BMC Medical Informatics and Decision Making

RESEARCH

Open Access

CRISP: A causal relationships-guided deep learning framework for advanced ICU mortality prediction

Linna Wang[1], Xinyu Guo[2], Haoyue Shi[4], Yuehang Ma[1], Han Bao[1], Lihua Jiang[2], Li Zhao[2], Ziliang Feng[1], Tao Zhu[3] and Li Lu[1*]

Abstract

Background Mortality prediction is critical in clinical care, particularly in intensive care units (ICUs), where early identification of high-risk patients can inform treatment decisions. While deep learning (DL) models have demonstrated significant potential in this task, most suffer from limited generalizability, which hinders their widespread clinical application. Additionally, the class imbalance in electronic health records (EHRs) complicates model training. This study aims to develop a causally-informed prediction model that incorporates underlying causal relationships to mitigate class imbalance, enabling more stable mortality predictions.

Methods This study introduces the CRISP model (Causal Relationship Informed Superior Prediction), which leverages native counterfactuals to augment the minority class and constructs patient representations by incorporating causal structures to enhance mortality prediction. Patient data were obtained from the public MIMIC-III and MIMIC-IV databases, as well as an additional dataset from the West China Hospital of Sichuan University (WCHSU).

Results A total of 69,190 ICU cases were included, with 30,844 cases from MIMIC-III, 27,362 cases from MIMIC-IV, and 10,984 cases from WCHSU. The CRISP model demonstrated stable performance in mortality prediction across the 3 datasets, achieving AUROC (0.9042–0.9480) and AUPRC (0.4771–0.7611). CRISP's data augmentation module showed predictive performance comparable to commonly used interpolation-based oversampling techniques.

Conclusion CRISP achieves better generalizability across different patient groups, compared to various baseline algorithms, thereby enhancing the practical application of DL in clinical decision support.

Trial registration: Trial registration information for the WCHSU data is available on the Chinese Clinical Trial Registry website (http://www.chictr.org.cn), with the registration number ChiCTR1900025160. The recruitment period for the data was from August 5, 2019, to August 31, 2021.

Keywords Mortality prediction, Class imbalance, Causal machine learning, Clinical decision support

*Correspondence:
Li Lu
luli@scu.edu.cn
[1]College of Computer Science, Sichuan University, 24 South Section 1, 1st Ring Road, Chengdu, Sichuan 610065, China
[2]Department of Health Policy and Management, West China School of Public Health and West China Fourth Hospital, Sichuan University, 37 Guoxue Alley, Chengdu, Sichuan 610041, China
[3]Department of Anesthesiology, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, 37 Guoxue Alley, Chengdu, Sichuan 610041, China
[4]College of Mechanical Engineering, Sichuan University, 24 South Section 1, 1st Ring Road, Chengdu, Sichuan 610065, China

Introduction

The intensive care units (ICUs) are specialized facilities that provide comprehensive care to the most severely ill patients, offering intensive medical and nursing services with advanced monitoring capabilities [1, 2]. Globally, ICU prognosis, particularly regarding mortality decision making, remains a prominent focus of clinical research [3]. In the United States, mortality rates in various ICUs ranged from 11.3% to 12.6% between 2001 and 2012 [4]. Additionally, in many countries, the average age of ICU patients now exceeds 65 years [5–7]. A global audit conducted in 2014 reported ICU mortality rates ranging from 9.3% to 26.2% [8], and a 2021 review of 129 studies focusing on elderly ICU patients revealed an even broader range, with mortality rates ranging from 1% to 51% [9]. In this clinical context, precise prediction of mortality for ICU patients is essential for timely risk assessment, which directly impacts patient outcomes, allocation of resources, and satisfaction with care.

Several traditional scoring systems have been employed to predict mortality in various patient groups, with moderate accuracy (AUROC ranging from approximately 0.65 to 0.85) [10–15]. These include systems such as the Acute Physiology and Chronic Health Evaluation (APACHE) II and III [16, 17], the Simplifed Acute Physiology Score (SAPS) II and III [18, 19], and the Sequential Oran Failure Assessment (SOFA) [20]. However, these tools have several limitations. First, factors such as increased life expectancy, shifts in public health conditions, and the emergence of new diseases can lead to a decline in their predictive accuracy over time [21]. For example, systems like APACHE, SOFA, and SAPS may experience calibration issues as patient populations and medical treatments evolve [22, 23]. Second, some existing scoring systems are static, largely relying on data collected only during the first day of ICU admission. This can compel clinicians to rely on subjective judgment, which is prone to bias [24, 25]. Third, when validated across different countries, the performance of these scoring systems tends to be inconsistent [21, 26]. This suggests that inadequate consideration of diverse patient cohorts may also contribute to performance discrepancies.

The limitations of traditional scoring systems have prompted a surge in research exploring Machine Learning (ML) and Deep Learning (DL) techniques for diagnosis and prognosis prediction in various medical fields [27–29]. The widespread adoption of automated electronic health records (EHRs) has enabled the extraction of vast amounts of clinical data, allowing models to be continuously updated and improved based on real-time clinical information. Early DL applications for mortality prediction primarily utilized simple feed-forward architectures, which showed comparable performance to traditional scoring systems [30]. Recent advancements have led to the development of more advanced DL models, which can be broadly categorized into 3 approaches based on the type of input data used: (1) Time series-based models, which leverage clinical time series data, such as vital signs and laboratory results, to predict patient outcomes [24, 31–34]; (2) Text-based models, which extract prognostic insights from clinical notes containing critical semantic information [35–37]; and (3) Hybrid models, which combine multimodal data, offering a more comprehensive approach to mortality prediction [38–40]. Recent studies using the publicly accessible MIMIC-III database for ICU mortality prediction have reported Area Under the Receiver Operating Characteristic (AUROC) values between approximately 0.8 and 0.9, and Area Under the Precision-Recall Curve (AUPRC) values ranging approximately from 0.3 to 0.7 [41–43].

While DL models have largely focused on improving predictive performance, the health sciences have often emphasized generalizability and the interpretation of domain-specific knowledge [44, 45]. Recent studies have underscored the need for enhanced capabilities to extract underlying causal structures that support clinical decision-making [46–48]. These causal structures [49] can facilitate interventions by evaluating counterfactual scenarios—for example, "Would the patient have survived if they had not developed a specific condition upon admission?" By analyzing the effects of alternative scenarios, referred to as counterfactuals [50], such models can provide actionable insights. However, manually constructing these complex and interdependent causal relationships is challenging. Recent advancements in causal discovery algorithms [51, 52] have made it possible to build causal models that can further support downstream tasks, such as guiding prognostic models [53]. Moreover, many DL models rely on high-dimensional features and are highly sensitive to input data variability [54], which presents significant challenges in identifying robust causal relationships that are generalizable across diverse datasets.

In addition to the concerns mentioned above regarding models, addressing class imbalance presents another significant challenge in clinical research. While mortality represents the most severe outcome, mortality rates in commonly available datasets typically range from 5.5% to 9.9% [55]. Although these rates are high from a human perspective, they may seem relatively low from a statistical viewpoint in the context of DL studies. Furthermore, strict privacy and ethical considerations limit the availability of mortality data. However, DL models rely on large volumes of data for effective training. Various techniques for handling imbalanced data have been explored, with benchmark oversampling methods such as the Synthetic Minority Oversampling Technique (SMOTE) [56] and its variants being widely employed in this context. Recent studies have shown that incorporating causal

knowledge can help mitigate data biases and improve generalization [47, 57]. There is potential in leveraging causal relationships to enhance data augmentation strategies [58]. Generating counterfactual instances for data augmentation [59, 60] is a promising approach, as counterfactuals offer a causal perspective on how alternative outcomes could arise. Existing research on synthetic counterfactual generation [59] can be broadly categorized into 2 types: endogenous counterfactuals, which are generated from naturally occurring feature values; and exogenous counterfactuals, which may not rely on actual feature values. Endogenous counterfactual methods adapt "native" counterfactuals to create plausible contrastive explanations. These native counterfactuals are derived directly from existing instances in the dataset, often represented by the nearest unlike neighbors of a target instance [61]. There is considerable potential for exploring class balancing methods based on native counterfactuals, as they ensure that the generated instances are both realistic and aligned with the underlying data distribution.

To address the aforementioned challenges and enhance in-hospital mortality prediction, this paper introduces CRISP, a deep learning framework that integrates the universal approximation capabilities of neural networks with causal relationships. Specifically, CRISP incorporates causality into 2 key aspects: data augmentation and the prediction model. We hypothesized that, with domain knowledge outlining a cause–effect pathway and sufficient data encompassing the causal path, it is possible to approximate causal effects. Additionally, to manage the complexity of causal graphs and address variability across datasets, we propose distilling the graph into a high-level version to reduce its complexity. This approach is inspired by the observation that, in complex systems, a small number of causal modules often dominate and substantially influence a significant portion of the causal pathways [62, 63].

This study utilized the MIMIC-III (Version 1.4) dataset [64] to develop the CRISP model. MIMIC-III is a widely recognized benchmark for ICU mortality prediction, comprising de-identified data from patients admitted to the emergency department or intensive care unit (ICU) at Beth Israel Deaconess Medical Center in Boston, USA. To externally validate the model, we used the publicly available MIMIC-IV (Version 3.1) dataset [65]. Given the elevated mortality risk among elderly ICU patients, we also trained and validated the model on a dataset from the West China Hospital of Sichuan University (WCHSU), China, focusing specifically on this demographic. The experiments demonstrated the model's strong performance across all 3 datasets. Below, the key contributions of this study are summarized:

- **Augmenting Data through Counterfactual Strategy:** This study presents a data augmentation strategy based on causal graphs to generate minority data by adapting native counterfactuals. By leveraging native counterfactuals as templates to create new minority class instances, our approach demonstrates competitive predictive performance compared to traditional interpolation-based oversampling techniques.
- **Enhancing Mortality Prediction Performance:** By incorporating causal graph to process feature groups alongside basic patient information, the proposed CRISP model demonstrates competitive performance in mortality prediction across 3 datasets.
- **Assessing Model Performance across Different Data Distributions:** This study evaluates the CRISP on the WCHSU dataset, which features distinct patient demographics, primarily consisting of older and Asian populations, compared to the MIMIC-III and MIMIC-IV datasets. Additionally, within the MIMIC-III dataset, the study tests CRISP's performance in predicting acute kidney injury, further demonstrating the model's generalizability to new datasets and tasks.

Preliminaries

This section outlines the core problem setting addressed in this study. For definitions related to causal theory, please refer to the Supplementary Material.

In our task, we define $D = \{(X_i, Y_i)|i = 1, 2, \ldots, N\}$ as a dataset containing N patient cases, where each case i consists of a single outcome value Y_i. Each case, denoted as X_i, can be represented as a sequence $X_i = [X_i^D, X_i^P, X_i^M, X_i^I, X_i^B]$, where X_i^D, X_i^P, X_i^M, X_i^I and X_i^B represent sets of diagnoses, procedures, medications, ICU indicators (such as vital signs and lab measurements) and basic demographic information, respectively. Let X represent the patient characteristics, and Y denote the mortality outcomes.

Problem 1: Insufficient minority class problem. In the datasets used in this study, instances with $Y = 1$ (representing mortality) are less frequent than those with $Y = 0$. How to leverage endogenous counterfactual methods, based on clear causal relationships, to enhance the minority class instances using native counterfactuals?

Problem 2: Incorporating Causal Structure into Deep Learning Model. The goal is to estimate the probability of mortality $\hat{Y}$ (a binary classification problem) for ICU patients prior to hospital discharge. The challenge lies in effectively integrating stable causal relationships and causal effects into deep learning models, thereby improving the robustness of predictions across datasets with varying distributions.

Methods

Identify the general causal structure

We assume a stable causal Directed Acyclic Graph (DAG [1]) $G = (V, E)$, where V is the set of nodes representing X and Y, and E is the set of directed edges. However, identifying robust causal relationships across diverse datasets presents significant challenges. First, collaboration with healthcare professionals to ensure that the resulting causal graphs are clinically meaningful is resource-intensive. Furthermore, the variable sets observed across multiple sources or domains are not entirely identical. While existing studies [66] have proposed methods to combine learned structures from multiple domains and obtain final structures over an integrated set of variables, we believe that this approach has inherent limitations (as explained in the Supplementary Material SFigure 1), such as the potential for missing or misidentified causal relationships.

In this study, rather than combining causal relationships derived from multiple databases, the focus is on investigating global causal relationships. Specifically, for the mortality outcome Y, a large number of both direct and indirect causal pathways are expected. Targeting only a small subset of these causal factors or pathways may lead to ineffective interventions [63]. Therefore, prioritizing the relationships between higher-level groups, where features are grouped based on their real-world significance, may offer an effective approach to addressing the problem. This strategy aligns with prior works, such as that by [62], which explains the final pathological phenotype by defining the network interactions between modular elements and identifies potential regulatory modules that could modify the phenotype. Similarly [67], introduced the concept of "think globally, act locally," which highlights that generating local interventions to cure a particular disease requires understanding the global organization. Building on these ideas, this study leverages this opportunity to distill global causal relations among distinct groups $X = [X^D, X^P, X^M]$ and Y. This systematic categorization not only enables us to incorporate a comprehensive set of features but also simplifies the exploration of overarching causal pathways within specific feature groups. For these distinct feature groups, the underlying logic of clinical outcomes is already evident based on the available information. Nevertheless, we seek to confirm these causal relationships using available causal discovery methods.

We employed a gradient-based algorithm known as the Graphical Autoencoder (GAE) [68], which builds casual DAG through graph convolutional neural networks. GAE is an extension of NOTEARS [69], which is widely regarded as the first approach to recast the combinatorial graph search problem as a continuous optimization problem for structure learning. This allows for the modeling of non-linear structural relationships and vector-valued variables. The function is defined as:

$$f(X_j, A) = f_2(A^T f_1(X_j)) \tag{1}$$

where A is the adjacency matrix of the graph, f_1 and f_2 are multilayer perceptrons (MLPs) [68]. Demonstrated that GAE outperforms NOTEARS, particularly as the number of vertices in the graph increases, and they also noted that GAE requires much shorter training times. Given the flexibility of the GAE framework, we feed our feature groups into MLPs to obtain input to GAE. The acyclicity constraint in GAE is:

$$h(A) = tr(e^A A) - d = 0 \tag{2}$$

where d is the number of vertices in the graph, tr denotes the trace operator, and e^A represents the element-wise exponential of the adjacency matrix A. In practice, $h(A)$ may be small but non-zero, and edges with small weights require a thresholding operation to filter out less significant connections. In this study, edges with weights below 0.5 are considered weak and are set to 0.

Figure 1 presents the causal graph, which illustrate the directional dependencies between variables and denote causal relationships. Patients with more severe conditions are likely to receive a more severe diagnosis. We aim to estimate the causal effect of the treatment variable D on the outcome variable Y. In the graph, D influences Y indirectly through P and M. We treat P and M as mediating variables. There are no backdoor paths from D to mediating variables, so no additional confounding adjustments are required. There are no other backdoor paths from mediating variables to Y. Therefore, the Front-door criterion[2] is satisfied, and the frontdoor adjustment can be applied to estimate the causal effect of D on Y, which can be formulated as:

$$P(y|do(d)) = \sum_p \sum_m P(p|d) P(m|d,p) \sum_{d'} P(y|d', p, m) P(d') \tag{3}$$

Building on this analysis, we now turn our attention to our approach for oversampling and the development of our mortality prediction model.

[1] Definition is provided in the Supplementary Material.

[2] Definition is provided in the Supplementary Material.

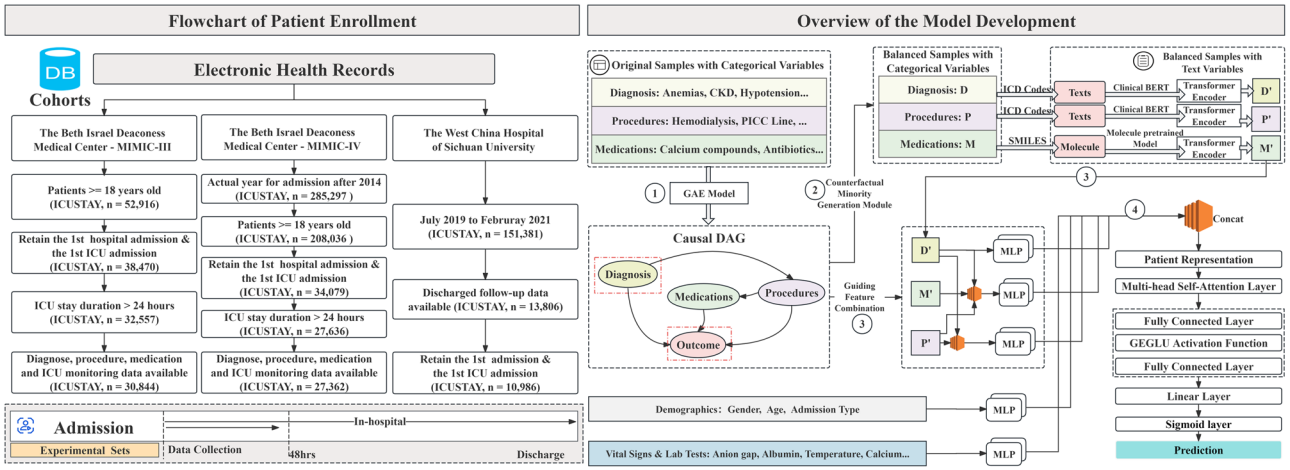

Fig. 1 Cohort Inclusion Process and Model Overview: This study includes 3 datasets, using MIMIC-III as the development dataset, MIMIC-IV as the external validation dataset, and WCHSU as a different source dataset to evaluate the model's performance. The causal graph guides data augmentation and input representation, forming the foundation for the primary prediction task. In the figure, "MLP" refers to basic Multi-layer Perceptrons, consisting of 2 fully connected layer with ReLU activation functions

Counterfactual minority generation module

In this section, the aim is to address the insufficient minority sample problem. In MIMIC III, the minority class consists of 3,265 records, while the majority class contains 27,579 records, resulting in a minority-to-majority ratio of approximately 1 to 10. In the validation set, MIMIC IV, there are 3,179 records with a mortality label of 1 and 24,183 records with a mortality label of 0, yielding a minority-to-majority ratio of approximately 1 to 10. Similarly, the WCHSU dataset exhibits extreme class imbalance, with only 99 records available in the minority class. The proportion of positive samples in WCHSU dataset is only approximately 0.90%. To address this imbalance, this study employed oversampling to preserve important information in the dataset while retaining all training data.

Previous studies have proposed methods for generating counterfactual instances. The seminal work of [70] proposed a method for generating counterfactual pairs by perturbing features until a shift in the target label is observed. Similar counterfactual generation methods [58–60] often leverage K-Nearest Neighbors (KNN) algorithms to identify the nearest counterfactual candidates across distinct subsets of the target variable, and implement strategies to enhance counterfactual coverage and similarity. Different from previous methods, this study introduces a novel Counterfactual Minority Generation Module (CMG) that integrates causal graphs into the counterfactual generation framework. The primary objective of CMG is to generate counterfactual instances for majority class observations, thereby ensuring class balance between the target groups. Specifically, during the counterfactual pair search process, CMG utilizes propensity scores[3] to match instances where the target class differs. The propensity score encapsulates multiple features into a single probability that reflects the likelihood of an individual belonging to the target class. Subsequently, for unmatched majority class instances, new counterfactual instances (new minority instances) are generated through causal pathways. In this study's causal graph, Diagnosis directly influences Procedures, Medications, and target outcome. The data generation process within CMG adheres to the causal ordering specified by the graph. The procedural steps for implementing CMG are outlined as follows (Fig. 2):

1. **Divide the dataset:** Let $D = \{(X_i, Y_i)|i = 1, 2, \ldots, N\}$ be the dataset consisting of N instances. Divide D into 2 subsets: the majority class subset $D_{majority} = \{(X_i, Y_i = 0)|i = 1, 2, \ldots, n\}$, and the minority class subset $D_{\min ority} = \{(X_i, Y_i = 1)|i = 1, 2, \ldots, N - n\}$.
2. **Estimate propensity scores:** Based on the causal graph, the intervention T corresponds to the diagnosis features in this study. The propensity score $e(X) = P(Y = 1|T, B)$, where B represents basic demographic information, is estimated. Apply Logistic Regression model to the data to compute the propensity scores $e(X)$.

[3] Definition is provided in the Supplementary Material.

Algorithm 1 Counterfactual Minority Generation Module

1: Input: Dataset $D = \{(X_i, Y_i)|i = 1, 2, \ldots, N\}$, intervention features T, basic covariant features B
2: Output: Generated minority instances X_i^*
3: Step 1: Divide the dataset into majority and minority subsets
4: $D_{majority} \leftarrow \{(X_i, Y_i)|Y_i = 0\}$
5: $D_{\min ority} \leftarrow \{(X_i, Y_i)|Y_i = 1\}$
6: Step 2: Estimate the propensity scores using Logistic Regression
7: **for** each $(X_i, Y_i) \in D_{majority} \cup D_{\min ority}$
8: $X_i^{T,B} \leftarrow X_i[T \cup B]$
9: Fit a Logistic Regression model on $(X_i^{T,B}, Y_i)$ to estimate the propensity score $e(X_i) = P(Y_i = 1|T, B)$
10: end for
11: Step 3: Construct counterfactual pairs using 1-Nearest Neighbor based on propensity scores
12: for each $(X_i, Y_i) \in D_{\min ority}$**do**
13: Find the nearest majority instance X_j such that the absolute difference in propensity scores $|e(X_i) - e(X_j)|$ is minimized
14: end for
15: Form counterfactual pairs $\{(X_i, X_j)\}$ where $i \in D_{\min ority}$ and $j \in D_{majority}$
16: Step 4: Randomly upsample minority instances based on features$T$
17: for each $(X_i, Y_i) \in D_{\min ority} \cup D_{unpairedmajority}$**do**
18: Generate new minority instances $X_i^{\min ority*}$ by randomly upsampling based on features T
19: end for
20: Step 5: Generate new minority instances by transferring features from majority instances
21: for each $X_i^{\min ority*}$ do
22: Find the nearest unpaired majority instance $X_j^{unpaired\_majority}$
23: Transfer features from the $T \rightarrow Y$ causal path of $X_j^{unpaired\_majority}$ to form the new minority instance $X_i^* = [X_i^{\min ority*}[T], X_j^{unpaired\_majority}]$
24: end for
25: Return: Generated minority instances X_i^*

3. **Construct counterfactual pairs:** Use the estimated propensity scores $e(X)$ to create counterfactual pairs between the majority and minority class instances using a 1-nearest neighbors(1-NN) approach, minimizing the difference in propensity scores. For each minority instance, a corresponding majority instance is paired. For unpaired majority instances, new minority instances will be generated.
4. **Randomly upsample minority instances:** Randomly upsample the minority instances based on the *T* features to generate new instances $X^T_{i^{\min ority*}}$. For each newly generated minority instance $X^T_{i^{\min ority*}}$, calculate its propensity score and match it to the nearest unpaired majority instance $X^{\mathrm{T}}_{j^{\mathrm{unpaired\_majority}}}$.
5. **Generate new minority instances:** Transfer the features along the *T–to–Y* causal path of the nearest unpaired majority instance to form new minority instances. The new minority instances are constructed as follows:

$$X_i^* = [X^{\mathrm{T}}_{i^{\mathrm{minority}*}}, X_{j^{\mathrm{unpaired\_majority}}}] \tag{4}$$

The procedure of the proposed module is outlined in Algorithm 1. We compared CMG with the benchmark oversampling methods in the class-imbalance task: SMOTE, ADASYN, SMOTE-Tomek and SOMTE-ENN [71]. The performance results of CMG are provided in Fig. 3.

Enhanced predictions through joint causal discovery and deep learning

CRISP is designed with a Transformer-based architecture comprising 2 main components: a prediction module for generating outputs based on transformed feature

representations, and a treatment effect estimation module. Our input data comprised a combination of tabular and text-based information, encompassing diagnoses, procedures, medications, demographics, and ICU observational indicators. Procedures refer to the interventions or operations performed during the ICU stay, while medications indicate the drug treatments administered after ICU discharge. For diagnoses, procedures, and medications, both categorical and textual data were utilized. First, tabular categorical data was used to represent these features. For each patient, if a specific diagnoses, procedure or medication was administered, the corresponding feature is assigned a value of 1; otherwise, if no record is available, it is assigned a value of 0. Textual data for diagnoses and procedures were tokenized using Clinical BERT [72], a language model fine-tuned on extensive medical text corpora. For medication-related data, we employed tokenization of SMILES (Simplified Molecular-Input Line-Entry System) strings using a dual-view molecule pretraining model [73]. Patient text representations for diagnoses, procedures, and medications were encoded through Transformer encoders. The embedding information was processed according to the identified causal graph structure. In this graph, X_i^D (diagnosis) is considered the intervention variable, while Y_i (mortality) serves as the outcome variable. Both procedures and medications are positioned within the path from diagnosis to outcome. Each of these features (procedures and medications) was concatenated with the nodes in their respective causal paths. The concatenated representations for procedures and medications were then passed through separate MLPs. Subsequently, these outputs were combined with the diagnosis representation, yielding the final representation, which is used as part of the patient representation. For ICU observational indicators, we included the minimum and maximum values [74, 75] recorded during the first 48 h of ICU admission. These multimodal features were integrated to form the final input for model analysis.

In the prediction module, the input sequence x is first processed by a multi-head self-attention mechanism, which projects x into query Q, key K, and value V tensors. Following the self-attention layer, a Feedforward Neural Network (FFN) is applied. The FFN consists of 2 fully connected layers, where the first layer projects the input to a higher-dimensional intermediate space, followed by a non-linear activation using the Gated Linear Unit with GELU activation (GEGLU) function [76]. The GEGLU function modifies the standard GLU by applying the Gaussian Error Linear Unit (GELU) activation to the gating mechanism. GEGLU enhances the non-linearity of the model by splitting the input into 2 parts, x_1 and x_2, and applying the GELU activation to x_2 while multiplying it with $x_1 x = [x_1, x_2]$. The functions are defined as follows:

$$\mathrm{GELU}(x) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}} \cdot (x + 0.044715x^3)\right)\right) \quad (5)$$

$$\mathrm{GEGLU}(x) = x_1 \cdot \mathrm{GELU}(x_2) \quad (6)$$

Following this, the output is further processed through a series of linear layers with ReLU activations and dropout regularization, ultimately leading to a binary classification.

In the treatment effect estimation module, CRISP utilze Average Treatment Effect [4] (ATE) [77–79] to evaluate the overall average effect of the intervention across the entire population, as shown in Eq. 3. Each patient has 2 potential outcomes: survival or death. If both potential outcomes could be fully observed for each patient, causal inference would be straightforward. However, we can never observe the outcomes for both states of the same patient simultaneously. During training, predictions are classified into positive and negative cases based on the ground truth. The ATE for positive cases is estimated as $Y'_{d=1} - Y'_{d=0}$, where d indicates the treatment status. CRISP integrates the ATE into the loss function. Previous research has explored incorporating causal inference into deep learning loss design to penalize the difference between the treated and untreated populations. For example [80, 81], took individual covariates as input to predict treatment assignment (control or treatment group). Their models were trained using a joint loss function, which combines binary cross-entropy (BCE) loss with a bias loss term that computes the mean squared error between inverse probability weight weighted means of treatment and control covariates. In this study, CRISP also uses BCE loss and ATE loss to penalize the model based on how accurately it predicts the difference in outcomes between counterfactual groups. This approach gives more weight to instances where the true outcome is near the decision boundary and where the difference in predicted probabilities is more pronounced. The overall training loss is defined as the combination of the BCE loss and the ATE loss, with α as the hyperparameter determining the weight of the ATE loss. This ensures that the model not only predicts mortality accurately but also learns to align with the causal relationships represented in the data. We select the value of α that maximizes the model's performance on the validation set, balancing the importance of accurately predicting mortality with capturing the causal relationships represented in the data. The overall loss function is given by:

[4] Definition is provided in the Supplementary Material.

$$L_{All} = \begin{array}{l} \alpha \cdot (-(y \cdot \log(y') + (1-y) \cdot \log(1-y'))) \\ +(1-\alpha) \cdot (-\log(\sigma(y'_{d=1} - y'_{d=0}))) \end{array} \quad (7)$$

Here, L_{All} is the total loss, α is the hyperparameter weight, $\sigma(\cdot)$ is the sigmoid function, y is the true label(0 or 1), and y' is the predicted probability.

Baselines

This study evaluated CRISP using 3 dataset and compared its performance against several baseline methods. (1) SAPS-II [18], a scoring system designed for predicting mortality in critically ill patients, considers 12 physiological features within the first 24 h of ICU admission, calculates score by summing the points for each feature, and then uses Eq. 8 to convert the score into a probability of mortality; (2) GRU-D (Deep learning model based on Gated Recurrent Unit) [82], a deep learning model that utilizes recurrent neural networks for analyzing multivariate time series data; (3) IPNET (Interpolation-Prediction NETworks) [83], a novel deep learning architecture based on the use of a semiparametric interpolation network; (4) MC (multitask channel-wise LSTM) [31], an enhancement of LSTM networks that capitalizes on channel-wise LSTMs to predict multiple tasks simultaneously with a single neural model; (5) MTRNN [34], a multi-task recurrent neural network with attention mechanisms specifically designed to predict hospital mortality; (6) IHM-AS [84], a deep learning model using Natural language processing (NLP) techniques to predict in-hospital mortality; (7) MultiModal-1DCNN [38], a deep neural network architecture that combines recurrent neural networks for processing time-series data with convolutional neural networks for analyzing clinical notes; (8) Vital + EntityEmb [85], a multimodal neural network that jointly trains time series signals and unstructured clinical text representations to predict the in-hospital mortality risk for ICU patients; (9) DECAF [25], a general deep cascading framework to predict the potential risks of all physiological functions at each clinical stage; (10) GAN (c-med GAN) [86], a variant of Generative Adversarial Network (GAN) called conditional medical GAN used to predict mortality among ICU inpatients; (11) MMDL (MultiModal Deep Learning) [87], a multimodal deep learning model composed of an ensemble of Feedforward Neural Networks and Gated Recurrent Unit networks; (12) conventional machine learning models including logistic regression (LR), Support Vector Machine (SVM), random forest (RF) and eXtreme Gradient Boosting (XGBoost) based on tabular features, respectively.

$$Mortality = \frac{1}{1 + e^{7.7631 - 0.07372 \times SAPS - 0.9971 \times \ln(SAPS+1)}} \quad (8)$$

Implementation details

The proposed model was implemented using Python 3.8 and PyTorch 1.12, and trained on NVIDIA Titan XP GPUs. Data instances, grouped by unique ICU stays, were split into training, validation, and testing sets with a ratio of 6:1:3. For the traditional ML models (LR, SVM, RF and XGBoost), default parameters were utilized, as extensive grid search revealed minimal performance variation with fine-tuning of base algorithms using different parameter settings. This setting aligns with related works such as [87, 88], where mortality prediction models also relied on default parameters to achieve their results.

For CRISP, hyperparameter tuning was performed on the validation set using Optuna [89], with the search range for each parameter as follows: the number of Transformer layers L was selected from [2, 4, **6**, 8], the number of attention heads from [2, 4, 6, **8**], the dropout rate from [0.1, 0.2, 0.3, **0.75**, 0.9] and hidden embedding size from [[8, 16, 32, 64, 128`]. The α hyperparameter in loss Eq. 7 was chosen from [0.1, 1]. Training was conducted with a learning rate of 0.0001, using the Adam optimizer. Model performance was evaluated on the test set using a comprehensive set of metrics. To estimate 95% confidence intervals (CIs), bootstrapping was applied with 1,000 resamples, sampling with replacement on mortality prediction probabilities.

Metrics

Using a combination of metrics is essential for obtaining a comprehensive evaluation of the model's performance. Metrics such as the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and the Matthews Correlation Coefficient (MCC) provide distinct insights into different aspects of the model's effectiveness, ensuring a well-rounded assessment across multiple performance dimensions.

Statistical analysis

The study population was characterized using descriptive statistics to compare ICU patients who died in the hospital with those who survived. We ran univariate statistics for patient demographics and the predictors of interest. Frequency and percentage were used to describe the categorical variables, and the Chi-square test was used to identify differences between groups. All continuous variables were described using the median and interquartile range, and the Mann–Whitney U-test was used to determine differences between different groups.

Data source and data preprocessing

This study incorporated 3 datasets from 2 large-scale general hospitals. The model was developed using the MIMIC-III (V 1.4) dataset [64] from the Beth Israel

Deaconess Medical Center, covering the period from 2001 to 2012, and validated using the MIMIC-IV dataset (V 3.1) [65], which includes patient data from the same center between 2014 and 2022. Both MIMIC-III and MIMIC-IV are publicly available critical care databases maintained by the Massachusetts Institute of Technology (MIT)'s Laboratory for Computational Physiology. As previously noted, with an aging population, the proportion of older patients admitted to the ICU is expected to continue rising. To rigorously evaluate the model's performance across different patient distributions, a dataset focusing on older patients (age ≥ 65 year) was also used, sourced from the West China Hospital of Sichuan University, one of China's largest and most prestigious medical centers, covering the period from 2019 to 2021. No sensitive information, such as patient identities or contact details, was included. Patient anonymization was performed prior to data extraction and analysis.

Cohort selection

For multiple admissions or ICU admissions of the same patient, only data from the first admission and ICU admission were included, to preserve the independence assumption of the dataset instances. For the MIMIC III and IV cohorts, this study focuses on adult patients admitted to ICU for any reason [31], adhering to the exclusion criteria detailed in Fig. 1. In the MIMIC-III dataset, 30,844 cases were included, with a median age of 66 years, a median ICU stay of 2.50 days (Q1–Q3: 1.59–4.71), 57.34% male patients, and an in-hospital mortality rate of 10.59%. The MIMIC-IV dataset included 27,362 cases, with a median age of 66 years, a median ICU stay of 2.55 days (Q1–Q3: 1.59–4.86), 58.18% male patients, and a mortality rate of 11.62%. Table 1 shows further details. For the external cohort WCHSU, cases lacking postoperative follow-up information were excluded as the true value of patients' postoperative death outcome could not be obtained. Following the exclusion criteria depicted in the Fig. 1, 10,984 unique cases were included in the validation cohort. The median age of this cohort was 71.98, with a median LOS in the ICU of 3.28 days. The patient characteristics of the WCHSU cohort are provided in the Supplementary Materials STable 1. Approximately 56% of these cases were male, and the in-hospital mortality rate was around 0.90%.

Feature extraction

This study delved into patients' clinical records, capturing a wide array of information such as demographics, diagnoses, vital signs, laboratory measurements, procedures, medications, and mortality indicators. All datasets underwent feature selection based on frequency of occurrence and missing rates. We focused on the most commonly encountered diagnoses, procedures, and medications. The retained diagnoses, procedures, and medications features are listed in Supplementary STable 2. To ensure consistency and international comparability, we used the International Classification of Diseases (ICD) codes. Additionally, we expanded the dataset by incorporating textual descriptions for diagnoses and procedures, derived from the corresponding ICD codes in the dictionary tables. The use of the first 48-h time series data is well-supported in the literature, with several studies demonstrating its effectiveness [32]. This study extracted 31 features related to vital signs and laboratory measurements from the ICU observational data (Supplementary STable 3).

Multiple pre-processing steps were undertaken to enhance the quality of the data extracted from both datasets. Inconsistencies in the recording, including units of certain variables, were addressed following [87]. Outlier detection and correction were conducted to ensure data accuracy, involving the identification of physician input errors based on medical common sense and the implementation of corrective measures. Outliers were detected by establishing acceptable ranges for each feature, with values outside these ranges removed (Supplementary STable 3). To normalize continuous features, we applied min-max normalization using the *sklearn.preprocessing.MinMaxScaler* [90]. Categorical features were encoded using label encoding with the *sklearn.preprocessing.LabelEncoder* [90], transforming target labels into values ranging from 0 to n_classes-1. For handling missing data, we employed the last observation carried forward method for time series data, while other missing values were imputed using the median.

Experiments and results

Performance on data sampling

Highly imbalanced data poses a significant challenge to obtaining reliable results, often leading to classifier bias in favor of the majority class. To address this, we conducted a thorough comparison of multiple data sampling methods on 3 datasets in Fig. 3. To compare the model's performance in class balance, 2 evaluation metrics were used: AUPRC and MCC. AUPRC evaluates the model's ability to distinguish between classes by considering both precision and recall, offering a comprehensive view of its performance, especially in identifying the minority class. MCC provides a balanced measure of classification accuracy, accounting for true positives, true negatives, false positives, and false negatives, and is particularly useful for imbalanced datasets.

Figure 3 shows the performance of models trained on MIMIC-III$_{train}$ and tested on MIMIC-III$_{test}$. The models' performance on the original data ranges from an AUPRC of 0.613 (LR) to 0.751 (CRISP), with MCC scores between

Table 1 Baseline characteristics of the included patients from MIMIC III and IV. Descriptive statistics of the patient cohort in the experimental set are provided. Part of ICU observational data variables are expressed as Median (InterQuartile Range Q1–Q3), and binary or categorical variables are shown as Count (%). Specifically, 'Length of Stay' is abbreviated as 'LOS', 'Heart Rate' is abbreviated as 'HR', 'Respiratory rate' as 'RR', 'Temperature' as 'Temp', 'Bicarbonate' as 'HCO_3', 'Blood urea nitrogen' as 'BUN', as 'BPm', 'Diastolic blood pressure' as 'DBP', 'Systolic blood pressure' as 'SBP', 'Glascow coma scale tota' as 'GCS', 'Red Cell Distribution Width' as 'RDW', and 'White blood cell count' as 'WBC'

	MIMIC III		**MIMIC IV**	
	Dead at hospital	**Alive at hospital**	**Dead at hospital**	**Alive at hospital**
ICUSTAY	3265	27579	3179	24183
LOS	4.23 (2.16–8.48)	2.32 (1.53–4.20)	4.02 (2.08–8.45)	2.40 (1.54–4.42)
Age	74.00 (60.00–82.00)	65.00 (52.00–76.00)	71.0 (59.00–81.00)	66.0 (55.00–76.00)
Gender				
Female	1519 (46.52%)	11638 (42.20%)	1409 (44.32%)	10033 (41.49%)
Male	1746 (53.48%)	15941 (57.80%)	1770 (55.68%)	14150 (58.51%)
Type				
Emergency	3075 (94.18%)	21762 (78.91%)	1604 (50.46%)	9238 (38.20%)
Elective	132 (4.04%)	5318 (19.28%)	20 (0.63%)	1029 (4.26%)
Urgent	58 (1.78%)	499 (1.81%)	853 (26.83%)	4949 (20.46%)
Features				
HR	89.09 (78.75–100.28)	83.73 (75.24–92.58)	88.5 (77.00–101.00)	81.0 (72.00–91.00)
RR	20.48 (17.69–23.43)	18.35 (16.34–20.69)	21.0 (18.00–24.00)	18.0 (16.00–20.00)
Temp	36.80 (36.31–37.25)	36.84 (36.55–37.17)	36.86 (36.61–37.11)	36.86 (36.72–37.06)
Albumin	3.10 (2.50–3.10)	3.10 (3.10–3.10)	3.10 (2.80–3.10)	3.10 (3.10–3.10)
Anion gap	15.0 (13.00–18.00)	13.0 (11.00–14.00)	15.0 (13.00–18.50)	13.0 (11.00–15.00)
HCO_3	22.79 (19.20–25.11)	24.67 (22.69–26.58)	21.0 (18.00–24.00)	23.0 (21.00–25.00)
BUN	28.0 (17.50–46.00)	16.5 (12.00–25.00)	29.0 (18.00–46.25)	16.0 (12.00–24.00)
Calcium	8.2 (7.70–8.65)	8.3 (7.95–8.65)	8.3 (7.80–8.75)	8.4 (8.00–8.80)
Creatinine	1.2 (0.80–2.05)	0.92 (0.70–1.55)	1.3 (0.90–2.25)	0.9 (0.70–1.20)
DBP	57.43 (51.18–64.03)	61.07 (55.07–67.72)	63.0 (55.00–69.00)	65.0 (58.00–73.00)
SBP	113.60 (102.64–127.83)	120.29 (111.02–131.50)	112.0 (101.50–123.00)	116.0 (107.00–129.00)
GCS	13.0 (8.00–15.00)	15.0 (15.00–15.00)	9.0 (4.00–14.00)	15.0 (12.00–15.00)
Glucose	134.5 (113.00–162.00)	122.5 (105.50–143.00)	136.0 (114.50–170.25)	124.0 (107.00–145.00)
Sodium	139.26 (136.37–142.30)	138.73 (136.64–140.80)	139.0 (135.00–142.50)	138.0 (136.00–140.50)
pH	7.26 (7.10–7.37)	7.28 (7.10–7.39)	7.37 (7.30–7.39)	7.375 (7.36–7.39)
Platelets	184.0 (113.00–252.00)	191.0 (143.00–250.00)	163.0 (101.50–228.25)	175.0 (133.00–231.00)
RDW	15.3 (14.30–17.05)	14.3 (13.50–15.60)	15.4 (14.10–17.60)	14.0 (13.10–15.30)
MCHC	33.7 (32.50–34.50)	34.15 (33.25–35.05)	32.5 (31.45–33.45)	32.9 (32.00–33.80)
WBC	12.32 (9.31–16.92)	10.77 (8.21–13.70)	12.55 (9.22–17.67)	11.0 (8.30–14.35)

0.491 (RF) and 0.552 (CRISP). After applying SMOTE, ADASYN, SMOTE-ENN, and SMOTE-Tomek for class balancing, the models' AUPRC ranged from 0.617 (LR with SMOTE-ENN) to 0.755 (CRISP with SMOTE), with MCC scores ranging from 0.459 (LR with SMOTE-ENN) to 0.664 (CRISP with SMOTE). All models trained on CMG-processed data outperformed the baseline, with MCC scores ranging from 0.528 (LR) to 0.668 (CRISP). Among the models, CRISP with CMG achieved the best overall performance, while SMOTE and ADASYN also delivered competitive results.

Figure 3 shows the performance of models trained on MIMIC-III$_{train}$ and tested on MIMIC-IV$_{test}$. The models' performance on the original data ranges from an AUPRC of 0.602 (LR) to 0.634 (CRISP). SOMTE outperforms other baseline algorithms in most scores, but it has not consistently surpassed the baseline scores. CMG's scores exceed the baseline scores, particularly achieving an AUPRC of 0.668 and an MCC of 0.584 under the CRISP model. This experiment demonstrates the performance of both CMG and the CRISP model, with CRISP and CMG offering more stable results.

Figure 3 shows the performance of models trained on WCHSU$_{train}$ and tested on WCHSU$_{test}$. Due to the more severe class imbalance in the WCHSU dataset, all models performed poorly on AUPRC and MCC, with scores below 0.5. The performance of SMOTE, ADASYN, SMOTE-ENN, and SMOTE-Tomek was similar across LR, SVM, RF, and XGBoost models. CMG outperformed these methods on the MCC metric, achieving scores ranging from 0.206 (LR) to 0.429 (CRISP). This experiment highlights that, under conditions of label

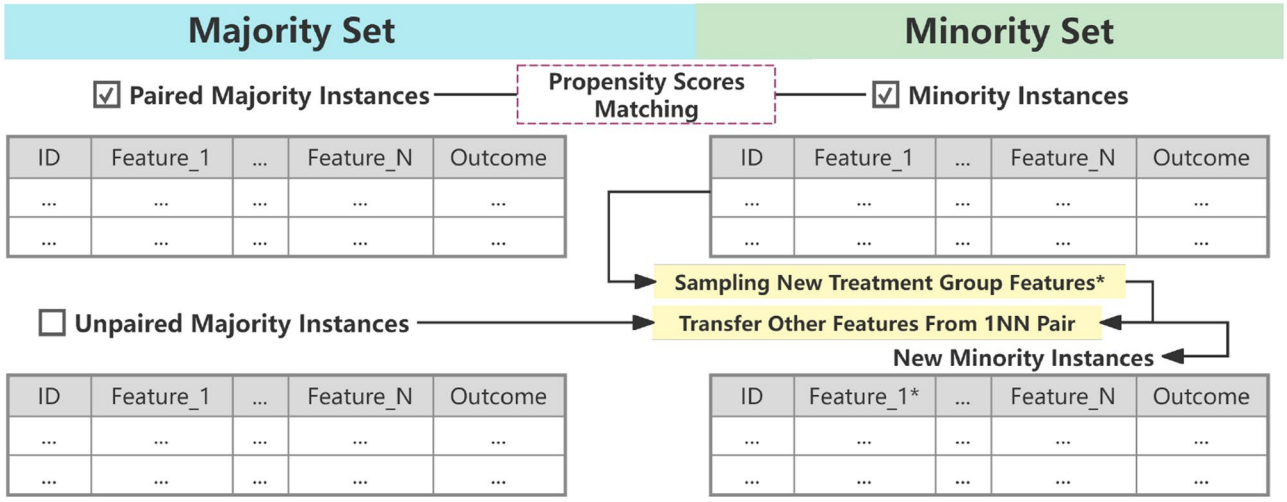

Fig. 2 The framework of Counterfactual Minority Generation Module (CMG). The goal is to generate new minority instances for unpaired majority instances

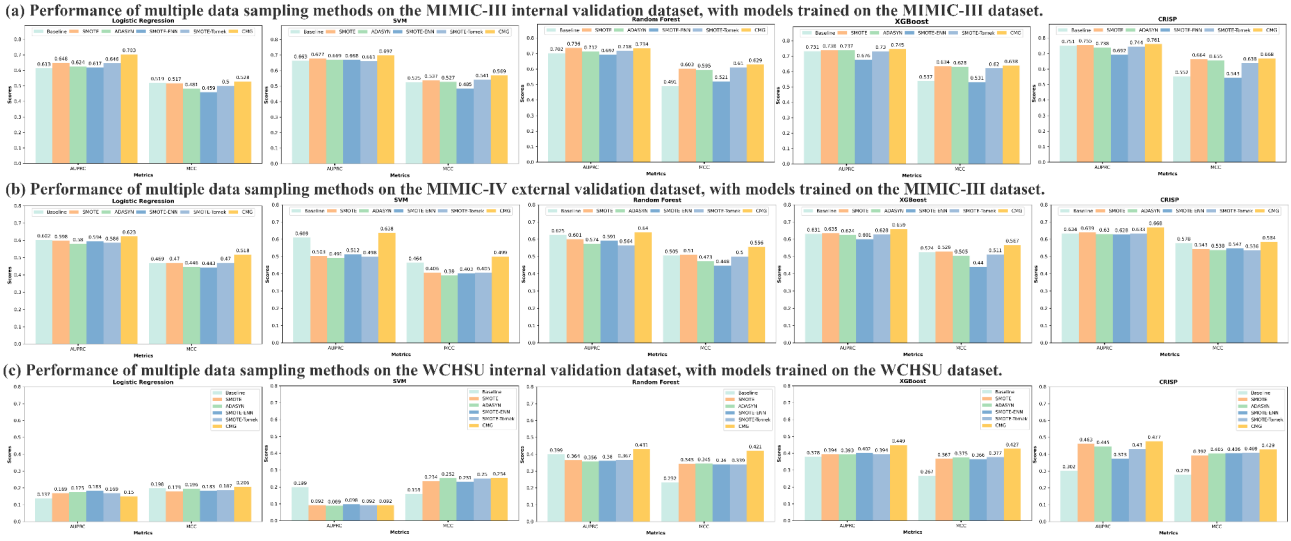

Fig. 3 Comparison of data sampling methods. (a) Performance of multiple data sampling methods on the MIMIC-III internal validation dataset, with models trained on the MIMIC-III dataset. (b) Performance of multiple data sampling methods on the MIMIC-IV external validation dataset, with models trained on the MIMIC-III dataset. (c) Performance of multiple data sampling methods on the WCHSU validation dataset, with models trained on the WCHSU dataset

imbalance, not all models benefit from label augmentation techniques.

Superior performance to traditional models

In this study, we conducted a comprehensive comparison of our model's performance with well-established traditional ML algorithms for predicting predefined outcomes across three datasets. We specifically selected algorithms commonly used in clinical settings for their robustness and efficiency. The WCHSU dataset consisted of 10,984 ICU cases from the West China Hospital of Sichuan University, with a median age of 71.98. While over 70% of the patients in MIMIC-III and MIMIC-IV are White, the 3 datasets introduced variability in data distribution.

Across all 3 experiments, the CRISP model achieved competitive results on all datasets. Taking the prediction on the CMG balanced datasets as an example, Table 2 illustrates the performance of each model. In the evaluation set MIMIC-III$_{test}$, CRISP model achieved an AUROC score of 0.9480, along with an AUPRC score of 0.7611 and MCC of 0.6678, surpassing those of XGBoost (AUROC: 0.9425, AUPRC: 0.7447, MCC: 0.6382). When validated on MIMIC-IV$_{test}$, all models achieve AUROC scores ranging from 0.9036 to 0.9171, with CRISP reaching the highest AUROC of 0.9171, an AUPRC of 0.6683,

Table 2 Performance of models trained on CMG sampling sets is presented, with mean values of corresponding 95% confidence interval for each metric. The best value for each metric is highlighted in bold

Models	Evaluation Set	AUROC	AUPRC	MCC
MIMIC-III_LR	MIMIC-IIItest	0.9249	0.7025	0.5281
MIMIC-III_SVM	MIMIC-IIItest	0.9169	0.6973	0.5689
MIMIC-III_RF	MIMIC-IIItest	0.9463	0.7335	0.6290
MIMIC-III_XGBoost	MIMIC-IIItest	0.9425	0.7447	0.6382
MIMIC-III_CRISP	MIMIC-IIItest	**0.9480**	**0.7611**	**0.6678**
MIMIC-III_LR	MIMIC-IVtest	0.9036	0.6231	0.5176
MIMIC-III_SVM	MIMIC-IVtest	0.9093	0.6376	0.4992
MIMIC-III_RF	MIMIC-IVtest	0.9124	0.6400	0.5560
MIMIC-III_XGBoost	MIMIC-IVtest	0.9127	0.6594	0.5666
MIMIC-III_CRISP	MIMIC-IVtest	**0.9171**	**0.6683**	**0.5838**
WCHSU_LR	WCHSUtest	0.8492	0.1500	0.2061
WCHSU_SVM	WCHSUtest	0.8366	0.0920	0.2543
WCHSU_RF	WCHSUtest	0.8930	0.4314	0.4211
WCHSU_XGBoost	WCHSUtest	0.8889	0.4487	0.4270
WCHSU_CRISP	WCHSUtest	**0.9042**	**0.4771**	**0.4289**

Table 3 Prediction results for the different models on MIMIC-III$_{test}$ dataset, where the best AUROC and AUPRC are in bold

Type	Models	AUROC	AUPRC
Score-based	SAPS II	0.805	0.337
Time series	GRU-D	0.835	0.359
Time series	IPNET	0.853	0.418
Time series	MC	0.870	0.533
Time series	MTRNN	0.895	0.520
Text	IHM-AS	0.918	0.625
Hybrid	MultiModal - 1DCNN	0.865	0.525
Hybrid	Vital + EntityEmb	0.874	0.529
Hybrid	DECAF	0.893	-
Hybrid	GAN (c-med GAN)	0.910	0.532
Hybrid	MMDL	0.940	**0.772**
Hybrid	CRISP	**0.948**	0.761

and an MCC score of 0.5838. In the evaluation set WCHSU$_{test}$, CRISP model achieved an AUROC score of 0.9042, along with an AUPRC score of 0.4771 and MCC of 0.4289, again outperforming XGBoost (AUROC: 0.8889, AUPRC:0.4487, MCC:0.4270). These findings demonstrate the stability of our model in predicting mortality. The AUROC curves and AUPRC curves on MIMIC-III$_{test}$ and WCHSU$_{test}$ are shown in Supplementary Material SFig. 2. Our model's performance across datasets indicates its potential for enhancing clinical decision-making and patient care.

Superior performance to deep learning models

This section compares the prediction results with recent works using the widely used public MIMIC-III dataset. Table 3 presents the performance of various prediction algorithms for the in-hospital mortality prediction task on the MIMIC-III dataset. SAPS II, as a traditional scoring system, performs the worst. Most models report AUROC scores below 0.9 and AUPRC scores below 0.7. CRISP and MMDL show a consistent advantage across both metrics. Notably, our proposed model achieves a slightly higher AUROC compared to all others. These results suggest that hybrid models, which integrate multimodal data such as time-series data and clinical notes, deliver superior predictive performance. This trend highlights the strength of deep learning models in capturing complex patterns and extracting meaningful representations from heterogeneous data, thereby improving prediction outcomes for in-hospital mortality benchmarks.

Table 4 Ablation study for CRISP on MIMIC III dataset. AUROC performance comparison over different feature sets. 'Tabular' denotes categorical variables and numerical variables. 'Text' denotes diagnoses, procedures, and medications texts. 'CasualI' indicates results with causal inference

	AUROC	AUPRC	Recall	F1
Tabular	0.9381 (0.9377–0.9386)	0.7523 (0.7516–0.7531)	0.7185 (0.7073–0.7297)	0.6832 (0.6821–0.6844)
Tabula|CasualI	0.9390 (0.9387–0.9393)	0.7570 (0.7544–0.7596)	0.7388 (0.7124–0.7652)	0.6911 (0.6886–0.6936)
Text	0.8963 (0.8958–0.8968)	0.5783 (0.5762–0.5803)	0.6677 (0.6514–0.6839)	0.5594 (0.5578–0.5609)
Text|CasualI	0.8973 (0.8972–0.8974)	0.5800 (0.5772–0.5828)	0.7033 (0.6738–0.7327)	0.5684 (0.5619–0.5748)
Tabular + Text	**0.9484 (0.9478–0.9490)**	0.7597 (0.7578–0.7616)	**0.7175 (0.7093–0.7256)**	0.6928 (0.6903–0.6953)
Tabular + Text|CasualI	0.9480 (0.9477–0.9483)	**0.7611 (0.7594–0.7628)**	**0.7612 (0.7154–0.8069)**	**0.6939 (0.6909–0.6968)**

Although the improvements in the CRISP model are relatively subtle, the enhancement in AUPRC compared to most other models is notable. Achieving significant improvements in AUROC above 0.9 is challenging, and while the increase may be small, the true value of this study lies in the underlying concept of the CRISP model. By incorporating causal structures into patient data representation, CRISP introduces a novel approach that provides fresh insights.

We conduct the ablation study for CRISP model, systematically comparing results across various feature inputs and assessing the impact of incorporating causal inference into the model. Table 4 demonstrates that models incorporating causal inference tend to perform better than those without, though the improvement is marginal. These results highlight the benefits of integrating causal relationships and diverse data types, which enhance the model's ability to capture a broader range of patterns

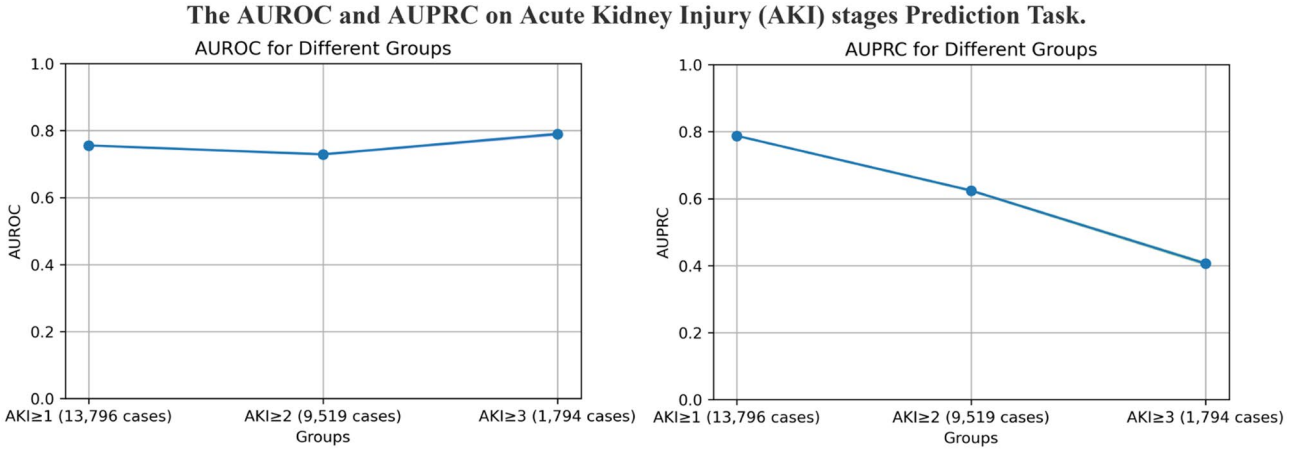

Fig. 4 The AUROC and AUPRC performance of CRISP on prediction of Acute Kidney Injury stages (1, 2, 3) within the initial 48 h of ICU admission

and relationships, ultimately leading to more reliable predictions.

Performance in other clinical outcome Scenario

Researchers are increasingly exploring the detection of risks related to severe medical conditions such as sepsis [91, 92], Acute Kidney Injury (AKI) [93, 94], and other clinical deterioration events [95]. Acute conditions are often reversible if detected and treated promptly. Therefore, to validate our CRISP's applicability in other scenarios, validating our prediction model on acute medical conditions is reasonable to check if it has the ability to recognize illness escalation to severe cases timely, reducing the risk of further damage and early mortality. In this study, we focused on AKI outcomes, as it is the most common and severe syndrome of renal failure [96], which is accompanied by high mortality and disease burden. Despite advancements in clinical treatment, the mortality rate linked with AKI remains concerning, especially in high-risk groups like sepsis patients, where it can reach as high as 41% [97]. It affects at least 30% of hospitalized patients [98] and is also associated with increased serious complications. The Kidney Disease: Improving Global Outcomes (KDIGO) criteria categorize AKI into 3 stages based on serum creatinine levels and urine output, offering a standardized approach for diagnosis and research [99]. Comparing to patients with AKI stage 1, patients with AKI stage 2 or 3 have higher in-hospital mortality and risk of progression to chronic kidney disease.

We implemented CRISP to predict AKI within the initial 48 h of ICU admission on MIMIC III dataset. We used the same group of patients from the MIMIC III dataset used for the mortality prediction task and adhered to the KDIGO's AKI definition to classify AKI into 3 severity stages (AKI stage $\geq 1, \geq 2$, and ≥ 3). We carefully excluded individuals with pre-existing AKI, renal failure, and chronic kidney disease prior to their ICU admission. The exclusion of patients with prior kidney-related conditions serves to minimize potential confounding factors influencing the escalation of AKI severity. This strategic exclusion enhances the internal validity of our study, allowing us to focus our assessment on the predictive capabilities of our models within a population devoid of pre-existing kidney disorders. Our finalized dataset comprises a total of 24,661 ICU stays (13,796 cases developed AKI stage ≥ 1, 9,519 cases developed AKI stage ≥ 2, 1,794 cases developed AKI stage ≥ 3).

As shown in Fig. 4, our prediction model achieved AUROCs ranging from 0.7557 to 0.7899 and AUPRCs ranging from 0.4061 to 0.7877. In comparison, a recent study on this dataset reported an AUROC of 0.7798 [100]. CRISP's AUROCs exhibited more stable performance with class imbalance. It emphasizes the importance of continuously evaluating and adapting the model to keep pace with evolving clinical landscapes.

Discussion

Numerous medical facilities are already utilizing EHR systems. Compared to traditional manual scoring systems, ML can offer superior predictive performance and can even automate clinical decision-making by leveraging the extensive data collected from EHRs [13]. Recent research highlights causal ML as a pivotal intersection between AI and statistics, marking a transition from mere prediction to a deeper understanding. This shift bridges the gap between traditional statistical inference and advanced predictive capabilities [101].

In this study, we not only proposed a novel causality based mortality prediction model but also demonstrated the effectiveness of various conventional ML algorithms in predicting in-hospital mortality among ICU patients. Causality is a new frontier in deep learning, capable of reducing data-driven errors by revealing hidden causal relationships within complex distributions. In this study,

CRISP's transportability is validated under a defined causal structure for the external validation of intervention models and counterfactual queries. By integrating causal structure, CRISP achieves enhanced stability across diverse settings. Clinicians and healthcare practitioners can modify the causal DAG graph using domain-specific knowledge to better guide the model's learning. Unlike traditional approaches that establish causal relationships based on individual features, CRISP builds relationships based on feature categories. This reduces the reliance on specific features within categories, making it easier to adapt the model to different settings. This generalizability greatly contribute to the utility of the model in real-world medical applications. Other studies are also integrating ML into causal inference methods. For instance [102], demonstrated that RF combined with the potential outcomes approach can effectively detect and estimate heterogeneity of treatment effects across multiple covariates considered simultaneously. Our work also demonstrates the potential of exploring the application of causal machine learning in clinical settings.

This study introduces the CMG Module, which integrates causality into the counterfactual generation process. Different from existing studies [58–60, 70], CMG incorporates propensity scores to guide the search for counterfactual pairs and generates new counterfactual instances for unmatched majority class observations through causal pathways. The implementation of the CMG method further enhanced the performance of the prediction models in the experimental results of this study. However, it's important to note CMG's limitations, such as its inability to generate instances based on time series or text features. Generating new counterfactual samples relies on a limited number of minority samples, potentially introducing bias. We anticipate conducting additional experiments on diverse datasets to gain deeper insights and address these challenges.

In this study, we used patient data from hospitals in 2 different countries to test the model's applicability across diverse source datasets. In the MIMIC-III and IV datasets, over 70% of the patients are white, while the WCHSU dataset comprises only Asian patients from the West China Hospital of Sichuan University. Our CRISP and CMG models demonstrate stable performance across these different datasets.

Interpretation. There are various methods for interpreting how a deep learning model works from different perspectives. One commonly used approach is feature importance estimation, which provides a straightforward understanding from the perspective of domain experts. This study uses Permutation Importance [103] to identify the features that contribute most to predictions, thereby enhancing interpretability (the top 15 most important features are ranked in Supplementary Material STable 4). This research includes a broad range of features predictive of mortality in the ICU setting, such as patient demographics, diagnoses, procedures, medications, vital signs, and lab tests. Consistent with existing literature, our findings highlight the importance of vital signs and laboratory tests in predicting ICU mortality [2]. For example, earlier studies have emphasized the critical role of parameters like blood pressure and oxygen saturation in assessing the risk of mortality in critically ill patients [104]. Procedures such as continuous invasive mechanical ventilation further underscore the importance of managing respiratory failure in determining outcomes [105].

Limitations. Firstly, although DL approaches can achieve superior performance, they may face challenges in standardizing clinical predictive indicators. In contrast, conventional scoring systems like SAPS II, despite having relatively lower predictive performance, excel in standardization and facilitate center-to-center comparisons. Therefore, using DL models alongside traditional scoring systems can provide more valuable insights for predicting the prognosis of critically ill patients and for comparing ICU performance. Secondly, future research should focus on exploring the causal relationships between temporal data during ICU stays, further enhancing the framework's ability to identify causal factors. Lastly, incorporating clinician feedback and investigating the design of more detailed causal graphs would help continually refine the model, improving its adaptability and resilience.

Implications To discuss the potential clinical use of our model and implications for future research, we applied our model for the comprehensive assessment of Acute Kidney Injury (AKI). Timely identification and proactive intervention in AKI are crucial, given its potential for prevention and reversibility within a relatively short timeframe, spanning from a few hours to several days [106]. Early detection plays a crucial role in mitigating the progression of AKI, leading to a reduction in elevated mortality rates among vulnerable ICU patients. Our prediction model achieved relatively good performance. Our research underscores the importance of vigilant monitoring of vital signs in ICU patients.

Conclusion

In conclusion, this study introduces CRISP, a causal deep neural network architecture that integrates causality into both data augmentation and the prediction model. The CMG module, which leverages causal graphs to generate new minority class instances through native counterfactuals, demonstrates predictive performance comparable to traditional oversampling techniques. CRISP, by incorporating causal graph with the processing of feature groups alongside basic patient information, shows competitive performance in mortality prediction. Its effectiveness is further validated on the MIMIC-IV dataset and a dataset from

West China Hospital, showcasing the model's generalizability across different data distributions, including diverse patient demographics and clinical outcomes. These findings highlight the potential of causality-based approaches in real-world clinical applications, emphasizing their flexibility and adaptability across various datasets and tasks.

Abbreviations

AKI	Acute Kidney Injury
AUP	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic
AT	Average Treatment Effect
BCE	Binary Cross-Entropy
CRISP	Clinical Representation and Inference for Structured Prediction
CMGM	Counterfactual Minority Generation Module
DAG	Directed Acyclic Graph
eGB	eXtreme Gradient Boosting
EHR	Electronic Health Record
FNN	Feedforward Neural Network
GELU	Gated Linear Unit with GELU activation
GA	Graphical Autoencoder
ICU	Intensive Care Units
ICD	International Classification of Diseases
IGO	The Kidney Disease: Improving Global Outcomes
K	K-nearest neighbors
LS	Length of Stay
LR	Logistic Regression
ML	Machine Learning
MIMIC	Medical Information Mart for Intensive Care
MP	Multilayer Perceptrons
NN	Neural Networks
RF	Random Forest
SAPS	Simplified Acute Physiology Score II
SM	Simplified Molecular-Input Line-Entry System
SMOT	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WCHSU	The West China Hospital of Sichuan University

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02981-1.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
L.W.: Conceptualization, Methodology, Data Curation, Validation, Visualization, Writing, Review and Editing. X.G.: Conceptualization, Formal analysis, Data Curation, Writing. H.S.: Data Curation, Writing. Y.M.: Data Curation, Writing. H.B.: Data Curation, Writing. L.J.: Conceptualization, Reviewing and Editing. L.Z.: Conceptualization, Reviewing and Editing. T.Z.: Resources, Reviewing and Editing, Funding acquisition. Z.F.: Conceptualization, Supervision, Reviewing and Editing.L.L.: Conceptualization, Investigation, Supervision, Reviewing and Editing, Funding acquisition. All authors read and approved the final manuscript.

Funding
This work was supported by the Science and Technology Department of Sichuan Province (Project 2019YFG0491), Chengdu Science and Technology Bureau (Project 2024-YF05-00900-SN).

Data availability
External data that support the findings of this study have been deposited in the Chinese Clinical Trial Registry website (https://www.chictr.org.cn/indexEN.html), with the registration number ChiCTR1900025160. The recruitment period for the data was from August 5, 2019, to August 31, 2021.

Code availability
https://github.com/Leyayaya251/CRISP

Materials availability
Not applicable.

Declarations

Ethics approval and consent to participate
The MIMIC-III and IV datasets were approved by the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center, with informed consent obtained for data collection. Data obtained from the West China Hospital of Sichuan University were covered by a Chinese clinical trial titled 'Study for Machine-Learning Based Perioperative Risk Assessment and Prediction System for Geriatric Patients'. The study protocol was approved by the ethics committee of the West China Hospital of Sichuan University (2019-473) with a waiver of informed consent. More information about the use of these data can be found at the Chinese Clinical Trial Registry website: (http://www.chictr.org.cn). The registration number is ChiCTR1900025160, and the data recruiting period is from August 5, 2019, to August 31, 2021.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 December 2024 / Accepted: 19 March 2025
Published online: 15 April 2025

References

1. Marshall JC, Bosco L, Adhikari NK, Connolly B, Diaz JV, Dorman T, Fowler RA, Meyfroidt G, Nakagawa S, Pelosi P, et al. What is an intensive care unit? a report of the task force of the world federation of societies of intensive and critical care medicine. J Crit Care. 2017;37:270–76.
2. Baker S, Xiang W, Atkinson I. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach. Sci Rep. 2020;10(1):21282.
3. Yeh Y-C, Kuo Y-T, Kuo K-C, Cheng Y-W, Liu D-S, Lai F, Kuo L-C, Lee T-J, Chan W-S, Chiu C-T, et al. Early prediction of mortality upon intensive care unit admission. BMC Med Inf Decis Making. 2024;24:394.
4. Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for united states intensive care unit admissions from 1988 to 2012. Critical Care. 2013;17:1–9.
5. Flaatten H, De Lange D, Artigas A, Bin D, Moreno R, Christensen S, Joynt G, Bagshaw SM, Sprung C, Benoit D, et al. The status of intensive care medicine research and a future agenda for very old patients in the ICU. Intensive Care Med. 2017;43:1319–28.
6. Rai S, Brace C, Ross P, Darvall J, Haines K, Mitchell I, Haren F, Pilcher D. Characteristics and outcomes of very elderly patients admitted to intensive care: a retrospective multicenter cohort analysis. Crit Care Med. 2023;51(10):1328–38.
7. Vincent J-L, Marshall JC, Ñamendys-Silva SA, François B, Martin-Loeches I, Lipman J, Reinhart K, Antonelli M, Pickkers P, Njimi H, et al. Assessment of the worldwide burden of critical illness: the intensive care over nations(icon) audit. Lancet Respir Med 2014;2:380–86.
8. Vallet H, Schwarz GL, Flaatten H, De Lange DW, Guidet B, Dechartres A.Mortality of older patients admitted to an ICU: a systematic review. Crit Care Med. 2021;49(2):324–34.
9. Kurtz P, Bastos LS, Salluh JI, Bozza FA, Soares M. Saps-3 performance for hospital mortality prediction in 30,571 patients with covid-19 admitted to ICUs in Brazil. Intensive Care Med. 2021;47:1047–49.
10. Tekin B, Kiliç J, Taskin G, Solmaz İ, Tezel O, Basgöz BB. The comparison of scoring systems: sofa, apache-ii, lods, mods, and saps-ii in critically ill elderly sepsis patients. J Infect Dev Ctries. 2024;18(01):122–30

11. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Apache ii: a severity of disease classification system. Crit Care Med. 1985;13(10):818–29.
12. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, et al. The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. Chest 1991;100:1619–36.
13. Le Gall J-R, Lemeshow S, Saulnier F. A new simplified acute physiology score(saps ii) based on a european/north american multicenter study. Jama. 1993;270(24):2957–63.
14. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall J-R, et al. Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med. 2005;31:1345–55.
15. Vincent J-L, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C, Suter P, Thijs LG. The SOFA(sepsis-related organ failure assessment) score to describe organ dysfunction/failure: on behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine(see contributors to the project in the appendix). Springer; 1996.
16. Sakr Y, Krauss C, Amaral AC, Rea-Neto A, Specht M, Reinhart K, Marx G. Comparison of the performance of saps ii, saps 3, apache ii, and their customized prognostic models in a surgical intensive care unit. Br J Anaesth. 2008;101(6):798–803.
17. Kramer AA. Predictive mortality models are not like fine wine. Critical Care. 2005;9:1–2.
18. Falcão ALE, Barros AGDA, Bezerra AAM, Ferreira NL, Logato CM, Silva FP, Monte ABFO, Tonella RM, Figueiredo LC, Moreno R, et al. The prognostic accuracy evaluation of saps 3, sofa and apache ii scores for mortality prediction in the surgical ICU: an external validation study and decision-making analysis. Ann Intens Care. 2019;9:1–10.
19. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heimann M, Dybdahl L, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digital Health 2020;2:179–91.
20. Jiang J, Yu X, Wang B, Ma L, Guan Y. Decaf: an interpretable deep cascading framework for ICU mortality prediction. Artif. Intell. Med. 2023;138:102437.
21. Lew CCH, Wong GJY, Tan CK, Miller M: Performance of the acute physiology and chronic health evaluation ii(apache ii) in the prediction of hospital mortality in a mixed ICU in singapore. Proceedings of Singapore Healthcare. 2019;28(3):147–52.
22. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, Xu S, Stub D, Smith K, Tacey M, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: a retrospective international registry study. PLoS Med 2018;15:1002709.
23. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Method. 2019;19:1–18.
24. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. Ann Surg. 2020;272(6):1133–39
25. Wong L, Young J. A comparison of ICU mortality prediction using the apache ii scoring system and artificial neural networks. Anaesthesia. 1999;54(11):1048–54.
26. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6(1):96.
27. Alves T, Laender A, Veloso A, Ziviani N: Dynamic prediction of ICU mortality risk using domain adaptation. In: 2018 IEEE International Conference on Big Data(Big Data), p. 1328–36(2018). IEEE
28. Yu K, Zhang M, Cui T, Hauskrecht M. Monitoring ICU mortality risk with a long short-term memory recurrent neural network. In: Pacific Symposium on Biocomputing 2020. World Scientific; 2019. p. 103–14.
29. Yu R, Zheng Y, Zhang R, Jiang Y, Poon CC. Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients. IEEE J Biomed Health Inform. 2019;24(2):486–92.
30. Grnarova P, Schmidt F, Hyland SL, Eickhoff C. Neural document embeddings for intensive care patient mortality prediction. 2016;arXiv preprint arXiv:1612.00467.
31. Liu N, Lu P, Zhang W, Wang J: Knowledge-aware deep dual networks for text-based mortality prediction. In: 2019 IEEE 35th International Conference on Data Engineering(ICDE), p. 1406–17(2019). IEEE
32. Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. 2019;arXiv preprint arXiv:1904.05342.
33. Khadanga S, Aggarwal K, Joty S, Srivastava J. Using clinical notes with time series data for ICU management. 2019;arXiv preprint arXiv:1909.09702.
34. Yang H, Kuang L, Xia F. Multimodal temporal-clinical note network for mortality prediction. J. Biomed. Semant. 2021;12:1–14.
35. Deznabi I, Iyyer M, Fiterau M. Predicting in-hospital mortality by combining clinical notes with time-series data. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021;4026–31.
36. Morid MA, Sheng ORL, Dunbar J. Time series prediction using deep learning methods in healthcare. ACM Trans Manage Inf Syst. 2023;14(1):1–29.
37. Lee Y, Jun E, Choi J, Suk H-I. Multi-view integrative attention-based deep representation learning for irregular clinical time-series data. IEEE J Biomed Health Inform. 2022;26(8):4270–80.
38. Liu M, Guo C, Guo S. An explainable knowledge distillation method with xgboost for ICU mortality prediction. Comput Biol Med. 2023;152:106466.
39. Leist AK, Klee M, Kim JH, Rehkopf DH, Bordas SP, Muniz-Terrera G, Wade S. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Sci Adv. 2022;8(42):1942.
40. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, Bauer S, Kilbertus N, Kohane IS, Schaar M. Causal machine learning for predicting treatment outcomes. Nature Med. 2024;30(4):958–68.
41. Maley JH, Wanis KN, Young JG, Celi LA. Mortality prediction models, causal effects, and end-of-life decision making in the intensive care unit. BMJ Health Care Inform. 2020;27(3).
42. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. Royal Soc Open Sci. 2022;9(8):220638
43. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun. 2020;11(1):3923.
44. Pearl, J: Causality. Cambridge: Cambridge University Press; 2009.
45. Höfler M. Causal inference based on counterfactuals. BMC Med. Res. Method. 2005;5:1–12.
46. Lagemann K, Lagemann C, Taschler B, Mukherjee S.Deep learning of causal structures in high dimensions under data limitations. Nature Mach Intell. 2023;5(11):1306–16.
47. Zanga A, Ozkirimli E, Stella F. A survey on causal discovery: theory and practice. Int J Approx Reason. 2022;151:101–29.
48. Jallbjørn S, Jarner SF, Hansen NR. Forecasting, interventions and selection: the benefits of a causal mortality model. Europ Actuar J. 2024;14(2):437–66.
49. Rajpurkar P, Chen E, Banerjee O, Topol EJ. Ai in health and medicine. Nature Med. 2022;28(1):31–38.
50. Sauer CM, Dam TA, Celi LA, Faltys M, Hoz MA, Adhikari L, Ziesemer KA, Girbes A, Thoral PJ, Elbers P. Systematic review and comparison of publicly available ICU data sets—a decision guide for clinicians and data scientists. Crit Care Med. 2022;50(6):581–88.
51. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
52. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health ai. Biostatistics. 2020;21(2):345–52.
53. Temraz M, Keane MT. Solving the class imbalance problem using a counterfactual method for data augmentation. Mach Learn Appl. 2022;9:100375.
54. Smyth B, Keane MT: a few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In: International Conference on Case-Based Reasoning, p. 18–32(2022). Springer
55. Keane MT, Smyth B: Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable ai(xai). In: Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28, p. 163–78(2020). Springer
56. Brughmans D, Leyman P, Martens D, Nice: an algorithm for nearest instance counterfactual explanations. Data Min Knowl Discov. 2024;38(5):2665–703
57. Loscalzo J, Kohane I, Barabasi A-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol. 2007;3(1):124.
58. Saxe GN, Ma S, Morales LJ, Galatzer-Levy IR, Aliferis C, Marmar CR. Computational causal discovery for post-traumatic stress in police officers. Transl Psychiatry. 2020;10(1):233.
59. Johnson AE, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. Mimic-iii, a freely accessible critical care database. Sci Data. 2016;3(1):1–9.

60. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, et al. Mimic-iv, a freely accessible electronic health record dataset. Scientific Data 2023;10:1.
61. Cao F, Wang Y, Yu K, Liang J: Causal discovery from unknown interventional datasets over overlapping variable sets. IEEE Transactions on Knowledge and Data Engineering(2024)
62. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
63. Ng I, Zhu S, Chen Z, Fang Z. A graph autoencoder approach to causal structure learning. 2019;arXiv preprint arXiv:1911.07420.
64. Zheng X, Aragam B, Ravikumar PK, Xing EP. Dags with no tears: continuous optimization for structure learning. Adv Neural Inf Process Syst. 31 2018.
65. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the gdpr. Harv JL & Tech. 2017;31:841
66. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, et al. Api design for machine learning software: experiences from the scikit-learn project. 2013;arXiv preprint arXiv:1309.0238.
67. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, McDermott M. Publicly available clinical bert embeddings. 2019;arXiv preprint arXiv:1904.03323.
68. Zhu J, Xia Y, Wu L, Xie S, Zhou W, Qin T, Li H, Liu T-Y: dual-view molecular pre-training. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p. 3615–27(2023)
69. Arzeno NM, Lawson KA, Duzinski SV, Vikalo H. Designing optimal mortality risk prediction scores that preserve clinical knowledge. J Biomed Informat. 2015;56:145–56.
70. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform. 2017;108:185–95.
71. Shazeer N. Glu variants improve transformer. 2020;arXiv preprint arXiv:2002.05202.
72. Angrist J, Imbens G. Identification and Estimation of Local Average Treatment Effects. Mass., USA: National Bureau of Economic Research Cambridge; 1995.
73. Becker SO, Caliendo M. Sensitivity analysis for average treatment effects. Stata J. 2007;7(1):71–83
74. Abdia Y, Kulasekera K, Datta S, Boakye M, Kong M. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: a comparative study. Biometrical J. 2017;59(5):967–85.
75. Belthangady C, Stedden W, Norgeot B. Minimizing bias in massive multi-arm observational studies with bcaus: balancing covariates automatically using supervision. BMC Med Res Method. 2021;21:1–10.
76. Belthangady C, Giampanis S, Jankovic I, Stedden W, Alves P, Chong S, Knott C, Norgeot B. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. Nat Commun. 2022;13(1):6921
77. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085.
78. Shukla SN, Marlin BM. Interpolation-prediction networks for irregularly sampled time series. 2019;arXiv preprint arXiv:1909.07782.
79. Nallabasannagari AR, Reddiboina M, Seltzer R, Zeffiro T, Sharma A, Bhandari M. All data inclusive, deep learning models to predict critical events in the medical information mart for intensive care iii database(mimic iii). 2020;arXiv preprint arXiv:2009.01366.
80. Jin M, Bahadori MT, Colak A, Bhatia P, Celikkaya B, Bhakta R, Senthivel S, Khalilia M, Navarro D, Zhang B, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. 2018;arXiv preprint arXiv:1811.12276.
81. Yang W, Zou H, Wang M, Zhang Q, Li S, Liang H. Mortality prediction among ICU inpatients based on mimic-iii database results from the conditional medical generative adversarial network. Heliyon. 2023;9(2).
82. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Informat. 2018;83:112–34.
83. Data MC, Pirracchio R. Mortality prediction in the ICU based on mimic-ii results from the super ICU learner algorithm(sicula) project. Secondary Analysis of Electronic Health Records. 2016;295–313.
84. Akiba T, Sano S, Yanase T, Ohta T, Koyama M: Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 2623–31(2019)
85. Pietiläinen L, Hästbacka J, Bäcklund M, Parviainen I, Pettilä V, Reinikainen M. Premorbid functional status as a predictor of 1-year mortality and functional status in intensive care patients aged 80 years or older. Intensive Care Med. 2018;44:1221–29.
86. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
87. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, Fuchs BD, Meadows L, Lynch M, Donnelly PJ, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med 2019;47:1485–92.
88. Fagerström J, Bång M, Wilhelms D, Chew MS.Lisep lstm: a machine learning algorithm for early detection of septic shock. Sci Rep. 2019;9(1):15132.
89. Kamel Rahimi A, Ghadimi M, Vegt AH, Canfell OJ, Pole JD, Sullivan C, Shrapnel S. Machine learning clinical prediction models for acute kidney injury: the impact of baseline creatinine on prediction efficacy. BMC Med Inf Decis Making. 2023;23(1):207.
90. Vagliano I, Chesnaye NC, Leopold JH, Jager KJ, Abu-Hanna A, Schut MC. Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal. Clin Kidney J. 2022;15(12):2266–80
91. Shamout FE, Zhu T, Sharma P, Watkinson PJ, Clifton DA. Deep interpretable early warning system for the detection of clinical deterioration. IEEE J Biomed Health Inform. 2019;24(2):437–46.
92. Kellum JA, Romagnani P, Ashuntantang G, Ronco C, Zarbock A, Anders H.-J.: Acute kidney injury. Nat Rev Dis Primers. 2021;7(1):52.
93. Zarbock A, Nadim MK, Pickkers P, Gomez H, Bell S, Joannidis M, Kashani K, Koyner JL, Pannu N, Meersch M, et al. Sepsis-associated acute kidney injury: consensus report of the 28th acute disease quality initiative workgroup. Nat Rev Nephrol. 2023;1–17.
94. Hoste EA, Kellum JA, Selby NM, Zarbock A, Palevsky PM, Bagshaw SM, Goldstein SL, Cerdá J, Chawla LS.Global epidemiology and outcomes of acute kidney injury. Nat Rev Nephrol. 2018;14(10):607–25.
95. Kellum JA, Lameire N, Group Kagw. Diagnosis, evaluation, and management of acute kidney injury: a kdigo summary(part 1). Critical Care. 2013;17:1–15.
96. Xu Z, Chou J, Zhang XS, Luo Y, Isakova T, Adekkanattu P, Ancker JS, Jiang G, Kiefer RC, Pacheco JA, et al. Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. J Biomed Informat. 2020;102:103361.
97. Choi MH, Kim D, Choi EJ, Jung YJ, Choi YJ, Cho JH, Jeong SH. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. Sci Rep. 2022;12(1):7180.
98. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. Int J Epidemiol. 2020;49(6):2058–64.
99. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228–42.
100. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–47.
101. Carson JL, Noveck H, Berlin JA, Gould SA. Mortality and morbidity in patients with very low postoperative hb levels who decline blood transfusion. Transfusion. 2002;42(7):812–18.
102. Teasdale G, Maas A, Lecky F, Manley G, Stocchetti N, Murray G. The glasgow coma scale at 40 years: standing the test of time. Lancet Neurol. 2014;13(8):844–54.
103. James MT, Bhatt M, Pannu N, Tonelli M. Long-term outcomes of acute kidney injury and strategies for improved care. Nat Rev Nephrol. 2020;16(4):193–205.
104. Ghazaly HF, Aly AAA, Sayed MH, Hassan MM. Apache iv, saps iii, and sofa scores for outcome prediction in a surgical/trauma critical care unit: an analytical cross-sectional study. Ain-Shams J Anesthesiol. 2023;15(1):101.
105. Czajka S, Ziebińska K, Marczenko K, Posmyk B, Szczepańska AJ, Krzych Łj. Validation of apache ii, apache iii and saps ii scores in in-hospital and one year mortality prediction in a mixed intensive care unit in poland: a cohort study. BMC Anesthesiol. 2020;20:1–8.
106. Pan X, Xie J, Zhang L, Wang X, Zhang S, Zhuang Y, Lin X, Shi S, Shi S, Lin W.Evaluate prognostic accuracy of sofa component score for mortality among adults with sepsis by machine learning method. BMC Infect Dis. 2023;23(1):76.

Publisher's Note