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

Anonymous authors
Paper under double-blind review

Abstract

Heart attack remain one of the greatest contributors to mortality in the United States and globally. Patients admitted to the intensive care unit (ICU) with diagnosed heart attack (myocardial infarction or MI) are more likely to suffer a secondary episode of MI and are at higher risk of death. In this study, we use two retrospective cohorts extracted from the eICU and MIMIC-IV databases, to develop a novel pseudo-dynamic machine learning framework for mortality and recurrent heart attack prediction in the ICU with interpretability and clinical risk analysis. The method provides accurate prediction of both outcomes for ICU patients up to 24 hours before the event and provide time-resolved interpretability results. The performance of the framework relying on extreme gradient boosting 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 achieving AUCs of 91.0 (balanced accuracy of 82.3) and 85.6 (balanced accuracy of 74.5) for 6-hour prediction of mortality and recurrent heart attack respectively. We show that our framework successfully leverages time-series physiological measurements by translating them into stacked static prediction problems to be robustly predictive through time in the ICU stay and can offer clinical insight from time-resolved intepretability.

1 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 Tsao et al. (2022). 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 Dégano et al. (2015). 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 Jayaraj et al. (2019); Roth et al. (2017); Soekhlal et al. (2013). 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 Chen et al. (2022b). 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 Than et al. (2019). 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 Nair et al. (2021). 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 Barrett et al. (2019); Venkataraman et al. (2018). Recent advances in tabular deep learning like TabNet and NODE have been a topic of lively conversation in the machine learning community but whether they can surpass classical machine learning models in different tasks is an ongoing debate Joseph (2021); Gorishniy et al. (2021). 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.

Here we present work done on two of the largest publicly available time-series electronic health records (EHR) datasets in the world which allow us to robustly train and test our models across a variety of ICUs across the United States. 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.

2 Methods

2.1 Study design and population

Full details on the data preprocessing can be found in the Appendix alongside a flowchart for patient cohort selection. 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 Zimmerman et al. (2006). 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. Details on MI outcome definition and patient cohort characteristics can be found in the Appendix under Data Description.

We externally validated our model on MIMIC-IV Johnson et al. (2020), 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 Appendix Figure 2 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. Due to lack of diagnosis and time annotation in clinical data collection, it was not possible to extract a label for recurrence of MI in this dataset. The data processing of time-series and static variables was completed in Python. Patient cohort characteristics can be seen in Appendix Table 5.

2.2 Machine learning methods

For details on the splits and hyperparamters, as well as metrics please consult the Appendix. We used Bayesian optimisation with inverse class weighting for the extreme gradient boosted decision trees to address class imbalance robustly and decrease optimisation costs. 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. A flowchart visualising the proposed framework for mortality and MI recurrence prediction in MI patients can be seen in Appendix Figure 3. It relies on dynamic feature extraction that links hospital-wide data with sliding time windows changing depending on the required prediction time and the time-series values being summarised using mean and standard deviations. The measurements are then concatenated with anamnesis, emergency department, and static variables to construct the feature matrix. Interpretability with Shapley values is then used to extract the most relevant features for external validation.

Table 1: eICU validation (Val: Mean ± SD) and test prediction results for secondary MI and mortality prediction 6 hours in advance. Details on the metric computations can be found in the Appendix Materials.

	AUC	Accuracy*	AP	AUC	Accuracy	AP
MI				**Mortality**		
XMI-ICU	**85.6**	**74.5**	**75.9**	**92.0**	**82.3**	**68.8**
TabNet	82.5	74.0	72.2	84.1	77.0	60.7
TabNet (pretrained)	82.9	72.8	71.8	82.2	76.0	64.1
NODE	74.6	74.0	62.1	85.4	67.6	62.3
Logistic Regression	74.6	67.9	54.0	89.6	73.5	61.5
Random Forest	82.2	64.0	68.6	90.6	78.2	64.4
SVM	76.4	72.4	56.8	89.3	77.0	58.1
SVM (linear)	74.9	74.0	51.7	87.7	78.8	63.8
LDA	65.8	50.6	37.0	78.7	51.0	29.3

3 RESULTS

3.1 eICU

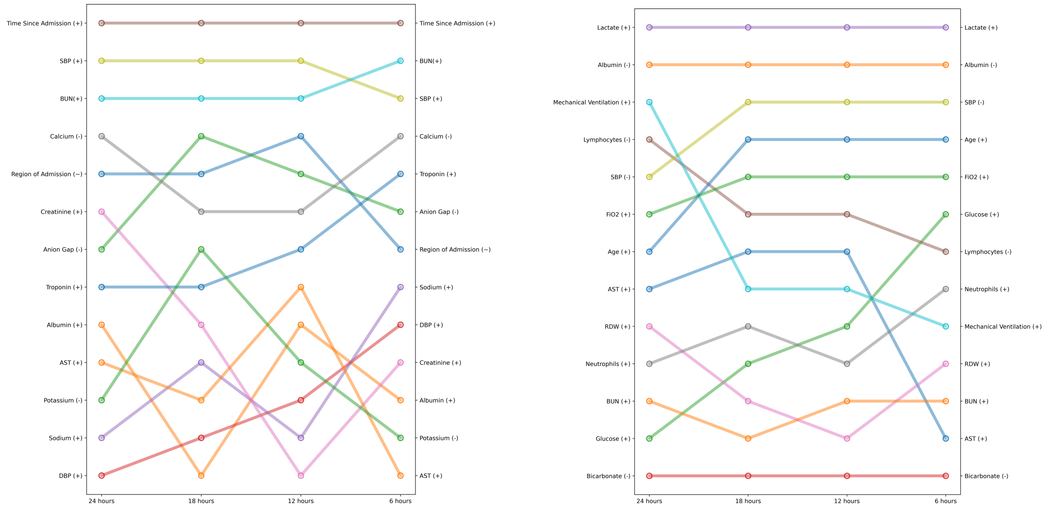
Applying the framework proposed in Appendix Figure 3, we compare our proposed XMI-ICU gradient-boosted model to listed alternatives. The first set of results concerning comparisons to other models including deep learning alternatives are in Table 1. All XMI-ICU results have been checked for statistical significance (n=1000; p<.001).

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 and secondary heart attack at any time prior. The results for XMI-ICU evaluated at 6, 12, 18, and 24 hour prediction for secondary MI and mortality in held-out test set of eICU can be seen in Appendix Table 9 and they continue to show reliable predictive performance across the different time windows. We also show XMI-ICU with low misclassification error across time for the same patient sample to check for temporal coherency. 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. Details on these results are included in the Appendix under ”Time-robustness Checks.”

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 further stratify Shapley values as a function of time in the ICU for mortality and secondary MI prediction. The time-graphs can be seen in Figure 1. 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 changes and how that is used by the model to learn underlying patters for disease outcome prediction.

3.2 EXTERNAL VALIDATION: MIMIC-IV

We evaluated XMI-ICU on the separate and independent MIMIC-IV dataset for mortality prediction in MI patients. XMI-ICU maintains high predictive performance across metrics when tested on this external dataset as can be seen in Appendix Table 9 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 Appendix Figure 6c. We also evaluate XMI-ICU for 6-hour prediction across subpopulations due to our multi-centre diverse dataset across sex and ethnicity



(a) Importance of clinical variables for secondary MI prediction across patient ICU stay

(b) Importance of clinical variables for mortality prediction across patient ICU stay

Figure 1: Ranking of most important features as identified by their relative SHAP values for XMI-ICU prediction of secondary MI and 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.

demographics as a fair robustness check. The results can be seen in Appendix Table 13 showing stable performance for XMI-ICU across different subcohorts for both eICU and MIMIC-IV held-out test sets.

4 Discussion

Our proposed XMI-ICU model shows superior predictive performance across both tasks compared to TabNet and NODE. 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 Fayaz et al. (2022); Shwartz-Ziv & Armon (2022). 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. Additionally, XMI-ICU maintains stable performance across all metrics during the 24 hours of ICU stay prior to secondary heart attack and 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.

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. 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.

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 outperforms 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.

REFERENCES

Laura A Barrett, Seyedeh Neelufar Payrovnaziri, Jiang Bian, and Zhe He. 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: 407, 2019.

Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14(1):1–16, 2013.

Wenyu Chen, Ming Yao, Lin Hu, Ye Zhang, Qinghe Zhou, Hongwei Ren, Yanbao Sun, Ming Zhang, and Yufen Xu. Development and validation of a clinical prediction model to estimate the risk of critical patients with covid-19. *Journal of Medical Virology*, 94(3):1104–1114, 2022a.

Zhihao Chen, Jixi Shi, Thibaut Pommier, Yves Cottin, Michel Salomon, Thomas Decourselle, Alain Lalande, and Raphaël Couturier. Prediction of myocardial infarction from patient features with machine learning. *Frontiers in cardiovascular medicine*, pp. 346, 2022b.

Irene R Dégano, Veikko Salomaa, Giovanni Veronesi, Jean Ferriéres, Inge Kirchberger, Toivo Laks, Aki S Havulinna, Jean-Bernard Ruidavets, Marco M Ferrario, Christa Meisinger, et al. Twenty-five-year trends in myocardial infarction attack and mortality rates, and case-fatality, in six european populations. *Heart*, 101(17):1413–1421, 2015.

Sheikh Amir Fayaz, Majid Zaman, Sameer Kaul, and Muheet Ahmed Butt. Is deep learning on tabular data enough? an assessment. *International Journal of Advanced Computer Science and Applications*, 13(4), 2022.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

Lujain Ibrahim, Munib Mesinovic, Kai-Wen Yang, and Mohamad A Eid. Explainable prediction of acute myocardial infarction using machine learning and shapley values. *IEEE Access*, 8:210410–210417, 2020.

Joshua Chadwick Jayaraj, Karapet Davatyan, SS Subramanian, and Jemmi Priya. Epidemiology of myocardial infarction. *Myocard. Infarct*, 3(10), 2019.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *version 0.4). PhysioNet. https://doi. org/10.13026/a3wn-hq05*, 2020.

Manu Joseph. Pytorch tabular: A framework for deep learning with tabular data. *arXiv preprint arXiv:2104.13638*, 2021.

Kathleen F Kerr, Marshall D Brown, Kehao Zhu, and Holly Janes. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *Journal of Clinical Oncology*, 34(21):2534, 2016.

Raunak Nair, Michael Johnson, Kathleen Kravitz, Chetan Huded, Jeevanantham Rajeswaran, Moses Anabila, Eugene Blackstone, Venu Menon, A Michael Lincoff, Samir Kapadia, et al. Characteristics and outcomes of early recurrent myocardial infarction after acute myocardial infarction. *Journal of the American Heart Association*, 10(16):e019270, 2021.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Emma Rocheteau, Pietro Liò, and Stephanie Hyland. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 58–68, 2021.

Gregory A Roth, Catherine Johnson, Amanuel Abajobir, Foad Abd-Allah, Semaw Ferede Abera, Gebre Abyu, Muktar Ahmed, Baran Aksut, Tahiya Alam, Khurshid Alam, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American college of cardiology*, 70(1):1–25, 2017.

Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

RR Soekhlal, LT Burgers, WK Redekop, and Siok Swan Tan. Treatment costs of acute myocardial infarction in the netherlands. *Netherlands Heart Journal*, 21(5):230–235, 2013.

Martin P Than, John W Pickering, Yader Sandoval, Anoop SV Shah, Athanasios Tsanas, Fred S Apple, Stefan Blankenberg, Louise Cullen, Christian Mueller, Johannes T Neumann, et al. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation*, 140(11):899–909, 2019.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Alvaro Alonso, Andrea Z Beaton, Marcio S Bittencourt, Amelia K Boehme, Alfred E Buxton, April P Carson, Yvonne Commodore-Mensah, et al. Heart disease and stroke statistics—2022 update: a report from the american heart association. *Circulation*, 145(8):e153–e639, 2022.

Ramesh Venkataraman, Vijayaprasad Gopichandran, Lakshmi Ranganathan, Senthilkumar Rajagopal, Babu K Abraham, and Nagarajan Ramakrishnan. 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*, 22(5):332, 2018.

Zhongheng Zhang. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(1), 2016.

Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.

Kelly H Zou, A James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657, 2007.

A Appendix

B Libraries and Packages

Statistical and clinical analyses were performed using R (version 3.6.2, R Foundation for Statistical Computing, Vienna, Austria) with packages including binom, Epi, ggplot2, lme4, sjstats, tableone, and tidyverse. Machine learning components were coded using Python (version 3.8.0) with packages including imblearn, matplotlib, skopt, xgboost, seaborn, shap, pandas, numpy, and sklearn.

We open-source our code at the following link: .

B.1 Data Description

The data used in this study is the eICU Collaborative Research Database is a public database available upon request and fulfillment of ethical training Pollard et al. (2018). 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 Pollard et al. (2018). 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 Appendix Materials.

We based this study on the data preprocessing workflow used in Rocheteau et al. (2021), but adapted it to our problem accordingly. Our inclusion criteria were patients of age>18 and <89 years with

Table 2: Diagnoses Taken for MI Outcome Definition

Diagnosis String
cardiovascular—chest pain / ASHD—acute coronary syndrome
ASHD—acute coronary syndrome—acute myocardial infarction (no ST elevation)
ASHD—acute coronary syndrome—acute myocardial infarction (with ST elevation)
ASHD—acute coronary syndrome—s/p PTCA / myocardial infarction
ASHD—coronary artery disease / myocardial infarction
ASHD—coronary artery disease—known / myocardial infarction
Acute MI location
Acute MI location—inferior
Acute MI location—non-Q
Non-operative—Diagnosis—Cardiovascular—Infarction, acute myocardial (MI)
Cardiovascular (R)—Myocardial Infarction
Cardiovascular (R)—Myocardial Infarction—MI - date unknown
Cardiovascular (R)—Myocardial Infarction—MI - remote
Cardiovascular (R)—Myocardial Infarction—MI - within 6 months

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 Sheikhalishahi et al. (2020). Our final subcohort consisted of 26,218 patients. We extracted diagnoses entered less than 5 hours after entering the ICU and diagnoses prior to admission as starting diagnosis or first diagnosis. Secondary diagnoses are those logged in 5 hours after being admitted to the ICU. A flowchart of the patients cohort selection can be seen in Appendix Figure 2.

We defined our outcome of interest using the most common diagnosis strings associated with the myocardial infarction diagnosis as can be seen in Table 2 below.

A detailed list of features used in the study and extracted from eICU and MIMIC-IV can be seen in Tables 3 and 4.

B.2 Machine Learning Setup

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 with Bayesian optimisation. 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 the parameters extracted only on the training set or the training folds respectively Zhang (2016). 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 Blagus & Lusa (2013). 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 Appendix Materials.

Table 3: Features extracted from the eICU database. The features include demographic data collected for all patients, ICU unit-specific information like type and number of beds, hospital information like regional location and teaching status, vital signs including respiratory rate and blood pressure, and biochemical measurements including troponin and levels of potassium and protein in blood.

Feature	**Type**	**Feature**	**Type**
Sex	binary	-basos	continuous
Age	integer	-eos	continuous
APACHE IV Score	continuous	SBP	continuous
Time Since Admission	continuous	DBP	continuous
Hour of Admission	integer	-lymphs	continuous
Height	continuous	-monos	continuous
Weight	continuous	-polys	continuous
Ethnicity	categorical	ALT	continuous
Unit Type	categorical	AST	continuous
Unit Admit Source	categorical	BUN	continuous
Unit Visit Number	categorical	Base Excess	continuous
Unit Stay Type	categorical	FiO2	continuous
Num Beds Category	categorical	HCO3	continuous
Region	categorical	Hct	continuous
Teaching Status	binary	Hgb	continuous
Physician Speciality	categorical	MCH	continuous
Unit Type	categorical	MCHC	continuous
Mechanical Ventilation	binary	MCV	continuous
Time-series (summary features)			
O2 Sat (%)	continuous	MPV	continuous
PT-INR	continuous	PT	continuous
RBC	continuous	PTT	continuous
RDW	continuous	WBC	continuous
Alkaline ph.	continuous	Albumin	continuous
Bedside Glucose	continuous	Anion Gap	continuous
Calcium	continuous	Bicarbonate	continuous
Creatinine	continuous	Glucose	continuous
Lactate	continuous	Magnesium	continuous
pH	continuous	paCO2	continuous
paO2	continuous	Phosphate	continuous
Platelets	continuous	Potassium	continuous
Sodium	continuous	Bilirubin	continuous
Protein	continuous	Troponin - I	continuous
Urinary s. Gravity	continuous	mean BP	continuous

Bayesian optimisation relies on using a Gaussian Process (GP) defined by the property that any finite set of N points $\{\mathbf{x}_n \in \mathcal{X}\}_{n=1}^{N}$ to induce a multivariate Gaussian distribution:

$$f : \mathcal{X} \to \mathbb{R}$$

With observations $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where $y_n \sim \mathcal{N}\left(f\left(\mathbf{x}_n\right), \nu\right)$ and ν is the variance of noise. The acquisition function is described as $a : \mathcal{X} \to \mathbb{R}^{+}$ and determines what point in $\mathcal{X}$ should be evaluated next via optimization $\mathbf{x}_{\text{next}} = \operatorname{argmax}_{\mathbf{x}} a(\mathbf{x})$. The acquisition functions depend on the previous observations, as well as the GP hyperparameters. The goal is then to maximize the expected improvement (EI) over the current best and use the highest utility hyperparameter values in computing the loss.

Table 4: Features extracted from the MIMIC-IV database. The features include demographic data collected for all patients, ICU unit-specific information like type of unit, hospital information like regional location, time since admission, vital signs including respiratory rate and blood pressure, and biochemical measurements including blood glucose and hemoglobin.

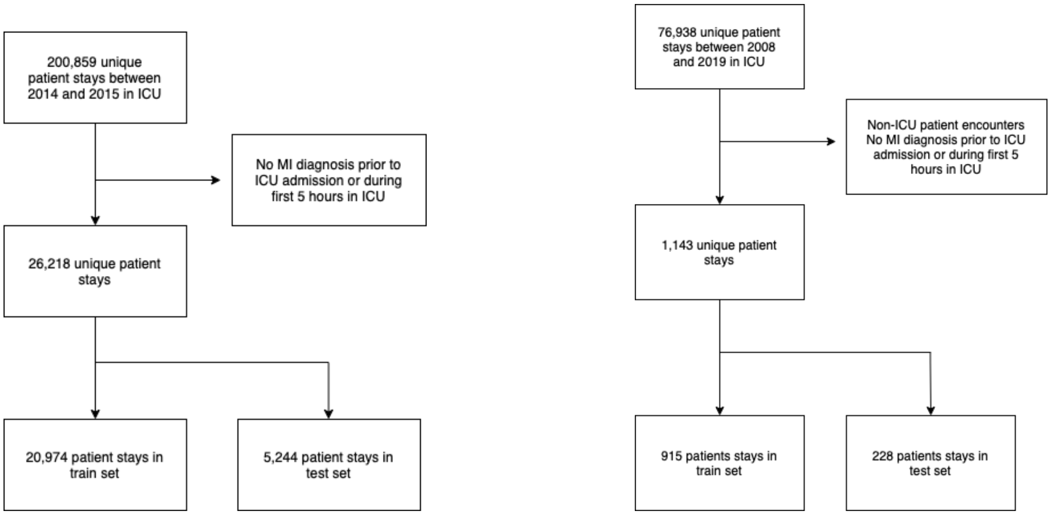
Feature	**Type**	**Feature**	**Type**
Sex	binary	Braden Score	continuous
Age	integer	Strength L Arm	continuous
Height	continuous	Strength R Arm	continuous
Weight	continuous	Strength L Leg	continuous
Hour of Admission	integer	Strength R Leg	continuous
Time Since Admission	continuous	GCS - Eye	continuous
Eye Response	continuous	GCS - Motor	continuous
Motor Response	continuous	GCS - Verbal	continuous
Verbal Response	continuous	Daily Weight	continuous
Ethnicity	categorical	ALT	continuous
Unit Type	categorical	AST	continuous
Admission Location	categorical	HCO3	continuous
Insurance	categorical	Hct	continuous
Time-series (summary features)			
ALT	continuous	Alkaline Phosphatase	continuous
Anion Gap	continuous	AST	continuous
Base Excess	continuous	Bicarbonate	continuous
Bilirubin	continuous	Calcium	continuous
Total CO2	continuous	Chloride	continuous
Creatinine	continuous	Glucose	continuous
Hematocrit	continuous	Hemoglobin	continuous
INR(PT)	continuous	Lactate	continuous
MCH	continuous	MCHC	continuous
MCV	continuous	Magnesium	continuous
PT	continuous	PTT	continuous
Phosphate	continuous	Platelet Count	continuous
Potassium	continuous	RDW	continuous
Red Blood Cells	continuous	Sodium	continuous
Urea Nitrogen	continuous	White Blood Cells	continuous
pCO2	continuous	pH	continuous
pO2	continuous	JH-HLM	continuous
Dyspnea Assessment	continuous	Daily Weight	continuous
Glucose	continuous	Heart Rate	continuous
DBP	continuous	SBP	continuous
O2 Flow	continuous	O2 Sat (%)	continuous
Pain Level	continuous	Pain Level Response	continuous
Phosphorous	continuous	Respiratory Rate	continuous
Richmond-RAS Scale	continuous	Temperature (°F)	continuous

When maximising the EI, we sample from the set of unexplored points without trying out all possible hyperparameter combinations. The algorithm can be shortly described as:

1. Given observed values $f(\mathbf{x})$, update the posterior using the GP model
2. Find $\mathbf{x}_{\text{new}}$ that maximises the EI: $\mathbf{x}_{\text{new}} = \arg\max EI(\mathbf{x})$
3. Compute the loss for the point $\mathbf{x}_{\text{new}}$

B.3 Metrics

The metrics used to evaluate the models include:


200,859 unique patient stays between 2014 and 2015 in ICU
No MI diagnosis prior to ICU admission or during first 5 hours in ICU
26,218 unique patient stays
20,974 patient stays in train set
5,244 patient stays in test set
76,938 unique patient stays between 2008 and 2019 in ICU
Non-ICU patient encounters No MI diagnosis prior to ICU admission or during first 5 hours in ICU
1,143 unique patient stays
915 patients stays in train set
228 patients stays in test set

(a) Cohort selection for eICU database

(b) Cohort selection for MIMIC-IV database

Figure 2: 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.

Table 5: 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. IQR used for secondary MI onset in hours after admission.

	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%)
MI onset (hours)	16.9 (10.0, 27.1)	15.9 (9.0, 26.0)	-	-

1. Area under receiver-operating-characteristic curve (AUROC): an ROC curve is a plot of true positives (TP) as a function of false positives (FP) where each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve is a summary measure of sensitivity and specificity Zou et al. (2007).

2. Sensitivity, the probability of a positive prediction for patients with disease (i.e. the conditional probability of correctly identifying diseased patients)

$$\frac{TP}{TP + FN}$$

3. Specificity, the probability of a negative prediction for patients without the condition

$$\frac{TN}{TN+FP}$$

4. Accuracy, ratio between correctly classified examples and the total number of cases in the dataset. In our case, can be misleading because of class imbalance where simply assigning all examples to the majority class is a way of achieving high accuracy, so instead we rely on using balanced accuracy as the average of sensitivity and specificity instead

$$\frac{Sensitivity + Specificity}{2}$$

5. Average Precision (AP) summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall at the nth threshold [1]. This implementation is not interpolated and is different from computing the area under the precision-recall curve with the trapezoidal rule, which uses linear interpolation and can be too optimistic. AP is then similar to using the midpoint rule for estimating the area (hence ”average” precision)

B.4 XMI-ICU Framework

The framework for pseudo-dynamic machine learning prediction can be seen in Figure 3.

B.5 Interpretability Methods

We use the *shap* library and built on the game-theoretic concept of treating features in the final model as players in a voting game. The method is applied on the entire test set and is based on ideas from game theory **?**Ibrahim et al. (2020). In short, the following equation is used to calculate the Shapley value φ for feature i:

$$\varphi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \bigcup \{x_i\}) - v(S)) \tag{1}$$

Where features have their value calculated by taking the difference between the results of the characteristic function v on N (the set of all features) and S (the subset of N without feature i). The Shapley value of a particular feature i is then calculated by taking the average of the marginal contributions of all possible combinations.

As Figure 4 highlights the application of this method for mortality prediction. This set of results relates only to the 6-hour prediction task. We then added a Gaussian distributed feature to the feature set to evaluate the susceptibility of the top variables as identified by Shapley values changing, and we can see in Figure 4 that the interpretability provided remains robust to noise.

B.6 Machine Learning Methods

C Machine Learning Model Specifications for Optimisation

C.1 Additional Results

We also evaluated our framework for internal validation with MIMIC-IV separately ie. training and tuning on MIMIC-IV train set and testing on a held-out MIMIC-IV test set. The results are a lot more robust than for external validation as we might expect and can be found in Figure 5.

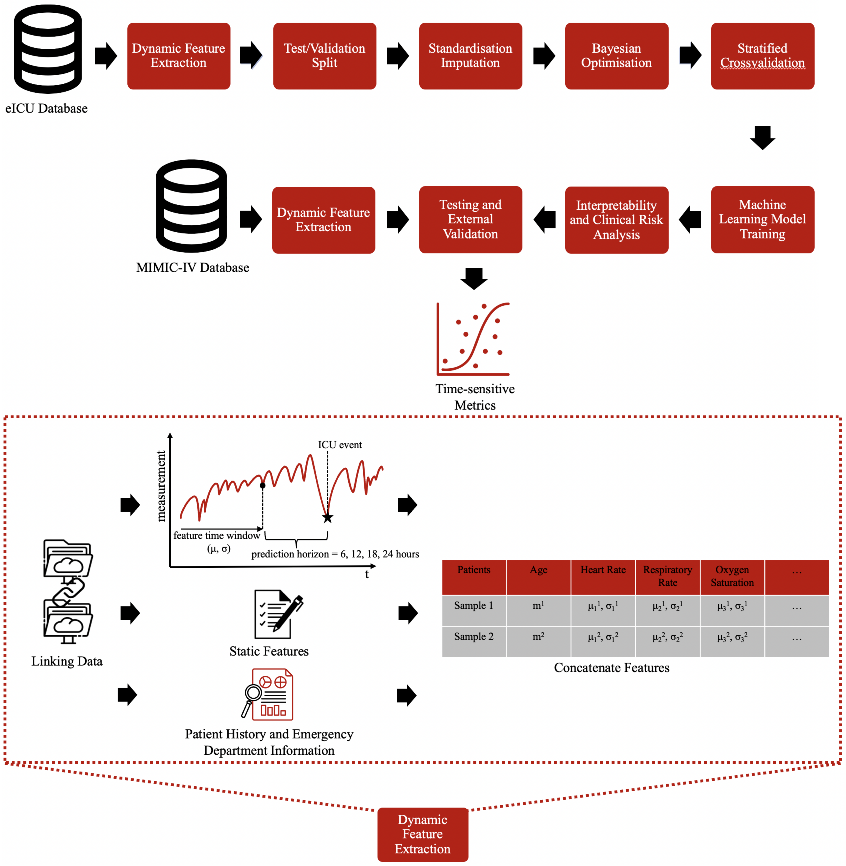

eICU Database
Dynamic Feature Extraction
Test/Validation Split
Standardisation Imputation
Bayesian Optimisation
Stratified Crossvalidation
Machine Learning Model Training
Interpretability and Clinical Risk Analysis
Testing and External Validation
Dynamic Feature Extraction
MIMIC-IV Database
Time-sensitive Metrics
measurement
ICU event
feature time window
(μ, σ)
prediction horizon = 6, 12, 18, 24 hours
t
Linking Data
Static Features
Patient History and Emergency Department Information
Patients
Age
Heart Rate
Respiratory Rate
Oxygen Saturation
…
Sample 1
m1
μ11, σ11
μ21, σ21
μ31, σ31
…
Sample 2
m2
μ12, σ12
μ22, σ22
μ32, σ32
…
Concatenate Features
Dynamic Feature Extraction

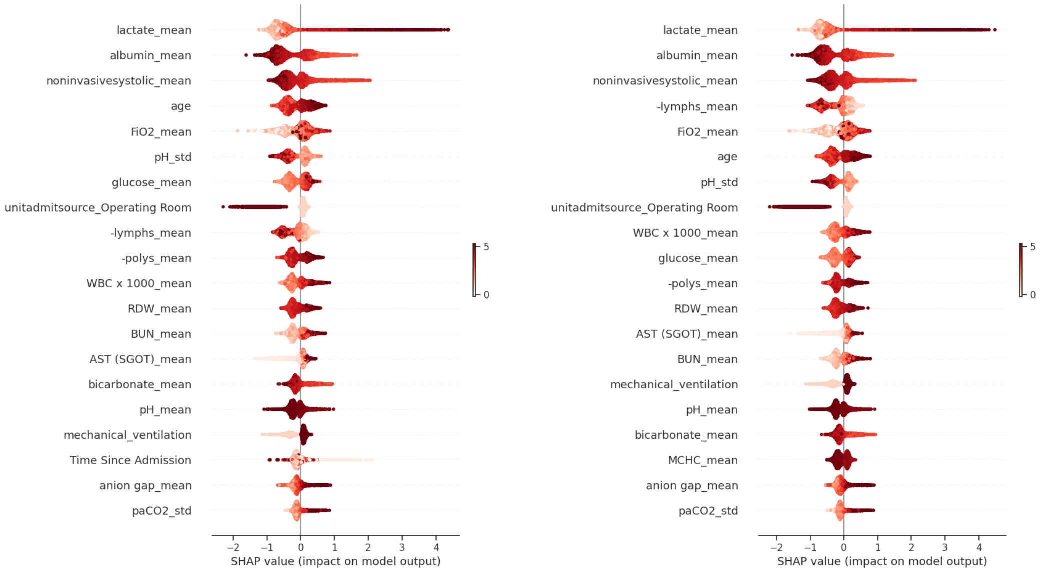
Figure 3: Proposed XMI-ICU framework for dynamic mortality and MI recurrence prediction in heart attack ICU Patients. The top part of the figure shows the dynamic feature extraction that links hospital-wide data including pre-admission information, ICU stay measurements, and emergency department variables. 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.

Results using APACHE-IV as a feature in trainin the models can be seen in Tables 10 and 11.

Examples of time-stratified prediction robustness for XMI-ICU can be seen in 6.

C.2 Time-robustness Checks

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 in the top of Table 12 where each indicates the group of patients correctly predicted at all previous time windows except one. The bottom of Table 12 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.



(a) Importance of clinical variables for secondary MI prediction across patient ICU stay

(b) Importance of clinical variables for mortality prediction across patient ICU stay

Figure 4: SHAP values of features for XMI-ICU prediction of mortality and with random noise added. No significant change appears in the top features satisfying the perturbation constraints. The relative vertical ranking of the features corresponds to higher importance of those features in making a correct prediction. The darker colours in the horizontal plane for each feature correspond to higher values of that feature contributing to either stronger positive prediction (if darker colour on the right side of the vertical line) or stronger negative prediction of outcome otherwise.

Decision curves, clinical risk calculations, and nomograms were computed and plotted in R.

C.3 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 Kerr et al. (2016). A mathematical representation can be seen in the equation bellow:

$$NB_R = TPR_R P - \frac{R}{1-R} FPR_R (1-P) \tag{2}$$

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) Chen et al. (2022a).

To communicate the clinical significance of the XMI-ICU model results to clinicians, we evaluated our model with clinical impact curves (Figures 7a and 7b) and decision curve estimates (Figures 8a and 8b) for robust risk evaluation. A 90 percent confidence interval was derived with 50 bootstrap

Dataset	Model		Parameters
eICU MIMIC-IV	Logistic Regression	C Regularisation Solver	0.1 Lasso (l1) liblinear
eICU MIMIC-IV	Naive Bayes	Smoothing	alpha = 0.0
eICU MIMIC-IV	Linear Discriminant Analysis	Shrinkage Solver	0.17 Eigen
eICU MIMIC-IV	Random Forest	Estimators Features Max Depth Minimum Splits Minimum Leaf Bootstrap	150 sqrt 10 5 10 False
eICU MIMIC-IV	XGBoost	Estimators Learning Rate Max Depth Minimum Splits Maximum Delta Tree Method	150 0.1 3 0.5 0 hist
eICU MIMIC-IV	AdaBoost Ensemble Ensemble (XGBoost)	Estimators Estimators	150 80

Table 6: Model Architecture Details for PE

iterations on the test set. As the clinical impact curves for MI and 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.

The nomogram in Figure 9 is an example of risk calculation where one first draws a line from each parameter value to the point line for the point for that feature, then the points for all the features are added up, after which a line from the total points line is drawn vertically to determine the risk of mortality on the lower line of the nomogram as defined by a linear transformation of risk probabilities.

Dataset	Model		Parameters
eICU MIMIC-IV	Logistic Regression	C Regularisation Solver	1.0 Lasso (l1) liblinear
eICU MIMIC-IV	Naive Bayes	Smoothing	alpha = 1e-5
eICU MIMIC-IV	Linear Discriminant Analysis	Shrinkage Solver	0.1 Eigen
eICU MIMIC-IV	Random Forest	Estimators Features Max Depth Minimum Splits Minimum Leaf Bootstrap	150 sqrt None 10 10 True
eICU MIMIC-IV	XGBoost	Estimators Learning Rate Max Depth Minimum Splits Maximum Delta Tree Method	200 0.3 2 0.06 0 hist

Table 7: Model Architecture Details for PE (With Undersampling)

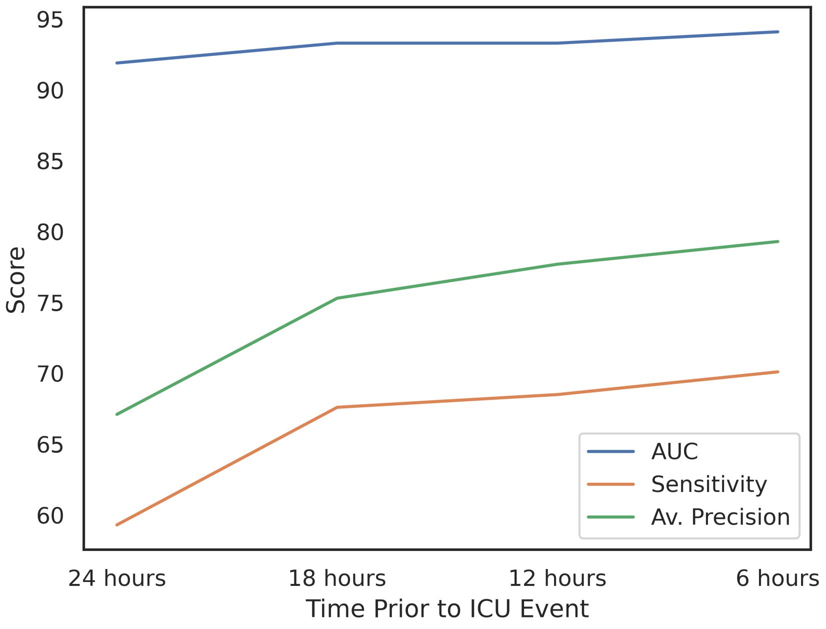


Figure 5: XMI-ICU performance across time for mortality prediction as evaluated on MIMIC-IV held-out test sets after training on an internal MIMIC-IV set

Dataset	Model		Parameters
eICU MIMIC-IV	Logistic Regression	C Regularisation Solver	0.01 Lasso (l1) liblinear
eICU MIMIC-IV	Naive Bayes	Smoothing	alpha = 0.0
eICU MIMIC-IV	Linear Discriminant Analysis	Shrinkage Solver	0.0 lsqr
eICU MIMIC-IV	Random Forest	Estimators Features Max Depth Minimum Splits Minimum Leaf Bootstrap	150 auto None 10 10 False
eICU MIMIC-IV	XGBoost	Estimators Learning Rate Max Depth Minimum Splits Maximum Delta Tree Method	350 0.1 4 0.45 1 hist
eICU MIMIC-IV	AdaBoost Ensemble Ensemble (XGBoost)	Estimators Estimators	20 50

Table 8: Model Architecture Details for Mortality

Table 9: eICU test with all features, eICU test only using top 8 features, and MIMIC-IV external validation (Val: Mean ± SD) prediction results for secondary MI and 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 Appendix Materials.

	Val AUC	AUC	Accuracy*	Average Precision
eICU Secondary MI				
6 hours	85.1 ± 0.3	85.6	79.0	75.9
12 hours	86.5 ± 0.8	85.4	78.7	73.0
18 hours	85.8 ± 1.0	85.5	78.6	70.3
24 hours	85.1 ± 1.2	84.3	75.7	70.3
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
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

Table 10: Validation (Val: Mean ± SD) and test prediction results for mortality prediction 6 hours in advance.

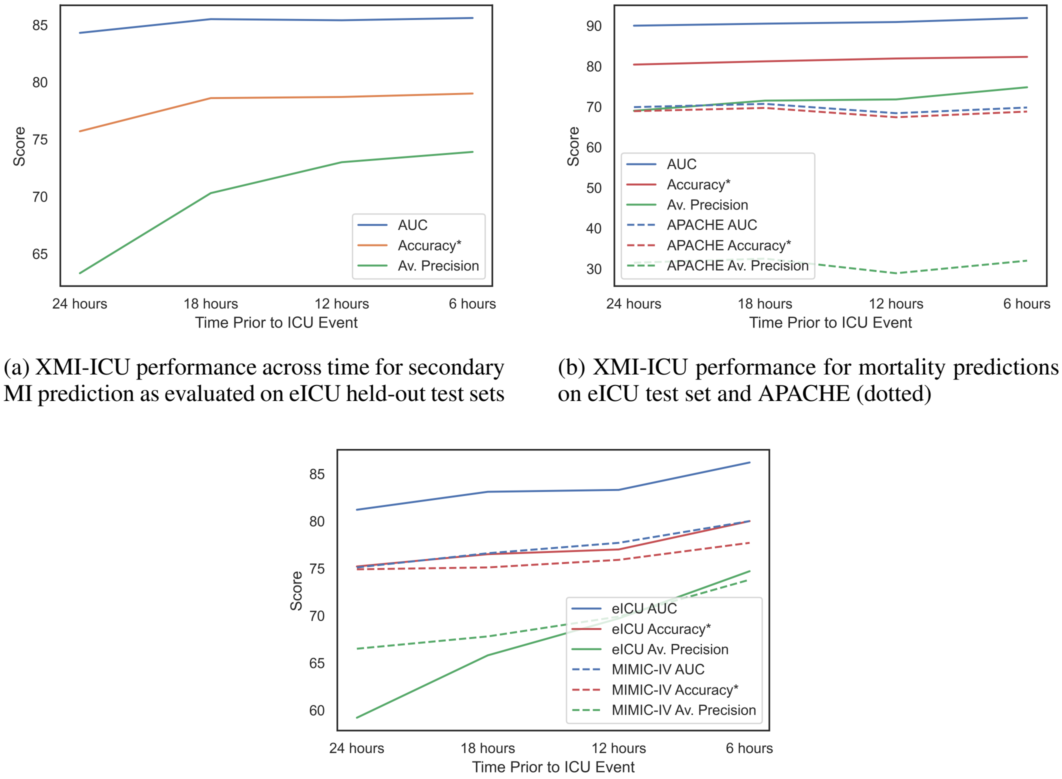
	Val AUC	AUC	Average Precision
XMI-ICU	**92.9 ± 0.4**	**91.9**	68.7
APACHE IV	-	69.8	31.9
TabNet	86.8 ± 2.1	85.0	64.7
TabNet (pretrained)	-	83.1	**82.1**
NODE	87.8 ± 0.7	86.3	66.3
Logistic Regression	91.3 ± 0.4	90.1	61.5
Random Forest	92.1 ± 0.5	91.1	64.4
SVM	91.3 ± 0.8	90.2	62.1
SVM (linear)	88.5 ± 0.7	88.6	67.8
LDA	80.5 ± 2.0	78.6	33.3

Table 11: Validation (Val: Mean ± SD) and test prediction results for secondary MI prediction stratified with time for XMI-ICU.

	Val AUC	AUC	Average Precision
6 hours	86.2 ± 0.4	86.0	75.9
12 hours	86.0 ± 0.8	84.1	71.1
18 hours	85.9 ± 1.0	86.6	68.7
24 hours	86.0 ± 1.3	86.5	63.2

Table 12: 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	
Secondary MI	8.4	8.4	5.8	
MIMIC-IV				
Mortality	6.4	6.3	4.7	



(a) XMI-ICU performance across time for secondary MI prediction as evaluated on eICU held-out test sets

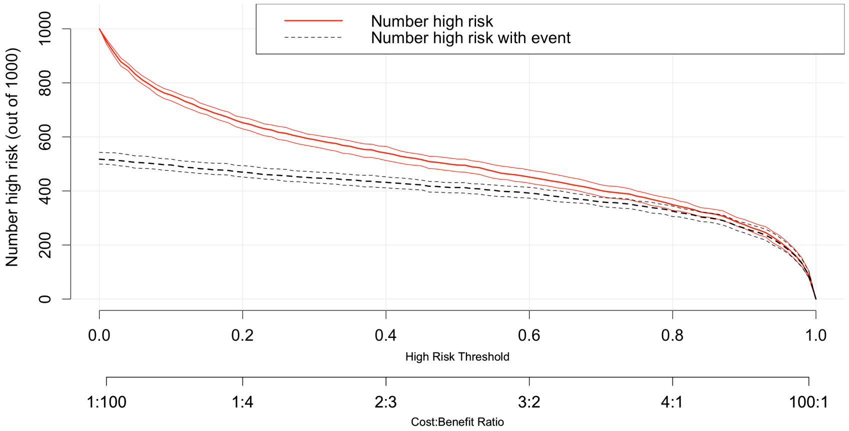
(b) XMI-ICU performance for mortality predictions on eICU test set and APACHE (dotted)

(c) XMI-ICU performance across time for mortality predictions on eICU held-out test set and external MIMIC-IV set (dotted) with only top 8 features used

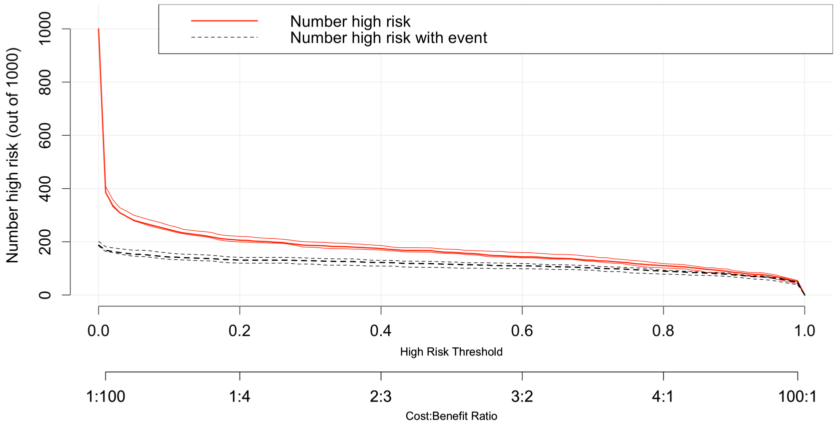
Figure 6: Robustness and reliability of XMI-ICU prediction performance over time in the ICU for secondary MI prediction (top left) and mortality prediction (top right) using all features available in eICU and as measured by a variety of metrics. The bottom figure contains results from eICU held-out test set and MIMIC-IV external cohort with only the top 8 features identified by Shapley value analysis.

Table 13: AUROC test results for XMI-ICU evaluated on subpopulations for 6 hour prediction.

	Secondary MI	Mortality	Mortality (external MIMIC-IV)
Men	85.2	90.2	81.9
Women	88.9	92.7	77.5
Caucasian	85.6	91.7	81.8
Black/Hispanic	84.8	92.3	75.6

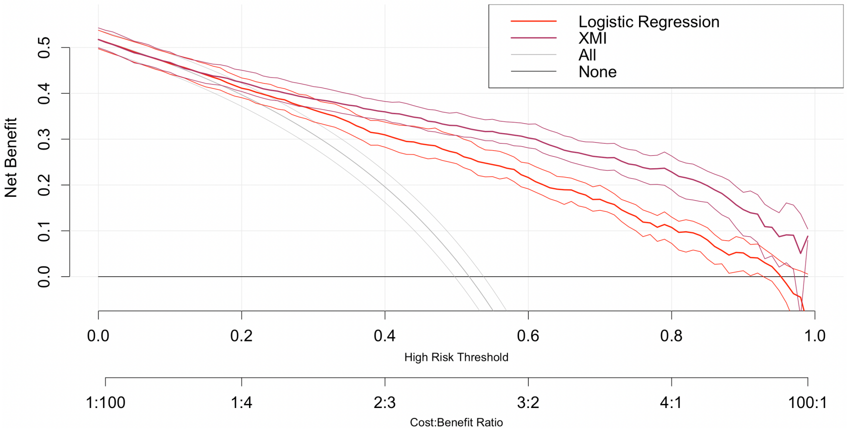


(a) Clinical Impact Curve of XMI-ICU for secondary MI risk

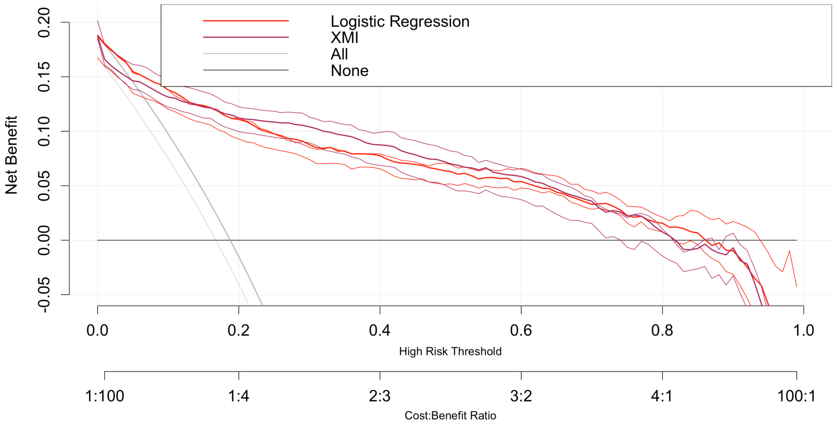


(b) Clinical Impact Curve of XMI-ICU for mortality risk

Figure 7: Clinical decision-making evaluation performance of XMI-ICU for secondary MI and mortality prediction using only the top 8 features on the entire eICU test set. Here we include 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 over- or underestimate risk of that event.



(a) Decision curve of XMI-ICU compared to a logistic regression for secondary MI across different risk thresholds



(b) Decision curve of XMI-ICU compared to a logistic regression for mortality across different risk thresholds

Figure 8: Clinical decision-making evaluation performance of XMI-ICU for secondary MI and mortality prediction using only the top 8 features on the entire eICU test set. Here we include 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. In the decision curves, the ”All” tag corresponds with the net benefit behaviour of having all patients predicted positive and ”None” with having no patients predicted positive.

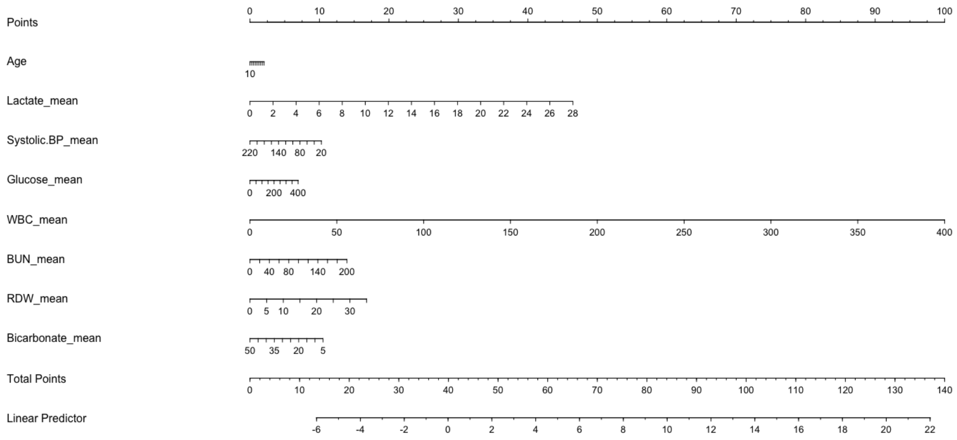


Figure 9: Nomogram to estimate the risk of mortality in MI patients in multi-centre ICUs from the eICU test set. The nomogram includes the top 8 features identified by the model as highly predictive for this patient population as well as the external cohort. The nomogram is used to provide insight into risk calculation based on these features using ranges measured for the patient. One simply draws a straight line from each feature value to the points line, then points are added on the total points line after which a straight line is drawn downward to the linear predictor for a risk estimate respectively. The risk score calculated through this nomogram is for 24-hour prediction.