

Highlights

Neuro-Symbolic Artificial Intelligence for Real-Time Sepsis Mortality Prediction

Fabrizio De Santis, Gyunam Park, Francesco Zanichelli

- Neuro-symbolic sepsis mortality prediction with 86.45% accuracy and full transparency
- Automated clinical guideline integration via medical knowledge graph and rule mining
- Three-tier explainability framework for concept, rule, and patient-level reasoning
- Outperforms machine learning baselines on sepsis patients
- Addresses artificial intelligence interpretability gap in critical care

Neuro-Symbolic Artificial Intelligence for Real-Time Sepsis Mortality Prediction

Fabrizio De Santis^a, Gyunam Park^b, Francesco Zanichelli^a

^a*University of Parma, Parma, Italy*

^b*Eindhoven University of Technology, Eindhoven, Netherlands*

Abstract

Sepsis mortality prediction in intensive care units represents a critical engineering challenge in healthcare systems, where timely intervention through intelligent decision support systems can significantly reduce mortality rates. Current clinical decision support systems rely on black-box machine learning models that lack the interpretability required for safe deployment in life-critical engineering applications. We develop a neuro-symbolic artificial intelligence framework that integrates Logic Tensor Networks with automated clinical knowledge extraction for interpretable mortality prediction. Our novel contributions include: (1) an automated pipeline for transforming clinical guidelines into first-order logic constraints using medical knowledge graphs and rule mining; (2) a dual axiom architecture with weak anchoring that enables context-adaptive learning while maintaining semantic grounding; and (3) a three-tier explainability framework providing concept-level, rule-level, and patient-level interpretability. Using 19,328 sepsis patients from a publicly available real-world clinical dataset, our system achieves 86.45% accuracy at 6-hour prediction windows, outperforming traditional machine learning baselines while providing full transparency in decision-making. Ex-

ternal validation on 6,735 patients demonstrates preserved ranking ability despite domain shift challenges.

Keywords:

Neuro-symbolic artificial intelligence, Clinical decision support systems, Healthcare systems engineering, Knowledge-based systems, Real-time intelligent monitoring, Explainable artificial intelligence

1. Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection that represents a global health challenge [26]. Each year, it is estimated to impact around 50 million individuals and cause approximately 11 million deaths across the globe [38]. Given its severity and high mortality, early antibiotic treatment and accurate prediction of outcomes is essential for effective patient management, enabling timely clinical interventions and the strategic allocation of healthcare resources [45, 42].

In recent years, the application of machine learning (ML) and deep learning (DL) models to predict sepsis onset and mortality has shown considerable promise, often outperforming traditional clinical scoring systems [44, 45]. However, the adoption of these predictive models in clinical practice is compromised by their inherent “black box” nature [12]. These models, while often accurate, provide predictions without transparent, human-understandable reasoning. This lack of interpretability poses a critical barrier to clinical trust and adoption, as physicians require a clear rationale to integrate artificial intelligence (AI)-driven recommendations into medical decision-making [1, 30]. Concerns about model reliability, data bias, and potential for incorrect con-

19 clusions due to a lack of transparent reasoning limit the effective use of these
20 tools at the patient’s bedside. Traditional engineering approaches, such as
21 rule-based expert systems, offer interpretability but lack the adaptability to
22 handle patient heterogeneity and the complexity of multivariate temporal
23 clinical data, while pure machine learning approaches fail to meet the ex-
24 plainability and safety requirements mandated for clinical decision support
25 systems.

26 To address this fundamental engineering trade-off between model per-
27 formance and system interpretability, this paper explores the paradigm of
28 neuro-symbolic artificial intelligence (NeSy AI) as a practical solution for
29 intelligent automation in critical care. Specifically, we propose NeSy-SMP
30 (Neuro-Symbolic Sepsis Mortality Prediction), a generalizable framework for
31 clinical mortality prediction based on Logic Tensor Networks (LTNs) [3], a
32 neuro-symbolic system that allows for learning from data while reasoning
33 with a set of explicit logical rules expressed in first-order logic (FOL). While
34 demonstrated for sepsis mortality prediction, our framework is designed to
35 be adaptable to other clinical prediction tasks and represents a replicable
36 methodology for developing interpretable AI systems in healthcare engineer-
37 ing applications.

38 We begin by constructing a clinical knowledge graph from established
39 guidelines, which serves as a structured and machine-readable representa-
40 tion of domain expertise. This knowledge graph (KG) enables the extraction
41 of logical rules through automated rule-mining techniques, capturing clini-
42 cally relevant patterns and dependencies between the entities and the main
43 outcome. These logical rules are then embedded into the LTN framework,

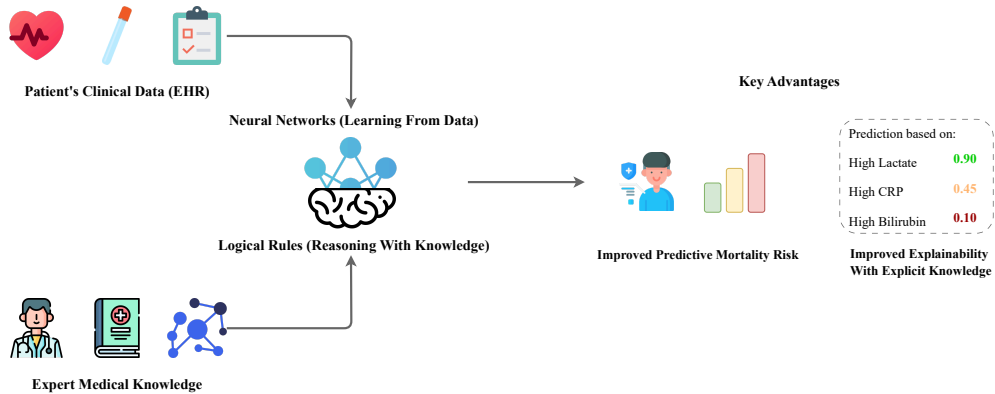


Figure 1: Overview of our contribution: integrating patient data with clinical knowledge in a neuro-symbolic artificial intelligence framework to provide interpretable and accurate sepsis mortality predictions.

allowing the model to incorporate symbolic knowledge as soft constraints during training using a weighted loss function. By doing so, the model learns not only from patient data but also from explicitly encoded clinical concepts, enhancing both its predictive accuracy and interpretability.

The primary contributions of this work are (see Figure 1 for an overview):

- 1. Novel Clinical Neuro-Symbolic Framework:** We introduce the first automated pipeline for sepsis mortality prediction that bridges the gap between clinical guidelines and neural networks. Unlike prior LTN applications that rely on manually crafted rules, we develop an end-to-end system that: (i) automatically extracts and formalizes clinical knowledge from SNOMED CT, (ii) learns context-adaptive clinical thresholds through weak anchoring (addressing the critical challenge of patient heterogeneity), and (iii) achieves superior performance (86.45% accuracy, 80.35% F1-score) while maintaining interpretability.

58 **2. Automated Knowledge Integration:** We develop a systematic ap-
59 proach for extracting and integrating clinical guidelines into neural
60 networks through automated rule mining from SNOMED CT-based
61 knowledge graphs, reducing manual knowledge engineering effort while
62 ensuring clinical validity through expert validation.

63 **3. Multi-Level Clinical Explainability:** We design a three-tier ex-
64 plainability framework that enables: (i) *clinical concept learning veri-*
65 *fication* to assess semantic alignment of learned predicates with clini-
66 cal definitions; (ii) *medical rule consistency validation* to evaluate rule
67 adherence across patient cohorts; and (iii) *individual patient reason-*
68 *ing analysis* to examine patient-specific decision patterns, providing
69 clinicians with actionable insights at multiple granularities for clinical
70 decision support.

71 **4. Validated Clinical Integration:** We demonstrate practical appli-
72 cability through comprehensive evaluation across multiple prediction
73 windows (6-48 hours) on 19,328 sepsis patients from MIMIC-IV, with
74 open-source implementation to facilitate reproducible research and clin-
75 ical adoption.

76 The paper is structured as follows. Section 2 provides an overview of
77 related work in sepsis mortality prediction and [NeSy](#) AI in healthcare. Sec-
78 tion 3 introduces the technical background and problem formulation. Sec-
79 tion 4 describes our neuro-symbolic framework, including knowledge graph
80 construction, rule mining, and model architecture. Section 5 presents a com-
81 prehensive case study using the MIMIC-IV dataset, including experimental

82 setup, performance evaluation, explainability analysis, and external valida-
83 tion using the eICU dataset. Section 6 presents the summary of key findings,
84 a discussion on clinical impact, and the analysis on limitations and threats to
85 validity. Finally, Section 7 summarizes the main contributions and outlines
86 directions for future research.

87 2. Related work

88 Artificial intelligence and machine learning have been increasingly adopted
89 to build clinical decision support systems to support risk and stratification
90 through ensemble methods, explainable artificial intelligence (XAI), and epi-
91 demiological markers to achieve high accuracy and mortality risk assessment.
92 For instance, ensemble learning approaches have been proven highly effective
93 in infectious disease management. Chadaga et al. [8] custom stacked ensem-
94 ble architectures utilizing haematological markers have been applied to di-
95 agnosing COVID-19, where explainability techniques were employed to iden-
96 tify key predictive features. In clinical decision-support systems, the lack
97 of transparency of ML models remains a major barrier to their adoption.
98 As a result, XAI techniques and model interpretability have gained increas-
99 ing attention as means to support clinicians’ trust and understanding [27].
100 Zhang et al. [52] developed an ensemble ML model to predict sepsis-associated
101 acute kidney injury demonstrating good performance in multicenter external
102 datasets. In the context of chronic conditions, Vivek et al. [46] employed en-
103 semble stacking for gestational diabetes prediction, combining data balancing
104 with XAI libraries such as SHAP and LIME to ensure interpretable clinical
105 decision support. Ghosh et al. [16] investigated the use of different XAI tech-

106 niques and ML models to predict chronic kidney disease, revealing creatinine,
107 glycosylated hemoglobin, and age as the three most influential features. Pad-
108 mal et al. [34] proposed a lightweight multi-path learning framework with XAI
109 capabilities for ovarian cancer detection, integrating DenseNet-based feature
110 extraction with attention and uncertainty estimation to achieve high diag-
111 nostic accuracy and model interpretability.

112 Mortality prediction has become a crucial focus in medical research, aim-
113 ing to identify patients at high risk of death and support timely clinical
114 interventions. Advancements in ML have enabled the development of pre-
115 dictive models that leverage diverse patient data to estimate mortality risks
116 in various diseases. Chadaga et al. [9] utilized ensemble algorithms to pre-
117 dict mortality in COVID-19 patients based on epidemiological parameters,
118 achieving high accuracy. Li et al. [25] evaluated the performance of ML mod-
119 els in predicting patients at high-mortality risk within 12h after admission in
120 the emergency department using demographics, vital signs, and laboratory
121 tests. While these studies demonstrate the growing effectiveness of ML and
122 XAI techniques in clinical settings, they rely on post-hoc explanations ap-
123 plied to black-box models rather than embedding domain knowledge directly
124 into the learning process, highlighting the need for approaches that integrate
125 clinical reasoning as a core component of the model architecture.

126 *2.1. Mortality Prediction in Sepsis*

127 The prediction of sepsis mortality has traditionally relied on clinical scor-
128 ing systems such as the Acute Physiology and Chronic Health Evaluation II
129 (APACHE II) and the Sequential Organ Failure Assessment (SOFA) score [31].
130 These systems are widely used and provide clinicians with a standardized

131 method for assessing disease severity. However, they have some limitations,
132 including reliance on a limited number of variables. Studies have shown that
133 while these scores have discriminative ability, their accuracy can be moder-
134 ate, with reported [AUC](#) values often ranging between 0.6 and 0.8, depending
135 on the specific model and patient [21].

136 The use of [ML and DL](#) models, especially those trained on massive
137 amounts of data, has led to significant advancements in many domains, in-
138 cluding healthcare. These models can effectively handle large amounts of
139 data and identify complex, non-linear patterns in clinical data, leading to
140 more accurate predictions. [Recent engineering applications have focused on](#)
141 [temporal pattern recognition in multivariate clinical data \[40\]](#). ML models,
142 such as Random Forests, Gradient Boosting, and Linear Regression, have
143 been employed for in-hospital mortality of sepsis patients in the [intensive](#)
144 [care unit \(ICU\)](#), demonstrating superior performance compared to tradi-
145 tional scoring systems [22]. Su et al. [43] applied machine learning techniques
146 to predict mortality, disease severity, and ICU length of stay using patient
147 data collected within the first 6 hours of ICU admission. Among the models
148 evaluated, such as logistic regression, random forest, and XGBoost, the ran-
149 dom forest classifier delivered the highest performance across all prediction
150 tasks, outperforming traditional tools like the SOFA score. Yong et al. [51]
151 proposed a deep learning model for sepsis mortality prediction combining a
152 deep neural network and a graph convolutional neural network, leveraging
153 similarity connectivity graphs of sepsis patients to connect patients with sim-
154 ilar indicators and outperforming common baseline methods. Feng et al. [14]
155 proposed an approach for sepsis mortality prediction that leverages Graph

156 Attention Networks (GAT) and Gated Recurrent Units (GRU). The GRU
157 models were employed to capture the temporal variations of single organs,
158 while the GAT model captures inter-organ dependencies over time using the
159 data collected 48 hours after the diagnosis. The model outperformed com-
160 mon baselines in mortality prediction, maintaining strong performance even
161 under data-scarce conditions.

162 However, a major disadvantage to their clinical adoption is their "black
163 box" nature. The reasoning process behind their predictions can undermine
164 trust and prevent clinical validation, as physicians cannot verify the rationale
165 behind life-or-death predictions [47]. This lack of transparency is a significant
166 ethical obstacle, creating a need for models that are not only accurate but also
167 explainable. For this reason, many authors have employed post-hoc explain-
168 ability techniques, such as SHAP [28], to identify the most important features
169 driving the model's predictions. Hu et al. [17] developed a machine learning
170 model for predicting in-hospital mortality in septic patients readmitted to
171 the ICU. After training the model, they applied SHAP techniques, identify-
172 ing [Glasgow Coma Scale \(GCS\)](#) score, urine output, lactate, platelet count,
173 and [systolic blood pressure](#) as the most influential features. Chang et al. [10]
174 proposed a two-stage approach that first trains a performance-optimized pre-
175 dictive model and then applies transfer learning to enhance fairness across
176 patient demographics. The authors introduce a permutation-based feature
177 importance method designed to explain each feature's contribution to fair-
178 ness rather than predictive accuracy. Their results highlight the potential
179 of combining fairness-aware model adaptation with tailored interpretability
180 techniques.

181 While ML methods surpass traditional scores like APACHE II and SOFA
182 in early sepsis mortality prediction, their black-box nature still limits clinical
183 adoption. Post-hoc explainability methods such as SHAP are commonly
184 used to identify key features, such as GCS, lactate, and urine output. However,
185 SHAP exhibits limitations with correlated features, potentially assigning
186 misleading importance scores by giving zero weight to predictive features
187 or inflating scores of irrelevant ones [18]. Consequently, there is increasing
188 demand for models that are both accurate and accompanied by simpler,
189 validation-friendly explanations.

190 2.2. Neuro-Symbolic AI in Healthcare

191 In response to the limitations of “black box” models, the field of NeSy AI
192 has gained significant attention. NeSy AI is a subfield of artificial intelligence
193 that seeks to integrate the advantages of neural networks with symbolic logic
194 and reasoning [5]. Connectionist systems (deep learning) excel at pattern
195 recognition in noisy, high-dimensional data, while symbolic systems provide
196 a framework for explicit reasoning and knowledge representation. In symbolic
197 AI, knowledge is represented using symbols that represent objects, concepts,
198 or relationships in the real world. These symbols are then manipulated using
199 a set of rules that define how they can be combined, transformed, or inferred
200 to produce new knowledge or solve problems. However, symbolic AI struggles
201 with uncertainty, ambiguity, and incomplete information [37]. NeSy AI aims
202 at combining the best of both worlds, creating more sophisticated systems
203 capable of both learning and logical inference.

204 Applications of NeSy AI in medicine are emerging, showing promise in
205 many tasks [24]. A recent comprehensive review and case study by [41]

206 systematically examined the integration of medical knowledge into ML mod-
207 els across the entire pipeline. By mapping various integration strategies onto
208 each stage of the ML process, the study demonstrated how domain knowledge
209 can be leveraged to improve both predictive performance and interpretabil-
210 ity. In a detailed case study on diabetes prediction, the authors showed that
211 embedding clinical knowledge not only improved model accuracy and data ef-
212 ficiency but also ensured better adherence to established medical guidelines.
213 Belmonte et al. [4] combined a symbolic approach to brain segmentation
214 and delineation of brain lesions with ML techniques to obtain a hybrid ap-
215 proach that is able to provide accurate and explainable segmentation results.
216 García-Barragán et al. [15] introduced a neuro-symbolic system which inte-
217 grates large language models (LLMs), named entity recognition (NER), and
218 entity linking with symbolic reasoning over medical vocabularies. Applied
219 to breast cancer clinical notes, they achieved substantial improvements in
220 entity recognition and linking, demonstrating the value of combining LLMs
221 with symbolic reasoning.

222 However, despite the potential advantages of NeSy AI, its application in
223 the healthcare domain remains relatively unexplored.

224 2.3. Knowledge Graphs in Clinical Decision Support

225 Together with propositional logic and FOL, knowledge graphs constitute
226 one of the most widely used forms of knowledge representation. In healthcare,
227 KGs are used to model complex biomedical information, linking concepts
228 such as diseases, symptoms, and drugs.

229 KGs have been applied in numerous use cases in clinical decision sup-
230 port [23]. By unifying data from diverse sources like electronic health records

(EHRs), clinical trial data, and scientific literature, KGs can reveal unseen patterns and connections [32]. Zhou et al. [53] proposed an approach to enhance clinical decision support in hypertension treatment, developing a domain-specific KG aligned with the Chinese hypertension guidelines. The proposed KG captures domain-specific medication knowledge and integrates traditional and graph-based representations to support automated reasoning. Evaluated on real-world patient data, the system delivers explainable medication recommendations, achieving 91% recall and 77% MRR in a single-drug recommendation task. Yang et al. [50] developed a sepsis KG leveraging large language models. They included clinical guidelines, public datasets, and real-world data, resulting in a reference for clinicians, enhancing the comprehension of sepsis and its treatments. Carvalho et al. [7] constructed a KG from EHR data, enriched with semantic annotations to enhance its expressiveness. They then generated graph-based embeddings that capture the clinical context of a patient’s ICU stay. These embeddings served as input features for ML models to predict 30-day ICU readmission, achieving a mean AUC of 0.827.

By enabling structured, explainable, and domain-specific reasoning, KGs enhance the accuracy and transparency of medical recommendations. Moreover, they can serve as a valuable and continuously updated knowledge base to be integrated into neuro-symbolic approaches.

Positioning Our Work: Despite growing interest in NeSy AI for healthcare, existing approaches remain largely unsuitable for real-world decision support. First, most systems rely on manually curated medical knowledge, which limits scalability, hinders adaptation to evolving clinical guidelines,

256 and introduces substantial engineering overhead. Second, prior NeSy models
257 lack temporal modeling, failing to capture highly dynamic and irregular tem-
258 poral patterns that characterize ICU patient trajectories [49], particularly in
259 sepsis. Third, current explainability methods provide single-level explana-
260 tions [2, 6], offering limited support for clinicians who require explanations
261 at multiple granularities, from physiological signals to symbolic medical con-
262 cepts. These limitations jointly prevent NeSy models from achieving clinical
263 relevance. Our work addresses these gaps introducing the following contri-
264 butions:

- 265 • A neuro-symbolic framework for ICU risk prediction that eliminates
266 manual knowledge engineering by automatically extracting clinically
267 grounded concepts and relations from SNOMED-CT, enabling scalable
268 and maintainable integration of medical knowledge.
- 269 • BiLSTM-based temporal modeling within LTN that embeds patient
270 trajectory modeling with neuro-symbolic reasoning, allowing the model
271 to capture irregular and evolving clinical observations.
- 272 • The first application of LTNs to sepsis mortality prediction, demon-
273 strating that LTN-based reasoning can be extended to high-stakes,
274 temporally dynamic clinical tasks.
- 275 • A weak anchoring mechanism that preserves the semantic grounding
276 of symbolic predicates, ensuring that the learned concepts remain in-
277 terpretable and aligned with clinical meaning throughout training and
278 inference.

- A three-tier explainability framework tailored to clinician needs, which demonstrates model interpretability by verifying that learned clinical concepts are represented correctly, validating that encoded medical rules are followed consistently, and examining the reasoning underlying individual patient predictions.

3. Preliminaries

3.1. Problem Formulation

In this study, we formulate the task as a supervised classification problem over temporal sequences of clinical events. For each patient, the electronic health record is represented as a chronologically ordered sequence of events

$$\sigma = \langle e_1, e_2, \dots, e_n \rangle$$

where each event e is defined as a tuple

$$e = (a, c, t, attr_1, \dots, attr_2)$$

Here, a denotes the activity (e.g., a laboratory test, medication administration, or vital sign measurement), c is the **de-identified** case identifier (such as the hospital admission ID), t is the timestamp indicating when the event occurred, and $attr_1, \dots, attr_2$ are additional attributes describing relevant information, such as measurement values, units, patient’s age, and admission location. **Importantly, c does not correspond to any real-world patient identifier and cannot be used to re-identify individuals. All data used in this study are fully de-identified and were obtained from the MIMIC-IV and the eICU databases, which are released under privacy protections and approved data**

299 use agreements, in compliance with HIPAA and relevant regulatory require-
 300 ments. This representation preserves both the temporal ordering and the
 301 contextual details of each clinical event, enabling the model to learn patterns
 302 that reflect the progression of a patient’s hospital stay. In our case, these se-
 303 quences serve as inputs to an LTN framework, which integrates the temporal
 304 data with domain-specific symbolic rules to predict the binary outcome of
 305 in-hospital mortality.

306 3.2. Logic Tensor Networks

307 Logic Tensor Networks are a neuro-symbolic learning and reasoning frame-
 308 work that integrates first-order logic with deep neural networks [3]. At its
 309 heart, LTN utilizes a fully differentiable first-order fuzzy logic language called
 310 “Real Logic”. Unlike classical logic, where statements are strictly true or false,
 311 Real Logic assigns a degree of truth to formulas, typically a real number in
 312 the interval $[0, 1]$. Real Logic is built upon FOL, which consists of

- 313 • a set $\mathcal{C}$ of constant symbols, representing specific entities in the domain,
 314 such as a particular vital sign, medication, or laboratory test;
- 315 • a set $\mathcal{X}$ of variable symbols, acting as placeholders for individuals and
 316 allowing statements to be made about any entity in a domain (e.g., x
 317 representing a patient);
- 318 • a set $\mathcal{F}$ of functional symbols, representing mappings that transform
 319 input tensors into output tensors using mathematical operations;
- 320 • a set $\mathcal{P}$ of relational symbols (predicates), representing properties of
 321 single entities or relationships between entities, such as *HighMortality*

322 for modeling the mortality risk probability.

323 • a set $\Diamond \in \{\neg, \vee, \wedge, \rightarrow\}$ of logical connectives, allowing to connect pred-
324 icates and logical formulas.

325 • a set $\mathcal{Q} \in \{\forall, \exists\}$ of quantifiers, allowing to aggregate the values of each
326 formula.

327 The fundamental idea behind LTNs is the *grounding* $\mathcal{G}$ of these symbolic
328 entities within a neural computational graph. LTN defines $\mathcal{G}$ as a formal
329 mapping of FOL elements into real-valued tensors and neural models so they
330 can be evaluated and learned using data. More in detail,

331 • Constants $\mathcal{C}$ are grounded as fixed tensors of real numbers.

332 • Variables $\mathcal{X}$ are grounded as sequences of individuals (tensors), rep-
333 resenting data samples (i.e., the sequences of events recorded for each
334 patient).

335 • Functions $\mathcal{F}$ are grounded as mathematical operations that take input
336 tensors and return output tensors.

337 • Predicates $\mathcal{P}$ are grounded as neural classifiers (e.g., an LSTM for
338 *HighMortality*) mapping tensor inputs to truth values in $[0, 1]$, typi-
339 cally with a sigmoid output. The truth value indicates the degree of
340 truth to which the predicate holds true.

341 The existential quantifier $\exists$ is implemented using the generalized mean (pMean),
342 while the universal quantifier $\forall$ is implemented using the generalized mean

343 of the deviations w.r.t. to the truth (pMeanError), defined as:

344

$$\text{pMean}(u_1, \dots, u_n) = \left(\frac{1}{n} \sum_{i=1}^n u_i^p \right)^{\frac{1}{p}}, \quad p \geq 1 \quad (1)$$

$$\text{pMeanError}(u_1, \dots, u_n) = 1 - \left(\frac{1}{n} \sum_{i=1}^n (1 - u_i)^p \right)^{\frac{1}{p}}, \quad p \geq 1 \quad (2)$$

345 where u is the vector of true values in $[0, 1]$ of logical axioms in the knowledge
 346 base, and the value of p can be seen as a hyperparameter used to write more
 347 or less strict formulas.

348 Logical connectives are also translated into differentiable operations within
 349 the tensor network using the following fuzzy operators:

- 350 • the standard negation $\neg u = 1 - u$
- 351 • the product t-norm $u \wedge v = uv$
- 352 • the product t-conorm $u \vee v = u + v - uv$
- 353 • the Reichenbach implication $u \rightarrow v = 1 - u + uv$

354 where u and v denote two truth values in $[0, 1]$.

355 The symbolic knowledge, expressed as logical axioms or rules, is incor-
 356 porated directly into the learning process. Instead of merely optimizing a
 357 standard loss function, LTNs convert formulas into computational graphs
 358 and reduce the learning problem to a maximum satisfiability problem. The
 359 objective during training is to find the optimal set of parameters θ of the
 360 neural networks that maximize the satisfiability (i.e., the truth values) of a

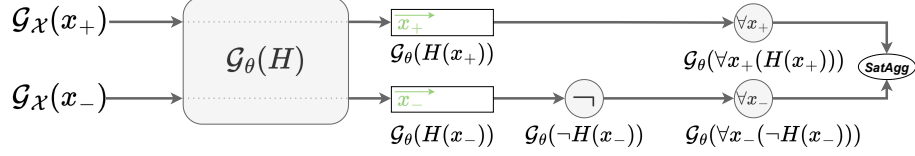


Figure 2: The computational graph representing the mortality prediction task modeled using the LTN framework.

361 given knowledge base $\mathcal{K}$. In practice, the optimizer uses the following loss
 362 function:

$$L = (1 - \text{SatAgg}_{\phi \in \mathcal{K}} \mathcal{G}_\theta(\phi)) \quad (3)$$

363 where SatAgg is an aggregation operator, which, in our case, is the pMean-
 364 Error, Θ is the hypothesis space for θ , and ϕ is the set of logical axioms in
 365 the knowledge base $\mathcal{K}$. The operator SatAgg aggregates the truth values of
 366 each logical formula to produce the overall satisfaction score.

367 We can model a mortality prediction task leveraging the LTN frame-
 368 work. Figure 2 depicts the computational graph of the mortality prediction
 369 task, where H is the *HighMortality* predicate modeled as a neural net-
 370 work and x_+ and x_- are the variables representing patients belonging to the
 371 positive class (i.e., dead patients) and the negative class (i.e., survivors), re-
 372 spectively. Predicate $H(x_+)$ returns values close to 1 for positive instances,
 373 while $\neg H(x_-) = 1 - H(x_-)$ returns values close to 1 for negative instances.
 374 The two quantifiers $\forall x_+$ and $\forall x_-$ aggregate the truth values for positive and
 375 negative instances, respectively, producing an overall satisfaction score that
 376 is used to compute the loss.

377 4. NeSy-SMP: A Neuro-Symbolic Framework for Sepsis Mortality 378 Prediction

379 This section introduces NeSy-SMP, a neuro-symbolic framework that op-
380 erates through a systematic five-stage pipeline, illustrated in Figure 3, that
381 seamlessly integrates clinical domain knowledge with data-driven learning.
382 The framework begins by extracting clinical knowledge from sepsis guide-
383 lines and constructing a formal knowledge graph using SNOMED CT ter-
384 minology. Automated rule mining then discovers logical patterns, which are
385 translated into FOL formulas. Simultaneously, patient data from electronic
386 health records is processed into structured temporal sequences with defined
387 observation windows. Finally, the logical rules and patient data are inte-
388 grated within the LTN framework, where clinical predicates are modeled as
389 neural networks and logical rules serve as soft constraints during training.
390 This systematic architecture enables the model to leverage both statistical
391 patterns in clinical data and explicit domain knowledge, producing mortality
392 predictions that are simultaneously accurate and clinically interpretable.

393 4.1. Clinical Knowledge Extraction

394 We systematically extract clinical concepts and relationships from evidence-
395 based sepsis treatment guidelines, particularly the Surviving Sepsis Cam-
396 paign [13]. This process involves identifying key clinical entities including
397 biomarkers (lactate, bilirubin, creatinine), vital signs (blood pressure, heart
398 rate, temperature), and organ dysfunction indicators (GCS, platelet count,
399 oxygen saturation). We also capture pathophysiological relationships includ-
400 ing causal dependencies (e.g., hypotension leading to organ failure), associa-

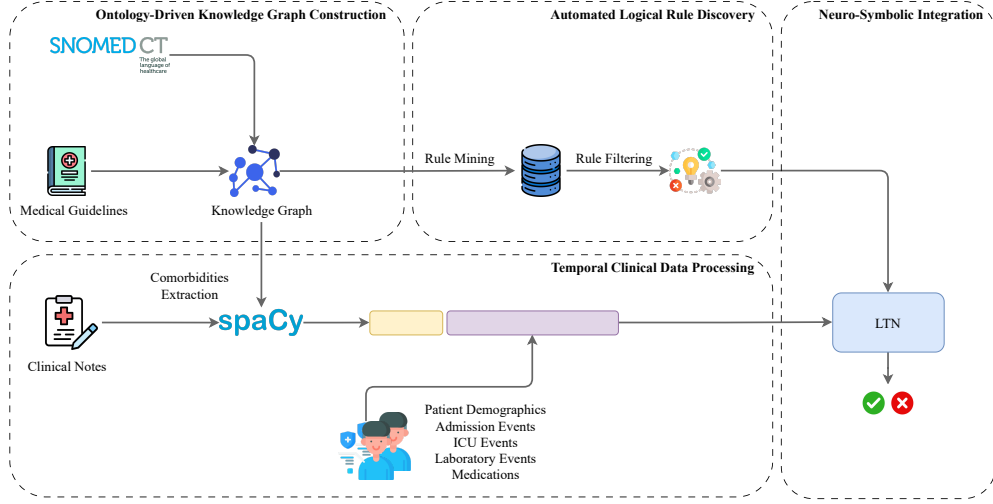


Figure 3: The pipeline followed in our approach, which consists of **ontology-driven** knowledge graph construction, **automated logical rule discovery**, **temporal clinical data processing**, and then **neuro-symbolic integration** leveraging the LTN framework.

401 tive patterns (e.g., elevated inflammatory markers correlating with severity),
 402 and temporal progressions (e.g., lactate clearance over time). The extraction
 403 creates a foundational knowledge repository that captures expert medical un-
 404 derstanding of sepsis progression and outcome determinants, ensuring that
 405 our computational model is grounded in established clinical evidence.

406 4.2. *Ontology-Driven Knowledge Graph Construction*

407 The extracted clinical concepts are formalized into a structured knowl-
 408 edge graph using SNOMED CT terminology as the semantic backbone, as
 409 depicted in Figure 4. Clinical entities are represented as **Resource Descrip-**
 410 **tion Framework (RDF)** nodes with standardized medical codes, ensuring un-
 411 ambiguous concept definitions and semantic interoperability across health-
 412 care systems. Relationships between entities are encoded as typed edges us-

ing Schema.org properties, specifically *increasesRiskOf*, *associatedWith*, and
causedBy predicates. The resulting graph structure follows RDF triple no-
tation (*subject, predicate, object*), enabling machine-readable representation
of complex clinical relationships.

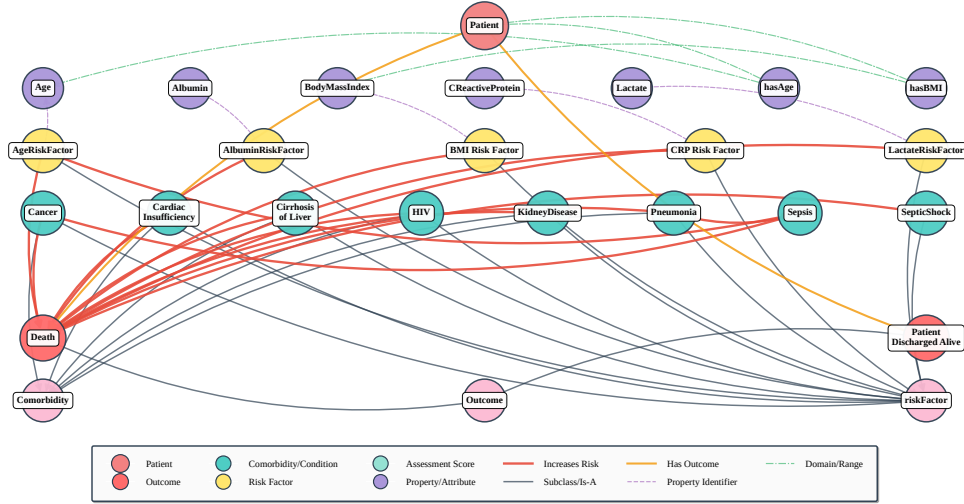


Figure 4: A subset of the knowledge graph constructed based on clinical guidelines for sepsis.

For example, key relationships in our sepsis knowledge graph include:

- (SNOMED.Lactate, RDFS.subClassOf, SCHEMA.RiskFactor)
- (SNOMED.Lactate, SCHEMA.increasesRiskOf, SNOMED.Death)
- (SNOMED.Hypotension, SCHEMA.causedBy, SNOMED.Sepsis)

The construction process involves three systematic steps: (1) concept
selection from SNOMED CT relevant to sepsis and mortality prediction, in-
cluding risk factors, clinical findings, and outcomes; (2) triple generation

424 representing each concept as an RDF node with medically plausible relation-
425 ships encoded as RDF triples; and (3) schema alignment linking SNOMED
426 concepts to Schema.org types for semantic consistency. All extracted con-
427 cepts and relationships are validated by domain experts, including experi-
428 enced clinicians in critical care and infectious diseases.

429 4.3. Automated Logical Rule Discovery

430 We employ AnyBURL, a state-of-the-art bottom-up rule mining algo-
431 rithm, to systematically discover logical patterns within the knowledge graph.
432 The algorithm samples random paths through the graph and generalizes them
433 into Horn clauses of the form $Head \leftarrow Body$, where the body represents con-
434 junctive clinical conditions and the head represents [a clinical condition that](#)
435 [holds when the body holds](#).

436 Rules are ranked by confidence scores based on statistical evidence in
437 the knowledge graph, computed as the ratio of positive instances to total
438 instances matching the rule pattern. To ensure high-quality knowledge inte-
439 gration, we retain only rules with confidence scores exceeding 0.8, indicating
440 strong statistical support. Additionally, we apply domain-specific filtering
441 to retain only rules where the head predicate represents patient outcomes
442 relevant to mortality prediction. [For instance, after filtering and processing,](#)
443 [a discovered rule might be: \$Death\(x\) \leftarrow Lactate\(x\) \wedge Platelets\(x\)\$, where](#)
444 [threshold values for vital signs and laboratory parameters can be systemati-](#)
445 [cally retrieved by executing queries over the knowledge graph.](#)

446 The discovered Horn clauses are subsequently translated into first-order
447 logic formulas suitable for neural-symbolic integration. Each clinical concept
448 in the rules is mapped to a corresponding FOL predicate, with universal

quantifiers applied to represent general clinical patterns: $\forall x : HighLactate(x) \wedge LowPlatelets(x) \rightarrow HighMortality(x)$, where *HighMortality* denotes the predicate formalizing the probability of death, and is semantically linked to the *Death* entity within the knowledge graph. This formalization enables direct integration into the LTN framework while preserving the logical structure of clinical relationships.

4.4. Temporal Clinical Data Processing

Patient data from EHRs undergoes comprehensive preprocessing to create structured temporal sequences compatible with the neuro-symbolic framework. Static features, including demographics and admission characteristics, are encoded as fixed-dimensional vectors.

Dynamic clinical measurements, encompassing vital signs and laboratory values, are organized into time-ordered sequences $\sigma = \langle e_1, e_2, \dots, e_n \rangle$ where each event $e_i = (activity, timestamp, value, metadata)$ captures the complete clinical context.

Comorbidities are extracted from unstructured clinical notes using spaCy-based NER combined with rule-based pattern matching. Our custom natural language-processing (NLP) pipeline focuses on *History of Present Illness* and *Past Medical History* sections to identify clinically relevant pre-existing conditions that influence mortality risk.

Observation windows are systematically defined with configurable lead times prior to outcome events, as illustrated in Figure 5. For deceased patients, observation windows are constructed ending exactly at the specified interval before death, capturing temporal progression while preventing data leakage. For survivors, observation windows are randomly selected from their

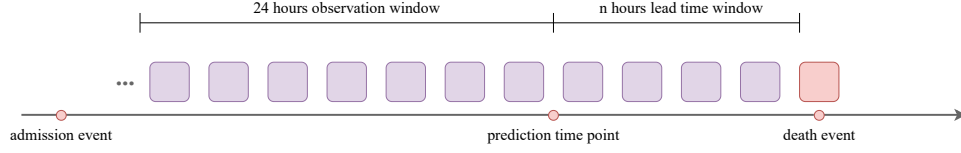


Figure 5: Observation window and lead time window definition for predicting in-hospital mortality. In our study, we use $n = 6, 12, 24$, and 48 hours.

clinical stay. This temporal design enables the model to learn predictive patterns at different time horizons while maintaining rigorous evaluation standards.

4.5. Neuro-Symbolic Integration via Logic Tensor Networks

The logical rules and processed patient sequences are integrated within the LTN framework through differentiable grounding. Clinical predicates (e.g., $HighLactate(x)$, $HighMortality(x)$) are modeled as neural networks with sigmoid activation functions, producing truth values in the interval $[0, 1]$ that represent the degree to which each predicate holds for a given patient.

The knowledge base $\mathcal{K}$ forms the core component of our neuro-symbolic approach, comprising logical rules extracted from the clinical knowledge graph and classification axioms. We distinguish between two categories of logical axioms: data axioms $\mathcal{K}_D$ and knowledge axioms $\mathcal{K}_K$.

Data axioms $\mathcal{K}_D$ comprise classification-related constraints that are optimized to enhance classifier performance:

- $\forall x_+ HighMortality(x_+)$ - deceased patients should have high mortality predictions
- $\forall x_- \neg HighMortality(x_-)$ - surviving patients should have low mortal-

492 ity predictions

493 Knowledge axioms $\mathcal{K}_K$ encode clinical domain expertise derived from sep-
494 sis guidelines through automated rule mining. These axioms identify key
495 prognostic indicators of systemic instability, ranging from hemodynamic met-
496 rics like mean arterial pressure (MAP) to inflammatory and immune markers
497 such as C-reactive protein (CRP) and white blood cell (WBC) count:

- 498 • $\forall x \text{ HighLactate}(x) \rightarrow \text{HighMortality}(x)$ - elevated lactate indicates
499 poor prognosis
- 500 • $\forall x \text{ HighBilirubin}(x) \rightarrow \text{HighMortality}(x)$ - hepatic dysfunction marker
- 501 • $\forall x \text{ HighCRP}(x) \rightarrow \text{HighMortality}(x)$ - inflammatory response indi-
502 cator
- 503 • $\forall x \text{ HighWBC}(x) \rightarrow \text{HighMortality}(x)$ - immune system activation
- 504 • $\forall x \text{ LowPlatelets}(x) \rightarrow \text{HighMortality}(x)$ - coagulation dysfunction
- 505 • $\forall x \text{ LowMAP}(x) \rightarrow \text{HighMortality}(x)$ - hemodynamic instability
- 506 • $\forall x \text{ AgeRisk}(x) \rightarrow \text{HighMortality}(x)$ - advanced age vulnerability
- 507 • $\forall x \text{ LactateNotClearing}(x) \rightarrow \text{HighMortality}(x)$ - persistent metabolic
508 dysfunction
- 509 • $\forall x \text{ HasChronicCondition}(x) \rightarrow \text{HighMortality}(x)$ - comorbidity bur-
510 den

511 While data axioms apply to specific patient groups (deceased x_+ and
512 surviving x_- patients), knowledge axioms hold universally across the entire

513 patient cohort, providing expert-guided constraints during learning. The
 514 logical rules serve as soft constraints through a weighted satisfiability loss
 515 function:

$$L = 1 - (w_D \cdot \text{SatAgg}(\mathcal{K}_D) + w_K \cdot \text{SatAgg}(\mathcal{K}_K)) \quad (4)$$

516 The operator $\text{SatAgg}(\cdot) \in [0, 1]$ denotes the satisfiability of a set of fuzzy
 517 FOL formulas modeled using the pMeanError aggregator (Eq. 2) with $p = 2$,
 518 which corresponds to the default setting used in [3]. This choice provides a
 519 balanced trade-off between smooth optimization and sensitivity to individual
 520 rule violations. For finite values of p , the aggregation behaves as a smooth
 521 mean over individual rule violations, allowing the learning process to balance
 522 partially satisfied constraints. In the limit $p \rightarrow \infty$, the mean converges to the
 523 maximum violation, and consequently the aggregation approaches the min-
 524 imum satisfiability among all rules, enforcing a strict worst-case semantics.

525
 526 Since satisfiability is differentiable with respect to the model parameters,
 527 minimizing L is equivalent to maximizing the joint satisfiability of data-
 528 driven and knowledge-driven axioms. The weights $w_D = 0.8$ and $w_K = 0.2$
 529 balance data-driven learning with domain knowledge integration, with $w_D +$
 530 $w_K = 1$ to ensure the total satisfaction score remains a truth value in $[0, 1]$.
 531 Assigning a higher weight to the satisfiability of data axioms scales their
 532 contribution to the loss accordingly, effectively prioritizing the fulfillment of
 533 data-driven constraints during optimization. As a result, violations of data
 534 axioms induce larger gradients than violations of knowledge axioms, biasing
 535 learning toward empirical evidence while still enforcing domain knowledge as
 536 a regularizing signal. The effect of assigning different weights to data and

537 knowledge axioms is analyzed through an ablation analysis in Section 5.4.

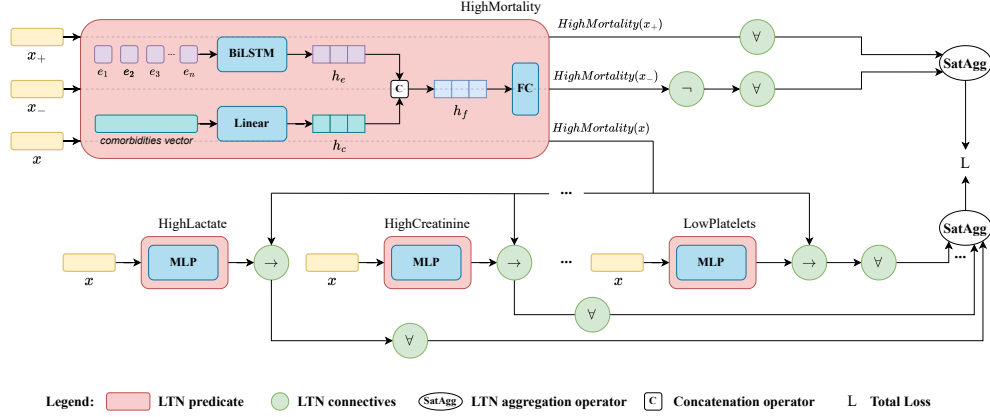


Figure 6: Overall architecture of the neuro-symbolic model for sepsis mortality prediction.

538 The neural architecture underlying NeSy-SMP consists of specialized com-
 539 ponents for different predicate types, as shown in Figure 6, where h_e is the
 540 embedding representation generated by the BiLSTM model, h_e is the embed-
 541 ding representation of the comorbidities vector generated by a linear layer,
 542 and h_f is the final representation that is fed into the output layer with a
 543 sigmoid activation function for binary classification. The main mortality
 544 prediction predicate ($HighMortality$) is modeled using a Bidirectional Long
 545 Short-Term Memory (BiLSTM) network that processes temporal sequences
 546 of clinical events, combined with a linear layer for static comorbidity features.
 547 Each clinical risk factor predicate (e.g., $HighLactate$, $LowPlatelets$) is im-
 548 plemented as a multi-layer perceptron (MLP) with two hidden linear layers
 549 with 64 hidden units, and a final output layer with a sigmoid activation
 550 function. Each risk factor predicate receives as input the corresponding clin-
 551 ical parameter sequences over the observation window, along with patient

demographics and comorbidities to provide essential clinical context. The architecture enables seamless integration of temporal dynamics with logical constraints.

To prevent predicate collapse and maintain semantic grounding, we employ weak anchoring techniques. Each clinical predicate is anchored to established clinical thresholds (e.g., lactate > 4.0 mmol/L for *HighLactate*) while allowing context-sensitive learning based on patient-specific factors including age, comorbidities, and concurrent clinical measurements. This approach ensures that predicates retain clinically meaningful interpretations while adapting to complex clinical scenarios.

4.6. *Technical Innovations Beyond Standard LTN*

While our framework builds upon Logic Tensor Networks, we introduce several critical innovations that enable clinical deployment:

Weak Anchoring for Clinical Predicates: Standard LTN predicates lack semantic grounding, potentially learning arbitrary functions. We introduce weak anchoring that initializes predicates with clinical thresholds while allowing context-sensitive refinement. This ensures *HighLactate*(x) remains interpretable (anchored to >4.0 mmol/L) while adapting to patient-specific factors (age, comorbidities).

Dual Axiom Architecture: Unlike standard LTN applications with uniform constraints, we separate data axioms (K_D) and knowledge axioms (K_K) with differentiated weights (0.8/0.2). This addresses the fundamental tension between fitting data and respecting domain knowledge in life-critical applications.

Automated Rule Discovery with Clinical Validation: We bridge

577 unstructured guidelines and formal logic through: (1) SNOMED CT-based
578 KG construction, (2) AnyBURL rule mining with confidence filtering, (3)
579 FOL translation with predicate mapping. This pipeline is reusable across
580 clinical domains, not just sepsis.

581 5. Case Study: MIMIC-IV Dataset

582 5.1. Dataset Description

583 The data utilized in this study were obtained from the MIMIC-IV v2.2, a
584 large, freely accessible, de-identified EHR database compiled from data col-
585 lected at the Beth Israel Deaconess Medical Center. It encompasses clinical
586 information from patients admitted to the emergency department or inten-
587 sive care units between 2008 and 2019 [19]. The data include a wealth of
588 clinical information, such as vital signs, laboratory measurements, medica-
589 tions, diagnoses, procedures, and de-identified clinical notes. One author of
590 this study has successfully completed the ethics training to gain access to the
591 database (certification number: 13158972).

592 The study population was selected from patients admitted to the ICU.
593 The initial cohort included all adult patients (age ≥ 18 years) who met
594 the Sepsis-3 criteria, defined as a suspected or documented infection and
595 an acute increase in the Sequential Organ Failure Assessment (SOFA) score
596 of ≥ 2 points. Patients were excluded if any of the following criteria were
597 met: (1) ICU length of stay less than 24 hours, (2) Multiple ICU admissions
598 during the hospital stay, or (3) Absence of significant clinical data required
599 for the analysis. The final cohort comprised 19,328 patients for training and
600 validating the neuro-symbolic model, with a mortality rate of 18%.

601 The final dataset comprises both static and dynamic patient features.
602 Static features include demographic and admission-related information, specif-
603 ically age and admission location. Dynamic features, representing time-
604 varying physiological and laboratory measurements, include: systolic blood
605 pressure, diastolic blood pressure, MAP, temperature, heart rate, respiratory
606 rate, creatinine, hemoglobin, platelet count, bilirubin, potassium, albumin,
607 CRP, glucose, lactate, lymphocytes, neutrophils, WBC count, alanine amino-
608 transferase, aspartate aminotransferase, and GCS scores.

609 To obtain patient comorbidities, we processed unstructured clinical notes
610 using spaCy, an open-source natural language processing framework. We de-
611 veloped a custom pipeline, incorporating NER and rule-based pattern match-
612 ing to identify clinically relevant comorbid conditions.

613 5.2. *Experimental Setup*

614 5.2.1. *Prediction Windows and Evaluation Protocol*

615 Following the approach proposed in [20], we defined the prediction task
616 based on multiple lead times: 6, 12, 24, and 48 hours prior to death. For each
617 lead time, we constructed a 24-hour observation window ending exactly at the
618 specified interval before the outcome event (death). This window captures
619 the temporal sequence of clinical events leading up to, but not including, the
620 lead time boundary, thus avoiding label leakage.

621 For negative cases (patients who survived the hospital stay), we randomly
622 selected a 24-hour window from their ICU stay to serve as the observation
623 period. To obtain robust estimates of model performance and minimize bias
624 introduced by any single train-test split, we employed a stratified 5-fold cross-
625 validation. We stratified on the binary target label to preserve the original

distribution. In each iteration, one fold served as the test set, while the remaining folds were pooled for training. Within each training fold, we further used 20% of the data as a validation set for hyperparameter tuning and early stopping.

5.2.2. Baseline Models and Implementation Details

We compared our NeSy-SMP framework against several established baseline approaches to demonstrate the effectiveness of neuro-symbolic integration:

- **Random Forest (RF)**: A tree-based ensemble method with 100 estimators and a maximum depth of 10
- **XGBoost**: A gradient boosting framework with a learning rate of 0.1, maximum depth of 6, and 100 estimators
- **BiLSTM**: A two-layer bidirectional LSTM network with a learning rate of 0.001, 128 hidden units, and a dropout rate of 0.1
- **LTN**: Standard Logic Tensor Network without weak anchoring and knowledge axioms. We used the same model backbone, a BiLSTM combined with a linear layer for the comorbidities vector, to model the main classification predicate *HighMortality*. The model is trained using only the data axioms $\mathcal{K}_{\mathcal{D}}$
- **NeSy-SMP**: Our proposed framework with weak anchoring and axiom-weighted loss.

The code to reproduce our experiments is available at <https://anonymous.4open.science/r/NeSy-SMP-68C8>.

649 5.2.3. Evaluation Metrics

650 Model performance was assessed using standard classification metrics,
651 including accuracy, precision, recall, F1-score, and area under the ROC curve
652 (AUC).

653 5.3. Performance Evaluation

654 The experimental results demonstrate the progressive benefit of incorpo-
655 rating domain knowledge into neural learning. Our comparison includes tra-
656 ditional machine learning baselines (Random Forest, XGBoost), deep learn-
657 ing approaches (BiLSTM), and neuro-symbolic variants (LTN, NeSy-SMP).

658 The results in Table 1 indicate that NeSy-SMP achieves competitive dis-
659 crimination and modest but consistent improvements on several clinically
660 relevant metrics across all prediction windows when evaluated at a fixed de-
661 cision threshold of 0.5. This threshold reflects a standard, model-agnostic
662 operating point and provides insight into how naturally each model sepa-
663 rates low- and high-risk patients without threshold optimization. Given the
664 imbalance of sepsis-related mortality and the clinical priority of not missing
665 high-risk patients, we focus on F1-score and recall for the mortality class,
666 while still reporting AUC as a key measure of overall discrimination. Un-
667 der this setting, our approach consistently achieves the highest F1-score and
668 recall values, reaching 80.35% F1-score at 6-hour prediction window, while
669 maintaining AUC values comparable to the strongest baselines. Although
670 traditional methods such as XGBoost and RF achieve higher precision and
671 accuracy, this comes at the cost of lower recall, implying more high-risk
672 patients may be missed at this default operating point. While different
673 trade-offs may be appropriate depending on deployment constraints, miss-

Table 1: Performance comparison of different models across different lead time windows using 5-fold cross-validation. The mean and the standard deviation [across folds](#) are reported for each metric. Bold values indicate the highest performing model for each lead time window.

Lead time window	Accuracy	F1-score	Precision	Recall	AUC
RF					
6h	84.78±0.19	74.77±0.56	83.07±0.44	71.48±0.62	88.11±0.50
12h	84.37±0.31	72.88±0.56	82.70±0.80	69.50±0.48	86.74±0.69
24h	83.59±0.19	69.46±0.61	80.30±1.00	66.35±0.63	85.13±0.70
48h	83.63±0.42	63.87±1.14	78.99±1.62	61.36±0.85	82.65±1.60
XGBoost					
6h	85.31±0.35	77.86±0.50	80.90±0.63	75.86±0.44	88.56±0.31
12h	84.39±0.27	75.78±0.43	79.12±0.54	73.72±0.45	86.49±0.23
24h	84.07±0.29	73.54±0.69	77.95±0.61	71.16±0.81	85.35±0.45
48h	83.62±0.50	69.27±1.18	74.61±1.10	66.81±1.12	82.34±0.84
BiLSTM					
6h	83.53±1.53	76.82±1.18	77.91±2.49	76.22±0.87	85.35±1.31
12h	83.37±0.68	75.19±0.57	77.35±1.41	73.77±0.75	83.91±0.81
24h	82.33±0.62	72.63±0.51	74.19±1.08	71.50±0.42	81.48±0.97
48h	81.54±1.28	69.10±1.32	70.57±1.98	68.19±1.50	79.15±2.17
LTN					
6h	85.65±0.48	79.20±0.38	80.63±1.18	78.06±0.52	88.10±1.07
12h	81.58±3.33	73.93±2.84	75.38±4.73	73.58±2.19	84.23±3.36
24h	81.99±3.04	73.58±1.86	74.73±3.45	73.57±1.04	83.91±0.91
48h	80.85±2.04	69.34±0.79	70.46±1.94	69.48±2.59	80.57±1.62
NeSy-SMP					
6h	86.45±0.47	80.35±0.45	81.86±0.97	79.14±0.47	88.33±0.74
12h	83.94±0.38	76.32±0.68	78.03±0.79	75.12±1.17	85.37±1.06
24h	82.93±0.59	74.71±1.21	74.98±0.89	74.55±1.71	84.58±0.71
48h	81.95±0.98	71.05±0.45	71.56±1.41	70.88±1.24	80.62±1.49

Table 2: Global performance of different models on all out-of-fold predictions from the 5-fold cross-validation on the MIMIC-IV dataset. Metrics were computed on the concatenated out-of-fold predictions, and 95% confidence intervals were estimated using 1,000 bootstrap resamples.

Lead time window	Accuracy	F1-score	Precision	Recall	AUC
XGBoost					
6h	85.31 [84.73-85.89]	64.99 [63.55-66.44]	72.20 [72.29-75.62]	58.10 [56.43-59.77]	88.43 [87.80-89.07]
12h	84.50 [83.91-85.09]	61.34 [59.86-62.83]	71.79 [70.01-73.58]	53.48 [51.74-55.22]	86.65 [85.95-87.36]
24h	84.00 [83.42-84.59]	56.67 [55.09-58.25]	69.04 [67.11-70.98]	48.02 [46.29-49.75]	85.28 [84.55-86.01]
48h	83.42 [82.79-84.05]	47.54 [45.67-49.41]	62.00 [59.71-64.30]	38.41 [36.52-40.30]	82.11 [81.17-83.04]
RF					
6h	84.07 [83.47-84.68]	54.12 [52.45-55.80]	84.04 [82.27-85.81]	39.94 [38.28-41.60]	87.60 [86.96-88.24]
12h	83.55 [82.96-84.15]	49.64 [47.98-51.31]	83.59 [81.62-85.56]	35.30 [33.71-36.89]	86.03 [85.33-86.74]
24h	82.94 [82.31-83.58]	42.44 [40.57-44.31]	79.84 [77.45-82.24]	28.92 [27.33-30.52]	84.43 [83.65-85.21]
48h	83.09 [82.48-83.70]	29.84 [27.71-31.97]	77.70 [74.46-80.95]	18.52 [17.00-20.04]	82.12 [81.22-83.01]
BiLSTM					
6h	84.35 [83.76-84.95]	64.30 [62.94-65.67]	69.42 [67.77-71.07]	59.90 [58.27-61.53]	85.68 [84.92-86.45]
12h	83.56 [82.91-84.21]	62.44 [61.06-63.83]	65.65 [64.00-67.31]	59.58 [57.91-61.26]	83.96 [83.13-84.79]
24h	81.49 [80.86-82.13]	54.40 [52.84-55.97]	58.56 [56.75-60.38]	50.84 [49.04-52.64]	80.03 [79.13-80.92]
48h	82.06 [81.37-82.74]	48.64 [46.91-50.37]	54.99 [52.96-57.02]	43.58 [41.70-45.47]	77.65 [76.62-78.68]
LTN					
6h	85.62 [85.04-86.21]	67.43 [66.14-68.72]	71.92 [70.30-73.55]	63.61 [62.07-65.16]	87.88 [87.24-88.53]
12h	84.69 [84.08-85.31]	63.85 [62.42-65.28]	69.59 [67.80-71.39]	59.06 [57.32-60.80]	86.34 [85.59-87.09]
24h	82.28 [81.65-82.91]	58.65 [57.18-60.13]	59.63 [57.85-61.41]	57.80 [56.08-59.53]	83.41 [82.59-84.22]
48h	81.53 [80.86-82.19]	52.17 [50.55-53.79]	52.60 [50.71-54.49]	51.80 [49.92-53.68]	80.23 [79.24-81.23]
NeSy-SMP					
6h	86.08 [85.54-86.62]	68.51 [67.08-69.95]	73.95 [72.29-75.62]	63.95 [62.41-65.49]	87.77 [87.08-88.47]
12h	84.42 [83.81-85.04]	65.68 [64.25-66.91]	66.61 [65.01-68.21]	64.75 [63.13-66.37]	86.34 [85.59-87.09]
24h	82.62 [82.00-83.25]	60.71 [59.34-62.07]	59.69 [58.04-61.33]	61.94 [60.34-63.54]	84.78 [84.04-85.51]
48h	82.06 [81.39-82.73]	54.59 [52.98-56.20]	53.86 [51.95-55.77]	55.33 [53.42-57.24]	81.66 [80.73-82.60]

ing high-risk patients is generally considered less acceptable in early-warning scenarios. These results suggest that NeSy-SMP offers a favorable balance between discrimination and sensitivity to sepsis-related mortality risk. The implications of these differences under clinically constrained operating points are further analyzed in Section 6.2.

To assess the robustness of these findings, we further estimate 95% con-

680 fidence intervals (CIs) for all evaluation metrics. Confidence intervals were
 681 estimated using non-parametric bootstrap resampling on pooled out-of-fold
 682 predictions (1,000 resamples). This procedure captures both data variability
 683 and model uncertainty while remaining consistent with the cross-validation
 684 protocol. The resulting CIs in Table 2 indicate that the observed improve-
 685 ments of NeSy-SMP in F1-score and recall are stable and not driven by a
 686 small subset of samples. In particular, for the 6-hour prediction window,
 687 NeSy-SMP shows stable advantages on F1-score (68.51 [67.08-69.95]) and
 688 recall (63.95 [62.41-65.49]) over XGBoost and baselines. These patterns per-
 689 sist with a 24-hour prediction window, where NeSy-SMP leads F1 (60.71
 690 [59.34-62.07]) and recall (61.94 [60.34-63.54]), with non-overlapping superi-
 691 ority to RF and competitive AUC across top models. While RF retains
 692 highest precision intervals, NeSy-SMP’s robust gains on mortality-sensitive
 693 metrics confirm its clinical value at the default threshold.

694 5.4. Ablation Analysis

695 To assess the effect of each component of NeSy-SMP, we conduct an
 696 ablation analysis considering three additional model configurations:

- 697 • **LTN-K**: The base LTN model augmented with the knowledge axioms
 698 $\mathcal{K}_K$ excluding both weak anchoring and weighted axiom loss
- 699 • **LTN-AK**: The base LTN model that incorporates the knowledge ax-
 700 ioms $\mathcal{K}_K$ together with weak anchoring, while omitting the weighted
 701 axiom loss
- 702 • **LTN-WK**: The base LTN model that combines the knowledge axioms
 703 $\mathcal{K}_K$ with the weighted axiom loss, but without weak anchoring

Table 3: Performance comparison of the main model and its ablated variants across different lead time windows using 5-fold cross-validation. The mean and standard deviation are reported for each metric. Bold values indicate the highest performing configuration.

Lead time window	Accuracy	F1-score	Precision	Recall	AUC
LTN					
6h	85.65±0.48	79.20±0.38	80.63±1.18	78.06±0.52	88.10±1.07
12h	81.58±3.33	73.93±2.84	75.38±4.73	73.58±2.19	84.23±3.36
24h	81.99±3.04	73.58±1.86	74.73±3.45	73.57±1.04	83.91±0.91
48h	80.85±2.04	69.34±0.79	70.46±1.94	69.48±2.59	80.57±1.62
LTN-K					
6h	84.92±0.63	78.17±0.79	79.63±1.34	77.16±1.21	85.08±1.40
12h	77.75±2.83	72.98±2.13	71.90±1.70	77.74±0.96	85.89±0.44
24h	81.28±0.79	72.70±1.15	72.65±1.16	72.95±1.79	80.82±2.37
48h	81.85±2.12	69.99±1.50	71.56±2.62	69.40±01	80.70±1.61
LTN-WK					
6h	85.19±0.82	77.35±1.67	81.23±1.75	75.25±2.26	87.21±0.60
12h	84.35±0.93	75.43±2.43	79.31±1.34	73.38±2.92	85.96±0.83
24h	82.93±0.99	73.86±0.68	75.27±1.87	72.96±0.94	83.74±0.40
48h	80.81±0.79	70.84±1.13	70.03±1.08	71.94±1.45	81.53±1.46
LTN-AK					
6h	23.51±0.01	19.03±0.01	11.75±0.01	50.00±0.00	77.37±1.18
12h	23.51±0.01	19.03±0.01	11.75±0.01	50.00±0.00	77.37±1.18
24h	23.51±0.01	19.03±0.01	11.75±0.01	50.00±0.00	77.37±1.18
48h	23.51±0.01	19.03±0.01	11.75±0.01	50.00±0.00	77.37±1.18
NeSy-SMP					
6h	86.45±0.47	80.35±0.45	81.86±0.97	79.14±0.47	88.33±0.74
12h	83.94±0.38	76.32±0.68	78.03±0.79	75.12±1.17	85.37±1.06
24h	82.93±0.59	74.71±1.21	74.98±0.89	74.55±1.71	84.58±0.71
48h	81.95±0.98	71.05±0.45	71.56±1.41	70.88±1.24	80.62±1.49

704 This analysis allows us to isolate the effects of weak anchoring and weighted
705 axioms, thereby clarifying their individual contributions to reasoning accu-
706 racy and learning performance. The results in Table 3 highlight the contribu-
707 tions of the different components of the proposed framework. The inclusion of
708 knowledge axioms without weak anchoring (LTN-K configuration) and with-
709 out weighted axiom loss leads to an overall decrease in performance compared
710 to the base LTN. Introducing predicates without semantic grounding that are
711 connected to the main predicate can inject noise that adversely affects the
712 classifier’s behavior. During training, such predicates may learn arbitrary
713 functions that either have no effect or decrease the classifier’s performance,
714 as reflected in the reported results.

715 When weighted axiom loss is introduced (LTN-WK configuration), data
716 axioms are assigned a higher weight than knowledge axioms, allowing them
717 to dominate the optimization process. This behavior is reflected in the re-
718 sults, where LTN-WK outperforms LTN-K and achieves performance com-
719 parable to the base LTN with 77.35% F1-score and 87.21% AUC for the
720 6-hour prediction window. However, due to the absence of weak anchoring,
721 the contribution of the knowledge axioms is insufficient to yield substantial
722 improvements over the base model. In this setting, the potential noise in-
723 troduced by the predicates modeling risk factors does not negatively impact
724 the main classifier, thanks to the higher weight assigned to the satisfaction
725 of data axioms.

726 Finally, when weak anchoring is introduced but equal weights are as-
727 signed to both data and knowledge axioms (LTN-AK configuration), an un-
728 expected behavior emerges. Since the optimization objective is to satisfy

the knowledge base, the model converges to the easiest solution that satisfies the constraints. In this case, this corresponds to trivially satisfying the knowledge axioms by collapsing the mortality predicate to a constant truth value of 1, independently of the truth degrees assigned to the predicates modeling risk factors. As evidenced by the results (19.03% F1-score and 50% recall), this leads to a degenerate solution in which the classifier always predicts the positive class, effectively neutralizing the classification-related constraints encoded in the data axioms.

This issue is resolved by prioritizing data axioms over knowledge axioms, resulting in the final NeSy-SMP configuration. Under this setting, weak anchoring enables the model to ground risk factor predicates in the data, while the weighted loss prevents trivial satisfaction of symbolic constraints. Consequently, the model is able to learn semantically meaningful representations of risk factors and exploit them as a structured inductive bias, improving classification performance without destabilizing the learning of the mortality predicate.

5.5. Explainability Analysis

Traditional AI models in healthcare often function as "black boxes," making it difficult for clinicians to understand why a particular prediction was made. While techniques like SHAP can show which features influenced a decision, they cannot explain the medical reasoning behind it. Our neuro-symbolic approach addresses this limitation by making the model's clinical reasoning transparent and interpretable.

Unlike conventional models, our LTN-based system learns explicit medical knowledge as logical rules, creating a transparent decision-making process.

754 Clinicians can "interrogate" the model to understand how clinical concepts
755 and logical rules contribute to each prediction. This interpretability is crucial
756 for building trust and enabling clinical adoption.

757 We demonstrate model transparency through three complementary anal-
758 yses: (1) verifying that clinical concepts are learned correctly, (2) validating
759 that medical rules are followed consistently, and (3) examining individual pa-
760 tient reasoning. These analyses ensure that the model's decisions align with
761 established clinical knowledge while revealing its decision-making process.

762 *5.5.1. Verifying Clinical Concept Learning*

763 The first level of explainability examines whether the model correctly
764 learns fundamental clinical concepts. We analyzed how the model interprets
765 key sepsis indicators: elevated lactate, high bilirubin, low platelet count, and
766 increased white blood cell count.

767 Figure 7 shows the model's confidence scores for these clinical concepts
768 before and after training. Initially, the model had no clinical understanding,
769 assigning random scores around 0.5 to all concepts. After training, the model
770 learned to recognize these clinical indicators with high accuracy, assigning
771 scores close to 1.0 when laboratory values exceeded dangerous thresholds.

772 Consider the HighLactate concept as a concrete example. Before train-
773 ing, the model assigned random truth values around 0.5 regardless of actual
774 lactate levels, shown as blue dots clustered horizontally. After training, the
775 model learned to distinguish dangerous lactate elevations. For deceased pa-
776 tients (top panel), lactate values above the clinical threshold of 4.0 mmol/L
777 consistently received high truth values (orange dots approaching 1.0), cor-
778 rectly identifying this as a high-risk indicator. Importantly, some deceased

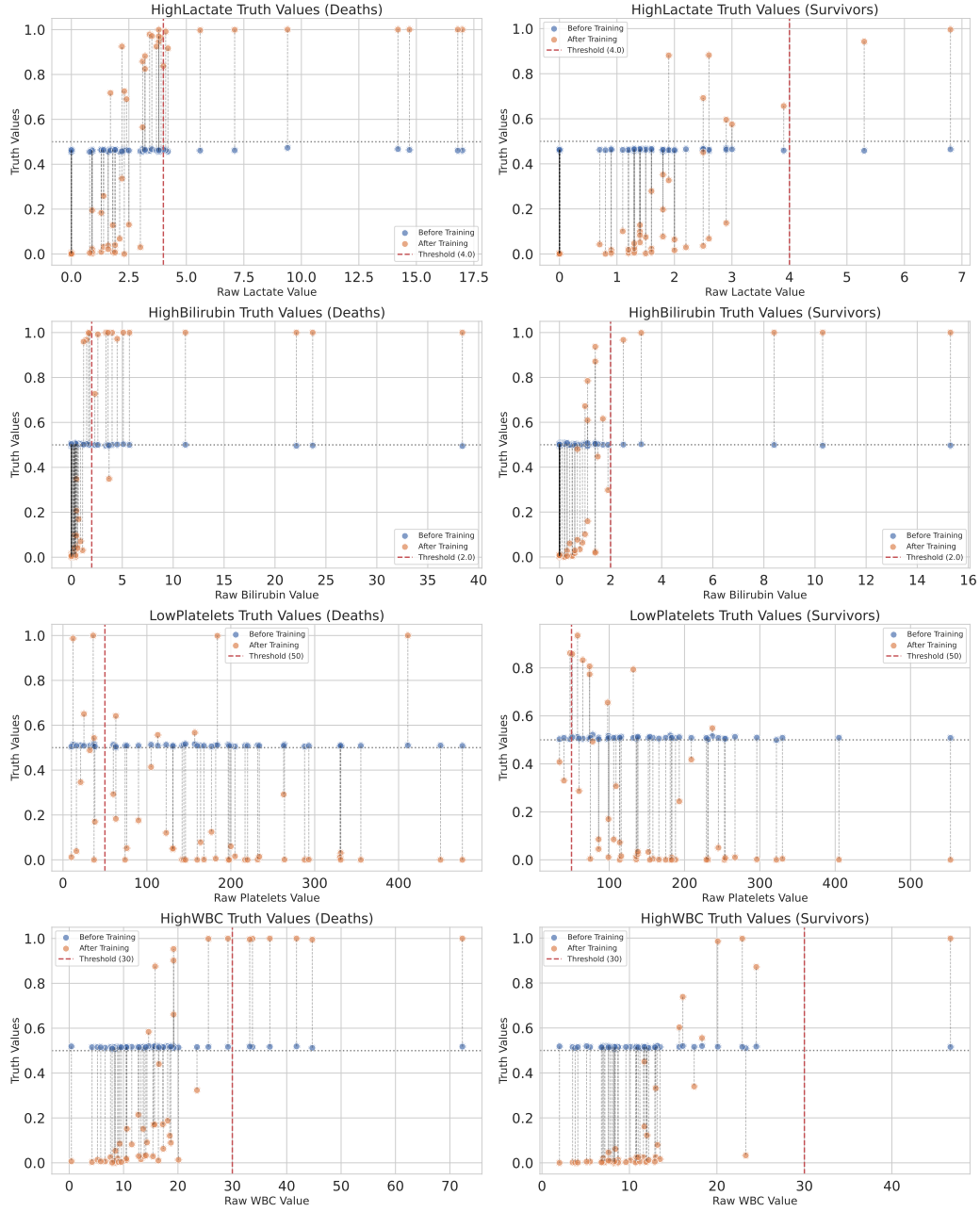


Figure 7: Clinical concept learning verification showing truth values for key sepsis indicators before and after training. The model successfully learns clinically meaningful thresholds for high lactate, high bilirubin, low platelets, and elevated white blood cells across 100 patients (50 deceased, 50 survivors).

779 patients with lactate levels near but below the 4.0 threshold still received
780 elevated truth values, demonstrating the model’s ability to consider patient-
781 specific context beyond rigid cutoffs. For survivors (bottom panel), most
782 patients received appropriately lower truth values, though some outliers with
783 high truth values appear even among survivors, suggesting the model consid-
784 ers additional patient-specific factors such as comorbidities, age, and other
785 clinical indicators when assessing lactate significance.

786 This example illustrates the model’s clinical sophistication, i.e., it learned
787 not just that “high lactate is bad”, but that lactate interpretation depends
788 on patient context and clinical trajectory. The model may flag moderately
789 elevated lactate as highly concerning in vulnerable patients while being more
790 tolerant of similar values in patients with better overall clinical profiles. This
791 adaptive behavior mirrors how experienced clinicians interpret laboratory
792 values in context, rather than relying solely on fixed cutoffs.

793 5.5.2. *Validating Medical Rule Consistency*

794 One of the key aspects of our explainability study is verifying that the
795 model has internalized the knowledge provided during training. In this way,
796 after training, we can check which rules the model believes in most strongly
797 and which ones are commonly violated. In this case, we can plot the truth
798 values of the knowledge axioms $\mathcal{K}_K$ for an entire test fold in relation to the
799 patient’s outcome to visualize the overall impact of injecting domain knowl-
800 edge on the main outcome. To achieve this, we can compute the truth values
801 for each patient and then [plot them](#) according to their ground truth labels.
802 Figure 8 shows the truth values of each knowledge axiom evaluated on each
803 patient of a test fold. We can observe a rather clear distinction between

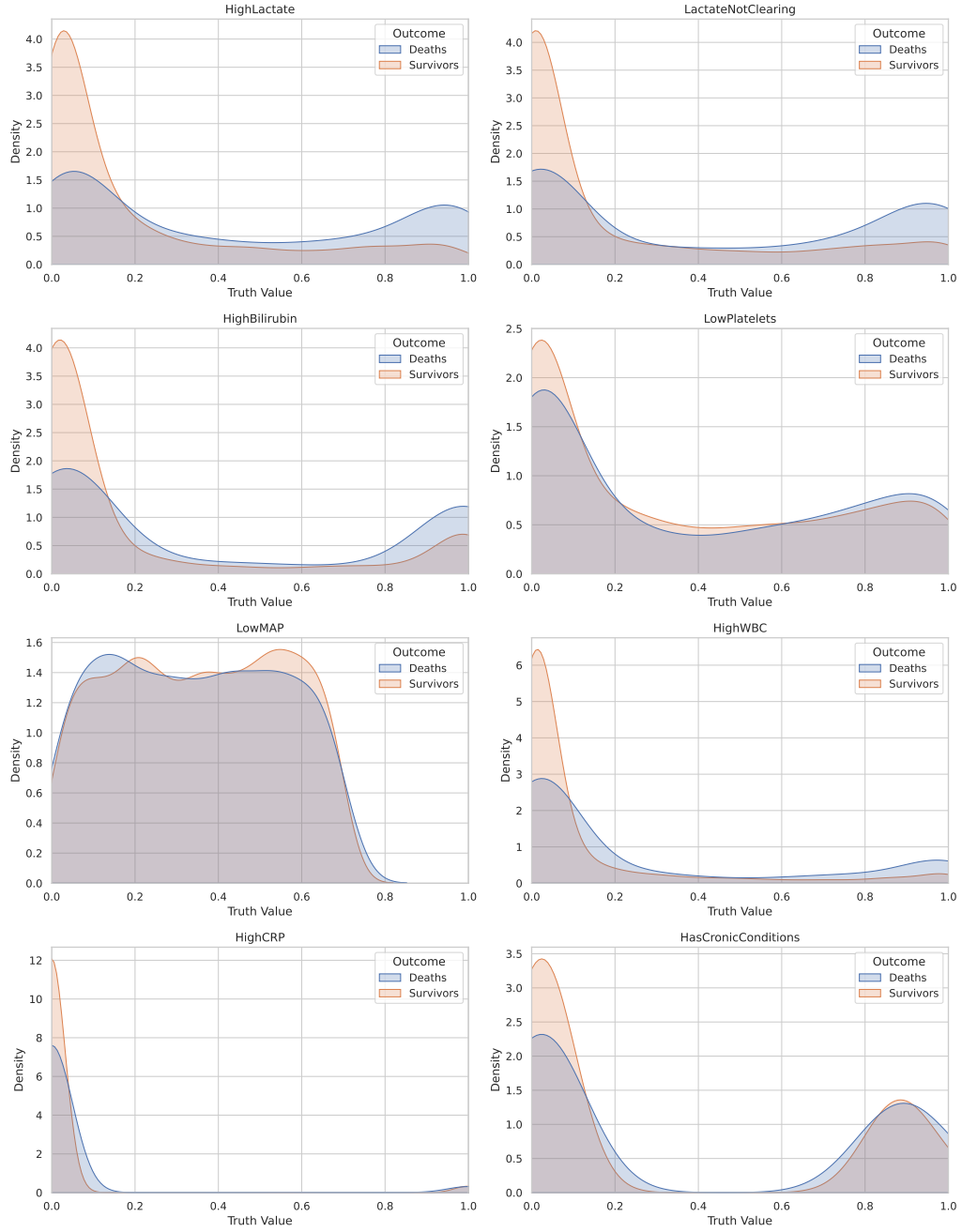


Figure 8: Medical rule validation showing how consistently the model follows clinical guidelines. The chart compares rule adherence between patients who died versus those who survived, demonstrating stronger rule activation in high-risk patients.

804 the two classes with respect to the risks associated with high lactate and
 805 bilirubin levels and high white blood cell count, while the difference is less
 806 pronounced in the case of low platelet counts and chronic conditions, but
 807 still visible. This suggests that the group of deceased patients in the ana-
 808 lyzed fold is generally characterized by higher levels of lactate and bilirubin
 809 in the observation window used to predict death, which is consistent with
 810 established clinical guidelines. However, *there are no instances in which the*
 811 *truth values of a risk factor are all, or nearly all, close to 1, which reflects*
 812 *the considerable* variability and heterogeneity *observed* among patients. Not
 813 all deceased patients exhibit elevated lactate levels or bilirubin levels, and
 814 conversely, patients with high lactate levels may still survive. However, this
 815 does not negatively impact the model’s behavior. Encoding these clinical
 816 concepts using neural networks rather than fixed hard thresholds allows the
 817 model to adapt their truth values based on patient-specific characteristics.
 818 Interestingly, no substantial differences were observed in the truth values of
 819 the risks associated with *high* levels of CRP and low levels of MAP, suggest-
 820 ing that, in this specified test fold, these factors did not significantly influence
 821 the final outcome.

822 *5.5.3. Understanding Individual Patient Reasoning*

823 The third level of explainability examines how the model makes decisions
 824 for individual patients, providing the detailed reasoning that clinicians need
 825 to trust and act on AI recommendations. We demonstrate this through two
 826 contrasting cases: a correctly predicted death and a misclassified survivor.

827 Figure 9 presents our clinical reasoning dashboard for two representa-
 828 tive patients. For the correctly predicted death, the patient had severely

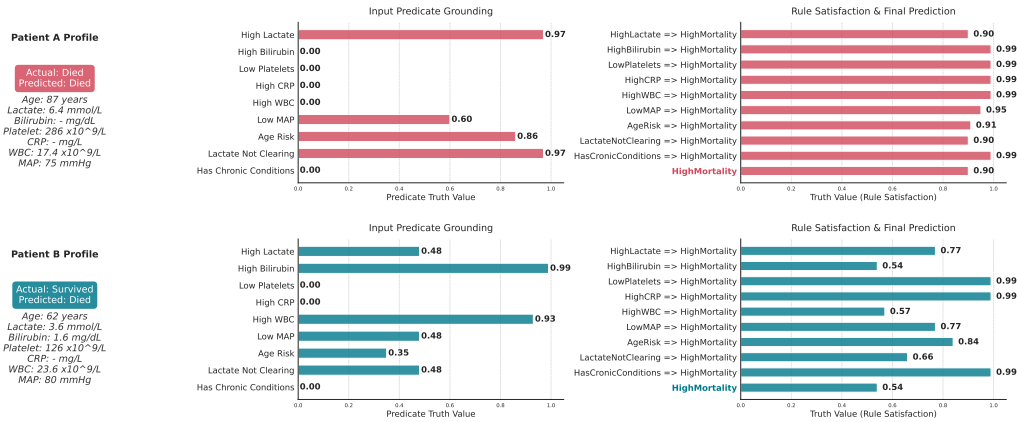


Figure 9: Individual patient reasoning analysis showing how clinical measurements translate into risk assessments and final predictions. The dashboard reveals the model’s step-by-step reasoning process for both correct and incorrect predictions.

829 elevated lactate (6.4 mmol/L), indicating significant metabolic dysfunction.
 830 The model recognized multiple high-risk indicators: lactate elevation (con-
 831 fidence: 97.3%), age-related risk (86.1%), and persistent lactate elevation
 832 (97.3%). These findings strongly supported the clinical rules linking organ
 833 dysfunction to mortality risk, resulting in a confident death prediction (90%
 834 probability).

835 The misclassified case reveals important insights about model limitations
 836 and clinical complexity. This patient had moderately elevated lactate (3.6
 837 mmol/L), high bilirubin (1.6 mg/dL), and very high white blood cells, but
 838 ultimately survived. The model identified several concerning indicators, but
 839 with lower confidence levels. The final prediction (54% death probability)
 840 was just above the decision threshold, indicating the model’s uncertainty.

841 This case-by-case analysis provides several clinical insights. Accurate pre-
 842 dictions typically show coherent patterns of high-risk indicators with strong

Table 4: Comparison of training time, inference time, and parameter complexity across models, illustrating the trade-off between computational efficiency and interpretability in clinical reasoning. Results are averaged across 5-fold cross-validation.

Model	Training time (s)	Inference time (s)	Number of parameters
XGBoost	0.12±0.03	0.0027±0.0002	-
RF	0.95±0.01	0.0206±0.0005	-
BiLSTM	64.47±6.19	0.2588±0.0030	1,355,010
LTN	105.98±35.61	0.2528±0.0057	1,355,010
NeSy-SMP	706.41±123.37	0.3022±0.0022	1,508,875

logical support. Uncertain cases often involve mixed signals or moderate risk factors that don’t clearly point toward a specific outcome. Importantly, even “incorrect” predictions can provide valuable clinical information - the second patient’s multiple abnormal values warranted close monitoring despite survival.

This transparent reasoning process enables clinicians to understand not just what the model predicted, but why it reached that conclusion, facilitating informed clinical decision-making that combines AI insights with human judgment.

5.6. Complexity Analysis

The integration of logical reasoning in clinical prediction models presents a fundamental trade-off between computational efficiency and interpretability. Table 4 shows training times, inference times computed on the held-out fold of each iteration, and the number of parameters of neural models. All experiments are conducted on a workstation equipped with an Intel Core i9-13900K CPU, an NVIDIA RTX 4090 with 24 GB of VRAM, and 128 GB of system RAM.

Traditional ML models like XGBoost and RF demonstrate remarkable training efficiency, but they provide limited insight into clinical decision-making processes. In contrast, NeSy-SMP requires substantially longer training times (706.41 ± 123.37 seconds). However, this computational cost enables the explicit encoding and learning of clinically meaningful rules, which can be directly interpreted and validated by healthcare practitioners. The standard BiLSTM architecture occupies an intermediate position (64.47 ± 6.19 seconds), offering deep learning capabilities without explicit logical reasoning, while LTN without knowledge demonstrates that the different training objective alone contributes minimally to training overhead. The primary computational burden stems from enforcing logical constraints during optimization.

When comparing LTN without knowledge to BiLSTM, the number of parameters remains the same, as both share the same backbone architecture. However, in NeSy-SMP, extra parameters are introduced by MLP components that model the clinical risk factor predicates. These additional parameters increase model capacity and allow domain rules to be explicitly represented in the learning process, but they also contribute to the higher computational load observed during training.

Despite these differences, inference time, computed on the held-out fold in the 5-fold cross-validation, remains relatively comparable across neural approaches (0.25-0.30 seconds), suggesting that the enhanced explainability does not compromise real-time clinical deployment. This trade-off is justified in high-stakes medical contexts where model interpretability, regulatory compliance, and clinician trust are fundamental, as the ability to

885 understand decisions through human-readable logical rules may justify the
886 one-time training cost.

887 Moreover, the construction of the clinical knowledge graph represents a
888 one-time upfront effort. Once created, the KG serves as a reusable and exten-
889 sible knowledge base, requiring updates only when new guidelines or clinical
890 evidence emerge. As a result, the computation and modeling effort associated
891 with knowledge modeling is amortized over time and should be viewed as an
892 initial investment that supports long-term model interpretability, adaptabil-
893 ity, and clinical relevance.

894 5.7. *External Validation on eICU*

895 To assess the generalizability of the proposed neuro-symbolic framework
896 beyond the development setting, we conduct external validation using the
897 eICU Collaborative Research Database [36]. We extract an independent
898 cohort using the same approach adopted for MIMIC-IV. The final cohort
899 includes 6,735 patients, of whom 5,913 are survivors, and 822 died (12.2%
900 mortality).

901 All models are trained exclusively on the MIMIC-IV cohort and eval-
902 uated on the independent eICU cohort without any retraining or domain
903 adaptation. This setting is challenging and should be interpreted as a test
904 of transportability rather than raw predictive ability.

905 The eICU dataset is a large multi-institutional critical care database that
906 aggregates records from over 200 hospitals, whereas MIMIC-IV represents
907 care from a single academic center. As a result, eICU encompasses greater
908 heterogeneity in clinical practice patterns, patient management protocols,
909 documentation practices, and data capture processes. Laboratory reporting

Table 5: Performance of all models trained exclusively on MIMIC-IV and evaluated without retraining on the independent eICU cohort. Reported values correspond to the mean performance across bootstrap resamples, with 95% confidence intervals estimated using non-parametric bootstrap resampling (1,000 iterations).

Lead time window	Accuracy	F1-score	Precision	Recall	AUC
XGBoost					
6h	62.82 [61.65-63.92]	52.70 [51.46-53.93]	56.63 [55.82-57.46]	65.05 [63.44-66.67]	72.00 [70.10-73.84]
12h	59.25 [58.17-60.46]	50.45 [49.30-51.60]	56.04 [55.25-56.86]	64.00 [62.35-65.72]	70.31 [68.38-72.25]
24h	69.79 [68.64-70.90]	56.35 [55.03-57.66]	57.35 [56.39-58.32]	65.04 [63.23-66.83]	70.18 [68.24-72.22]
48h	67.60 [66.50-68.67]	54.55 [53.35-55.74]	56.30 [55.40-57.14]	63.26 [61.45-64.96]	68.35 [66.40-70.17]
RF					
6h	64.25 [63.15-65.33]	53.52 [52.35-54.72]	56.81 [56.00-57.62]	65.24 [63.56-66.88]	72.18 [70.23-74.06]
12h	74.79 [73.75-75.77]	59.49 [58.06-60.77]	58.90 [57.80-59.91]	66.16 [64.32-67.78]	71.97 [69.95-73.85]
24h	78.89 [77.91-79.87]	60.80 [59.32-62.25]	59.60 [58.32-60.87]	64.30 [62.55-66.11]	69.99 [67.99-72.04]
48h	81.87 [80.95-82.78]	60.98 [59.48-62.47]	60.26 [58.85-61.66]	62.01 [60.30-63.70]	68.58 [66.58-70.70]
BiLSTM					
6h	78.31 [77.36-79.23]	58.52 [57.14-59.86]	57.67 [56.46-58.81]	60.92 [59.12-62.57]	67.58 [65.58-69.41]
12h	81.59 [80.76-82.46]	59.17 [57.65-60.66]	58.76 [57.34-60.24]	59.69 [58.08-61.31]	65.17 [63.15-67.30]
24h	86.64 [85.82-87.50]	55.81 [54.13-57.41]	62.60 [59.71-65.29]	54.84 [53.66-55.96]	63.04 [60.94-65.14]
48h	84.44 [83.56-85.33]	54.24 [52.72-55.70]	56.51 [54.45-58.60]	53.74 [52.53-54.91]	62.48 [60.46-64.49]
LTN					
6h	60.36 [59.26-61.51]	50.65 [49.49-51.80]	55.52 [54.69-56.30]	62.68 [60.94-64.37]	68.61 [66.64-70.54]
12h	76.50 [75.44-77.45]	57.12 [55.80-58.50]	56.51 [55.38-57.68]	59.96 [58.31-61.63]	64.98 [62.90-66.97]
24h	80.80 [79.85-81.80]	57.46 [56.08-58.92]	57.12 [55.79-58.61]	57.90 [56.39-59.54]	62.71 [60.70-64.72]
48h	76.13 [75.12-77.18]	58.71 [57.36-60.14]	57.95 [56.86-59.14]	63.11 [61.37-64.94]	68.14 [65.96-70.13]
NeSy-SMP					
6h	80.00 [79.06-80.97]	61.26 [59.76-62.67]	60.05 [58.72-61.30]	64.03 [62.17-65.79]	69.46 [67.10-71.59]
12h	78.45 [77.49-79.39]	59.88 [58.50-61.39]	58.82 [57.62-60.09]	63.09 [61.39-64.84]	65.93 [63.57-68.18]
24h	82.20 [81.31-83.13]	57.09 [55.61-58.49]	57.34 [55.83-58.92]	56.88 [55.47-58.28]	64.09 [62.01-66.18]
48h	74.55 [73.48-75.61]	56.40 [55.10-57.83]	56.10 [55.05-57.24]	60.19 [58.45-62.00]	65.35 [63.25-67.40]

910 frequency is generally sparser and more variable in eICU than in MIMIC-IV,
911 and overall mortality and intervention rates differ across these data sources
912 partly because of case mix and variation of case intensity [39].

913 Table 5 reports the results of the external validation. Across all models
914 and lead times, discrimination and precision-recall decrease compared to in-
915 ternal validation, which is expected given the substantial differences between
916 MIMIC-IV and eICU. Importantly, all models maintain AUC values largely
917 in the 62-72% range, indicating preserved ranking ability despite calibration
918 and threshold-related degradation.

919 XGBoost and RF consistently achieve the highest AUC across most lead
920 times, with RF showing robust F1-scores at longer horizons (12-48h). This
921 suggests that ensemble methods relying on aggregated tabular patterns gen-
922 eralize better under the domain shift, likely due to their ability to capture
923 coarse but stable feature interactions that are less sensitive to site-specific
924 temporal dynamics.

925 Temporal neural models exhibit a sharp drop in discrimination and precision-
926 recall performance as lead time increases. This pattern suggests that models
927 relying on richly sampled sequential signals are more susceptible to differ-
928 ences in measurement frequency and documentation patterns characteristic of
929 multicenter datasets like eICU. NeSy-SMP maintains competitive F1-scores,
930 achieving the highest external F1 at a 6-hour prediction horizon. This sug-
931 gests that the rule-based components may provide decision stability under
932 distributional change even when global ranking ability is attenuated. Unlike
933 internal validation trends, performance does not uniformly deteriorate with
934 increasing lead time, reflecting differences in care pathways and interven-

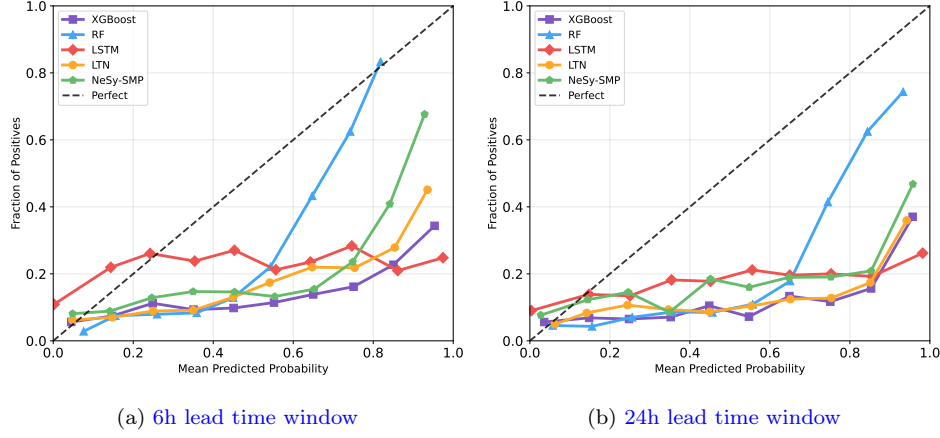


Figure 10: Calibration curves computed on the eICU dataset for 6h and 24h prediction horizons.

tion timing in eICU, where early physiological signals may be noisier or less standardized.

In addition to the observed decline in discrimination, external validation reveals systematic miscalibration across all models. As shown in the calibration curves for the 6-hour and 24-hour prediction horizons (Figure 10a and Figure 10b), predicted risks are consistently overestimated, with calibration curves lying below the diagonal across most probability ranges. This behavior indicates a lack of calibration from MIMIC-IV to eICU. Importantly, this miscalibration is coherent across models and lead times, suggesting that it is primarily driven by differences in outcome prevalence, case, mix, and measurement practices between the two datasets rather than by model-specific failure. As no recalibration was applied, these results reflect the true out-of-distribution behavior of the trained models and highlight the challenges of deploying probability estimates across heterogeneous clinical settings. Re-

949 calibration techniques such as Platt scaling or isotonic regression were in-
950 tentionally omitted to preserve the unmodified out-of-distribution behavior
951 of each model; however, applying such methods in a prospective deployment
952 setting could substantially improve the reliability of predicted risk estimates.

953
954 The rule activation analysis provides additional insight into the observed
955 degradation in external performance and clarifies which clinical signs are
956 most affected by the dataset shift. In the MIMIC-IV cohort, rules such as
957 *AgeRisk* and *HasChronicCondition* are frequently activated, reflecting well-
958 established associations between age, comorbidity burden, and mortality that
959 are readily observable in a single-center setting with dense clinical documen-
960 tation. In contrast, in the eICU dataset, the activation frequency of several
961 rules, particularly those relying on chronic conditions, inflammatory markers
962 (e.g., *HighCRP*), and dynamic laboratory trends such as *LactateNotClearing*,
963 is substantially reduced compared to MIMIC-IV, with *HasChronicCondi-*
964 *tion* decreasing from 32.6% to 3.0% of samples and *LactateNotClearing* from
965 16.1% to 5.1%. The latter is especially sensitive to the availability of lac-
966 tate measurements, which are less consistently recorded across institutions in
967 eICU. This divergence in rule activation patterns is likely driven by structural
968 characteristics of the eICU dataset, including heterogeneous coding practices
969 and variability in laboratory sampling frequency. Consequently, the logical
970 components encoded in NeSy-SMP are unable to fully express their intended
971 inductive bias, limiting their potential advantages during external validation
972 and reflecting constraints imposed by the structure and completeness of the
973 external dataset.

974 6. Discussion

975 6.1. Summary of Key Findings

976 This study evaluated LTNs for real-time sepsis mortality prediction across
977 multiple time horizons (6h, 12h, 24h, 48h), comparing the standard LTN,
978 knowledge-augmented LTN (NeSy-SMP), and established baselines (XGBoost,
979 RF, BiLSTM).

980 When evaluated on MIMIC-IV using 5-fold cross-validation at a fixed de-
981 cision threshold of 0.5 (Table 1), NeSy-SMP consistently achieved the highest
982 F1-score and recall for the mortality class, reaching 80.35% F1-score at the
983 6-hour prediction window, while maintaining AUC values comparable to the
984 strongest baselines and logical interpretability through rules encoding clin-
985 ical knowledge. This demonstrates that the knowledge-augmented model
986 naturally separates patients without threshold tuning. Tree-based models
987 achieved higher precision, but at the expense of recall, meaning more high-
988 risk patients would be missed at this default operating point. While these
989 approaches achieve high discrimination, their outputs provide limited guid-
990 ance for clinicians.

991 NeSy-SMP addresses this gap by integrating domain knowledge using log-
992 ical rules, yielding competitive or superior recall and F1 at standard thresh-
993 olds. Unlike prior methods, it also provides transparent justification for each
994 prediction, improving trust and potential adoption in ICU settings.

995 To further analyze the contribution of symbolic components, we con-
996 ducted an ablation analysis to isolate the contribution of weak anchoring
997 and weighted axiom loss (Table 3). The results reveal that naively augment-
998 ing the base LTN with knowledge axioms does not improve and can even

999 degrade performance, as unanchored predicates may inject noise and lead
1000 to less stable learning dynamics. Introducing a weighted axiom loss miti-
1001 gates this issue by allowing data axioms to dominate optimization. However,
1002 without weak anchoring, the contribution of symbolic knowledge remains
1003 limited. Conversely, adding weak anchoring but no weighting causes the
1004 model to overprioritize satisfaction of the knowledge axioms, collapsing to a
1005 degenerate solution that always predicts the positive class. These findings
1006 highlight that effective integration of symbolic knowledge requires a careful
1007 balance between data fidelity and logical consistency.

1008 Beyond internal validation, we evaluated all models under a strict exter-
1009 nal validation setting, training exclusively on MIMIC-IV and testing without
1010 retraining on the independent eICU cohort. This scenario represents a chal-
1011 lenging but realistic evaluation of transportability across datasets that differ
1012 substantially in institutional heterogeneity, outcome prevalence, and data
1013 collection practices. As expected, performance decreased across all models
1014 relative to internal validation. However, tree-based models and NeSy-SMP
1015 retained moderate discrimination and competitive F1-scores. Importantly,
1016 several logical rules that contribute during training on MIMIC-IV are ac-
1017 tivated less frequently in eICU. As a consequence, the logical components
1018 embedded in NeSy-SMP cannot be systematically leveraged to improve per-
1019 formance in the evaluated external setting. However, the existing logical rules
1020 continue to provide valuable explanations for individual predictions, allowing
1021 clinicians to trace model outputs back to recognizable clinical concepts.

6.2. Clinical Impact

To translate statistical performance into bedside relevance, we evaluated models using pooled out-of-fold predictions from 5-fold cross-validation at four operating points reflecting real-world trade-offs for a medium-sized hospital managing 500 sepsis patients per year (100 deaths, 400 survivors) using the 24-hour prediction window [29, 35, 48]. For each model independently, we identified the threshold that achieves fixed 85% sensitivity (recall), prioritizing high-mortality detection [33], or fixed specificity at 70%, 88%, and 90% levels, corresponding to moderate/strict alert budgets of 120, 47, and 40 false positives/year among 400 survivors in a 500-patient/year hospital. These operating points represent plausible clinical policies ranging from aggressive early warning to conservative alerting designed to avoid fatigue. Specifically, the 88% and 90% specificity thresholds address documented alarm fatigue concerns by limiting false alerts to maintain clinician responsiveness and engagement [11].

In the high-sensitivity regime (fixed sensitivity = 85%, identifying approximately 85 of 100 patients who will die within 24 hours), all models achieve comparable death detection but differ substantially in false alert burden. The BiLSTM generates 190 false positives per year, while NeSy-SMP reduces this to 137 false positives per year. This represents 53 fewer unnecessary alerts annually compared to BiLSTM and 4 fewer than RF, nearly matching XGBoost’s 129 false positives per year. For clinicians, this translates to fewer workflow interruptions per year compared to traditional deep learning models while maintaining equivalent mortality detection, directly addressing alarm fatigue concerns prevalent in ICU settings.

1047 Under constrained alert budgets, NeSy-SMP demonstrates particular strength.
1048 At 70% specificity (tolerable 120 false positives per year), XGBoost identi-
1049 fies 84 true high-risk deaths while NeSy-SMP identifies 82, a difference of 2
1050 patients annually. NeSy-SMP substantially outperforms BiLSTM (75 true
1051 positives, +7 patients per year) and LTN (80 positives, +2 patients), achiev-
1052 ing competitive detection with the added benefit of rule-based explanations
1053 at manageable alert rates.

1054 This advantage becomes more pronounced under stricter budgets. At
1055 88% specificity (allowing up to 48 false positives per year), NeSy-SMP cor-
1056 rectly identifies 63 true high-risk deaths compared to XGBoost and RF (61
1057 patients, a gain of 2), BiLSTM (55 patients, a gain of 8), and LTN (60 pa-
1058 tients, a gain of 3). When configured to maintain 90% specificity (allowing
1059 at most 40 false positives), NeSy-SMP identifies 58 true high-risk patients,
1060 matching the performance of XGBoost and Random Forest, which identifies
1061 57 patients. BiLSTM and LTN identify fewer high-risk patients (51 and 56
1062 high-risk patients, respectively). Under conservative alerting policies, these
1063 gains translate into the identification of additional high-risk patients without
1064 increasing alert burden.

1065 NeSy-SMP’s graceful performance degradation under increasingly strict
1066 constraints, combined with its logical explanations, can provide superior ac-
1067 tionability.

1068 In a validation case where standard guidelines indicated moderate risk
1069 (SOFA score=9, APACHE II=13) and other models failed to flag high mor-
1070 tality risk, NeSy-SMP correctly classified this patient as high-risk (predicted
1071 probability 0.70). The model activated rules such as *HasChronicCondition*

and *AgeRisk*, justifying escalation despite relatively modest guideline scores. The patient then died within the next 24 hours, validating NeSy-SMP’s prediction and demonstrating how its clinical knowledge integration identifies high-risk patients that both traditional scoring systems and black-box models miss.

Overall, NeSy-SMP achieves competitive performance with tree-based baselines across operating points, substantially reducing deep learning alert burden, and incorporates transparent clinical reasoning. In practical terms, this translates to dozens more correctly identified high-risk patients and hundreds fewer alarms per year compared to alternative approaches, while providing clinicians with logical rationales that enhance trust and support timely decision-making in safety-critical ICU environments.

6.3. Limitations and Threats to Validity

While our study demonstrates promising results, several limitations must be acknowledged. First, although we report an external validation, the models are trained on a single-center dataset (MIMIC-IV), which may limit generalizability across institutions with different patient populations, documentation practices, and data availability. Moreover, the temporal distribution of the MIMIC-IV cohort (2008-2019) may not fully reflect current clinical protocols and treatment standards. Second, the manual curation and expert validation of the knowledge graph, while ensuring clinical accuracy, introduces potential bias and limits scalability to other domains or updated clinical guidelines. Third, the current framework does not address regulatory considerations for AI deployment in clinical settings, including FDA approval pathways or integration with existing hospital information systems.

1097 **Fourth**, the computational complexity of the LTN framework, while manage-
1098 able for research purposes, may pose challenges for real-time deployment in
1099 resource-constrained clinical environments.

1100 Internal validity is also affected by sources of noise and bias in both the
1101 data and modeling strategy. EHR-derived sepsis labels may contain record-
1102 ing errors, missing timestamps, or inconsistent criteria across cases. Simi-
1103 larly, rule-based components in the LTN framework may encode expert bias
1104 or oversimplify complex clinical reasoning, potentially influencing model be-
1105 havior and interpretability. Reported metrics also depend on fixed thresholds
1106 and operating points, and performance could vary under alternative decision
1107 criteria. Furthermore, the assumption that clinicians consistently interpret
1108 and act upon rule-based explanations may not hold in practice, potentially
1109 altering model impact.

1110 7. Conclusion

1111 In this work, we presented a neuro-symbolic approach based on LTNs for
1112 sepsis mortality prediction, leveraging logical rules extracted from a clinical
1113 knowledge graph encoding fundamental clinical concepts. Experimental re-
1114 sults obtained on a retrospective cohort from MIMIC-IV demonstrated that
1115 the injection of domain knowledge can enhance the classification performance
1116 compared to pure data-driven models, achieving 86.45% accuracy and 80.35%
1117 F1-score for a 6-hour prediction window.

1118 A key advantage of the proposed framework lies in its ability to jointly
1119 enhance predictive performance and model interpretability. Through the ex-
1120 plainability analysis, we showed how we can query the model to retrieve

1121 grounded representations of clinical concepts and inspect how knowledge
1122 axioms contribute to the final outcome. From a clinical perspective, our
1123 framework addresses the critical need for interpretable AI in critical care set-
1124 tings. The ability to trace predictions back to specific clinical concepts (e.g.,
1125 elevated lactate, organ dysfunction) enables clinicians to validate AI recom-
1126 mendations against their domain expertise, potentially accelerating adoption
1127 of AI-assisted decision support in sepsis management. These findings demon-
1128 strate the potential of neuro-symbolic artificial intelligence in mortality pre-
1129 diction tasks and suggest that this approach can encourage the integration
1130 of domain knowledge in other clinical prediction tasks as well.

1131 Despite these advantages, this work has several limitations. The knowl-
1132 edge graph and logical rules were constructed based on available clinical
1133 knowledge and retrospectively curated data, which may limit completeness
1134 and introduce bias. Moreover, the training and the internal validation were
1135 conducted on a single-center dataset, potentially affecting the generalizability
1136 of the results to other clinical settings.

1137 Future work will focus on addressing these limitations by enhancing the
1138 automatic construction of the knowledge graph from clinical guidelines and
1139 other relevant sources. To this end, integrating LLMs may be promising, as
1140 they have already demonstrated significant potential when combined with
1141 knowledge graphs. Additionally, we plan to explore automatic optimization
1142 strategies for the knowledge base, where logical rules are selectively retained
1143 or discarded based on their contribution to performance improvements.

1144 From an explainability perspective, an important future direction consists
1145 in combining the intrinsic interpretability of the proposed neuro-symbolic

framework with classical post-hoc XAI techniques, such as SHAP. This combination would enable a systematic comparison between knowledge-driven explanations and data-driven explanations, as well as their joint use to provide complementary insights into model behavior. Finally, a key future direction will be to train the model across multiple clinical centers, leveraging techniques such as federated learning or domain adaptation to mitigate center-specific biases and improve robustness and generalizability across heterogeneous clinical settings.

Ethics statement

This study does not require ethics approval as it involves retrospective analysis of the publicly available, de-identified MIMIC-IV and eICU datasets. Both datasets have been previously approved for research use and contain no identifiable patient information. Access to the datasets was obtained through appropriate credentialing and completion of required ethics training.

References

- [1] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., 2020a. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics Decis. Mak. 20, 310. URL: <https://doi.org/10.1186/s12911-020-01332-6>, doi:10.1186/S12911-020-01332-6.
- [2] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., 2020b. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics Decis. Mak. 20,

- 1169 310. URL: <https://doi.org/10.1186/s12911-020-01332-6>, doi:10.
1170 1186/S12911-020-01332-6.
- 1171 [3] Badreddine, S., d’Avila Garcez, A.S., Serafini, L., Spranger, M., 2022.
1172 Logic tensor networks. *Artif. Intell.* 303, 103649. URL: <https://doi.org/10.1016/j.artint.2021.103649>, doi:10.1016/J.ARTINT.
1173 //doi.org/10.1016/j.artint.2021.103649, doi:10.1016/J.ARTINT.
1174 2021.103649.
- 1175 [4] Belmonte, G., Ciancia, V., Massink, M., 2025. Symbolic and hybrid AI
1176 for brain tissue segmentation using spatial model checking. *Artif. Intell.*
1177 *Medicine* 167, 103154. URL: [https://doi.org/10.1016/j.artmed.](https://doi.org/10.1016/j.artmed.2025.103154)
1178 2025.103154, doi:10.1016/J.ARTMED.2025.103154.
- 1179 [5] Bhuyan, B.P., Ramdane-Cherif, A., Tomar, R., Singh, T.P., 2024.
1180 Neuro-symbolic artificial intelligence: a survey. *Neural Com-*
1181 *put. Appl.* 36, 12809–12844. URL: [https://doi.org/10.1007/](https://doi.org/10.1007/s00521-024-09960-z)
1182 s00521-024-09960-z, doi:10.1007/S00521-024-09960-Z.
- 1183 [6] Carriero, A., de Hond, A., Cappers, B., Paulovich, F., Abeln, S., Moons,
1184 K.G., van Smeden, M., 2025. Explainable ai in healthcare: to explain,
1185 to predict, or to describe? *Diagnostic and prognostic research* 9, 29.
- 1186 [7] Carvalho, R.M.S., Oliveira, D., Pesquita, C., 2023. Knowledge
1187 graph embeddings for ICU readmission prediction. *BMC Medical*
1188 *Informatics Decis. Mak.* 23, 12. URL: [https://doi.org/10.1186/](https://doi.org/10.1186/s12911-022-02070-7)
1189 s12911-022-02070-7, doi:10.1186/S12911-022-02070-7.
- 1190 [8] Chadaga, K., Prabhu, S., Bhat, V., Sampathila, N., Umakanth, S.,

- 1191 Chadaga, R., 2023. Artificial intelligence for diagnosis of mild–moderate
1192 covid-19 using haematological markers. *Annals of medicine* 55, 2233541.
- 1193 [9] Chadaga, K., Prabhu, S., Umakanth, S., Sampathila, N., Chadaga, R.,
1194 et al., 2021. Covid-19 mortality prediction among patients using epi-
1195 demiological parameters: an ensemble machine learning approach. *En-
1196 gineered Science* 16, 221–233.
- 1197 [10] Chang, C., Wang, X., Yang, C.C., 2024. Explainable AI for fair sepsis
1198 mortality predictive model, in: Finkelstein, J., Moskovitch, R., Parim-
1199 belli, E. (Eds.), *Artificial Intelligence in Medicine - 22nd International
1200 Conference, AIME 2024, Salt Lake City, UT, USA, July 9-12, 2024, Pro-
1201 ceedings, Part II*, Springer. pp. 267–276. URL: [https://doi.org/10.](https://doi.org/10.1007/978-3-031-66535-6_29)
1202 [1007/978-3-031-66535-6_29](https://doi.org/10.1007/978-3-031-66535-6_29), doi:10.1007/978-3-031-66535-6_29.
- 1203 [11] Chromik, J., Klopfenstein, S.A.I., Pfitzner, B., Sinno, Z., Arnrich, B.,
1204 Balzer, F., Poncette, A., 2022. Computational approaches to alleviate
1205 alarm fatigue in intensive care medicine: A systematic literature review.
1206 *Frontiers Digit. Health* 4. URL: [https://doi.org/10.3389/fdgth.](https://doi.org/10.3389/fdgth.2022.843747)
1207 [2022.843747](https://doi.org/10.3389/fdgth.2022.843747), doi:10.3389/FDGTH.2022.843747.
- 1208 [12] Chung, D., 2025. Artificial intelligence in healthcare and medicine
1209 technology development review. *Eng. Appl. Artif. Intell.* 143, 109801.
1210 URL: <https://doi.org/10.1016/j.engappai.2024.109801>, doi:10.
1211 [1016/J.ENGAPPAI.2024.109801](https://doi.org/10.1016/j.engappai.2024.109801).
- 1212 [13] Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith,
1213 C.M., French, C., Machado, F.R., McIntyre, L., Ostermann, M.,

- 1214 Prescott, H.C., et al., 2021. Surviving sepsis campaign: international
1215 guidelines for management of sepsis and septic shock 2021. Critical care
1216 medicine 49, e1063–e1143.
- 1217 [14] Feng, X., Zhu, S., Shen, Y., Zhu, H., Yan, M., Cai, G., Ning, G., 2024.
1218 Multi-organ spatiotemporal information aware model for sepsis mor-
1219 tality prediction. Artif. Intell. Medicine 147, 102746. URL: <https://doi.org/10.1016/j.artmed.2023.102746>, doi:10.1016/J.ARTMED.
1220 2023.102746.
1221
- 1222 [15] Garcia-Barragán, A., Sakor, A., Vidal, M., Menasalvas, E., Gonzalez,
1223 J.C.S., Provencio, M., Robles, V., 2025. NSSC: a neuro-symbolic AI
1224 system for enhancing accuracy of named entity recognition and linking
1225 from oncologic clinical notes. Medical Biol. Eng. Comput. 63, 749–
1226 772. URL: <https://doi.org/10.1007/s11517-024-03227-4>, doi:10.
1227 1007/S11517-024-03227-4.
- 1228 [16] Ghosh, S.K., Khandoker, A.H., 2024. Investigation on explainable ma-
1229 chine learning models to predict chronic kidney diseases. Scientific Re-
1230 ports 14, 3687.
- 1231 [17] Hu, C., Li, L., Li, Y., Wang, F., Hu, B., Peng, Z., 2022. Explainable
1232 machine-learning model for prediction of in-hospital mortality in septic
1233 patients requiring intensive care unit readmission. Infectious Diseases
1234 and Therapy 11, 1695–1713.
- 1235 [18] Huang, X., Marques-Silva, J., 2024. On the failings of shap-
1236 ley values for explainability. Int. J. Approx. Reason. 171, 109112.

- 1237 URL: <https://doi.org/10.1016/j.ijar.2023.109112>, doi:10.1016/
1238 J.IJAR.2023.109112.
- 1239 [19] Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A.,
1240 Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al., 2023.
1241 Mimic-iv, a freely accessible electronic health record dataset. Scientific
1242 data 10, 1.
- 1243 [20] Kim, S.Y., Kim, S., Cho, J., Kim, Y.S., Sol, I.S., Sung, Y., Cho, I.,
1244 Park, M., Jang, H., Kim, Y.H., et al., 2019. A deep learning model for
1245 real-time mortality prediction in critically ill children. Critical care 23,
1246 1–10.
- 1247 [21] Koch, C., Edinger, F., Fischer, T., Brenck, F., Hecker, A., Katzer, C.,
1248 Markmann, M., Sander, M., Schneck, E., 2020. Comparison of qsofa
1249 score, sofa score, and sirs criteria for the prediction of infection and mor-
1250 tality among surgical intermediate and intensive care patients. World
1251 Journal of Emergency Surgery 15, 1–10.
- 1252 [22] Kong, G., Lin, K., Hu, Y., 2020. Using machine learning methods to
1253 predict in-hospital mortality of sepsis patients in the ICU. BMC Medical
1254 Informatics Decis. Mak. 20, 251. URL: <https://doi.org/10.1186/s12911-020-01271-2>, doi:10.1186/S12911-020-01271-2.
1255
- 1256 [23] Lan, G., Hu, M., Li, Y., Zhang, Y., 2023. Contrastive knowledge in-
1257 tegrated graph neural networks for chinese medical text classification.
1258 Eng. Appl. Artif. Intell. 122, 106057. URL: <https://doi.org/10.1016/j.engappai.2023.106057>, doi:10.1016/J.ENGAPPAI.2023.106057.
1259

- 1260 [24] Leiser, F., Rank, S., Schmidt-Kraepelin, M., Thiebes, S., Sunyaev, A.,
1261 2023. Medical informed machine learning: A scoping review and future
1262 research directions. *Artif. Intell. Medicine* 145, 102676. URL: <https://doi.org/10.1016/j.artmed.2023.102676>, doi:10.1016/J.ARTMED.
1263 2023.102676.
1264
- 1265 [25] Li, C., Zhang, Z., Ren, Y., Nie, H., Lei, Y., Qiu, H., Xu, Z., Pu, X.,
1266 2021. Machine learning based early mortality prediction in the emer-
1267 gency department. *International journal of medical informatics* 155,
1268 104570.
- 1269 [26] Liu, H., Xiangfei, Z., Hang, Y., Zhang, Q., 2025. Dual-channel batch
1270 constrained deep q-learning for sepsis treatment. *Eng. Appl. Artif. In-*
1271 *tell.* 159, 111545. URL: [https://doi.org/10.1016/j.engappai.2025.](https://doi.org/10.1016/j.engappai.2025.111545)
1272 111545, doi:10.1016/J.ENGAPPAI.2025.111545.
- 1273 [27] Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., Acharya,
1274 U.R., 2022. Application of explainable artificial intelligence for health-
1275 care: A systematic review of the last decade (2011-2022). *Comput.*
1276 *Methods Programs Biomed.* 226, 107161. URL: [https://doi.org/10.](https://doi.org/10.1016/j.cmpb.2022.107161)
1277 1016/j.cmpb.2022.107161, doi:10.1016/J.CMPB.2022.107161.
- 1278 [28] Lundberg, S.M., Lee, S., 2017. A unified approach to interpreting model
1279 predictions , 4765–4774 URL: [https://proceedings.neurips.cc/](https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html)
1280 [paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.](https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html)
1281 [html](https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html).
- 1282 [29] Markwart, R., Saito, H., Harder, T., Tomczyk, S., Cassini, A.,

- 1283 Fleischmann-Struzek, C., Reichert, F., Eckmanns, T., Allegranzi, B.,
1284 2020. Epidemiology and burden of sepsis acquired in hospitals and in-
1285 tensive care units: a systematic review and meta-analysis. *Intensive care*
1286 *medicine* 46, 1536–1551.
- 1287 [30] Mehdiyev, N., Majlatow, M., Fettke, P., 2025. Integrating permuta-
1288 tion feature importance with conformal prediction for robust explainable
1289 artificial intelligence in predictive process monitoring. *Eng. Appl. Ar-*
1290 *tif. Intell.* 149, 110363. URL: [https://doi.org/10.1016/j.engappai.](https://doi.org/10.1016/j.engappai.2025.110363)
1291 [2025.110363](https://doi.org/10.1016/j.engappai.2025.110363), doi:10.1016/J.ENGAPPAI.2025.110363.
- 1292 [31] Minne, L., Abu-Hanna, A., de Jonge, E., 2008. Evaluation of sofa-based
1293 models for predicting mortality in the icu: A systematic review. *Critical*
1294 *care* 12, R161.
- 1295 [32] Murali, L., Gopakumar, G., Viswanathan, D.M., Nedungadi, P., 2023.
1296 Towards electronic health record-based medical knowledge graph con-
1297 struction, completion, and applications: A literature study. *J. Biomed.*
1298 *Informatics* 143, 104403. URL: [https://doi.org/10.1016/j.jbi.](https://doi.org/10.1016/j.jbi.2023.104403)
1299 [2023.104403](https://doi.org/10.1016/j.jbi.2023.104403), doi:10.1016/J.JBI.2023.104403.
- 1300 [33] Nemati, S., Holder, A., Razmi, F., Stanley, M.D., Clifford, G.D., Buch-
1301 man, T.G., 2018. An interpretable machine learning model for accurate
1302 prediction of sepsis in the icu. *Critical care medicine* 46, 547–553.
- 1303 [34] Padmalal, S., Babu, C.N.K., 2026. A novel lightweight multi-path
1304 learning with explainable artificial intelligence technique based on ovar-
1305 ian cancer detection framework. *Eng. Appl. Artif. Intell.* 163, 112772.

1306 URL: <https://doi.org/10.1016/j.engappai.2025.112772>, doi:10.
1307 1016/J.ENGAPPAI.2025.112772.

1308 [35] Paoli, C.J., Reynolds, M.A., Sinha, M., Gitlin, M., Crouser, E., 2018.
1309 Epidemiology and costs of sepsis in the united states—an analysis based
1310 on timing of diagnosis and severity level. *Critical care medicine* 46, 1889–
1311 1897.

1312 [36] Pollard, T.J., Johnson, A.E., Raffa, J.D., Celi, L.A., Mark, R.G.,
1313 Badawi, O., 2018. The eicu collaborative research database, a freely
1314 available multi-center database for critical care research. *Scientific data*
1315 5, 1–13.

1316 [37] Reed, S.K., Pease, A., 2017. Reasoning from imperfect knowledge. *Cogn.*
1317 *Syst. Res.* 41, 56–72. URL: [https://doi.org/10.1016/j.cogsys.](https://doi.org/10.1016/j.cogsys.2016.09.006)
1318 2016.09.006, doi:10.1016/J.COGSYS.2016.09.006.

1319 [38] Rudd, K.E., Johnson, S.C., Agesa, K.M., Shackelford, K.A., Tsoi, D.,
1320 Kievlan, D.R., Colombara, D.V., Ikuta, K.S., Kissoon, N., Finfer, S.,
1321 et al., 2020. Global, regional, and national sepsis incidence and mor-
1322 tality, 1990–2017: analysis for the global burden of disease study. *The*
1323 *Lancet* 395, 200–211.

1324 [39] Sauer, C.M., Dam, T.A., Celi, L.A., Faltys, M., de la Hoz, M.A., Ad-
1325 hikari, L., Ziesemer, K.A., Girbes, A., Thorat, P.J., Elbers, P., 2022.
1326 Systematic review and comparison of publicly available icu data sets—a
1327 decision guide for clinicians and data scientists. *Critical care medicine*
1328 50, e581–e588.

- 1329 [40] Shenderovitz, G., Sheetrit, E., Nissim, N., 2024. Patterns of time-
1330 interval based patterns for improved multivariate time series data classi-
1331 fication. Eng. Appl. Artif. Intell. 133, 108171. URL: [https://doi.org/](https://doi.org/10.1016/j.engappai.2024.108171)
1332 [10.1016/j.engappai.2024.108171](https://doi.org/10.1016/j.engappai.2024.108171), doi:10.1016/J.ENGAPPAI.2024.
1333 108171.
- 1334 [41] Sirocchi, C., Bogliolo, A., Montagna, S., 2024. Medical-
1335 informed machine learning: integrating prior knowledge into med-
1336 ical decision systems. BMC Medical Informatics Decis. Mak. 24,
1337 186. URL: <https://doi.org/10.1186/s12911-024-02582-4>, doi:10.
1338 1186/S12911-024-02582-4.
- 1339 [42] Smedt, J.D., Weerdt, J.D., 2024. Predictive process model monitoring
1340 using long short-term memory networks. Eng. Appl. Artif. Intell. 133,
1341 108295. URL: <https://doi.org/10.1016/j.engappai.2024.108295>,
1342 doi:10.1016/J.ENGAPPAI.2024.108295.
- 1343 [43] Su, L., Xu, Z., Chang, F., Ma, Y., Liu, S., Jiang, H., Wang, H., Li, D.,
1344 Chen, H., Zhou, X., et al., 2021. Early prediction of mortality, severity,
1345 and length of stay in the intensive care unit of sepsis patients based on
1346 sepsis 3.0 by machine learning models. Frontiers in Medicine 8, 664966.
- 1347 [44] Valik, J.K., Ward, L., Tanushi, H., Johansson, A.F., Färnert, A., Mo-
1348 gensen, M.L., Pickering, B.W., Herasevich, V., Dalianis, H., Henriksson,
1349 A., et al., 2023. Predicting sepsis onset using a machine learned causal
1350 probabilistic network algorithm based on electronic health records data.
1351 Scientific reports 13, 11760.

- [45] Vazifehdoostirani, M., Buliga, A., Genga, L., Francescomarino, C.D., Ronzani, M., Ghidini, C., Dijkman, R.M., 2025. Outcome-oriented local explanation using context-aware process patterns. *Eng. Appl. Artif. Intell.* 159, 111387. URL: <https://doi.org/10.1016/j.engappai.2025.111387>, doi:10.1016/J.ENGAPPAI.2025.111387.
- [46] Vivek Khanna, V., Chadaga, K., Sampathila, N., Prabhu, S., Chadaga P, R., Bhat, D., KS, S., 2024. Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers. *Cogent Engineering* 11, 2330266.
- [47] Wadden, J.J., 2022. Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics* 48, 764–768.
- [48] Walkey, A.J., Wiener, R.S., 2014. Hospital case volume and outcomes among patients hospitalized with severe sepsis. *American journal of respiratory and critical care medicine* 189, 548–555.
- [49] Xie, F., Yuan, H., Ning, Y., Ong, M.E.H., Feng, M., Hsu, W., Chakraborty, B., Liu, N., 2022. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *J. Biomed. Informatics* 126, 103980. URL: <https://doi.org/10.1016/j.jbi.2021.103980>, doi:10.1016/J.JBI.2021.103980.
- [50] Yang, H., Li, J., Zhang, C., Sierra, A.P., Shen, B., 2025. Large language model-driven knowledge graph construction in sepsis care using multi-

- 1374 center clinical databases: Development and usability study. Journal of
1375 Medical Internet Research 27, e65537.
- 1376 [51] Yong, L., Zhenzhou, L., 2024. Deep learning-based prediction of in-
1377 hospital mortality for sepsis. Scientific Reports 14, 372.
- 1378 [52] Zhang, L., Wang, Z., Zhou, Z., Li, S., Huang, T., Yin, H., Lyu, J., 2022.
1379 Developing an ensemble machine learning model for early prediction of
1380 sepsis-associated acute kidney injury. Iscience 25.
- 1381 [53] Zhou, G., E, H., Kuang, Z., Tan, L., Xie, X., Li, J., Luo, H., 2022.
1382 Clinical decision support system for hypertension medication based on
1383 knowledge graph. Comput. Methods Programs Biomed. 227, 107220.
1384 URL: <https://doi.org/10.1016/j.cmpb.2022.107220>, doi:10.1016/
1385 J.CMPB.2022.107220.