

Evaluating Ontologically-Aware Large Language Models: An Experiment in Sepsis Prediction

Lucas Gomes Maddalena^{a,*}, Fernanda Araujo Baião^a, Tiago Prince Sales^b and Giancarlo Guizzardi^b

^a *Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil*

E-mails: lucasmadda@aluno.puc-rio.br, fbaiiao@puc-rio.br

^b *Semantics, Cybersecurity and Services, University of Twente, Enschede, The Netherlands*

E-mails: t.princesales@utwente.nl, g.guizzardi@utwente.nl

Abstract. Early and accurate detection of sepsis during hospitalization is critical, as it is a life-threatening condition with significant implications for patient outcomes. Electronic Health Records (EHRs) offer a wealth of information, including unstructured textual data, often containing more nuanced insights than regular structured data. To process such textual data, a variety of Natural Language Processing (NLP) methods have been employed with limited effectiveness. Recent advancements in computational resources have led to the development of Large Language Models (LLMs), which can effectively process vast amounts of text to identify relationships and patterns between words and structure them into embeddings. This enables LLMs to extract meaningful insights within specific domains. Despite these advances, LLMs face challenges in capturing the real-world semantics of clinical texts, which are critical for understanding the complex interconnections among terms and ensuring terminological precision. This work proposes a case study using Clinical KB BERT, an approach for embedding clinical notes of ICU patients that incorporates semantic information from the Unified Medical Language System (UMLS) ontology. By integrating domain-specific knowledge from UMLS, Clinical KB BERT aims to improve the semantic understanding of clinical data, thus enhancing the predictive performance of the resulting models. The present study compares Clinical KB BERT against Clinical BERT, a widely used model in the healthcare domain. The experimental results demonstrate that semantically enriched embeddings produced a more accurate and less uncertain model for the early prediction of sepsis. Specifically, it increased the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) from 0.826 to 0.853, while the mean predictive entropy for the entire test dataset decreased from 0.159 to 0.142. Furthermore, the reduction in mean predictive entropy was even more pronounced in cases where both models made correct predictions, decreasing from 0.148 to 0.129. Noteworthy, the practical impacts of these improvements include a substantial decrease in the number of false negatives (from 162 to 128, out of 227 septic cases), emphasizing the ability of the semantically aware model in reducing missed early diagnoses, and improving patient outcomes.

Keywords: LLM, BERT, Semantically Enriched LLMs

*Corresponding author. E-mail: lucasmadda@aluno.puc-rio.br.

1. Introduction

Sepsis is a life-threatening condition that occurs when the body’s immune system overreacts to an infection, leading to multi-organ failure [1]. Research has demonstrated that early detection of sepsis, followed by the prompt initiation of antibiotic treatment, significantly improves patient outcomes [2]. Yet, despite recent advances, its early detection remains a challenging and unresolved problem in clinical practice and research. The use of Machine and Deep Learning (ML) methods to build prediction models reflecting recurrent patterns in hospitalized patients that further developed sepsis from historical data has been effective; however, existing models often struggle with issues of accuracy and understandability, highlighting the need for further progress in this field [3].

Electronic Health Records (EHRs) provide a valuable source of information for sepsis prediction, encompassing both structured data (such as vital signs and laboratory results) and unstructured clinical notes from healthcare professionals. While structured data is easier to process, unstructured free-text notes require advanced natural language processing (NLP) techniques and, more recently, large language models (LLMs) have emerged as powerful tools to extract insights from these rich textual sources [3].

Textual embedding techniques are a vital component of Large Language Models. They provide a numerical representation of text that is supposed to capture its meaning, allowing the model to understand and generate human-like language. By converting words, sentences, or even entire documents into vectors in a continuous vector space, these embeddings enable LLMs to process and analyze text more efficiently. These embeddings are then used as input for various downstream tasks, such as summarization, question answering, and text classification [4].

The different techniques for producing text embeddings differ in the size of the vocabulary, their approach for building vector representations (frequency-based and/or prediction-based), and their ability to address sub-word info and out-of-vocabulary (OOV) words [5]. The size of the vocabulary and the dimensionality of the embeddings greatly influence the quality and utility of the resultant vectors. These techniques may be broadly classified as frequency-based or prediction-based. Frequency-based methods focus on word counts and document co-occurrence to build vector representations. In contrast, prediction-based methods, such as neural network models, use the surrounding words to predict a target word, thus learning the representation. Subword information handling accounts for parts of words—such as character n-grams—to better capture the meaning of morphologically complex words and effectively deal with out-of-vocabulary (OOV) terms. This approach is particularly valuable in languages with rich morphology or specialized domains, where new terms frequently emerge. Some notable advances in this context include Bidirectional Encoder Representations from Transformers (BERT) [6] - which attempt to incorporate contextual nuances by considering the surrounding words in a sentence – and the Clinical BERT [7] - which is trained in the Medical Information Mart for Intensive Care (MIMIC) III database clinical notes [8].

Despite the significant advances in applying Language Models, they still face challenges related to semantics, which is the focus of this work. Text embeddings may not effectively convey the complex semantic connections between terms, especially in specialized fields like biomedicine or customer analysis, where precise terminology is crucial [9]. Additionally, addressing the occurrence of hallucinations in the context of Language Models remains a challenge. Hallucinations refer to instances where a model generates outputs that deviate from factual accuracy, resulting in nonsensical or irrelevant responses. Integrating structured real-world knowledge from ontologies into Language Models can help reduce these hallucinations. By grounding the model’s generative processes in a network of factual relationships and

entities, we can guide the model towards more factually consistent and contextually relevant responses, ultimately enhancing the trustworthiness of its output.

Recent literature evaluates the benefits of enriching semantics not only with textual data but also leveraging structured knowledge representations, such as Knowledge Graphs, RDFs, taxonomies, and ontologies, with findings pointing toward significant enhancements in various machine learning applications [10–12]. Such studies have shown promising results in improving the accuracy and depth of data interpretation in several domains, including risk assessment [13–16], emergence capability [17], and knowledge-intensive processes [18].

To face the challenges of leveraging structured and unstructured clinical data for sepsis prediction, this study explores the integration of ontologies into language models by means of three research questions (RQ). First, it investigates how embedding structured domain knowledge, such as ontologies, can enhance the performance and reliability of predictive models (RQ1). Second, it suggests how to measure the benefits of using semantically enriched embeddings, particularly in terms of improving predictive accuracy and reducing uncertainty (RQ2). Third, the study examines whether incorporating structured semantic information can help mitigate critical misclassifications, such as false negatives, which are particularly detrimental for sepsis prediction (RQ3).

Table 1
Research Questions

RQ1
How embedding structured domain knowledge, such as ontologies, can enhance the performance and reliability of predictive models?
RQ2
How to measure the benefits of using semantically enriched embeddings, particularly in terms of improving predictive accuracy and reducing uncertainty?
RQ3
Does the incorporation of structured semantic information help mitigate critical misclassifications, such as false negatives, which are particularly detrimental for sepsis prediction?

These research questions, summarized in Table 1, aim to assess the potential of ontologically enriched embeddings to address current limitations in clinical NLP tasks.

To address the three research questions, this work describes an experiment to evaluate the impact of semantically enriched embeddings on the early prediction of sepsis in ICU patients. Specifically, we compare Clinical BERT, a baseline language model for clinical text, with our proposed pipeline (henceforth named Clinical KB BERT), which incorporates domain-specific knowledge from the Unified Medical Language System (UMLS) ontology. By embedding semantic information, Clinical KB BERT addresses the limitations of capturing the semantics of medical terms mentioned in EHRs. Our findings demonstrate that the proposed approach leads to significant improvements, including higher prediction accuracy, measured in terms of both AUC-ROC and Matthews correlation coefficient (MCC) and reduced prediction uncertainty, measured using entropy-based methods [19].

This work is structured as follows. Section 3 introduces Clinical BERT and Clinical KB BERT, highlighting their respective architectures and features. Section 4 details the Materials and Methods applied to implement the gated recurrent unit (GRU) architecture used as the foundation of both predictive models, alongside an explanation of the MIMIC-III dataset and data preprocessing steps that were performed.

Section 5 presents the experimental findings, including a detailed analysis of specific case groups, providing insight into how semantic enrichment improves prediction performance and model interpretability.

2. Theoretical Background

2.1. Sepsis Prediction

Sepsis is a life-threatening condition that arises when the body’s immune system overreacts to an infection, causing widespread inflammation and potentially leading to multi-organ failure and death. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) define sepsis as a “life-threatening organ dysfunction caused by a dysregulated host response to infection” [1]. Organ dysfunction is typically quantified using the Sequential Organ Failure Assessment (SOFA) score, with an increase of 2 or more points indicating a higher risk of mortality.

Early detection and intervention are critical for improving patient outcomes, as studies have demonstrated that prompt administration of antibiotics and other treatments significantly reduces mortality rates [2]. Despite advances in medical practice, the timely prediction of sepsis remains a challenging problem, owing to its heterogeneous presentation and the complex interplay of physiological, laboratory, and clinical factors involved in its diagnosis.

2.2. Sepsis Labeling Criteria and Temporal Considerations

To label sepsis cases in this study, we follow the Sepsis-3 definition using a SOFA score increase criterion [8]. The labeling process consists of the following steps:

- *SOFA Score Calculation per Hour*: For each intensive care unit (ICU) stay, SOFA scores are computed hourly based on organ dysfunction categories: respiratory, coagulation, liver, cardiovascular, central nervous system, and renal functions.
- *Sepsis Onset Time Determination*: A patient is labeled as septic if their SOFA score increases by at least 2 points within the observation window. The exact sepsis onset time is calculated as:

$$\text{Admission Time} + \text{Hours from Admission when SOFA} \geq 2 \quad (1)$$

- *Time Window Constraints*: Only SOFA scores within the suspicion of infection (SI) window are considered to ensure proper alignment with clinical sepsis diagnosis. SOFA calculations exclude ICU stays with negative or invalid hospital length of stay values ($\text{SOFA}_{\text{hlos}} \geq 0$).
- *First Sepsis Onset Filtering*: If multiple sepsis events are detected in a single ICU stay, only the first occurrence is retained to avoid redundant labels.
- *Data Storage and Usage*: The processed sepsis labels are integrated into the study’s database for further analysis and model training.

This Sepsis-3-compliant labeling strategy ensures the dataset aligns with clinical diagnostic standards, allowing for accurate and meaningful machine learning model training and evaluation.

2.3. Sepsis Prediction and Challenges

Machine learning models have emerged as promising tools for sepsis prediction, leveraging structured data such as vital signs and laboratory results alongside unstructured data from clinical notes. However, existing approaches often struggle with high false negative rates, which can delay critical interventions. Furthermore, many models lack the capacity to fully utilize the semantic richness of clinical text, limiting their ability to capture complex temporal and contextual relationships that are crucial for accurate predictions.

This study focuses on improving sepsis prediction by integrating ontological knowledge into the model, aiming to enhance its semantic understanding of clinical data. By leveraging the Unified Medical Language System (UMLS) ontology, the approach seeks to bridge the gap between structured and unstructured data, providing a robust framework for early and reliable sepsis detection.

There is a growing body of work focused on augmenting Large Language Models (LLMs) with structured knowledge, reflecting the increasing importance and momentum of this rapidly evolving field [?]. Notable examples include CoDEX [20], which introduces an innovative model-agnostic loss function enabling Knowledge Graph Embedding (KGE) models to leverage the textual context of LLMs for more accurate triple likelihood estimation. Similarly, in [21], embeddings are enhanced by combining hierarchical LLM-generated representations—spanning word, sentence, and document levels—with graph embeddings using Quaternion-based 4D hypercomplex vectors. Another significant contribution involves integrating vision and graph encoders to create multimodal knowledge graph embeddings, generating substantial improvements in downstream task performance [22]. Additionally, shared transformer-based encoders jointly optimize knowledge graph embedding objectives and masked token pretraining, aligning language processing with knowledge graph tasks [23].

Despite all these advances, Sepsis prediction continues to be a challenging problem, as shown in [?], where the authors review machine learning approaches combining structured and unstructured data for this purpose. While these approaches improved the identification and early detection of sepsis, none of them explores the potential of structured knowledge integration, such as ontologies, to address this issue. Instead, the reliance on standard NLP techniques limits their ability to capture the semantic relationships present in clinical text. This gap creates an opportunity for approaches like the one presented in this study to leverage ontological knowledge for enhanced interpretability and predictive performance.

The above-mentioned works illustrate the potential of integrating structured knowledge with LLMs. The contribution of our study is twofold: first, while they primarily target general-purpose applications or multimodal knowledge integration, we emphasize the integration of domain-specific ontologies to enhance semantic understanding in predicting sepsis; second, while existing studies often utilize knowledge graphs, our study leverages the Unified Medical Language System (UMLS) ontology to embed structured clinical knowledge into the language model directly. Thus, our approach applies ontological integration specifically to improve predictive accuracy and reduce uncertainty in clinical NLP tasks.

3. Related Work

There is a growing body of work focused on augmenting Large Language Models (LLMs) with structured knowledge, reflecting the increasing importance and momentum of this rapidly evolving field [12]. Notable examples include CoDEX [20], which introduces an innovative model-agnostic loss function enabling Knowledge Graph Embedding (KGE) models to leverage the textual context of LLMs for more

accurate triple likelihood estimation. Similarly, in [21], embeddings are enhanced by combining hierarchical LLM-generated representations—spanning word, sentence, and document levels—with graph embeddings using Quaternion-based 4D hypercomplex vectors. Another significant contribution involves integrating vision and graph encoders to create multimodal knowledge graph embeddings, generating substantial improvements in downstream task performance [22]. Additionally, shared transformer-based encoders jointly optimize knowledge graph embedding objectives and masked token pretraining, aligning language processing with knowledge graph tasks [23].

Beyond knowledge graph integration, structured domain knowledge has been explored in machine learning pipelines to improve model generalization and adherence to clinical guidelines. (author?) [24] highlight the necessity of incorporating clinical knowledge into medical machine learning models, demonstrating that such integration enhances accuracy, interpretability, and data efficiency. Their study maps knowledge integration strategies across different stages of the machine learning pipeline, underscoring the benefits of incorporating medical domain rules, causal networks, and structured intervals in predictive modeling. Similarly, (author?) [25] propose a framework for embedding conceptual models into data science workflows, illustrating how structured knowledge can support machine learning applications in healthcare. By combining conceptual modeling with machine learning, they emphasize the potential of ontologies to provide meaningful constraints, improving the interpretability and reliability of AI-driven predictions.

Despite all these advances, Sepsis prediction continues to be a challenging problem, as shown in [3], where the authors review machine learning approaches combining structured and unstructured data for this purpose. While these approaches improved the identification and early detection of sepsis, none of them explores the potential of structured knowledge integration, such as ontologies, to address this issue. Instead, the reliance on standard NLP techniques limits their ability to capture the semantic relationships present in clinical text. This gap creates an opportunity for approaches like the one presented in this study to leverage ontological knowledge for enhanced interpretability and predictive performance.

The above-mentioned works illustrate the potential of integrating structured knowledge with LLMs. The contribution of our study is twofold: first, while they primarily target general-purpose applications or multimodal knowledge integration, we emphasize the integration of domain-specific ontologies to enhance semantic understanding in predicting sepsis; second, while existing studies often utilize knowledge graphs, our study leverages the Unified Medical Language System (UMLS) ontology to embed structured clinical knowledge into the language model directly. Thus, our approach applies ontological integration specifically to improve predictive accuracy and reduce uncertainty in clinical NLP tasks.

4. Methods and Materials

As Mealy [26] stated way back in the late 60s, "Data are fragments of a theory of the real world", highlighting the need to connect data to their real-world referents. Controversially, data are still (and surprisingly increasingly) often perceived without the essential semantic commitment it requires.

One promising way of approaching this connection requires structuring the entities (and their relationships) of the real world that the data refers to, and representing them in a machine-readable format. Ontologies address this need by providing "a formal, explicit specification of a shared conceptualization" [27]. As an artifact, it offers a structured representation of interrelated concepts and formal axioms that define entities within a domain and the relationships between them. When enriched with constructs from a foundational ontology, ontologies establish an ontological commitment, enabling semantically precise definitions and ensuring that the knowledge represented is both contextually relevant and interpretable.

Furthermore, a systematic approach to delineate the essence of various domains is crucial for LLMs to grasp the underlying semantics of the data they process. Using structured knowledge representations, LLMs could achieve a more precise understanding of data, shedding light on potentially ambiguous definitions and discovering more profound and accurate correlations that reflect the true nature of the entities and events represented within the data.

In this context, together with the advances of LLMs, Clinical KB BERT emerges [28] as a significant advancement by integrating domain knowledge from the Unified Medical Language System (UMLS). The model introduces a novel joint training method for incorporating entities and relations of the UMLS ontology [29] into the Clinical BERT model; a detailed pipeline of this model is depicted in Figure 1.

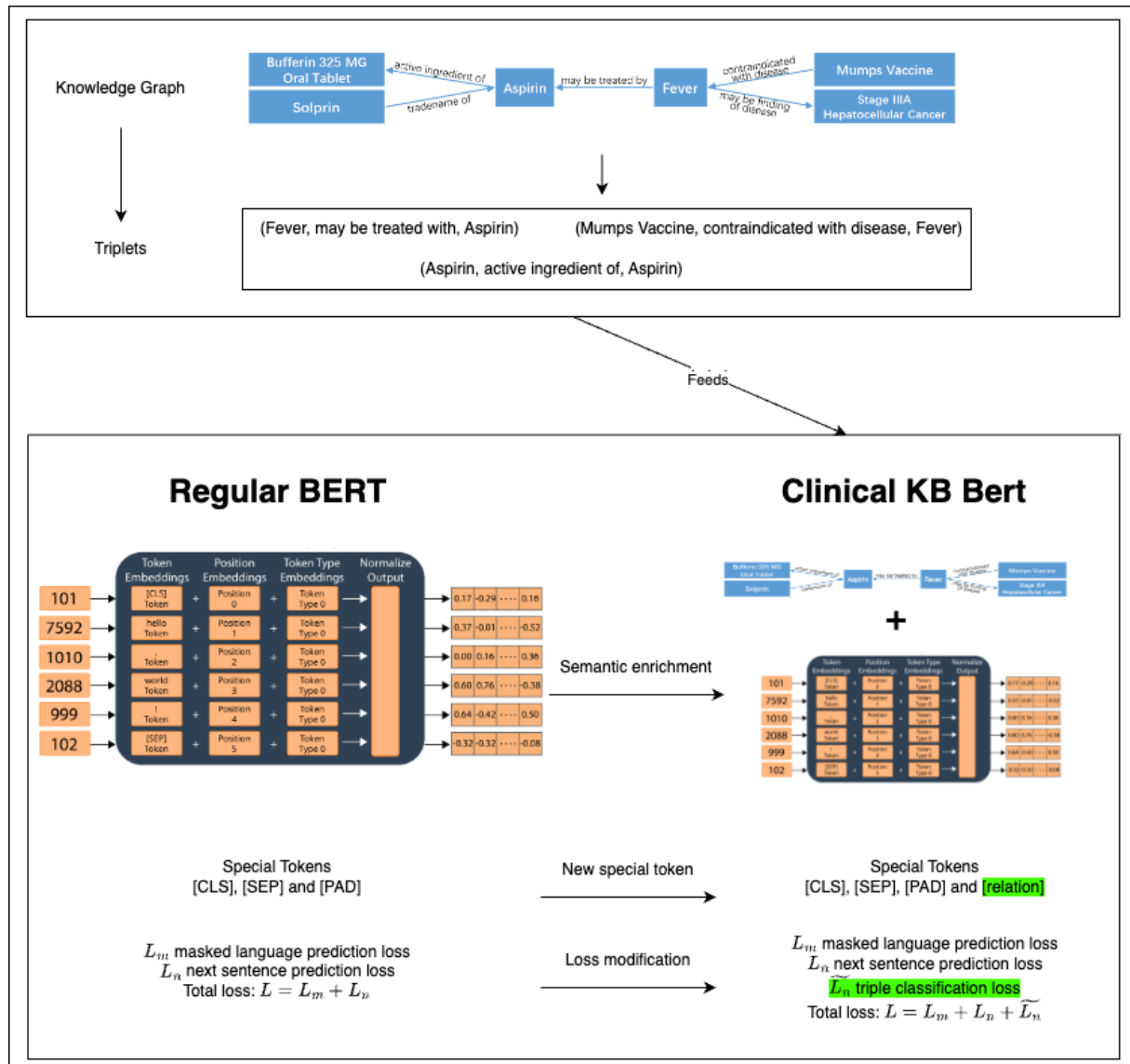


Fig. 1. Clinical KB Bert Pipeline. Source: Authors.

Methodologically, Clinical KB BERT integrates domain knowledge from the Unified Medical Language System (UMLS) into the pre-training process of the Clinical BERT model [28]. This integration is achieved through a novel joint training approach, where both masked language modeling (MLM) and knowledge base (KB) modeling tasks are optimized simultaneously. The key steps in the pipeline can be summarized as follows:

- (1) **Dataset Preparation:** The model uses two primary data sources: the MIMIC-III dataset, which provides approximately two million de-identified clinical notes, and the UMLS knowledge base, which contains over 27 million relations across 127 semantic types [29]. UMLS is represented in a multi-relational graph, with nodes corresponding to medical concepts and edges denoting relationships such as "may be treated by" or "associated with."
- (2) **Encoding UMLS Triplets:** UMLS relations are transformed into triplets of the form (*concept1*, *relation*, *concept2*). These triplets are then encoded into input sequences for the transformer model, formatted as: [CLS] *concept1* [relation] *concept2* [SEP]. Special tokens are used to represent the relations, enabling the model to learn their contextual significance.
- (3) **Joint Training Objective:** During pre-training, the model optimizes three objectives: the masked language model loss (L_m), the next sentence prediction loss (L_n), and a triplet classification loss (\tilde{L}_t). The triplet classification loss is specifically designed to model the relational structure in the UMLS graph. The final loss function combines these objectives as $L = 4L_m + L_n + \tilde{L}_t$, ensuring balanced learning of linguistic and knowledge-based representations.
- (4) **Negative Sampling for Triplets:** To address the sparsity of valid UMLS triplets, a negative sampling method is employed. Negative triplets are constructed by randomly replacing one of the concepts in a valid triplet while ensuring that the new triplet respects the semantic types of the original concepts. This method prevents the classification task from becoming trivial and enhances the robustness of the model.
- (5) **Fine-Tuning for Downstream Tasks:** After pre-training, Clinical KB BERT is fine-tuned on specific clinical NLP tasks, such as named entity recognition (NER) and natural language inference (NLI) [28]. The model demonstrated improvements in both accuracy and F1-scores compared to baseline methods, highlighting the effectiveness of integrating structured domain knowledge.

This multi-faceted training methodology allows Clinical KB BERT to leverage the rich relational structure of the UMLS ontology while maintaining strong performance on conventional language modeling tasks. The approach sets a precedent for integrating external knowledge bases into language models in low-resource domains such as clinical NLP.

The results were better in different downstream clinical NLP tasks: in a natural language inference task applied to clinical texts, it improved accuracy by up to 1.7%, and in clinical named entity recognition tasks, the F1-score improved by up to 1.0% [28]. These results highlight the potential of integrating structured knowledge to enhance the performance of NLP systems in low-resource - yet critical - settings, such as the clinical domain.

Building on this foundation, this research extends the application of Clinical KB BERT by incorporating it into a predictive modeling framework for sepsis prediction. Specifically, the study leverages the model's semantically enriched embeddings to enhance the representation of clinical notes, integrating these with structured physiological and demographic data to improve prediction accuracy. By focusing on the early detection of sepsis, a critical challenge in healthcare, this work not only validates the utility of Clinical KB BERT in a new context but also demonstrates its potential for addressing high-stakes predictive tasks in clinical decision-making.

This approach contributes to the literature by demonstrating how structured ontological knowledge, when integrated with advanced language modeling techniques, can improve model interpretability, reduce uncertainty, and ultimately lead to more reliable outcomes in resource-constrained domains like healthcare. By expanding the scope of Clinical KB BERT’s application, this research sets a precedent for leveraging semantic enrichment in other predictive healthcare challenges.

4.1. Definition of the performance metrics

The metrics used to evaluate model performance in this study were the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and the Matthews Correlation Coefficient (MCC). As highlighted by [30], ROC-AUC measures a model’s discriminatory ability, summarizing its capacity to differentiate between positive and negative classes across all potential decision thresholds. ROC-AUC values range from 0.5 (random chance) to 1.0 (perfect discrimination), with higher values indicating superior diagnostic performance. This metric was used both for selecting the best-performing model during 8-fold cross-validation and for comparing the overall performance of the models.

The Matthews Correlation Coefficient (MCC) [31] was used to determine the optimal prediction threshold for each model. MCC is especially valuable in binary classification tasks with imbalanced datasets, as it considers all elements of the confusion matrix (true positives, true negatives, false positives, and false negatives). This ensures a less biased and more robust evaluation compared to metrics like accuracy or F1 score, which can be misleading under imbalance conditions. For each model, the threshold that maximized the MCC was selected, ensuring the best balance between true positives and false positives.

In addition to these metrics, model uncertainty was assessed using predictive entropy. As discussed by [19], predictive entropy quantifies the uncertainty of a model’s predictions, where higher entropy indicates greater uncertainty. This metric is particularly important in high-stakes domains such as healthcare, where understanding a model’s confidence is as crucial as its accuracy. Furthermore, as emphasized by [32], prediction model reliability is closely tied to its accuracy and can be effectively evaluated using measures like predictive entropy. Lower entropy values signify more reliable predictions, as they reflect reduced uncertainty and a higher likelihood that the model’s outputs align with the ground truth. In this context, the reduction in predictive entropy observed in the semantically enriched model not only indicates greater confidence but also underscores its improved reliability, a critical attribute in clinical applications where dependable decision-making can have direct implications for patient care.

Beyond classification performance and uncertainty measures, this study also evaluated model calibration, which assesses how well predicted probabilities align with actual outcomes [33]. Proper calibration is essential in clinical decision-making, as it ensures that predicted risk scores correspond to the true likelihood of an event occurring. Unlike purely discriminative metrics like AUC-ROC, which measure a model’s ability to differentiate between positive and negative cases, calibration provides insights into whether a model systematically overestimates or underestimates risk.

In the context of sepsis prediction, miscalibrated models can have serious consequences. Overestimated risks may lead to unnecessary interventions, increasing healthcare costs and resource allocation inefficiencies, while underestimated risks can delay crucial treatments, worsening patient outcomes. As highlighted by (author?) [33], calibration impacts clinical usefulness, which considers the practical implications of deploying a predictive model, including its effects on costs, utilities, and potential harms associated with decision-making. Poorly calibrated models may contribute to excessive or insufficient medical responses, undermining their real-world applicability.

1 To ensure that the predictions made by both models are reliable, this study examines their calibration 1
2 across different probability ranges, identifying patterns of overconfidence or underconfidence in their 2
3 predictions. A well-calibrated model improves not only interpretability but also the cost-effectiveness of 3
4 early sepsis detection, aligning probability estimates with meaningful clinical thresholds. By integrat- 4
5 ing structured knowledge, the semantically aware model demonstrates improved predictive confidence, 5
6 which, in turn, enhances the practical value of AI-driven sepsis risk assessment in intensive care settings. 6

7 Ultimately, the combination of ROC-AUC, MCC, predictive entropy provides a comprehensive as- 7
8 sessment of model performance, balancing discriminatory power, reliability, and confidence in predic- 8
9 tions. The improvements observed in the semantically enriched model further support the integration 9
10 of structured domain knowledge into machine learning frameworks to enhance predictive accuracy and 10
11 trustworthiness in clinical settings. 11

12 4.2. Dataset 12

13 This study used MIMIC-III, a publicly available dataset that contains information on 46,520 patient 13
14 admissions in three different hospitals in the USA. It provides data on the demographics of patients, 14
15 vital signs, laboratory tests, and clinical electronic records (EHR) written by nurses, physicians, and 15
16 specialists, which will be the main type of data further analyzed in the present study. 16
17

18 We analyzed ICU stays for patients between the ages of 18 and 89. We excluded stays if the patient 18
19 developed sepsis within the first 8 hours of ICU admission or if the stay exceeded 30 days. Sepsis was 19
20 defined according to the Sepsis-3 criteria [1], with the objective of predicting sepsis development within 20
21 the next 4 hours. 21

22 The curated dataset consisted of 29,967 ICU stays, among which 1,135 cases developed sepsis. To 22
23 ensure a robust and clinically relevant feature selection, we adopted the approach used in the Phys- 23
24 ioNet/Computing in Cardiology Challenge 2019 [34], which focused on early sepsis prediction. Specif- 24
25 ically, we selected variables that were both considered in the challenge and available in the MIMIC-III 25
26 dataset, ensuring consistency with prior benchmarking efforts while leveraging a widely used real-world 26
27 clinical dataset. The final set of structured variables comprised 27 features related to patient vital signs, 27
28 laboratory values, and demographic characteristics, as detailed in Table 2, in addition to 768 features 28
29 representing text embeddings derived from electronic health records (EHR). 29

30 At this point, two versions of the dataset were created: for the semantically unaware scenario, the text 30
31 embeddings were generated using Clinical BERT, while for the semantically aware scenario, Clinical 31
32 KB BERT was used. 32

33 Due to the significant rate of missing values, the input data was formatted into matrices with 795 33
34 columns and T rows, where T represents the number of observations during a patient’s stay. We used a 34
35 Gated Recurrent Unit (GRU) model to classify whether a patient’s observations up to a certain time point 35
36 would indicate the likelihood of developing sepsis within the next 4 hours, given the varying lengths of 36
37 ICU stays. 37

38 All variables were standardized by subtracting the global mean and dividing by the standard deviation. 38
39 Missing values were imputed with zeros, and a masking layer was incorporated to indicate the presence 39
40 of missing data. 40

41 4.3. Model Architecture Details 41

42 The GRU-based architecture used in this study was implemented with TensorFlow [35], leveraging its 42
43 robust tools for building and optimizing deep learning models. The training process employed the Adam 43
44 44
45 45
46 46

Table 2
Selected Structured Variables for the Study

Variable	Variable name	Unit
Heart Rate	heartrate	beats per minute (bpm)
Pulse Oximetry	spo2	%
Body Temperature	tempc	Celsius Degree (°C)
Systolic Blood Pressure	sysbp	millimeters of mercury (mmHg)
Mean Arterial Pressure	meanbp	millimeters of mercury (mmHg)
Diastolic Blood Pressure	diasbp	millimeters of mercury (mmHg)
Respiration Rate	resprate	breaths per minute (bpm)
Base Excess	baseexcess	millimoles per liter (mmol/L)
Bicarbonate	bicarbonate	millimoles per liter (mmol/L)
Fraction of Inspired Oxygen	fio2	%
Partial pressure of carbon dioxide from arterial blood	pco2	millimeters of mercury (mmHg)
Oxygen Saturation from arterial blood	so2	%
Blood Urea Nitrogen	bun	milligrams per deciliter (mg/dL)
Calcium	calcium	milligrams per deciliter (mg/dL)
Chloride	chloride	millimoles per liter (mmol/L)
Creatinine	creatinine	milligrams per deciliter (mg/dL)
Serum Glucose	glucose	milligrams per deciliter (mg/dL)
Lactic Acid	lactate	milligrams per deciliter (mg/dL)
Potassium	potassium	millimoles per liter (mmol/L)
Total Bilirubin	bilirubin_total	milligrams per deciliter (mg/dL)
Hematocrit	hct	%
Hemoglobin	hgb	grams per deciliter (g/dL)
Partial thromboplastin time	ptt	seconds (s)
Leukocyt Time	wbc	count per liter (count/L)
Platelets	platelets	count per milliliter (count/mL)
Admission Age	admission_age	years

optimizer [36] with a learning rate of 0.0005, chosen for its ability to adaptively adjust learning rates during training, ensuring efficient convergence.

The model architecture consisted of two Gated Recurrent Unit (GRU) layers, with 64 and 32 units respectively, designed to capture sequential dependencies in the input data. These layers were followed by two fully connected dense layers, with 32 and 16 units, applying non-linear transformations to the extracted features. The output layer was configured with a sigmoid activation.

This design ensured a balance between model complexity and computational efficiency, allowing the model to effectively process sequential data while minimizing the risk of overfitting. By combining GRU layers with dense layers, the architecture was well-suited to integrate temporal dependencies from clinical notes with additional structured data for sepsis prediction.

4.4. Experiment setup

We trained two models: one using Clinical BERT (henceforth referred to as *semantically unaware*) and the other using Clinical KB BERT (*semantically aware*), to assess the impact of incorporating struc-

1 tured medical knowledge. The training dataset constituted 80% of the total data, with stratification to 1
2 preserve the proportion of septic patients. This 80% was further divided into 8 folds for training, with 2
3 the remaining fold serving as the validation set for each training set partition. 3

4 We assessed the ability of both the *semantically aware model* and the *semantically unaware model* 4
5 to predict sepsis onset within a 4-hour window for ICU patients. For each model, we performed cross- 5
6 validation by training 8 different models, each using a distinct fold of the data as the validation set 6
7 while the remaining folds were used for training. This approach ensured a robust evaluation of model 7
8 performance across different subsets of the dataset. 8

9 The evaluation process consisted of three key stages: 9

- 10 (1) **Cross-validation:** Models were trained using an 8-fold cross-validation approach to ensure robust 10
11 performance estimation. For each fold, a model was trained and validated on separate subsets of 11
12 the training data, enabling the identification of the best-performing model in each scenario. 12
- 13 (2) **Test set evaluation:** The best-performing models from cross-validation were evaluated on a hold- 13
14 out test set to measure their generalization capabilities. This step provided a realistic assessment of 14
15 how the models might perform in clinical settings. 15
- 16 (3) **Stratified analysis:** To provide deeper insights into the strengths and limitations of each model, 16
17 the test set predictions were further divided into distinct strata based on prediction correctness and 17
18 entropy values. This stratification enabled a nuanced interpretation of model behavior in different 18
19 clinical scenarios. 19
20

21 The code and implementation details for this research are publicly available on GitHub¹, ensuring 21
22 transparency and facilitating reproducibility for future studies. 22

23 5. Results and Discussion 23

24 This chapter presents a detailed evaluation of the models developed in this study, comparing their 24
25 performance on the task of early sepsis prediction. Following the methodology presented in section 4, the 25
26 evaluation was designed to comprehensively assess model performance using various metrics, such as 26
27 Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Matthews Correlation Coefficient 27
28 (MCC), and predictive entropy. These metrics were selected to ensure that the analysis captured both the 28
29 models' overall accuracy and their ability to manage uncertainty, particularly in a domain where missed 29
30 predictions can have severe clinical consequences. 30
31

32 Additionally, model calibration was assessed using the Giviti Calibration Belt, which evaluates the 32
33 alignment between predicted and observed probabilities. This analysis provides insights into whether the 33
34 models systematically overestimate or underestimate sepsis risk, ensuring that their probability outputs 34
35 are meaningful and reliable for clinical applications. 35
36
37

38 5.1. Performance Overview 38

39 For the *semantically aware model*, the best performance was achieved when fold 0 was used as the 39
40 validation set, resulting in an AUC-ROC of 0.833. Similarly, the *semantically unaware model* performed 40
41 best with fold 2 as the validation set, achieving an AUC-ROC of 0.805. 41
42
43

44
45 ¹<https://github.com/lucasmadda/EvaluatingOntologicallyAwareLargeLanguageModels/> 45
46

Table 3

ROC-AUC scores for the *Semantically Aware* and *Semantically Unaware* models using the dataset with 8-fold cross-validation.

Fold	Validation (V) or Test (T) set	Model	
		Semantically Aware	Semantically Unaware
0	V	0.833	0.797
	T	0.853	0.834
1	V	0.807	0.803
	T	0.839	0.833
2	V	0.822	0.805
	T	0.872	0.826
3	V	0.764	0.765
	T	0.827	0.838
4	V	0.795	0.803
	T	0.845	0.843
5	V	0.780	0.793
	T	0.860	0.813
6	V	0.786	0.794
	T	0.873	0.881
7	V	0.786	0.780
	T	0.845	0.827
8	V	0.797	0.793
	T	0.852	0.837

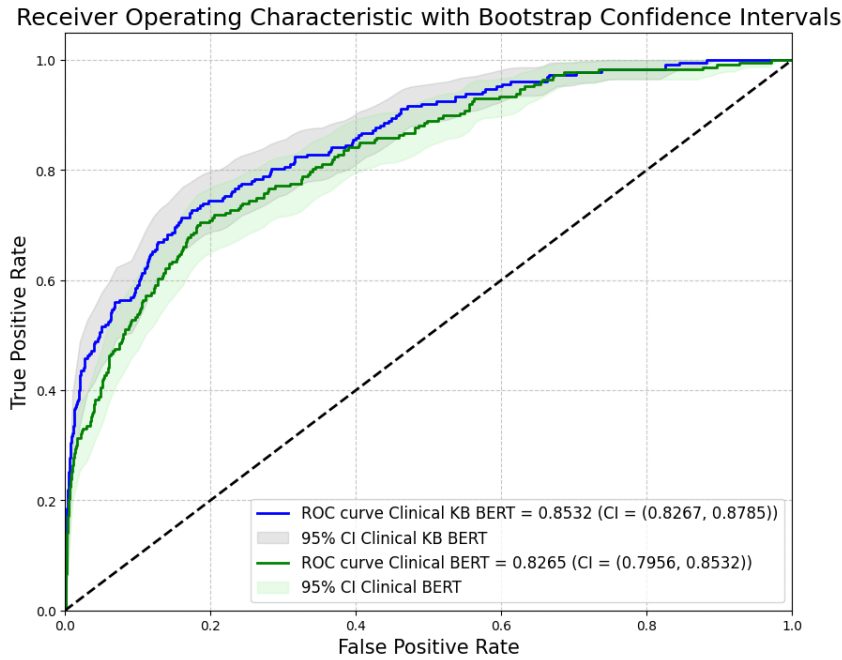
These results, summarized in table 3, highlight the importance of cross-validation in systematically evaluating and selecting models that generalize well across different subsets of data. The selected models were subsequently evaluated on the test set to ensure their generalizability extended beyond the validation folds. Also, figure 2 shows how the ROC curve behaves for each of the chosen trained models.

To determine an optimal decision threshold, we analyzed the ROC (Receiver Operating Characteristic) curve for each model, selecting thresholds that maximized the MCC for predictive accuracy. For Clinical BERT, this threshold was set at 0.485, yielding an MCC of 0.356. In contrast, the Clinical KB BERT model required a lower threshold of 0.321 to maximize its MCC, resulting in an improved MCC score of 0.424. These thresholds allowed each model to classify ICU stays with a probability of sepsis within the next 4 hours as positive when exceeding the specified threshold. The definition of specific thresholds enabled each model to perform its best possible classification, according to its learning phase.

To further assess model reliability, I analyzed the models' calibration, which provides a visual and statistical evaluation of the agreement between predicted probabilities and observed outcomes. Figure 3 and figure 4 illustrate the calibration curves for the semantically unaware and semantically aware models, respectively.

The calibration belt for the semantically unaware model (Figure 3) exhibits significant deviations from the ideal calibration line (bisector), particularly in the mid-range probability region. The confidence intervals indicate that the model tends to underestimate the probability of sepsis onset for certain cases, as the observed probabilities fall below the diagonal. The p-value of <0.001 further suggests that the model's predictions are not well-calibrated.

Conversely, the semantically aware model (Figure 4) demonstrates improved calibration, with a p-value of 0.031, suggesting better alignment between predicted and observed probabilities. Although some miscalibration remains, particularly for lower probability predictions, the reduction in the range



11

Fig. 2. Receiver Operating Characteristic with Bootstrap Confidence Intervals Using the Test Set. Source: Author.

of deviations indicates that incorporating structured knowledge improves the model's ability to generate more reliable probability estimates.

The confusion matrices (see table 4 and table 5) provide a detailed breakdown of predictions considering the selected thresholds.

True Class	Predicted Class	
	Non-Sepsis	Sepsis
Non-Sepsis	5277	117
Sepsis	128	99

Table 4

Confusion Matrix for Sepsis Prediction with *Semantically Aware Model*

True Class	Predicted Class	
	Non-Sepsis	Sepsis
Non-Sepsis	5327	67
Sepsis	162	65

Table 5

Confusion Matrix for Sepsis Prediction with *Semantically Unaware Model*

The classification metrics presented in table 6 provide a comprehensive comparison of the *semantically aware* and *semantically unaware models*. While the semantically unaware model achieves slightly higher accuracy (0.959 vs. 0.956) and precision (0.492 vs. 0.458), the semantically aware model demonstrates superior performance in other critical metrics that are particularly important in the context of sepsis prediction, such as recall, F1-score, and Matthews Correlation Coefficient (MCC).

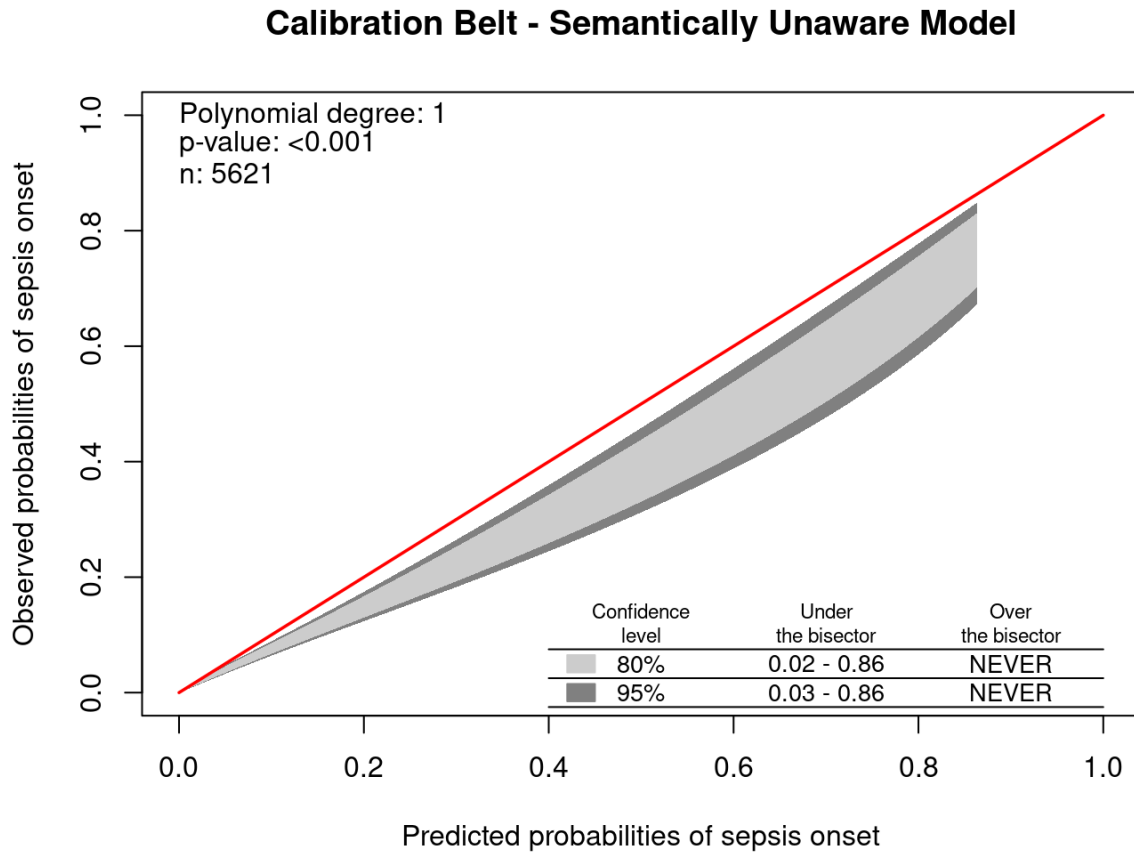


Fig. 3. Calibration Belt for the Semantically Unaware Model on the Test Set. Source: Author.

The marginally higher accuracy and precision of the *semantically unaware model* may seem advantageous at first glance. However, it is notorious that accuracy can be misleading in imbalanced datasets, as it is heavily influenced by the majority class, which in this case is non-septic patients. Similarly, the higher precision of the *semantically unaware model* indicates fewer False Positives, but this comes at the expense of a much lower recall. For a task as critical as sepsis prediction, the cost of missing True Positives (septic cases) far outweighs the cost of flagging False Positives, as missing a septic case can delay life-saving interventions. Thus, while precision and accuracy are important, their clinical impact is overshadowed by metrics like recall and MCC, where the *semantically aware model* excels.

The *semantically aware model's* substantially higher recall (0.436 vs. 0.286) and F1-score (0.446 vs. 0.362) show that it is better equipped to identify septic cases while maintaining a reasonable trade-off between sensitivity and specificity. Recall, in particular, is crucial in sepsis prediction because it directly correlates with the model's ability to minimize False Negatives—cases where septic patients are misclassified as non-septic. Additionally, the *semantically aware model's* superior MCC (0.424 vs. 0.356) highlights its overall robustness and balanced performance in the imbalanced scenario, making it more reliable for real-world clinical use. By leveraging domain-specific knowledge through ontological

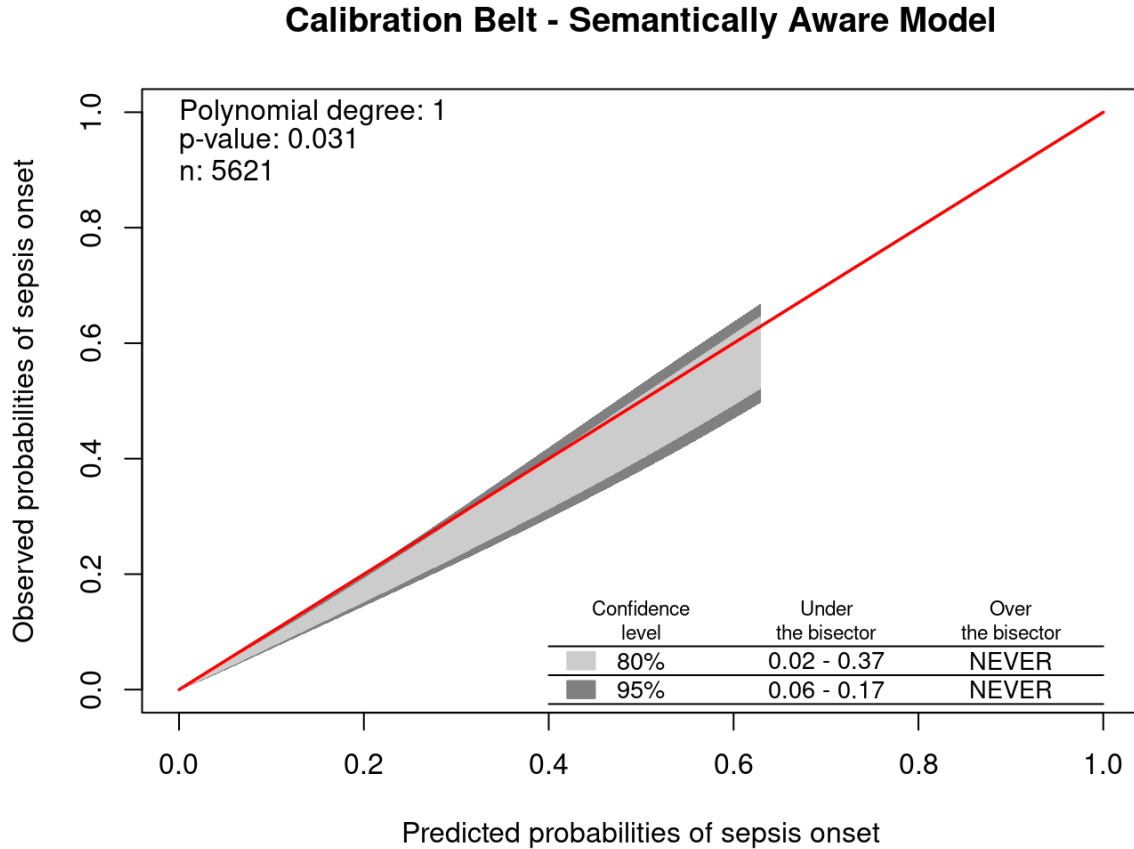


Fig. 4. Calibration Belt for the Semantically Aware Model on the Test Set. Source: Author.

embeddings, the *semantically aware model* captures nuanced relationships in clinical notes, leading to better-informed predictions and ultimately improving patient outcomes.

5.2. Model Performance Comparison and Interpretative Insights

The predictive performance of Clinical KB BERT shows a notable improvement over Clinical BERT, as reflected in the ROC-AUC values. Clinical BERT achieved an ROC-AUC of 0.8265 (95% CI: 0.7956, 0.8532), while Clinical KB BERT reached an enhanced ROC-AUC of 0.8532 (95% CI: 0.8267, 0.8785). This increase demonstrates Clinical KB BERT's ability to leverage domain-specific ontological knowledge, improving contextually accurate predictions and particularly benefiting cases with complex risk profiles.

Further analysis of individual cases highlights Clinical KB BERT's advantage in capturing high-risk sepsis cases missed by Clinical BERT. Specifically, Clinical KB BERT correctly identified sepsis in 37 instances that Clinical BERT failed to predict, while only three cases were correctly identified by Clinical BERT and missed by Clinical KB BERT.

Table 6
Classification metrics results on the Test Set and 95% confidence interval for each metric and model.

Classification Metric	Semantically Aware		Semantically Unaware	
	Result	Bootstrapped 95% CI	Result	Bootstrapped 95% CI
Accuracy	0.956	(0.951, 0.961)	0.959	(0.953, 0.964)
Precision	0.458	(0.393, 0.528)	0.492	(0.407, 0.570)
Recall	0.436	(0.373, 0.500)	0.286	(0.229, 0.344)
F1-Score	0.446	(0.390, 0.503)	0.362	(0.296, 0.422)
MCC	0.424	(0.367, 0.483)	0.356	(0.290, 0.418)

5.3. Diving into Relevant Strata

To provide a comprehensive comparison of model performance, we analyzed the Electronic Health Records (EHRs) clinical notes from the test set, dividing the cases into five distinct strata that represent different subsets of data, thus providing a better way of analyzing the impact of the semantic enrichment:

- (1) **Correct predictions by both models:** Cases where both *semantically aware* and *semantically unaware models* correctly predicted sepsis. The number of such cases was $n = 5310$.
- (2) **Semantically-aware gains:** Cases where the *semantically aware model* correctly predicted sepsis (true positive and true negative), while the *semantically unaware model* failed to identify it (false negative and false positive). There were $n = 66$ such cases.
- (3) **Positive semantically-aware misses:** Cases where the *semantically aware model* misclassified a sepsis case (false negative), but the *semantically unaware model* succeeded (true positive). The number of such cases was $n = 3$.
- (4) **Incorrect predictions by both models:** These are cases where both the *semantically aware model* and the *semantically unaware model* failed to correctly predict sepsis. There were $n = 163$ such cases observed in the test dataset.
- (5) **Semantically-aware misses:** Cases where the *semantically unaware model* correctly identified positive and negative cases at the same time that the *semantically aware model* misclassified these cases. There were $n = 82$ cases in the test dataset.

In each scenario, we analyzed the predictive entropy of sepsis predictions for both models. This analysis aimed not only to evaluate whether the models made correct or incorrect classifications but also to assess their confidence levels as reflected by the predictive entropy.

5.4. Correct predictions by both models

In the strata where both models correctly predicted outcomes (septic and non-septic cases) (Figure 5), the violin plot highlights differences in the distribution of predictive entropies between the two models. The *semantically unaware model* exhibits higher predictive entropy values, which suggests greater variability and less certainty in its predictions. This variability may be linked to its embedding process of clinical texts, which does not incorporate external semantic relationships, potentially leading to noisier internal representations.

In contrast, the *semantically aware model* produces lower predictive entropy values, indicating more consistent predictions across cases. The distribution for this model shows a tighter interquartile range

(IQR) and fewer high-entropy outliers, reflecting a more concentrated range of uncertainty. These observations are consistent with the structured information included during its training, which could contribute to more stable internal representations.

The differences in entropy distributions between the models suggest that the inclusion of external knowledge in the training process may affect the variability of predictions. This variability is observed not only in the central tendency of entropy values but also in their spread, as seen in the violin plot.

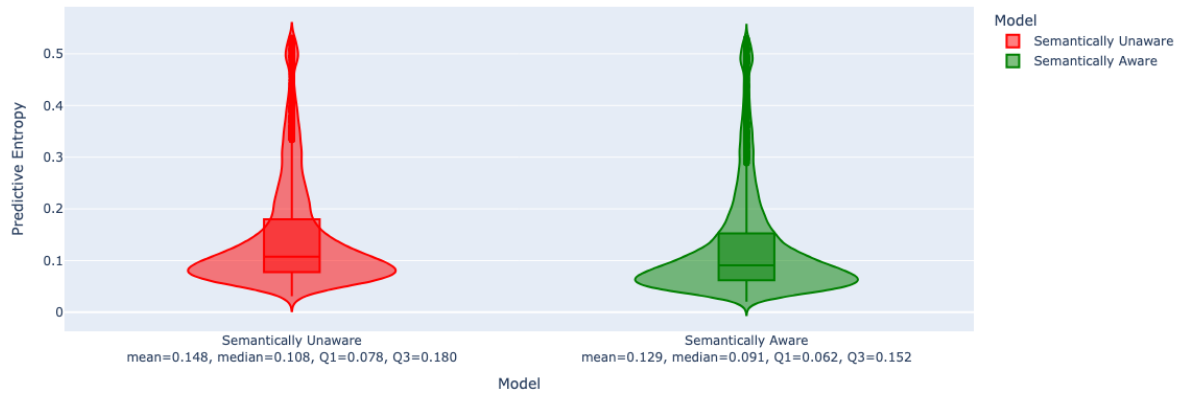


Fig. 5. Comparison of predictive entropies for correct predictions (septic and non-septic cases) by both models. Source: Authors.

5.5. Semantically-aware gains

When analyzing the strata of *semantically-aware gains*, we observed 66 critical cases where the *semantically aware model* correctly predicted sepsis, while the *semantically unaware model* misclassified these same cases.

In the scenario illustrated in Figure 6, we analyzed the predictive entropy for cases where the *semantically aware model* correctly identified both septic and non-septic outcomes, while the *semantically unaware model* failed. This subset contained a total of 66 cases, of which 37 were septic. This is notable, as septic cases in the entire dataset accounted for 227 out of 5,621 cases, reflecting a proportion of approximately 4.04%. In this subset, however, septic cases represented 56.06% (37/66), highlighting the *semantically aware model's* ability to excel in identifying a disproportionately higher number of septic cases in this scenario.

The violin plot in Figure 6 displays the distribution of predictive entropies for these cases. The *semantically aware model* exhibits a tighter and more consistent distribution of entropies, with a mean of 0.511, a median of 0.524, and an interquartile range (IQR) spanning from 0.512 to 0.529. This narrow IQR and the absence of high-entropy outliers suggest that the model made its predictions with high confidence and consistency. In contrast, the *semantically unaware model*, which failed in these cases, demonstrated higher overall entropy values (mean of 0.381, median of 0.406), with a much wider IQR (0.237 to 0.513). This broader distribution indicates greater uncertainty and variability in its predictions, which may have contributed to its failure to correctly classify these cases.

The enrichment of septic cases within this subset further emphasizes the differences in model behavior. The *semantically aware model's* ability to produce confident and correct predictions in cases where

septic outcomes are overrepresented suggests that its embeddings, enriched with structured semantic knowledge, may provide more reliable representations for detecting challenging cases. This performance stands in stark contrast to the variability and uncertainty seen in the *semantically unaware model's* entropy distribution.

Overall, the analysis of predictive entropy distributions in this subset reveals a clear distinction between the models' performance, particularly in scenarios where the correct identification of septic cases is critical.

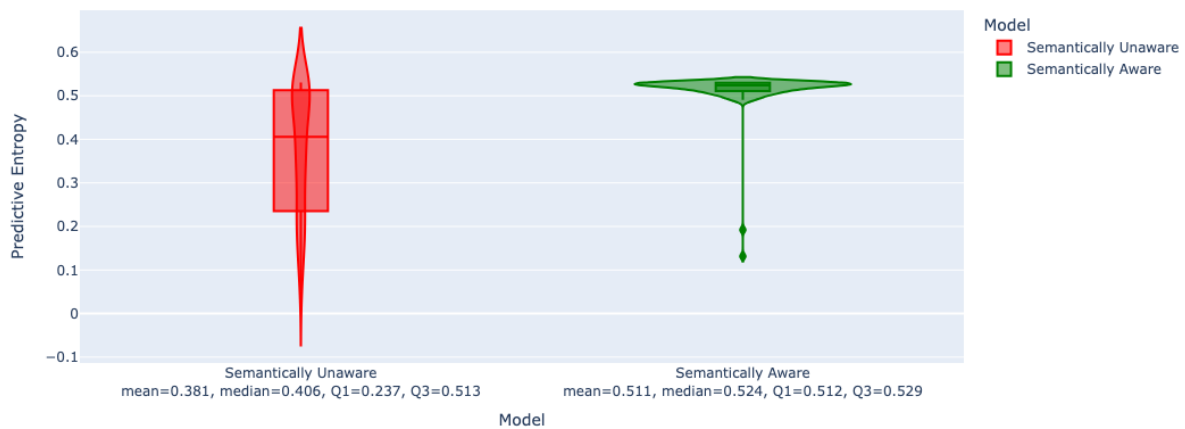


Fig. 6. Comparison of predictive entropies for cases which only the *semantically aware model* predicted correctly. Source: Authors.

In the scenario illustrated in Figure 6, we analyzed the predictive entropy for cases where the *semantically aware model* correctly identified both septic and non-septic outcomes, while the *semantically unaware model* failed. This subset contained a total of 66 cases, of which 37 were septic. This is notable, as septic cases in the entire dataset accounted for 227 out of 5,621 cases, reflecting a proportion of approximately 4.04%. In this subset, however, septic cases represented 56.06% (37/66), highlighting the overrepresentation of septic cases in scenarios where the *semantically aware model* excelled.

The violin plot in Figure 6 displays the distribution of predictive entropies for these cases. The *semantically aware model* exhibits a tighter and more consistent distribution of entropies, with a mean of 0.511, a median of 0.524, and an interquartile range (IQR) spanning from 0.512 to 0.529. This narrow IQR and the absence of extreme entropy values suggest that the model made its predictions with higher confidence and consistency. In contrast, the *semantically unaware model*, which failed in these cases, demonstrated lower overall entropy values (mean of 0.381, median of 0.406), but with a much wider IQR (0.237 to 0.513). This broader distribution indicates greater variability in its predictive entropy, reflecting a lack of consistency.

5.6. Positive semantically-aware misses

We refer to the cases where the *semantically aware model* misclassified a sepsis case (false negative) while the *semantically unaware model* succeeded (true positive) as **Positive Semantically-Aware**

Misses. Such instances were extremely rare, with only $n = 3$ cases observed across the entire dataset. Given the limited number of occurrences, it was not meaningful to include a visualization for this subset.

The small number of **Positive Semantically-Aware Misses** highlights how infrequently the *semantically unaware model* outperformed the *semantically aware model* in sepsis predictions. Table 7 provides the predicted probabilities (\hat{y}) and predictive entropy for each case. Here, \hat{y} represents the probability predicted by the respective model. As seen in the table, the predictive entropy for the *semantically aware model* was consistently higher than that of the *semantically unaware model*, indicating greater uncertainty in the former’s incorrect predictions.

Table 7

Positive semantically-aware misses predicted probabilities and predictive entropies.

Case	\hat{y}		Predictive Entropy	
	Semantically Aware	Semantically Unaware	Semantically Aware	Semantically Unaware
1	0.042	0.664	0.389	0.192
2	0.025	0.614	0.432	0.132
3	0.284	0.489	0.505	0.516

5.7. Incorrect predictions by both models

In the subset of $n = 163$ cases where both the *semantically aware model* and the *semantically unaware model* failed to correctly predict sepsis, the violin plot in Figure 7 shows the distribution of predictive entropies for these shared incorrect predictions. While the overall distributions appear similar, the *semantically aware model* demonstrates slightly higher predictive entropy values compared to the *semantically unaware model*.

The *semantically aware model* has a mean predictive entropy of 0.293 and an interquartile range (IQR) spanning from 0.131 to 0.483. In comparison, the *semantically unaware model* has a slightly lower mean predictive entropy of 0.287 and a narrower IQR of 0.136 to 0.429.

These results suggest that, while both models made incorrect predictions, the *semantically aware model* tended to exhibit higher uncertainty in its decisions, as reflected by the elevated entropy values. This indicates that the *semantically aware model* was less confident in its incorrect classifications, which could be seen as a preferable behavior compared to confidently making errors.

Higher predictive entropy values for the *semantically aware model* might reflect a recognition of the complexity or ambiguity inherent in these cases, where the model’s semantic knowledge influenced its decision-making process, albeit without reaching a correct conclusion. Conversely, the lower entropy values for the *semantically unaware model* suggest that it was more confident in its incorrect predictions, which could indicate a lack of sensitivity to nuanced patterns present in the data.

5.8. Semantically Aware Misses

The subset of Semantically Aware Misses refers to cases where the *semantically aware model* misclassified sepsis outcomes, while the *semantically unaware model* succeeded. In total, there were $n = 82$ such cases in the test dataset, encompassing both septic and non-septic instances. As depicted in Figure 8, the predictive entropy for the *semantically aware model* in this subset was notably higher, with a mean of 0.511 and a median of 0.524, compared to the *semantically unaware model*, which exhibited a lower



Fig. 7. Comparison of predictive entropies for incorrect predictions (septic and non-septic cases) by both models. Source: Authors.

mean entropy of 0.381 and a median of 0.406. The interquartile range (IQR) was also narrower for the *semantically aware model* (0.512 to 0.529), indicating a more concentrated distribution of uncertainty.

This higher predictive entropy suggests that, in cases where the *semantically aware model* produced incorrect predictions, it demonstrated a higher degree of uncertainty compared to the *semantically unaware model*. This behavior aligns with the expectations of a model that integrates semantic knowledge, as the increased uncertainty may reflect the model's recognition of the complexity or ambiguity in these cases. In contrast, the lower entropy values for the *semantically unaware model* indicate overconfidence in its incorrect predictions, highlighting its inability to discern the nuances present in these challenging cases.

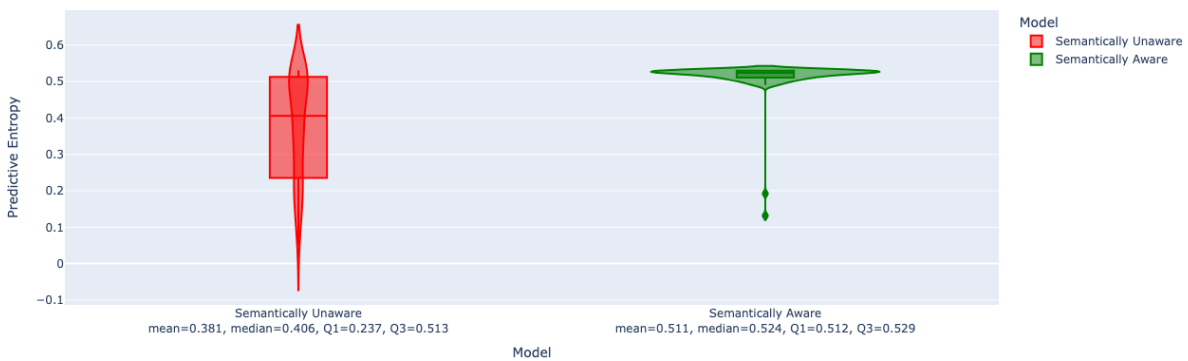


Fig. 8. Comparison of predictive entropies for Semantically Aware Misses. Source: Authors.

5.9. Threats to Validity

This study faces several threats to validity that should be acknowledged. First, the choice of the compared models—Clinical KB BERT and Clinical BERT—introduces the potential for model selection bias. While these models were selected for their relevance to the clinical domain, other large language models or techniques for incorporating structured knowledge might yield different results. Future work could explore a broader range of models to assess whether the observed benefits of semantic enrichment are consistent across architectures.

Second, the dependency on the MIMIC-III dataset poses a challenge to external validity. Although widely adopted to evidence successful advances in healthcare data analysis, since it encompasses a diverse and very large population of real ICU patients and contains highly granular (both structured and unstructured) data, MIMIC-III is specific to a single healthcare institution and reflects documentation practices and patient demographics from a (considerably long, yet still particular) time period. This may limit the generalizability of the findings to other healthcare systems or datasets with different populations and practices.

Furthermore, the study’s focus on sepsis prediction raises concerns about domain-specific generalizability. While the integration of ontologies demonstrated clear benefits for this task, it is unclear whether similar gains would be observed in other clinical or non-clinical predictive tasks. Evaluating the performance of semantically enriched embeddings across a broader range of domains could provide more robust evidence of their utility.

Finally, the dataset’s class imbalance, with a significantly smaller proportion of septic cases compared to non-septic cases, could influence the results. Although this imbalance reflects clinical reality, it may lead to overly optimistic metrics for the majority class and reduce sensitivity to the minority class. We mitigated this limitation by using evaluation metrics and techniques more suitable for imbalanced datasets, such as MCC. Other approaches to balance the data or improve model sensitivity to underrepresented classes could also be experimented in future studies.

6. Conclusion

This study experimentally demonstrates that ontologically enriched embeddings improve the prediction of sepsis in ICU patients. By integrating domain-specific knowledge from the Unified Medical Language System (UMLS) into the embedding process, the proposed BERT model (Clinical KB BERT) achieved more accurate and less uncertain predictions over a baseline model (Clinical BERT). Specifically, the semantically enriched model increased the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) from 0.826 to 0.853, while the mean predictive entropy for the entire test dataset decreased from 0.159 to 0.142. Furthermore, the reduction in mean predictive entropy was even more pronounced in cases where both models made correct predictions, decreasing from 0.148 to 0.129. Noteworthy, the practical impacts of these improvements include a substantial decrease in the number of false negatives (from 162 to 128, out of 227 septic cases), emphasizing the ability of the *semantically aware model* to reduce missed early diagnoses and improve patient outcomes.

The stratified analysis further revealed that the semantically aware model excelled in identifying septic cases that the baseline model failed to predict. Analyzing the strata where both models made incorrect predictions, the semantically aware model exhibited higher predictive entropy, reflecting a more cautious approach to these challenging cases. This cautious behavior aligns with the requirements of clinical contexts, where overconfidence in incorrect predictions could lead to adverse outcomes.

1 This work successfully addressed the research questions presented in the Introduction. Regarding 1
2 RQ1, the integration of structured domain knowledge, in the form of UMLS ontologies, demonstrably 2
3 enhanced both the performance and reliability of predictive models by means of higher F1-Score, Recall 3
4 and MCC, as well as lower predictive entropy. 4

5 For RQ2, the benefits of using semantically enriched embeddings were quantified through improve- 5
6 ments in AUC-ROC and reductions in predictive entropy. 6

7 Finally, concerning RQ3, the incorporation of structured semantic information notably reduced false 7
8 negatives, underscoring its critical role in mitigating harmful misclassifications in sepsis prediction. 8

9 These results reinforce the potential of ontological integration in advancing predictive modeling across 9
10 high-stakes domains. 10

11 The results of this study highlight the potential of integrating structured knowledge into predictive 11
12 models, yet several directions for future research remain open. 12

13 A crucial avenue is the exploration of neuro-symbolic AI approaches for improving explainability and 13
14 reasoning in clinical decision-making. While current deep learning models, including those used in this 14
15 study, excel at pattern recognition, they struggle with abstract reasoning and logical inference. Neuro- 15
16 symbolic AI aims to bridge this gap by combining the strengths of both symbolic reasoning and neural 16
17 networks [37]. Future work could investigate hybrid architectures that incorporate structured knowledge 17
18 — such as ontologies — into learning pipelines, allowing models to engage in explicit reasoning over 18
19 medical concepts, rather than merely learning implicit associations. 19

20 Additionally, this study demonstrated that semantically enriched embeddings improve predictive accu- 20
21 racy, yet post-hoc explainability remains a challenge. Techniques such as rule extraction, counterfactual 21
22 reasoning, and logical constraints could be explored to interpret model decisions in a clinically meaning- 22
23 ful way. In high-stakes applications like sepsis prediction, understanding why a model makes a particular 23
24 prediction is just as important as the prediction itself. 24

25 Another key question for future research is the impact of different ontological structures on model 25
26 performance. The UMLS ontology, while extensive, is relatively shallow and lacks well-defined axioms 26
27 to support robust inferencing. An open question is whether richer, more foundational ontologies—such 27
28 as UFO [38]—could further enhance performance. One hypothesis worth testing is: Does the depth and 28
29 richness of an ontology improve downstream predictive accuracy and explainability? 29

30 Moreover, expanding the ontology’s graph structure before embedding generation presents an intrigu- 30
31 ing possibility. Instead of directly encoding UMLS relationships, one could experiment with reifying 31
32 transitive relations and hierarchical structures prior to embedding computation. This would allow mod- 32
33 els to better leverage indirect semantic relationships that may not be explicitly represented in the raw 33
34 textual data. 34

35 Beyond clinical applications, these methods should be evaluated in other domains where structured 35
36 knowledge is crucial. Domains such as cybersecurity, finance, and legal reasoning require predictive 36
37 models that can not only classify but also justify their decisions in interpretable ways. Investigating how 37
38 ontology-based embeddings generalize to diverse settings could provide further evidence of their broad 38
39 applicability. 39

40 Lastly, while this study focused on predictive modeling, future work could extend towards generative 40
41 approaches that construct interpretable explanations for model predictions. By integrating ontological 41
42 knowledge with natural language generation, it may be possible to create models that not only predict 42
43 sepsis but also generate human-readable justifications, enhancing trust and adoption in clinical settings. 43

44 In conclusion, the integration of structured domain knowledge in machine learning has shown sig- 44
45 nificant promise, but several key questions remain. Future research should explore richer ontological 45
46 46

1 representations, neuro-symbolic architectures, post-hoc explainability techniques, and generalizability 1
2 across domains. Addressing these challenges will further strengthen the role of ontologies in enhancing 2
3 AI interpretability, reliability, and trustworthiness. 3

4 In conclusion, this study highlights the value of structured knowledge in enhancing language models 4
5 for critical domains. The integration of ontologies advances the performance of predictive algorithms, 5
6 while providing a foundation for building more trustworthy and interpretable AI systems, paving the 6
7 way for innovations in data-driven decision-making applications across domains. 7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

References

- [1] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.-D. Chiche, C.M. Coopersmith, R.S. Hotchkiss, M.M. Levy, J.C. Marshall, G.S. Martin, S.M. Opal, G.D. Rubenfeld, T. van der Poll, J.-L. Vincent and D.C. Angus, The third international consensus definitions for sepsis and septic shock (sepsis-3), *JAMA* **315**(8) (2016), 801–810.
- [2] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson and M. Tomlanovich, Early goal-directed therapy in the treatment of severe sepsis and septic shock, *N. Engl. J. Med.* **345**(19) (2001), 1368–1377.
- [3] L. Gustad, Nytrø and M. Yan, Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review, *Journal of the American Medical Informatics Association* **00** (2021), 1–17. doi:10.1093/jamia/ocab236.
- [4] I. Jahan, M.T.R. Laskar, C. Peng and J.X. Huang, A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks, *Computers in Biology and Medicine* **171** (2024), 108189. doi:https://doi.org/10.1016/j.combiomed.2024.108189. https://www.sciencedirect.com/science/article/pii/S0010482524002737.
- [5] F.K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney and F. Rudzicz, A survey of word embeddings for clinical text, *Journal of Biomedical Informatics* **100** (2019), 100057, Articles initially published in Journal of Biomedical Informatics: X 1–4, 2019. doi:https://doi.org/10.1016/j.yjbix.2019.100057. https://www.sciencedirect.com/science/article/pii/S2590177X19300563.
- [6] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.
- [7] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann and M. McDermott, Publicly Available Clinical BERT Embeddings, in: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, A. Rumshisky, K. Roberts, S. Bethard and T. Naumann, eds, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. doi:10.18653/v1/W19-1909. https://aclanthology.org/W19-1909.
- [8] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R.G. Mark, MIMIC-III, a freely accessible critical care database, *Scientific Data* **3**(1) (2016). doi:10.1038/sdata.2016.35.
- [9] W. Yao, L. Jin, H. Zhang, X. Pan, K. Song, D. Yu, D. Yu and J. Chen, How do Words Contribute to Sentence Semantics? Revisiting Sentence Embeddings with a Perturbation Method, in: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, eds, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3001–3010. doi:10.18653/v1/2023.eacl-main.218. https://aclanthology.org/2023.eacl-main.218/.
- [10] S. Jiang, W. Wu, N. Tomita, C. Ganoe and S. Hassanpour, Multi-Ontology Refined Embeddings (MORE): A hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts, *Journal of Biomedical Informatics* **111** (2020), 103581. doi:https://doi.org/10.1016/j.jbi.2020.103581. https://www.sciencedirect.com/science/article/pii/S1532046420302094.
- [11] L. Maddalena and F. Baião, An Application of the Disease Ontology (DO) for Clustering COVID-19 Hospitalizations in Rio de Janeiro., in: *ONTOBRAS*, 2022, pp. 9–22.
- [12] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, Unifying Large Language Models and Knowledge Graphs: A Roadmap, *IEEE Transactions on Knowledge and Data Engineering* **36**(7) (2024), 3580–3599. doi:10.1109/TKDE.2024.3352100.
- [13] T.P. Sales, F. Baião, G. Guizzardi, J.P.A. Almeida, N. Guarino and J. Mylopoulos, The common ontology of value and risk, in: *Conceptual Modeling: 37th International Conference, ER 2018, Xi'an, China, October 22–25, 2018, Proceedings 37*, Springer, 2018, pp. 121–135.
- [14] M. Fumagalli, G. Engelberg, T.P. Sales, Í. Oliveira, D. Klein, P. Soffer, R. Baratella and G. Guizzardi, On the semantics of risk propagation, in: *International Conference on Research Challenges in Information Science*, Springer, 2023, pp. 69–86.
- [15] G. Amaral, F. Baião and G. Guizzardi, Foundational ontologies, ontology-driven conceptual modeling, and their multiple benefits to data mining, *WIREs Data Mining and Knowledge Discovery* **11**(4) (2021), e1408. doi:https://doi.org/10.1002/widm.1408. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1408.
- [16] D. Porello, G. Guizzardi, T.P. Sales and G. Amaral, A core ontology for economic exchanges, in: *International Conference on Conceptual Modeling*, Springer, 2020, pp. 364–374.
- [17] R.F. Calhau, T. Prince Sales, Í. Oliveira, S. Kokkula, L. Ferreira Pires, D. Cameron, G. Guizzardi and J.P.A. Almeida, A system core ontology for capability emergence modeling, in: *International Conference on Enterprise Design, Operations, and Computing*, Springer, 2023, pp. 3–20.

- [18] J.C. de Almeida Rodrigues Gonçalves, F.A. Baiao, F.M. Santoro and G. Guizzardi, A cognitive BPM theory for knowledge-intensive processes, *Business Process Management Journal* **29**(2) (2023), 465–488.
- [19] M. Weiss and P. Tonella, Uncertainty-wizard: Fast and user-friendly neural network uncertainty quantification, in: *2021 14th IEEE Conference on Software Testing, Verification and Validation (ICST)*, IEEE, 2021, pp. 436–441.
- [20] M.M. Alam, M.R.A.H. Rony, M. Nayyeri, K. Mohiuddin, M.S.T.M. Akter, S. Vahdati and J. Lehmann, Language Model Guided Knowledge Graph Embeddings, *IEEE Access* **10** (2022), 76008–76020. doi:10.1109/ACCESS.2022.3191666.
- [21] M. Nayyeri, Z. Wang, M.M. Akter, M.M. Alam, M.R.A.H. Rony, J. Lehmann and S. Staab, Integrating Knowledge Graph embedding and pretrained Language Models in Hypercomplex Spaces (2022).
- [22] N. Huang, Y.R. Deshpande, Y. Liu, H. Alberts, K. Cho, C. Vania and I. Calixto, Endowing language models with multi-modal knowledge graph representations (2022).
- [23] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li and J. Tang, KEPLER: A unified model for Knowledge Embedding and pre-trained Language Representation (2019).
- [24] C. Sirocchi, A. Bogliolo and S. Montagna, Medical-informed machine learning: integrating prior knowledge into medical decision systems, *BMC Medical Informatics and Decision Making* **24**(S4) (2024). doi:10.1186/s12911-024-02582-4.
- [25] W. Maass and V.C. Storey, Pairing conceptual modeling with machine learning, *Data amp; Knowledge Engineering* **134** (2021), 101909. doi:10.1016/j.datak.2021.101909.
- [26] G.H. Mealy, Another Look at Data , in: *Managing Requirements Knowledge, International Workshop on*, Vol. 1, IEEE Computer Society, Los Alamitos, CA, USA, 1967, p. 525. doi:10.1109/AFIPS.1967.112.
- [27] R. Studer, V.R. Benjamins and D. Fensel, Knowledge engineering: Principles and methods, *Data Knowledge Engineering* **25**(1) (1998), 161–197. doi:https://doi.org/10.1016/S0169-023X(97)00056-6. https://www.sciencedirect.com/science/article/pii/S0169023X97000566.
- [28] B. Hao, H. Zhu and I. Paschalidis, Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base, in: *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, 2020. doi:10.18653/v1/2020.coling-main.57.
- [29] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Research* **32**(90001) (2004), 267D – 270–. doi:10.1093/nar/gkh061.
- [30] K. Çorbacıoğlu and G. Aksel, Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value, *Turkish Journal of Emergency Medicine* **23**(4) (2023), 195–198–. doi:10.4103/tjem.tjem18223.
- [31] D. Chicco and G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics* **21**(1) (2020). doi:10.1186/s12864-019-6413-7.
- [32] E.D. Gireesh and V.P. Gurupur, Information Entropy Measures for Evaluation of Reliability of Deep Neural Network Results, *Entropy* **25**(4) (2023), 573. doi:10.3390/e25040573.
- [33] C.G. Walsh, K. Sharman and G. Hripesak, Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk, *Journal of Biomedical Informatics* **76** (2017), 9–18. doi:https://doi.org/10.1016/j.jbi.2017.10.008. https://www.sciencedirect.com/science/article/pii/S1532046417302277.
- [34] M.A. Reyna, C.S. Josef, R. Jeter, S.P. Shashikumar, M.B. Westover, S. Nemati, G.D. Clifford and A. Sharma, Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019, *Critical Care Medicine* **48**(2) (2020), 210–217–. doi:10.1097/ccm.0000000000004145. http://dx.doi.org/10.1097/CCM.0000000000004145.
- [35] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015, Software available from tensorflow.org. https://www.tensorflow.org/.
- [36] D.P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, arXiv, 2014. doi:10.48550/ARXIV.1412.6980. https://arxiv.org/abs/1412.6980.
- [37] K. Hamilton, A. Nayak, B. Božić and L. Longo, Is neuro-symbolic AI meeting its promises in natural language processing? A structured review, *Semantic Web* **15**(4) (2024), 1265–1306–. doi:10.3233/sw-223228. http://dx.doi.org/10.3233/SW-223228.
- [38] G. Guizzardi, A.B. Benevides, C.M. Fonseca, J. ao Paulo A. Almeida, T.P. Sales and D. Porello, Ufo: Unified Foundational Ontology, *Applied ontology* **1**(17) (2022), 167–210. doi:10.3233/ao-210256.
- [39] J. Lazar, J.H. Feng and H. Hochheiser, *Research Methods in Human-Computer Interaction, Second Edition*, 2 edition edn, Morgan Kaufmann, Cambridge, MA, 2017. ISBN 978-0-12-805390-4.
- [40] G. Guizzardi and N. Guarino, Explanation, semantics, and ontology, *Data Knowledge Engineering* **153** (2024), 102325. doi:https://doi.org/10.1016/j.datak.2024.102325. https://www.sciencedirect.com/science/article/pii/S0169023X24000491.

- [41] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey and S. Linkman, Systematic Literature Reviews in Software Engineering – A Systematic Literature Review, *Information and Software Technology* **51**(1) (2009), 7–15. doi:10.1016/j.infsof.2008.09.009.
- [42] P.A. Chalmers, The Role of Cognitive Theory in Human–Computer Interface, *Computers in Human Behavior* **19**(5) (2003), 593–607. doi:10.1016/S0747-5632(02)00086-9.
- [43] E.K. Choe, B. Lee, H. Zhu, N.H. Riche and D. Baur, Understanding Self-Reflection: How People Reflect on Personal Data through Visual Data Exploration, in: *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, Association for Computing Machinery, Barcelona, Spain, 2017, pp. 173–182. ISBN 978-1-4503-6363-1. doi:10.1145/3154862.3154881.
- [44] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014. <https://arxiv.org/abs/1406.1078>.
- [45] I. Bibi, A. Akhunzada, J. Malik, J. Iqbal, A. Musaddiq and S. Kim, A Dynamic DL-Driven Architecture to Combat Sophisticated Android Malware, *IEEE Access* **8** (2020), 129600–129612–. doi:10.1109/access.2020.3009819. <http://dx.doi.org/10.1109/ACCESS.2020.3009819>.
- [46] P.H. Cleverley and S. Burnett, The Best of Both Worlds: Highlighting the Synergies of Combining Manual and Automatic Knowledge Organization Methods to Improve Information Search and Discovery, *Knowledge Organization* **42**(6) (2015), 428–444. doi:10.5771/0943-7444-2015-6-428.
- [47] C. Cunha and L. Cintra, *Nova Gramática do Português Contemporâneo*, Edição: 7ª edn, Lexikon, Rio de Janeiro, 2016. ISBN 978-85-8300-026-6.
- [48] W. D'Alessandro, Viewing-as Explanations and Ontic Dependence, *Philosophical Studies* **177**(3) (2020), 769–792. doi:10.1007/s11098-018-1205-5.
- [49] D.P.S. Dickinson, J. Teather, S. Gallina and E. Newsom-Davis, Medicine Package Leaflets – Does Good Design Matter?, *Information Design Journal* **18**(3) (2010), 225–240. doi:10.1075/idj.18.3.05dic.
- [50] A. Dix, Human–Computer Interaction, Foundations and New Paradigms, *Journal of Visual Languages & Computing* **42** (2017), 122–134. doi:10.1016/j.jvlc.2016.04.001.
- [51] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.-D. Chiche, C.M. Coopersmith, R.S. Hotchkiss, M.M. Levy, J.C. Marshall, G.S. Martin, S.M. Opal, G.D. Rubinfeld, T. van der Poll, J.-L. Vincent and D.C. Angus, The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), *JAMA* **315**(8) (2016), 801. doi:10.1001/jama.2016.0287.
- [52] M.Y. Yan, L.T. Gustad and Nytrø, Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review, *Journal of the American Medical Informatics Association* **29**(3) (2021), 559–575. doi:10.1093/jamia/ocab236.
- [53] L. Maddalena and F. Baião, OntoCovid: Aplicando SABiO para a modelagem conceitual bem fundamentada no domínio da COVID-19.(OntoCovid: Applying SABiO to conceptual modeling well grounded in the COVID-19 domain)., in: *ONTOBRAS*, 2021, pp. 259–266.
- [54] H.N. Hastenreiter Filho, I.T. Peres, L.G. Maddalena, F.A. Baião, O.T. Ranzani, S. Hamacher, P.M. Maçaira and F.A. Bozza, What we talk about when we talk about COVID-19 vaccination campaign impact: a narrative review, *Frontiers in Public Health* **11** (2023). doi:10.3389/fpubh.2023.1126461.
- [55] J. Villar, L. Maddalena, T. de Abreu Camargo, P. Medina Maçaira, F. Baião and F.L. Cyrino Oliveira, A statistical analysis of COVID-19 mortality dynamics: Unraveling the interplay between vaccination trends, socioeconomic factors, and government interventions in Brazilian states, *Socio-Economic Planning Sciences* **92** (2024), 101855. doi:<https://doi.org/10.1016/j.seps.2024.101855>. <https://www.sciencedirect.com/science/article/pii/S0038012124000545>.
- [56] J.N. Otte, J. Beverley and A. Ruttenberg, Bfo: Basic Formal Ontology, *Applied ontology* **17**(1) (2022), 17–43. doi:10.3233/ao-220262.
- [57] N. Guarino, D. Oberle and S. Staab, *What Is an Ontology?*, 2009, pp. 1–17. ISBN 9783540709992. doi:10.1007/978-3-540-92673-3_0.
- [58] M. Taye, Understanding Semantic Web and Ontologies: Theory and Applications, *Journal of Computing* **2** (2010).
- [59] N. Guarino, Formal Ontology in Information Systems (FOIS), 2000, pp. 3–15.
- [60] G. Amaral, T.P. Sales and G. Guizzardi, Towards an Ontology Network in Finance and Economics, in: *Advances in Enterprise Engineering XV*, Vol. 441, D. Aveiro, H.A. Proper, S. Guerreiro and M. De Vries, eds, Springer International Publishing, pp. 42–57, Series Title: Lecture Notes in Business Information Processing. ISBN 978-3-031-11519-6 978-3-031-11520-2. doi:10.1007/978-3-031-11520-2_4.
- [61] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng and H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* **101**(23) (2000), E215–20.
- [62] A. Johnson, T. Pollard, S. Horng, L.A. Celi and R. Mark, MIMIC-IV-Note: Deidentified free-text clinical notes, PhysioNet. doi:10.13026/1N74-NE17. <https://physionet.org/content/mimic-iv-note/2.2/>.

- [63] V.A. Carvalho, J.P.A. Almeida, C.M. Fonseca and G. Guizzardi, Multi-level ontology-based conceptual modeling **109**, 3–24. doi:10.1016/j.datak.2017.03.002. <https://linkinghub.elsevier.com/retrieve/pii/S0169023X17301052>.
- [64] K. Denecke, Does Enrichment of Clinical Texts by Ontology Concepts Increases Classification Accuracy?, in: *Studies in Health Technology and Informatics*, P. Otero, P. Scott, S.Z. Martin and E. Huesing, eds, IOS Press. ISBN 978-1-64368-264-8 978-1-64368-265-5. doi:10.3233/SHTI220148.
- [65] N. Guarino, The Ontological Level: Revisiting 30 Years of Knowledge Representation, in: *Conceptual Modeling: Foundations and Applications*, Vol. 5600, A.T. Borgida, V.K. Chaudhri, P. Giorgini and E.S. Yu, eds, Springer Berlin Heidelberg, pp. 52–67, Series Title: Lecture Notes in Computer Science. ISBN 978-3-642-02462-7 978-3-642-02463-4. doi:10.1007/978-3-642-02463-4₄.
- [66] N. Guarino (ed.), Formal ontology in information systems: proceedings of the first international conference (FOIS '98), June 6 - 8, Trento, Italy, *Frontiers in artificial intelligence and applications*, IOS-Press [u.a.], Meeting Name: FOIS. ISBN 978-90-5199-399-8 978-4-274-90223-9.
- [67] N. Guarino and G. Guizzardi, Relationships and Events: Towards a General Theory of Reification and Truthmaking, in: *AI*IA 2016 Advances in Artificial Intelligence*, Vol. 10037, G. Adorni, S. Cagnoni, M. Gori and M. Maratea, eds, Springer International Publishing, pp. 237–249, Series Title: Lecture Notes in Computer Science. ISBN 978-3-319-49129-5 978-3-319-49130-1. doi:10.1007/978-3-319-49130-1₁₈.
- [68] N. Guarino and G. Guizzardi, Relationships and Events: Towards a General Theory of Reification and Truthmaking, in: *AI*IA 2016 Advances in Artificial Intelligence*, Vol. 10037, G. Adorni, S. Cagnoni, M. Gori and M. Maratea, eds, Springer International Publishing, pp. 237–249, Series Title: Lecture Notes in Computer Science. ISBN 978-3-319-49129-5 978-3-319-49130-1. doi:10.1007/978-3-319-49130-1₁₈.
- [69] G. Guizzardi, C.M. Fonseca, J.P.A. Almeida, T.P. Sales, A.B. Benevides and D. Porello, Types and taxonomic structures in conceptual modeling: A novel ontological theory and engineering support **134**, 101891. doi:10.1016/j.datak.2021.101891. <https://linkinghub.elsevier.com/retrieve/pii/S0169023X21000185>.
- [70] G. Guizzardi, Ontological foundations for structural conceptual models..
- [71] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R.G. Mark, MIMIC-III, a freely accessible critical care database **3**(1), 160035. doi:10.1038/sdata.2016.35. <https://www.nature.com/articles/sdata201635>.
- [72] L. Maddalena and F. Baião, OntoCovid: Applying SABiO to conceptual modeling well grounded in the COVID-19 domain, pp. 259–266. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123294670&partnerID=40&md5=d1876b5a1b28e2f989ae2188a00ea0d8>.
- [73] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan and D.S.W. Ting, Large language models in medicine **29**(8), 1930–1940. doi:10.1038/s41591-023-02448-8. <https://www.nature.com/articles/s41591-023-02448-8>.
- [74] Z. Zhuang, Q. Chen, L. Ma, M. Li, Y. Han, Y. Qian, H. Bai, Z. Feng, W. Zhang and T. Liu, Through the Lens of Core Competency: Survey on Evaluation of Large Language Models, Publisher: arXiv Version Number: 1. doi:10.48550/ARXIV.2308.07902. <https://arxiv.org/abs/2308.07902>.
- [75] F. Amrollahi, S.P. Shashikumar, F. Razmi and S. Nemati, Contextual Embeddings from Clinical Notes Improves Prediction of Sepsis. *AMIA, Annual Symposium proceedings. AMIA Symposium 2020* (2021), 197–202.
- [76] Z. Zhang, X. Liu, Y. Zhang, Q. Su, X. Sun and B. He, Pretrain-KGE: Learning knowledge representation from pretrained language models, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020.
- [77] X. Wang, Q. He, J. Liang and Y. Xiao, Language Models as Knowledge Embeddings (2022).