İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (2): 770–789

https://doi.org/10.26650/acin.1835775
Submitted: 04.12.2025
Revision Requested: 18.12.2025
Last Revision Received: 19.12.2025
Accepted: 23.12.2025

Research Article

Open Access

Explainable Graph Neural Networks in Intensive Care Unit Mortality Prediction: Edge-Level and Motif-Level Analysis

Şebnem Akal [1]

[1] Istanbul University, Faculty of Science, İstanbul, Türkiye

Abstract Accurate forecasting of ICU patient outcomes is essential for clinical decision support. However, most high-performing machine learning models function as black boxes, limiting their interpretability and clinical adoption. This study introduces a graph-based explainable risk-prediction framework, in which patient–patient relations are modeled through a diagnosis-based similarity network. An undirected graph was derived from the eICU-CRD demo subset (PhysioNet v2.0) by linking individuals sharing three-digit ICD-9 categories, and a GCN was trained for in-hospital mortality prediction. Despite the dataset's modest size and imbalance, meaningful discrimination was achieved (AUROC = 0.708; AUPRC = 0.308). A two-layer explainability analysis was applied to clarify the model's decision process. Each prediction was driven by a combination of patient-specific clinical attributes and signals from a small number of influential neighbors, according to GNNExplainer. SubgraphX, a Shapley-value-based motif discovery method, identified compact and clinically coherent subgraphs with strong causal influence on the prediction. Consistency between edge- and motif-level explanations indicated that the model relies on stable relational patterns with clinical relevance. These findings suggest that integrating GNNs with structured explainability methods can transform a single risk score into a transparent, data-driven decision-support mechanism that provides interpretable and hypothesis-generating insights to clinicians.

Keywords Explainable Graph Neural Networks · GNNExplainer · SubgraphX

Citation: Akal, Ş. (2025). Explainable Graph Neural Networks in Intensive Care Unit Mortality Prediction: Edge-Level and Motif-Level Analysis. Acta Infologica, Advance Online Publication. . *Acta Infologica*, *9*(2), 770-789. https://doi.org/10.26650/acin.1835775

Corresponding author: Şebnem Akal : sebnem.akal@istanbul.edu.tr

INTRODUCTION

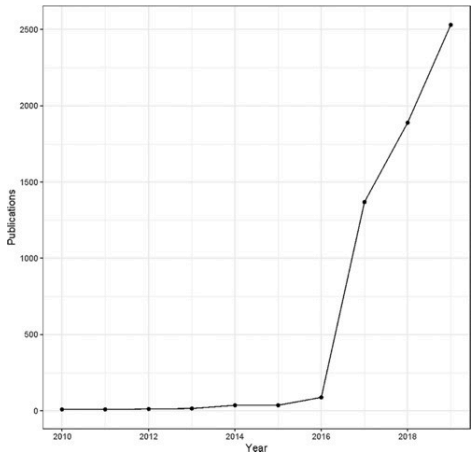
Deep learning–based methods are widely used in healthcare and often achieve high predictive accuracy; nevertheless, they are commonly characterized as opaque models whose internal workings are difficult to interpret. Although human cognition is complex, understanding the world is tightly bound to cause–effect relations. Thus, significant trust issues arise when AI systems do not clarify causal relationships, especially in critical clinical choices. Explainable Artificial Intelligence (XAI) is crucial in healthcare applications, where transparency, trust, and accountability are vital.

In recent years, high predictive accuracy has been reported for artificial intelligence (AI) in domains such as medical imaging, clinical decision-support systems, and early diagnosis. Nevertheless, many underlying models are characterized as black boxes because their decision-making processes are not readily comprehensible to humans. Consequently, concerns about transparency and trust are raised, particularly for high-stakes clinical decisions. A survey of 1,600 scientists published in Nature reported that 69% of respondents were concerned about excessive reliance on AI's ability to recognize patterns without understanding them, and 66% were concerned about these systems' potential to generate misinformation and bias (Nature, 2023). Accordingly, in XAI, solutions that focus not only on predictive performance but also on ensuring that decision processes are auditable to human experts are pursued (Van Noorden & Perkel, 2023).

As noted by Adadi and Berrada, the interpretability and explainability of ML algorithms are critical for three reasons. First, in high-stakes domains such as healthcare, law, and finance, the identification of the party responsible in the event of system failures is required, thereby enabling accountability. Second, understanding and articulating the causes of errors are essential for improving system reliability and preventing recurrence. Third, knowing why a system performs well when it does and why it performs poorly when it does promotes broader acceptance and deployment across domains (Adadi & Berrada, 2018). For these reasons, XAI is regarded not only as a technical property but also as an ethical, legal, and operational necessity (Tjoa & Guan, 2021). XAI approaches are developed to render model decisions interpretable to human experts.

Figure 1

Rising Trend in Research Interest



Source: Watson, 2022

Figure 1 shows the number of academic publications from 2010 to 2019 whose titles, abstracts, or keywords contain "interpretable machine learning" or "explainable artificial intelligence" (Watson, 2022).

Conventional AI models are largely constrained to flat (tabular) data representations. Healthcare data are characterized by complex relational structures. Accordingly, graph representations and, in particular, graph neural network (GNN)–based models—are considered better suited to capturing these multidimensional dependencies.

However, because GNNs learn over graph-structured domains that carry explicit topology, their interpretation is more challenging than that of conventional DN networks. In this context, GNNExplainer and PGExplainer are used to visualize and explain node-level decision mechanisms.

This study examines the ways in which GNN-based models can be interpreted on healthcare data. In conjunction with XAI approaches, an assessment is conducted to enhance the explainability of GNN models. The role of modern graph-based models in healthcare applications is discussed by drawing on the explainability of early systems such as MYCIN. This study first introduces the concept of XAI, then reviews the explanation methods developed for GNNs, and finally presents visualizations and an interpretability analysis on an illustrative dataset.

Background and Related Work

The origins of explainable AI predate the rise of deep learning and trace back to the knowledge-based expert systems developed in the late 1960s and early 1970s. These systems were explainable because an inference engine could traverse the underlying knowledge base to expose the chain of reasoning that led to a conclusion. Although significant advances in explainability were made during this period, the problem was not fully resolved (Adadi & Berrada, 2018).

Among the pioneering systems, Dendral was developed to perform structural elucidation of unknown organic molecules. It operated on domain-specific production rules and was therefore regarded as one of the first AI systems that emphasized expert knowledge over general problem-solving strategies. Its knowledge-intensive architecture subsequently served as a foundation for medical systems (Buchanan & Shortliffe, 1984).

Dendral provided evidence that domain-specific production rules could be more effective than general-purpose problem-solving methods for addressing complex scientific problems. Mycin, which was developed as a rule-based inference system to offer diagnostic and therapeutic recommendations for bacterial infections One of Mycin's most innovative features was its ability to explain its recommendations in human-understandable terms. In contrast to today's black-box "just-works" models, early AI in medicine was designed so that decision processes could be traced and interpreted. Therefore, this is regarded as one of the earliest examples of explainable AI. Today, XAI approaches to pursue interpretability alongside performance in deep learning and GNNs play a critical role in making clinical decisions verifiable and trustworthy.

While prior studies have demonstrated the potential of applying GNNs to patient networks for clinical prediction tasks, further research on model interpretability is needed. This study addresses the gap by training a simple yet effective GNN on the eICU-CRD demo subset (PhysioNet, v2.0) and by providing patient-level (local) explanations of model decisions using an XAI technique (GNNExplainer).

In this context, XAI and related notions, such as interpretable machine learning, have long been investigated under the "explanation problem" rubric for expert systems (Sağıroğlu & Demirezen, 2022). In contemporary practice, high-accuracy yet opaque black-box models have become prevalent, particularly with the proliferation of deep learning in healthcare. Although the inherently intuitive and complex nature

of human cognition is acknowledged, the inability of computer-implemented AI systems to articulate clear cause–effect relations raises serious trust concerns, especially for high-stakes medical decisions.

XAI has been cited to enable reliable diagnostic applications in healthcare, more efficient treatment planning, and more trustworthy drug discovery. Physicians and other healthcare professionals can use it to obtain dependable information and decision support and facilitate knowledge discovery. It has also been reported that an explainable-AI approach was employed within the "Turkish Brain Project," conducted in collaboration between the Digital Transformation Office of the Presidency of the Republic of Türkiye and Gazi University (Sağıroğlu & Demirezen, 2022).

In recent years, machine learning models used in healthcare have achieved high predictive performance; however, the mechanisms underlying their decisions often remain opaque. TreeExplainer was introduced by Lundberg et al. (2020) to enhance the explainability of tree-based models used in clinical applications. The method goes beyond individual predictions to enable a more global understanding of model behavior. Effective explainability can be provided for critical clinical decision problems such as mortality risk, progression of chronic kidney disease, and length of hospital stay. Analogous explainability techniques are also needed for graph-based deep learning models (Lundberg et al., 2020).

An artificial neural network trained to predict whether patients with pneumonia should be hospitalized or managed as outpatients was found to exploit a pattern that yielded a mistaken clinical recommendation. This case underscores the importance of interpretability of the model. Following this finding, the clinical use of the system was stopped (Caruana et al., 2015).[1]

Errors in transportation, healthcare, law, finance, and the military can have serious ramifications; however, healthcare is one of the domains in which consequences may be most severe. This study aims to elucidate and exemplify the use of Explainable Artificial Intelligence (XAI) to make the decision processes of GNNs used in healthcare more interpretable and transparent. Understanding how such models operate is critical for developing reliable, explainable decision-support systems for the diagnosis of respiratory diseases such as asthma and pneumonia. In this context, GNNExplainer was applied to a GCN trained on a patient graph derived from a publicly available, comprehensive dataset.

In line with this approach, the principal contributions are as follows: (i) a patient–patient graph is constructed from ICU patients by linking shared diagnosis codes (three-digit ICD-9 categories), and a GNN-based classification model is trained on this network; (ii) the GNNExplainer XAI technique is applied to the constructed patient graph to provide visual and statistical accounts of the salient patient features and influential neighborhood relations underlying model decisions; and (iii) a detailed single-patient case analysis demonstrates the potential value of these explanations for clinical decision support. Accordingly, a concrete framework is offered for the use of GNN and XAI techniques to understand and interpret patient risk in the ICU setting.

[1] An early case—dating to the mid-1990s and detailed by Caruana et al. (2015)—underscored the critical importance of XAI in healthcare. An artificial neural network was trained to decide whether patients with pneumonia should be hospitalized or treated as outpatients; the model learned that patients with pneumonia and asthma had a low mortality risk and therefore recommended against admission. This counterintuitive pattern arose from the failure to capture the causal chain in the training data. The potential danger was recognized by clinicians, and the system was withdrawn from clinical use. This case demonstrates that explainability in healthcare is not only a technical requirement but also an ethical and safety imperative.

Definition

Van Lent et al. first introduced the term explainable AI (XAI) in 2004. In that work, the term was used to denote a system's capability to explain the behavior of AI-controlled entities within a simulation-game application. (Adadi & Berrada, 2018).

Interest in XAI first appeared in the 1970s with the need to justify expert systems' decisions. In the 1990s, mistaken inferences arising from artificial neural networks used in medical decision-making highlighted the importance of interpretability (Caruana et al., 2015). The term XAI was first defined in 2004, and the field was revived in 2017 by DARPA's XAI program (Adadi & Berrada, 2018).

Interpretability in machine learning was not added ex post; rather, it originated with linear regression and statistical modeling in the 19th century. Pioneers, such as Gauss and Legendre, enabled prediction via transparent, closed-form expressions. In other words, the notion of interpretability arose from statistics and not from AI. In the latter half of the 20th century, interpretability remained integral to models through decision trees and rule-based systems (e.g., CART, ID3), alongside margin-based methods such as SVM. During this period, the traceability of rules was emphasized more than maximizing the predictive accuracy. With the success of random forests and boosting, the accuracy–interpretability trade-off became more salient. Debate then intensified over models that achieved accurate results while failing to explain their reasoning. This issue is discussed in Breiman's "Statistical Modeling: The Two Cultures" (Breiman, 2001). As DL models entered safety-critical domains such as medicine and finance, an interpretability crisis ensued. In response, model-agnostic approaches were developed to produce post hoc explanations, exemplified by LIME and SHAP.

Table 1

Emergence and development of XAI

Year	Event
1970s	The first mechanisms of explanation were developed in expert systems.
1990s	The erroneous inference made by the ANN-based pneumonia system demonstrated the need for XAI (Caruana et al., 2015)
2004	The term XAI was first defined with its currently accepted meaning.
2015	Interpretable models for healthcare
2016	Local interpretable model-agnostic explanations (LIME) method
2017	Launch of the DARPA XAI program
2018	The right to explanation under the GDPR as a legal obligation
2020-	The combined use of GNN and XAI in healthcare through SHAP, GNNExplainer, and PGExplainer.

Today, XAI is understood not only as a technical need but also as a prerequisite for trust in ethical, legal, and societal institutions. Regulatory frameworks (e.g., the EU General Data Protection Regulation, GDPR), the risks of clinical error, and the deployment of autonomous systems have rendered this need an obligation. Both model-agnostic and model-specific explanation methods, including SHAP and GNNExplainer, have been developed.

Paradigm Shift and the Black-Box Era's Rise

As AI reached an inflection point with striking advances in machine learning, the pace of resolving the interpretability problem slowed. The focus of AI research was shifted toward deploying models and

algorithms that prioritize predictive power, while the capacity to explain decision processes was relegated to the background. This shift laid the groundwork for DL. Although these models achieve exceptional predictive performance, they have been regarded as opaque rather than transparent, hindering insight into their internal mechanisms. (Adadi & Berrada, 2018).

Methods of XAI

XAI methods have been classified along several axes in the literature—including complexity (intrinsic/ante hoc vs. post hoc), scope (global vs. local), and model dependence (model-specific vs. model-agnostic) (Adadi & Berrada, 2018). In this study, a post hoc, local interpretability approach was adopted to make the decisions of the developed GNN transparent while preserving its predictive accuracy. Accordingly, GNNExplainer was employed to identify the most influential subgraph structure and feature subset underlying the model's prediction for a given patient (Ying, Bourgeois, You, Zitnik, & Leskovec, 2019).

Within this scope, XAI methods often remain superficial in perception-oriented approaches—such as saliency maps, signal-processing techniques, or language-based explanations—and that they fail to make concrete contributions to model-development workflows. At the mathematical level, difficulties have been emphasized, including the challenge of simplifying highly parameterized models, global–local mismatches, and explanations that amount to mere numerical artifacts. One of the most salient critiques is that medical XAI systems have not been adequately evaluated in terms of how well they meet clinical requirements. Accordingly, XAI research should be assessed not only for technical correctness but also for the interpretability and usability of its explanations by healthcare professionals (Tjoa & Guan, 2021).

In response to the frequently noted shortcomings of medical XAI—namely, superficiality and detachment from clinical context (Tjoa & Guan, 2021), this study not only reports technical metrics but also substantiates the practical significance of the generated explanations through a case study and an ablation study, thereby underscoring the model's potential value for clinical decision-making.

Challenges in the Medical XAI

Data, such as medical images, are typically available in far smaller quantities than ordinary images. Their acquisition often necessitates patient consent and administrative approval. Moreover, the high dimensionality of medical data increases processing complexity, and due to substantial memory demands, such data may not be used as direct inputs without modification, random subsampling, or dimensionality reduction, which can jeopardize analytical accuracy. It cannot be guaranteed that an algorithm captures the pertinent features without firm assurance that fine-grained details have not been inadvertently discarded (Tjoa & Guan, 2021).

Another concern is that some attention/saliency maps may highlight seemingly irrelevant regions, indicating that predictions may be driven by clinically irrelevant cues—even when the final decision is correct. The context dependence of verbal interpretability can also be misleading; for example, rule-like statements such as "if asthma, then lower risk" can be misleading in the absence of an account of the underlying reasoning process (Tjoa & Guan, 2021).

METHOD

Predicting patient outcomes on complex electronic health records (EHRs), such as ICU data, has been recognized as an important research problem in machine learning. In recent years, graph neural network (GNN) architectures have been increasingly employed to capture latent interpatient relations and similar-

ities. Below, several representative studies that construct patient–patient graphs on ICU datasets and adopt GNN-based approaches are briefly reviewed.

In one of the pioneering studies in this area, the eICU dataset was used to construct a patient–patient network based on diagnostic similarity (Tong, Rocheteau, Veličković, Lane, & Liò, 2021). A GNN (GraphSAGE) was trained on this graph after time-series data were processed with a Long Short-Term Memory (LSTM) network to predict ICU length of stay. The hybrid LSTM–GNN model outperformed approaches that relied on LSTM alone, demonstrating the applicability of diagnosis-based patient graphs for clinical prediction tasks. Similarly, co-diagnosis–based patient networks have been constructed.

In a study on the MIMIC-III dataset, the SBSCGM was proposed, in which a dynamic graph was constructed from patient similarities across multimodal EHRs. A hybrid architecture combining GCN, GraphSAGE, and GAT layers was employed, and high AUROC was achieved for mortality prediction. Unlike this work, interpretability was pursued via attention mechanisms rather than GNNExplainer (Sahu & Roy, 2025).

Studies employing alternative graph-construction strategies have also been reported. In the WPN study, a bipartite patient–disease network was first constructed from hospital records. This network was then projected into a patient–patient network to obtain a weighted graph. Training a GNN on this graph yielded high accuracy for predicting chronic diseases. In another study using the MIMIC-III dataset, embedding vectors derived from diagnosis, medication, and procedure codes were employed, and the data were analyzed using a k-nearest neighbors (k-NN) algorithm (Lu & Uddin, 2021).

Graph neural networks (GNNs) have increasingly been used in healthcare data to model inter-patient relations and complex interactions. After reviewing the relevant literature, we devised an interpretable risk-prediction framework for ICU patients. A patient–patient graph was constructed from the eICU-CRD demo subset (PhysioNet, v2.0) based on diagnostic similarity (three-digit ICD-9 categories). A Graph Convolutional Network (GCN) was trained to predict a critical clinical endpoint—mortality—on this graph, and the model's predictions were analyzed using GNNExplainer, an XAI technique. This section presents the full methodology in sequence: from dataset preparation and graph construction to the GCN architecture and training, followed by interpretation of the resulting explanations.

Objective

The eICU-CRD demo subset (PhysioNet, v2.0) was used. This demo is a small illustrative sample of the eICU Collaborative Research Database, whose full release contains records for over 200,000 ICU admissions from 208 U.S. hospitals during 2014–2015. The aim is: (i) to construct a diagnosis-based patient–patient graph by linking individuals who share three-digit ICD-9 categories and to train an appropriate GNN to predict a binary endpoint (mortality); and (ii) to apply GNNExplainer to attribute the prediction to both node features and influential neighbor connections.

Dataset

The eICU Collaborative Research Database (eICU-CRD) is a large-scale, multicenter, de-identified healthcare dataset collected from ICUs via Philips Healthcare's eICU program and has been made available to the research community by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT). It is used for observational studies of ICU patient care, treatment protocols, and clinical outcomes (Pollard et al., 2018).

The dataset comprises records collected between 2014 and 2015 from 335 ICUs across 208 hospitals across multiple U.S. regions. It encompasses more than 200,000 ICU admissions and more than 139,000 unique patients. Longitudinal physiological measurements were recorded at regular intervals throughout each ICU stay. All patient and provider identifiers have been fully de-identified in accordance with ethical and privacy guidelines, and each patient is represented by a unique anonymized identifier.

The dataset is organized as a structured schema of 31 interrelated tables. Among the key tables are those containing patient information, laboratory test results, medications, diagnoses, treatments/therapies, and clinical assessments (Goldberger et al., 2000; Pollard et al., 2018).

The eICU-CRD can be used to address numerous critical care research questions, including the following:

- Epidemiology and outcomes of sepsis, ARDS, and other critical illnesses
- Comparison of the effects of alternative treatment strategies and medications on patient outcomes
- Development of machine learning and artificial intelligence models to predict clinical deterioration
- Analysis of resource utilization and care quality in the ICU
- Development and evaluation of clinical decision-support systems

In the eICU-CRD demo subset, 2,520 ICU stays were included, with an approximate positive-class prevalence of 5%. Graph nodes represent patients, and features encode basic demographic and temporal attributes. Four numeric variables are available: age, length of ICU stay (hours), weight at ICU admission, and weight at ICU discharge.

Graph Construction

A patient–patient network was constructed to model inter-patient relations using diagnostic information extracted from the dataset. The pipeline comprised three stages: data preprocessing, similarity computation, and sparsification.

- Data Preprocessing and Node Definition

Diagnosis codes were extracted from the diagnostic table and truncated to the three-digit ICD-9 categories so that clinically related diagnoses could be grouped under the same category. To balance the influence of extremely rare and extremely prevalent codes, frequency-based filtering was applied: three-digit ICD-9 categories with corpus frequency < 5 or > 300 were excluded, and only codes within the range [5, 300] were retained for analysis.

- Data Preprocessing and Node Definition

Diagnosis codes were extracted from the diagnostic table and truncated to the three-digit ICD-9 categories so that clinically related diagnoses could be grouped under the same category. To balance the influence of extremely rare and extremely prevalent codes, frequency-based filtering was applied: three-digit ICD-9 categories with corpus frequency < 5 or > 300 were excluded, and only codes within the range [5, 300] were retained for analysis.

- Edge construction and weighting

G (V, E) denotes the patient graph, with V representing the set of patients. The nodes (V) and edges (E) of the graph were constructed by linking patients who shared the same three-digit ICD-9 codes. The similarity between the two patients was computed based on the rarity of their shared diagnosis codes. The Adamic–Adar similarity metric was employed to connect patients sharing identical ICD codes. Each shared diagnosis

code contributes to the similarity score in inverse proportion to its global frequency in the dataset. Thus, patients with a rare diagnosis receive a much higher similarity weight than those with a common diagnosis.

Under a same-hospital filter, pairwise similarities were computed using the Adamic–Adar metric. For patients with multiple ICU admissions, additional intra-patient edges were introduced to link different stays of the same individual. To indicate temporal continuity without dominating the structural topology, these edges were assigned a fixed, low weight. The graph was designed to capture the longitudinal progression of patient trajectories by incorporating distinct admissions of the same patient across different time points.

To reduce computational overhead, the graph was sparsified by retaining only the top k = 15 neighbors with the highest similarity scores for each node. This procedure ensured that each patient was connected only to the most strongly associated individuals, thereby maintaining graph sparsity while preserving structural relations.

- Graph Properties

The node degree distribution was analyzed to understand the structural characteristics of the patient–patient network (Figure 2). The distribution follows a long-tail pattern, indicative of a scale-free network, which is a common feature observed in real-world graphs, as shown in the figure.

This structure reveals the presence of some highly connected nodes (hubs) that serve as central points in the network topology and maintain a disproportionate number of edges.

Figure 2

Distribution of Node Degree of the Patient Graph

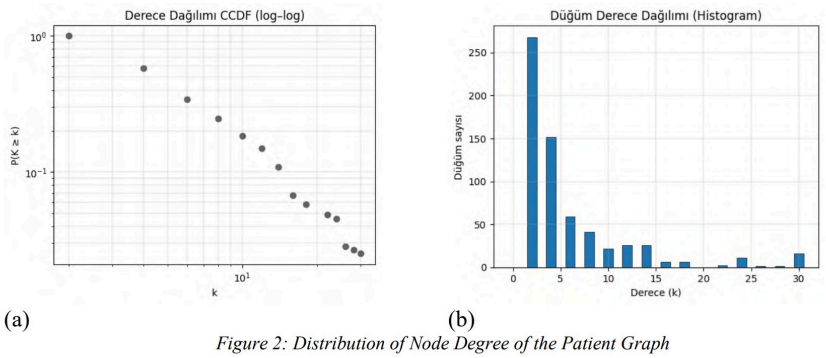


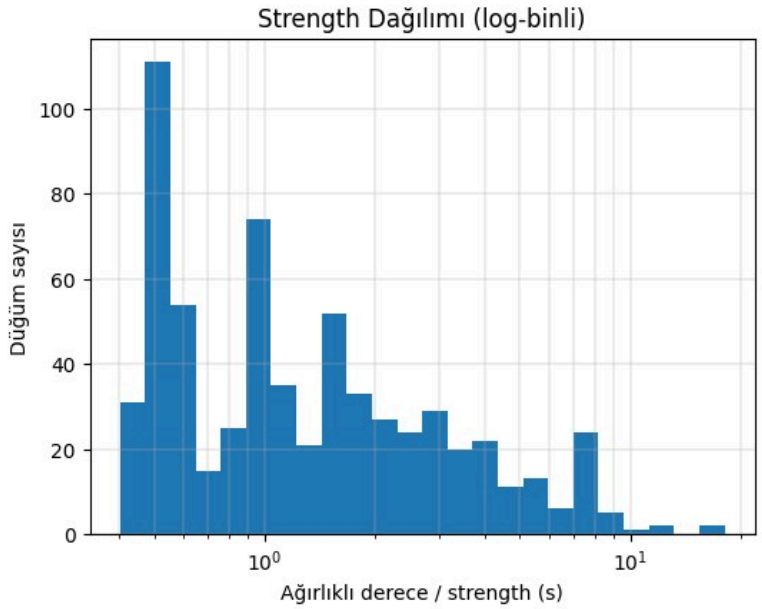
Figure 2: Distribution of Node Degree of the Patient Graph

The resulting graph, constructed using the methods described above, comprises 637 nodes and 1,884 edges. To assess the structural quality of the graph, the homophily coefficient was computed and found to be approximately 0.93. This high level of homophily indicates that patients with similar diagnostic profiles are strongly clustered within the network, providing evidence of meaningful structural coherence. Figure 3 shows an alternative degree distribution based on the weighted degree (s) of each node. While most of the distribution is concentrated at low s values, some nodes are connected via numerous and high-strength edges. This pattern supports the right-skewed structure observed in the unweighted degree distribution and reflects the quality of the connections. Edge weights were computed as the sum of the Adamic–Adar similarities.

A high s value indicates not only the number of connections but also the intensity of the interactions with the neighbors. Based on this distribution, strategies such as degree normalization or dropout were considered during model construction to mitigate the dominance of high-degree nodes.

Figure 3

Comparison of network degree centrality measures according to the predicted risk groups



- Model architecture and training procedure

In this study, a GCN was employed for the classification task on the constructed patient–patient graph. The model was used to learn patient representations based on diagnostic similarities among patients.

The model architecture comprises two successive graph convolution layers. These layers update the feature vector of each node by aggregating information from its neighbors, weighted by the edge strengths. Incorporating edge weights into the model allows the Adamic–Adar similarity measure between patients to influence the learning process. Patients connected via stronger links exert a greater impact on each 's representations.

A dropout rate of 0.5 was applied between the GCN layers to mitigate overfitting and improve generalization. (Overfitting was observed in earlier model iterations due to the limited size and narrow scope of the dataset.) Class weights were incorporated into the loss function to address potential class imbalance in the training data to increase the influence of minority-class samples during training. This adjustment was intended to encourage the model to learn all classes more equitably.

The nodes in the graph (a total of 637 patients) were partitioned into three disjoint subsets: training, validation, and test to objectively evaluate model performance. To ensure consistent class distribution across all subsets, stratified sampling was employed. Accordingly, the dataset was split into a training set with 445 samples, a validation set with 96 samples, and a test set with 96 samples. The validation set was used for hyperparameter tuning and model selection, whereas the test set was used for final performance metrics, which remained entirely unseen during the training and validation phases.

This study used a graph structure, representing patients and diagnoses as separate node types interconnected by edges that signify diagnostic events. This design acknowledges the clinical data's intrinsic heterogeneity and facilitates relational modeling through message passing among conceptually distinct entities. We chose this schema because it is a common way to represent hospital interactions in earlier works, such as Mao, Yao, and Luo (2022), who used patient–diagnosis bipartite graphs, and Lin, Kuo, Wang,

and Tseng (2025), who used temporal visit-diagnosis graphs to show sequential information. The graph structure explicitly models the hospital–patient–diagnosis triad, reflecting real-world EHR semantics and facilitating inductive reasoning for patients not encountered during training.

Explainability Framework

We employed two complementary explainability methods that represent fundamentally different paradigms in graph-level attribution to interpret the predictions of the GCN model. GNNExplainer provides an optimization-based mechanism that learns continuous masks over edges and node features to identify the substructure most responsible for a target node's prediction, maximizing the mutual information between the explanatory subgraph and the model output (Ying et al., 2019). However, optimization-driven masks may show start sensitivity and sometimes return fragmented structures that do not capture higher-order clinical motifs. SubgraphX, a game-theoretic explainability algorithm that uses Shapley values to look at separate candidate subgraphs, was added to solve these problems (Shapley, 1952; Yuan, Yu, Wang, Li, & Ji, 2021). SubgraphX uses Monte Carlo Tree Search (MCTS) to quickly search the combinatorial subgraph space and find small, high-fidelity motifs using a reward function that balances predictive contribution and parsimony (Yuan et al., 2021). These two methods work together to create a single explainability framework: GNNExplainer picks up fine-grained relational signals at the edge level, whereas SubgraphX finds stable, clinically coherent multi-node diagnostic motifs. This gives us a better picture of how the GNN makes decisions.

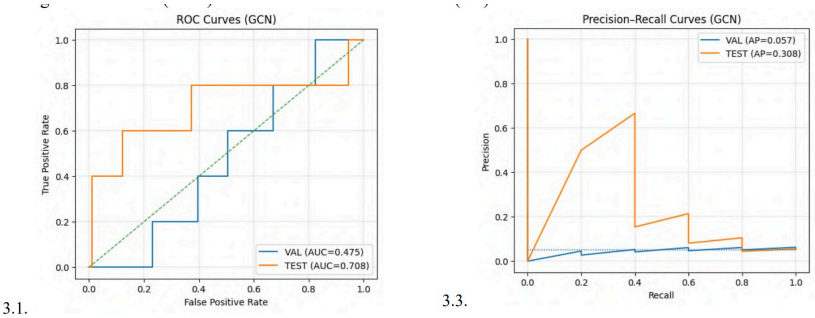
RESULTS

The training process was watched over 50 epochs using the training and validation sets. The model had a high predictive performance at the end of training, with an AUC of 0.815 and an AUPRC of 0.157 on the training set. The model performed much worse on the validation set (AUC = 0.475, AUPRC = 0.057), which suggests that it was overfitting, probably because the training data were too small.

The model was evaluated on a test set that had not been used during training or validation. It achieved an AUC of 0.708 and an AUPRC of 0.308, which can be considered meaningful given the model's structure and dataset characteristics. Standard evaluation metrics were employed to assess model performance, including the Receiver Operating Characteristic (ROC) curve and the Precision–Recall (PR) curve.

Figure 4

(a) Receiver operating characteristic curve and area under the curve value of the model; (b) precision-recall curve and AUPRC value



3.1. 3.2. (a) 3.3. 3.4. (b)

Figure 4-a displays the receiver operating characteristic (ROC) curve for the model evaluated on the test set, where an AUC of 0.708 was achieved. This score indicates that the proposed model possesses greater

discriminative power than random guessing. It also suggests that the representations learned via the patient–patient graph can effectively distinguish between high-risk and low-risk patients without overfitting to the training data.

Figure 4-b presents the PR curve, where the area under the curve (AUPRC) reaches 0.308. This score is approximately six times higher than a no-skill classifier's baseline performance (dashed line). Despite the low prevalence of the positive class (5.1%), this result demonstrates the strong ability of the model to identify rare cases. These findings suggst that the model holds promise as a useful and discriminative clinical tool when trained on a more comprehensive dataset.

Figure 5

Community Detection Based on Node Degree, Size, Colored by Community Membership, and GCN-Predicted Risk Probability

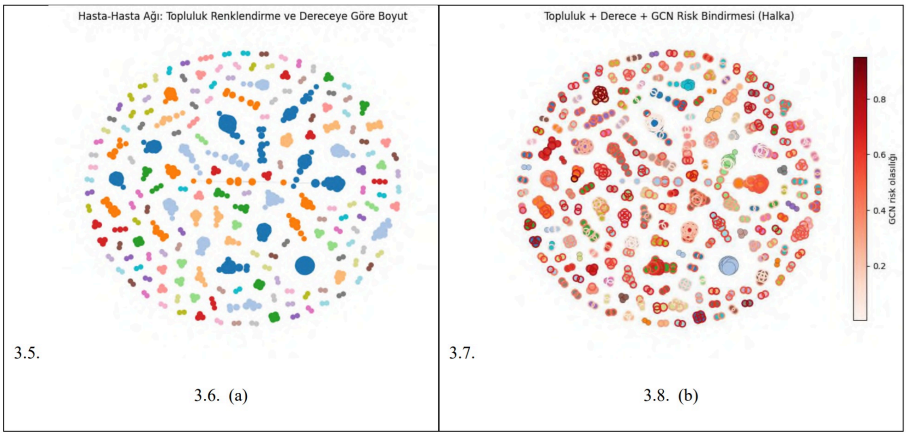
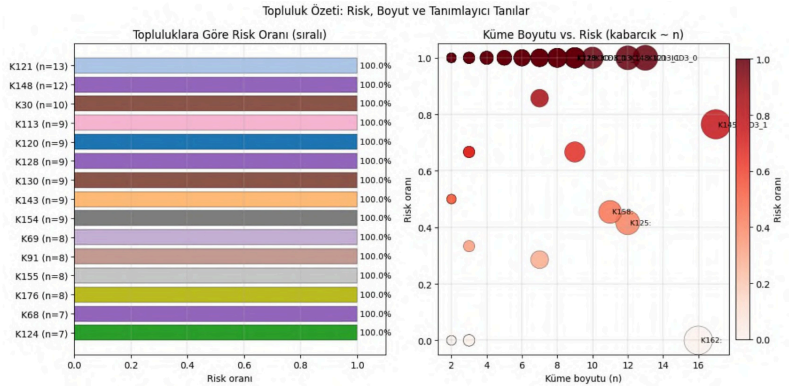


Figure 6 shows that high-risk contribution ($\hat{r}_c$) values are predominantly concentrated within small- to medium-sized groups, indicating that the risk signal emerges primarily within communities rather than across the global network. This suggests that local diagnostic homophily, rather than global centrality, is the key predictor of risk. These findings reinforce the interpretation presented in Figure 5b.

Using weighted edges s(W()), we obtained Louvain communities (C = 184, resolution = 1.0) from the patient graph. For each community, the risk ratio was calculated using the following equation:

Figure 6

GCN-Predicted Risk Distribution Across Detected Patient Communities



$\hat{r}_c = \frac{\#\{\text{pred\_high} \in c\}}{\lvert c \rvert}$	(1)

Node-level risk labels were generated using a threshold optimized to maximize $F_{\beta=2}$ m on the validation set. For each community, the Top 3 most representative ICD-9 codes (based on the first three digits) were identified according to their within-community patient share.

Figure 6 illustrates the presence of several communities with high-risk ratios, particularly those in the size range of (n$\geq$ 9–13). These high-risk communities cluster around specific three-digit ICD-9 code groups. The emergence of elevated risk is independent of community size, indicating that local diagnostic homophily, rather than broader structural or size-related factors, primarily drive high-risk scores.

Figure 7

Centrality distributions across risk groups

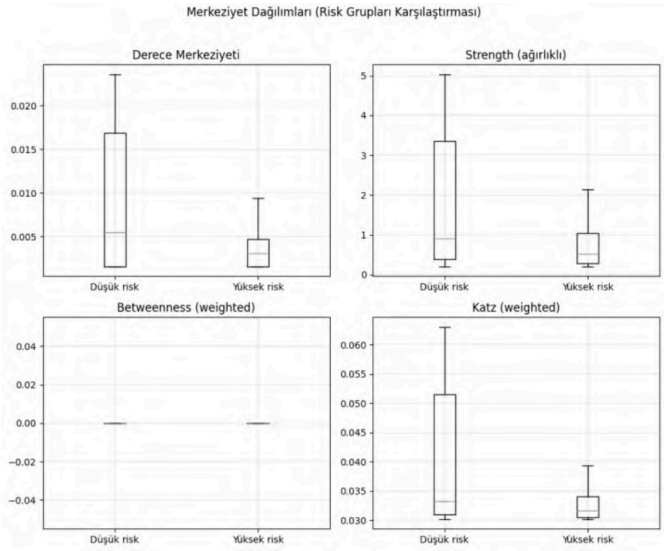


Figure 7 shows that high risk does not come from centrality but from diagnostic similarity and a shared hospital context, as shown by comparisons made using centrality measures.

The degree and weighted degree distributions are shown at the top, and the medians are noticeably lower in the high-risk group, indicating no tendency toward central nodes. Risk-associated nodes are located at the network's periphery and are embedded within low edge-density regions.

In terms of Katz centrality, the likelihood of being connected to important neighbors is also weak. Betweenness centrality is zero for both groups, which is expected given that the graph is highly fragmented and composed of small modules.

Explainable AI Results

Figure 8 shows a waterfall chart that illustrates how the risk score was derived for a single patient. The chart begins with the baseline risk level across the entire population. Then, it shows how the patient's own contributions, along with the influence of the most impactful neighbor nodes, increase or decrease this baseline—depicted in green (positive contribution) and red (negative contribution).

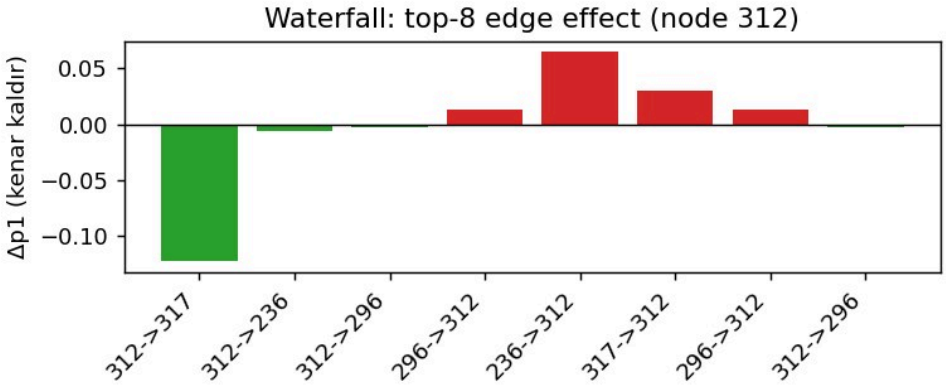
The sum of all positive and negative contributions results in the model's final predicted risk score for this patient is 0.86. The model incorporates both the baseline risk and the combined effects of node features and neighbors' relational signals. In this instance, a strong positive effect from the patient's own attributes,

coupled with mixed positive and negative influences from neighboring patients, yields a high overall risk score.

This example highlights graph-based models performing dynamic computations that integrate several interrelated factors, emphasizing the relational and distributed nature of decision-making in GNNs.

Figure 8

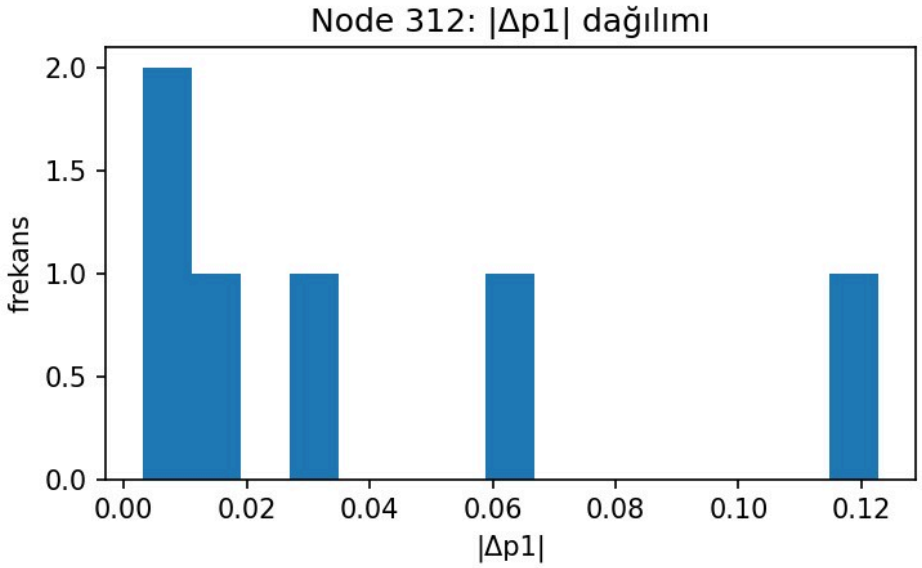
Explanation of the Risk Score for Node-312



To analyze the magnitude of each neighbor's influence on the model's prediction for the same patient, a histogram was plotted showing the absolute change in predicted probability ($|\Delta p1|$) resulting from the removal of each edge (Figure 9).

Figure 9

$Node\ 312\ |\Delta p1|$ *histogram*



The histogram exhibits a distribution that resembles a long-tail pattern. Most of the neighboring patients have a negligible or near-zero impact on the risk prediction of the target patient. However, the tall bars on the left side of the histogram confirm that some neighbors exert disproportionately large effects on the prediction of the model.

This finding supports the earlier ablation analysis results and provides insight into the decision-making mechanism of the model. Rather than using all neighboring nodes equally when determining a patient's risk, the model selectively focuses on a small set of diagnostically relevant and highly influential neighbors. This result indicates that the model is capable of filtering out noise within the graph and attending to the most informative signals.

Figure 10

$Node\ 312$ $|\Delta p1|$ *histogram.*

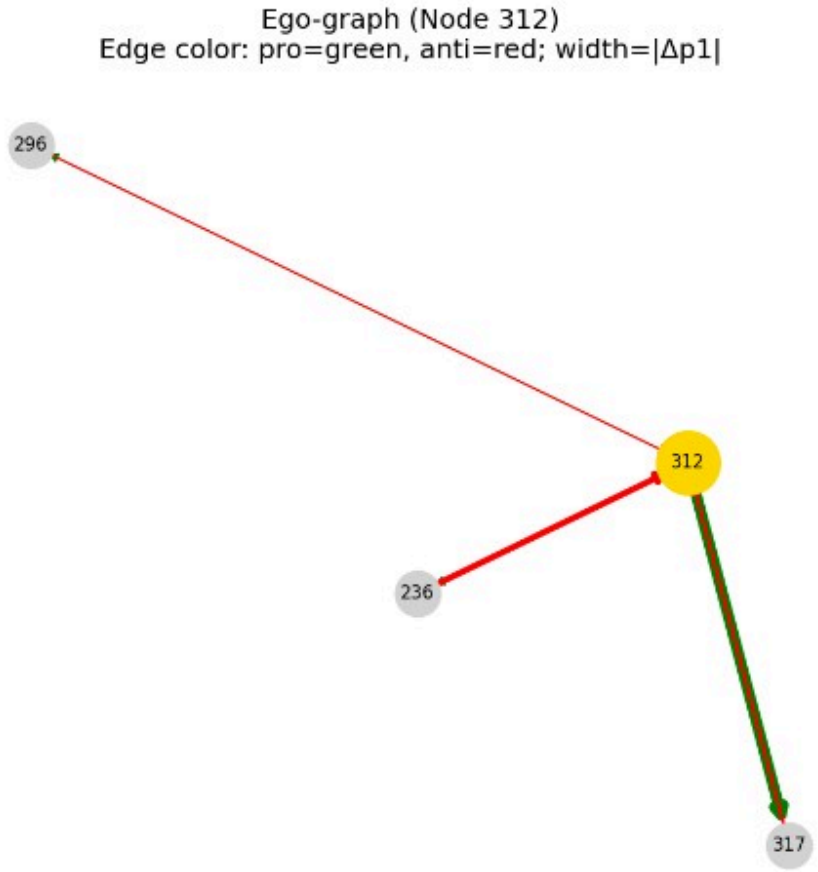


Figure 10 displays the local subgraph generated by GNNExplainer for node 312, which was predicted by the model to be high risk. The colors of the edges represent their effect on the model's prediction: red and blue edges indicate neighbor relationships that increase and decrease risk, respectively. Edge thickness is proportional to the GNNExplainer's importance scores.

The figure shows that the model's high-risk prediction is primarily driven by some strong neighbor relationships, represented by thick red edges. The decision is not based on a single connection but instead arises from multiple neighbors with diagnostic similarity and elevated risk. The overall assessment reflects the cumulative effect of these interactions. Additionally, edges with negative (risk-reducing) influence have lower weights than those with positive (risk-increasing) influence.

Figure 11
Ablation analysis results.

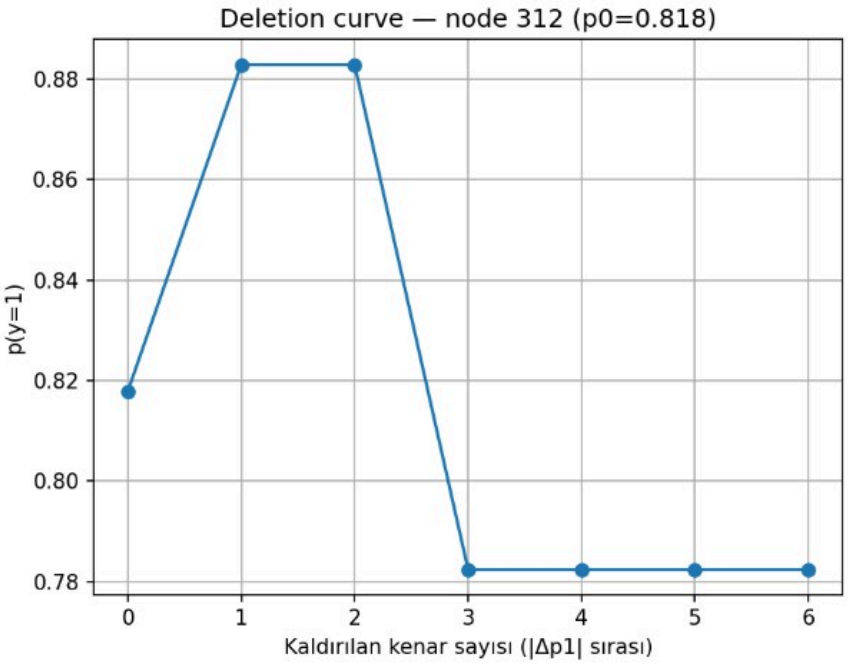


Figure 11 presents the ablation analysis, which measures the causal impact of the neighbor relationships identified by GNNExplainer as most important to the model's prediction. Through ablation, it becomes possible to examine how the probability of positive prediction changes as the most influential connections are removed from the graph.

For node 312, removing approximately six key neighbors is sufficient to reduce the high-risk prediction. These explanations enhance the trustworthiness of the model by providing transparency into how specific relational factors influence the outcome.

This analysis causally substantiates the following two critical findings:

The model's prediction is not equally dependent on all neighbors but instead focuses on some nodes that are diagnostically relevant and influential.

The model's ability to consistently exhibit this behavior across different patients strengthens confidence in the decision mechanism's reliability.

SubgraphX

To complement the edge-level explanations obtained from GNNExplainer, we applied SubgraphX-lite to quantify the causal contribution of local graph motifs to the prediction for node 312 in the same high-risk patient. SubgraphX-lite generates a discrete set of candidate subgraphs around the target node through an MCTS-based search procedure and assigns each motif a Shapley-inspired Δp score, representing the causal change in predicted mortality probability when that motif is removed from the computational graph.

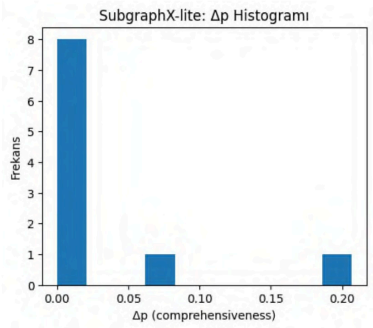
Across K=10 candidate motifs, the Δp distribution for node 312 was strongly left-skewed (Figure 12a). Most motifs exhibited a near-zero influence, with a median Δp = 0.001 and an interquartile range of 0.000–0.011.

Only a small subset of motifs produced substantial causal shifts—Δp ≥ 0.10 in 10% of motifs and Δp ≥ 0.20 in 10%. This pattern confirms that the prediction of the GCN model is driven not by the entire neighborhood but by a compact set of diagnostically coherent substructures.

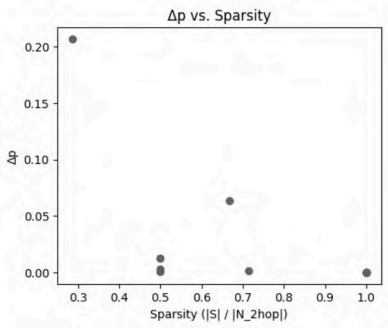
Motif sparsity further supported this finding (Figure 12b). The extracted motifs were generally compact, with a median sparsity of 0.690, and Δp showed a strong negative correlation with sparsity (Pearson = −0.61; Spearman = −0.87). Thus, denser motifs corresponded to larger causal effects, indicating that clinically meaningful similarity clusters—not isolated edges—account for the model's risk estimation's major drivers.

Figure 12

Subgraph X-lite explanation results.



(a) Histogram of Δp causal effects across K = 10 motifs, illustrating the long-tailed distribution in which only few motifs exert substantial influence.



(b) Relationship between Δp and sparsity for the same motifs, showing that denser motifs exhibit stronger causal impact.

Comparison: GNNExplainer vs. SubgraphX

These SubgraphX-lite findings complement and extend the earlier GNNExplainer analysis. Whereas GNNExplainer highlights individual influential neighbors using continuous edge-importance masks, SubgraphX-lite identifies multi-node diagnostic motifs that drive higher-order relational effects. The motifs with the highest Δp values contain the same influential neighbors identified by GNNExplainer, demonstrating strong methodological consistency.

Together, the two explainability approaches provide a coherent picture of the decision mechanism of the model:

- Only a small subset of neighbors drives the prediction
- These neighbors form dense and clinically interpretable subgraphs,
- The relational reasoning of the model is stable, selective, and clinically plausible.

SubgraphX-lite thus offers a causally grounded and structurally coherent complement to GNNExplainer, yielding a deeper understanding of how the GCN integrates local patient similarity patterns to generate its mortality prediction.

During our evaluations, SubgraphX outperformed GNNExplainer in generating clinically plausible and concise explanations. GNNExplainer occasionally produced subgraphs that were disconnected or lacked clarity, but SubgraphX's Monte Carlo Tree Search identified compact yet cohesive node sets. When analyzed quantitatively, SubgraphX exhibited superior fidelity ratings and reduced explanation sizes. However, GNNExplainer demonstrated superior speed and adaptability for data cluster analysis. We recommend the use of both GNNExplainer for preliminary analysis and SubgraphX for comprehensive insight.

DISCUSSION AND CONCLUSIONS

Artificial intelligence (AI) is now used in virtually every domain—from product recommendations on platforms like Netflix and Amazon to personalized advertising on Google. However, when it comes to disease diagnosis, knowing the reasons behind critical decisions is vital. Entrusting such decisions to an AI system that cannot explain its reasoning may lead to dangerous consequences.

Explainability is necessary not only for justifying decisions but also for preventing things from going wrong. It plays a central role in ensuring transparency, safety, and trustworthiness in high-stakes healthcare applications.

In healthcare, AI models cannot achieve high predictive accuracy alone. In real clinical scenarios, understanding why the model made a particular decision, which clinical variables it relied upon, and whether these decisions are interpretable by physicians is equally important. Interpretability directly affects the usability, reliability, and clinical acceptance of AI systems in practice.

A GCN model was developed on a patient–patient graph constructed from the eICU-demo dataset, enabling prediction over ICU patient networks. The model can perform binary classification with an AUC of 0.708 and an AUPRC of 0.308. Using XAI methods, the model's predictive success relies not only on individual patient features but also on the structural properties of the graph, highlighting the importance of relational context in clinical outcome prediction.

Centrality analysis revealed that patients predicted to be at high risk occupy more central positions within the network. Network visualization and community detection further showed that these central, high-risk patients tend to cluster around specific clinical profiles. Patient-level analyses performed using GNNExplainer confirmed this mechanism, demonstrating that the patient's features and strong signals from a small number of critical neighbors combine to drive the model's decision.

Beyond GNNExplainer, the integration of SubgraphX provided a higher-order understanding of the relational structures driving the model's predictions. While GNNExplainer highlighted a small set of influential neighbors, SubgraphX revealed that these neighbors form compact, clinically coherent diagnostic motifs with a strong causal impact. The presence of dense, high-Δp subgraphs around high-risk patients suggests that mortality predictions emerge not only from individual edges but also from clusters of multi-node clinical similarity. This finding strengthens the model's interpretability by demonstrating consistency across two fundamentally different XAI paradigms—optimization-based (GNNExplainer) and game-theoretic (SubgraphX). Together, these findings demonstrate that the decision mechanism of the GCN is stable, localized, and grounded in clinically meaningful structures, which is an essential property for real-world adoption in critical care settings.

Centrality analysis showed that patients predicted to be at high risk tended to occupy more central roles within the network. Network visualization and community analysis further revealed that these central, high-risk patients cluster around specific clinical profiles. Single-patient analyses using GNNExplainer confirmed this mechanism, demonstrating that the model's decision results from a combination of the patient's own features and strong signals from a few critical neighbors.

The findings of this study support previous research conducted on the eICU dataset, which has demonstrated the potential of diagnosis-based patient graphs in clinical prediction tasks. (Pollard et al., 2018). However, the objective of this study was not only to develop a prediction model but also to generate explanations for those predictions by integrating XAI techniques such as GNNExplainer, ablation analysis,

and community characterization. This approach directly addresses the critiques of Tjoa and Guan (2021), who emphasized that XAI applications in healthcare must be clinically meaningful and verifiable (Tjoa & Guan, 2021).

The primary limitation of this study is the use of the eICU-demo dataset, which is a small and imbalanced subset of the full eICU-CRD. This may affect the stability of evaluation metrics and limit the ability to draw definitive conclusions regarding model performance. Second, the patient graph was constructed solely based on three-digit ICD-9 codes. Although the full dataset contains 31 interlinked tables with rich clinical information, additional modalities such as medications, procedures, and other temporal or contextual features were not incorporated into the graph construction. Finally, the study was limited to a single GNN architecture (GCN). The results may be further enriched by comparison with more advanced architectures, such as GAT or GraphSAGE, which could offer improved performance or interpretability.

For future work, the same framework should be tested on the full eICU-CRD or MIMIC datasets. Additionally, a heterogeneous graph structure incorporating multiple relation types—such as diagnoses, medications, and procedures—should be constructed, and predictive performance should be compared across more advanced graph neural network architectures. Furthermore, other XAI tools applicable to graph-based models could be integrated to produce richer and more comprehensive explanations. Studies should also be conducted through expert review by clinicians to evaluate the clinical validity of the generated explanations. Such extensions are essential for real-world clinical integration of GNN and XAI methodologies.

In addition to extending the presented framework to larger and more heterogeneous datasets, advanced graph-specific XAI methods that provide complementary perspectives to SubgraphX could be explored in future work. For example, PGExplainer offers a probabilistic modeling approach that learns a parametric distribution over explanatory subgraphs and may yield more stable motif discovery across patients. XGNN, a generative explanation model, can synthesize prototypical graph structures that maximize model confidence, potentially enabling clinicians to visualize “idealized” diagnostic patterns learned by the GNN. GraphLIME provides node-level feature attributions through local linear approximations and can be used to disentangle the contributions of clinical variables from relational signals. Integrating these methods would allow the construction of a multi-perspective, clinically grounded explanation ecosystem, strengthening both the interpretability and robustness of graph-based models in real-world critical care applications.

Peer Review Externally peer-reviewed.
Conflict of Interest The author have no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.
Acknowledgment A preliminary version of this work, focusing solely on GNNExplainer-based patient-level interpretations, was presented at the IMISC 2025 Conference (October 2025). The present manuscript substantially extends that initial study by incorporating a dual-paradigm explainability framework, introducing SubgraphX for motif-level causal analysis, and providing a comprehensive integration of edge-level and subgraph-level reasoning across the PSN.

Author Details Şebnem Akal
[1] Istanbul University, Faculty of Science, İstanbul, Türkiye
0000-0001-8239-2957 : sebnem.akal@istanbul.edu.tr

References

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. doi: 10.1214/SS/1009213726

Buchanan, B. G., & Shortliffe, E. H. (Eds.). (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project* (Addison Wesley, Rea...). Retrieved from https://www.shortliffe.net/Buchanan-Shortliffe-1984/MYCIN%20Book.htm

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting the risk of pneumonia and 30-day hospital readmission !emph[Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining];, *2015-August*, 1721–1730. https://doi.org/10.1145/2783258.2788613/SUPPL_FILE/P1721.MP4

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, *101*(23). https://doi.org/10.1161/01.CIR.101.23.E215

Lin, K.-W., Kuo, Y.-C., Wang, H.-Y., & Tseng, Y.-J. (2025). *KAT-GNN: A Knowledge-Augmented Temporal Graph Neural Network for Risk Prediction in Electronic Health Record*s Retrieved from https://arxiv.org/pdf/2511.01249

Lu, H., & Uddin, S. (2021). A weighted patient network-based framework for predicting chronic diseases using GNs *Scientific Reports*, *11*(1), 22607. DOI: 10.1038/S41598-021-01964-2

Lundberg, S. M., Erion, G., Chen, H., Degrave, A., Prutkin, J. M., Nair, B., ... Lee, S.-I. (n.d.). *From Local Explanations to Global Understanding with Explainable Tree AI* https://github.com/suinleelab/treeexplainer-study

Mao, C., Yao, L., & Luo, Y. (2022). MedGCN: Medication recommendation and lab test imputation via graph convolutional networks *Journal of Biomedical Informatics*, *127*, 104000. DOI: 10.1016/J.JBI.2022.104000

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU collaborative research database is a freely available multi-center database for critical care research. *Scientific Data*, *5*, https://doi.org/10.1038/SDATA.2018.178,

Sağıroğlu, Ş., & Demirezen, M. U. (2022). *Yapay Zekâ ve Büyük Veri Kitap Serisi 4: Yorumlanabilir ve Açıklanabilir Yapay Zekâ ve Güncel Konular*.

Sahu, M. K., & Roy, P. (2025). *Similarity-Based Self-Construct Graph Model for Predicting Patient Criticalness Using Graph Neural Networks and Electronic Health Record Data* Retrieved from https://arxiv.org/pdf/2508.00615

Shapley, L. S. (1952). Value for n-person games *The Shapley Value*, (28), 307–317. https://doi.org/10.1017/CBO9780511528446.003

Tjoa, E., & Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Tong, C., Rocheteau, E., Veličković, P., Lane, N., & Liò, P., 2021. Predicting Patient Outcomes Using Graph Representation Learning *Studies in Computational Intelligence*, *1013*, 281–293. DOI: 10.1007/978-3-030-93080-6_20

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature*, *621*(7980), 672–675. https://doi.org/10.1038/D41586-023-02980-0https://doi.org/10.1038/D41586-023-02980-0

Watson, D. S. (2022). Statistics of Interpretable Machine Learning *Digital Ethics Lab Yearbook*, 133–155. doi: 10.1007/978-3-031-09846-8_10

Ying R, Bourgeois D, You J, Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, *32*. Retrieved from https://arxiv.org/pdf/1903.03894

The yuan H, Yu H, Wang J, Li, K., & Ji, S. (2021). *Explainability of Graph Neural Networks via Subgraph Exploration* The yuan H, Yu H, Wang J, Li, K., & Ji, S. (2021). *Explainability of Graph Neural Networks via Subgraph Exploration*