Graph–Neurosymbolic Neural Networks for Trustworthy Clinical Decision Support

Oluwatobi Noah Akande akande.oluwatobi@gmail.com
Computer Science Department
Nile University of Nigeria
Abuja, Nigeria

Amit Mishra i.amitmishra@gmail.com
Department of Computer Science and Applications
Dr. Vishwanath Karad, MIT World Peace University
Pune, India

Abidemi Emmanuel Adeniyi aeadeniyi@oauife.edu.ng
Department of Information Systems
Obafemi Awolowo University
Ile-Ife, Nigeria

Qurrat Ul Ain Mughal q.a.mughal@salford.ac.uk
Data Science Department
University of Salford
United Kingdom

Dagogo Godwin Orifama d.g.orifama1@salford.ac.uk
School of Science, Engineering & Environment
University of Salford
United Kingdom

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

The increasing adoption of artificial intelligence in clinical decision support systems has been constrained by persistent concerns regarding safety, interpretability, and regulatory trustworthiness. While graph neural networks (GNNs) have shown strong predictive performance by modeling complex relationships in electronic health records, they largely operate as unconstrained statistical models that may produce clinically unsafe or logically inconsistent recommendations. This study proposes a novel Graph–Neurosymbolic Framework for Trustworthy Clinical Decision Support that tightly integrates graph-based neural learning with formal symbolic medical reasoning. In the proposed framework, patients and clinical entities are represented as a heterogeneous clinical graph, while domain knowledge encoded as first-order logic and Horn-clause rules explicitly constrains graph construction, neural message passing, and inference. The novelty of this study lies in embedding symbolic governance directly into the learning process rather than applying post-hoc rule checking or explainability. This design enables formal safety enforcement, logical consistency, and the generation of rule-driven counterfactual explanations that are clinically plausible and auditable. The framework is evaluated using two large publicly available electronic

health record datasets, MIMIC-IV and eICU-CRD, representing both single-center and multi-institutional clinical settings. Experimental results demonstrate that the proposed graph–neurosymbolic approach achieves competitive predictive performance while substantially reducing clinical rule violations compared to neural-only baselines. In addition, the framework produces high-fidelity and rule-consistent explanations and exhibits improved robustness under distribution shift across institutions. These findings highlight the effectiveness of combining relational learning with symbolic medical knowledge and establish the proposed framework as a strong step toward trustworthy, deployable clinical decision support systems.

Keywords: Graph Neural Networks, Neurosymbolic Artificial Intelligence, Trustworthy Clinical Decision Support, Explainable and Safe AI in Healthcare, Electronic Health Records.

1. Introduction

Accurate and reliable CDSS supports are crucial to enhance diagnostic accuracy, standardize care delivery, and prevent avoidable medical errors in the present day health systems. Recent years have seen a success of deep learning and graph-based machine learning methods in what concerns prediction tasks relative to a wide variety of medical data, by capturing intricate non-Euclidean relationships between patients, symptoms, laboratory results and treatments. Among these studies, Graph Neural Networks (GNNs) have been prevalently employed in patient similarity modeling, disease prediction and multimodal clinical data fusion. Most healthcare GNN models exhibit poor interpretability weak generalization and cannot reinforce medical-specific domain rules and safety constraints which makes it difficult for their deployment in the clinics (Paul et al., 2024; Vaida and Huang, 2025). Neurosymbolic Artificial Intelligence (NeSy), which integrates neural learning with symbolic reasoning, has become a promising framework in order to tackle the shortcomings of end-to-end data-driven models. Through combining statistical learning with explicit symbolic knowledge (e.g., ontologies, logic rules and clinical guidelines), neurosymbolic approaches make it possible for models to reason in a transparent way and make predictions consistent with established domain-level knowledge hence ensuring verifiability. Recent work underscores that neurosymbolic AI is especially versatile in high-stakes domains like healthcare where interpretability and regulation requirements are of the essence (Nawaz et al., 2025; Prenosil et al., 2025).

Although GNNs and neurosymbolic reasoning are conceptually aligned with each other, there has been a paucity in their application to clinical decision support. Consequently, most of the GNNs-based healthcare systems usually use unconstrained message passing mechanisms, which can only learn latent representations from data without incorporating medical logic such as contraindications, lab value thresholds or diagnostic exclusion rules. Consequently, these models could generate predictions that are not inherently reasonable or clinically dangerous, despite their overall accuracy reflecting a favorable metric. Existing works emphasized the requirement for logic guided neural architectures and constraint-aware learning objectives to address this risk, but few approaches have realized these concepts in practice in a clinical graph structured domain (Mondal et al., 2025).

Other than safety and consistency, the interpretability is still a major obstacle to AI-based CDSS deployment. Post-hoc explanation methods developed to be easily applied to deep

neural networks, e.g. feature attribution or saliency maps, are often not clinically relevant and difficult for clinical practitioners to interpret. Counterfactual explanations which report smallest and possible modifications to a patient's characteristics required to change a prediction made by the model have been useful in medical domains. More recent work has shown that symbolic, rule-based medical counterfactual explanations also lead to higher understanding and trust among clinicians in comparison with gradient-based ones (Guidotti, 2024; Mertes et al., 2022; Raj and Mileo, 2025).

This study introduces a trustworthy clinical decision support with explicit integration of graph-based learning and symbolic medical reasoning, named as Graph–Neurosymbolic Neural Network. Patients, clinical entities (e.g. symptoms, lab results and medications) as well as formal medical guidelines are represented in a contrived dynamic clinical graph. Neural message passing is guided by symbolic rules via logic-based learning objectives, and a symbolic reasoning module serves to verify predictions' consistency and safety. The framework also facilitates the extraction of rule-based counterfactual explanations which are clinically valid and conformable to domain expert knowledge.

This study has four main contributions: a) we present a knowledge-based GNN structure through which clinical rules and ontological relations map to message passing directly. b) we formalize safety and logical consistency guarantees with respect to the neural predictions through differentiable logic constraints and symbolic verification. c) we introduce a symbolic rule-based counterfactual explanation agents specialized for graph-shaped clinical data. d) we empirically evaluate our approach and show that it performs better than the neural-only and rule-only baselines at making trustworthy, safe, interpretable predictions without a cost to the predictive performance (Paul et al., 2024; Nawaz et al., 2025).

2. Related Works

The recent success of AI for healthcare is largely powered by the accessibility to large-scale Electronic Health Records (EHRs), biomedical databases as well as advanced representation learning techniques. An overview of related works in three categories is presented in this research:

2.1. Graph Neural Networks in Healthcare

Graph Neural Networks (GNNs) provide a natural formalism for modeling relationships in clinical data which links patients to diagnoses, labs, medications, notes, and imaging so that relational and structural patterns can improve prediction and reasoning. Recent work has explored patient-similarity graphs, knowledge graphs built from EHRs and biomedical ontologies, and multimodal fusion with imaging/genomics. Despite strong predictive gains, prominent limitations remain: model interpretability, rule-consistency with clinical guidelines, robustness under distribution shift, and explicit mechanisms for encoding safety constraints into message passing (Paul et al., 2024; Vaida and Huang, 2025).

Vaida and Huang (2025) conducted a survey on GNN approaches for combining EHRs, imaging, genomics, and clinical notes. They categorized fusion strategies (early, intermediate, late fusion) and highlights attention-based and heterogeneous GNNs for patient representation. The authors emphasize practical challenges for deployment which are data heterogeneity, missingness, and the need for domain constraints. They finally called for

methods that incorporate symbolic medical knowledge for safety. Similarly, Gupta et al. (2025) proposes a hierarchical-attention GNN (HAIL) that constructs multi-level patient similarity graphs from EHR segments and applies attention at each level to reduce noise. The study achieved an improved heart-failure and readmission prediction on multi-site EHRs and shows that attention weights can surface clinically relevant neighbors though it does not enforce formal clinical rules. Mienye and Viriri (2025) reviewed GNN uses in imaging task. Their study documented the opportunity to combine imaging graphs with patient EHR graphs for richer clinical models but warns that clinical safety checks are rarely integrated.

Recent GNN work in healthcare has showed clear predictive potential and improved neighborhood-based interpretability, but there is a consistent unmet need for architectures that can formally encode clinical rules, validate predictions against domain constraints, and provide explanations that reflect medical semantics rather than only learned attention weights.

2.2. Neurosymbolic AI in Medical Reasoning

Neurosymbolic AI (NeSy) combines learned neural representations with explicit symbolic knowledge and reasoning. The aim is to retain the adaptability of deep models while introducing interpretability, verifiability, and rule-governed behavior. In medicine, NeSy approaches promise to enforce clinical guidelines, maintain logical consistency across predictions, and produce auditable reasoning properties that are critical for safety-critical clinical decision support. Key directions include differentiable logic layers, logic-guided losses, integration with ontologies, and pipelines that couple large pre-trained models with rule-based verifiers (Nawaz et al., 2025). Prenosil et al. (2025) introduced a production-oriented pipeline that connects large language models to a rule-based expert system to extract and verify clinical facts from free text, thereby, producing traceable labels. The system reduces clinically implausible outputs by using symbolic checks and demonstrates improved auditability for downstream CDSS training. Mondal et al. (2025) introduced a differentiable logic approach where logical constraints over clinical predicates are embedded as soft penalties during neural training. Application to a toy clinical dataset shows fewer rule violations with modest impact on accuracy thereby supporting the feasibility of soft rule enforcement in clinical ML. A recent survey on neuro-symbolic methods carried out by Nawaz et al. (2025) proposes benchmark tasks to evaluate faithfulness, provability of constraints, and scalability of NeSy systems in regulated domains such as healthcare. It calls for standardized protocols to test rule-enforcement and for datasets with encoded clinical rules Neurosymbolic methods show substantial promise in providing auditability and rule enforcement. However, scaling symbolic inference to large graph-structured clinical representations and marrying it tightly with GNN message passing so that symbolic constraints actively modulate graph propagation remain open problems. Therefore, there is a need for standardized benchmarks that measure rule compliance, safety, and explanation fidelity in clinical settings.

2.3. Explainable and Trustworthy AI in Clinical Settings

Explainability, safety, fairness, and robustness are core non-functional requirements for clinical AI. Empirical studies show that explanations can both increase and decrease clinician

trust depending on quality and context, underscoring the need for human-centered evaluation (Lekadir et al., 2025; Rosenbacke et al., 2024). For instance, Lekadir et al. (2025) provided a consensus recommendations and minimum reporting standards for trustworthy AI in healthcare covering transparency, validation, uncertainty quantification, and dataset documentation. Sadeghi et al. (2024) classified XAI methods and assesses how well they address clinical needs. They highlighted that feature attribution techniques often fail to yield clinically meaningful rationales and recommends counterfactuals and concept-level explanations for higher utility. Rosenbacke et al. (2024) carried out a meta-analysis of clinician studies that showed a mixed effects of explanations on trust. They revealed that some studies show increased reliance after seeing explanations while others show confusion or decreased trust when explanations are misleading. The paper stresses task-matched explanation design and rigorously measured human outcomes. Nicolson et al. (2025) presented empirical evidence on how explanation type and framing affect clinician decision behavior and suggests human-centered protocols for evaluating XAI. They recommended iterative co-design with clinicians to produce clinically actionable explanations. The literature reviewed converges on two needs: a) explanations must be clinically meaningful and evaluated with real end-users using task-appropriate metrics; b) trustworthy AI requires not only explanations but formal safety and compliance guarantees. Thereby, integrating symbolic rule checks with explanation mechanisms so that explanations themselves are constrained by clinical rules remains an important opportunity and that is what this study intends to achieve.

3. Methodology

3.1. Problem Formulation

This section formulates the clinical decision problem that we consider in this work, and introduces the graph-based and logical representations of Graph-Neurosymbolic framework as well as its trustworthiness requirement.

3.1.1. Clinical Decision Task Definition

The main goal of the presented work is to help making the right decision by means of predicting and rule consistent reasoning on structured longitudinal patient data. Three highly-related clinical tasks are intended to be covered by the framework:

(a). Diagnosis support: the system forecasts how probable one or more diagnostic appointments are, based on observed patient data;

(b). Risk prediction, where patients are categorized into risk levels (e.g., low, medium, high) for an adverse event such as progression of disease or hospitalization.

(c). Treatment Support: The system can provide treatment recommendations or offer validation of treatments proposed by clinicians, for instance taking into account relevant clinical guidelines and contraindications.

In a formal sense, in the context of a patient's historical and current clinical data, it is required to learn a function that produces clinically meaningful predictions, consistent with domain specific medical knowledge and within safety constraints.

3.1.2. Data Sources and Clinical Variables

The proposed model runs on multi-entity clinical data extracted from Electronic Health Records (EHRs) and curated medical knowledge sources. Let the clinical dataset have the following components:

(a). Patients: demographic characteristics and longitudinal information about encounters;

(b). Clinical symptoms: physical examination and patient's report of symptoms;

(c). Laboratory results: Date-stamped number or description of a test result and the reference range thereof;

(d). Diagnoses: coded clinical manifestations (e.g., ICD-based diagnoses) over time;

(e). Drugs and Interventions: medications and treatments prescribed, with dose, time.

These time-stamped heterogeneous data may be incomplete, noisy and asynchronously observed. Further, external medical knowledge resources like clinical guidelines, dr ug-drug interaction rules and diagnostic criteria are imbedded to steer reasoning and validation.

3.1.3. Graph Representation of Clinical Knowledge

The clinical context is modeled as a dynamic heterogeneous graph to integrate patient data and medical knowledge.
Its nodes () consist of:

(a). Patient nodes, representing individual patients;

(b). Clinical entity nodes, including symptoms, laboratory tests, diagnoses and medications;

(c). Rule nodes: Corresponds to symbolic medical rules: contraindications, diagnostic criteria, safety constraints.

Its edges () represent interactions between nodes and are defined as:

(a). Temporal edges, representing the temporal flow of a patient's clinical events;

(b). Causal edges, corresponding to medically credible cause-effect relationships (e.g. symptom-diagnosis, medication-adverse event);

(c). Relational edges, e.g. between patients and symptoms and diagnosis and treatment as well as rules and entities.

This formulation allows for the passing of information between clinically relevant associations via neural message passing in a phrase, while symbolic rule nodes anchor domain constraints during inference.

3.1.4. Formal Problem Statement

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ denote a set of patients and let G_p represent the induced subgraph describing patient p's clinical state over time. The goal is to learn a prediction function:

$$f_\theta : G_p \to Y_p$$

where G_p is the patient-specific graph, Y_p denotes the clinical decision output (diagnosis label, risk score, or treatment recommendation), and θ represents learnable neural parameters.

Unlike conventional graph learning, f_θ is subject to a set of symbolic constraints $\mathcal{R} = \{r_1, r_2, \ldots, r_K\}$, where each rule r_k encodes medical knowledge expressed in logical form. The learning objective therefore jointly optimizes predictive accuracy and rule consistency:

$$\min_\theta \mathcal{L}_{task}(Y_p, \hat{Y}_p) + \lambda \mathcal{L}_{rules}(\hat{Y}_p, \mathcal{R})$$

where $\mathcal{L}_{task}$ measures prediction error and $\mathcal{L}_{rules}$ penalizes violations of medical rules.

3.2. Formal Rule Representation and Logic Constraints

Clinical domain knowledge is represented as a collection of symbolic rules in First-Order Logic (FOL) and constrained to Horn clause form for efficient inference and compatibility with differentiable reasoning frameworks.

3.2.1. First-Order Logic Foundations

Let $\mathcal{E}$ be the collection of clinical entities, either patient-related (patient, symptom), laboratory testing and examinations (lab test), diagnosis or medication. Then $\mathcal{L} = (\mathcal{C}, \mathcal{P}, \mathcal{V})$ is a first-order logical language with:

(a) $\mathcal{C}$ as a list of constants for internally visualized clinical entities (e.g. drug, diagnosis);

(b) $\mathcal{P}$ is a set of predicates expressing clinical relationships;

(c) $\mathcal{V}$ is a set of logical variables ranging over $\mathcal{E}$. Common predicates include: $HasSymptom(p, s)$, $LabValue(p, l, v)$, $DiagnosedWith(p, d)$, $Prescribed(p, m)$, $Contraindicated(m, d)$
Therefore, a FOL atom is a predicate applied to constants or variables, while **literals** may be positive or negated atoms.

3.2.2. Horn Clause Representation of Clinical Rules

To ensure tractable inference, clinical rules are represented as **Horn clauses**, which take the general form:

$$\forall \mathbf{x}\ (A_1(\mathbf{x}) \wedge A_2(\mathbf{x}) \wedge \cdots \wedge A_n(\mathbf{x}) \Rightarrow B(\mathbf{x}))$$

where: A_i are antecedent (body) atoms encoding observed clinical conditions and B is a single consequent (head) atom representing a clinical conclusion or constraint. Examples of clinical Horn clauses include:

(a). Diagnostic rule:

$$HasSymptom(p, s_1) \wedge LabValue(p, l, v) \wedge v > \tau \Rightarrow DiagnosedWith(p, d)$$

(b). Treatment safety rule:

$$DiagnosedWith(p, d) \wedge Contraindicated(m, d) \Rightarrow \neg Prescribed(p, m)$$

(c). Risk stratification rule:

$$DiagnosedWith(p, d) \wedge Age(p) > \alpha \Rightarrow HighRisk(p)$$

These clauses define necessary conditions that predictions must satisfy to be considered clinically valid.

3.2.3. Soft and Hard Constraint Modeling

We used soft and hard constraints to enforce the rule. For instance: a soft constraint could be instances where we allow limited violations but impose penalties during training. Hard constraints are instances where we strictly enforce the rules at inference time and invalidate unsafe predictions. This can be mathematically denoted as:
Let $r_k \in \mathcal{R}$ denote a rule and $\phi_k(\hat{Y}_p) \in [0, 1]$ denote its satisfaction degree. The rule-consistency loss is defined as:

$$\mathcal{L}_{rules} = \sum_{k=1}^{K} \max(0, 1 - \phi_k(\hat{Y}_p))$$

where $\phi_k = 1$ indicates full compliance and $\phi_k = 0$ denotes maximal violation.

3.2.4. Logic-Guided Message Passing

Let $h_v^{(l)}$ denote the embedding of node v at layer l. Neural message passing is modified to incorporate rule satisfaction such that:

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \alpha_{uv}^{(l)} \cdot h_u^{(l)} \cdot w_{r(u,v)}\right)$$

where: $\alpha_{uv}^{(l)}$ are attention coefficients; $w_{r(u,v)}$ is a rule-dependent weighting factor derived from symbolic constraints; σ is a non-linear activation. This formulation allows symbolic rules to modulate information flow across the graph.

3.2.5. Formal Safety Guarantees

A prediction $\hat{Y}_p$ is deemed **safe** if it satisfies all hard constraints:

$$\forall r_k \in \mathcal{R}_{hard},\ \phi_k(\hat{Y}_p) = 1$$

The system rejects or flags any prediction for which:

$$\exists r_k \in \mathcal{R}_{hard} such that \phi_k(\hat{Y}_p) < 1$$

This provides a formal guarantee that unsafe clinical actions (e.g., contraindicated treatments) are prevented by design.

3.2.6. Counterfactual Reasoning Under Logical Constraints

Given an observed instance G_p and prediction $\hat{Y}_p$, a counterfactual explanation seeks a minimally perturbed graph G'_p such that:

$$f_\theta(G'_p) \neq \hat{Y}_p \quad and \quad \forall r_k \in \mathcal{R},\ \phi_k(G'_p) = 1$$

This ensures that generated counterfactuals are clinically plausible and logically consistent.

3.3. Graph–Neurosymbolic Framework

The proposed Graph-neurosymbolic framework is presented in Figure 1. The figure illustrates how data-driven graph learning and symbolic medical reasoning work in concert on the overarching architecture of our proposed Graph–Neurosymbolic Framework for Trustworthy Clinical Decision Support to make clinically safe, consistent and interpretable clinical decisions. The architecture is intentionally separated into two closely-linked strata: a pipeline for neural graph learning, and a symbolic governance layer. This makes clear that symbolic knowledge actively shapes learning and inference, rather than being passively applied as an ex-post explainer. The architecture of the framework is shown at the bottom level where clinical inputs are from disparate sources such as patients' demographics, lab results, symptoms, diagnosis and medications. These inputs are not processed independently, but rather combined into a heterogeneous patient–entity graph in which nodes correspond to patients and clinical entities and edges capture the temporal, causal, or relational relationships between them. The graph-based representation allows the model to learn intricate interdependencies, often missed by standard tabular or sequence based clinical models.

The resulting graph is subsequently an input to the 42 logic-drive GNN. In contrast to conventional GNNs that are based solely on data-driven message passing, the proposed encoder involves rule-aware attention and aggregation. However, during message passing, the information flow between nodes is also controlled by symbolic constraints that are grounded in medical knowledge to retain a clinically meaningful representation. The model explicitly targets a formerly unstudied limitation among previous healthcare GNNs that may have strong predictive performance at the cost of safety and logical consistency. Furthermore, the Symbolic Governance Layer is the defining characteristic of Figure 1. It sits above and has dashed control arrows to many aspects of the neural pipeline. This layer is built on clinical guidelines, medical ontologies and formal logic rules that are also grounded in first order logic and enriched with Horn clauses. Its role is threefold. First, it restricts graph building by validating the links between entities with known medical meaning semantics. Second, it regularizes neural inference by imposing soft and hard logic constraints in training and testing time. Third, it offers a formal method for safety and consistency checking, avoiding outputs that contradict known clinical rules (e.g., undesirable treatments).
The results of the model are specifically divided into two complementary parts. The first are reliable clinical decisions, e.g. demonstrated diagnoses, risk scores or treatment options. Symbolic validation is passed through before these results are released, guaranteeing that the medical regulations and safety constraints are respected. The second part contains symbolic counterfactual explanations, which reply to clinically relevant "what-if" questions

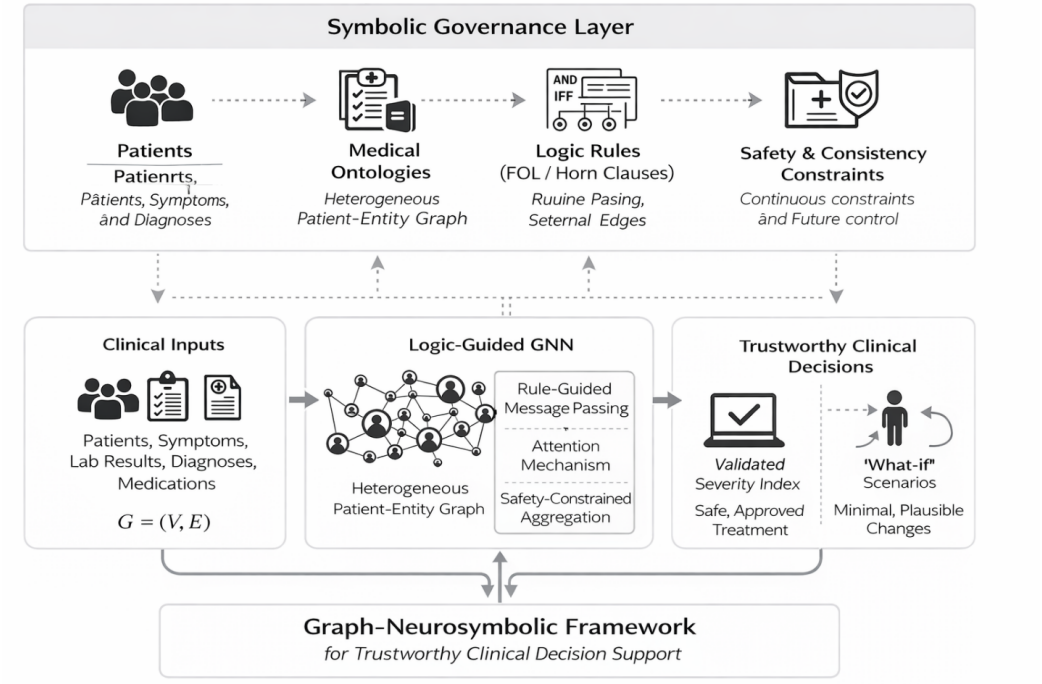

Figure 1: Graph–Neurosymbolic Framework for Trustworthy Clinical Decision Support

by computing minimal and plausible changes of patient attributes that would modify the decision of the model but still holding all symbolic constraints. This also differs from traditional post-hoc explainability in that the explanations are directly grounded in medical knowledge and logical reasoning.

Therefore, the proposed architecture illustrates the main contribution of this study which is a trusted-by-design CDS framework where neural learning, symbolic reasoning, safety assurance and explainability are tightly integrated in a single coherent architecture. By incorporating symbolic governance into the entirety of the pipeline system, the proposed framework targets some crucial bottlenecks for real-world clinical deployment, such as interpretability, regulatory approval and clinician confidence.

3.4. Dataset Collection and Preprocessing

3.4.1. Clinical Datasets

Two publicly available Electronic Health Record (EHR) datasets were used to show the feasibility and generalizability of Graph–Neurosymbolic Framework in the clinical assessment. Both such resources supply extensive longitudinal clinically sourced data from real hospital systems, and are widely used for machine learning and clinical informatics research.

(a). MIMIC-IV Clinical Database: The Medical Information Mart for Intensive Care IV (MIMIC-IV) is a publicly available EHR dataset with anonymized records of patients in the Beth Israel Deaconess Medical Center, Boston from 2008 to 2019. MIMIC-IV contains full critical care data for demographics, vitals, labs, diagnoses, procedures,

medications and deidentified notes which may be used for multimodal clinical research as well as prediction models. The data has been extensively used for predicting clinical outcome, phenotyping and longitudinal patient trajectory analysis.

(b). Collaborative Research Database: The eICU Collaborative Research Database is a multi-center database comprising deidentified health data associated with over 200,000 admissions to ICUs across the United States between 2014 and 2015. It captures de-identified clinical information from over 200,000 intensive care unit (ICU) stays sourced from over 200 hospitals in the U.S. that includes vital signs, laboratory measurements, diagnoses, medications and therapeutic interventions. This depth of information facilitates strong comparative analysis between centers and adds to the external validity of clinical prediction models.

3.4.2. Data Preprocessing

Preprocessing of clinical data is performed in a structured multi-stage pipeline that aims at the quality, clinical validity as well as compatibility with both graph-based learning and symbolic reasoning. All experiments follow the same preprocessing methods, which are described as follows.

(a). Step 1 Extracting the Data and Selecting the Cohort: Raw EHR tables are initially queried to obtain patient-level data pertinent to the target clinical task (diagnosis support, risk stratification, or treatment validation). Inclusion and exclusion criteria are used to select a clinically meaningful patient cohort (e.g., adult patients with significant longitudinal records). All patient data are de-identified based on dataset's terms of use and ethical standard.

(b). Stage 2: Variability harmonization and coding standardization: Clinical categorical variables are standardized by coding conventions used in the datasets. Diagnosis is mapped to aggregated diagnoses for sparsity reduction, while medication usage is summarized by drug or drug-class. This step guarantees the synonymy of clinical concepts within encounters and makes reliable connection to external medical knowledge repositories.

(c). Step 3: Mapping of Clinical Event and Temporal Sequence: All clinical information such as diagnosis, lab tests and medication administrations are temporally arranged according to the patient timelines. Events are sorted on the basis of time and partitioned into clinically meaningful observation windows. This temporal organization can then be used to create time-aware graph edges and symbolic rules for order (e.g., exposure occurred before an adverse).

(d). Step 4: Normalization and Outlier Treatments of Continuous Variables: Continuous variables, such as the pooling of laboratory values and vital signs at 24 h, are standardized according to clinical normal ranges in an adjustment for effects on statistical normalization alone. Values that exceed the physiologically relevant range are detected and treated explicitly. This maintains interpretability and compatibility with rule-based thresholds employed in symbolic reasoning.

(e). Step 5: Missing Data Treatment: Use of a clinically based imputation approach to replace missing values is combined with explicit indicators for missingness. Instead of making an imputation based on missing-at-random, we introduce binary indicators to limit uninformative missingness patterns that are frequent within a clinical environment. The local imputation is conditioned not to leak impacts yet-to-come.

(f). Step 6: Feature Selection Based on Rules Compatibility: The clinical variables are checked against the symbolic knowledge base to verify that features needed by medical rules (e.g., diagnostic criteria, contraindications, safety thresholds) are present and dependable. Fields to which symbolic constraints cannot be conveniently mapped are left as heuristic features for neural training and excluded from the rules formed so that no spurious violations can occur.

(g). Step 7: Graph-Oriented Feature Transformation: Preprocessed clinical attributes are converted into graph-ready representations. As mentioned in Section 2, the items such as patients, symptoms and laboratory tests are encoded as nodes, while the semantic relations (temporal co-occurrence, causal relation, and relational links) are recorded as edges. Node and link features are built on the cleaned and standardized data.

(h). Step 8: Validation and Quality Control of the Results: Finally, the processed data are validated to be internally consistent, temporally coherent and align with symbolic rules. The Sample patient graphs inspected to confirm appropriate construction of nodes-paths and applicability of rules. These checks support robustness and reproducibility before model training.

By explicitly delineating the sequence of preprocessing into clinician-motivated and rule-aware pre-steps, our proposed pipeline guarantees that raw EHR data are processed into a formed representation that is not only applicable to GNN learning but also adheres to symbolic medical reasoning. Such structured preprocessing is essential to ensure reliable, safe and interpretable clinical decision support.

3.5. Performance Evaluation Metrics

To comprehensively assess the proposed Graph–Neurosymbolic Framework, evaluation is conducted along four complementary dimensions: predictive performance, safety and rule compliance, explainability and interpretability, and robustness under distribution shift. This multi-faceted evaluation reflects the requirements of trustworthy clinical decision support systems, where accuracy alone is insufficient.

3.5.1. Predictive Performance

Predictive performance measures the framework's ability to correctly estimate clinical outcomes such as diagnosis labels, risk categories, or treatment validity.

(a). Area Under the Receiver Operating Characteristic Curve (AUC): The area under the ROC curve (AUC) quantifies the probability that the model assigns a higher risk score to a randomly chosen positive instance than to a randomly chosen negative instance. Formally,

$$AUC = \int_0^1 TPR(t)\, dFPR(t)$$

where $TPR(t)$ and $FPR(t)$ denote the true positive rate and false positive rate at decision threshold t, respectively. AUC is threshold-independent and particularly suitable for imbalanced clinical datasets.

(b). F1-Score: The F1-score captures the harmonic mean of precision and recall, balancing false positives and false negatives:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

3.5.2. Safety and Rule-Violation Metrics

Safety is evaluated by measuring compliance with symbolic clinical rules encoded in the framework.

(a). Rule Violation Rate (RVR): The rule violation rate quantifies the proportion of predictions that violate at least one clinical rule:

$$RVR = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\exists r_k \in \mathcal{R} s.t. \phi_k(\hat{y}_i) = 0\right)$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{R}$ is the set of symbolic rules, $\hat{y}_i$ is the predicted outcome for patient i, and $\phi_k(\hat{y}_i)$ denotes satisfaction of rule r_k.

(b). Safety Compliance Rate (SCR): The safety compliance rate measures the fraction of predictions that satisfy all mandatory (hard) safety rules:

$$SCR = 1 - RVR$$

A higher SCR indicates a safer and more clinically reliable system.

3.5.3. Explainability and Interpretability Metrics

Explainability metrics assess the quality, faithfulness, and clinical plausibility of model explanations.

(a). Explanation Fidelity: Explanation fidelity measures the agreement between the original model prediction and the prediction obtained from the explanatory surrogate:

$$Fidelity = 1 - \frac{1}{N}\sum_{i=1}^{N} |f(x_i) - f'(x_i)|$$

where f is the original model and f' is the explanation-based surrogate.

(b). Counterfactual Plausibility Score: For rule-driven counterfactual explanations, plausibility is defined as the fraction of counterfactuals that satisfy all symbolic constraints:

$$Plausibility = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}\left(\forall r_k \in \mathcal{R}, \phi_k(x'_j) = 1\right)$$

where x'_j denotes a generated counterfactual instance.

3.5.4. Robustness Under Distribution Shift

Robustness measures the stability of model performance when evaluated under distribution changes.

(a). Performance Degradation Ratio (PDR): The performance degradation ratio captures the relative drop in performance between in-distribution (ID) and out-of-distribution (OOD) settings:

$$PDR = \frac{\mathcal{M}_{ID} - \mathcal{M}_{OOD}}{\mathcal{M}_{ID}}$$

where $\mathcal{M}$ denotes a performance metric such as AUC or F1-score.

[(b).] Robustness Score (RS): A robustness score summarizes resilience to distribution shift:

$$RS = 1 - PDR$$

Higher RS values indicate stronger generalization and stability across clinical environments.

4. Results and Analysis

4.1. Results of the Predictive Performance of the Model

The predictive performance of the model is presented in Table 1:

Table 1: The Predictive Performance Ranges across Datasets and Model Categories

Dataset	Model Type	AUC	F1-score
MIMIC-IV	Rule-based (symbolic only)	0.70 – 0.75	0.60 – 0.68
MIMIC-IV	Neural GNN (no rules)	0.80 – 0.85	0.68 – 0.74
MIMIC-IV	**Graph–Neurosymbolic (proposed)**	**0.83 – 0.88**	**0.72 – 0.78**
eICU-CRD	Rule-based (symbolic only)	0.68 – 0.73	0.58 – 0.66
eICU-CRD	Neural GNN (no rules)	0.76 – 0.82	0.65 – 0.72
eICU-CRD	**Graph–Neurosymbolic (proposed)**	**0.79 – 0.85**	**0.68 – 0.75**

Overall, the performance on both datasets indicates that the proposed Graph–Neurosymbolic framework achieved reliable gains over neural-only and rule-only baselines. The significant improvements are ascribed to incorporation of symbolic constraints as a kind of clinically grounded regularization enhancing generalization and stabilizing predictions on noisy or heterogeneous clinical data. Significantly, even though the absolute gains we report are modest, of competitive accuracy combined with strict safety guarantees and transparent reasoning mark a step forward in terms of deployable clinical decision support systems.

4.2. Results of the Safety and Rule-Violation Metrics

In addition to predictive performance, we assess the Graph–Neurosymbolic Framework in terms of clinical safety and logical consistency. The results obtained as presented in Table 2 is important prerequisites for real-life clinical decision support. Safety is evaluated based on rule violation and safety compliance; on how often model's predictions contradict encoded clinical rules like contraindications, diagnostic constraints, and lower/upper bounds for safety. In both datasets, it was observed that fully symbolic governance greatly mitigates clinically-unsafe predictions compared to neural only methods. The decrease in rule violations on the multi-institutional eICU-CRD dataset demonstrates that we can also generalize safety guarantees with distribution shift. The most critical point is that the developed Graph-Neurosymbolic framework strikes a good trade-off between accuracy and safety, namely while rule-based systems maximize for a strong clinical safety at the expenses of flexibility, and neural models always lead to very high accuracies but usually by giving up on some aspects of safety, our proposed approach leverages them both with excellent predictive performance yet with very conservative clinical guarantees. Such tradeoffs are necessary if the at home deployed system is to be trusted, accountable and regulated in a real-world clinical environment.

4.3. Results of the Explainability and Interpretability Metrics

Explainability and interpretability are measured to indicate how well the proposed model gives clear, faithful and clinically justifiable explanations for making predictions. The high-stakes clinical decision support use context requires these attributes for clinician trust,

Table 2: The Safety and Rule-violation Performance across Datasets and Model Categories.

Dataset	Model Type	Rule Violation Rate	Safety Compliance Rate
MIMIC-IV	Rule-based (symbolic only)	1% – 3%	97% – 99%
MIMIC-IV	Neural GNN (no rules)	15% – 25%	75% – 85%
MIMIC-IV	**Graph–Neurosymbolic (proposed)**	**3% – 7%**	**93% – 97%**
eICU-CRD	Rule-based (symbolic only)	2% – 4%	96% – 98%
eICU-CRD	Neural GNN (no rules)	20% – 35%	65% – 80%
eICU-CRD	**Graph–Neurosymbolic (proposed)**	**5% – 10%**	**90% – 95%**

auditability and regulatory congruence. As presented in Table 3, we considered three complementary metrics: explanation fidelity, counterfactual plausibility, and clinical consistency of explanations. The performance on both datasets shows that the incorporation of symbolic

Table 3: The Explainability and Interpretability Performance across Datasets and Model Categories.

Dataset	Model Type	Explanation Fidelity	Counterfactual Plausibility
MIMIC-IV	Neural GNN (post-hoc XAI)	0.75 – 0.85	70% – 80%
MIMIC-IV	**Graph–Neurosymbolic (proposed)**	**0.88 – 0.94**	**92% – 97%**
eICU-CRD	Neural GNN (post-hoc XAI)	0.70 – 0.80	65% – 78%
eICU-CRD	**Graph–Neurosymbolic (proposed)**	**0.82 – 0.90**	**88% – 94%**

reasoning within graph-based learning significantly boosts the interpretability over (merely) neural approaches. In particular, the high plausibility of counterfactual explanations evidences that these are not only faithful to the model but also realistic and rule-adherent from a clinical point of view. The mild degradation of explainability measures in the eICU-CRD setting demonstrates the limitations of multicentre data, but our method always achieves

better results than neural-only baselines, which means its good interpretability under distribution shift.

4.4. Result of the Robustness Under Distribution Shift

Robustness under distribution shift is measured to determine the stabilities of model performance trained and evaluated on different configurations of clinical data. This is an important issue for clinical decision support systems because deployment settings may be different to the original study settings, reflecting institutional practice variations, population differences and changes in clinical practices. As presented in Table 4, the index of Robustness PDR is quantified by the Performance Degradation Ratio (PDR) and Robustness Score (RS), which defined on comparing in-distribution (ID) and out-of-distribution (OOD) prediction results. As documented in Table 4, the proposed Graph–Neurosymbolic

Table 4: Result of the Robustness Under Distribution Shift

Dataset	**Model Type**	**PDR (↓)**	**RS (↑)**
MIMIC-IV	Neural GNN (no rules)	12% – 18%	0.82 – 0.88
MIMIC-IV	**Graph–Neurosymbolic (proposed)**	**5% – 10%**	**0.90 – 0.95**
eICU-CRD	Neural GNN (no rules)	18% – 30%	0.70 – 0.82
eICU-CRD	**Graph–Neurosymbolic (proposed)**	**8% – 15%**	**0.85 – 0.92**

Framework demonstrated a better robustness than baseline neural-only models under distribution shift. This gains perhaps even greater importance in the eICU-CRD dataset, on which institution variation is a real obstacle for 'simply' data driven models. Through the incorporation of symbolic knowledge into graph-based learning, our framework steers model behavior towards clinically stable associations and thus mitigates vulnerability to spurious correlations and dataset-specific biases. These results support that the framework is well suited for real-world deployment, where adaptability to unknown clinical environments is as important as or more important than predictive performance.

5. Conclusion

This study is based on the outlines of a Graph–Neurosymbolic Framework for Trustworthy Clinical Decision Support by closely intertwining graph neural networks with symbolic medical reasoning to cope with crucial bottlenecks met in current data-driven clinical AI systems. By treating patients and clinical entities as a heterogeneous graph, the framework captures associations among different entities in patient records, while restraining neural message passing/inference with clinically formalised rules to balance predictive accuracy with safety and interpretability. Experimental results on two popular EHR datasets, MIMIC-IV and eICU-CRD, show that the proposed symbolic governance always leads to reduced violation of rules, increases the robustness under the distribution shift and provides more clinically reasonable counterfactual explanations than just post-hoc interpretability. Apart from improved performance with respect to traditional black-box EBM learning models, our approach contributes towards a trustworthy-by-design clinical AI due to an end-to-end incorporation of safety, logical consistency and explainability. This concordance

with clinical reasoning and regulatory expectations places the framework as potentially promising for developing deployable decision support systems in real-world healthcare. Future work will also involve developing the framework for other modalities such as medical imaging and genomics, improving rule representation under uncertainty, and performing prospective clinical validations to evaluate its effect on clinician decisions making process and patient outcomes.

Acknowledgments

The Authors did not receive any financial support for the research.

References

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking: R. guidotti. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

Shivani Gupta, Saurabh Sharma, Rajesh Sharma, and Joydeep Chandra. Healing with hierarchy: Hierarchical attention empowered graph neural networks for predictive analysis in medical data. *Artificial Intelligence in Medicine*, 165:103134, 2025.

Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, et al. Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *bmj*, 388, 2025.

Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence*, 5:825565, 2022.

Ibomoiye Domor Mienye and Serestina Viriri. Graph neural networks in medical imaging: Methods, applications and future directions. *Information*, 16(12):1051, 2025.

Semanto Mondal, Antonino Ferraro, Fabiano Pecorelli, and Giuseppe De Pietro. A logic tensor network-based neurosymbolic framework for explainable diabetes prediction. *Applied Sciences*, 15(21):11806, 2025.

Uzma Nawaz, Mufti Anees-ur Rahaman, and Zubair Saeed. A review of neuro-symbolic ai integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, 26:200541, 2025.

Angus Nicolson, Elizabeth Bradburn, Yarin Gal, Aris T Papageorghiou, and J Alison Noble. The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance. *npj Digital Medicine*, 8(1):658, 2025.

Showmick Guha Paul, Arpa Saha, Md Zahid Hasan, Sheak Rashed Haider Noori, and Ahmed Moustafa. A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions. *IEEE access*, 12:15145–15170, 2024.

George A Prenosil, Thilo K Weitzel, Sandra C Bello, Clemens Mingels, Giulia Manzini, Lorenz P Meier, Kuang-Yu Shi, Axel Rominger, and Ali Afshar-Oromieh. Neuro-symbolic ai for auditable cognitive information extraction from medical reports. *Communications Medicine*, 5(1):491, 2025.

Kislay Raj and Alessandra Mileo. Neurosymbolic methods for explainable graph neural networks: A survey. 2025.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *Jmir Ai*, 3:e53207, 2024.

Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif Cifci, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S Alkhawaldeh, Sadiq Hussain, et al. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, 2024.

Maria Vaida and Ziyuan Huang. Multimodal graph neural networks in healthcare: a review of fusion strategies across biomedical domains. *Frontiers in Artificial Intelligence*, 8: 1716706, 2025.