XMI-ICU: Explainable Machine Learning Model for Pseudo-Dynamic Prediction of Mortality in the ICU for Heart Attack Patients

Munib Mesinovic*[1], Peter Watkinson[2], Tingting Zhu[1]

Abstract

a) Objective: Heart attack remain one of the greatest contributors to mortality and these patients admitted to the intensive care unit (ICU) are at higher risk of death. In this study, we use two retrospective cohorts extracted from two US-based ICU databases, eICU and MIMIC-IV, to develop an explainable pseudo-dynamic machine learning framework for mortality prediction in the ICU.

b) Materials and Methods: The method provides accurate prediction for ICU patients up to 24 hours before the event and provides time-resolved interpretability. We compare standard supervised machine learning algorithms with novel tabular deep learning approaches and find that an integrated XGBoost model in our EHR time-series extraction framework (XMI-ICU) performs best. The framework was evaluated on a held-out test set from eICU and externally validated on the MIMIC-IV cohort using the most important features identified by time-resolved Shapley values.

c) Results: XMI-ICU or XMI achieved AUCs of 92.0 (balanced accuracy of 82.3) for 6-hour prediction of mortality. We show that it maintains reliable predictive performance over time in the ICU while also being externally validated in a separate patient cohort from MIMIC-IV without any previous training on that dataset. We also evaluated the model for clinical risk analysis by comparing it to the standard APACHE IV system in active use.

d) Conclusion and Future Work: We show that our framework successfully leverages time-series physiological measurements from ICU health records by translating them into stacked static prediction problems for mortality in heart attack patients and can offer clinical insight from time-resolved interpretability through the use of Shapley values.

*Corresponding author: Munib Mesinovic, Department of Engineering Science, Oxford, Turl Street, Oxford, UK; munib.mesinovic@jesus.ox.ac.uk

[1]Department of Engineering Science, University of Oxford, Oxford, UK

[2]Critical Care Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

Index Terms

explainability, intensive care, machine learning, myocardial infarction, prediction

I. INTRODUCTION

Acute myocardial infarction (AMI) or heart attack is one of the greatest contributors to cardiovascular deaths in the world whose incidence remains critically high with approximately every 40 seconds someone in the United States suffering an episode [1]. Cardiovascular diseases (CVDs) also represent a major cost burden globally with MI in the ICU being one of the most common CVD-related conditions in the critical care system [2]. In 2015, there were more than 18 million CVD-related deaths with MI accounting for over 15% of overall mortality and research showing that healthcare costs skyrocket with longer and more inefficient treatment in the ICU [3]–[5]. A considerable amount of previous work was concerned with the classification and diagnosis of MI in the ICU with measurements using ECG signals or subtypes of MRI, but due to the acute nature of the condition and its urgent need for immediate therapy, these proposals have done little to proactively forecast the disease prior to occurrence, a task of high clinical relevance [6]. Even the use of time-granular troponin assays, a biological marker for myocardial injury and thus infarction only helps with diagnosing the occurrence of an MI event faster but not with its prediction a priori [7]. Therefore, prediction and timely treatment of MI as well as its risk factors in a high-risk population such as previous survivors is urgently needed and will not just help treat these vulnerable patients but will also help streamline the costs and burdens of the critical care system.

Patients who exhibit MI are usually referred to the ICU, however, they are 10% more likely to suffer another episode in the days following and are at higher risk of death, especially the elderly [8]. Mortality prediction models can help design treatment plans and reduce costs and mortality rates but existing mortality prediction tools like the APACHE system deployed in US critical care centres have been criticised as too general and inaccurate for specific populations and diseases [9], [10]. One reason for machine learning's rise is its predictive performance compared to existing statistical and simple linear tools as well as its ability, in different cases and methods, of learning complex non-linear behaviours among variables [11]. When combined with interpretability methods, machine learning can be a useful tool for clinical guidance and

decision-making.

Deep learning has been the core focus of the research community as it has shown incredible success in imaging and text problems, including in healthcare [12]. The great advantage being that it does not require user-defined features and instead uses representational learning for tasks. Recent advances in tabular deep learning like TabNet and NODE have been a topic of lively conversation in the machine learning community with their interesting algorithmic and modular compositions but whether they can surpass classical machine learning models in different tasks is an ongoing debate [13], [14]. One of the drawbacks of such models is their opaqueness, lack of familiarity with tuning parameters, costs of training, and a dependency on a large amount of data being available. While deep learning models are the current standard in time-series EHR processing, we hope to show that by transforming the problem into connected and stacked static prediction problems, more reliable and low-cost models like extreme gradient boosted ensembles can be used instead and achieve superior performance to the deep learning alternatives.

Due to the recent nature of the proposed tabular deep learning models, research applying them to different healthcare challenges has been limited with only one recent paper looking at ICU mortality prediction with TabNet in COVID-19 patients specifically [15]. Prior work, particularly research that used deep learning, has largely limited the populations studied in the ICU settings to either general admission populations, acute kidney injury, or sepsis patients [16]–[18]. Other, more related work, has looked at survival and MI occurrence after ICU discharge, but again at a general patient population and without addressing the needs for prediction during ICU stay where the immediate risks of death for this population are considerable [19], [20].

It is, therefore, both of interest and need to propose a machine learning framework that can reliably predict negative outcomes for heart attack patients in the ICU, test it independently, validate it externally, and provide useful interpretability of its predictions for clinicians.

Our contributions are as follows:

1) We propose XMI-ICU, **X**GBoost for **M**yocardial **I**nfarction in the **ICU**, a novel gradient-boosted machine learning framework for the ICU that predicts mortality in MI patients, and that beats both existing prediction tools in active use as well as complex deep learning

models recently proposed for tabular data

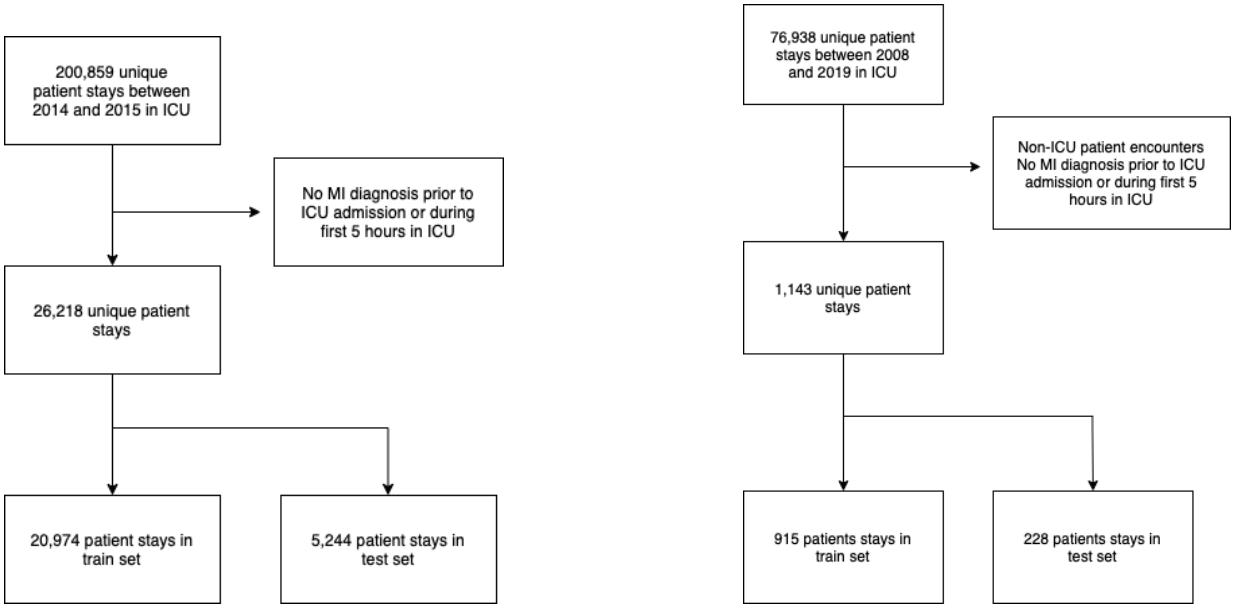
2) We evaluate our model the high-risk ICU population of MI survivors using multiple ICU centres and externally validated in a separate ICU cohort
3) We show the robustness of our model for varying time prediction in the ICU, including 6, 12, 18, and 24 hours in advance of the event of death
4) We investigate the clinical risk factors most informative to the prediction and stratify them across time of stay using time-varying Shapley values in the ICU to show how different clinical attributes indicate risk at different times prior to the event
5) We verify our model's clinical significance by a combination of interpretability methods like Shapley value analysis and clinical risk benefit and decision curves

II. MATERIALS AND METHODS

A. Study design and population

The data used in this study is the eICU Collaborative Research Database is a public database available upon request and fulfillment of ethical training [21]. The eICU database was processed using postgreSQL and the *pandas* package. eICU is a multi-center ICU database with over 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 to one of 335 ICUs at 208 hospitals located throughout the United States [21]. The database is de-identified and includes vital sign measurements, demographic data, and diagnosis information. For a full list of features used in our study please consult the relevant tables in the Supplementary Materials.

We based this study on the data preprocessing workflow used in [22], but adapted it to our problem accordingly. Our inclusion criteria were patients of age$>$18 and $<$89 years with an ICU length of stay of at least 5 hours to remove transient patients. We also include those with at least one recorded observation and excluded those without any laboratory measurements. Patients on respiratory support had a separate set of measurements which we included with a mechanical ventilation tag feature for this patient subgroup. We included variables present in at least 12.5% of patient stays, or 25% for lab variables due to their relative sparsity. We then removed those patients without any diagnosis information after 5 hours of stay because they might be inactive ICU patients logged for longer than was the case. A similar approach was taken by [23]. Our final subcohort consisted of 26,218 patients. We extracted diagnoses entered less than 5 hours



(a) Cohort selection for eICU database

(b) Cohort selection for MIMIC-IV database

Fig. 1: MI patient cohort selection. The exclusion criteria were listed here as they were implemented in PostgreSQL and Pandas. The final exclusion criteria is to extract the relevant subcohort at the end which is MI admitted patients to the ICU.

after entering the ICU and diagnoses prior to admission as starting diagnosis or first diagnosis. A flowchart of the patients cohort selection can be seen in Figure 1. The minimum length of stay changes depending on what model prediction time one enters into the framework. For example, evaluating the model for 6-hour or 24-hour prediction time means that the minimum length of stay for those patients would have to be 6 or 24 hours for there to be existent measurements for analysis.

The eICU database as well as many of the ICUs in the United States use the APACHE IV system for mortality risk prediction. The Acute Physiology, Age, and Chronic Health Evaluation (APACHE) IV system is a tool used to risk-adjust ICU patients which provides estimates of the probability that a patient dies given data from the first 24 hours [24]. The APACHE IV score is the result of a multivariate logistic regression which uses demographic, laboratory measurement, and diagnosis data to make a mortality risk assessment. It is the standard benchmark for mortality prediction tools in the ICU. Here we will use it both as a feature and as a benchmark as it was

highly important for us to evaluate the score's feature importance downstream in our models. We will provide XMI-ICU prediction performance for 24 hours which is the most directly comparable to APACHE-IV. APACHE-IV is only present in the eICU dataset.

We define our myocardial infarction tag as a collection of the diagnosis strings and respective ICD-10 codes described in the Supplementary Material. Once we have the defined outcome as MI, we are left with 26,218 samples that have had a diagnosis of MI prior to admission to the ICU with 3,139 (12.0%) having died during their stay.

We externally validated our model on the most recent release of Medical Information Mart for Intensive Care (MIMIC-IV v. 2.0, July 2022) which includes discharge information for over 15,000 additional ICU patients compared to the previous release [25]. Similar to eICU, MIMIC-IV is a de-identified and real world intensive care database using data from the Beth Israel Deaconess Medical Center for the years 2008 - 2019. We use similar cohort selection criteria as illustrated in Figure 1 and label definition as in eICU resulting in 1,143 unique patient ICU stays with confirmed MI out of 76,938. 131, or 12.0%, have died during their stay. Regarding missingness of variables, we mimic the steps taken for eICU processing.

The data processing of time-series and static variables was completed in Python. Patient cohort characteristics can be seen in Table I.

B. Machine learning methods

Following extraction of patients, we split the dataset into training and testing (20%) with the test set being used as hold-out for reporting only the final results. The training set was used for hyperparameter tuning of different machine learning and deep learning models using either Bayesian optimisation (to help reduce the overall computational costs of the framework) or the traditional grid search with 5-fold stratified cross validation. The validation scores in the results section represent the results of this cross validation. The next step in the framework is to pad the missing measurements for the time-windows using imputation with Multivariate Imputation by Chained Equation (MICE) and for feature standardisation or normalisation where necessary to avoid any data leakage either inside the validation folds or, at the end, the held-out test set with

TABLE I: Summary of demographics and variables used for external validation across training and testing datasets. MIMIC-IV has been used separately as an external validation source with the summary statistics for the entire dataset being a compound average of its train and test set statistics listed here individually.

	eICU (N = 26,218)		MIMIC-IV (N = 1,143)	
Attributes	**Train (N = 20,974)**	**Test (N = 5,244)**	**Train (N = 915)**	**Test (N = 228)**
Age (mean ± SD)	66.8 (± 12.7)	67.2 (± 12.4)	68.1 (± 13.2)	68.0 (± 13.1)
Sex (male)	13,369 (63.7%)	3,385 (64.5%)	585 (51.9%)	156 (55.4%)
LoS (days)	4.1 (± 2.7)	4.0 (± 2.3)	3.7 (± 2.9)	3.2 (± 3.1)
Lactate	2.9 (± 2.8)	2.5 (± 2.3)	2.0 (± 1.5)	1.9 (± 1.5)
SBP	120.2 (± 17.9)	120.0 (± 16.3)	126.3 (± 18.8)	124.5 (± 13.1)
Glucose	150.4 (± 61.7)	147.3 (± 56.7)	136.5 (± 49.3)	133.7 (± 45.1)
WBC	15.5 (± 10.5)	15.1 (± 9.3)	10.6 (± 7.4)	10.5 (± 7.4)
RDW	15.1 (± 2.2)	15.0 (± 2.0)	14.4 (± 2.1)	14.2 (± 2.0)
Urea Nitrogen	27.4 (± 19.5)	22.8 (± 13.4)	22.8 (± 17.0)	21.3 (± 14.6)
Bicarbonate	24.7 (± 4.2)	24.8 (± 4.4)	23.3 (± 3.1)	23.0 (± 3.0)
Mortality (dead)	2,511 (12.0%)	628 (12.0%)	105 (11.5%)	26 (11.3%)

the parameters extracted only on the training set or the training folds respectively [26]. Instead of using resampling techniques like SMOTE which can incur bias, we use inverse class-weighting in the training phase of the models which successfully allows it to generalise to an imbalanced prediction scenario [27]. Once the models were optimised, they were compared using their average validation set performance and finally their generalisation capability as evidenced by the test set metrics. The metrics used included Area-Under-Receiver-Operating-Curve (AUROC or AUC), Sensitivity, and Average Precision (AP) as they most completely capture the predictive performance of these binary classifiers even in cases of class imbalance. Details on how the metrics are calculated can be seen in the Supplementary Materials.

The XMI-ICU framework uses an extreme gradient-boosting approach with rolling time windows to extract the relevant features at defined times. This is a low-cost, time-efficient, imbalance-robust, and interpretable framework of dynamically predicting outcomes without relying on

complex transfomer models for time-series analysis. The benefits of transforming a time-series dynamic prediction into n-time-window static predictions for each time point are highlighted in the Supplementary.

For time-series measurements, we leverage the advantage of gradient boosting performance previously established on tabular data. We extracted summary features like the mean and standard deviation (to preserve the units necessary for later interpretability) for each patient stay for each feature for each time window. A time window is defined as all time-series measurements for the patient during the specific stay from the time since admission to the x-time prior to the event of interest. For example, for 24-hour prediction, we use all measurements since admission into the ICU stay until 24 hours prior to the event to model a prediction scenario. Using inverse class-weighting and a sliding time window, this framework enables users or clinicians to obtain estimates for risk of death at varying times in the future through dynamic feature extraction. For each time window, features are defined with means and standard deviations. For our experiments, the time-varying performance during different time window trials of 6, 12, 18, and 24 hours prior to death was evaluated, and Shapley values were used to ascertain interpretability of the model for each of the time windows and comment on clinical significance of risk factors (also evaluated over time periods) [28]. A flowchart visualising the proposed framework for mortality prediction in MI patients can be seen in Figure 2.

As far as the methodology of the deep learning models applied, TabNet and NODE, is concerned, more information can be found in the Supplementary Materials.

We used the same training, validation, and held-out test sets across all of our models with the hyperparameter search space included in the Supplementary Material with the selected best-performing parameters in bold. We use standard deviation to denote the variance in our validation results across the board. Our analysis was completed in Python 3.8 using Jupyter, pandas, numpy, SHAP, the original TabNet implementation, and the extension of the NODE implementation in PyTorch Tabular with some modifications. Decision curves, clinical risk calculations, and nomograms were computed and plotted in R.

C. Clinical risk analysis

To provide additional analysis of the model, we used clinical impact and decision curves in estimating the performance of the model at various risk thresholds. While decision curves are mostly used in cases of intervention effect on prognosis, they can also be used to diagnose the performance of predictive models albeit their adoption in machine learning has not been widespread, possibly due to applied machine learning work in healthcare being based more on advances in computer science rather than clinical significance. Decision curves account for both the benefits of higher risk estimation and the costs of overestimating risk to a patient who cannot benefit from the prediction. They are suggested to be an improvement over measures of performance such as AUROC. The intuition behind them is if a risk model tends to identify cases as high risk without falsely identifying too many negatives as high risk, then the net benefit of the risk model to the population will be positive [29]. A mathematical representation can be seen in the equation bellow:

$$NB_R = TPR_R P - \frac{R}{1-R} FPR_R (1-P) \tag{1}$$

Where NB is the net benefit, TPR and FPR are the positive rates, and P is the prevalence and R is the risk threshold respectively. They allow us to evaluate the models across a range of risk thresholds and observing tendencies of the model to overestimate risk. A clinical impact curve is simpler in that it displays the estimated number of people declared high-risk for each risk threshold, and visually displays the proportion of cases (true positives) [30].

III. RESULTS

A. eICU

Applying the framework proposed in Figure 2, we compare our proposed XMI-ICU gradient-boosted model to standard supervised learning methods. For some of the methods like support vector machines (SVMs), we have standardised features. All features listed in the supplementary section on data processing were used for the eICU test results while only the most important (top 8) features identified by Shapley values analysis used for external validation on MIMIC-IV. APACHE IV score was not used as a feature in these models, albeit experiments doing so are included in the Supplementary Materials for those curious. The first set of results in Table II

TABLE II: eICU validation (Val: Mean ± SD) and test prediction results for mortality prediction 6 hours in advance. Details on the metric computations can be found in the Supplementary Materials.

	Val AUC	AUC	Accuracy*	Average Precision
XMI-ICU	**91.8 ± 0.4**	**92.0**	**82.3**	**68.8**
TabNet	85.7 ± 2.1	84.1	77.0	60.7
TabNet (pretrained)	-	82.2	76.0	64.1
NODE	86.7 ± 0.7	85.4	67.6	62.3
Logistic Regression	90.2 ± 0.4	89.6	73.5	61.5
Random Forest	91.0 ± 0.5	90.6	78.2	64.4
SVM	90.2 ± 0.8	89.3	77.0	58.1
SVM (linear)	87.4 ± 0.7	87.7	78.8	63.8
LDA	79.4 ± 2.0	78.7	51.0	29.3

concerns prediction of mortality in MI patients with several hours prior to the event. It is clear that XMI-ICU maintains superior performance across all metrics for a priori prediction beating state-of-the-art tabular deep learning models. For AUROC and average precision, we evaluated the model at the default risk threshold in the results presented in the tables. Validation results include the mean and standard deviation of the 5 stratified folds. All XMI-ICU results have been checked for statistical significance (n=1000; $p<.001$). The results can also be seen visualised in Figure 4 which highlights the superior performance of XMI-ICU compared to alternative supervised learning models as measured by both AUROC and average precision.

After the XMI-ICU model was evaluated at 6 hour prediction prior to death, we extend to a more dynamic prediction evaluation by adapting the framework to arbitrarily predict the events of death at any time prior, and the framework will automatically extract, preprocess, standardise existing measurements, optimise respective hyperparameters, and deploy the XGBoost model for test prediction. The results for XMI-ICU evaluated at 6, 12, 18, and 24 hour prediction for

TABLE III: eICU test with all features, eICU test only using top 8 features, and MIMIC-IV external validation (Val: Mean ± SD) prediction results for mortality prediction stratified with time for XMI-ICU. External validation uses all eICU data as train set and MIMIC-IV data as test set with only the top 8 features included as identified by Shapley value analysis. Accuracy stands for balanced accuracy, details on the metric computations can be found in the Supplementary Materials.

	Val AUC	AUC	Accuracy*	Average Precision
eICU Mortality				
6 hours	91.8 ± 0.4	92.0	82.3	68.8
12 hours	90.5 ± 0.7	89.9	81.9	65.8
18 hours	89.1 ± 1.0	89.8	81.2	65.5
24 hours	87.7 ± 1.0	88.2	80.4	63.0
APACHE IV	-	69.9	69.3	31.5
Top-8 eICU Mortality				
6 hours	86.7 ± 1.1	86.2	80.0	74.7
12 hours	85.2 ± 1.2	83.3	77.0	69.7
18 hours	83.4 ± 1.3	83.1	76.5	65.8
24 hours	81.9 ± 1.4	81.2	75.2	59.2
External MIMIC-IV Mortality				
6 hours	-	80.0	77.7	73.8
12 hours	-	77.7	75.9	69.9
18 hours	-	76.6	75.1	67.8
24 hours	-	75.1	74.9	66.5

mortality in held-out test set of eICU can be seen in Table III and they continue to show reliable predictive performance across the different time windows. The table also includes the results for APACHE-IV as a matter of comparison for 24 hour prediction of mortality.

A plot showing the stability of predictive performance across different metrics for XMI-ICU as

a function of time in the ICU prior to death can be seen in Figure 4a. Figure 4a indicates stable performance for mortality prediction and its superiority compared to APACHE-IV (included for times beyond just 24 hours). The right figure shows the generalisation ability of the model to perform on a completely external test set which is MIMIC-IV using only the top 8 features from eICU for training. Evaluating a time-prediction model like XMI-ICU also requires showing coherent prediction across time and not just consistency of prediction accuracy and robustness. In the next set of results, we show XMI-ICU with low misclassification error across time for the same patient sample. A patient is deemed misclassified if they are predicted incorrectly at time x in advance when they have been previously predicted correctly at times $>$x. For example, a patient might be predicted to die at the 24 and 18-hour prediction windows correctly but at 12 hours in advance, they are predicted (incorrectly) to survive. These instabilities in prediction across time need to be measured if the model is to sustain reliable performance throughout the ICU stay. We define three patient subcohorts as illustrated at the top of Table IV where each indicates the group of patients correctly predicted at all previous time windows except one. The bottom of Table IV presents these results for both death and heart attack prediction indicating the low levels of misclassification most likely indicate sensitivity to noise rather than predictive weakness.

To understand how XMI-ICU is making these predictions and obtain further analysis for clinical significance testing, we applied Shapley value analysis on the held-out test set and observe relative feature importance. We did so across all time window prediction problems with the 6 hour example found in the Supplementary. We also subjected our interpretability results to random perturbation tests by adding a Gaussian distributed feature to the feature set to evaluate the susceptibility of change in the top variables identified and we observe no significant changes by introduction of noise variables. An example of this test is also included in the Supplementary.

We further stratify Shapley values as a function of time in the ICU for mortality prediction. A feature ranking at each time-point corresponds to the relative importance of that feature at that point in time in the ICU stay prior to the event in question. The time-graphs can be seen in Figure 5. These values were extracted for each of the time windows, in effect converting a static interpretability method to a dynamic explainability framework that shows how at different times closer to the event (death or heart attack) different values of features and their importance

TABLE IV: TOP: Defined patient cohorts for evaluating XMI-ICU predictive robustness across time windows. Each patient cohort corresponds to a grouping of patients who have been wrongly predicted at time x after being correctly predicted at all times before. BOTTOM: Misclassification rate (in percentage) is defined as number of wrong classifications divided by total patient sample present in cohorts for 6, 12, 18, and 24 hours prediction windows. A misclassification example is one where a patient is wrongly predicted in a time prediction window after being correctly predicted at previous windows.

Patient Cohort	24 hours	18 hours	12 hours	6 hours
P_1	✓	✓	✓	X
P_2	✓	✓	X	
P_3	✓	X		
	P_3	P_2	P_1	
eICU				
Mortality	7.9	8.2	5.5	
MIMIC-IV				
Mortality	6.4	6.3	4.7	

changes and how that is used by the model to learn underlying patters for disease outcome prediction.

B. External Validation: MIMIC-IV

We evaluated XMI-ICU on the separate and independent MIMIC-IV dataset for mortality prediction in MI patients. The features identified as most important by Shapley values analysis were used to create a new training set of the entirety of eICU and test on the entirety of MIMIC-IV cohorts only using the top 8 features whose statistical distributions for the different sets are included in Table I. XMI-ICU maintains high predictive performance across metrics when tested on this external dataset as can be seen in Table III without any training or tuning on it using only the top 8 features identified by Shapley value analysis from eICU test set. The results immediately above correspond to held-out test set performance for eICU using those same 8

features.

A plot showing predictive performance across different metrics for XMI-ICU evaluated on the MIMIC-IV cohort can be seen in the bottom Figure 4b. We also evaluate XMI-ICU for 6-hour prediction across subpopulations due to our multi-centre diverse dataset across sex and ethnicity demographics as a fair robustness check. The results can be seen in Table VIII in the Supplementary showing stable performance for XMI-ICU across different subcohorts for both eICU and MIMIC-IV held-out test sets.

C. Clinical Risk Benefit Analysis

To communicate the clinical significance of the XMI-ICU model results to clinicians, we evaluated our model with clinical impact curves (Figure 6a) and decision curve estimates (Figure 6b) for robust risk evaluation. A 90 percent confidence interval was derived with 50 bootstrap iterations on the test set. As the clinical impact curves for mortality show, XMI-ICU consistently identifies patients at risk across different risk thresholds showing robustness to false negatives. For those at highest risk ($>$75%), XMI-ICU has very low tendencies for false positives or ”over-risking” in its predictions, learning to focus on those most at risk with higher specificity and sensitivity. The decision curves indicate XMI-ICU’s approximated net benefit outperforming logistic regression (underlying model used in APACHE) using only top features identified from Shapley values analysis.

To assist clinicians more readily in their decision making process, the top features of our XMI-ICU model were used to construct a nomogram which is included in the Supplementary Materials and shows a simple representation of what could be an automatic calculator for 24-hour risk calculation in the ICU.

IV. DISCUSSION

Our proposed XMI-ICU model shows superior predictive performance for mortality prediction. Since our prediction time for mortality is at least 6 and up to 24 hours before the events but works for any arbitrary time that leaves clinicians with flexible extra time to prioritise high

risk patients and administer preventative measures. Interestingly, both TabNet and NODE do not achieve high performance when compared to XMI-ICU. These results directly contribute to the ongoing debate on the comparisons of tabular deep learning with classical methods which have shown mixed results over the last year in published research [31], [32]. We provide the first set of comparisons for these models in healthcare data known to be noisy, sparse, and presenting unique challenges. From our results it is clear that both complex and costly training and optimisation and longer deployment times combined with lackluster performance makes these tabular deep learning models currently incomparable to gradient-boosted methods like XMI-ICU.

XMI-ICU also beats the existing prediction tool in use across ICUs in the United States, APACHE IV, by 18.3% in test AUROC and 11.1% in test accuracy at 24-hour prediction. It requires miliseconds to be deployed once trained which also only takes a couple of seconds allowing for rapid response times in the ICU. Additionally, as Figure 4a shows, XMI-ICU maintains stable performance across all metrics during the 24 hours of ICU stay prior to death for MI patients. The model also successfully performs mortality prediction across different prediction time-windows in an external patient cohort obtained from MIMIC-IV using only the 8 most important features identified by Shapley values analysis on eICU as seen in Figure 4b. The drop in predictive performance compared to using MIMIC-IV as part of the training set with all features included is expected as we now use only 8 features without any training on the MIMIC-IV dataset itself. Despite this challenge, XMI-ICU maintains relatively high external predictive performance.

XMI-ICU combined with interpretability provides clinical risk factor importance which can aid physicians in both relying on the model but also investigating what aspects of the physiological measurements are more informative at what time during the ICU stay. For mortality prediction, we see that the closer a patient gets to highest risk of death close to the event, whether they are mechanically ventilated drops in importance compared to blood measurements like higher lactate, lower albumin, and systolic blood pressure. Hyperlactatemia has been found to be highly associated with in-hospital mortality in a relatively small and isolated heterogeneous ICU population [33]. Our findings on a much larger multi-centre patient cohort provide predictive evidence for this case. Previous research has established that lower serum albumin levels are good predictors of higher risk of death in ICU patients with sepsis and COVID-19 while our work seems to suggest a similar predictive pattern for heart attack patients as well [34], [35].

While this is currently a matter of debate we are glad to contribute to in medical sciences, lower albumin levels may be a marker of persistent injury to arteries and progression of atherosclerosis and thrombosis. The more time passes from low albumin levels the higher the risk of further acute injury in the myocardium which is what would make it useful for tracking risk of MI as our results suggest [36], [37].

Prior work has shed light on the hypothesis that hypotension as measured by the lowering of systolic blood pressure can be an indicator of higher risks of death in ICU patients, specifically those with acute kidney injury [38]. Some have suggested that myocardial injury is also more likely in cases of lower SBP values but here we provide early indication of the high prediction value of lower SBP levels for heart attack in the ICU [39]. The sudden rise in high glucose levels and variability (as captured by our standard deviation measure for blood glucose) being strong predictors of mortality have been confirmed with several retrospective cohort studies in the ICU [40], [41]. Work on prioritising those patients with such measurements and controlling for blood glucose and albumin can more easily be extended to preventive care for MI patients as well.

Comparing the framework to existing deep learning time-series models that tend to be costly and complex, our system with its simple embedded gradient boosted model sensitive to class imbalance and with dynamic feature extraction maintains prediction fidelity at varying time points while being faster, more interpretable, and less environmentally and financially costly to train and deploy. The XMI-ICU dynamic framework also offers an alternative to the rush in clinical machine learning in applying costly and less intepretable transformer and time-series models to these types of problems while still providing a dynamic prediction framework.

V. CONCLUSION

In conclusion, we developed a highly predictive machine learning framework that trains on time-series ICU ward data without requiring complex deep learning models. Instead, it relies on dynamic feature extraction and takes advantage of the predictive power of static models like XGBoost which outperformed other models including state-of-the-art tabular deep learning. The framework offers time-resolved interpretability that allows tracking changes in vital sign and blood measurement importance across the ICU stay for heart attack patients whose conclusions

seek to provide medical insight. The framework could be integrated into ICU systems to predict negative outcomes in heart attack patients with real-time patient measurements.

VI. AUTHOR CONTRIBUTIONS

MM, PW, and TZ designed the experiments, defined the patient criteria and outcomes. MM obtained access to, preprocessed, and cleaned the data; MM developed code for proposed model and conducted the experiments; MM, PW, and TZ contributed to the analyses of the data; MM and TZ wrote the manuscript and TZ provided overall supervision to MM.

VII. CONFLICT OF INTEREST STATEMENT

None declared.

VIII. DATA AVAILABILITY

The data is available from public access requests for eICU and MIMIC-IV. Accessing the data requires ethics module training and certification.

IX. FUNDING

M. Mesinovic appreciates the support of the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1) and the Rhodes Trust.

REFERENCES

[1] Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation. 2022;145(8):e153-639.

[2] Dégano IR, Salomaa V, Veronesi G, Ferriéres J, Kirchberger I, Laks T, et al. Twenty-five-year trends in myocardial infarction attack and mortality rates, and case-fatality, in six European populations. Heart. 2015;101(17):1413-21.

[3] Jayaraj JC, Davatyan K, Subramanian S, Priya J. Epidemiology of myocardial infarction. Myocard Infarct. 2019;3(10).

[4] Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. Journal of the American college of cardiology. 2017;70(1):1-25.

[5] Soekhlal R, Burgers L, Redekop W, Tan SS. Treatment costs of acute myocardial infarction in the Netherlands. Netherlands Heart Journal. 2013;21(5):230-5.

[6] Chen Z, Shi J, Pommier T, Cottin Y, Salomon M, Decourselle T, et al. Prediction of Myocardial Infarction From Patient Features With Machine Learning. Frontiers in cardiovascular medicine. 2022:346.

[7] Than MP, Pickering JW, Sandoval Y, Shah AS, Tsanas A, Apple FS, et al. Machine learning to predict the likelihood of acute myocardial infarction. Circulation. 2019;140(11):899-909.

[8] Nair R, Johnson M, Kravitz K, Huded C, Rajeswaran J, Anabila M, et al. Characteristics and outcomes of early recurrent myocardial infarction after acute myocardial infarction. Journal of the American Heart Association. 2021;10(16):e019270.

[9] Barrett LA, Payrovnaziri SN, Bian J, He Z. Building computational models to predict one-year mortality in ICU patients with acute myocardial infarction and post myocardial infarction syndrome. AMIA Summits on Translational Science Proceedings. 2019;2019:407.

[10] Venkataraman R, Gopichandran V, Ranganathan L, Rajagopal S, Abraham BK, Ramakrishnan N. Mortality prediction using acute physiology and chronic health evaluation II and acute physiology and chronic health evaluation IV scoring systems: Is there a difference? Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine. 2018;22(5):332.

[11] Mandair D, Tiwari P, Simon S, Colborn KL, Rosenberg MA. Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data. BMC medical informatics and decision making. 2020;20(1):1-10.

[12] Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health. 2019;1(6):e271-97.

[13] Joseph M. Pytorch tabular: A framework for deep learning with tabular data. arXiv preprint arXiv:210413638. 2021.

[14] Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems. 2021;34:18932-43.

[15] Nazir A, Ampadu HK. Interpretable deep learning for the prediction of ICU admission likelihood and mortality of COVID-19 patients. PeerJ Computer Science. 2022;8:e889.

[16] Shillan D, Sterne JA, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Critical care. 2019;23(1):1-11.

[17] Parreco J, Soe-Lin H, Parks JJ, Byerly S, Chatoor M, Buicko JL, et al. Comparing machine learning algorithms for predicting acute kidney injury. The American Surgeon. 2019;85(7):725-9.

[18] Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Frontiers in medicine. 2021;8:607952.

[19] Olsson De Capretz P, Bjorkelund A, Mokhtari A, Bjork J, Ohlsson M, Ekelund U, et al. Prediction of acute myocardial infarction or death in acute chest pain patients with machine learning models or first troponin T alone. European Heart Journal. 2021;42(Supplement_1):ehab724-3066.

[20] Law MR, Watt HC, Wald NJ. The underlying risk of death after myocardial infarction in the absence of treatment. Archives of internal medicine. 2002;162(21):2405-10.

[21] Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific data. 2018;5(1):1-13.

[22] Rocheteau E, Liò P, Hyland S. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In: Proceedings of the Conference on Health, Inference, and Learning; 2021. p. 58-68.

[23] Sheikhalishahi S, Balaraman V, Osmani V. Benchmarking machine learning models on multi-centre eICU critical care dataset. Plos one. 2020;15(7):e0235424.

[24] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Critical care medicine. 2006;34(5):1297-310.

[25] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. version 04) PhysioNet https://doi org/1013026/a3wn-hq05. 2020.

[26] Zhang Z. Missing data imputation: focusing on single imputation. Annals of translational medicine. 2016;4(1).

[27] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics. 2013;14(1):1-16.

[28] Ibrahim L, Mesinovic M, Yang KW, Eid MA. Explainable prediction of acute myocardial infarction using machine learning and shapley values. IEEE Access. 2020;8:210410-7.

[29] Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. Journal of Clinical Oncology. 2016;34(21):2534.

[30] Chen W, Yao M, Hu L, Zhang Y, Zhou Q, Ren H, et al. Development and validation of a clinical prediction model to estimate the risk of critical patients with COVID-19. Journal of Medical Virology. 2022;94(3):1104-14.

[31] Fayaz SA, Zaman M, Kaul S, Butt MA. Is Deep Learning on Tabular Data Enough? An Assessment. International Journal of Advanced Computer Science and Applications. 2022;13(4).

[32] Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Information Fusion. 2022;81:84-90.

[33] Van Beest PA, Brander L, Jansen S, Rommes JH, Kuiper MA, Spronk PE. Cumulative lactate and hospital mortality in ICU patients. Annals of intensive care. 2013;3(1):1-7.

[34] Kendall H, Abreu E, Cheng AL. Serum albumin trend is a predictor of mortality in ICU patients with sepsis. Biological research for nursing. 2019;21(3):237-44.

[35] Aziz M, Fatima R, Lee-Smith W, Assaly R. The association of low serum albumin level with severe COVID-19: a systematic review and meta-analysis. Critical Care. 2020;24(1):1-4.

[36] Djoussé L, Rothman KJ, Cupples LA, Levy D, Ellison RC. Serum albumin and risk of myocardial infarction and all-cause mortality in the Framingham Offspring Study. Circulation. 2002;106(23):2919-24.

[37] Oduncu V, Erkol A, Karabay CY, Kurt M, Akgün T, Bulut M, et al. The prognostic value of serum albumin levels on admission in patients with acute ST-segment elevation myocardial infarction undergoing a primary percutaneous coronary intervention. Coronary artery disease. 2013;24(2):88-94.

[38] Li-wei HL, Saeed M, Talmor D, Mark R, Malhotra A. Methods of blood pressure measurement in the ICU. Critical care medicine. 2013;41(1):34.

[39] Maheshwari K, Nathanson BH, Munson SH, Khangulov V, Stevens M, Badani H, et al. The relationship between ICU hypotension and in-hospital mortality and morbidity in septic patients. Intensive care medicine. 2018;44(6):857-67.

[40] Hermanides J, Vriesendorp TM, Bosman RJ, Zandstra DF, Hoekstra JB, DeVries JH. Glucose variability is associated with intensive care unit mortality. Critical care medicine. 2010;38(3):838-42.

[41] Gunst J, De Bruyn A, Van den Berghe G. Glucose control in the ICU. Current opinion in anaesthesiology. 2019;32(2):156.

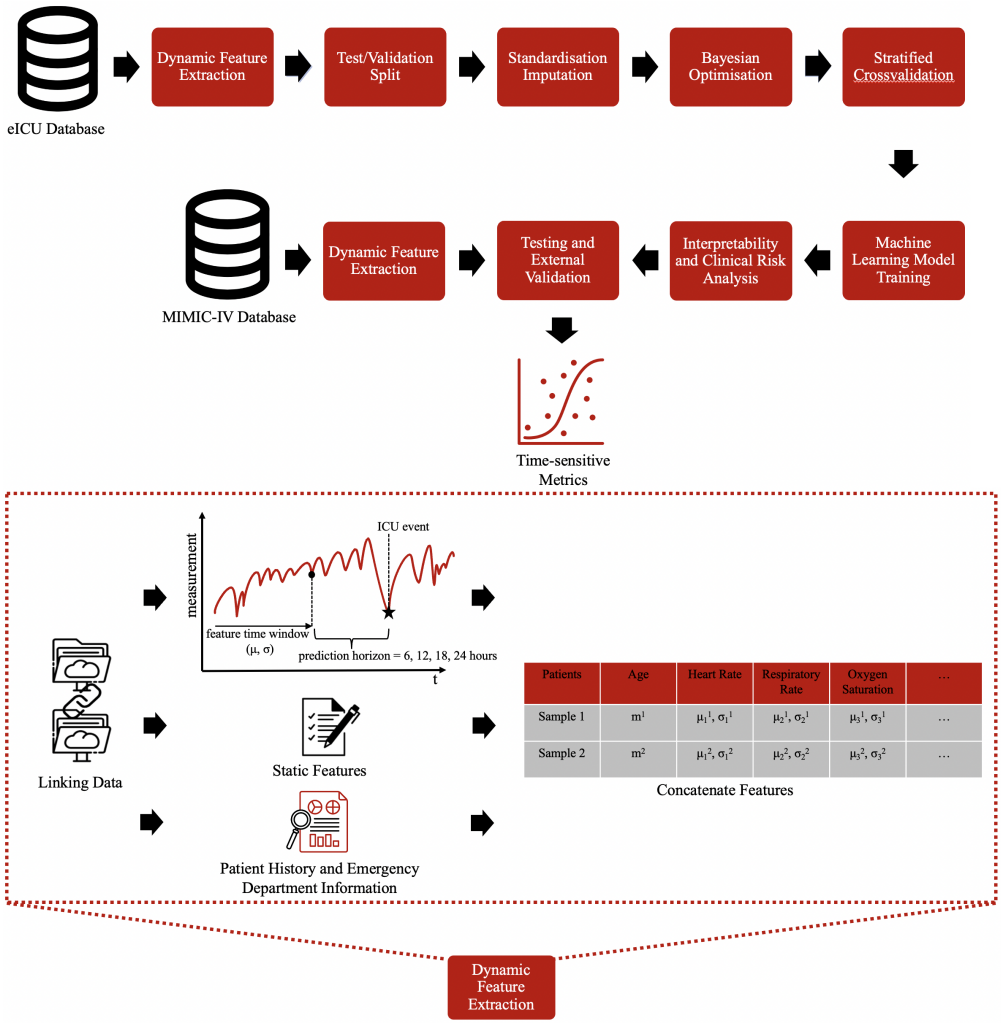
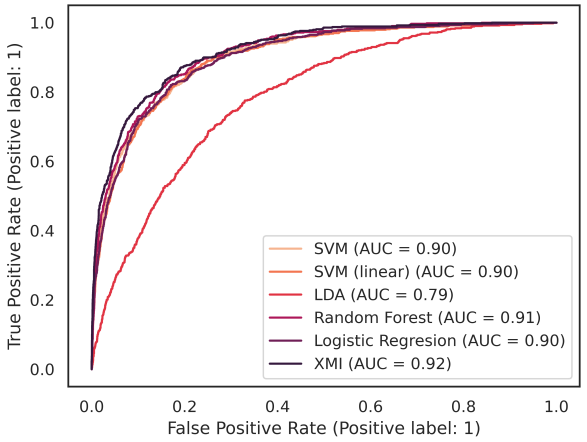
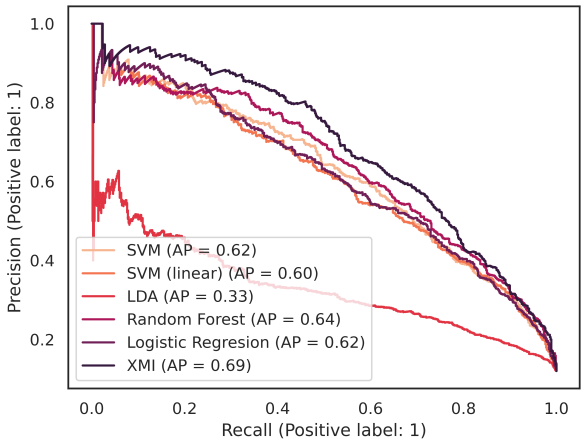


Fig. 2: Proposed XMI-ICU framework For dynamic mortality prediction in heart attack ICU Patients. The top part of the figure shows the dynamic feature extraction that links hospital-wide data including pre-admission information, ICU stay measurements, and emergency department variables. The sliding time windows change depending on the required prediction time and the time-series values are summarised using mean and standard deviations. For example, for 24-hour prediction, we use all time-series measurements since time of admission until 24 hours prior to the event as our feature time window to be summarised. The measurements are then concatenated with anamnesis, emergency department, and static variables to construct the feature matrix. The bottom half of the figure showcases the framework and how the dynamic feature extraction integrates with other components.

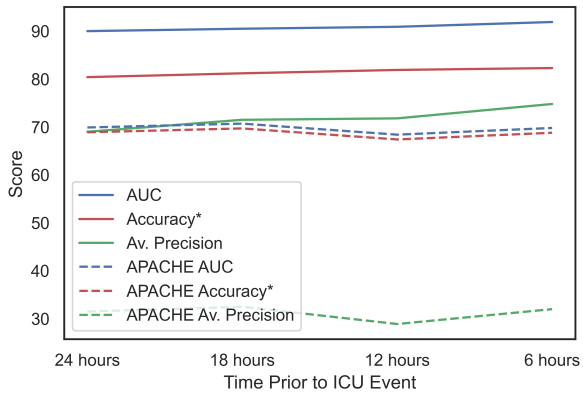


(a) AUROC performance of models for mortality prediction

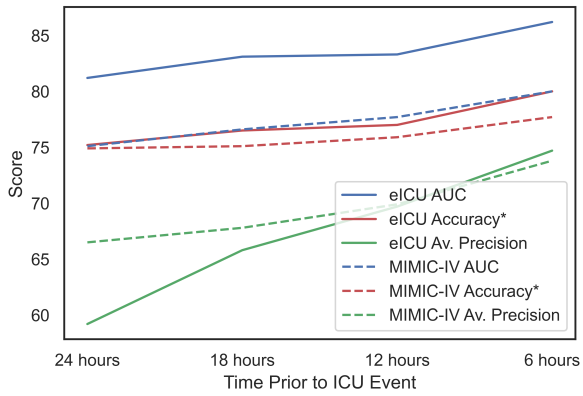


(b) AP of models for mortality prediction

Fig. 3: Evaluation performance of XMI-ICU to predict mortality 6 hours in advance compared to other models for different metrics on eICU held-out test. AUROC and average precision results are measured across prediction thresholds and they support the claim that XMI-ICU is a successful prediction model beyond default threshold values.



(a) XMI-ICU performance across time for mortality predictions on eICU held-out test set and APACHE performance (dotted)



(b) Performance across time for mortality prediction on eICU test set and external MIMIC-IV set (dotted) with only top 8 features

Fig. 4: Robustness and reliability of XMI-ICU prediction performance over time in the ICU for mortality prediction (left) using all features available in eICU and as measured by a variety of metrics. The right figure contains results from eICU held-out test set and MIMIC-IV external cohort with only the top 8 features identified by Shapley value analysis.

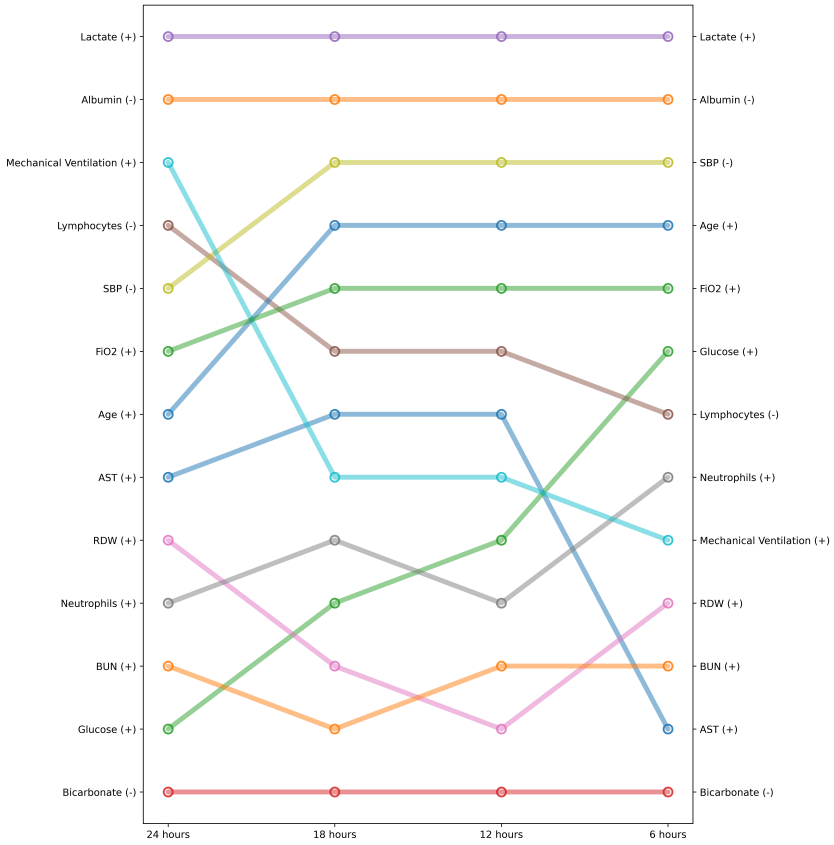
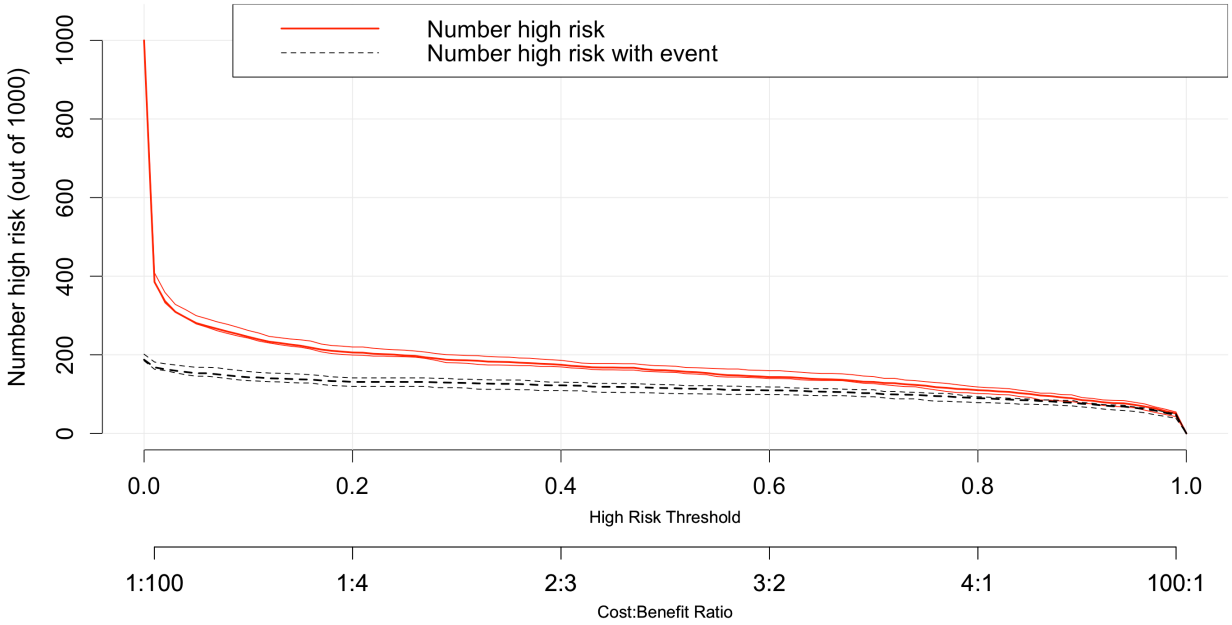
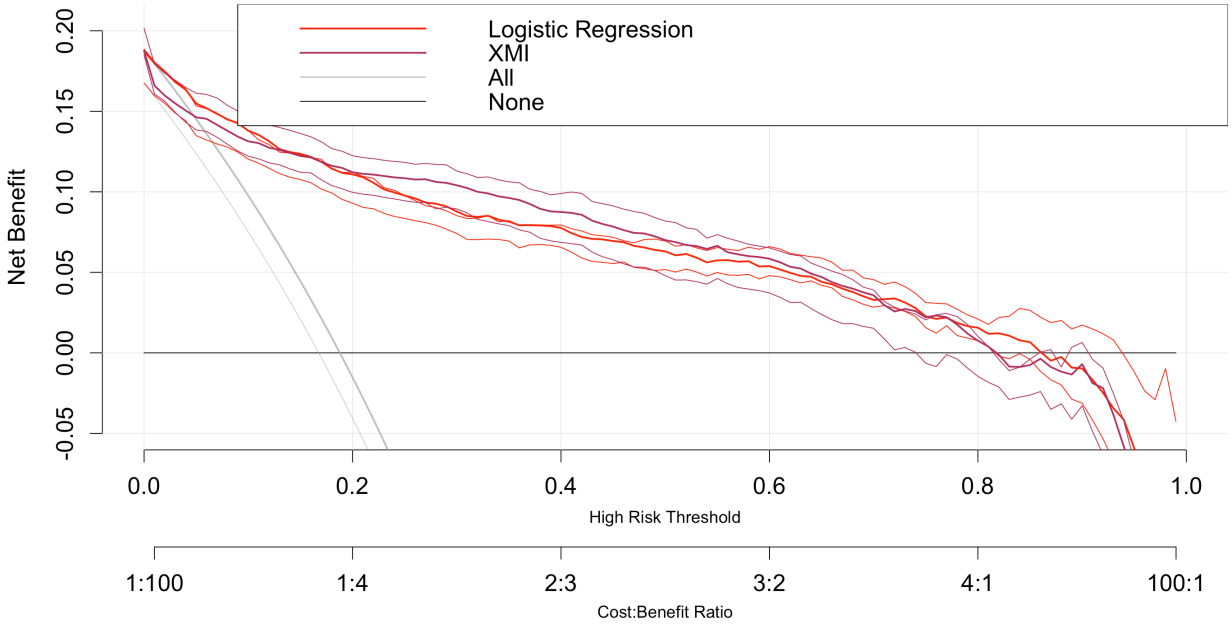


Fig. 5: Ranking of most important features as identified by their relative SHAP values for XMI-ICU prediction of mortality varied across time during ICU stay prior to event. For the time windows in the 6, 12, 18, 24 hour intervals, the top 13 features in each of the windows are presented as extracted from eICU thereby showcasing how the most important features for correct prediction of mortality changes through time or closer to the prediction event.



(a) The clinical impact curve measuring the risk predicted by XMI-ICU across different risk groups relative to the actual risk. For each risk threshold, we see the propensity of our prediction model to overestimate risk of death (prevalent behavior in machine learning risk prediction models) that helps us identify better risk thresholds for prediction.



(b) The decision curves comparing the net benefit of XMI-ICU to logistic regression (analog to APACHE IV) models across risk groups as defined by the risk thresholds. We see that the net benefit remains consistent with and is especially larger for XMI in those with moderate risk. The ”All” tag corresponds with the net benefit behaviour of having all patients predicted positive and ”None” with having no patients predicted positive.

Fig. 6: Clinical decision-making evaluation performance of XMI-ICU for mortality prediction using only the top 8 features on the entire eICU test set.