Causal Hierarchy-Guided Feature Reconstruction: Enhancing Interpretability and Performance of Postoperative Delirium Prediction in ICU Patients

Jiunan Wen[*1], **Linna Wang**[*1], **Peng Yuan**[1], **Li Lu**[1†]
[1] College of Computer Science, Sichuan University, Chengdu, Sichuan, China
[*] Equal contribution; [†] Corresponding author: luli@scu.edu.cn

Abstract

Postoperative delirium is a prevalent acute neurocognitive dysfunction in ICU patients, impairing key cognitive functions and deteriorating clinical prognosis, which is closely related to cognitive science research on neurocognitive regulation. Traditional machine learning prediction methods depend on statistical correlations and suffer from poor interpretability, failing to reflect the inherent causal hierarchy of delirium. To solve this issue, this paper integrates causal inference and proposes a Causal Hierarchy-Guided Feature Reconstruction Model (CDM). It adopts Greedy Equivalence Search and graph neural networks for causal feature reconstruction, followed by Bayesian-optimized XGBoost. Evaluated on the MIMIC-IV dataset, our model achieves superior performance with an AUROC of 0.9054, effectively enhancing model interpretability and offering a novel paradigm for cognitive disorder prediction.

Keywords: Postoperative delirium; Causal hierarchy; Feature reconstruction; Neurocognitive dysfunction

Introduction

Cognitive function—encompassing attention, consciousness, decision-making, and information processing—is the foundation of human adaptive behaviour, and its preservation under pathological conditions is a core research theme in cognitive science. Postoperative delirium, an acute neurocognitive syndrome with a prevalence of 7.4–87% in ICU patients, represents a severe disruption of these cognitive processes, leading to prolonged hospital stays, increased mortality, and irreversible long-term cognitive decline e.g., dementia (Thom et al., 2019). From a cognitive science perspective, investigating delirium's causal mechanisms and developing early prediction tools addresses a critical gap: understanding how acute stressors (e.g., inflammation, hypoxia, sleep disruption) perturb dynamic cognitive regulation, advancing knowledge of context-dependent neurocognitive dysfunction.

Societally, the aging population and rising ICU admissions have amplified the demand for effective delirium prevention and intervention (Fong & Inouye, 2022). Machine learning (ML) has emerged as the mainstream approach for delirium prediction, with numerous studies leveraging clinical data (vital signs, laboratory results, clinical scores) to build predictive models. While these ML frameworks outperform traditional statistical methods in scalability, they suffer from two critical limitations that misalign with cognitive science's mechanism-driven focus: first, they rely on statistical correlations rather than causal relationships, failing to distinguish direct drivers of cognitive dysfunction from spurious associations (e.g., confounding by disease severity (Duprey et al., 2022)); second, they operate as "black boxes," providing limited interpretability regarding which cognitive pathways or factors are most influential—hindering the translation of predictions into targeted interventions to protect cognitive function.

Causal inference resolves these limitations by explicitly modeling directional relationships between variables, enabling the identification of direct causal pathways (e.g., hypoxia-induced attention impairment (Thom et al., 2019)) and disentangling complex clinical confounders. This shift from "what correlates with delirium" to "what causes delirium" aligns predictive modeling with cognitive science's goal of elucidating the intrinsic mechanisms of cognitive disruption, making predictions not only accurate but also actionable and interpretable.

To bridge this gap, we propose the Causal Feature Reconstruction Model (CDM), a novel framework integrating causal discovery and graph neural networks to enhance delirium prediction's performance and interpretability. CDM follows three core steps: (1) GES-based causal structure learning to construct a DAG modeling relationships between clinical variables and delirium; (2) DoWhy-driven causal effect quantification to mitigate bias; (3) GNN-based feature reconstruction to fuse raw clinical data with causal information, generating causal features. These features are input into a Bayesian-optimized XGBoost model for prediction.

On the MIMIC-IV dataset (Johnson et al., 2023), CDM achieves an AUROC of 0.9054 (outperforming ML baselines by 4.39–16.30%), with SHAP analysis identifying core causal features (minimum GCS score, invasive ventilation) linked to cognitive dysfunction. This study contributes: (1) a causal-GNN framework addressing ML's "black-box" limitation in cognitive disorder prediction; (2) actionable causal drivers for clinical interventions targeting cognitive protection; (3) a paradigm bridging cognitive science and critical care via mechanism-driven predictive modeling.

Related Work

Traditional delirium prediction relies on statistical models and conventional risk scores, which lack expressiveness for complex clinical data (Van den Boogaard et al., 2012). Modern machine learning methods achieve stronger performance using EHR data but often act as black boxes without interpretability (Gong et al., 2023; Song et al., 2024). Causal

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

inference addresses spurious correlations and confounding by modeling directional relationships between variables (Pearl & Mackenzie, 2018; Spirtes et al., 2000), while graph neural networks enable structured representation of clinical features (Fey & Lenssen, 2019; Velickovic et al., 2017). These methods have been successfully applied to risk screening for sepsis, respiratory failure, and postoperative complications, yet they rarely integrate cognitive science principles into model design (Gu et al., 2023; Li et al., 2025).

Despite these advances, most delirium prediction models still focus on correlation-driven pattern recognition rather than mechanism-aware causal modeling. They rarely incorporate hierarchical causal structures to distinguish core driving factors from indirect confounding indicators, leading to unstable predictions and limited clinical explainability (Gu et al., 2023; Tang et al., 2024). Furthermore, few methods use causal graphs to guide feature learning and reconstruction, leaving a gap between model performance and clinical interpretability. Even state-of-the-art predictive models often overlook the cognitive mechanisms underlying delirium, such as disturbances in arousal, attention, and consciousness (Fong & Inouye, 2022; Thom et al., 2019), further weakening their clinical relevance and interpretability.

However, few studies unify causal hierarchy learning and feature reconstruction for delirium forecasting. This work integrates causal discovery and graph representation to improve both performance and interpretability for postoperative delirium prediction.

Materials

Ethical Review

The MIMIC-IV database was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Centre and Massachusetts Institute of Technology (Johnson et al., 2023). As this study was retrospective and used de-identified data without direct patient intervention, the requirement for informed consent was waived. This study followed the TRIPOD statement to ensure methodological rigor and transparency (Collins et al., 2015). All data processing, exclusion criteria, and sample size procedures were fully reported without selective omission. **Experimental Transparency Statement**:We report all experimental procedures, inclusion/exclusion criteria, data preprocessing thresholds, and model implementation details. No conditions or measures were selectively omitted. Sample size was determined by all eligible cases in MIMIC-IV that met inclusion criteria. In accordance with open science transparency guidelines, we report all measures, conditions, data exclusions, and sample size determination procedures without selective omission.

Study Population

The study population was sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database (Johnson et al., 2023), which contains comprehensive electronic health record data of ICU admissions at Beth Israel Deaconess Medical Centre spanning from 2008 to 2019. The eligibility criteria were defined as follows:

(1) Postoperative ICU admission (identified using ICD procedure codes with prefixes);
(2) First-time ICU admission to avoid repeated measurement bias;
(3) ICU length of stay ≥24 hours to ensure adequate clinical data collection;
(4) Availability of at least one valid Confusion Assessment Method for the ICU (CAM-ICU) assessment during hospitalization (Ely et al., 2001);
(5) Age between 1 and 120 years (only extreme age outliers were excluded, with no upper age limit).

Patients were excluded if they had missing key identifiers (subject_id, hadm_id, stay_id), invalid admission time records, or lacked gender/age information. A detailed flowchart of cohort selection is presented in Figure 1.

Table 1: Variables Included in the Prediction Models

Demographic data: Age, gender, race
Vital signs: Heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, temperature, respiratory rate, oxygen saturation
Laboratory results: Hematocrit, hemoglobin, platelets, white blood cell, anion gap, bicarbonate, blood urea nitrogen, chloride, creatinine, sodium, potassium, international normalized ratio, prothrombin time, partial thromboplastin time, urine output, glucose
Clinical scores: Sequential Organ Failure Assessment (SOFA), minimum Glasgow Coma Scale (GCS)
Comorbidities: Myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, peptic ulcer disease, renal disease, Acquired Immunodeficiency Syndrome, liver disease, diabetes
Treatment measures: Invasive ventilation, renal replacement therapy, acetaminophen, anticholinergic, anticoagulants, antihistamines, antipsychotics, benzodiazepines, diuretics, anesthetics, Nonsteroidal Antiinflammatory Drugs, opioids, vasopressors
Other key information: First 24h delirium assessment result, first care unit type

Delirium Assessment

Delirium status was evaluated using the CAM-ICU, a well-validated tool specifically designed for delirium assessment in critically ill patients. The observation window was set as the first 24 hours after ICU admission, during which clinical data were collected for model development (Gong et al., 2023). The primary outcome was the occurrence of delirium after the observation window: Delirium was diagnosed if at least one valid positive CAM-ICU value (1) was recorded following the first 24 hours of ICU admission; patients with only negative CAM-ICU values (0) or no valid assessments after the observation window were categorized as non-delirium (Song

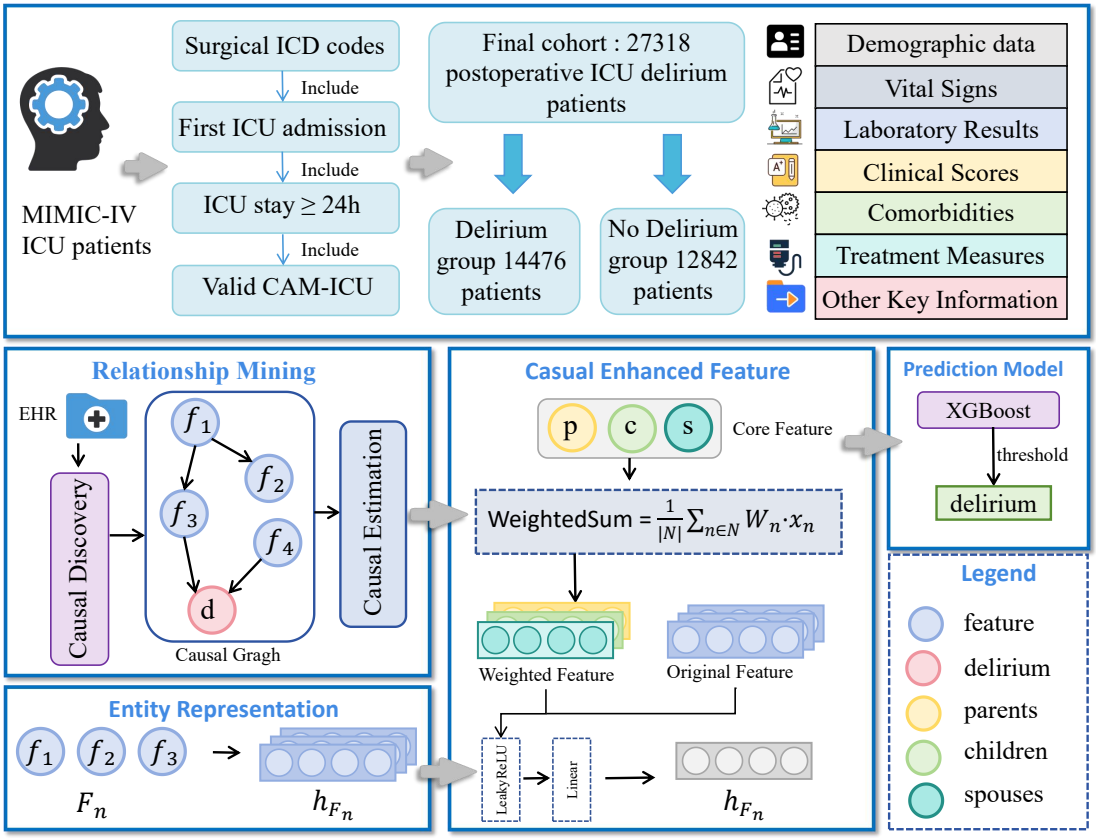

Figure 1: Overall workflow of the postoperative delirium prediction model. The chart integrates clinical data processing, causal structure learning, and XGBoost classification, with detailed module explanations in Section 4.1–4.3. *Abbreviations*: - f: Feature (clinical variable); - d: Delirium (outcome variable); - p: Parent node (feature that directly causes other features/core features); - c: Child node (feature directly caused by other features/core features); - s: Spouse node (feature sharing a common causal relationship with other features).

et al., 2024). Only CAM-ICU values of 0 or 1 were considered valid, and all other values were excluded during data cleaning.

Data Extraction and Processing

A total of 54 categorical and numerical variables were selected from MIMIC-IV (Johnson et al., 2023), covering demographics, vital signs, laboratory results, scores, comorbidities, and treatments (Table 1). All data were extracted within the first 24 hours of ICU admission. Mean values were computed for repeated measurements. Variables with >15% missing values were excluded (Van den Boogaard et al., 2012), and remaining missing values were imputed using median (for numerical features) and 0 (for binary features). Continuous features were standardized, and all samples were shuffled before training. The Boruta algorithm (Kursa & Rudnicki, 2010) was used for feature selection. It identifies valid predictors by comparing real features with synthetic "shadow features" via Random Forest (Breiman, 2001), retaining only robust and clinically meaningful variables for subsequent modeling.

Methods

Causal Feature Reconstruction with CDM

The Causal Feature Reconstruction Model (CDM) integrates Greedy Equivalence Search (GES)-based causal graph construction and graph neural network-driven feature reconstruction to extract discriminative causal features from clinical data, eliminating spurious correlations between variables and the delirium outcome.

Causal Graph Construction with GES The global causal directed acyclic graph (DAG) was constructed using the preprocessed dataset after feature selection, with continuous features standardized to $[-1, 1]$ to stabilize GES (Spirtes et al., 2000).

Causal effects between clinical features and the delirium label were quantified using the DoWhy library (Sharma & Kiciman, 2020). For continuous-to-binary relationships, backdoor propensity score matching was employed to reduce confounding bias (Rosenbaum & Rubin, 1983). Ordinary Least Squares

(OLS) regression was used when causal identification was not feasible. All causal effect values were normalized to $[0, 1]$ for consistent weighting.

Causal Feature Reconstruction with Graph Neural Network Based on the causal DAG, a graph neural network was designed to reconstruct causal features, consisting of three components: Feature Embedding Layer, Causal Graph Integration Module, and Sequence Encoder. The forward propagation is shown in Equation 1:

$$\mathbf{z} = \text{Dropout}\left(\text{SeqEncoder}\left(\text{CGIM}\left(\text{FE}\left(\mathbf{x}\right)\right)\right)\right) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^F$ is the input feature vector (F: number of selected features), FE$(\cdot)$, CGIM$(\cdot)$, SeqEncoder$(\cdot)$ denote the three modules, and $\mathbf{z} \in \mathbb{R}^{64}$ is the final causal feature.

Feature Embedding Layer This layer maps the raw feature vector to a high-dimensional embedding space using a two-layer fully connected network with learnable weights (Equation 2):

$$\begin{aligned} \mathbf{h} &= \text{LayerNorm}\left(\text{Linear}_2\left(\text{Dropout}\left(\mathbf{h}_1\right)\right)\right) \\ \mathbf{h}_1 &= \text{LayerNorm}\left(\text{ReLU}\left(\text{Linear}_1\left(\mathbf{x} \odot \mathbf{w}\right)\right)\right) \end{aligned} \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^F$ is initialized to 1, $\odot$ denotes element-wise multiplication, Linear_1 and Linear_2 have dimensions 128 and 64. ReLU, LayerNorm, and dropout (rate=0.2) stabilize training. Hyperparameters were selected according to common architectural practices and fixed during training.

Causal Graph Integration Module This module refines embeddings using causal relationships in the DAG, focusing on nodes directly linked to delirium. For each core feature f_i, parent, child, and spouse nodes were aggregated via weighted summation:

$$\mathbf{h}_{neigh}(f_i) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{w}_k \odot \mathbf{h}(f_k) \tag{3}$$

where K is the number of valid neighbors, $\mathbf{w}_k$ is a learnable weight initialized to 0.1, and $\mathbf{h}(f_k)$ is the neighbor embedding. The original and neighbor embeddings are fused to form the reconstructed feature:

$$\mathbf{h}_{\text{recon}}(f_i) = (1 - \alpha) \cdot \mathbf{h}(f_i) + \alpha \cdot \mathbf{h}'_{\text{neigh}}(f_i) \tag{4}$$

$$\mathbf{h}'_{\text{neigh}}(f_i) = \text{LeakyReLU}\left(\text{Linear}\left(\text{Concat}\left(\mathbf{h}(f_i), \mathbf{h}_{neigh}(f_i)\right)\right)\right) \tag{5}$$

where $\alpha = \text{sigmoid}(\omega)$ with ω initialized to 0.1.

Sequence Encoder Reconstructed embeddings are aggregated and compressed into the final causal feature:

$$\mathbf{z} = \text{Dropout}\left(\text{ReLU}\left(\text{Linear}\left(\sum_{i=1}^{M} \mathbf{h}_{recon}(f_i)\right)\right)\right) \tag{6}$$

where M is the number of core feature nodes. The Linear layer compresses the output to 64 dimensions.

CDM was implemented in Python, PyTorch, and PyTorch Geometric (Fey & Lenssen, 2019; Paszke et al., 2019) following the preprocessing scheme in Section 2.5. It outputs only the 64-dimensional causal vector $\mathbf{z}$ and does not perform prediction.

Prediction Model and Evaluation

The 64-dimensional causal features $\mathbf{z}$ were input to an XGBoost model (Chen & Guestrin, 2016) suitable for nonlinear clinical data. The dataset was split stratifying into 70% training and 30% test sets. Feature selection, standardization, and hyperparameter optimization were performed **only on the training set** to avoid data leakage. Thresholds were optimized via Bayesian methods to maximize F1.

Performance was assessed using AUROC, AUPRC, accuracy, precision, recall, F1, Brier score, and MCC. Calibration curves, decision curve analysis (Vickers & Elkin, 2006), and SHAP (Lundberg & Lee, 2017) were used to evaluate model consistency, clinical utility, and feature importance.

Result

Causal Analysis of Clinical Variables

Through GES-based causal structure learning, direct causal edges pointing to the delirium outcome were identified (Figure 2), and the core nodes (features) in the causal diagram are consistent with the baseline characteristics of the study population (Table 2).

Key findings of the causal analysis included: 1. Core causal features (direct drivers of delirium): The following features were identified as having direct causal edges to the delirium label: - Minimum Glasgow Coma Scale (GCS) score, which reflects disturbances in arousal and awareness central to delirium; - Invasive ventilation, which disrupts sleep and brainstem arousal regulation; - First care unit type; - First 24h delirium assessment result; - Renal replacement therapy; - Acetaminophen. These features directly drive the occurrence of postoperative delirium, consistent with the significant differences between the delirium and non-delirium groups in Table 2 (all $P < 0.001$).

2. Parent/child/spouse nodes relative to core features: - Parent nodes (upstream drivers of core features): Anesthetics drive the core feature ‘First 24h delirium assessment result‘, while Nonsteroidal Antiinflammatory Drugs (NSAIDs) indirectly affect features via vasopressors. - Child nodes (downstream effects of core features): ‘First care unit type‘ drives platelets, blood urea nitrogen (BUN), anticoagulants, and creatinine; ‘Invasive ventilation‘ drives partial thromboplastin time (PTT) and renal replacement therapy; ‘Minimum GCS score‘ drives temperature, linking altered arousal to physiological dysregulation; ‘Acetaminophen‘ drives hematocrit. - Spouse nodes (features sharing causal pathways with core features): Anticoagulants and ‘First care unit type‘ co-drive platelets; vasopressors and ‘First care unit type‘ co-drive BUN; benzodiazepines and ‘First care unit type‘ co-drive platelets.

Table 2: Baseline Characteristics of Parents Nodes in Causal Diagram (Delirium Group vs. Non-Delirium Group)

Parents Nodes in Causal Diagram	**Total Sample (n=N) mean±std**	**Non-Delirium Group (n=N_0) mean±std**	**Delirium Group (n=N_1) mean±std**	***P* Value**
GCS	1.29 ± 0.99	1.02 ± 0.27	1.53 ± 1.29	< 0.001
Invasive Ventilation	0.42 ± 0.49	0.27 ± 0.44	0.51 ± 0.50	< 0.001
First Care Unit Type	1.24 ± 0.94	0.75 ± 0.95	1.68 ± 0.67	< 0.001
First 24h Delirium Assessment	0.91 ± 0.28	1.00 ± 0.07	0.84 ± 0.37	< 0.001
Renal Replacement Therapy	0.85 ± 0.36	0.88 ± 0.33	0.84 ± 0.37	< 0.001
Acetaminophen	0.65 ± 0.48	0.80 ± 0.40	0.52 ± 0.50	< 0.001

Note: All variables were non-normally distributed by Shapiro-Wilk normality test, and Mann-Whitney U nonparametric test was used for inter-group comparison 1,2. n represents the sample size; $P < 0.05$ was considered statistically significant, and $P < 0.001$ was considered extremely statistically significant. GCS and First Care Unit Type are ordinal categorical variables, which were described by mean±standard deviation 3.
Abbreviations: GCS = Glasgow Comal Scale. Total sample $N = 27,318$; non-delirium group $N_0 = 12,842$; delirium group $N_1 = 14,476$.

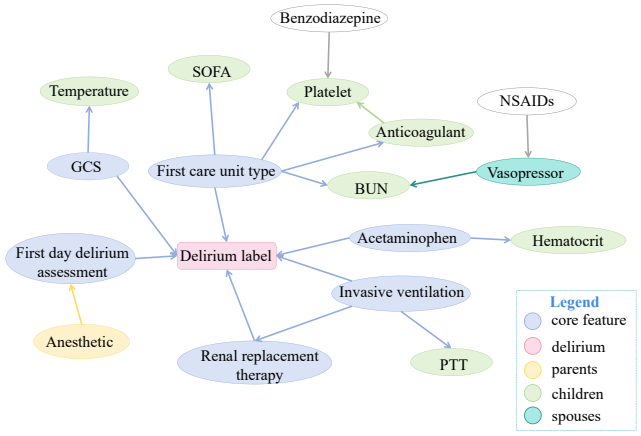

Figure 2: Direct causal edges pointing to delirium outcome (identified via GES algorithm). *Legend*: - Core feature (light blue nodes): Features with direct causal edges to the delirium label (direct drivers of delirium); - Delirium (pink node): Outcome variable (postoperative delirium); - Parents (yellow nodes): Features that directly cause core features; - Children (light green nodes): Features directly caused by core features; - Spouses (teal nodes): Features sharing a common causal pathway with core features. *Abbreviations*: - SOFA: Sequential Organ Failure Assessment; - GCS: Glasgow Coma Scale; - BUN: Blood urea nitrogen; - PTT: Partial thromboplastin time; - NSAIDs: Nonsteroidal Antiinflammatory Drugs.

3. Non-direct causal features: Features such as temperature, benzodiazepines, NSAIDs, and vasopressors have no direct edges to the delirium label, suggesting their association with delirium is mediated by core features linked to cognitive and arousal function, such as ‘Minimum GCS score‘ and ‘First care unit type‘.

These causal relationships (Figure 2) clarify the direct/indirect associations between clinical features and delirium: core features (e.g., ‘first care unit type‘) directly drive delirium, while other nodes act as upstream/downstream factors—avoiding spurious correlations and providing a basis for CDM’s feature extraction.

Performance Comparison of Baseline and CDM-Enhanced XGBoost Models

All baseline models were trained on the same raw clinical features used in our pipeline, while the CDM-enhanced model used the causal features extracted by our framework.Table 3 presents the comprehensive performance metrics of five models (LR, SVM, RF, XGBoost, and CDM) on the MIMIC-IV dataset (Johnson et al., 2023), and Figure 3 visualizes the metric trends.

The CDM model outperformed all baselines: its ROC-AUC (0.9054) was 4.39% higher than XGBoost, and Recall (0.9233) increased by 15.29%, with the lowest Brier score (0.1225) indicating better calibration. XGBoost was the best baseline (ROC-AUC=0.8615), while traditional models (LR/SVM) had limited performance (ROC-AUC < 0.8).

Table 3: Performance metrics of models on MIMIC-IV dataset

Model	ROC_AUC	PRC_AUC	Accuracy	Precision	Recall	F1	Brier
LR	0.7424	0.8148	0.6743	0.6977	0.6743	0.674	0.2079
SVM	0.7704	0.8386	0.6906	0.7163	0.6906	0.6905	0.1919
RF	0.8452	0.8987	0.7674	0.7821	0.7674	0.767	0.1718
XGBoost	0.8615	0.9087	0.7736	0.7814	0.7736	0.765	0.1516
CDM	0.9054	0.9057	0.8219	0.7807	0.9233	0.8460	0.1225

Abbreviations: LR = logistic regression; SVM = support vector classifier; RF = random forest classifier; XGBoost = extreme gradient boosting; CDM = Causal Feature Reconstruction Model.

SHAP Interpretability Analysis

Figure 4 compares the feature contribution rankings of the baseline XGBoost model and the CDM-enhanced model, focusing on the rank changes of core causal features (marked by pink arrows in the diagram).

Key insights from the SHAP analysis (strictly based on Figure 4 and updated feature names) included: 1. Core causal feature prioritization: In the baseline XGBoost model, ‘First 24h delirium assessment result‘ (a core feature with direct causal edges to delirium) was ranked 5th; after CDM’s causal feature extraction, this feature rose to 2nd place—aligning with its status as a direct driver of delirium. 2. Stable ranking of high-impact core features: ‘First care unit type‘ (a core feature directly linked to delirium) maintained the top rank in both models, confirming its high predictive value as a direct

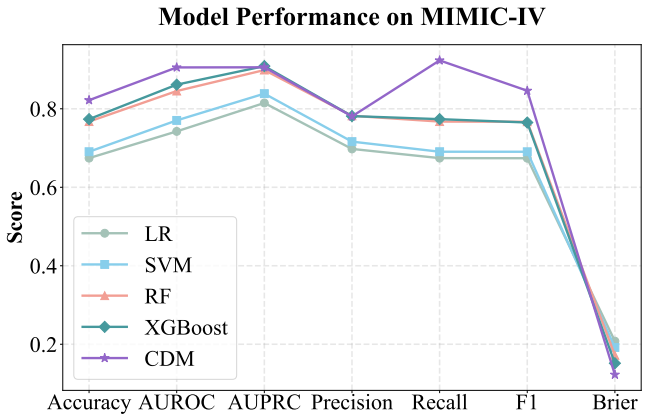

Figure 3: Performance comparison of models on MIMIC-IV dataset. *Abbreviations*: LR = logistic regression; SVM = support vector classifier; RF = random forest classifier; XGBoost = extreme gradient boosting; CDM = Causal Feature Reconstruction Model.

driver of delirium (consistent with Table 2's significant group difference). 3. Balanced core feature relevance: CDM adjusted the ranking order to ensure all core features (e.g., 'Acetaminophen', 'Renal replacement therapy') are prioritized, avoiding the baseline model's underweighting of some direct causal factors.

This adjustment demonstrates that CDM enhances interpretability: it aligns feature contribution rankings with the actual causal structure (prioritizing all core features directly linked to delirium), rather than relying solely on statistical correlation.

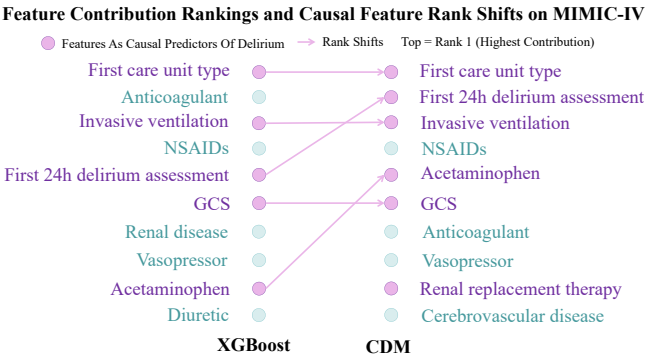

Figure 4: Feature contribution rankings and causal feature rank shifts between XGBoost and Causal Feature Reconstruction Model (CDM) on MIMIC-IV (Medical Information Mart for Intensive Care IV) database. *Abbreviations*: - GCS: Glasgow Coma Scale; - NSAIDs: Nonsteroidal Antiinflammatory Drugs.

Discussion

Our key causal findings align with cognitive science theories of delirium as a disorder of arousal, attention, and cognitive control (Fong & Inouye, 2022; Thom et al., 2019). We identify minimum GCS score, invasive ventilation, first care unit type, first 24h delirium assessment, renal replacement therapy, and acetaminophen as direct drivers, consistent with clinical and mechanistic accounts (Bodien et al., 2021; Gu et al., 2023; Mart et al., 2021). These results provide empirical causal evidence for how physiological disruptions impair core cognitive systems, offering theoretical insights for cognitive science research on acute neurocognitive dysfunction.

CDM improved prediction by better prioritizing causally relevant features rather than purely correlational signals. By emphasizing causal structure, the model provides more interpretable, mechanism-based insights that better connect to the theoretical focus of cognitive science (Gao et al., 2024), supporting more targeted clinical decision-making.

Limitations include the single-center MIMIC-IV design (Pollard et al., 2018) and static features, which may be addressed in future work with sequential models (Choi et al., 2016). Overall, this work shows how causal representation learning can improve delirium prediction, offering a generalizable approach for interpretable modeling in cognitive and critical care research.

Conclusion

We constructed a high-performance, interpretable delirium prediction model by integrating causal inference and ML, addressing traditional models' spurious correlation/black-box limitations (Gao et al., 2024; Pearl & Mackenzie, 2018). Using MIMIC-IV data, we: 1) screened 54 variables via Boruta (Kursa & Rudnicki, 2010); 2) identified 6 causal edges directly pointing to the label and other causal edges that indirectly affect delirium prediction via the GES algorithm; 3) fused data/causal graphs via CDM to generate features for Bayesian-optimized XGBoost. The model outperformed baselines (AUROC=0.9054, Recall=0.9233, Brier=0.1225), with SHAP prioritizing clinical causal factors (SOFA, GCS, etc. (Song et al., 2024)) linked to arousal and attentional dysfunction. Contributions: 1) a causal-GNN framework for interpretable ML relevant to cognitive dysfunction prediction; 2) clarified delirium causal drivers for interventions; 3) a tool for early risk stratification.

References

Bodien, Y. G., van den Boogaard, M., & Slooter, A. J. C. e. a. (2021). Glasgow coma scale subscores are better predictors of delirium than total scores in critically ill patients. *Critical Care*, *25*(1), 364. https://doi.org/10.1186/s13054-021-03815-8

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Choi, E., Bahadori, M. T., & Sun, J. e. a. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, *29*, 3504–3512.

Collins, G. S., Reitsma, J. B., & Altman, D. G. e. a. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. *BMJ*, *350*, g7594. https://doi.org/10.1136/bmj.g7594

Duprey, M. S., Devlin, J. W., & Griffith, J. L. e. a. (2022). Association between perioperative medication use and postoperative delirium and cognition in older adults undergoing elective non-cardiac surgery. *Anesthesia & Analgesia*, *134*(6), 1154–1163. https://doi.org/10.1213/ANE.0000000000005664

Ely, E. W., Gautam, S., & Margolin, R. e. a. (2001). Evaluation of delirium in critically ill patients: Validation of the confusion assessment method for the icu (cam-icu). *Critical Care Medicine*, *29*(7), 1370–1379. https://doi.org/10.1097/00003246-200107000-00019

Fey, M., & Lenssen, J. E. L. (2019). Fast graph representation learning with pytorch geometric. In *Workshop on representation learning on graphs and manifolds*. NeurIPS.

Fong, T. G., & Inouye, S. K. (2022). The inter-relationship between delirium and dementia: The importance of delirium prevention. *Nature Reviews Neurology*, *18*(10), 579–596. https://doi.org/10.1038/s41582-022-00656-2

Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., & Li, Y. (2024). Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems*, *42*(4), 1–32. https://doi.org/10.1145/3626772

Gong, K. D., Lu, R., & Bergamaschi, T. S. e. a. (2023). Predicting intensive care delirium with machine learning: Model development and external validation. *Anesthesiology*, *138*(3), 299–311. https://doi.org/10.1097/ALN.0000000000004478

Gu, Q., Yang, S., Fei, D. T., Lu, Y., & Yu, H. (2023). A nomogram for predicting sepsis-associated delirium: A retrospective study in mimic iii. *BMC Medical Informatics and Decision Making*, *23*(1), 184. https://doi.org/10.1186/s12911-023-02282-5

Johnson, A. E. W., Bulgarelli, L., & Shen, L. e. a. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1), 1. https://doi.org/10.1038/s41597-023-02090-4

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*, 1–13. https://doi.org/10.18637/jss.v036.i11

Li, R., Zhang, Y., & Wang, H. e. a. (2025). Development of a machine learning-based predictive model for postoperative delirium in older adult intensive care unit patients: Retrospective study. *Journal of Medical Internet Research*, *27*(1), e67258. https://doi.org/10.2196/67258

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

Mart, M. F., Williams Roberson, S., & Salas, B. e. a. (2021). Prevention and management of delirium in the intensive care unit. *Seminars in Respiratory and Critical Care Medicine*, *42*(1), 112–126. https://doi.org/10.1055/s-0040-1710572

Paszke, A., Gross, S., & Massa, F. e. a. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8024–8035.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pollard, T. J., Johnson, A. E. W., & Raffa, J. D. e. a. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, *5*, 180178. https://doi.org/10.1038/sdata.2018.178

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Sharma, A., & Kiciman, E. (2020). Dowhy: An end-to-end library for causal inference. *arXiv Preprint arXiv:2011.04216*.

Song, Y., Zhang, D., & Wang, Q. e. a. (2024). Prediction models for postoperative delirium in elderly patients with machine-learning algorithms and shapley additive explanations. *Translational Psychiatry*, *14*(1), 57. https://doi.org/10.1038/s41398-024-02762-w

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.

Tang, D., Ma, C., & Xu, Y. (2024). Interpretable machine learning model for early prediction of delirium in elderly patients following intensive care unit admission: A derivation and validation study. *Frontiers in Medicine*, *11*, 1399848. https://doi.org/10.3389/fmed.2024.1399848

Thom, R. P., Levy-Carrick, N. C., Bui, M., & Silbersweig, D. A. (2019). Delirium. *American Journal of Psychiatry*, *176*(10), 785–793. https://doi.org/10.1176/appi.ajp.2018.18070893

Van den Boogaard, M., Pickkers, P., & Slooter, A. J. C. e. a. (2012). Development and validation of pre-deliric (prediction of delirium in icu patients) delirium prediction model for intensive care patients: Observational multicentre study. *BMJ*, *344*, e420. https://doi.org/10.1136/bmj.e420

Velickovic, P., Cucurull, G., & Casanova, A. e. a. (2017). Graph attention networks. *arXiv Preprint arXiv:1710.10903*.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361