and has room for improvements to model performance and for adaptions to predicting RFD into the future.

In this study, the modelling methods are adapted to use features based on measurement gradients over the entirety of ICU admissions. These are hypothesised to improve performance, as suggested by the findings when using 24-hour gradients in [29]. To predict readiness for discharge several days into the future, a set of similar models is developed and assessed. For classification, EBMs are used for their high explainability, interpretability, editability and performance [24].

3 Data and Features

In this section, the processes from data acquisition through to feature construction are explained. This entails some data exploration and preprocessing that is based on the methods from [21], with some alterations and improvements.

Due to its accessibility and ease of comparison with the literature, the MIMIC database is used for this study. The implementation of this work is on a local installation of the dataset using Python 3.11, the pandas package to handle table operations, the matplotlib package for plotting, and scikit learn and interpretML for the methods described in Section 4. For full details of implementation see the repository for this project: `https://github.com/RyanPort/Explainable-Intensive-Care-Discharge-Prediction`. Some of the methods in this code implementation are adaptations from code in the repository associated with [21]: `https://github.com/UHBristolDataScience/towards-decision-support-icu-discharge`.

3.1 Dataset Description

The MIMIC database contains deidentified health data for patients at the Beth Israel Deaconess Medical Center between 2001 and 2012 [17]. Particularly of interest for this study are the bedside vital measurements, lab results and demographic data for patients in the Medical Intensive Care Unit (MICU) and Surgical Intensive Care Unit (SICU). Lab results such as blood tests typically occur with frequency in the order of days, whereas vital measurements such as temperature readings are recorded approximately hourly, as seen in Figure 1. It is important to note that these frequencies of the measurements are not exact as vital measurements are not all necessarily taken regularly or at all in this dataset. Encoded in these measurements, or lack thereof, is an element of human decision making. This is where the aforementioned informative missingness emerges for this study. While hospitals in more modern times may be able to remove this effect by automatically recording measurements to EHR, there is no thorough way to remove it in this instance. Instead, missingness is observed and its effect on the study discussed.

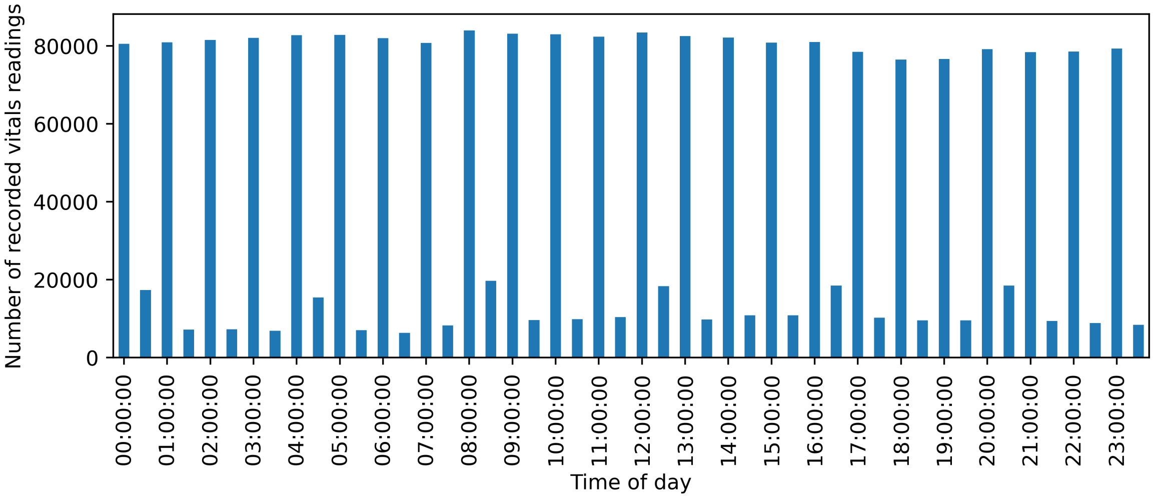

Figure 1: Number of vitals readings vs time of recording, grouped by 30 minute periods for ICU admissions in MIMIC-III.

All data used for this study is from the MIMIC-III edition dataset [17]. While it is no longer the most comprehensive nor recent edition of the MIMIC database, being superseded by MIMIC-IV in 2023, it is the most applicable to this use-case [16]. This is because, unlike MIMIC-IV, MIMIC-III contains a callout table which can be used to find the exact time that a given ICU patient was declared RFD by a doctor. Callout time is required to establish a point in time after which the outcome of a discharge decision can be classified as positive or negative. A potential work-around could be to approximate callout time with ICU out-times. However, there typically are delays between callout and discharge. This can be seen in Figure 2, which shows the distribution of callout-to-discharge delay times for ICU patients in the MIMIC-III dataset. On average these delays are more than 7 hours, with some spanning multiple days in extreme cases. These delays are significant compared to the time-scale of the physiological measurements in the ICU, so it would not be not an adequate approximation.

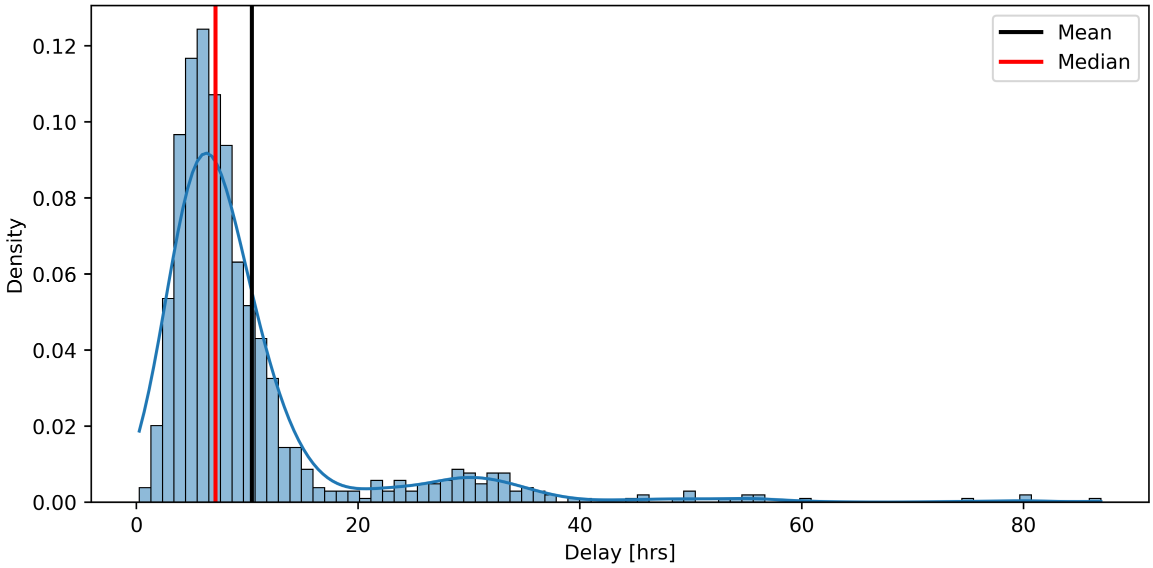

Figure 2: Delay in hours between callout and discharge for a sample of 1000 medical and surgical intensive care patients from MIMIC. Mean = 10.427 (3 s.f.), median = 7.126 (3 s.f.).

The database contains no pediatric or neonatal patients and has a minimum patient age of 15 years old. Therefore, any modelling on this data would need to be extrapolated for patients under the age of 15, thus it is not necessarily expected to give reliable predictions for children. There are two sources of data in MIMIC-III; Metavision and Carevue. In [21] it was found that the labelling of the relevant measurements in the Carevue subset is much less consistent, so analysis was restricted to only the Metavision subset. Likewise, only Metavision is used in this study for the same reasons.

The physiological measurements and lab results on which the features are derived are based on the nurse-led discharge criteria [18]. These measurements can be considered standard for EHRs from ICUs and are present in most ICU data studies. While more sophisticated measurements may be collected at some hospitals, using standard measurements increases the transferability of developed tools between hospitals. The physiological measurements are shown in Table 1. The test conditions shown are the NLD criteria [18].

For this study, 'vitals' denotes the approximately hourly recorded respiratory rate, heart rate, blood pressure, temperature, blood oxygen and Glasgow Coma Scale score (GCS) variables. 'Lab results' denotes the haemoglobin, potassium, sodium, creatinine, blood urea nitrogen, bicarbonate, fraction of inspired oxygen variables. In the following methods the pain and airway variables are handled similarly to the lab results due to their similar frequency of measurement. In [21], the admission level variables included age, sex, length of stay (LOS) and body mass index (BMI).

Test ID	Test name	Variable	Test condition
R0	Respiratory: airway	airway	airway patent
R1	Respiratory: Fio2	fio2	fio2 ≤ 0.6
R2	Respiratory: blood oxygen	spo2	spo2 ≥ 95 (%)
R3	Respiratory: bicarbonate	hco3	hco3 ≥ 19 (mmol/L)
R4	Respiratory: rate	resp (rate)	$10 \leq$ resp ≤ 30 (bpm)
C0	Cardiovascular: blood pressure	bp (systolic)	bp ≥ 100 (mm Hg)
C1	Cardiovascular: heart rate	hr	$60 \leq$ hr ≤ 100 (bpm)
P	Pain	pain	$0 \leq$ pain ≤ 1
CNS	Central nervous system	gcs	gcs ≥ 14
T	Temperature	temp	$36 \leq$ temp ≤ 37.5 (°C)
B0	Bloods: haemoglobin	haemoglobin	haemoglobin ≥ 90 (g/L)
B1	Bloods: potassium	k	$3.5 \leq$ k ≤ 6.0 (mmol/L)
B2	Bloods: sodium	na	$130 \leq$ na ≤ 150 (mmol/L)
B3	Bloods: creatinine	creatinine	$59 \leq$ creatinine ≤ 104 (umol/L)
B4	Bloods: urea	bun	$2.5 \leq$ bun ≤ 7.8 (mmol/L)

Table 1: Codified version of the nurse-led discharge criteria for application to electronic health record data. Here the 15 criteria have been grouped into intuitive subsets and each assigned a test ID ('R0' to 'B4'). According to the original specification, if all 15 criteria are met for a period of at least 4 hours the patient can be safely discharged.
Table and caption reproduced from [21].

3.2 Data Extraction

In addition to the aforementioned callout table, the required tables include 'patients', 'admissions', 'icustays', 'chartevents' and 'd_items'. These contain hospital admission-level data, demographic data, ICU admission-level data, individual events such as measurements/results and a lookup table for further information on these events, respectively. These tables are linked by patient IDs, hospital admission IDs and ICU admission IDs, which allow for convenient cross-tabulation.

Once filtered to only contain MICU and SICU data from Metavision, the chartevents table is queried for events with item-ids belonging to the list of relevant measurements shown in Table 1. This extraction is then merged with the ICU admission and patient tables to add sex, date of birth (DOB), ICU in-time, out-time and length of stay (LOS). Measurement labels and units are added from the d_items table.

3.3 Inclusion Criteria and Class Labels

For this study, the outcome class labels and cohort inclusion criteria from [21] are adopted, since the methods aim to classify the ICU admissions by the same outcomes. As such, each ICU admission is assigned a binary cohort label by checking if it is the first instance of ICU admission for the linked patient ID, and if it has a recorded callout time. Binary outcome class labels are assigned as positive (class 1) if there is only one ICU admission linked to the same hospital admission ID and if there is no record of in hospital death, according to the admissions table and negative (class 0) otherwise [21].

3.4 Data Exploration and Preprocessing

To map the events in the extracted data to physiological variables, an item-variable mapping is provided in the code repository for [21]. According to this mapping, some variables are captured by a single type of measurement event, whereas others are described by multiple, often only slightly different measurement events. For example, blood pressure can be measured from the left arm or the right arm. To validate the mapping for this study, the measurements items and associated units of measurement (UOM) present in the data are checked, as shown in Table 2.

Variable	UOM	Description
fio2	nan	'Inspired O2 Fraction'
resp	'insp/min'	'Respiratory Rate', 'Respiratory Rate (spontaneous)', 'Respiratory Rate (Set)', 'Respiratory Rate (Total)'
temp	'°F', '°C'	'Temperature Fahrenheit', 'Temperature Celsius'
hr	'bpm'	'Heart Rate'
bp	'mmHg'	'Non Invasive Blood Pressure systolic', 'Arterial Blood Pressure systolic', 'ART BP Systolic'
k	'mEq/L'	'Potassium (serum)', 'Potassium (whole blood)'
na	'mEq/L'	'Sodium (serum)', 'Sodium (whole blood)'
hco3	'mEq/L'	'HCO3 (serum)'
spo2	'%'	'O2 saturation pulseoxymetry', 'Arterial O2 Saturation', 'PAR-Oxygen saturation'
bun	'mg/dL'	'BUN'
airway	nan	'ETT Mark (location)', 'ETT Type', 'ETT Size (ID)', 'ETT Location' 'Trach Tube Type', 'ETT Re-taped'
gcs	nan	'GCS - Eye Opening', 'GCS - Verbal Response', 'GCS - Motor Response'
creatinine	'mg/dL'	'Creatinine'
pain	nan	'Pain Level'
haemoglobin	'g/dl'	'Hemoglobin'
peep	'cmH2O'	'PEEP set', 'Total PEEP Level'
weight	'kg'	'Admission Weight (Kg)'
height	'cm'	'Height (cm)'

Table 2: Variable mapping of measurement items in the MIMIC-III dataset, with associated units of measurement and measurement descriptions.
UOM, units of measurement.

All mappings have consistent UOMs, except for temperature which has measurements in both degrees Celsius and degrees Fahrenheit. After converting all temperature measurements to Celsius, further unit conversions are made such that all measurements are in equivalent metric units, as used in the UK. Specifically, *creatinine* is converted from mg/dl to umol/L and *bun* (blood urea nitrogen) is converted from mg/dl to mmol/L [21].

The airway variable, which is a binary representation of whether or not airway is patient, is simply derived from the presence of an endotracheal tube, as in [21]. As described in Section 2, The Glasgow Coma Scale score is comprised of an eye opening score, a motor response score and a verbal response score. The GCS score is the sum of these components and ranges from 3 (worst) to 15 (best) [33]. Since all three components are required to form the score, only GCS measurements for which all 3 exist are included from the data. The GCS variable represents the summed score for these components [21].

Date of birth and admission time are used together to calculate age for each patient. However, patients in MIMIC-III who were more than 89 years old have had their date of birth adjusted such that their true age is hidden, in compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations. The median age of patients over the age of 89 is 91.4 [17]. Since this is the only information available regarding their age, all patients over 89 are assigned age 91 for this study. Height and weight are used as admission-level for this study. That is, only a single measurement (the first) of height and of weight are required per patient per ICU admission. In doing this, change in these variables, particularly weight, that may occur over the course of an ICU admission is not explored for its predictive power. Instead, these measurements are combined to calculate body mass index (BMI) as a stationary variable for each patient's entire admission. The inclusion of BMI may be useful for comparison with other studies. However, it should be noted that the medical relevance of BMI is an area of contentious debate [5,12,25].

Next, all variables are subject to anomaly removal according to value ranges that are physically possible. These ranges were provided by clinicians [21].

3.5 Feature Extraction

In this section, the construction of features from the data is detailed. These features are the inputs upon which ML models may classify patients as RFD or NRFD. The features are split into two main groups. There are the base features from [21] which are based on the NLD criteria, using the last 4 hours of data, with the addition of admission-level features. Then, there are the gradient-based features which are novel to this study. Lastly, a missingness meta-feature is created, but is excluded from the main results due to the issues surrounding informative missingness discussed in Section 2.

3.5.1 Base Features

As in [21], feature matrices are produced by taking measurements from the 4-hour window prior to callout time for each patient. The choice of time window size as well as the variables (Table 1) is based on the NLD criteria for comparative reasons [18]. Frequent vital measurements are split into minimum and maximum readings over the time window, whereas for the less frequent lab results only a single reading is taken. However, due to the infrequency of these lab results many admissions lack data for the 4-hour window. To overcome this, a maximum look-back window of 36-hours is queried for lab results and only the last recorded results are added to the feature matrix. To extend on the NLD features age, sex and length of stay (LOS) are added as admission level features. For this study, this feature creation process is reproduced from [21] to establish base level performance, as well as for validation. However, only around 50% of patients had a recorded height during their ICU admission in this edition of MIMIC. For this reason, the BMI feature is omitted, since it requires height in its calculation.

3.5.2 Gradient Features

The concepts of recovery and deterioration describe a net change in health condition over time. For patients in intensive care recovery or deterioration may be characterised by changes in measured physiological values over the duration of their admission. In mathematical terms this can be interpreted as a gradient: $\frac{\Delta\text{measured value}}{\Delta\text{time}}$. There is evidence to suggest that features based on the gradients of measurements may enhance predictive power in relation to ICU outcomes, as discussed in Section 2. Particularly, some 24-hour gradients of vital measurement values were found to outperform their non-temporal counterparts when predicting ICU readmission [29]. Inspired by these findings, the addition of gradient-based features to the previously defined feature matrix is investigated.

The rationale behind the investigation into gradient features is also based in the concepts of temporal vs static data, as well as relative vs absolute measurements. The NLD-based features described in Section 3.5.1 are non-temporal as they include only the absolute values of the most recent measurements. For some measurements, absolute values are very indicative of physical condition. For example, healthy body temperature is largely uniform across the human population regardless of other physiological differences [26], so a significant deviation from the norm in absolute temperature values can be expected to reflect health decline. However, other measurements have natural variation in their typical absolute values between different patients, such as heart rate or respiration rate, which may be due to physiological differences such as body size or age rather than unhealthiness. Since the calculation of gradients requires taking the difference of absolute measurement values, gradient features are effectively normalised such that they do not account for natural variation between different patients. Instead, it is how drastic the changes in these physiological measurements are that provide the predictive information to a model. Using both absolute, static features as well as temporal, gradient-based features gives a model the opportunity to learn from both types of data where they are important.

In order to observe gradients for variables in the ICU data there must be defined start and end points in time, relative to each ICU admission, over which the rate of change is calculated. Features using absolute measurement values from the final 4 hours of ICU admission were demonstrably informative for classifying RFD with supervised learning [21]. This is likely because they capture the condition of the patient in the most recent moment, so they indicate if the patient is RFD in their current condition. Therefore, using the measurements from the current 4 hour window as the end point for gradient features is likely to offer the most predictive power.

There are many options for the start point for the gradient features. This study bases this decision on the findings from [2]. As discussed in Section 2, these findings suggest that data from the initial period of an ICU admission are useful for predicting eventual outcome. In line with this, a 4-hour window of the initial period of data after admission is used as the start point for the gradient features. The start and end points for the gradient features are now defined. However, the initial and current 4 hours of admission may occur at completely different times of day. In the clinical context, this may lead to detrimental implications for modelling due to effects that arise from the time of day.

Time of day effects refer to patterns which are bound by a 24-hour cycle. In the medical context, there are normal and natural variations in physiological measurements that occur due to biological mechanisms such as circadian rhythms [37]. If these patterns are not accounted for, a ML model may mistakenly learn to make predictions of health based on these variations, despite the fact that they are benign. There are also time of day effects that arise due to the scheduling that hospitals adhere to. The most relevant of these to this study is the scheduling of callout decisions. It is typical for doctors to assess SICU patients for readiness for discharge at some point in the morning to free up beds to receive incoming patients from operations or surgeries that are scheduled later that day. This may vary between hospitals. The distribution of callout times for the MIMIC-III cohort is shown in Figure 3, with most callouts happening around 11am. There is even some evidence that time of day effects also exist in the behavior and decision making of the clinicians, particularly in relation to work hours [36]. Since the presence of measurements in the MIMIC data reflects the decisions of clinicians to take physiological measurements of the ICU patients, this effect may be relevant to this study.

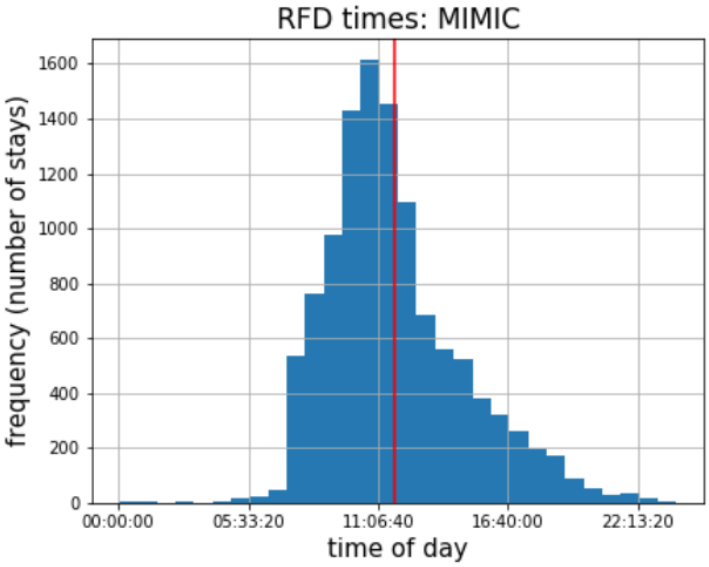

Figure 3: Histogram of the time of day at which patients were declared ready for discharge (callout time). The red line indicates the mean of the distribution.
Figure and caption reproduced from [21].

To mitigate these time of day effects when creating the gradient features, the initial 4 hour data window is placed at a position within the first 24 hours of the ICU admission such that the resultant length of time between the initial and final data periods is a multiple of 24 hours. This is a relaxation of the justification to use the initial 4 hours of data from the findings in [2]. However, model performance on data up to the initial 24 hours was still higher than that of up to 72 hours and a week, so it is still reasonable [2]. For convenience, this 4-hour window that occurs within the first 24 hours of an ICU admission is still referred to as the initial data window.

With start and end points established, three gradient features are calculated for each of the vital measurements. For a given variable with initial period values, $\mathbf{U} = [u_1, ..., u_n]$ recorded at times, $\mathbf{T}_u = [t_{u_1}, ..., t_{u_n}]$ and final period values $\mathbf{V} = [v_1, ..., v_m]$ recorded at times, $\mathbf{T}_v = [t_{v_1}, ..., t_{v_m}]$ the mean gradient is given by

$$\text{mean grad} = \frac{\text{mean}(\mathbf{V}) - \text{mean}(\mathbf{U})}{\text{mean}(\mathbf{T}_v) - \text{mean}(\mathbf{T}_u)}. \tag{1}$$

The 'min-max' gradient is given by

$$\text{min-max grad} = \frac{\min(\mathbf{V}) - \max(\mathbf{U})}{t_{\min(\mathbf{V})} - t_{\max(\mathbf{U})}}, \tag{2}$$

and the 'max-min' gradient is given by

$$\text{min-max grad} = \frac{\max(\mathbf{V}) - \min(\mathbf{U})}{t_{\max(\mathbf{V})} - t_{\min(\mathbf{U})}}, \tag{3}$$

where $t_{\min(\mathbf{V})}$ is the recorded time of the measurement with minimum value in $\mathbf{V}$, et cetera.

These gradient features are best explained with an example. Figure 4a) shows the data points on which these three gradient features would be calculated for the blood pressure measurements from an exemplary ICU admission. For this example, the mean gradient feature describes a change of $\sim +1.32$ mmHg/hour in Blood pressure.

The min-max and max-min gradients are able to capture the steepest and shallowest changes in variables for cases that have either increasing (Figure 4a)) or decreasing (Figure 4b)) trajectories over the course of the ICU admission. They also preserve the direction of change as well as the magnitude. This is important since taking just the magnitude of change may lose predictive information. For example, an incline in blood pressure of 1.32 mmHg/hour might be more indicative of deterioration than a decline of 1.32 mmHg/hour, despite the magnitude being the same. A mean gradient is calculated using the difference between the mean of values in the final 4 hour window and the mean of values in the initial window. The averaging makes this feature more robust to noise and thus more likely to capture the overall trend.

Since these gradient features use measurements from only the initial and final time periods of a given ICU admission they ignore the behavior of the trajectory between these periods. In some cases this means that drastic deviations from the overall trajectory of measurements are completely ignored in the calculation of the gradient features. Figure 4c) shows an example of this, with a substantial rise in blood pressure which returns to values that are similar to the initial period before the end of the admission, resulting in relatively shallow gradients. This example highlights a compromise in the design of these gradient features. That is, potentially informative deviations from the overall trajectory of a patient's recovery in the ICU are ignored in favour of the comparison between just their initial and current conditions, as reflected in their measurements.

Since these deviations from overall trajectory are temporary, it is justifiable that they are ignored. In some cases, the deviation may represent a temporary deterioration in health from which the patient recovers by the time the discharge decision is made. In others, the initial and final measurements indicate poor health condition, in which case a deviation may represent a temporary recovery before deterioration back poor health condition. The absolute minimum and maximum values for vitals are included within the base features alongside the gradient features for modelling. Their combination may allow a clinician to distinguish between these two deviation scenarios when assessing the explanation of the model predictions.

These gradient features are easily available in deployment, as patients only need to have been in the ICU for a minimum of 24 hours to have this data readily available for use in model predictions. These features are also not too complicated to have meaningful clinical interpretations. Therefore, inclusion of gradient features is not likely to negatively impact the clinical deployability of the decision support tool.

3.5.3 Missingness Feature

The last additional feature is not based on the values of measurements but instead their frequency. As discussed in Section 3.1, missing data is likely to be informative in the intensive care context. This means that the performance of a classifier may benefit from using meta-features based on the missingness of data. In this study, a missingness meta-feature that is the proportion of missing entries (NaN values) across all the aforementioned features is calculated for each ICU admission in the cohort. However, modelling with meta-features that are based on missingness risks the transportability of a model to future applications, if

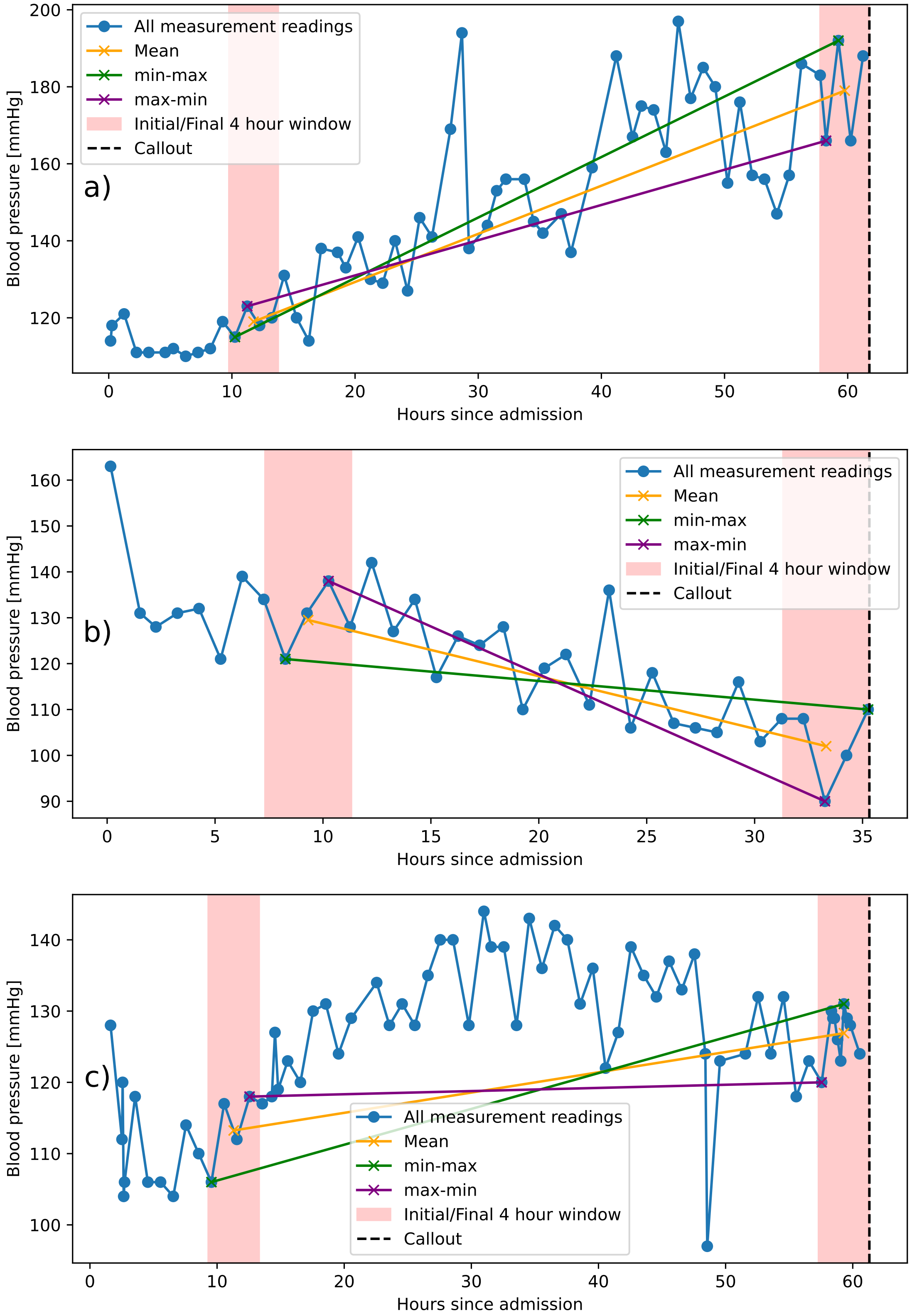

Figure 4: Graphs of the blood pressure readings from three randomly selected ICU admissions. The initial window is shifted such that it is a multiple of 24 hours from the final window. Here, 'Mean' refers to the mean of the blood pressure readings in each of the initial and current 4 hour windows. These mean values are connected with a line, the gradient of which is the 'mean grad' feature for blood pressure. 'min-max' refers to the minimum value in the initial window and the maximum value in the final window, and vice versa for 'max-min'.

the mechanisms by which missing data occurs change [11]. For this reason, the missingness meta-feature is used only as an opportunity to separately measure informative missingness for this dataset, but is omitted from the final model and results.

4 Methodology

The methodology details how the features described in Section 3 are modelled to create a discharge decision support tool. The method by which predictions of readiness for discharge in the future are achievable requires the creation of multiple sets of similar feature matrices, but with varying data query conditions. The feature matrices are then balanced in terms of class, before missing data is imputed. Then, the construction of an EBM and a RF model, the training methodology and the choice of performance metrics are described.

4.1 Predicting Future Outcomes

Using the features discussed so far, a classifier model can output a prediction of whether or not a patient is ready for discharge at a given point in time, and so far this point is recorded callout. The model does not predict the length of time that the patient will take to become RFD. Predicting a length of time is a regression task which is far more difficult than classifying RFD at a single point in time. However, by moving the point in time at which a classifier is predicting RFD, it may be able to predict if a patient will be RFD at a fixed point in the future. By training multiple instances of classifiers at different points in time, a set of models can be created which are able to separately predict RFD at several discrete time steps into the future. A tool consisting of the combined set of models implicitly offers a prediction of when a patient will be RFD, the time-granularity of which is limited by the size and number of discrete time steps at which classification occurs.

The granularity and scope of the composite model tool is a design choice that is subject to constraints of the dataset as well as the clinical deployment context. Time of day effects must again be controlled here. For the composite model tool, these time of day effects are most elegantly mitigated by choosing a granularity of multiples of 24 hours for predicting when a patient is RFD. That is, the composite model tool is composed of individual models which predict readiness for discharge at around 11am (Figure 3) and integer multiples of 24 hours after then.

Each binary classification model in the composite model tool is identical in architecture. Instead, it is the preparation of data that each model is trained on that is varied in order to achieve these future RFD predictions. To create the feature matrix for the 24-hour future prediction model, each ICU admission in the cohort is queried again to produce the same features described in Section 3.5 and with the same outcome class labels. However, the final 4 hour window from which the features are calculated shifted back so that it is set at exactly 24 hours before the recorded callout time. This effectively moves the 'current time' back by 24 hours for each patient. Therefore, the class labels correspond to the outcomes for patients if they were called out in 24 hours time from the 'current time' at which the patients data is being sampled. Training a classifier on this data with the recorded outcomes as the ground truth gives an approximation of the question of will a given patient be ready for discharge in 24 hours time. This process is repeated for integer multiples of 24 hours before callout. Figure 5 illustrates how the final data window is shifted when resampling patients to create each of these feature matrices. To avoid data leakage the admission-level length of stay feature must be adjusted by the multiple of 24 hours also, since it contains information from the future relative to the current time at which the model predictions are being made.

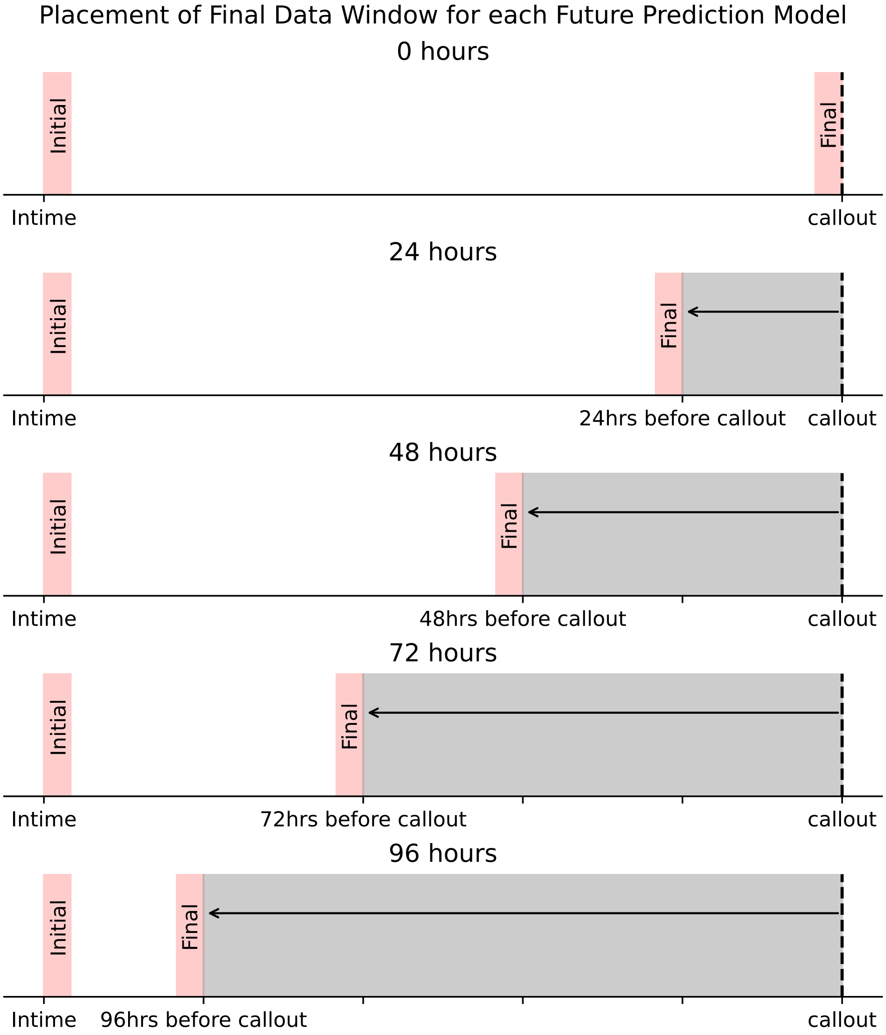

Figure 5: Diagram showing how the final 4-hour data window is shifted relative to the timeline of ICU admission when resampling patients in order to create feature matrices for each of the future prediction models. The grey area represents the length of time into the future at the end of which callout is proposed.

One drawback of this resampling process is it that shortens the length of stay of each patient between the initial and current window. In order to have well defined initial and current windows there is a minimum requirement of 24 hours of ICU admission from which the data can be queried. This means that the larger the multiple of 24 hours that is taken between prediction and outcome, the fewer the number of ICU admissions that are available in the dataset to train and test the model on. This combined with the fact that the majority of ICU admissions last from 1-3 days causes a limiting factor on how far into the future predictions can be made due to data availability. Due to this limitation, 96 hours is chosen as the furthest prediction time. Table 3 shows a summary of feature matrices created using this method, including the size (number of admissions) and number of instances of each class.

	0 hours	24 hours	48 hours	72 hours	96 hours
Total ICU admissions	7033	3884	2458	1711	1248
Positive outcomes (class = 1)	6299	3373	2075	1414	1019
Negative outcomes (class = 0)	734	511	383	297	229
% Negative outcomes	10.4	13.2	15.6	17.4	18.3

Table 3: Table showing the class imbalance of the feature matrices for each of RFD prediction time models.

4.2 Class Imbalance

The number of instances of each class in the cohort is unbalanced; there are many more positive outcomes post-discharge than negative outcomes. In context this suggests that the doctors who gave these callout verdicts were 'correct' much more often than not. However, in order to achieve good performance on predicting both positive and negative outcomes using a ML classifier it is optimal to train and test on a balanced number of instances of both classes. A simple and typical method is to repeat instances of the under-represented class to provide as many as are required to achieve balance. In the context of this study, doing this would mean that individuals from the negative class are significantly over-represented, because each individual would have to be repeated multiple times to achieve balance. This over-representation is likely to cause a ML model to over-fit to these individuals, leading to loss of generality [27].

There are more sophisticated methods, such as the popular Synthetic Minority Oversampling Technique (SMOTE) which creates new samples using the feature distributions from the existing minority class data. As can be seen in Table 3 the number of instances of the negative class for our data is between 5-10 times less than that of the positive class. This limits the usefulness of SMOTE, as the performance of SMOTE sharply declines as the number of samples of the minority class decreases [7]. Furthermore, samples created by SMOTE are entirely synthetic, and it is not necessarily the case that these synthetic samples are considered realistic in the clinical context, regardless of the fact that they fit the distributions of our particular dataset. There may be causal relationships between different physiological parameters that determine their range of possible values, ans SMOTE may be unable to capture this. While there exist adaptations of SMOTE which take the context of medical data specifically into account [23], it is ideal to avoid synthetic data, if possible.

In [21], it is argued that any patient who's data is resampled from more than 3 days before callout can be assumed to be not RFD at that point in time, regardless of eventual outcome [21]. Under this assumption, resampling the cohort data from 3-8 days before callout provides enough instances of the negative class to achieve class balance. While this involves resampling the same patients, it is data from an entirely different portion of their ICU admission that is used to build the resultant feature matrix. Therefore, the classes are balanced using real data that is also not repeated. The data is sampled from time windows that are multiples of 24-hours before callout. This circumvents the time of day effects described in Section 3.5.2. For this study, data from the initial period that contributes to the gradient features will be repeated when resampling using this technique. The final periods will be different, so the gradient features will not be identical. Still, this may contribute to a degree of over fitting to the resampled patients.

For this study, this class balancing method is considered the most ideal and is adopted. For the future prediction feature matrices, this means that patients would be resampled 3-8 days before the end of the adjusted final data window. While this solution largely overcomes the limitations of the previously mentioned class balancing methods, it is also not perfect. The approximation that any patient was not RFD 3 days or more before callout relies on the underlying assumption that all patients from this dataset were called-out within 3 days of being truly RFD [21]. In reality, patients may have been called out late due to constraints on the hospital, or simply due to the tendency for conservative decision making in medical contexts. If there are a significant number of patients who were called-out more than 3 days late then this resampling technique may introduce noise in the form of mislabeled data, which will negatively impact the performance of ML classifiers. It may be possible to classify patients as being late-discharged, however that is outside the scope of this project.

In an attempt to minimise the possibility of mislabelled data due to late discharge in the dataset, the resampling window for negative class is pushed back to 4-8 days prior to callout. This provides more than enough instances of the negative class to balance all feature matrices except for the 0 hours subset. The class counts after the resampling process is shown in Table 4. After this, the larger class for each feature matrix is resampled to give a 50:50 ratio between class 1 and class 0.

	0 hours	24 hours	48 hours	72 hours	96 hours
Total ICU admissions	11211	7220	5141	3873	3006
Positive outcomes (class = 1)	6299	3373	2075	1414	1019
Negative outcomes (class = 0)	4912	3847	3066	2459	1987
% Negative outcomes	43.8	53.3	59.6	63.5	66.1

Table 4: Table showing the class imbalance of the feature matrices after resampling from 4-8 days prior to the end of the final data window to create more instances of the negative class.

4.3 Missing Data

Typically, ML methods which are used for classification are not robust to missing data, so it must be filled in by a process called imputing. Imputing involves the synthesis of values where there are currently none to yield an artificially complete dataset. RF are not robust to missing data. EBM are robust to missing data, since they treat NaN as its own legitimate value. However, this allows the EBM to learn from informative missingness, as it can make predictions based off the occurrence of NaNs. The opportunity to learn from informative missingness is dropped in favor of transportability of the models, as discussed in Section 2. Therefore, imputing is still required for this study.

The missing data is imputed using a k nearest neighbours (K-NN) imputer [22]. The imputer is first fit to complete case data [21]. It is then used to impute the missing values in the rest of the data. Imputation is not applied at this stage. Instead, it is applied as a part of each training run of the classifiers. Before training, the data is partitioned into a training set and test set. This partition is random and varies between runs of the model. In order to fairly evaluate the performance of a supervised classifier it must not have been exposed to the test set data. If the imputer is fit using the data prior to being split into test and train, then it is possible for the test data to influence the imputed values in the train data. This is known as data leakage. To avoid this, the imputer is re-fit to the train partition of the data for each run of the model. Before the imputation is applied, patients with 3 or more missing values are removed, as in [21].

There are many methods available for imputation, however none can be as truthful as the data being complete in the first place. For this reason, complete case versions of each feature matrix are also prepared. Comparison of results between imputed and complete case analysis may give an indication of the impact of the imputation method.

4.4 Summary of Data After Preparation

The number of ICU admissions in the feature matrices for each prediction time are shown in Table 5. Each ICU admission corresponds to a row of data in the feature matrix.

	0 hours	24 hours	48 hours	72 hours	96 hours
Total ICU admissions	7664	5258	3280	2260	1652
Age, median years	62	62	62	61	60
Sex, % female	46.8	46.6	45.8	45.0	44.3
Length of stay, median days	2.847	3.385	3.732	4.212	4.529
Negative outcome, %	50.0	50.0	50.0	50.0	50.0

Table 5: Summary of cohort feature matrices for each prediction time after class balancing and removal of ICU admissions with 3 or more missing data entries.

The full feature set as well as the proportions of missing entries in each feature matrix before imputation are shown in Table 6. These features are are the input on which a binary classifier is trained to predict the class label (RFD or NRFD) as output.

Feature set	Variable	0 hours	24 hours	48 hours	72 hours	96 hours
Base features	bp MIN	0.009	0.007	0.005	0.006	0.008
	bp MAX	0.009	0.007	0.005	0.006	0.008
	gcs MIN	0.099	0.084	0.074	0.073	0.070
	gcs MAX	0.099	0.084	0.074	0.073	0.070
	hr MIN	0.004	0.004	0.003	0.004	0.005
	hr MAX	0.004	0.004	0.003	0.004	0.005
	resp MIN	0.008	0.008	0.005	0.006	0.008
	resp MAX	0.008	0.008	0.005	0.006	0.008
	spo2 MIN	0.013	0.008	0.007	0.006	0.008
	spo2 MAX	0.013	0.008	0.007	0.006	0.008
	temp MIN	0.070	0.056	0.053	0.051	0.050
	temp MAX	0.070	0.056	0.053	0.051	0.050
	k	0.010	0.011	0.014	0.016	0.019
	na	0.010	0.011	0.014	0.016	0.019
	bun	0.010	0.012	0.015	0.018	0.021
	creatinine	0.010	0.012	0.015	0.018	0.021
	hco3	0.010	0.012	0.015	0.018	0.021
	haemoglobin	0.011	0.013	0.016	0.018	0.021
	fio2	0.507	0.450	0.425	0.407	0.393
	airway	0.000	0.000	0.000	0.000	0.000
	pain	0.089	0.082	0.083	0.083	0.083
	age	0.000	0.000	0.000	0.000	0.000
	sex	0.000	0.000	0.000	0.000	0.000
	LOS	0.000	0.000	0.000	0.000	0.000
Gradient features	bp MEAN GRAD	0.016	0.016	0.016	0.018	0.020
	bp MIN MAX GRAD	0.016	0.016	0.016	0.018	0.020
	bp MAX MIN GRAD	0.016	0.016	0.016	0.018	0.020
	gcs MEAN GRAD	0.167	0.156	0.150	0.149	0.151
	gcs MIN MAX GRAD	0.167	0.156	0.150	0.149	0.151
	gcs MAX MIN GRAD	0.167	0.156	0.150	0.149	0.151
	hr MEAN GRAD	0.011	0.013	0.013	0.015	0.017
	hr MIN MAX GRAD	0.011	0.013	0.013	0.015	0.017
	hr MAX MIN GRAD	0.011	0.013	0.013	0.015	0.017
	resp MEAN GRAD	0.015	0.016	0.016	0.017	0.019
	resp MIN MAX GRAD	0.015	0.016	0.016	0.017	0.019
	resp MAX MIN GRAD	0.015	0.016	0.016	0.017	0.019
	spo2 MEAN GRAD	0.019	0.015	0.016	0.015	0.018
	spo2 MIN MAX GRAD	0.019	0.015	0.016	0.015	0.018
	spo2 MAX MIN GRAD	0.019	0.015	0.016	0.015	0.018
	temp MEAN GRAD	0.133	0.122	0.119	0.120	0.119
	temp MIN MAX GRAD	0.133	0.122	0.119	0.120	0.119
	temp MAX MIN GRAD	0.133	0.122	0.119	0.120	0.119
Missingness features	missingness	0.000	0.000	0.000	0.000	0.000

Table 6: Table of features with proportion of missing entries for each feature matrix prepared for 0, 24, 48, 72 and 96 hours predictions.

4.5 Machine Learning Classifiers

There is a huge variety of machine learning methods which offer the ability to learn the underlying patterns in data for binary classification tasks. In general, there is a trade-off between model explainability and model performance. The Explainable Boosting Machine (EBM) is a recent (2019) supervised ML method that defies this trade-off [24]. EBMs have accuracy comparable to state-of-the-art ML methods while having highly interpretable and explainable output. This makes EBM a great choice for this study, since interpretability of model predictions is essential in the medical decision-making context (Section 2).

Generalized additive models (GAM) are generalized linear models which have a linear response to smooth feature functions [13]. EBM is a generalized additive model of the form:

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j) \tag{4}$$

where f_j are feature functions for each feature x_j, g is the link function, the inverse of which is a sigmoid function in the case of classification, and β_0 is an intercept term. The $f_{i,j}(x_i, x_j)$ terms are functions that capture pairwise interactions between features which EBM can detect automatically. Each feature function is learnt by bagging and gradient boosting methods which are configured such that the impacts of co-linearity are minimal and feature order is unimportant [24]. Algorithmic Details for the training procedure and selection of pairwise interaction terms can be found in [19].

When making individual (local) predictions with the EBM, the feature functions each return term contributions which are summed and input into the link function to give the output. It is this additivity that gives rise to the interpretability of the model predictions, since term contributions and thus features can be sorted by importance on any individual prediction [24]. Furthermore, since EBM is a tree-based ML method it does not require the input data to be scaled. This is advantageous as it means the explanations of the model are directly interpretable in terms of real units of measurement values.

After training, a global explanation of the EBM can be explored. Each of the single and pairwise feature function terms can be plotted to give a visual representation of their individual contributions to prediction outputs. Similarly to the local case, these can be ranked to give feature importance. When visually exploring the functions for each feature there may be behaviors of functions that appear anomalous in context of what the feature represents. These artifacts can occur for a variety of reasons, typically it is a reflection of the limitations of the data set in use. For a given feature function, if an anomaly or artifact is detected it can be easily edited directly from the visualisation, which immediately edits the contributions from that feature to any EBM predictions. Such a clear communication of feature contributions allows the model to be reviewed by clinicians who may perform edits to individual feature functions according to reason in the medical context.

For this study, the composite model tool is made up of 5 individual EBM models, one for each of the [0, 24, 48, 72, 96] hours-into-the-future predictions. It's local output will be binary predictions of RFD for each of these points in time relative to the current point in time of a given patient's recorded data. Each of these predictions will be explainable by separate feature contribution rankings. The feature functions are separately editable for each of the EBM models, however they are left unedited in this study since this requires clinical expertise. Another composite model that consists of 5 random forest models is created for comparison of performance metrics against the EBM based model as well as the results from [21].

4.6 Training Methodology

For each of the five prediction times an instance of an EBM and RF are created and trained on their respective feature matrix. The process for training each individual EBM (RF) begins with splitting the feature matrix into train and test subsets with a ratio of 70:30. An instance of a K-NN imputer is fit to only the train set and then used to impute missing values on both the train and test sets. For the complete-case version this imputation step is not required. Next, the EBM (RF) is optimised for the F1-score on the train set via gridsearch cross-validation with five folds over a hyperparameter space. The performance of the optimised classifier is then evaluated on the test set. This process is repeated for 100 random train-test splits [21].

The metrics used to evaluate performance on the test set are F1-score, accuracy, sensitivity, area under the receiver operating characteristic (AUROC), partialAUROC (pAUROC) and Brier score. Except for the AUROC and Brier score, metrics are evaluated at a specificity of 0.7 by estimating this point in receiver-operator-characteristic space via linear interpolation [21]. To create a results table, the averages and standard deviations of all metrics over the 100 repeats are taken. The choice of performance metrics as well as the specificity threshold is based on the most similar studies in the literature [10, 21], for comparative reasons. Furthermore, it is important to report on a range of performance metrics for classification models since this reduces the risk of misrepresenting the performance of the models [15].

Having 20 hyperparameters, the hyperparameter space for EBM is very large. Computing gridsearch over the entire hyperparameter space would be prohibitively expensive with the resources available for this study. The EBM documentation claims that it performs well for most datasets without any hyperparameter tuning [6]. However, the documentation recommends varying a subset of the hyperparameters which are better understood. Specifically, it is recommended to tune two hyperparameters over the following ranges: $max\_bins \in [1024, 4096, 16384, 65536]$, $max\_leaves \in [3, 4]$. For the random forest models, three hyperparameters are tuned over the following ranges: $n\_estimators \in [20, 50, 100]$, $max\_features \in [20, 50, 100]$, $max\_depth \in [4, 5, 6, 7]$, as in [21].

The training and evaluation process is repeated firstly for the base feature set, and secondly for the same feature set but with the addition of the gradient features that are original to this study. This allows improvements to model performance that may arise due to the addition of the gradient-based features to be measured. For comparison, it is necessary to reproduce the results on the same features as [21] because the dataset contained data from another source which was not available for this study. This means that direct comparison to those results are not fair. The training and evaluation processes are also repeated with and without the inclusion of the missingness feature, and on the complete case data preparation.

5 Results

This section entails a comparison of the performance metrics evaluated for the EBM and RF models both with and without the gradient features. The complete-case and imputed results are compared. Continuing with the EBM with gradient features models, the decrease in performance with further forward looking predictions is shown. The global feature importances for each of the EBM models are compared, and some exemplary feature functions are visualised to demonstrate the interpretability and editability of the models. Lastly, the explainability of local predictions is demonstrated with examples from randomly selected patients.

5.1 Performance Metrics

		0 hours	24 hours	48 hours	72 hours	96 hours
EBM	F1	0.8322 (0.0056)	**0.7929** (0.0063)	**0.7192** (0.0160)	0.6349 (0.0343)	**0.5834** (0.0278)
	Accuracy	0.8128 (0.0048)	**0.7769** (0.0062)	**0.7150** (0.0117)	0.6513 (0.0245)	**0.6157** (0.0189)
	Sensitivity	0.9272 (0.0111)	**0.8561** (0.0137)	**0.7331** (0.0239)	0.6068 (0.0490)	**0.5371** (0.0352)
	pAUROC	0.1993 (0.0034)	**0.1770** (0.0072)	**0.1424** (0.0067)	0.1072 (0.0085)	0.0981 (0.0072)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.8902 (0.0041)	**0.8561** (0.0087)	**0.7905** (0.0107)	0.7120 (0.0164)	0.6701 (0.0181)
	Brier	0.1295 (0.0029)	**0.1526** (0.0043)	**0.1863** (0.0049)	0.2179 (0.0063)	0.2297 (0.0067)
EBM + gradient features	F1	**0.8353** (0.0042)	0.7782 (0.0095)	0.7159 (0.0147)	0.6443 (0.0190)	0.5745 (0.0241)
	Accuracy	**0.8158** (0.0033)	0.7639 (0.0077)	0.7127 (0.0083)	0.6572 (0.0127)	0.6089 (0.0132)
	Sensitivity	**0.9324** (0.0062)	0.8306 (0.0162)	0.7269 (0.0161)	0.6199 (0.0259)	0.5270 (0.0285)
	pAUROC	**0.2027** (0.0033)	0.1665 (0.0037)	0.1400 (0.0070)	0.1125 (0.0080)	0.0973 (0.0089)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	**0.8944** (0.0037)	0.8401 (0.0074)	0.7847 (0.0098)	0.7143 (0.0118)	0.6745 (0.0111)
	Brier	**0.1271** (0.0021)	0.1611 (0.0047)	0.1891 (0.0045)	0.2168 (0.0050)	0.2288 (0.0041)
RF	F1	0.8289 (0.0076)	0.7771 (0.0136)	0.7133 (0.0174)	**0.6538** (0.0224)	0.5804 (0.0256)
	Accuracy	0.8098 (0.0064)	0.7631 (0.0110)	0.7106 (0.0120)	**0.6640** (0.0149)	0.6129 (0.0176)
	Sensitivity	0.9199 (0.0118)	0.8284 (0.0211)	0.7232 (0.0252)	**0.6337** (0.0321)	0.5343 (0.0330)
	pAUROC	0.1889 (0.0052)	0.1631 (0.0087)	0.1361 (0.0059)	**0.1218** (0.0074)	0.0982 (0.0072)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.8792 (0.0055)	0.8368 (0.0117)	0.7800 (0.0106)	**0.7281** (0.0139)	**0.6679** (0.0153)
	Brier	0.1348 (0.0029)	0.1626 (0.0053)	0.1908 (0.0047)	**0.2112** (0.0046)	**0.2286** (0.0033)
RF + gradient features	F1	0.8276 (0.0091)	0.7780 (0.0148)	0.7154 (0.0160)	0.6344 (0.0190)	0.5821 (0.0339)
	Accuracy	0.8085 (0.0072)	0.7640 (0.0118)	0.7121 (0.0137)	0.6504 (0.0133)	0.6145 (0.0180)
	Sensitivity	0.9183 (0.0119)	0.8298 (0.0224)	0.7269 (0.0271)	0.6056 (0.0276)	0.5368 (0.0413)
	pAUROC	0.1887 (0.0063)	0.1611 (0.0086)	0.1338 (0.0069)	0.1146 (0.0043)	**0.1019** (0.0071)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.8793 (0.0073)	0.8347 (0.0107)	0.7758 (0.0135)	0.7149 (0.0106)	0.6632 (0.0182)
	Brier	0.1354 (0.0039)	0.1632 (0.0051)	0.1926 (0.0052)	0.2159 (0.0027)	0.2295 (0.0041)

Table 7: Performance metrics for EBM and RF models predicting readiness for discharge with and without additional gradient features at 0, 24, 48, 72 and 96 hours into the future. Scores are given as: mean (standard deviation) over 100 random train-test splits. All scores other than AUROC and Brier are evaluated at a specificity of 0.7. The best metric scores for each column (prediction time) are shown in bold.
EBM, explainable boosting machine; RF, random forest; AUROC, area under the receiver operating characteristic; pAUROC, partial AUROC.

Table 7 shows a comparison of the mean performance metric scores for both the set of EBM models and the set of RF models, with and without the addition of the gradient features. All models were able to capture predictive information from the data, however none of the models achieve accuracy in excess of 82%. The 0 hours RF model is nearly identical to the RF model from [21] in that they both predict RFD at the current time (0 hours into the future), the feature matrix it is trained on is prepared in the same manner, and the training methodology and hyperparameter space are the same. However, the RF model in [21] was also trained on data from a secondary dataset. At specificity of 0.7, the 0-hours RF model performs less well than the RF model from [21] according to all metrics except sensitivity and Brier score. These scores are generally similar, with the largest discrepancy occurring between the accuracy scores; 0.8098 vs 0.8531 for [21]. The difference in performance of the RF models may be due to the difference in training data. Ultimately, it is not possible to determine the exact cause.

Assuming the RF model from this study is representative of the methods from [21], the impact of the addition of the gradient features on performance can be observed. The RF model with additional gradient features performed marginally worse for the 0 hours prediction. This suggests that the gradient features do not improve the ability of a model to predict RFD at the current time of classification. Conversely, the 0 hours EBM models performed better with the gradient features than without them, according to all metrics. This suggests that the EBM model architecture is able to leverage more predictive power from gradient features, resulting in a slightly improved performance. However, the addition of the gradient features only improved the performance of the EBM by less than a single percentage point for each metric, suggesting that they have little influence to the performance of the model.

The addition of the gradient features to the EBM models only improves performance for the 0 hours ad 72 hours predictions, as seen in Table 7. For 24, 48 and 96 hours predictions, the EBM models performed better without the gradient features. In both cases, the better performing models only gain at most two percentage points on metrics from the addition or lack of the gradient features. This suggests that the gradient features provide little to no improvement to performance when predicting RFD at multiples of 24 hours into the future, up to 4 days.

For all models the performance decreases for each 24-hour increment into the future, across all metrics. In context this means the further into the future predictions are made, the less likely they are to be accurate. Figure 6 visualises this decrease in performance across the set of EBM with gradient feature models. From figure 6 it can be seen that the performances of models decrease almost linearly from 0 to 96 hours. It should be noted that these are discrete changes in performance between each of the separate EBM models, but are plotted as lines here to show the trends. It cannot be expected that this relationship would fit for models trained to predict RFD at times which are not multiples of 24 hours, since time of day effects would likely interfere with performance.

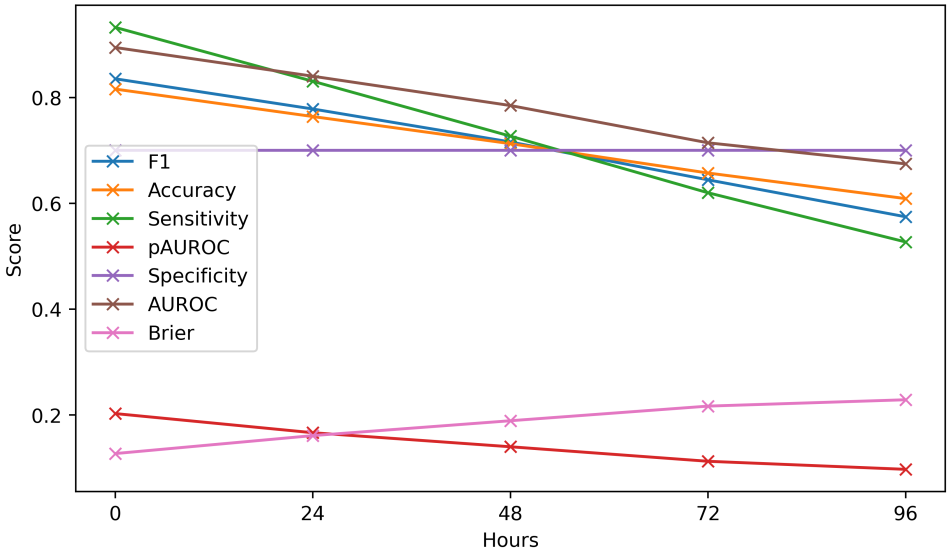

Figure 6: Performance metrics for each of the EBM with gradient features models predicting readiness for discharge at 0, 24, 48, 72 and 96 hours into the future. Scores are means over 100 random train-test splits. All scores other than AUROC and Brier are evaluated at a specificity of 0.7.
EBM, explainable boosting machine; RF, random forest; AUROC, area under the receiver operating characteristic; pAUROC, partial AUROC.

The further addition of the missingness feature to models with and without the gradient feature set yields little change to model performance (Table 9 in Appendix A.1). This suggests that informative missingness is not captured by this feature, or that it is not found to be important in model training. This may be a reflection of the nature of this dataset when used to predict readiness for discharge with EBM and RF. Alternatively, the proportion of missing entries on this feature set may be too simplistic for a feature indented to capture informative missingness. To reiterate, the missingness feature is excluded from the models in the main results (Table 7) as it is not suitable for deployment in clinical context, as described in Section 2.

Results for the models with the complete case data preparation are similar to that of the imputed data preparation (Table 10 in Appendix A.2). The missingness feature is excluded from the complete case results, as it has a value of zero for all admissions by definition. The best performing models between the imputed and complete case results have a gap in performance that generally widens for predictions further into the future. That is, the 0 hours and 24 hours results from the best models achieve similar performance between the imputed and complete case, whereas for the 48 hours, 72 hours and 96 hours models there is a greater discrepancy. The largest difference occurs for the 72 hours models, where the best F1 score for the imputed case is 0.654 as opposed to 0.605 for the complete case (3 s.f.).

Overall, these results suggest that predicting RFD at multiples of 24 hours into the future up to 4 days is possible with this methodology. There is little evidence that EBMs perform better than RFs at this task. The accuracy of predictions decrease the further into the future they are made. The results suggest that the addition of the gradient features did not garner improvements to overall model performance in general. This does not necessarily mean that the gradient features did not capture information with predictive power for readiness for discharge. Instead, it may mean that the predictive power of the gradient features is less than that of the base features.

5.2 Feature Importance

The fifteen most globally important features are shown for each of the EBM with gradient features models in Figure 7. These features are ranked by weighted mean absolute score across predictions on the cohort. Consistently, *gcs_MAX* and *gcs_MIN* are the two most important features, meaning they contribute most to the RFD predictions. With each 24 hour time step, the importance of the *gcs_MAX* and *gcs_MIN* features decreases relative to the rest of the features, but their rank stays the same. The *bun* (blood urea nitrogen) and *resp_MIN* (minimum respiration rate) features are frequently high ranking for importance across the models. *haemoglobin* ranks comparatively lowly for the 0-hour and 24-hour models, however it places in the top six most important features for the 48-hour, 72-hour and 96-hour models. This suggests that haemoglobin lab results are generally more useful when predicting further into the future, compared to the other features.

Some of the gradient-based features rank within the top fifteen for importance. This suggests that they do give predictive power, despite the fact that they don't generally improve overall model performance. Particularly, gradient features based on the GCS scores, heart rate and temperature readings are frequently important contributors. However, gradient features never rank in the top five most important, and rank in the top fifteen less frequently than the base features.

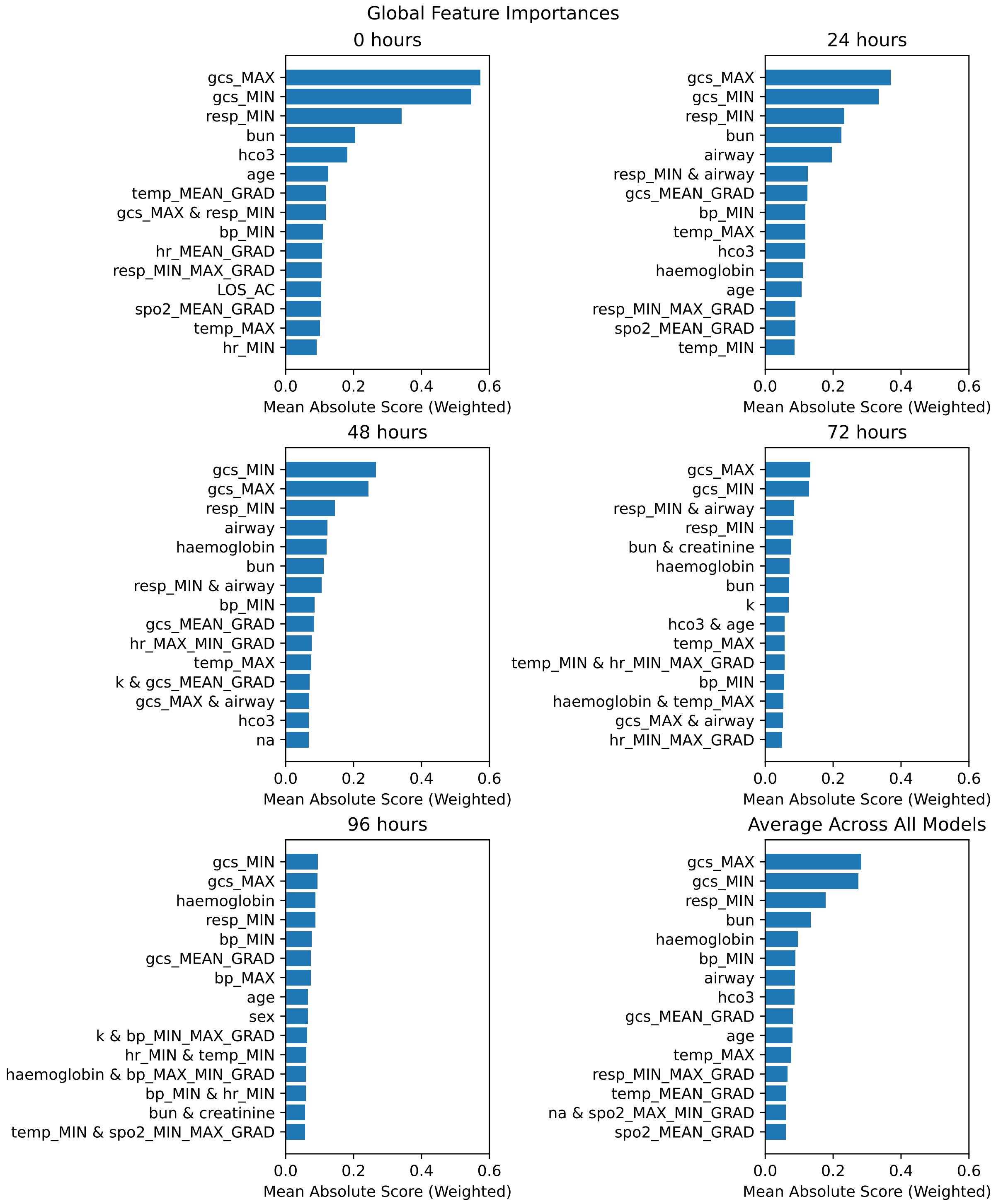

Figure 7: Global importances of single and pairwise features for prediction of readiness for discharge 0, 24, 48, 72 and 96 hours into the future using the EBM with gradient features models. The lower right plot is the average importance calculated as mean across the 5 models.

As discussed in Section 4.5, feature functions learnt by the EBM can be visualised and edited. Some examples of these feature functions are discussed here, with the intention of demonstrating how the explainability and editability of the EBM models is appropriate for deployment in clinical context.

Figure 8 shows an example of one such feature function for *gcs_max* from the 0 hour prediction EBM with gradient features model. The EBM has learnt that the lower a patient's maximum GCS score the less likely the patient is to be ready for discharge, as is to be expected from the definition of the GCS scoring system [33]. However, there is an anomaly to this trend that occurs between scores of 12 and 13. The feature function assigns a greater contribution to RFD (class 1) for a GCS score of 12 than it does for a GCS score of 13. Clinical expertise is likely to inform that this is erroneous. From the data perspective, this artifact may have arisen due to the fact that only a small number of instances of 12 scores were present in the training data compared to the neighbouring scores, as seen from the density plot. In the deployment of a tool that uses this model, this anomaly may be edited out if deemed appropriate with medical expertise. The error bars in Figure 8 are estimates of the uncertainty of the model for each region of the feature space [24]. The size of the error is sensitive to the amount of training data for each region, as indicated for the GCS score of 12.

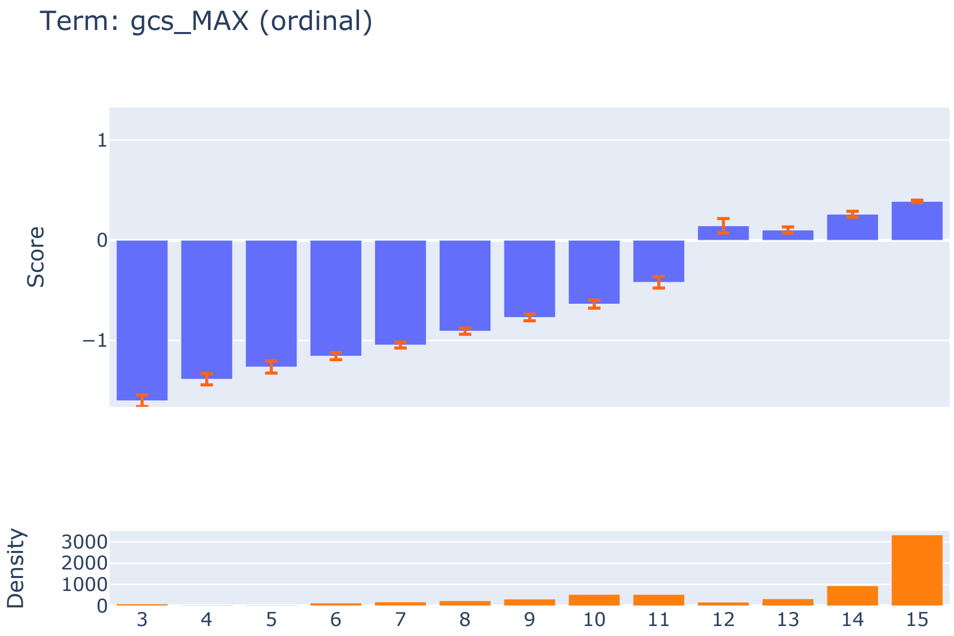

Figure 8: Feature function for the *gcs_MAX* feature learnt by the 0 hours EBM with gradient features model. Score represents the contribution of this feature to classifications from this model. The error bars represent an estimate of uncertainty in the score calculated by the EBM.

Figure 9 shows another example of a feature function from the same model but for a continuous variable; maximum blood pressure. The EBM has learnt to recognise what could be interpreted as a 'healthy' range of blood pressures, outside of which the likelihood that a patient is RFD decreases. There is a small notch in the function that might be considered an anomaly with clinical expertise. This notch occurs at a value of 90, so it may reflect the impact of measurements that were rounded when recorded in the dataset. Similarly, this feature function could be edited to smooth out this anomaly, if deemed appropriate in clinical context. Uncertainty is indicated by the width of the grey regions above and below the function.

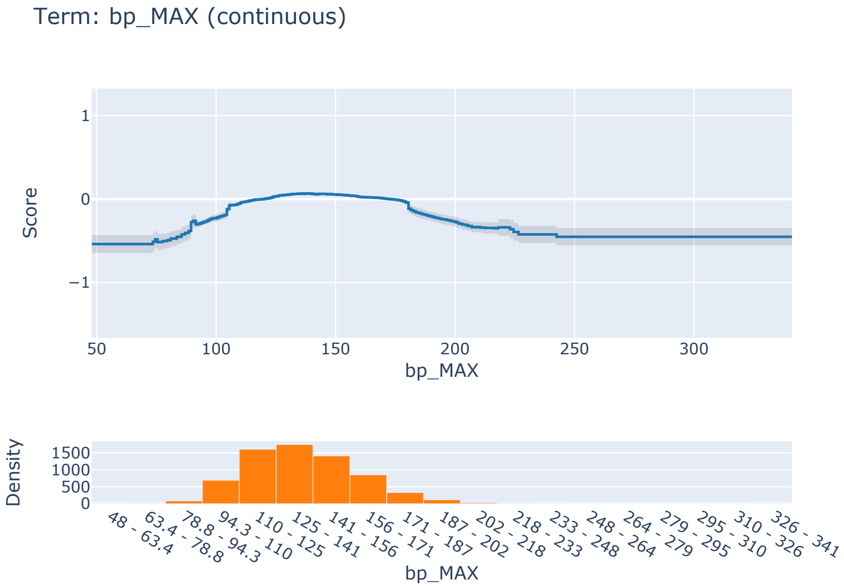

Figure 9: Feature function for the *bp_MAX* feature learnt by the 0 hours EBM with gradient features model. Score represents the contribution of this feature to classifications from this model. The grey error regions represent an estimate of uncertainty in the score calculated by the EBM.

An example of a pairwise interaction function from the 0-hours EBM model can be seen in Figure 10. The function takes the *hr_MEAN_GRAD* and *age* features as input. This interaction has been automatically detected by the EBM during training. In classification, this acts as a lookup table with 2 features as input [24]. For this example, if a patient's heart rate had a mean gradient of -20 BPM per hour from their initial to their current readings and they were 85 years old, this feature function would contribute a value of -0.468 (3 d.p.) to their classification. Again, this function may be edited if necessary.

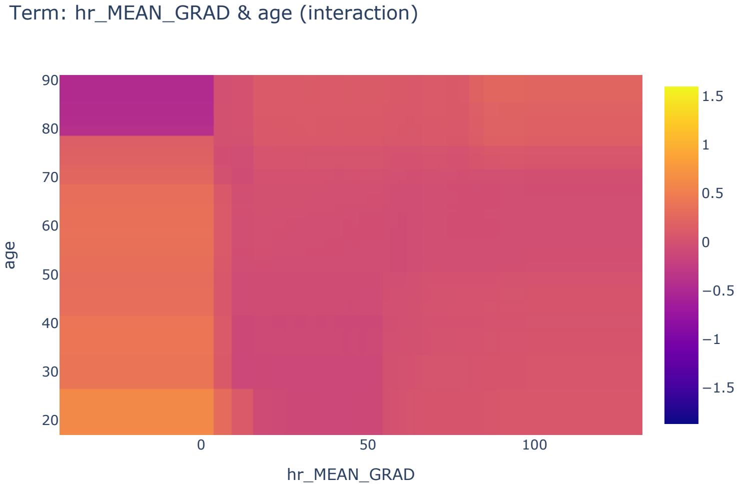

Figure 10: Pairwise interaction function between the *hr_MEAN_GRAD* and *age* features learnt by the 0 hours EBM model. The color represents the contribution of the combination of these features to classifications from this model.

5.3 Example of Predictions on a Patient

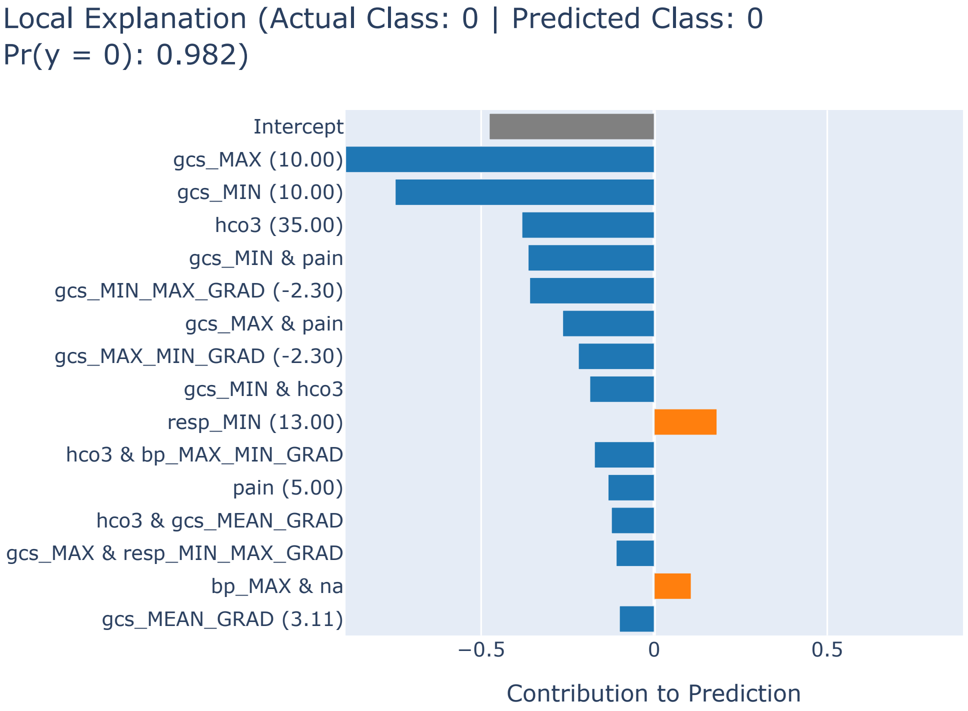

Figure 11: Local explanation of a prediction that a randomly selected patient is not ready for discharge. The importance of each feature to the predicted class label is shown, ranked by contribution. The estimated confidence in this prediction is denoted by Pr(y = 0).

Figure 11 shows an example of a single classification explanation by the 0 hours EBM with gradient features model for a random patient from the cohort. It shows which features contributed to the classification decision as well as a probability score which can be interpreted as the confidence in the prediction. There is a separate explanation like this for each of the models that make predictions for multiples of 24 hours into the future. For the bedside decision support tool, this would be the visualisation upon which an ICU doctor may evaluate their own confidence in the prediction.

The predictions of each model within the set of models that make up a composite tool are shown for an exemplary ICU admission in Table 8. The predicted outcomes and associated confidences for a single day in Table 8 represents the summary of the predictions of the decision support tool upon which an ICU doctor may schedule assessments. Over a timeline of several days, the ability of the tool to recognise when the patient will be RFD varies as the input data for each model varies. The patient is first ready for discharge 96 hours from the initial time, day 1. At day 1, the model does not correctly predict that the patient will be RFD in 96 hours time. For day 2, one day closer to the patient being RFD, the model correctly predicts RFD for 72 hours and 96 hours time, but with low confidence. By day 4, the model incorrectly predicts that the patient is RFD at the current point in time, with relatively high confidence (0.736).

		0 hours	24 hours	48 hours	72 hours	96 hours
	Actual	0	0	0	0	1
Day 1	Predicted	0	0	0	0	0
	confidence	0.999	0.985	0.861	0.673	0.593
	Actual	0	0	0	1	1
Day 2	Predicted	0	0	0	1	1
	confidence	0.982	0.850	0.588	0.583	0.582
	Actual	0	0	1	1	1
Day 3	Predicted	0	1	1	1	1
	confidence	0.923	0.537	0.690	0.707	0.674
	Actual	0	1	1	1	1
Day 4	Predicted	1	1	1	1	1
	confidence	0.736	0.756	0.647	0.840	0.633

Table 8: Predictions of readiness for discharge with confidence scores by the EBM with gradient features models for an example patient from the cohort. The predictions are made at 11:00am on days 1-4, where day 1 represents 1 full day since admission to the ICU. The outcomes are predicted 0, 24, 48 and 72 hours into to future from each current day.

6 Discussion

This study has proposed two expansions to a discharge decision support tool model that is based on an important method in the literature [21]. Firstly, the engineering of gradient features which aim to capture trajectories of recovery or deterioration in ICU patients. Secondly, the training of a series of similar models which enable prediction of readiness for discharge several days into the future. In this section, the achievements and limitations of these expansions as well as their implications in this field of research are discussed. The suitability for a bedside tool using these methods in deployment is analysed. Lastly, the limitations of the design choices in this study and the scope for further work are considered.

6.1 Gradient Features

The addition of gradient features did not garner overall improvements to the performance of the models on this dataset. This suggests that the rate of change in value for vital measurements between the initial and current periods of ICU admission may not be very informative to prediction of RFD status. The high predictive power that the results from [2] indicated was contained in the initial data did not translate to improved model performance through these gradient features. However, similarly to the 24-hour slope features in [29], the gradient features did capture information that had some predictive power. This was indicated by the presence of gradient features in the top 15 features ranked for global model performance.

These findings do not rule out the possibility for gradient-based features to significantly improve the accuracy of ICU outcome prediction models. There are many other ways to construct features that capture deterioration or recovery over time in ICU data. For example, the middle portion of ICU admissions was ignored in the creation of these gradient features, and the windows from which the initial and final data was sampled were restricted to 4 hours in length. Further exploration into these design choices may lead to the development of gradient features that achieve meaningful enhancements in the ability of ML models to predict readiness for discharge.

6.2 Predicting Future Outcomes

Predicting readiness for discharge for multiples of 24 hours into the future proved to be plausible, within a range of 4 days. The accuracy of predictions decreased the further into the future they were made. This is to be expected, as the most recent data available for a patient is likely to be more indicative of their condition at the current time then it is in the future. The overall performances of the models were not high by modern standards, particularly for the further forward looking predictions. However, this aspect of the study was a proof of concept that may inform further development of discharge decision support tools.

A strength of this concept is that it offers an indication of when a patient will be ready for discharge, rather than only whether they are at the current moment. It manages to do this without having to predict a time-based estimate as model output. It is ideal to avoid doing this since it is a much more complicated task, likely requiring many more considerations, extensive modelling and computational power. For example, to give a continuous estimate of time at which a patient is RFD using time-series modelling, time of day effects would have to be carefully modelled. For this study, time of day effects were elegantly mitigated by simply choosing a granularity of 24 hours between the individual prediction models. This limits the use of the models from this study to one prediction per day, at around 11am. However, this is appropriate given that discharge decisions are typically only made once per day in hospitals.

It is likely that the future prediction models were overfit to individuals, due to the low numbers of patients available in the dataset who were in intensive care for long enough to be eligible for resampling. The resampling method for class balancing further increased the likelihood that the same patients were resampled many times. This is a data availability issue, which may be overcome in further work by using a cohort from one or more larger datasets. The impact of having different amounts of data for each of the future prediction models to train and test on was not explored in this study. Due to this, it is not strictly possible to conclude that the decrease in accuracy for each multiple of 24 hours into the future is solely a reflection of the underlying patterns of informativeness in the data. It may be the case that the decrease in performance was caused by the decrease in data available for each model. Intuition suggests that it is likely a combination of both of these factors. Further work should aim to separate and inspect the impact of these two factors.

6.3 Suitability for Deployment

The explainable boosting machine architecture was greatly beneficial for this study to meet the requirements for potential model trust. The explainability of local predictions from the model was exemplified for a single patient. The interpretability of the models was demonstrated with the global feature importances and some examples of feature functions. Particularly, the fact that these feature functions are editable in a manner that is accessible to doctors and clinicians makes EBM highly suitable in the clinical context. This editability makes EBM a better choice than the other explainable ML methods that were discussed in the literature [14, 30].

While the explainability, interpretability and editability were demonstrated, the specific explanations from this study were not reviewed with clinical expertise. This means that the usefulness of the decision support tool models was not able to be measured beyond a comparison of performance metrics. To see if the explanations of the models were meaningful in the clinical context, they should be reviewed by ICU doctors. One such method is for doctors to classify a sample of patients as RFD or not using the same features that the models were trained on and compare this with the model output [38]. Furthermore, there was not the option to review and edit each of the feature functions with clinical expertise. Doing this would likely improve the generalisability of the models to other datasets and patient cohorts.

The addition of the gradient features is unlikely to have negatively impacted the explainability of the models. This is because their calculation process was intentionally simple such that the units of measurement were aimed to be still directly understandable in clinical context. The gradient features are given as change in measurement value per hour. The usefulness of these units to the decision process of doctors should be tested to further explore its impact on model trust.

The models in this study attempt to fit only a subset of patients. Pediatric patients were not available in the dataset, therefore they are excluded form the cohort. For future work, a separate tool for pediatric discharge decision support could be investigated. Patients who died in the ICU were excluded from the cohort, meaning that the models should not be expected to fit such patients. It cannot be known before the fact whether a patient will die in the ICU. Therefore it is not possible to know whether a given patient is expected to be fit by the model. This is detrimental to model trust, since the performance scores were evaluated on only patients who did not die in the ICU. To make the model more robust, in ICU mortality should be included. This could be as a separate outcome, making it a multiclass classification problem. A simpler method might be to include a separate mortality prediction model, such as the one made in [2]. The mortality model should be queried alongside the RFD models, to catch patients who are not fit by the RFD models due to the fact they are predicted to die in ICU.

The change in predictions over time from the set of models as shown in Table 8 may be useful to ICU doctors for a number of purposes. Discharge assessment scheduling as well as ward planning may benefit from these insights. There is also scope for the predictions of RFD status to influence more than discharge decisions alone. For example, an unexpected change in the future predictions from RFD to NRFD over the course of the admission may indicate that a patient has rapidly deteriorated since the previous day. This has potential to inform the decisions to administer treatments and interventions for ICU patient. Again, testing in clinical context is required to explore the usefulness of the decision support tool to these purposes.

At the performance level shown in Table 7 there will be numerous misclassifications of patients as RFD or not. Despite this, the tool can still be useful since it is intended for decision support rather than automated discharge. However, there is a danger that ICU doctors may become complacent and start to blindly trust model predictions, subconsciously or otherwise. This potential issue is rooted in the behavior of doctors, perhaps due to human nature. It is not necessarily possible to account for this at the level of modelling. Instead, the interaction between doctors and decision support tools must be studied before a model like this could be implemented.

6.4 Other Considerations

Glasgow Coma Scale score was found to be the most informative measure for readiness for discharge in this study. This further attests to its continued usefulness and relevance in the field of intensive care [34].

Only one data source was used for this study. As a result, the generalisability of these methods to other datasets and cohorts has not been investigated. Missingness was not shown to be clearly informative for this study. However, the mechanisms by which missing data occurs in the MIMIC-III dataset may be entirely different for other hospitals and their EHR. The comparison between the imputed and complete case results suggested that the imputation method did not have a large impact on the study. However, the implementation of a more sophisticated imputation method would likely improve the model performance to some degree [21].

This framework makes no effort to predict if or by how much a patient is overdue discharge. As discussed in Section 4, the assumption is made that patients in the dataset were called out either as soon as they were RFD or erroneously early, but not a significant time after the point at which they first became RFD. This is an important area for further work to validate how appropriate this assumption is, and whether using this assumption to balance classes via resampling introduces significant noise into the data.

Patient subtyping was not implemented in this study. Subtyping is likely to improve model performance by decreasing the need for each model to be generalisable to all types of patients [2,10]. There is scope for subtyping to further the usefulness of a decision support tool in bedside systems, so long as the method by which patients are subtyped is inductive and achievable in deployment.

7 Conclusion

This study has investigated the contribution of gradient-based vital measurement features to the ongoing field of ICU predictions with machine learning. It found that such features modelled on the difference of measurements from the initial to the relative current time points of ICU admissions did not improve the overall performance of prior readiness for discharge classification methods. However, some predictive information was captured by these gradient features, suggesting that they may be useful in further work. This study also expanded predictions of readiness for discharge to several days into the future, providing a simplistic and elegant way to avoid complicated time-series modelling and time of day effects when doing so. It was found that predicting readiness for discharge is possible for 24, 48, 72 and 96 hours into the future. The accuracy of predictions decreased the further into the future they were made. While the performances of models on these predictions were not comparatively high for modern ML standards, they may still be useful for their intended purpose of supporting the discharge decision making process of ICU doctors.

The methods used in this study were demonstrated to be compliant with requirements of explainability, such that there is potential for models to be reviewed and even edited by clinicians. This was achieved by using explainable boosting machines, and by maintaining clinically interpretable units of measurement for the novel gradient features. The clinical meaning of the model responses to features was not explored or validated in this study. This is highlighted as an important area for further work to increase generalisability of the models as well as build trust from doctors.

The study was limited by data availability, and was not validated against other data sources. Further testing is required to establish whether the methods are generalisable enough to be useful on other patient cohorts. The increasing availability of EHR from ICUs will further enable validation of methods like the ones from this study as well as others. Furthermore, it can be expected that the research area of ML methods for ICU predictions will continue to grow with the availability of data, with increasing scope for these findings to be built on.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Thea Barnes. Towards personalised patient risk prediction using routinely collected ward data. Unpublished thesis, Department of Engineering Mathematics and Technology - University of Bristol, 2023.

[3] Stijn Blot, Etienne Ruppé, Stephan Harbarth, et al. Healthcare-associated infections in adult intensive care unit patients: Changes in epidemiology, diagnosis, prevention and contributions of new technologies. *Intensive and Critical Care Nursing*, 70:103227, June 2022.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Richard V. Burkhauser and John Cawley. Beyond bmi: The value of more accurate measures of fatness and obesity in social science research. *Journal of Health Economics*, 27(2):519–529, 2008.

[6] The InterpretML Contributors. Interpretml documentation. `https://interpret.ml/docs/faq.html`. Accessed: 22/04/24.

[7] Dina Elreedy and Amir F. Atiya. A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, 505:32–64, 2019.

[8] Brian Fiani, Ryan Arthur Figueras, Patrick Samones, et al. Long-term intensive care unit (icu) stays can lead to long-term cognitive impairment (ltci): Neurosurgery nursing strategies to minimize risk. *Cureus*, September 2022.

[9] Ognjen Gajic, Michael Malinchoc, Thomas B. Comfere, et al. The stability and workload index for transfer score predicts unplanned intensive care unit patient readmission: Initial development and validation*. *Critical Care Medicine*, 36(3):676–682, March 2008.

[10] José A. González-Nóvoa, Laura Busto, Juan J. Rodríguez-Andina, José Fariña, Marta Segura, Vanesa Gómez, Dolores Vila, and César Veiga. Using explainable machine learning to improve intensive care unit alarm systems. *Sensors*, 21(21), 2021.

[11] Rolf H. H. Groenwold. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*, 4(1), July 2020.

[12] D M B Hall. What use is the bmi? *Archives of Disease in Childhood*, 91(4):283–286, January 2006.

[13] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

[14] Monica Hernandez, Ubaldo Ramon-Julvez, Elisa Vilades, Beatriz Cordon, Elvira Mayordomo, and Elena Garcia-Martin. Explainable artificial intelligence toward usable and trustworthy computer-aided diagnosis of multiple sclerosis from optical coherence tomography. *PLOS ONE*, 18(8):e0289495, August 2023.

[15] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1), April 2022.

[16] A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. Mimic-iv (version 1.0), 2020.

[17] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, 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.

[18] Gillian Knight. Nurse-led discharge from high dependency unit. *Nursing in Critical Care*, 8(2):56–61, April 2003.

[19] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD' 13. ACM, August 2013.

[20] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[21] Christopher J McWilliams, Daniel J Lawson, Raul Santos-Rodriguez, et al. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from mimic-iii and bristol, uk. *BMJ Open*, 9(3), 2019.

[22] Della Murbarani Prawidya Murti, Utomo Pujianto, Aji Prasetya Wibawa, and Muhammad Iqbal Akbar. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88, 2019.

[23] Mehdi Naseriparsa, Ahmed Al-Shammari, Ming Sheng, Yong Zhang, and Rui Zhou. Rsmote: improving classification performance over imbalanced medical datasets. *Health Information Science and Systems*, 8(1), June 2020.

[24] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability, 2019.

[25] Frank Q. Nuttall. Body mass index: Obesity, bmi, and health a critical review. *Nutrition Today*, 50(3):117–128, May 2015.

[26] Ziad Obermeyer, Jasmeet K Samra, and Sendhil Mullainathan. Individual differences in normal body temperature: longitudinal big data analysis of patient records. *BMJ*, page j5468, December 2017.

[27] Nadeem Qazi and Kamran Raza. Effect of feature selection, smote and under sampling on class imbalance classification. In *2012 UKSim 14th International Conference on Computer Modelling and Simulation*, pages 145–150, 2012.

[28] Kirsi Reini, Mats Fredrikson, and Anna Oscarsson. The prognostic value of the modified early warning score in critically ill patients: a prospective, observational study. *European Journal of Anaesthesiology*, 29(3):152–157, March 2012.

[29] Juan C. Rojas, Kyle A. Carey, Dana P. Edelson, et al. Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society*, 15(7):846–853, 2018. PMID: 29787309.

[30] Alessia Sarica, Andrea Quattrone, and Aldo Quattrone. Explainable boosting machine for predicting alzheimer's disease from mri hippocampal subfields. In Mufti Mahmud, M. Shamim Kaiser, Stefano Vassanelli, Qionghai Dai, and Ning Zhong, editors, *Brain Informatics*, pages 341–350, Cham, 2021. Springer International Publishing.

[31] Eline Stenwig, Giampiero Salvi, Pierluigi Salvo Rossi, and Nils Kristian Skjærvold. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Medical Research Methodology*, 22(1), February 2022.

[32] Kieran Stone, Reyer Zwiggelaar, Phil Jones, and Neil Mac Parthaláin. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4):e0000017, April 2022.

[33] Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness. *The Lancet*, 304(7872):81–84, July 1974.

[34] Graham Teasdale, Andrew Maas, Fiona Lecky, Geoffrey Manley, Nino Stocchetti, and Gordon Murray. The glasgow coma scale at 40 years: standing the test of time. *The Lancet Neurology*, 13(8):844–854, August 2014.

[35] Patrick J. Thoral, Mattia Fornasa, Daan P. de Bruin, et al. Explainable machine learning on amsterdamumcdb for icu discharge decision support: Uniting intensivists and data scientists. *Critical Care Explorations*, 3(9):e0529, September 2021.

[36] Peter Trinh, Donald R. Hoover, and Frank A. Sonnenberg. Time-of-day changes in physician clinical decision making: A retrospective study. *PLOS ONE*, 16(9):e0257500, September 2021.

[37] M H Vitaterna, J S Takahashi, and F W Turek. Overview of circadian rhythms. *Alcohol Res. Health*, 25(2):85–93, 2001.

[38] Enrico Werner, Jeffrey N. Clark, Ranjeet S. Bhamber, et al. *Identification, Explanation and Clinical Evaluation of Hospital Patient Subtypes*, pages 137–149. Springer Nature Switzerland, Cham, 2023.

[39] Bryan Williams. The national early warning score: from concept to nhs implementation. *Clinical Medicine*, 22(6):499–505, 2022.

A Additional Results

A.1 Missingness Feature

		0 hours	24 hours	48 hours	72 hours	96 hours
EBM + missingness feature	F1	**0.8364** (0.0068)	**0.7822** (0.0134)	0.7141 (0.0164)	0.6394 (0.0232)	0.5709 (0.0243)
	Accuracy	**0.8168** (0.0059)	**0.7676** (0.0108)	0.7109 (0.0115)	0.6544 (0.0164)	0.6075 (0.0168)
	Sensitivity	**0.9353** (0.0096)	**0.8375** (0.0209)	0.7253 (0.0234)	0.6120 (0.0319)	0.5212 (0.0345)
	pAUROC	**0.2029** (0.0038)	0.1692 (0.0073)	0.1351 (0.0080)	0.1131 (0.0070)	0.0933 (0.0052)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	**0.8950** (0.0045)	0.8445 (0.0093)	0.7778 (0.0116)	0.7173 (0.0159)	0.6657 (0.0113)
	Brier	**0.1263** (0.0033)	0.1588 (0.0050)	0.1915 (0.0048)	0.2156 (0.0064)	0.2322 (0.0046)
EBM + gradient features + missingness feature	F1	0.8344 (0.0062)	0.7816 (0.0116)	0.7096 (0.0153)	0.6361 (0.0312)	0.6060 (0.0217)
	Accuracy	0.8150 (0.0050)	0.7670 (0.0083)	0.7076 (0.0115)	0.6524 (0.0225)	0.6302 (0.0158)
	Sensitivity	0.9311 (0.0090)	0.8361 (0.0143)	0.7175 (0.0237)	0.6078 (0.0447)	0.5672 (0.0313)
	pAUROC	0.2013 (0.0066)	**0.1695** (0.0059)	0.1379 (0.0065)	**0.1135** (0.0083)	0.1085 (0.0100)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.8938 (0.0072)	**0.8459** (0.0077)	0.7788 (0.0119)	0.7179 (0.0201)	0.6793 (0.0205)
	Brier	0.1271 (0.0042)	**0.1583** (0.0042)	0.1919 (0.0055)	0.2166 (0.0084)	0.2282 (0.0078)
RF + missingness feature	F1	0.8252 (0.0068)	0.7781 (0.0060)	**0.7302** (0.0149)	0.6465 (0.0125)	0.5931 (0.0251)
	Accuracy	0.8062 (0.0057)	0.7639 (0.0054)	**0.7236** (0.0105)	0.6591 (0.0093)	0.6214 (0.0159)
	Sensitivity	0.9139 (0.0103)	0.8299 (0.0119)	**0.7514** (0.0209)	0.6218 (0.0183)	0.5510 (0.0344)
	pAUROC	0.1890 (0.0053)	0.1607 (0.0055)	**0.1441** (0.0034)	0.1121 (0.0048)	0.1028 (0.0089)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.8783 (0.0062)	0.8350 (0.0067)	**0.7934** (0.0061)	0.7195 (0.0105)	0.6807 (0.0176)
	Brier	0.1371 (0.0037)	0.1636 (0.0026)	**0.1860** (0.0025)	0.2142 (0.0037)	0.2256 (0.0043)
RF + gradient features + missingness feature	F1	0.8276 (0.0052)	0.7779 (0.0098)	0.7179 (0.0225)	**0.6476** (0.0211)	**0.6241** (0.0390)
	Accuracy	0.8084 (0.0045)	0.7636 (0.0083)	0.7141 (0.0176)	**0.6600** (0.0143)	**0.6433** (0.0254)
	Sensitivity	0.9183 (0.0084)	0.8299 (0.0183)	0.7314 (0.0351)	**0.6237** (0.0296)	**0.5926** (0.0509)
	pAUROC	0.1898 (0.0049)	0.1641 (0.0068)	0.1350 (0.0100)	0.1124 (0.0071)	**0.1085** (0.0162)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.8798 (0.0051)	0.8392 (0.0083)	0.7819 (0.0153)	**0.7212** (0.0144)	**0.6897** (0.0288)
	Brier	0.1355 (0.0025)	0.1615 (0.0033)	0.1895 (0.0059)	**0.2141** (0.0049)	**0.2242** (0.0077)

Table 9: Performance metrics for EBM and RF models with the missingness feature predicting readiness for discharge with and without additional gradient features at 0, 24, 48, 72 and 96 hours into the future. Scores are given as: mean (standard deviation) over 100 random train-test splits. All scores other than AUROC and Brier are evaluated at a specificity of 0.7. The best metric scores for each column (prediction time) are shown in bold.
EBM, explainable boosting machine; RF, random forest; AUROC, area under the receiver operating characteristic; pAUROC, partial AUROC.

A.2 Complete Case

		0 hours	24 hours	48 hours	72 hours	96 hours
EBM	F1	**0.8441** (0.0078)	**0.7829** (0.0142)	**0.6626** (0.0287)	**0.6049** (0.0448)	0.5052 (0.0566)
	Accuracy	**0.8220** (0.0059)	**0.7653** (0.0108)	**0.6718** (0.0181)	**0.6337** (0.0277)	0.5679 (0.0344)
	Sensitivity	**0.9564** (0.0133)	**0.8441** (0.0287)	**0.6527** (0.0333)	**0.5766** (0.0589)	0.4456 (0.0662)
	pAUROC	**0.2298** (0.0081)	0.1724 (0.0145)	**0.1224** (0.0101)	**0.1030** (0.0129)	0.0788 (0.0139)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	**0.9214** (0.0098)	**0.8435** (0.0170)	**0.7341** (0.0206)	**0.6891** (0.0278)	0.6238 (0.0203)
	Brier	**0.1059** (0.0089)	**0.1600** (0.0086)	**0.2110** (0.0091)	**0.2282** (0.0110)	0.2478 (0.0090)
EBM + gradient features	F1	0.8402 (0.0068)	0.7712 (0.0254)	0.6509 (0.0326)	0.5642 (0.0227)	0.5447 (0.0644)
	Accuracy	0.8185 (0.0066)	0.7567 (0.0200)	0.6625 (0.0171)	0.6062 (0.0183)	0.5887 (0.0415)
	Sensitivity	0.9467 (0.0187)	0.8196 (0.0386)	0.6374 (0.0332)	0.5215 (0.0288)	0.4982 (0.0795)
	pAUROC	0.2278 (0.0077)	**0.1770** (0.0071)	0.1189 (0.0121)	0.0948 (0.0115)	0.0847 (0.0127)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.9183 (0.0104)	0.8426 (0.0169)	0.7271 (0.0249)	0.6593 (0.0219)	0.6405 (0.0387)
	Brier	0.1110 (0.0099)	0.1636 (0.0108)	0.2170 (0.0116)	0.2418 (0.0117)	0.2489 (0.0181)
RF	F1	0.8330 (0.0111)	0.7605 (0.0134)	0.6504 (0.0329)	0.5741 (0.0324)	**0.5554** (0.0428)
	Accuracy	0.8099 (0.0097)	0.7470 (0.0107)	0.6630 (0.0275)	0.6117 (0.0256)	**0.5976** (0.0214)
	Sensitivity	0.9409 (0.0170)	0.8008 (0.0217)	0.6354 (0.0533)	0.5366 (0.0475)	**0.5054** (0.0450)
	pAUROC	0.2174 (0.0130)	0.1611 (0.0060)	0.1224 (0.0092)	0.0963 (0.0149)	**0.0907** (0.0088)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.9060 (0.0160)	0.8222 (0.0081)	0.7285 (0.0192)	0.6708 (0.0266)	**0.6487** (0.0182)
	Brier	0.1176 (0.0100)	0.1720 (0.0034)	0.2128 (0.0056)	0.2305 (0.0070)	**0.2352** (0.0047)
RF + gradient features	F1	0.8369 (0.0069)	0.7552 (0.0000)	0.6247 (0.0384)	0.5712 (0.0326)	0.5362 (0.0570)
	Accuracy	0.8159 (0.0055)	0.7407 (0.0000)	0.6460 (0.0233)	0.6099 (0.0234)	0.5873 (0.0292)
	Sensitivity	0.9369 (0.0122)	0.7811 (0.0000)	0.5979 (0.0486)	0.5326 (0.0440)	0.4819 (0.0640)
	pAUROC	0.2183 (0.0090)	0.1513 (0.0000)	0.1120 (0.0069)	0.0978 (0.0067)	0.0865 (0.0146)
	Specificity	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)	0.7000 (0.0000)
	AUROC	0.9053 (0.0115)	0.8121 (0.0000)	0.7106 (0.0248)	0.6582 (0.0258)	0.6453 (0.0256)
	Brier	0.1160 (0.0079)	0.1763 (0.0000)	0.2182 (0.0078)	0.2326 (0.0081)	0.2359 (0.0066)

Table 10: Performance metrics for complete case EBM and RF models predicting readiness for discharge with and without additional gradient features at 0, 24, 48, 72 and 96 hours into the future. Scores are given as: mean (standard deviation) over 100 random train-test splits. All scores other than AUROC and Brier are evaluated at a specificity of 0.7. The best metric scores for each column (prediction time) are shown in bold.

EBM, explainable boosting machine; RF, random forest; AUROC, area under the receiver operating characteristic; pAUROC, partial AUROC.