An Explainable AI Approach using Graph Learning to Predict ICU Length of Stay

Tianjian Guo[*], Indranil R. Bardhan[*], Ying Ding[**], Shichang Zhang[***]
[*] McCombs School of Business, The University of Texas at Austin, [**] School of Information, The University of Texas at Austin, [***] Harvard Business School, Harvard University.

Abstract

Intensive care units (ICU) are critical for treating severe health conditions but represent significant hospital expenditures. Accurate prediction of ICU length of stay (LoS) can enhance hospital resource management, reduce readmissions, and improve patient care. In recent years, widespread adoption of electronic health records (EHR) and advancements in artificial intelligence (AI) have facilitated accurate prediction of ICU LoS. However, there is a notable gap in the literature on explainable AI methods which identify and include interactions between model input features in developing accurate predictions of health outcomes. This gap is especially noteworthy as the medical literature suggests that complex interactions between clinical features are likely to significantly impact patient health outcomes. We propose a novel graph learning-based approach that offers state-of-the-art prediction and greater interpretability for ICU LoS prediction. Specifically, our explainable AI (XAI) graph model can generate interaction-based explanations, supported by evidence-based medicine, which provide rich patient-level insights compared to existing XAI methods. We test the statistical significance of our XAI approach using a distance-based separation index and utilize perturbation analyses to examine the sensitivity of our model explanations to changes in input features. Finally, we validate the explanations of our graph learning model using the Co-12 framework and a small-scale user study of ICU clinicians. Our approach offers interpretable predictions of ICU LoS grounded in design science research which can facilitate greater integration of AI-enabled decision support systems in clinical workflows, thereby enabling clinicians to derive greater value.

Key words: Length of stay, intensive care units, prediction, machine learning, deep learning, graph learning, explainable AI, perturbation analyses, user study.

1. Introduction

ICUs provide life-saving capabilities to patients hospitalized for severe diseases, comorbidities, and other life-threatening conditions. However, ICUs also consume significant resources in their utilization of clinical staff and equipment. Prior studies have shown that as much as a third of hospital budgets are spent on ICUs, and a third of inpatient costs can be attributed to ICU stays (Multz et al. 1998, Shweta et al. 2013). Hence, it is in the best interest of hospitals, taxpayers, and insurers to reduce ICU costs while ensuring delivery of high-quality patient care. Since hospitals use LoS to measure the effectiveness of treatments, schedule resources and make staffing decisions, accurate LoS prediction for ICU patients should lead to better ways of managing scarce ICU resources (Romano et al. 2014). Furthermore, since LoS serves as an early indicator of future hospital readmissions, effective LoS prediction can allow healthcare practitioners to manage ICU resources better by reducing readmission rates (Singh and Terwiesch 2012, Oh et al. 2018).

With widespread adoption and use of EHR systems in recent years, researchers can use AI to analyze clinical and administrative claims data in developing more accurate predictions of patient health outcomes. However, extant research has often prioritized predictive performance over actionable and interpretable insights, a gap that undermines the practical utility of predictive models for clinical decision-making (Chen et al. 2023). Computer scientists and healthcare professionals have advocated for integrating intrinsic explanations within predictive models in healthcare settings, especially to promote greater utilization of AI-based, clinical decision support tools (Rudin 2019, Petch et al. 2022). AI systems with intrinsic explanation capabilities are designed to inherently explain the prediction process. Unlike post-hoc explanation methods, which generate explanations by approximating the inner working of black-box AI models, intrinsic explanations accurately reflect the prediction process with no approximation (Molnar 2022). In healthcare, this enhanced transparency is critical to increase physician trust in the prediction and underlying logic of AI-based clinical decision support tools.

Furthermore, recent research suggests that simple feature-based explanations are inadequate to explain the complex relationships that AI-based models utilize to generate accurate predictions (Fernández-

Loría et al. 2022, Carmichael and Scheirer 2023; Jiang et al. 2023). Evidence-based medicine also emphasizes the importance of recognizing the complex interactions among clinical factors in understanding patient health outcomes (Singbartl and Kellum 2012, Jankovic et al. 2018). Hence, it is important to provide explanations that accurately identify key interactions among features to better represent the underlying prediction process and align with clinical domain knowledge (Ahmad et al. 2018).

Yet, there remains a significant gap in the development and application of *intrinsically interpretable* models which effectively identify key feature interactions, particularly in healthcare settings. To address such a gap, we develop a novel graph learning-based prediction model to intrinsically identify complex *interactions* between patient attributes and their impact on LoS prediction. Our model constructs patient-level relational graphs that serve as instruments to predict ICU LoS with high accuracy and interpret the contribution of salient features and feature interactions toward LoS prediction. We compare our graph learning model against alternative state-of-the-art, interaction-based XAI methods. Such methods either provide interaction-based explanations of complex prediction methods in a *post-hoc* manner or construct intrinsically explainable models that offer interaction-based explanations. Our results indicate that prior XAI methods fail to generate meaningful explanations based on feature interactions and are computationally less efficient. In comparison, our model not only identifies the importance of feature interactions but does so more efficiently than existing XAI methods, demonstrating its superiority in providing more transparent explanations, while offering comparable predictive accuracy.

We further validate our interaction-based explanations through multiple tests to evaluate our model properties based on the Co-12 framework, which defines a set of conceptual properties for evaluation of XAI methods (Nauta et al. 2023). Utilizing perturbation analysis, we demonstrate that modifications to input features result in appropriate changes in model explanation based on the importance of the perturbed features. We deploy a distance-based separation index to test the significance of feature interactions identified by our model and confirm their relevance for ICU LoS prediction. Finally, we validate the coherence of explanations generated by our model by ensuring that salient feature interactions identified by our model are medically relevant and corroborated by prior medical research. We conduct a small-scale

user study with ICU physicians on the insights generated by our XAI approach, and their feedback further supports the practical usability and explanations generated by our model (Abbasi et al. 2024).

Hence, we develop a novel graph-learning, intrinsically explainable prediction model to predict ICU LoS. Our intrinsic approach provides richer patient-level insights compared to existing XAI methods, by generating medically-relevant explanations of interactions between patient attributes. Hence, from a methodological and application perspective, our approach represents a significant contribution in terms of its ability to identify and generate patient-specific explanations of salient features and interactions that contribute to LoS prediction and are validated by our user study of ICU physicians. Although designed for ICU LoS prediction, our framework is generalizable to other types of health risk prediction, thereby enabling clinicians to make better-informed decisions using AI-enabled clinical decision support tools (Petch et al. 2022). Our approach provides pathways that illustrate our contributions to design science research based on specific traits, such as the richness of domain adaptation and use of a sociotechnical lens, to develop and validate a novel explainable AI prediction approach (Abbasi et al. 2024).

2. Background

In this section, we review the extant research on applications of machine learning (ML) and deep learning (DL) techniques for prediction of health outcomes, with a focus on LoS. Subsequently, we discuss advancements in graph learning algorithms—a subset of deep learning that processes data with graph structures—and their use in healthcare. We also identify and discuss significant gaps and limitations in the current XAI literature and describe how our proposed graph learning model addresses these challenges.

2.1. LoS Prediction

Length of stay in the ICU is one of the most important measures of patient health, a proxy for resource allocation decisions, and an indicator of future readmissions. Hence, accurate prediction of LoS is critical for effective ICU management and care delivery, especially among high-risk patients with severe complications (Singh and Terwiesch 2012, Romano et al. 2014, Oh et al. 2018). Severity score-based

measures, such as Acute Physiology and Chronic Health Evaluation IV (APACHE IV), have been deployed in ICUs to predict patient outcomes such as mortality and LoS (Zimmerman et al. 2006). These scores were derived from regression models using patient characteristics and vital signs as independent variables. However, the efficacy of risk-score-based systems, such as APACHE IV, has come under greater scrutiny due to their limited selection of independent variables and over-reliance on the underlying statistical assumptions of logistic regression models (Zangmo and Khwannimit 2023).

In recent years, medical researchers have deployed ensemble-based ML techniques, such as random forests and gradient boosting, to predict LoS in ICU settings. These methodologies have been applied to diverse patient populations, ranging from general ICU patients to those with specific conditions such as lung cancer and COVID-19 (Alsinglawi et al. 2022, Saadatmand et al. 2023). Information systems researchers have also utilized these techniques to study preventable readmissions among patients with chronic conditions (Ben-Assuli and Padman 2020). While ensemble methods outperform statistical approaches, they are unable to exploit latent relationships, such as temporal dependencies, in clinical data.

In contrast, recent computational advancements have enabled development of increasingly sophisticated DL models that utilize latent relationships within healthcare datasets (Morid et al. 2023). For instance, researchers have studied the application of Temporal Pointwise Convolutional Neural Networks to predict ICU LoS (Rocheteau et al. 2021, Al-Dailami et al. 2022b). These models represent distinct variations of temporal convolutional neural networks (T-CNNs) that were developed to analyze time-varying data. Alternatively, attention mechanisms have also been utilized to improve LoS prediction. These mechanisms allow neural networks to focus on relevant input data segments and have established healthcare applications, such as the Reverse Time Attention (RETAIN) model (Choi et al. 2016). Recent innovations in this stream of literature have applied variants of attention mechanisms specifically designed to handle complex healthcare data. These innovations include additional designs to process multi-modal and time series data. Examples include the Attention-Based Memory Fusion Network and Temporal-Spatial Correlation Attention Network (Al-Dailami et al. 2022a, Nie et al. 2023).

2.2. Graph Learning

Graph learning, or deep graph learning, has emerged as a powerful approach to analyze and model data with complex interactions between entities. Traditional deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), excel at handling structured data, such as images and numeric or text sequences, but struggle with complex, unstructured interactions. Graphs naturally depict these interactions through nodes and edges, making them suitable to represent a variety of real-world phenomena, including social networks, molecular structures, and transportation systems (Wu et al. 2020). In turn, graph learning methods, also known as graph neural networks (GNNs), offer a general framework for learning representations of graph data. These models aggregate and process information from the neighbors of nodes in graphs and capture complex interaction patterns within the data.

Prior studies have explored applications of graph learning methods in healthcare, particularly for ICU risk prediction and chronic disease management. For example, Ma et al. (2023) constructed a patient graph to predict mortality risk in ICU patients, where the graph edges are weighted by patient similarity. The patient graph was used to identify missing patient features, and a dynamic attention mechanism was used to learn additional structural features for each patient. Carvalho et al. (2023) predicted 30-day ICU readmission risk by enriching electronic health record (EHR) data with a knowledge graph (KG) and used KG embeddings to integrate ontology information. Similarly, Sun et al. (2024) addressed EHR data heterogeneity by using multi-view graphs to encode diagnosis and medication co-occurrence and analyzed their impact on ICU outcomes. Tong et al. (2021) proposed an ICU LoS prediction model that combines Long Short-Term Memory (LSTMs) networks to extract temporal features and GNNs for exploiting similarity in patient diagnoses.

2.3. Explainable AI

Despite a rapid increase in deployment of AI applications in healthcare, the “black box” nature of ensemble and DL models poses a barrier to clinical use and integration, as they lack transparency in decision-making. Without being able to interpret the recommendations proposed by AI models, the adoption of AI in clinical

practice has sparked criticism and raised questions about numerous legal, ethical, equity, and medical concerns (Rai 2020, Bauer et al. 2023, Bauer and Gill 2024). Due to these challenges, there has been greater emphasis in recent years on the role of XAI methods in enhancing the transparency and acceptance of AI models in healthcare. The field of XAI seeks to develop methods that explain AI-based models to enhance model interpretability, fairness, and transparency. Such explainability allows for better human understanding of AI decision-making and fosters greater trust in model outputs (Chaddad et al. 2023).

Previous studies have classified XAI methods based on various attributes (Chaddad et al. 2023). Table 1 compares various XAI methods based on two defining characteristics: type of explanation—*intrinsic or post-hoc*, and attribute type—*feature-based or interaction-based*. In Appendix A, we provide a comprehensive comparison of the XAI methods discussed. Extant research on XAI methods has mainly focused on developing and using post-hoc, feature-based explanations to explain deep learning models. Prominent examples include Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and Integrated Gradient (IG) (Bach et al. 2015, Sundararajan et al. 2017, Selvaraju et al. 2020). These methods determine the importance of input features by examining the gradient associated with each input. Applications in healthcare include identification of critical regions for medical imaging, such as chest CT scans for COVID-19 detection (Zhang et al. 2021), developing explainable early warning scores for conditions such as sepsis (Lauritsen et al. 2020), and evaluating the significance of various input features in predicting ICU LoS (Rocheteau et al. 2021).

Outside the scope of deep learning models, post-hoc, feature-based XAI techniques are more model agnostic, offering interpretability irrespective of their underlying model architecture. Perturbation-based methods, which modify specific features to evaluate their impact on model output, have been particularly useful to identify vulnerabilities in prediction models (Finlayson et al. 2019). Model distillation techniques, such as Local Interpretable Model-agnostic Explanations (LIME), create localized linear models to interpret more complex models and have been used to explain predictions of heart failure incidents during hospitalization (Khedkar et al. 2020). Building on LIME, the robust local explanations (ROLEX) method was developed to provide locally faithful explanations and used to explain predictions of fragility-related

fractures in patients (Kim et al. 2023). Shapley Additive Explanations (SHAP) utilize Shapley values to assign importance to individual features. SHAP has been particularly useful in explaining and predicting hospital LoS for lung cancer patients (Alsinglawi et al. 2022).

Beyond feature-based explanations, there is an emerging body of research aimed at developing techniques that provide explanations based on *interactions between features*. A prominent stream of work involves extending SHAP values to account for feature interactions, such as the Shapley Interaction Index (SII), Shapley Taylor Index (STI), and Faith Shapley Index (FSI) (Grabisch and Roubens 1999, Sundararajan et al. 2020, Tsai et al. 2023). These methods extend the SHAP framework to include subsets of features, thereby calculating the importance of feature interactions. Another significant stream of research in building post-hoc, interaction-based XAI methods for DL models involves extending gradient-based explanation methods to calculate the gradient of feature interactions. A notable example is Integrated Hessian (IH), an extension of IG, which utilizes the gradient of IG values to identify the importance of interactions between pairs of features. The effectiveness of IH has been studied in the context of identifying drug-drug interactions for treatment of leukemia (Janizek et al. 2021).

Last, a related stream of literature focuses on developing intrinsic, interaction-based XAI methods by expanding Generalized Additive Models (GAMs) to include pairwise interaction terms. Initial work in this area started with GA^2M, later evolving to explainable boosting machines (EBM), which constructs GAMs using ensemble decision trees, and more recently NODE-GAM, which employs deep neural networks to build GAMs (Lou et al. 2013, Nori et al. 2019, Chang et al. 2021). These methods excel at generating intrinsically explainable models but are limited by the additive nature of the GAM framework. Our graph learning model (lower right quadrant of Table 1) provides a novel approach to generate intrinsic explanations that emphasize both features and feature interactions and their contributions to explainability.

2.4. Research Gaps

Computer scientists and medical professionals have increasingly advocated for using *inherently interpretable* models in healthcare, highlighting significant concerns with the limitations of post-hoc

explanation methods. Yet, most applications of XAI in healthcare focus on utilizing feature-based, post-hoc XAI methods, such as SHAP and CAM (Chaddad et al. 2023). Such post-hoc methods primarily rely on approximations and often fail to accurately represent the nuances of black box models. Information lost in this approximation process can potentially erode users' trust in the prediction model or lead to misinterpretation of predictions (Rudin 2019, Petch et al. 2022). Petch et al. (2022, p. 211) articulated the challenges associated with post-hoc explanations and argued for inherently interpretable prediction models:

> *"…. The most notable limitation of explainability techniques is that most of them are approximations of black-box models and therefore do not precisely account for the inner workings of those models […]. A key advantage of many ML methodologies is that they can model nonlinear relationships, but the strategy of explaining black-box models through approximations may be particularly limiting […]. Even with nonlinear explainability techniques such as decision trees, the relative simplicity of explanations compared with the black-box models means that any nonlinear relationships surfaced through the explanation are likely to be oversimplifications and thus should be interpreted with caution […]. If there is no meaningful difference in accuracy between an interpretable model and a black box, an interpretable method should be used …."*

This perspective highlights the critical importance of deploying *intrinsically interpretable* models in healthcare. Such models ensure that physicians can rely on the accuracy of the explanations provided, avoiding error-prone decisions based on prior beliefs and superficial information (Jussupow et al. 2021). Similarly, we argue that feature-based explanations alone are insufficient to understand the complex relationships that AI models exploit to generate predictions. Instead, it is important to offer explanations that highlight key interactions between features, especially in real-world healthcare settings.

Extant research has shown that feature-based XAI methods fail to accurately explain complex AI-based models, often providing misleading or incorrect explanations (Fernández-Loría et al. 2022, Carmichael and Scheirer 2023; Jiang et al. 2023). Providing misleading or incorrect interpretations can significantly impair the effectiveness of AI-based tools. Modern evidence-based medical research suggests that understanding patient health outcomes requires recognizing the complex interactions of multiple factors. In ICUs, for example, acute kidney injuries—which affect up to 25% of ICU patients—occur due to complex interactions of clinical conditions instead of individual factors (Singbartl and Kellum 2012). Similarly, drug-drug interactions in ICUs may have unexpected synergistic or antagonistic effects, further

complicating patient outcomes (Jankovic et al. 2018). Therefore, it is critical to offer explanations that account for interactions among input features. We posit that such explanations may not only represent the prediction process more accurately but are also aligned with domain knowledge, making them accessible to practitioners who often lack a background in machine learning (Ahmad et al. 2018).

Despite the importance of intrinsic, interaction-based explanations, a significant gap persists in the development and application of AI models with these capabilities. We address this gap by developing a graph learning-based XAI approach that provides *intrinsic, interaction-based explanations* to predict patient health outcomes, using ICU LoS as our research domain. By focusing on the *design artifact* of feature interactions and their impact on LoS prediction and explainability, we ground our XAI approach in the principles of design science research based on rich domain knowledge (provided by ICU physicians in our user study) and highlight the importance of nuanced interactions among patient attributes that contribute toward greater interpretability of our graph-based predictive model (Abbasi et al. 2024, Liao et al. 2020).

2.5. Research Contributions

Our graph learning model constructs patient-specific, relational graphs that not only serve as predictive instruments of ICU LoS but also explain the relationships between patient attributes that contribute to the predicted outcome. In comparison, prior studies on ICU LoS prediction primarily seek to improve prediction capabilities without explaining these models. Furthermore, prior studies that attempt to offer explanatory insights into their prediction models utilize post-hoc, feature-based XAI methods that exhibit major limitations as discussed in the previous section (Rocheteau et al. 2021, Al-Dailami et al. 2022a).

Our model is different from existing graph learning models as it aims to address the task of constructing patient-level graphs to provide intrinsic, interaction-based explanations. Unlike previous graph learning applications in healthcare which analyze cohort-level graphs to predict patient outcomes, our approach utilizes patient-level graphs, thereby enhancing both the accuracy and explainability of predictions (Tong et al. 2021, Carvalho et al. 2023, Ma et al. 2023). Furthermore, our approach autonomously constructs graph structures from data without a predefined graph format, identifying key

interactions or edges, that are unobserved in the initial data. This approach is superior to prior graph learning models which rely on pre-defined graph structures or are limited to exploring only a subset of potential unobserved interactions (Kreuzer et al. 2021, Zhu et al. 2021).

We also introduce an innovative, attention-based method to assess the importance of both nodes and edges for graph-level prediction tasks. Existing methods, such as Graph Attention Networks (GAT), primarily focus on studying edge importance at the node or edge level (Veličković et al. 2018). In contrast, our model evaluates the hierarchical significance of both nodes and edges, providing a deeper and more nuanced understanding of the final prediction at the graph level. Our comprehensive approach to assess node and edge importance, combined with an ability to generate and analyze unique graphs for individual patients, facilitates the development of intrinsically interpretable and accurate predictive models.

Compared to existing XAI methods, our proposed graph learning model offers the distinct advantage of providing *intrinsic, interaction-based explanations*. By representing each patient as a relational graph, where nodes correspond to clinical features and edges denote their interactions, our model can accurately identify nuanced interactions between features that contribute to the predicted outcome. In contrast, existing post-hoc interaction-based explanation methods rely on approximation of the internal mechanism of complex black-box models. While these techniques can offer some insights, they are limited in their ability to faithfully represent the intricate feature interactions within the prediction model.

Although recent advances in intrinsically interpretable models provide interaction-level explanations, the family of GA^2M models, which include EBM and NODE-GAM, prioritize an optimal GAM based on features alone, before identifying and ranking potential feature interactions within the residuals. Their design treats interactions as less important than individual features and limits the magnitude of their contribution to the final prediction. We empirically demonstrate that our graph learning model provides explanations that are computationally more efficient compared to post-hoc, interaction-based XAI methods and offers more insightful interaction-based explanations. From an application and methodological perspective, the enhancement in computational efficiency and explanatory power establishes our model as a superior approach in understanding the key features and interactions that influence ICU LoS.

Our graph learning-based XAI approach is distinct from extant studies that apply XAI methods to graph learning models. While XAI techniques have been developed and applied to graph learning models, these applications focus on post-hoc interpretations that illuminate the internal mechanism of black-box graph learning models (Ying et al. 2019, Zhang et al. 2022). These methods identify critical nodes and edges within predefined graph structures based on pre-trained graph learning models. In contrast, we build an intrinsically explainable model which constructs graphs from data that do not have a graph-structure format. Our approach then uses the constructed graph to predict and explain predictions of patient LoS, integrating graph construction directly into the prediction and explanation process.

Hence, we develop a novel graph learning-based model to generate explainable predictions that highlight important interactions between input features that are not easily observable in the underlying data. Our model is unique in its ability to provide *intrinsic* and *interaction-based explanations*. We address the challenge of generating intrinsic, interaction-based explanations by transforming it into a graph-based task. Specifically, the objective is to construct graphs from data that initially lack graph structure and identify the significance of features and their interactions utilizing the structure of the constructed graph. To accomplish this, we extend graph learning techniques that were not originally designed for this purpose. Table 2 summarizes the salient differences between our approach and related XAI methods.

3. Explainable AI Framework

In this section, we first describe the specific task of predicting patient ICU LoS, research data utilized in this study, followed by the design and evaluation of our model using a design science framework.

3.1. Prediction Task

Previous studies have primarily focused on predicting the numeric value of ICU LoS by calculating the exact duration between patient admission and discharge from the ICU. However, current state-of-the-art models have revealed limitations with this approach, as prediction errors are measured in days, rendering them less useful in real-life clinical settings (Al-Dailami et al. 2022b, Sun et al. 2024). This drawback has

prompted a shift toward more accurate and interpretable LoS prediction methodologies. For example, Harutyunyan et al. (2019) transformed the task of predicting the numeric value of LoS into a multi-label classification problem and predicted the specific day of discharge for a patient after admission, with each label corresponding to a different discharge date. Alsinglawi et al. (2022) and Saadatmand et al. (2023) utilized binary predictions based on whether a patient is likely to be discharged from the ICU within a specific time window, such as within seven days of admission. Hence, the design of our prediction model needs to take into consideration the salient artifacts of the contextual setting such as the prediction window, interplay between features, and other attributes.

In this study, we embrace prior research and adopt a binary prediction strategy based on the likelihood of patient discharge within seven days following ICU admission. Identifying patients with predicted ICU stay exceeding one week enables early intervention of specialized care management teams, enhancing the quality of care, especially for at-risk patients (Dahl et al. 2012). We also conduct robustness tests using alternate prediction tasks, specifically the binary prediction of ICU discharge within 3 days as well as numeric prediction of ICU LoS, as discussed in Appendix C.

3.2. Research Data and Domain

We utilize data collected from MIMIC III, a publicly accessible database provided by the MIT Lab for Computational Physiology, to assess the prediction and explanation capability of our proposed model. MIMIC III encompasses de-identified health records from 61,532 ICU admissions, compiled between 2001 and 2012 at a large academic medical center in Boston (Johnson et al. 2016). Since the MIMIC III data spans twelve years, some patients have multiple records from recurrent ICU admissions from one or more hospital visits. We only consider the first ICU stay of each patient as a qualifying stay and eliminate successive ICU visits (if applicable) to limit our research scope to LoS prediction based on clinical data from their first ICU visit. This preempts the potential for serial correlation across multiple visits, since LoS on a later visit may depend on treatments performed during a prior ICU visit.

We further refine our dataset by excluding ICU admissions with LoS less than two days—the data

collection period—to ensure that comprehensive and relevant data is used for model training. Selecting the data collection period is critical since an excessive duration can hamper model operability, while a shorter period might not offer sufficient data for training, culminating in suboptimal prediction. Our choice of a 48-hour window is consistent with prior research and addresses the relative scarcity of clinical data for selected input variables within the first 24 hours of ICU admission (Rotar et al. 2022). Our final data set contains 22,243 ICU stays which provide the relevant data for our predictive models. Since our data has a one-to-one correspondence between patients and ICU stays, we refer to them interchangeably in the following discussion. For reference, we do not remove patients who passed away during their ICU stay.

3.3 Graph Learning-Based Model

Next, we propose and design a novel graph learning model to generate intrinsically explainable predictions of ICU LoS, capable of highlighting key interactions between features. This model predicts ICU LoS by constructing patient-level graphs that illustrate the importance of individual features and interactions between features at the patient level. Figure 1 provides a visual representation of our design approach.

First, during *node attribute generation* (step 1), each type of input feature is transformed into a fixed-length vector within a unified feature space utilizing different projection layers, each corresponding to a specific type of input feature. These projection layers are customized based on defining characteristics of the associated type of input features—LSTM units for processing temporal data and feed-forward neural layers for the remaining types of features. Let x be the input for a given patient. In step 1, x is transformed into h, a combination of transformed feature representations $h_{temporal}$ and h_{static}, as defined in equation (1).

$$h = [h_{temporal}, h_{static}] \tag{1}$$

Specifically, projections for temporal data are generated through LSTMs as shown in equation (2),

$$h_{temporal} = LSTM(x_{temporal}) \tag{2}$$

and a feed-forward layer for other feature types as shown in equation (3).

$$h_{static} = LSTM(W_{static}x_{static} + b_{static}) \tag{3}$$

where W_{static} and b_{static} are the weights and biases of the feed-forward layers.

Step 2 involves *graph construction*, where a fully connected directed graph, *G=(V, E),* is constructed based on the projected input features *h*. In this graph, each node *i* in *V* corresponds to a specific type of input feature with the associated projection h_i encapsulated as the node attribute. Each edge e_{ji} = *(j, i)* in *E* represents the potential flow of information, or interaction, from node *j* to node *i*.

In step 3, we calculate *edge importance* where we leverage a GAT to refine the node attributes in the constructed graph *G*. GAT is a specialized type of message-passing GNN that utilizes attention mechanisms to selectively focus on and aggregate relevant node-level information (Veličković et al. 2018). Specifically, the attributes of each node are updated with the weighted sum of attributes of its neighboring nodes. These weights are dynamically determined by an edge-level attention mechanism, which assesses the relevance of each neighbor in relation to the attribute vector of the focal node. For each node *i* in *G*, its updated attribute h'_i is computed as shown in equation (4),

$$h'_i = \sigma(\sum_{j \in V} \alpha_{ji} W h_j) \tag{4}$$

where *W* is a learnable weight matrix, and α_{ij} are attention coefficients computed as shown in equation (5),

$$\alpha_{ji} = \frac{\exp\left(LeakyReLU(a^T[Wh_i || Wh_j])\right)}{\sum_{k \in V} \exp\left(LeakyReLU((a^T[Wh_i || Wh_k])\right)} \tag{5}$$

with *a* being a learnable weight vector of the attention mechanism. The calculated edge-level attention coefficients, α_{ji}*'s,* describe the relevance of edge e_{ji} for updating the attributes of node *i*. Given the fully connected nature of the patient-level graph constructed in step 2, *α's* are calculated for all possible combinations of $i, j \in V$, enabling the model to comprehensively assess all potential interactions.

Next, in step 4, the *node importance calculation*, an attention-based read-out mechanism is utilized to generate a vector representation, h_g, of the entire graph. This process involves creating a weighted sum of the updated node attributes h' shown in equation (6), where β_i's are attention weights computed similarly to the *α's* by evaluating the relevance of each node's transformed attributes h'_i for the prediction task.

$$h_G = \sum_{i \in V} \beta_i h'_i \tag{6}$$

The *β*'s assigned to each node not only determine h_g but also serve as indicators for the importance of the updated node attributes.

Finally, in the *graph-based prediction* step (step 5a), the graph-level representation h_g obtained from the attention-based read-out in step 4 is processed using a multi-layer perceptron, consisting of multiple feed-forward layers, to generate the final prediction for ICU LoS, as shown in equation (7).

$$Y = MLP(h_G) \tag{7}$$

Simultaneously, in the *graph-based explanation* step (step 5b), we construct a patient-level directed graph utilizing the attention values from steps 3 and 4. This graph encapsulates the importance of individual types of features and their interactions in contributing to the ICU LoS prediction for each patient. Specifically, we define the importance of the node *i*, $FeatImp_i$, as the corresponding node-level attention value, β_i, which represents the importance of the feature type represented by node *i*, shown in equation (8)

$$FeatImp_i = \beta_i \tag{8}$$

We then define the importance of the edge e_{ji}, $InteractionImp_{j,i}$, as the product of the attention value attributed to the edge, α_{ji}, with the attention value assigned to the destination node, β_i. Its value is equal to the proportion of importance assigned to node *i* (in step 4) attributed to the flow of information from the feature represented by node *j* to the feature represented by node *i*.[1] We interpret this value as the importance of the interaction between the features represented by node *i* and node *j*, as shown in equation (9).

$$InteractionImp_{j,i} = \alpha_{ji}\beta_i \tag{9}$$

The product term in equation (9) is particularly important in representing the true importance of a given edge or feature interaction to the overall prediction process. While the attention values generated by the GAT represent the relative importance of an edge for information flow to a particular node, such values are assigned at the node level and do not measure the global relevance of that edge for graph-level prediction tasks. By multiplying the edge-specific attention with node-specific attention, we derive a measure of the *overall importance* of the edge (or interaction). Based on effective integration of the GAT with an attention-

[1] It should be noted that the constructed graph is directional in nature, $InteractionImp_{i,j}$ and $InteractionImp_{j,i}$ represent different values. $InteractionImp_{i,j}$ represent the importance of the information flow from node *i* to node *j*, while $InteractionImp_{j,i}$ represent the importance of the information flow from node *j* to node *i*.

based read-out mechanism, our model can identify the contributions of the interactions between features to the attention allocated to each feature. The sum of both feature- and edge-level importance scores is equal to one, as shown in equation (10).

$$\sum_{j,i} InteractionImp_{j,i} = \sum_{j,i} \alpha_{ji}\beta_i = \sum_i FeatImp_i = \sum_i \beta_i = 1 \quad (10)$$

3.4. Model Adaptation

Based on the model design in section 3.3, we discuss our model adaptation in the context of LoS prediction using the MIMIC III data. While we apply the model in the context of ICU LoS, our model design can be adapted to other health risk prediction tasks by simply modifying the process of transforming input features into a unified vector space.

For each patient, we utilize 47 types of features across four categories: patient administrative data, diagnosis, medication data, and vital signs. Table 3 provides descriptive statistics of selected input features. Specifically, we utilize 7 types of patient administrative data: patient age, gender, ethnicity, marital status, type of hospital admission, insurance status, and ICU admission type, which together form a 1x71 vector. Patient diagnosis is represented as a 1x18 vector, indicating the presence or absence of disease diagnoses based on 18 top-level ICD-9 categories. We include 8 types of vital sign measures: heart rate, glucose level, body temperature level, oxygen level, respiration rate, systolic blood pressure, diastolic blood pressure, and mean blood pressure. Each type of vital sign is represented as a 1x24 vector, based on the average readings of the corresponding vital signs, organized in 24 two-hour intervals during the initial two days of ICU admission. Intervals with no readings are filled with a value of -1.[2] Medications administered to patients are represented using seven principal components derived from daily dosage data across the two-day data collection period, for a total of 14 distinct values.

The original daily dosage data spans 178 medication categories classified under level 3 of the Anatomical Therapeutic Chemical (ATC) system. These data are factorized into seven principal

[2] Alternative filling approaches, such as backward/forward filling, mean filling, or 0 filling, were examined and did not significantly influence the results.

components using principal component analysis (PCA). We categorize and label the seven principal components based on their corresponding loading values, as (a) Metabolic and Anti-infective Agents, (b) Cardiovascular and Blood Agents, (c) Gastrointestinal and Hormonal Agents, (d) Nutritional and Anti-inflammatory Agents, (e) Antineoplastic and Immunomodulating Agents, (f) Dermatological and Respiratory Agents, and (g) Analgesics and Central Nervous System Agents. Appendix G provides a detailed description of our classification approach. Utilizing principal components, instead of raw medication dosage data, as inputs, helps to reduce noise and ensures consistent node count in the patient-level graphs, thereby standardizing input features across patients.

In the next step, 47 projection layers map each of the feature types into the same 64-dimensional vector space. The eight vital signs are processed through unidirectional LSTM layers with a hidden size of 64 to exploit the temporal relationships inherent in the data, while the remaining 39 feature types are mapped via feed-forward layers, reflecting their simplicity. Subsequently, a fully connected directed graph is constructed for each patient comprising 47 nodes and 2209 edges. Each node corresponds to one of the 47 types of features. The 64-dimensional vectors, generated by the projection layers, are included in the graph as node attributes, aligning with their respective nodes. A GAT with a hidden dimension of 64, 4 attention heads, and *ReLU* as the activation function, is utilized to generate the edge-level attention values and update the node attributes. A global attention pooling layer then computes a weighted sum of the updated node attributes across the 47 nodes based on the node-level attention values. This computation yields a 64-dimensional vector representing the entire graph. This vector is subsequently processed through a feed-forward layer with a sigmoid activation function to yield the predicted likelihood of 7-day ICU stay. The node- and edge-level attention values are then utilized to construct the patient-specific relational graph and explain the corresponding ICU LoS prediction.

4. Results

Due to the intrinsically explanatory nature of our graph learning model, it is imperative to assess its predictive capabilities and the quality of explanations generated. This two-pronged evaluation ensures a

comprehensive understanding of the ability of our graph learning model to not only predict accurately but also explain the prediction. We first compare the predictive performance of our graph learning-based model (henceforth referred to as our graph model) with EBM, a custom-built DL model (henceforth referred to as the DL model), and other widely used ML algorithms.[3] This comparative evaluation is designed to validate the reliability and efficacy of our graph-centric approach for accurate prediction (Li et al. 2020, Liu et al. 2020). Next, we shift our focus to the explanation dimension of our model and compare the explanations generated by various XAI techniques. Due to our interest in generating interaction-based explanations, we compare the following XAI approaches: our graph model, EBM, IH, and FSI, where the latter two methods explain the DL model (henceforth referred to as DL-IH and DL-FS, respectively). These alternative XAI methods are included as benchmarking targets based on the classification in Table 1. Implementations details of these methods are provided in Appendix B.

4.1 Prediction Comparison

Table 4 provides a detailed comparison of the predictive performance of various models, including our graph learning model, the DL model, EBM, and conventional ML models such as XGBoost, random forests, and logistic regressions, in predicting the likelihood of ICU discharge within 7 days (of admission) across 10 cross-validation runs with an 80/20 split of training/test data. Notably, our graph model and DL models demonstrate identical performance in terms of the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC), with scores of 0.824 and 0.899, respectively. The EBM model also reports comparable AUROC and AUPRC values of 0.824 and 0.898, respectively, and exhibits the highest F1 score of 0.839 with a prediction accuracy of 0.771, matching that of the DL model. Since approximately 30% of patients in our data remain in the ICU for more than seven days, metrics such as AUROC and AUPRC are important as they are less prone to the effects of class imbalance, compared to accuracy or F1 scores. The predictive performance of our graph learning model is superior to conventional ML models and comparable to the DL model and EBM.

[3] Microsoft actively supports the EBM package, which is more up-to-date and accessible compared to NODE-GAM.

While the focus of our study is to predict 7-day ICU discharge, Appendix C broadens the scope of our analysis by evaluating both prediction and explanation capabilities of our graph learning model and the EBM, to predict 3-day ICU discharge and numeric LoS. Although EBM demonstrates slightly higher precision in predicting 3-day discharge, our model offers superior capability in predicting numeric LoS. However, with a mean average error exceeding 5 days, we observe that numeric LoS predictions lack practical relevance in a clinical setting.

4.2 Explanation Comparison

After assessing the predictive accuracy of our graph learning model, our focus shifts to its ability to explain the relationships identified by the model. Our goal is to assess whether our graph learning model and various interaction-based XAI techniques can generate meaningful explanations based on the interactions among patient attributes. In the ensuing analysis, we visually contrast the explanations generated by our graph learning model with other interaction-based XAI methods. This includes comparison of individual patients, patient cohorts, and evaluation of the significance of interaction-based explanations. We also explore their computational efficiency based on time required to provide explanations for an individual patient.

4.2.1 Computation Time

Before we present explanations provided by various XAI methods, we first evaluate the computation time required by each method to generate explanations for an individual patient. This is particularly pertinent for interaction-based explanations that necessitate computing the importance of at least N^2 pair-wise interactions—compared to feature-based explanations that only require computing the significance of N features. Table 5 compares the computational efficiency of four types of XAI methods deployed to explain the predicted outcome for a single patient. These include two intrinsic (EBM and our graph learning model) and two post-hoc (DL-FS and DL-IH) methods.

We observe a notable discrepancy between intrinsic and post-hoc methods with respect to the computation time to generate explanations. Specifically, both EBM and our graph learning model can produce explanations in under 0.1 seconds, whereas post-hoc methods require significantly more time. For

instance, the DL-FS method averages 20 seconds, while DL-IH requires up to 4 minutes to generate explanations for a single patient. Due to the considerably poorer performance of DL-IH, we exclude it from subsequent analysis.[4] We observe that generating explanations is significantly quicker for intrinsic XAI methods, since a simple forward pass through the neural network (for our graph learning model) or the ensemble of decision trees (for EBM) is sufficient to generate the relevant explanation. On the contrary, both post-hoc techniques must, by design, calculate the relevance of each feature and their interactions at the time of generating the explanation, a process that is more computationally intensive.

4.2.2 Patient-level Explanation

In this analysis, we compare the explanations generated by DL-FS, EBM, and our graph learning model, to predict the LoS of a 46-year-old male patient, admitted through the emergency department and treated in the surgical ICU. The patient had an ICU stay exceeding seven days which was accurately predicted by all models (i.e., binary prediction LoS > 7 days). Figure D1 in Appendix D displays the explanations from the EBM model for the top 15 terms—either a feature or interaction between two specific features—that impact LoS prediction for this patient. Based on the EBM results, there is a 34.2% chance of this patient being discharged within 7 days. It identifies *respiratory system-related diagnosis* as a critical factor, suggesting its presence decreases the likelihood of ICU discharge within seven days by 27.3% (=1-e^-0.32). We note that, among the top 15 terms (in Figure D1), all are features and do not include any feature interactions.

Figure D2 in Appendix D provides a graphical illustration of the explanation of the DL-FS model, showing salient features and interactions that explain the LoS prediction. It estimates a 4.6% likelihood of ICU discharge within 7 days and identifies the prevalence of respiratory system-related diagnosis in reducing the likelihood of 7-day discharge by 12.29%. Vital signs, such as mean blood pressure and glucose levels, are also noteworthy as predictive indicators. DL-FS does not assign significant importance to feature interactions, with the most significant interaction only having a -0.007% impact on LoS prediction.

On the other hand, Figure 2 displays the explanations associated with our graph learning model,

[4] In Appendix B, we explain why the DL-IH method is much slower compared to other methods.

which predicts an 8.5% likelihood of discharge within 7 days. The graph representation visually emphasizes the importance of features and feature interactions through the size of nodes and width of edges in the personalized graph. While our model identifies *respiratory system-related diagnosis* as a prominent feature that receives 71.81% of the total attention, a closer examination reveals that significant components of this attention—specifically, 17.954%, 17.954%, and 15.629%—can be attributed to the interactions between *respiratory system diagnosis* and *patient age*, *nutritional and anti-inflammatory agents*, and *analgesics and central nervous system agents,* respectively. This result suggests that the attention assigned to respiratory system diagnosis can be attributed to its *interactions* with patient age and medications.

Compared to EBM and DL-FS, our graph learning model generates a more comprehensive explanation of ICU LoS that identifies the nuanced impact of feature interactions. For instance, it identifies the interaction between patient age and respiratory system diagnosis as an important explanatory attribute. Such an interaction is medically sound, since prior research has observed gradual deterioration in lung function as patients age, emphasizing the need to consider patient age when diagnosing and treating respiratory system conditions (Sharma and Goodwin 2006).

4.2.3 Population-level Explanation

Next, we aggregate patient-level explanations generated by the three XAI methods—DL-FS, EBM, and our graph learning model—at the cohort level to demonstrate the importance of features and feature interactions across the patient population. By averaging the attention scores of nodes (representing features) and edges (representing interactions) from our graph learning model across patients, we identify key features and interactions that are relevant for LoS prediction across the patient population. Similarly, the EBM model calculates global term importance for each feature and pairwise interaction as the mean of absolute importance values across all patients. We apply a similar approach to calculate the mean absolute values of Faithful Shapley scores which measure the average influence of each feature or interaction on prediction of LoS. Tables 6 and 7 provide a comparative evaluation of the salient features and interactions determined by the three XAI methods. Although the specific importance scores are not directly comparable

across the three methods, the relative rankings of the features and interactions can be compared.

Based on the DL-FS results in the left panel of Table 6, we observe that patient *diagnoses*, *age*, and *blood pressure* during the final two hours of the ICU stay, are significant predictors of LoS. Similarly, the EBM model identifies different types of *diagnosis*, *ICU type*, and *medication*s, as salient factors in predicting LoS, with *respiratory system diagnosis* emerging as the most influential factor. Our graph learning model also reports patient *diagnoses*, *age*, *vital signs*, and *ICU type*, as important features. On average, 11.5% of the attention is assigned to respiratory system diagnosis, demonstrating its role as the most important feature. Overall, our results indicate a high level of consistency across the three XAI methods with respect to salient features for LoS prediction.

On the other hand, the results reported in Table 7 provide a different perspective on the explanations related to feature interactions. A closer examination of mean absolute Faith SHAP scores shows that the DL-FS method overlooks the importance of feature interactions. For example, admission to a medical intensive care unit (MICU) is ranked as the tenth most important feature by Faith SHAP, influencing the predicted likelihood of an extended ICU stay by an average of 0.69%. In comparison, the top-rated interaction between systolic and mean blood pressure values (during hours 46-48) has a trivial importance score of 0.02%, implying an almost negligible impact on the likelihood of LoS prediction. Our results suggest that DL-FS struggles to assign substantive importance to feature interactions. In contrast, EBM and our graph learning model attribute meaningful significance to feature interactions. The mean importance scores of feature interactions identified by these methods are substantial compared to individual features. Specifically, we observe greater magnitude of importance scores associated with the interactions between patient age, diagnosis, medications, and vitals, such as body temperature, heart rate, and blood pressure.

4.2.4. Distance-based Separation

Quantitative evaluation of explanations generated by various interaction-based XAI methods poses a significant challenge. Existing evaluation techniques deployed in prior XAI research primarily focus on feature-level analysis or utilize synthetic data with predetermined underlying relationships (Janizek et al. 2021, Kim et al. 2023). Consequently, these methods are not directly applicable to assess the quality of

interaction-based explanations in our research context. Drawing on the literature on concept-based explanations, we develop an approach to evaluate the efficacy of interaction-based explanations by measuring their ability to distinguish between different outcomes (Crabbé and van der Schaar 2022). We posit that including interactions in explanations, as opposed to utilizing only features, should improve the ability to distinguish patients with different LoS outcomes. We utilize t-distributed stochastic neighbor embedding (T-SNE), a well-known technique for dimension reduction and visualization, to process and visualize the high-dimensional explanations generated. Specifically, we generate 2-dimensional T-SNE plots for the patient cohort using our explanations from the DL-FS, EBM, and the graph learning model. For each XAI method, two plots are generated: one based on the importance scores of the top features, and another based on the scores of top features and interactions. Each patient is then classified based on their LoS outcomes, specifically whether they are discharged within 7 days.

Our proposition suggests that T-SNE plots generated from feature and interaction importance scores should offer better separation between patients with different outcomes compared to T-SNE plots that only include feature importance scores. We assess the separation within the T-SNE plots using the Distance-based Separation Index (DSI) (Guan and Loew 2022). DSI measures the degree of separability between two sets of data, with a value between 0 to 1, with a higher DSI value indicating a greater degree of separation. By examining the DSI values, we can determine the effectiveness of including interaction-based elements in explanations, with the expectation of observing greater separation based on explanations provided by including feature- and interaction-based importance scores.

Table 8 shows the average DSI improvement for the DL-FS, EBM, and graph learning models, highlighting the impact of interaction-based explanations in T-SNE plots. These improvements are calculated across ten cross-validation runs, utilizing different training/test splits generated randomly with each pair of T-SNE plots derived from the respective test dataset. We evaluate the statistical significance of these improvements using a paired t-test. The results suggest that inclusion of interaction terms in the graph learning model enhances patient separation in the corresponding T-SNE plots. This improvement is statistically significant across all T-SNE configurations for the graph learning model, as shown in Table 8.

However, inclusion of interaction terms in the EBM and DL-FS models does not significantly enhance their ability to distinguish patients based on LoS. This indicates that the interaction-based components of model explanation from EBM and DL-FS do not augment the insights provided by the feature-based components.

The design philosophy of the GA^2M family of algorithms prioritizes building an optimal GAM-based on features before identifying and ranking potential feature interactions within the residuals. Only top-ranked feature pairs determined through cross-validation are included in the final model. This sequential estimation approach which focuses initially on individual features, and subsequently on their interactions, may explain why EBM yields high predictive accuracy without providing interaction-based explanations. Similarly, while FSI does not employ a sequential process to identify the importance of features and interactions, it relies on linear regressions to estimate their contributions simultaneously. This approach can obscure the true impact of interaction terms because the importance of feature interactions may be overshadowed by the magnitude of the importance scores of individual features (as shown in Tables 6 and 7). Hence, FSI may fail to accurately represent the distinct impact of interaction terms, thereby limiting the interpretability and insights derived from interaction-based explanations. In contrast, our graph learning model adopts a novel methodology by first assessing the significance of feature interactions using GAT before evaluating the importance of individual features. This ensures that our interaction-based explanations can consistently enhance separation between patients with different LoS outcomes.

We illustrate this separation in Figures 3 and 4 where we present the T-SNE plots derived from explanations generated by our graph learning model. These plots compare the visual clustering of patients based on 30 features versus a combination of 30 features and 30 interactions, with both plots subjected to a perplexity of 100 and 10,000 iterations. Figure 3 includes feature interactions and reveals three clusters: patients in the bottom left cluster generally have longer ICU stays, those in the top center cluster have shorter stays, and a gradient from left to right in the lower right cluster indicates increasing LoS. Figure 4 represents a T-SNE plot without interactions which does not indicate a clear separation between clusters. The DSI scores are 0.171 and 0.097 for the T-SNE plots with and without interactions, respectively, which suggest that interaction-based explanations contribute to an improvement of 0.074 in DSI.

In Appendix E, we further evaluate the utility of the interaction-based explanations of our graph learning model, focusing on the degree of separation enabled by feature interactions. We demonstrate that the attention values attributed to two interactions, exhibit significantly different distributions across patient groups in Figures E1 and E2. Specifically, patients with ICU stays longer than seven days are more likely to exhibit salient attention on the interaction between patient age and respiratory system diagnoses. In contrast, patients with shorter stays are more likely to exhibit salient attention on the interaction between patient age and mental disorders. Although we highlight only two of the top ten interactions, the importance of all interactions in Table 7 underscore significant differences between the two patient groups. Our XAI approach, with its emphasis on interaction-based explanations and impact on predictive performance, contributes to the emergent research on explainable AI that requires novel ways of abstraction and formulation of new design principles and insights (Gregor and Hevner, 2013).

5. Evaluation of Model Explainability

In this section, we further validate the explanations generated by our graph learning model, based on the Co-12 framework (Nauta et al. 2023). The Co-12 framework is a collection of 12 key properties that can be used to systematically evaluate explanations generated by machine learning models. Evaluation of XAI methods is a nascent area of research and early studies have primarily focused on the assessment of post-hoc, feature-based XAI methods (Janizek et al. 2021, Kim et al. 2023). However, there is a notable research gap with respect to the systematic evaluation of intrinsic, interaction-based XAI methods. Our evaluation approach includes several tests designed to assess whether the explanations generated by the graph learning model adhere to the Co-12 framework. We provide a summary of our evaluation approach in Table 9.

5.1. Correctness, Completeness, and Compactness

The first two properties, correctness and completeness, are foundational to assess the quality of explanations provided by XAI methods. *Correctness* ensures that explanations accurately reflect the prediction of the underlying model, while *completeness* emphasizes the need for explanations to fully represent the model

decision-making process. Since our model explanations are intrinsically derived from the prediction model, the correctness of explanations is inherently assured. Since these explanations are generated directly by the model, they offer a complete view of the decision process by design. This relationship between model prediction and explanation distinguishes intrinsic methods from post-hoc alternatives. Hence, we argue that our graph learning model satisfies both correctness and completeness properties of the Co-12 framework.

The *compactness* property states that explanations should be succinct and sparse. Figure 2 suggests that the explanations of our graph learning model are compact, focusing on a limited set of key features and interactions. Furthermore, the results described in Appendix F demonstrate that the distribution of node and edge importance scores adheres to a zero-inflated pattern, indicating that our model assigns substantial attention only to a limited number of nodes and edges across all patients.

5.2 Consistency, Continuity, and Contrastivity

Next, we evaluate the explanations provided by our graph learning model through the lens of *consistency*, *continuity*, and *contrastivity*. These properties suggest that the ability of the XAI model to provide explanations should accurately reflect the importance of input features and its ability to generate reliable and meaningful insights. We evaluate these properties by perturbing the most and least significant diagnoses, i.e., respiratory system diagnosis as the most important and blood-forming organs as the least important. For patients diagnosed with both categories of conditions, we generate explanations using the original data, a perturbed version excluding blood-forming, organ-related diagnosis, and another excluding respiratory system-related diagnoses. The impact of perturbation analyses using aggregated graph-based explanations is presented in Figures 5a, 5b, and 5c.

While Figure 5a represents the explanations based on the original data, Figure 5b demonstrates that perturbation of a *less critical* diagnosis category—*blood-forming organs*—across the patient cohort, does not significantly alter the explanation or attention values of the nodes (features). For instance, the prominence of respiratory system diagnosis remains unaffected as do other important nodes such as *patient age*, *mental disorder*, and *injury and poisoning diagnosis*. This stability indicates that variations in less

critical diagnostic categories have negligible impact on model explanation. Conversely, Figure 5c demonstrates that perturbation of an *important* feature—respiratory system diagnosis—leads to a drastic change in the importance of other features and interactions. For example, patient age emerges as the most significant node in explaining LoS prediction when respiratory diagnosis is perturbed. Overall, perturbation analysis confirms our model adherence to the principles of consistency, continuity, and contrastivity.

5.3. Confidence

Next, we utilize logit regressions to assess the property of *confidence,* which is related to probability-based confidence measures of model explanation. Specifically, we focus on statistical confidence of the significance of interactions identified by our model. We compare two logit models: one with only salient features identified by our graph learning model as independent variables, and another which includes both salient features and interactions. The goal is to determine whether inclusion of the interaction terms improves the goodness of fit for predicting ICU stays. We present a comparison of the two logit regression models in Table 10. The logit regression utilizing interactions increases McFadden's R-square from 0.218 to 0.228, statistically significant based on the likelihood ratio test and reduces the Akaike Information Criterion (AIC) from 22,156 to 22,069. These results confirm the importance of salient interactions identified by our graph learning model and establish their statistical significance with high confidence.

5.4 Coherence and Covariate Complexity

Next, we evaluate the coherence of our model explanations with the medical literature. *Coherence* ensures that explanations align with domain knowledge while *covariate complexity* requires explanations to be understandable to the target audience. Such properties are particularly important as prior research suggests that XAI methods can enhance user trust in algorithms and confidence in decision making when designed using task-specific domain knowledge (Lee and Ram 2023). We assess these properties by cross-referencing salient interactions identified by our model with extant clinical research. Table 11 augments the top 10 salient interactions identified by our graph learning model, with clinical evidence which provide evidence-based support for the validity of our model explanations. For example, our model highlights the

interaction between "*nutritional and anti-inflammatory agents*" and *heart rate* as important predictors of ICU LoS. This is supported by the medical literature, which suggest that short-term usage of corticosteroids, a type of anti-inflammatory agent, is associated with significant decrease in heart rate and can lead to bradycardia (Brotman et al. 2005). Similarly, the interaction between respiratory system diagnosis and patient age reflects the impact of age on lung capacity and increased risk of respiratory failure, which may affect recovery time and LoS (Sharma and Goodwin 2006).

These results indicate that the explanations generated by our graph learning model are not only consistent with evidence-based medicine but also provide insights that are unavailable using traditional XAI methods. By using a socio-technical lens to evaluate explainability, we test and validate the insights from our approach using the rich domain knowledge on ICU LoS in the medical literature.

6. Discussion

In this section, we discuss the results of a small-scale user study based on our XAI approach and the implications of our proposed XAI approach for research and practice.

6.1 User-based Evaluation

To evaluate the usefulness of our graph learning model in clinical settings, we designed a small-scale user study based on a survey of ICU clinicians (Kim et al. 2023). The questionnaire was comprised of five statements based on the explanations between patient attributes and ICU LoS, as identified by our graph learning XAI approach. The respondents include six practicing ICU physicians in central Texas, who rated the statements on a 5-point Likert scale, where 1 = "Strongly Disagree," 2 = "Disagree," 3 = "Neither Agree nor Disagree," 4 = "Agree," and 5 = "Strongly Agree." The survey also included an open-ended question to elicit physician feedback regarding the feasibility of using our XAI approach to improve care delivery.

Tables H1 and H2 in Appendix H present the survey statements, mean Likert scores from respondents, and their written responses to the final question. The results show that ICU clinicians generally disagreed with the cohort-level statements (i.e., Q1 and Q2), with mean Likert scores below 3. However, they generally agreed with individual patient-level statements (i.e., Q3). The salient interaction between

patient age and *respiratory system diagnoses*, as illustrated in Figure 2, is discussed in statement 3a. Statements 3b and 3c provide additional explanations on the role of skin issues and mental disorders.

We anticipated that physicians would concur with our model prediction based on statement 3a, with statements 3b and 3c having minimal impact on physician opinion. We observe that physicians generally agreed with statement 3a, with a mean Likert score of 3.33, with similar responses to statements 3b and 3c. Hence, their feedback with respect to the individual patient-level insights of our XAI model indicates a valuable role in clinical settings, especially for care personalization based on unique patient characteristics. Four of the six physicians observed that our model can improve staffing and resource management efficiency, enabling better planning for patient scheduling. Our user study of ICU physicians highlights the impact of (a) salient design artifacts (i.e., feature interactions) on users who may otherwise ignore their importance for LoS prediction when using AI-enabled decision tools, and (b) sociotechnical considerations related to the use of our model for patient-centered care (Abbasi et al. 2024).

6.2 Research Implications

Our research demonstrates the importance of intrinsically generated explanations that identify important interactions between patient attributes for accurate prediction of ICU LoS. Compared to interaction-based XAI methods, our graph learning model accurately identifies complex, non-linear relationships in the underlying data, thereby offering a more nuanced understanding of their impact on LoS prediction. Our approach demonstrates that interaction-based explanations can provide more accurate and comprehensive understanding of the underlying prediction model, which is particularly important for risk prediction in healthcare. Furthermore, the results of our user study indicate general agreement among ICU clinicians with the patient-level explanations offered by our XAI approach and underscores the practical relevance of our graph learning model. Our model can enhance clinical decision-making and improve ICU operational efficiency by providing physicians with more transparent and comprehensible insights into LoS prediction not only in the ICU but other hospital settings as well.

From a methodology perspective, our model provides a unique solution to the challenge of

designing prediction models that are accurate and capable of providing intrinsic, interaction-based explanations. By operationalizing interaction-based explanations as a patient-level graph that describes the relationships between patient attributes, our model learns the structure of patient-level graphs by deploying an end-to-end, attention-based learning approach. Our approach provides accurate identification of the underlying feature interactions that explain predicted outcomes, and thereby, bridges extant research on graph learning and XAI methods. In other words, our research expands the application space of graph learning techniques and provides new tools for developing XAI methods. Overall, our graph learning approach provides a novel contribution from methodological and application perspectives to address the problem of generating intrinsically interpretable solutions for risk prediction.

Our results also provide a solution to the computation complexity of $O(N^2)$ associated with examining all potential interactions between model features. By adding intrinsic explanation capabilities to the prediction model, computation complexity is internalized in model training, thereby enabling more computationally efficient explanations compared to extant post-hoc methods. Our model also generates interaction-based explanations significantly faster than alternate post-hoc methods.

Our approach uses node attributes derived from structured numerical data. However, the healthcare sector also generates vast amounts of unstructured multi-modal data, including medical images and clinical notes. Integrating unstructured data using new types of AI methods into our XAI approach could significantly enhance its predictive and explainability. For instance, medical images, such as radiographs and MRIs, can be processed using CNNs to extract high-level representations. These CNNs, pre-trained on large medical image datasets, generate vector representations that represent important visual features and influence prediction of patient outcomes (Salehi et al. 2023). Similarly, recent advances in large language models have enabled synthesis of vector embeddings which represent large volumes of clinical notes. Transformer-based models, such as Med-PaLM, can generate contextualized vector embeddings encapsulating the semantic nuances in clinical notes (Singhal et al. 2024, Nazi et al. 2024).

It is also possible to align vector representations from distinct data modalities within a unified vector space, thereby facilitating integration into a common graph structure. Techniques such as projection

layers can map each the embeddings of each modality into a common dimensional space. Similarly, contrastive learning techniques, exemplified by models like Contrastive Language-Image Pre-training, can facilitate semantic alignment of representations across different modalities (Radford et al. 2021). By generating and storing unified vector representations of multimodal data, our approach can integrate vector representations as node attributes when constructing patient graphs. This integration improves model predictive accuracy and explanatory power by enabling cross-modality explanations.

The ability to include multi-modal data as nodes enhances the versatility of our graph learning approach. It can be adapted to other predictive tasks in healthcare, including prediction of readmission rates, disease progression, and risk of adverse events, which often share similar input data and necessitate similar types of explanations (Mahmoudi et al. 2020, Tjendra et al. 2020). For example, understanding how drug-drug interactions contribute to adverse events highlights the need for interaction-based explanations when predicting health outcomes (David et al. 2021, Magro et al. 2012).

7. Conclusions

In this study, we propose and test a novel graph learning-based explainable AI model to address the challenge of developing explainable predictions of ICU LoS. Our model intrinsically constructs a patient-level graph which identifies the importance of features and feature interactions during prediction. Our model demonstrates superior explanation capability based on identification of important feature interactions, compared to traditional XAI methods for predicting LoS. We supplement our XAI approach with a small-scale user study which demonstrates that our model provides accurate explanations that can lead to greater user acceptance of AI model-based decisions by contributing to greater interpretability of the predictive artifacts (Abbasi et al. 2024). Our model lays the foundation to develop interpretable, predictive tools which healthcare professionals can utilize to improve ICU resource allocation and enhance the clinical relevance of AI systems in providing effective patient care.

Although our primary research setting is the ICU, our graph learning model can be generalized to other healthcare contexts to accurately identify key feature interactions for prediction of other health

outcomes such as mortality, readmission risk, and hospitalizations. As experts increasingly recognize the limitations of post-hoc, feature-based explanations, our model offers a novel approach to generate intrinsic, interaction-based explanations that are promising in contexts that require an understanding of complex feature interactions and their impact on risk prediction (Petch et al. 2022, Carmichael and Scheirer, 2023).

7.1. Limitations and Future Research

In recognizing the limitations of our research, we also identify several avenues for future research. Our graph learning model is designed with broader applicability in mind. It can be adapted to various healthcare settings and potentially extended beyond healthcare applications. Future research may validate our model using data collected across a diverse group of hospitals, such as non-teaching institutions or safety-net hospitals, for various health outcome prediction tasks. We acknowledge that clinical features, such as diagnosis and medications, are represented in an abstract form, while clinical notes, radiology images, and lab results, are omitted due to data sparsity and quality challenges. Future research can extend our model to include such multi-modal data or evaluate the potential to utilize unstructured clinical notes by integrating LLMs with our graph learning model. While the user study offers preliminary insights into the practical relevance of our model, future studies may adopt a comprehensive approach, using randomized field experiments to study how XAI models can help practitioners improve care delivery. Last, although we provide a review of extant deep learning and XAI methods in Table A1 of the Appendix, we acknowledge the challenges and limitations in conducting an exhaustive review across multiple streams of literature.

References

Abbasi A, Parsons J, Pant G, Sheng ORL, Sarker S (2024) Pathways for design research on artificial intelligence. *Information Systems Research* 35(2):441-459.

Acharya UR, Kannathal N, Sing OW, Ping LY, Chua T (2004) Heart rate analysis in normal subjects of various age groups. *Biomedical Engineering Online* 3, (1):24.

Ahmad MA, Eckert C, Teredesai A, Acm (2018) Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*:559-560.

Al-Dailami A, Kuang H, Wang J (2022a) Attention-based memory fusion network for clinical outcome prediction using electronic medical records. *2022 IEEE International Conference on Bioinformatics and Biomedicine,* 902-907.

Al-Dailami A, Kuang HL, Wang JX (2022b) Predicting length of stay in ICU and mortality with temporal dilated separable convolution and context-aware feature fusion. *Computers in Biology and Medicine* 151:106278.
Alsinglawi B, Alshari O, Alorjani M, Mubin O, Alnajjar F, Novoa M, Darwish O (2022) An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific Reports* 12(1):607.
Alturaymi MA, Almadhi OF, Alageel YS, Bin Dayel M, Alsubayyil MS, Alkhateeb BF (2023) The association between prolonged use of oral corticosteroids and mental disorders: do steroids have a role in developing mental disorders? *Cureus* 15(4):e37627.
Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One* 10(7):e0130140.
Bauer K, von Zahn M, Hinz O (2023) Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research* 34(4):1582-1602.
Bauer K, Gill A (2024) Mirror, mirror on the wall: algorithmic assessments, transparency, and self-fulfilling prophecies. *Information Systems Research* 35(1):226-248.
Ben-Assuli O, Padman R (2020) Trajectories of repeated readmissions of chronic disease patients: risk stratification, profiling, and prediction. *MIS Quarterly* 44(1):201-226.
Boom M, Niesters M, Sarton E, Aarts L, Smith TW, Dahan A (2012) Non-analgesic effects of opioids: opioid-induced respiratory depression. *Current Pharmaceutical Design* 18(37):5994-6004.
Brotman DJ, Girod JP, Garcia MJ, Patel JV, Gupta M, Posch A, Saunders S, Lip GYH, Worley S, Reddy S (2005) Effects of short-term glucocorticoids on cardiovascular biomarkers. *Journal of Clinical Endocrinology & Metabolism* 90(6):3202-3208.
Carmichael Z, Scheirer W (2023) How well do feature-additive explainers explain feature-additive predictors? *XAI in Action: Past, Present, and Future Applications. https://openreview.net/forum?id=iqXixXrMKa*
Carvalho RM, Oliveira D, Pesquita C (2023) Knowledge graph embeddings for ICU readmission prediction. *BMC Medical Informatics and Decision Making* 23(1):12.
Chaddad A, Peng JH, Xu J, Bouridane A (2023) Survey of explainable AI techniques in healthcare. *Sensors* 23(2):634.
Chang C-H, Caruana R, Goldenberg A (2021) NODE-GAM: neural generalized additive model for interpretable deep learning. *International Conference on Learning Representations*. https://openreview.net/forum?id=g8NJR6fCCl8
Chen Z, Liang N, Zhang HL, Li HZ, Yang YJ, Zong XY, Chen YX, Wang YP, Shi NN (2023) Harnessing the power of clinical decision support systems. *Open Heart* 10(2):e002432.
Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun JM (2016) RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems 29*:3512-3520.
Crabbé J, van der Schaar M (2022) Concept activation regions: a generalized framework for concept-based explanations. *Advances in Neural Information Processing Systems* 35:2590-2607.
Dahl D, Wojtal GG, Breslow MJ, Holl R, Huguez D, Stone D, Korpi G (2012) The high cost of low-acuity ICU outliers. *Journal of Healthcare Management* 57(6):421-433.
David SP, Singh L, Pruitt J, Hensing A, Hulick P, Meltzer DO, O'Donnell PH, Dunnenberger HM (2021) The contribution of pharmacogenetic drug interactions to 90-day hospital readmissions: preliminary results from a real-world healthcare system. *Journal of Personalized Medicine* 11(12):1242.
Fernández-Loría C, Provost F, Han XT (2022) Explaining data driven decisions made by AI systems the counterfactual approach. *MIS Quarterly* 46(3):1635-1660.
Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS (2019) Adversarial attacks on medical machine learning. *Science* 363(6433):1287-1289.
Geneva II, Cuzzo B, Fazili T, Javaid W (2019) Normal body temperature: a systematic review. *Open Forum Infectious Diseases* 6(4).

Grabisch M, Roubens M (1999) An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory* 28(4):547-565.
Gregor S, Hevner AR (2013) Positioning and presenting design science researcg for maximum impact. *MIS Quarterly* 37(1): 337-355.
Guan S, Loew M (2022) A novel intrinsic measure of data separability. *Applied Intelligence* 52(15):17734-17750.
Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A (2019) Multitask learning and benchmarking with clinical time series data. *Scientific Data* 6(1):96.
Janizek JD, Sturmfels P, Lee SI (2021) Explaining explanations: axiomatic feature interactions for deep networks. *Journal of Machine Learning Research* 22(104): 1-54.
Jankovic SM, Pejcic AV, Milosavljevic MN, Opancina VD, Pesic NV, Nedeljkovic TT, Babic GM (2018) Risk factors for potential drug-drug interactions in intensive care unit patients. *Journal of Critical Care* 43:1-6.
Jiang G, Zhuang F, Song B, Zhang T, Wang D (2023) PriSHAP: prior-guided Shapley value explanations for correlated features. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 955-964.
Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) MIMIC-III, a freely accessible critical care database. *Scientific Data* 3(1):1-9.
Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* 32(3):713-735.
Khedkar S, Gandhi P, Shinde G, Subramanian V (2020) Deep learning and explainable AI in healthcare using EHR. *Deep Learning Techniques for Biomedical and Health Informatics*:129-148.
Kim BR, Srinivasan K, Kong SH, Kim JH, Shin CS, Ram S (2023) ROLEX: A novel method for interpretable machine learning using robust local explanations. *MIS Quarterly* 47(3): 1303-1332.
Kreuzer D, Beaini D, Hamilton W, Létourneau V, Tossou P (2021) Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems* 34:21618-21629.
Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jorgensen MJ, Lange J, Thiesson B (2020) Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications* 11(1):3852.
Lee HH, Cho JS, Lim YS, Hyun SY, Woo JH, Jang JH, Yang HJ (2019) Relationship between age and injury severity in traffic accidents involving elderly pedestrians. *Clinical and Experimental Emergency Medicine* 6(3):235-241.
Lee K, Ram S (2023) Explainable deep learning for false information identification: an argumentation theory approach. *Information Systems Research*. ePub ahead of print, https://doi.org/10.1287/isre.2020.0097
Li J, Larsen K, Abbasi A (2020) TheoryOn: A design framework and system for unlocking behvaiorla knowledge through ontology learning. *MIS Quarterly* 44(4):1733-1772.
Liao V, Gruen D, Miller S (2020) Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376590
Liu Y, Pant G, Sheng ORL (2020) Predicting labor market competition: Levergaing interfirm network and employee skills. *Information Systems Research* 44(4): 1443-1466.
Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623-631.
Ma M, Sun P, Li Y, Huo W (2023) Predicting the risk of mortality in ICU patients based on dynamic graph attention network of patient similarity. *Mathematical Biosciences and Engineering* 20(8):15326-15345.
Magro L, Moretti U, Leone R (2012) Epidemiology and characteristics of adverse drug reactions caused by drug–drug interactions. *Expert opinion on drug safety* 11(1):83-94.

Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK (2020) Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ* 369.
McKinnon SA, Holloway BM, Santoro MS, May AC, Cronan TA (2016) The effects of age, mental health, and comorbidity on the perceived likelihood of hiring a healthcare advocate. *Californian Journal of Health Promotion* 14(3):45-57.
Molnar C (2022) *Interpretable Machine Learning*, https://christophm.github.io/interpretable-ml-book/
Morid MA, Sheng ORL, Dunbar J (2023) Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems* 14(1):1-29.
Multz AS, Chalfin DB, Samson IM, Dantzker DR, Fein AM, Steinberg HN, Niederman MS, Scharf SM (1998) A "closed" medical intensive care unit (MICU) improves resource utilization when compared with an "open" MICU. *American Journal of Respiratory and Critical Care Medicine* 157(5):1468-1473.
Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, Schlötterer J, Keulen Mv, Seifert C (2023) From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Computing Surveys* 55(13s):1-42.
Nazi ZA, Peng W (2024) Large language models in healthcare and medical domain: A review. *Informatics*, 57.
Nie W, Yu Y, Zhang C, Song D, Zhao L, Bai Y (2023) Temporal-spatial correlation attention network for clinical data analysis in intensive care unit. *arXiv preprint arXiv:2306.01970*.
Nori H, Jenkins S, Koch P, Caruana R (2019) InterpretML: a unified framework for machine learning Interpretability. https://ui.adsabs.harvard.edu/abs/2019arXiv190909223N.
Oh JH, Zheng ZQ, Bardhan IR (2018) Sooner or later? health information technology, length of stay, and readmission risk. *Production and Operations Management* 27(11):2038-2053.
Petch J, Di S, Nelson W (2022) Opening the black Box: the promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology* 38(2):204-213.
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J (2021) Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748-8763.
Rai A (2020) Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science* 48(1):137-141.
Rocheteau E, Liò P, Hyland S (2021) Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. *Proceedings of the Conference on Health, Inference, and Learning*, 58-68.
Romano P, Hussey PD, Ritley D. 2014. *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives*: Agency for Healthcare Research and Quality. https://www.ahrq.gov/sites/default/files/publications/files/perfmeas.pdf.
Rotar EP, Beller JP, Smolkin ME, Chancellor WZ, Ailawadi G, Yarboro LT, Hulse M, Ratcliffe SJ, Teman NR (2022) Prediction of prolonged intensive care unit length of stay following cardiac surgery. *Seminars in Thoracic and Cardiovascular Surgery* 34(1):172-179.
Rubini A, Bosco G (2013) The effect of body temperature on the dynamic respiratory system compliance-breathing frequency relationship in the rat. *Journal of Biological Physics* 39(3):411-418.
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206-215.
Saadatmand S, Salimifard K, Mohammadi R, Kuiper A, Marzban M, Farhadi A (2023) Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients. *Annals of Operations Research* 328(1):1043-1071.
Salehi AW, Khan S, Gupta G, Alabduallah BI, Almjally A, Alsolai H, Siddiqui T, Mellit A (2023) A study of CNN and transfer learning in medical imaging: advantages, challenges, future scope. *Sustainability* 15(7):5930.
Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: visual

explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128(2):336-359.
Sharma G, Goodwin J (2006) Effect of aging on respiratory system physiology and immunology. *Clinical Interventions in Aging* 1(3):253-260.
Shweta K, Kumar S, Gupta AK, Jindal SK, Kumar A (2013) Economic analysis of costs associated with a Respiratory Intensive Care Unit in a tertiary care teaching hospital in Northern India. *Indian Journal of Critical Care Medicine* 17(2):76-81.
Singbartl K, Kellum JA (2012) AKI in the ICU: definition, epidemiology, risk stratification, and outcomes. *Kidney International* 81(9):819-825.
Singh KCD, Terwiesch C (2012) An cconometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50-65.
Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S (2023) Large language models encode clinical knowledge. *Nature* 620(7972):172-180.
Sun MX, Yang XB, Niu JH, Gu YF, Wang CT, Zhang WS (2024) A cross-modal clinical prediction system for intensive care unit patient outcome. *Knowledge-Based Systems* 283:111160.
Sundararajan M, Dhamdhere K, Agarwal A (2020) The shapley taylor interaction index. *37th International Conference on Machine Learning*, 9259-9268.
Sundararajan M, Taly A, Yan QQ (2017) Axiomatic attribution for deep networks. *34th International Conference on Machine Learning*, 3319-3328.
Tjendra Y, Al Mana AF, Espejo AP, Akgun Y, Millan NC, Gomez-Fernandez C, Cray C (2020) Predicting disease severity and outcome in COVID-19 patients: a review of multiple biomarkers. *Archives of pathology & laboratory medicine* 144(12):1465-1474.
Tong C, Rocheteau E, Veličković P, Lane N, Liò P (2021) Predicting patient outcomes with graph representation learning. *International Workshop on Health Intelligence*, 281-293.
Tsai CP, Yeh CK, Ravikumar P (2023) Faith-Shap: the faithful shapley interaction index. *Journal of Machine Learning Research* 24(1):4326–4367.
Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. *International Conference on Learning Representations*. https://arxiv.org/abs/1710.10903.
Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY (2020) A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 32(1):4-24.
Yasir M, Goyal A, Sonthalia S. (2023) Corticosteroid adverse effects. StatPearls Publishing, Accessed June 5th, 2024, https://www.ncbi.nlm.nih.gov/books/NBK531462/.
Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J (2019) Gnnexplainer: generating explanations for graph neural networks. *Advances in Neural Information Processing Systems* 32.
Zangmo K, Khwannimit B (2023) Validating the APACHE IV score in predicting length of stay in the intensive care unit among patients with sepsis. *Scientific Reports* 13(1):5899.
Zhang S, Liu Y, Shah N, Sun Y (2022) Gstarx: Explaining graph neural networks with structure-aware cooperative games. *Advances in Neural Information Processing Systems* 35:19810-19823.
Zhang YD, Zhang X, Zhu WG (2021) ANC: attention network for COVID-19 cxplainable diagnosis based on convolutional block attention module. *Computer Modeling in Engineering & Sciences* 127(3):1037-1058.
Zhu Y, Xu W, Zhang J, Du Y, Zhang J, Liu Q, Yang C, Wu S (2021) A Survey on graph structure learning: progress and opportunities. https://ui.adsabs.harvard.edu/abs/2021arXiv210303036Z.
Zimmerman JE, Kramer AA, McNair DS, Malila FM (2006) Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine* 34(5):1297-1310.

Figures and Tables

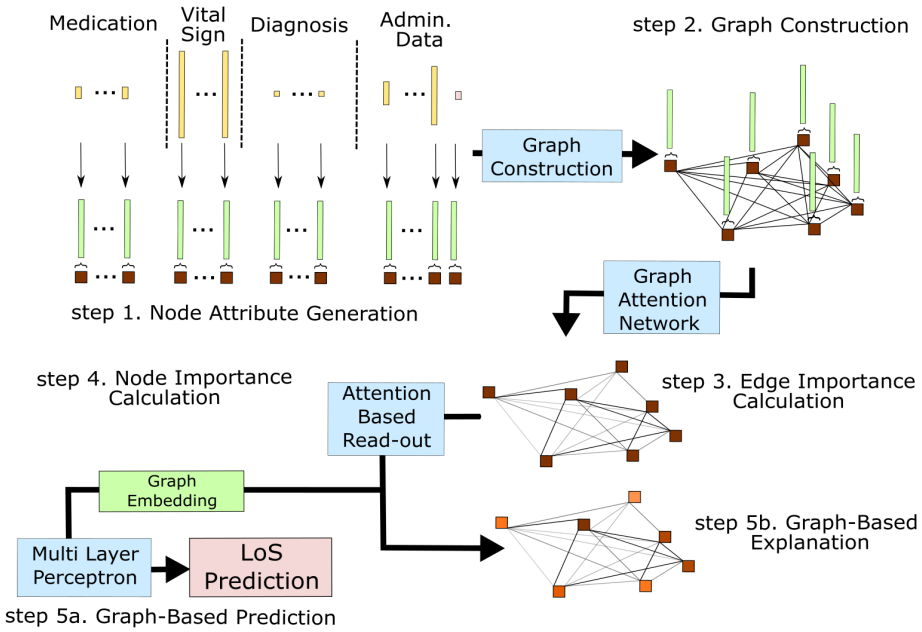

Figure 1. Design Framework of Graph Learning Model

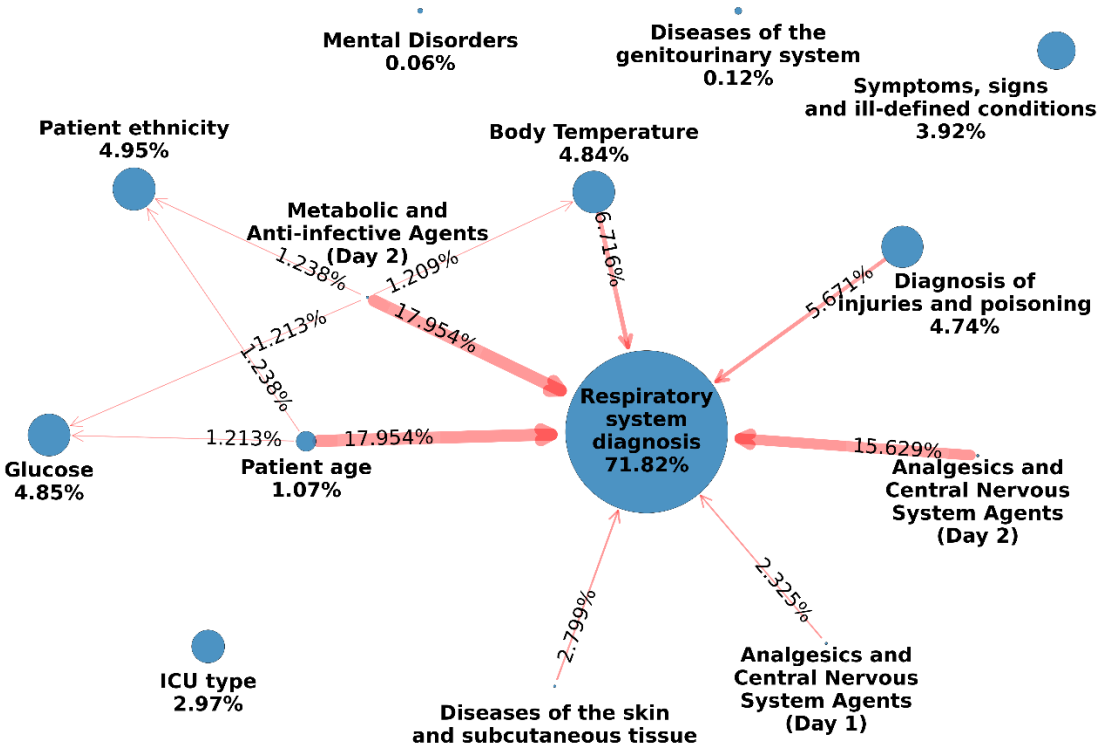

Figure 2. Personalized Explanations of Graph Learning Model

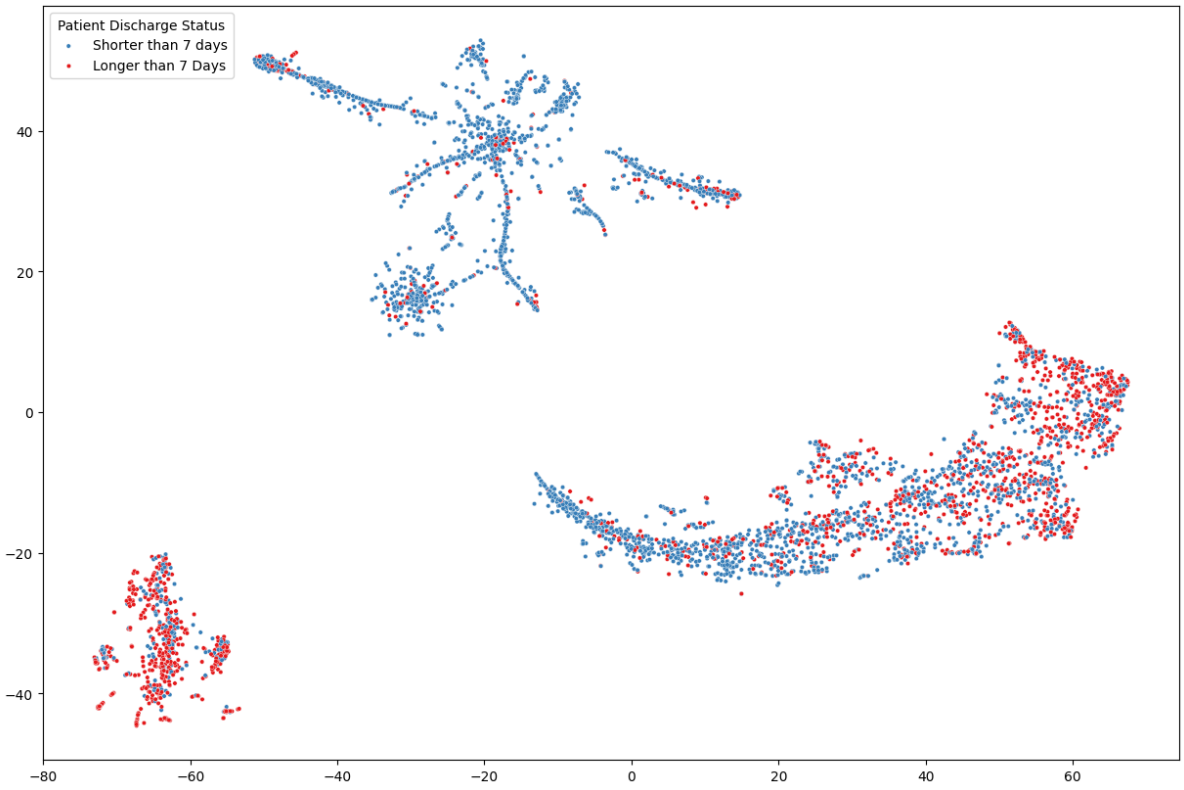

Figure 3. Visualization of T-SNE Separation with Feature Interactions.

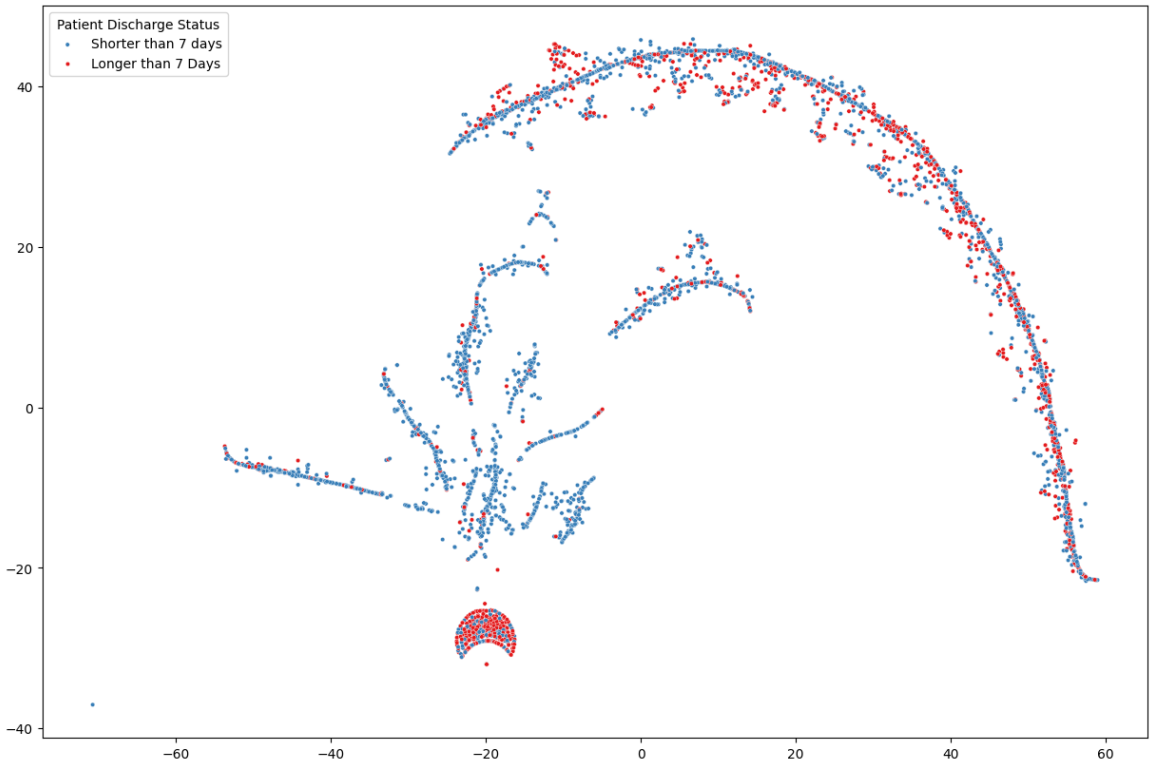

Figure 4. Visualization of T-SNE Separation without Feature Interactions

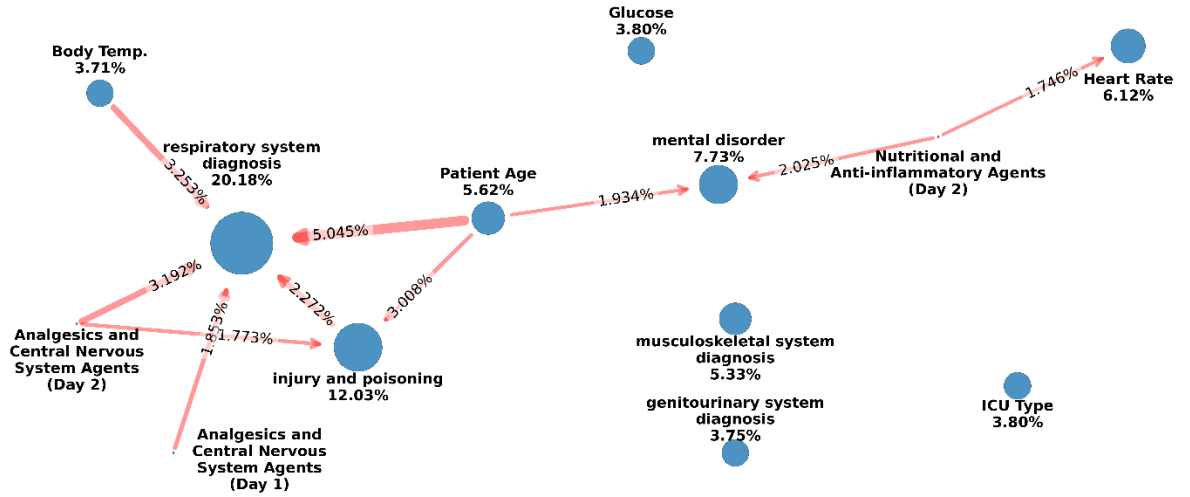

Figure 5a. Original Data

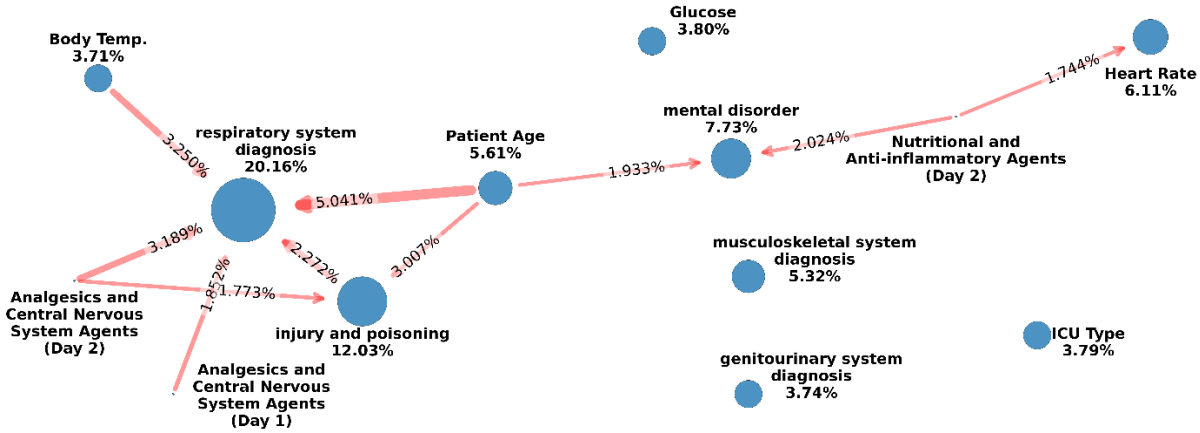

Figure 5b. Perturbation of Diagnosis associated with Blood-forming organs

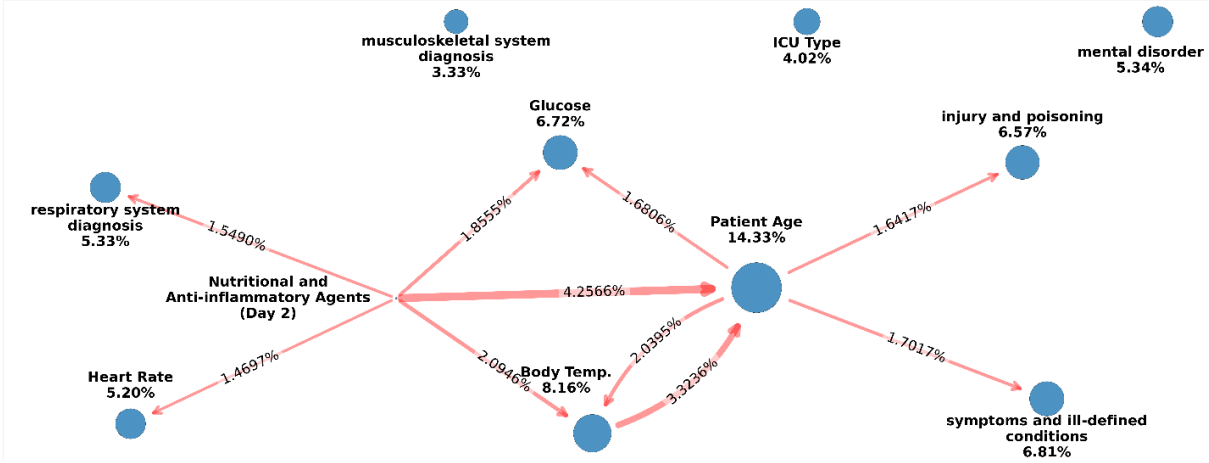

Figure 5c. Perturbation of Diagnosis associated with Respiratory system

Figure 5. Results of Perturbation Analysis

Table 1. Comparison of XAI Methods based on Explanation Type

	Feature-Based	Interaction-Based
Post-Hoc	Integrated Gradient (IG), Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive Explanations (SHAP)	Gradient based: Integrated Hessian (IH) SHAP based: Shapley Interaction Index (SII), Shapley Taylor Index (STI), Faithful Shapley Index (FSI)
Intrinsic	Regressions, simple decision trees	Generalized Additive Model with interactions (GA^2M) * Our graph learning-based model

Table 2. Summary of Research Contributions

XAI Method	LoS Prediction	Graph Learning	XAI
Our Graph Learning Approach	Predicts ICU LoS using *intrinsically explainable* predictions by identifying key interaction-based explanations.	Extends graph learning to XAI by synthesizing existing methods to construct interpretable graphs, and handle data without inherent graph structures.	Interaction-based explanations that enhance the explanatory power and interpretability of feature-based explanations.
EBM	Not used for LoS prediction.	Does not use graph structure for prediction or explanation.	Explanations based on feature attributes with less emphasis on feature interactions.
FSI	Not used for LoS prediction.	Does not utilize graph structure for prediction or explanation of outcomes.	Explanations based on feature attributes with less emphasis on feature interactions.

Table 3. Descriptive Statistics of Model Variables

Binary Output Variable

Variable Name	Description and Unit of Measure	Distribution
7-day discharge	Binary indicator of ICU patients discharged after the seventh day	28.79%

Selected Inputs

Variables Vital Signs

Heart Rate	Heart rate of the patient measured in beats per minute	94.06 (25.22)
Mean BP	Mean blood arterial pressure of the patient in mmHg	78.76 (14.02)
Respiration Rate	Patient respiration rate in breaths per minute	19.36 (5.20)
Body Temperature	Body temperature of the patient in degree Celsius	37.02 (0.83)
Glucose	Concentration of glucose present in the blood of patients in mg/dl	141.60 (55.82)
Bihourly Entry Count	Count of vital sign recordings during the 2-hour period	2.97 (4.32)

Administrative Variables

Age	Patient age in years	55.47 (27.59)
Gender*F*	Binary (1 = patient is female)	44.10%
Ins_Medicare	Binary (1 = patient insurance is Medicare)	47.65%
Adm_Elective	Binary (1 = patient from elective admission)	12.35%

Diagnosis Variables

Respiratory	Binary (1 = patient has respiratory system related diagnosis under ICD 9)	50.39%
Circulatory	Binary (1 = patient has circulatory system related diagnosis under ICD 9)	73.74%
Injury	Binary (1 = patient has injuries and poisoning diagnosis under ICD 9)	43.51%

Standard deviations (if applicable) are shown in parentheses.

Table 4. Predictive Performance Comparison

	Method	Accuracy	AUROC	AUPRC	F1 score
1	Graph Learning-based Model	0.767 (0.003)	**0.824 (0.002)**	**0.899 (0.003)**	0.829 (0.003)
2	DL Model	**0.771 (0.004)**	**0.824 (0.004)**	**0.899 (0.003)**	0.831 (0.004)
3	EBM	**0.771 (0.005)**	**0.824 (0.006)**	0.898 (0.005)	**0.839 (0.003)**
4	XGBoost	0.762 (0.004)	0.810 (0.006)	0.890 (0.005)	0.831 (0.003)
5	Random Forest	0.755 (0.05)	0.803 (0.007)	0.880 (0.005)	0.836 (0.003)
6	Logistic Regression	0.732 (0.005)	0.729 (0.007)	0.826 (0.007)	0.818 (0.004)

Standard deviations are shown in parentheses.

Table 5. Computation Time for Patient-Level Explanation

	Intrinsic Methods		Post-Hoc Methods	
XAI Method	EBM	Graph Learning Approach	DL-FS	DL-IH
Time (in seconds)	0.05	0.1	20	240

Table 6. Salient Features identified by XAI Methods

DL-FS Model		*EBM Model*		*Graph Learning Model*	
Feature Name	***Avg. Absolute FSI Score***	***Feature Name***	***Global Term Importance***	***Feature Name***	***Avg. Node Attention***
Respiratory system related diagnosis	0.03537	Respiratory system related diagnosis	0.56907	Respiratory system related diagnosis	0.11531
Patient Age	0.02792	Infectious and parasitic diseases	0.18134	Heart Rate	0.08628
Mean BP (Hr 46-48)	0.02719	Injuries and poisoning	0.15509	Patient Age	0.07790
Systolic BP (Hr 46-48)	0.01418	Diseases of the genitourinary system	0.10952	Injuries and poisoning	0.06780
Injuries and poisoning	0.01313	Symptoms, signs, and ill-defined conditions	0.10670	Mental Disorders	0.05178
Diastolic BP (Hr 46-48)	0.01080	Nervous system and sense organs diagnosis	0.10099	Body Temperature	0.04150
Glucose (Hr 46-48)	0.00951	Nutritional and Anti-inflammatory Agents (Day 2)	0.09521	Glucose	0.03807
Infectious and parasitic disease	0.00899	ICU type (CSRU)	0.08764	MSK and Connective Tissue Diagnosis	0.03585
Glucose (Hr 44-46)	0.00702	Analgesics and Central Nervous System Agents (Day 2)	0.07710	ICU Type	0.03502
ICU type (MICU)	0.00694	ICU type (MICU)	0.07092	Symptoms, signs, and ill-defined conditions	0.03161

CSRU: Cardiac Surgery Recovery Unit, MICU: Medical Intensive Care Unit.

Table 7. Salient Interactions identified by XAI Methods

DL-FS Model		*EBM Model*		*Graph Learning Model*	
Interaction Name	***Average Absolute FSI Score***	***Interaction Name***	***Global Term Importance***	***Interaction Name***	***Average Edge Attention***
Systolic BP × Mean BP (Hr 46-48)	0.0002	Metabolic and Anti-infective × Analgesics and Central Nervous System Agent (Day 2)	0.0399	Nutritional and Anti-inflammatory Agents (Day 2) -> Heart Rate	0.0301
Mean BP (Hour 46-48) × Insurance (Government Subsidy)	0.0002	Patient Age × Admission (elective)	0.0377	Patient Age -> Respiratory system related diagnosis	0.0285
Oxygen Level (Hr 46-48) × neoplasms	0.0002	Oxygen Level (Hr 12-14) × Vital Sign Count (Hr 34-36)	0.0337	Nutritional and Anti-inflammatory Agents (Day 2) -> Patient Age	0.0228
Oxygen Level (Hr 46-48) × Ethnicity (Eastern European)	0.0002	ICU type (CSRU) × Diseases of the genitourinary system	0.0297	Body Temperature -> Respiratory system related diagnosis	0.0199
Oxygen Level (Hr 46-48) × ethnicity (Filipino)	0.0002	Mean BP (Hr 44-46) × Diastolic BP (Hr 46-48)	0.0294	Body Temperature -> Patient Age	0.0169
Oxygen Level (Hr 46-48) + Dermatological and Respiratory Agents (Day 1)	0.0002	Diastolic BP (Hr 4-6) × Antineoplastic and Immunomodulating Agents (Day 2)	0.0248	Patient Age -> Injuries and poisoning	0.0166
Mean BP (Hr 46-48) × complications of pregnancy	0.0001	ICU type (CSRU) × Injuries and poisoning	0.0242	Analgesics and Central Nervous System Agents (Day 2)-> Respiratory system related diagnosis	0.0164
Glucose (Hr 44-46) × ethnicity (Thai)	6E-05	Heart Rate (Hr 20-22) × Nutritional and Anti-inflammatory Agents (Day 2)	0.0228	Nutritional and Anti-inflam Agents (Day 2) -> Mental Disorders	0.0142
Oxygen Level (Hr 38-40) × symptoms, signs, and ill-defined conditions	4E-05	Heart Rate (Hr 12-14) × Antineoplastic & Immuno Agents (Day 2)	0.0205	Patient Age -> Heart Rate	0.0131
Respiration Rate (Hr 38-40) × Diastolic BP (Hr 40-42)	3E-05	ICU type (MICU) × Respiratory system related diagnosis	0.0203	Patient Age -> Mental Disorders	0.0128

CSRU: Cardiac Surgery Recovery Unit, MICU: Medical Intensive Care Unit.

Table 8. Improvement in DSI with Inclusion of Salient Interactions

Hyperparameters			Top X features versus Top X features and Interactions		
Top X	**T-SNE Perplexity**	**T-SNE Iteration**	**DL-FS**	**EBM**	**Graph-Learning Model**
10	100	5000	-0.00061	-0.00332	***0.02935******
20	100	5000	-0.00061	***-0.01671****	***0.03470****
30	100	5000	0.00007	0.00295	***0.04210******
10	100	10000	0.00049	-0.00352	***0.03572******
20	100	10000	0.00169	***-0.01561****	***0.04246****
30	100	10000	-0.00072	-0.00059	***0.05194******

* p-value < 0.1, ** p-value < 0.05, *** p-value < 0.01

Table 9. Evaluation of Co-12 Properties

Co-12 Property	Interpretation	Evaluation Approach
Correctness	The explanation should correctly describe the behavior of the underlying black box model	*Section 5.1*
Completeness	The explanation should comprehensively describe the behavior of the underlying black box model	
Compactness	Offer sparse but meaningful explanation	
Consistency	Identical inputs should have identical explanations	*Section 5.2*
Continuity	Similar inputs should have similar explanations	
Contrastivity	Different inputs should have different explanations	
Confidence	Explanation should contain accurate probability information	*Section 5.3*
Covariate complexity	Explanations should offer appropriate feature complexity that are comprehensible	*Section 5.4*
Coherence	Explanation should align with prior knowledge and beliefs	
Composition	Explanations should be similar to real counterparts	*Section 6.1*
Context	User should be able to understand the explanation and act upon it	
Controllability	User should be able to influence the explanation through interactions	

Table 10. Goodness of Fit of Logistic Regressions

	Top 10 Features with highest node attention	Top 10 interactions with highest edge attention	AIC	McFadden's R-square	P-Value of Likelihood Ratio Test
1	✓		22156	0.218	< 2.2e-16
2	✓	✓	22069	0.228	

Note: Dependent variable: ICU LoS <= 7 days.

Table 11. Evidence-based Support from Medical Literature

Salient Interactions	***Relevant Clinical Findings***
Nutritional and Anti-inflammatory Agents (Day 2) -> Heart Rate	Short term utilization of corticosteroids is associated with significant decrease in heart rate and can lead to bradycardia (Brotman et al. 2005)
Patient Age -> Respiratory system related diagnosis	Elder patients are known to have reduced lung capacity, which can contribute to respiratory failure (Sharma and Goodwin 2006).
Nutritional and Anti-inflammatory Agents (Day 2) -> Patient Age	Patient age is known to influence the potential adverse effects of corticosteroids (Yasir et al. 2023)
Body Temperature -> Respiratory system related diagnosis	Body temperature is known to influences breathing patterns and respiratory mechanics (Rubini and Bosco 2013).
Body Temperature -> Patient Age	Normal body temperature differs based on age (Geneva et al. 2019).
Patient Age -> Injuries and poisoning	Injury severity increased as age increased (Lee et al. 2019).
Analgesics and Central Nervous System Agents (Day 2) -> Respiratory system related diagnosis	Opioids utilization can lead to opioid-induced respiratory depression (Boom et al. 2012)
Nutritional and Anti-inflammatory Agents (Day 2) -> Mental Disorders	Corticosteroids utilization can lead to a variety of mental health problems, such as such as anxiety, depression, and psychosis (Alturaymi et al. 2023)
Patient Age -> Heart Rate	Heart rate variability, a reliable indicator of heart condition, becomes less random and more predictable with aging (Acharya et al. 2004).
Patient Age -> Mental Disorders	Older adults are more prone to cognitive and mood disorders, with late-life depression linked to increased disability, poorer physical health, and higher mortality rate (McKinnon et al. 2016).