Semantic-based Data Augmentation for Machine Learning Prediction Enhancement

Majlinda Llugiqi [a,*], Fajar J. Ekaputra [a] and Marta Sabou [a]
[a] *Institute for Data, Process and Knowledge Management, Vienna University of Economics and Business, Vienna, Austria*
E-mails: majlinda.llugiqi@wu.ac.at, fajar.ekaptura@wu.ac.at, marta.sabou@wu.ac.at

Abstract.
Machine learning (ML) methods have demonstrated strong predictive capabilities when trained on large datasets. However, in domains where data is scarce or sensitive, ML models often exhibit suboptimal performance. Our hypothesis is that semantically enriching the available training dataset can enhance the predictive power of ML models, particularly in data-scarce scenarios. To investigate this hypothesis, we propose novel neuro-symbolic approaches that augment tabular data with KG information, providing additional context and structure to improve model performance. Concretely, we introduce and examine several integration techniques of KG information through embeddings and explore how different KG embedding algorithms affect model performance, with a specific focus on accuracy and F2 scores. Our evaluation involves four distinct ML algorithms and four KG embedding techniques. We apply our approach to binary classification tasks on tabular data, including heart disease and chronic kidney disease. Our experimental results show improvements in performance particularly when tabular data is augmented with distance features computed in the embedding space. Notably, we achieve gains in F2 scores, such as an increase in XGBoost performance from 75.19% to 90.85% for heart disease prediction. These findings demonstrate the potential of KG-based augmentation to enhance ML performance.

Keywords: Neuro-symbolic AI, Knowledge Graph Embeddings, Machine Learning, Data Augmentation

1. Introduction

Machine learning (ML) has revolutionized various domains by providing powerful tools for pattern recognition, predictive analytics, and data-driven decision-making. Techniques such as deep learning have achieved remarkable success in fields ranging from computer vision [13, 55] to natural language processing [31, 41]. These advancements have been largely driven by the availability of large datasets and the computational power to process them.

However, ML methods often face significant challenges related to data quality and availability. Data sparsity, imbalance, and sensitivity can severely hinder the performance of ML models [39]. In the medical domain, one important task is predicting patient outcomes, for instance, determining the presence or absence of a disease based on clinical observations. This task often suffers from an insufficient amount of labeled data due to privacy concerns [27]. Although advances have been made, models trained solely on tabular data fail to fully capture the domain's complexity and semantics, limiting their ability to generalize effectively [44].

To overcome these limitations, neuro-symbolic (NeSy) AI has emerged as a promising approach to integrate domain knowledge into ML models. NeSy AI combines the strengths of symbolic AI—known for logical reasoning

*Corresponding author. E-mail: majlinda.llugiqi@wu.ac.at.

and explainability—with sub-symbolic methods such as deep learning [16, 23, 45]. In particular, structured semantic knowledge such as knowledge graphs (KGs) has emerged as a key element in bridging the gap, providing a structured way to represent relationships between entities and capture domain-specific semantics [4, 17, 22, 59]. KGs have been widely used in tasks such as knowledge graph completion [34] and link prediction [53]. However, their potential to enhance ML predictions on tabular data by incorporating semantic knowledge through embeddings remains underexplored.

We propose integrating KGs into ML pipelines to enhance tabular data with structured, domain-specific information. Drawing upon techniques from the Semantic Web community, our approach begins by utilizing ontologies to formalize domain semantics. We then construct KGs based on these ontologies, enriching the datasets with structured knowledge specific to the medical domain. Subsequently, we employ knowledge graph embeddings to transform the KGs into numerical vector representations suitable for ML algorithms. By embedding relationships and domain knowledge from KGs into these vectors, our methodology enhances the ML pipeline by augmenting the datasets with semantic knowledge, aiming to improve predictive performance—especially in data-scarce domains. This study specifically explores binary classification tasks in both medical predictions (heart disease and chronic kidney disease) where domain-specific structure is crucial for robust prediction. Our research is guided by the following research questions:

- RQ1: How can KGs be optimally infused into an ML pipeline to enhance performance in terms of accuracy and F2 score?
- RQ2: How does the choice of knowledge graph embedding algorithms affect the performance of machine learning models when used to augment tabular data?
- RQ3: How do different ML algorithms perform when KG-based information is integrated into the input data?

To address these research questions, we took an exploratory approach, systematically investigating each aspect step by step. For RQ1, we derived five sub-hypotheses to examine how knowledge graphs can be optimally integrated into ML pipelines to enhance performance metrics such as accuracy and F2 score (with the reason for selecting these metrics explained in Section 6.2). We tested these hypotheses using eight different approaches, each incorporating knowledge graphs and embeddings in various ways. For RQ2 and RQ3, we empirically evaluated the impact of different knowledge graph embedding algorithms and ML models across two medical domains—heart disease and chronic kidney disease prediction.

Building on our previous work [35], we extend and formalize our methodology for integrating KG embeddings into ML pipelines. We employ two additional embedding techniques alongside those used previously to transform the KGs into numerical vector representations suitable for ML algorithms. We developed and tested different approaches based on five sub-hypotheses derived from our first research question, providing a comprehensive evaluation of their impact on model performance in heart and kidney disease prediction. Our study demonstrates the effectiveness of incorporating ontological knowledge into the ML training process, highlighting the potential for improved predictive performance in data-scarce domains and its applicability across various fields where ontologies can be developed or expanded.

The remainder of this paper is organized as follows: In Section 2, we define the key concepts that we use in our work. This is followed by an overview of related work in Section 3. In Section 4, we present an overview of our proposed approach, with more detailed explanations about our approaches provided in Section 5. Our experimental analysis is discussed in Section 6, where we outline the goals and setup of our experiments. In Section 7, we present and analyze the outcomes of our experiments. Finally, we summarize our findings and outline directions for future work in Section 8.

2. Problem Description and Background Information

In this section, we outline the problem we aim to address, followed by introducing the key concepts that we use throughout the paper, beginning with ontologies, knowledge graphs, and knowledge graph embeddings.

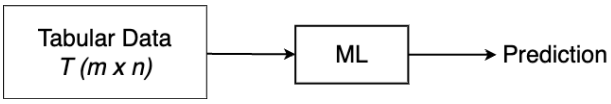

Fig. 1. Baseline for ML prediction on tabular data.

2.1. Problem Description

In this study, we address the challenge of predicting heart disease and chronic kidney disease using patient medical records in tabular data format. Each dataset can be represented as a table $T \in \mathbb{R}^{n \times m}$, where n is the number of patient instances and m is the number of features or attributes. These features capture patient demographics, clinical measurements, and diagnostic information essential for disease prediction. For heart disease prediction, we consider features such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, and resting electrocardiogram results. The kidney disease dataset similarly includes essential attributes, including age, blood pressure, specific gravity, albumin, blood glucose, blood urea, serum creatinine, hemoglobin, and red and white blood cell counts.

We focus on binary classification to predict the presence or absence of these diseases. Formally, given the dataset T, the goal is to learn a function $f : \mathbb{R}^m \rightarrow \{0, 1\}$ that maps a patient's feature vector to a binary outcome indicating disease presence (1) or absence (0). As illustrated in Figure 1, the tabular data T serves as input to machine learning models, which then output predictions regarding disease presence.

For example, given a patient's data (e.g., age 62, female, asymptomatic, resting blood pressure 140, cholesterol 268, no fasting blood sugar, max heart rate 160, downsloping slope and thalassemia), our model aims to determine the likelihood of heart disease. Similarly, a record for kidney disease might involve attributes such as age 68, blood pressure 70, specific gravity 1.01, and blood urea 54. The objective is to accurately predict disease presence.

Due to the sensitive nature of medical data, datasets in this domain are often limited or partially incomplete, impacting model performance. This scarcity of data, combined with varying data quality, presents a challenge to achieving optimal prediction accuracy, necessitating robust preprocessing and, potentially, data augmentation strategies to improve model generalizability and reliability.

2.2. Background Information

Given the problem definition described in the previous subsection, our approach aims to augment these datasets by integrating semantic information to enhance predictive capabilities. To achieve this, we leverage *ontologies* to capture the domain knowledge, and then we use *knowledge graphs* to enrich the datasets with ontologies. We then need *knowledge graph embeddings* to transform the knowledge graphs into a vector space suitable for machine learning models. In the following we discuss each of these concepts in detail.

Ontology: Originally a philosophical term, ontology refers to the study of existence and the nature of being. In computer science, Gruber [20] redefined ontology as "explicit specifications of conceptualizations", where a conceptualization represents a simplified, abstract view of a domain to capture essential aspects. An ontology establishes a standardized vocabulary for knowledge sharing within a specific domain. Formally, an ontology represented as $O = (C, R, H^C)$ encompasses a collection of concepts C, a set of relations R, and a hierarchical structure of concepts H^C. Each relation $r \in R$ indicates an association between pairs of concepts, such that $R \subseteq C \times C$. The concept hierarchy H^C is a subset of $C \times C$, illustrating the relationships among concepts.

Knowledge Graphs: Knowledge graphs (KGs) expand on ontologies by capturing not only the structured relationships between concepts but also the specific instances and values within a domain. Originally popularized by Google in 2012 [48] to enhance search understanding, KGs have since become integral in a range of applications, providing a structured, machine-readable format to represent knowledge. We define the KG as $KG = (E, R', L, Tr)$ where:

- E represents the set of entities in the knowledge graph. Each entity $e \in E$ can represent a real-world concept, object or idea, such as 'Person' or 'City'.

- R' represents the set of instantiated relations between entities within the KG such as 'hasAge' or 'worksAt'.
- L represents the set of literals, which are attributes associated with entities, such as numerical values or textual descriptions (e.g., '30' or 'Alice').
- Tr denotes a set of triples, where each triple $tr = (e_1, e, e_2) \in Tr$ represents a fact or statement in the knowledge graph.

KG embeddings: While KGs provide a structured representation of entities and their relationships, they can become highly complex as the number of entities and relations grows. To enable efficient computation, learning, and reasoning over KGs, knowledge graph embeddings (KGEs) are commonly used [5, 34, 54]. KG embeddings transform entities and relations from a discrete symbolic space into a continuous vector space, capturing the structure and semantics of the KG in a form that is compatible with ML algorithms. KGE algorithms can be broadly categorized into three main types based on their methodology and objectives: translational distance models, semantic matching models, and random walk-based models. In the following, we briefly describe the embedding algorithms used in our experiments: Node2Vec [19] and Rdf2Vec [43] as random-walk based models that leverage the graph structure, DistMult [57] as a semantic matching model, and TransH [54] as a translational model.

- Node2Vec uses a flexible random walk strategy to combine depth-first and breadth-first sampling, allowing it to capture various structural features of the graph whether they are labeled or unlabeled, directed or undirected. Node2Vec employs random walks, incorporating an adjustable bias parameter that allows for targeted exploration of local neighborhoods as well as a broader global search.
- RDF2Vec is designed specifically for RDF (Resource Description Framework) graphs within the Semantic Web, RDF2Vec generates embeddings for entities and relations by leveraging random walks to create sequences from the graph. These sequences are then transformed into embeddings using Word2Vec, making RDF2Vec particularly effective at capturing the semantic and relational attributes present in RDF data. While both RDF2Vec and Node2Vec utilize random walks, RDF2Vec focuses more on semantic relationships within the context of the Semantic Web, whereas Node2Vec emphasizes structural characteristics applicable to a wider range of graph types.
- DistMult is a semantic matching model that uses a bilinear scoring function to evaluate the interactions between entities and relations in a knowledge graph. In this model, each relation is represented as a diagonal matrix, simplifying the bilinear form to a weighted element-wise multiplication of entity embeddings. While this approach effectively captures pairwise relationships, it inherently assumes that all relations are symmetric, which may restrict its expressiveness for datasets containing asymmetric relations.
- TransH is a translational model that represents entities as vectors and relations as hyperplanes in the embedding space. Each relation is associated with a specific hyperplane and a translation vector on that hyperplane. Entities are projected onto the hyperplane of a relation before the translation operation is applied. This method allows entities to have different representations in the context of different relations, enabling the model to capture complex and diverse relationships thereby improving its ability to represent multiple types of relationships in a knowledge graph.

3. Related work

We review related work on (i) the categorization of neuro-symbolic approaches, positioning our work within these categories, (ii) we discuss the use of ML models in disease prediction and (iii) enhancing ML predictions with semantic knowledge, and we conclude by discussing the novelty of our approach.

Categorization of Neuro-Symbolic Approaches In recent years, the field of neuro-symbolic AI has gained significant attention due to its potential to combine the strengths of both symbolic and sub-symbolic AI [16, 23, 45]. Symbolic AI excels at logical reasoning and explainability, while sub-symbolic approaches, such as deep learning, have proven effective in pattern recognition and data-driven decision-making. Combining these approaches, neuro-symbolic AI seeks to leverage the best of both worlds: the learning capability of sub-symbolic methods and

the structured, interpretable reasoning of symbolic methods. Several efforts have focused on categorizing neuro-symbolic approaches. Kautz et al. [30] classify neurosymbolic systems into six types based on the interaction between neural networks and symbolic reasoning. *Type 1* employs standard deep learning with symbolic inputs and outputs, while *Type 2* combines neural networks with symbolic solvers, as seen in systems such as AlphaGo. *Type 3* uses neural networks for tasks such as object detection, while symbolic systems handle complementary tasks such as query answering. In *Type 4*, symbolic knowledge is embedded into neural network training, whereas *Type 5* incorporates symbolic rules as constraints in the loss function. Finally, *Type 6* aims for fully integrated systems, merging symbolic reasoning with neural architectures, although fully mature combinatorial reasoning within such systems remains a challenge. Our approach belongs to *Type 4* of Kautz's classification, where symbolic knowledge is incorporated into the training process.

Similarly, Sheth et al. [33, 47] identify three levels of knowledge infusion in neural models: shallow, semi-deep, and deep. *Shallow infusion* introduces syntactic and symbolic knowledge at the input level, *semi-deep infusion* introduces external knowledge into intermediate layers via attention mechanisms or constraints, and *deep infusion* embeds structured, multi-layered knowledge into the network itself, aligning abstraction layers with learning stages. Our work adopts the *shallow infusion* approach by enriching input data with syntactic and symbolic knowledge, enhancing the model's performance.

Dash et al. [12] categorize methods for integrating domain-specific knowledge into deep neural networks into three main approaches: enhancing input data, modifying the loss function, and adjusting the network architecture. Our research aligns with the *input transformation* category, where domain-specific knowledge is integrated by enriching the input data provided to the ML models.

Van Harmelen and ten Teije [51] introduced a conceptual framework known as "boxology," which outlines various patterns for integrating machine learning with semantic web technologies. Breit et al. [6] expanded this framework by identifying 44 distinct patterns used in hybrid learning and reasoning techniques, based on a review of around 500 papers from 2010 to 2020. Our approach falls under the *T* patterns, specifically *T4*, where input transformations using symbolic knowledge are applied to improve model performance.

As a summary, our approach falls under the *shallow infusion* category as described by Sheth et al. [47], where syntactic and symbolic knowledge is introduced at the input level. It aligns with *Type 4* in Kautz's classification [30], as symbolic knowledge is embedded into the training process. Furthermore, it belongs to the *input transformation* approach discussed by Dash et al. [12], where domain-specific knowledge enhances the input data provided to machine learning models. Finally, our work corresponds to the *T4* pattern in the "boxology" framework, focusing on input transformations to improve model performance.

Machine Learning Models in Disease Prediction The application of ML in healthcare has attracted significant research interest due to its potential. Kraivsnikovic et al. [32] proposed an approach leveraging fine-tuned BERT models to analyze German pathology reports. Their work highlights how domain-specific adaptations can enhance the interpretability and utility of ML models in medical diagnostics by effectively capturing contextual representations. Additionally, ML algorithms have been successfully employed in predicting diseases such as heart disease [29, 42, 46, 56] and kidney disease [8, 40, 52, 58], using various techniques such as data preprocessing, feature selection, and hyperparameter tuning to enhance prediction accuracy.

Several studies have also explored combining ML methods to further improve performance. For instance, Mohan et al. [36] combined random forest and linear methods to enhance heart disease prediction, while Ali et al. [2] introduced a framework for heart failure prediction using dual support vector machine (SVM) models—one for feature selection and the other for the prediction task.

Although these models demonstrate good predictive performance, they often rely on extensive preprocessing [21], feature selection, and hyperparameter tuning to achieve optimal results. Moreover, the effectiveness of ML models can be limited by insufficient or sub-optimal quality data. In this context, healthcare ontologies [9, 14, 26, 28, 38] offer a structured, semantically rich layer of information that can enhance the contextual understanding of ML models, which is further explored in this work.

Enhancing ML Predictions with Semantic Knowledge Recent research has increasingly focused on integrating semantic knowledge, such as KGs and ontologies, into ML models to enhance their performance. KGs have

been widely applied in various domains, notably improving feature extraction and entity representation in natural language processing tasks. For instance, Moussallem et al. [37] demonstrated how augmenting neural machine translation systems with KGs improved the translation quality by enhancing the semantic understanding of terminological expressions. Similarly, KG-based input enhancement has been shown to improve recommendation systems and community detection, enhancing both accuracy and explainability [4].

Table 1
Summary of related work on integrating semantic knowledge into ML models

Paper	Domain/Task	KG/Ontology	ML Algorithm	KGE	Method of Including the KG
[24]	Classification on tabular and image data	DBPedia, Wikidata, YAGO3, ConceptNet	Neural networks		Logic rules mining from KGs during learning
[18]	Healthcare / Predicting hospitalization	DBpedia, Wikidata domain specific[1]	SVM, RF, LogReg		Enriching EMR with features from ontologies
[49]	Smart building management	custom	Neural networks		Integrating data from sensors with knowledge
[44]	High-dimensional tabular learning	Custom auxiliary KG	Multilayer Perceptron (MLP)		KG to regularize a MLP for tabular datasets
[37]	Neural machine translation	DBPedia	RNN and Transformer	✓	1) Entity Linking + KGE 2) Semantic enrichment of KGE via entity labels
[3]	Text classification and natural language inference	Freebase, WordNet	LSTM	✓	KG embeddings injected into the model for enriched representation learning
[60]	Credit card fraud detection	DBPedia	Deep neural network	✓	Augmentation of dataset with semantic vector representation of countries and public holidays information
Our paper	Disease prediction Healthcare	SNOMED Custom KG	KNN, NN, SVM, XGBoost	✓	Augmenting tabular data with KG embeddings

Moreover, KG-augmented neural networks have demonstrated improved performance in text classification and natural language inference tasks. Annervaz et al. [3] showed that integrating structured knowledge from KGs not only improved model accuracy but also allowed models to perform well with less labeled data, addressing the common issue of data sparsity. Ziegler et al. [60] adopted a similar approach by incorporating semantic knowledge through graph embeddings for credit card fraud detection, demonstrating how the injection of background knowledge—such as public holidays from DBpedia into neural models could enhance classification outcomes.

In addition to NLP and fraud detection, Szilagyi et al. [49] applied semantic knowledge in smart building management by integrating taxonomies, schemas, and logic rules with ML models. This hybrid system optimized building management by combining data-driven insights with rule-based reasoning, showing the potential of semantic knowledge in enhancing decision-making processes. Huang et al. [24] introduced an Abductive Learning with KG approach that automatically mines logic rules from KGs and integrates them into ML models using a knowledge-forgetting mechanism to filter irrelevant information, thereby improving model performance even with limited labeled data.

In healthcare, Gazzotti et al. [18] demonstrated how augmenting sparse electronic medical records (EMRs) with ontological resources improved the predictive capabilities of ML algorithms, specifically in hospitalization prediction. Ontologies such as DBPedia, Wikidata and the more domain specific ones provide structured medical knowledge, enabling a richer representation of patient data. Similarly, Ruiz et al. [44] introduced the PLATO method, which uses a KG to regularize a multilayer perceptron for tabular datasets, showing that semantic knowledge can help ML models handle high-dimensional and low-sample-size data more effectively.

These studies highlight the growing importance of integrating semantic knowledge into ML models to address challenges such as data quality, sparsity, and explainability across different domains. Table 1 provides an overview

[1] Anatomical Therapeutic Chemical Classification, National Drug File - Reference Terminology, International Primary Care Classification

of these studies, outlining the types of semantic knowledge used, the domains or tasks covered, the ML models applied, the incorporation of KGEs, and the integration methods employed. These approaches generally fall into two main categories: (i) direct integration of structured knowledge through explicit rules or ontological features and (ii) representation learning via KGEs, where entities and relations are embedded into a continuous vector space, allowing downstream ML models to leverage the semantic structure.

Although this prior research has demonstrated that embedding-based methods can improve ML performance, most existing work relies on large, general-purpose KGs, such as Wikidata or DBpedia. In contrast, our recent study [35] introduced four approaches for augmenting tabular data with KGEs using two embedding algorithms, focusing on smaller, domain-specific ontologies. These initial methods illustrated the potential of embedding-driven enrichment exploring how semantic context could be systematically incorporated into tabular datasets.

Building on that foundation, this paper proposes four additional approaches that calculate various metrics in the embedding space to enrich tabular data with additional semantic context. Furthermore, we employ two more KG embedding algorithms, extending the methodology and formalizing our approaches. We also perform a more thorough evaluation of the proposed techniques, applying them to the prediction of heart disease and chronic kidney disease.

Compared to other studies shown in Table 1, our method goes beyond simply embedding entities and relations, we exploit the embedding space itself to derive meaningful metrics that further enrich tabular features. Moreover, instead of relying on extensive, generic KGs, our approach leverages small, existing domain-specific ontologies (or select subsets of existing big ontologies), which we populate with relevant tabular data to form task-specific KGs. Such domain-focused strategies remain underexplored within medical prediction and systems. By emphasizing smaller ontologies and extracting deeper semantic insights from the embedding space, our work aims to advance semantic knowledge integration for ML in medical and other specialized domains.

4. Knowledge Graph Embedding-Based Augmentation for Tabular Data

To improve the performance of ML models, we leverage KGs to enrich tabular datasets with semantic information. This section outlines the two core steps of our approach. First, in Section 4.1, we explain the construction of KGs using instances from the tabular data, as shown in the top part of Figure 2. This step focuses on building KGs that capture deeper relationships within the data. Next, in Section 4.2, we focus on integrating KG embeddings into the ML pipeline. This includes various augmentation strategies designed to enhance model performance by incorporating structural and relational information from the KG into the training data, as depicted in the bottom part of Figure 2.

4.1. Knowledge Graph Construction

For our approach to enrich ML input data with supplementary knowledge, constructing KGs is essential. They serve as structured representations of domain knowledge, capturing the semantics of the data and allowing for the integration of ontological information into datasets. This enrichment allows ML models to leverage contextual and relational information, enhancing their predictive capabilities. The upper part of Figure 2 illustrates the methodology used for building these KGs, which represent data that was initially captured in tabular form. The following steps provide a formal description of this construction process.

Step 1: Ontology Definition The first step in constructing the KG is defining an ontology, which is used to capture domain semantics and provide a structured framework for enriching the datasets. There are different ways to develop an ontology, represented as $O = (C, R, H^C)$. We considered (i) creating a new ontology from scratch, (ii) extending and reusing existing ontologies to include additional domain-specific information, or (iii) extracting relevant components from a more extensive ontology (see Section 6.2).

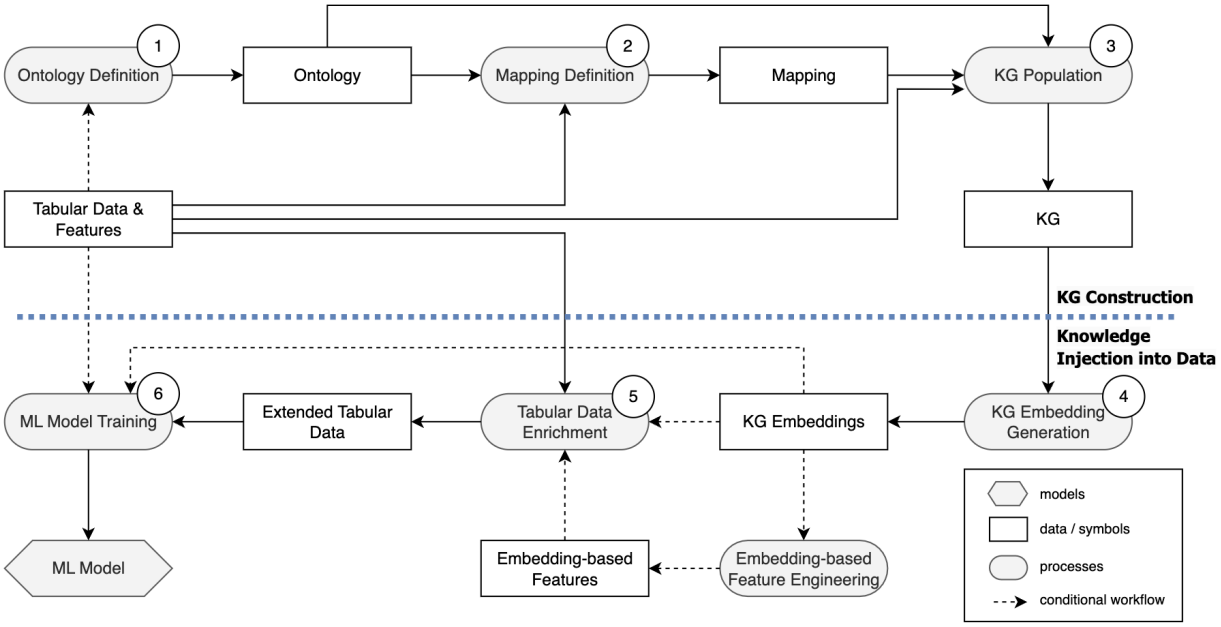

Fig. 2. Overview of the proposed approach including (i) KG construction (top) and (ii) knowledge injection into data (bottom) (adapted from [35] following the boxology notation [50]).

Step 2: Mapping Definition The process of mapping dataset features to the concepts in the ontology is crucial for using instances from tabular data to populate the ontology and, consequently, construct a KG. A key aspect of this mapping process is the **mapping function** $\psi : F \rightarrow C$, where $F = \{f_1, f_2, \ldots, f_n\}$ represents the features within the tabular dataset T, defined as a matrix of dimensions $m \times n$. This function entails the manual mapping of each feature f_i with a corresponding concept C in the ontology O.

Step 3: Knowledge Graph Population The knowledge graph KG is constructed by utilizing the ontology O along with the instances from the tabular data T and applying the mapping function $\psi : F \rightarrow C$. This process is automated through a Python script. We define the KG as $KG = (E, R', L, Tr)$ where:

- E signifies the set of entities, with each entity $e_i \in E$ corresponding to an instance in the tabular data derived from each row m_i in T,
- R' represents the set of instantiated relations within the KG, which includes relations from R through the mapping ψ, and illustrates direct relationships between entities E or between an entity and a literal value,
- L represents the set of literals, which are attribute values associated with entities, such as numerical data (e.g., 30) from T,
- Tr consists of triples generated for each feature value in an instance row m_i, following the mapping ψ. For instance, if a feature $f_{\text{glucoseLevel}}$ corresponds to an instance e_i with a blood pressure value of 95, the associated triple would be $(e_i, r_{\text{hasGlucoseLevel}}, 95)$, indicating the relationship $r_{\text{hasGlucoseLevel}}$ between entity e_i and the literal value 95.

This preprocessing phase ensures that features from the tabular data T are semantically represented within the Knowledge Graph KG using the defined ontology O.

4.2. Integrating KG Embeddings into ML Pipeline

In Section 4.1, we outlined the construction of enriched data structures that capture deeper semantics beyond the raw data. This section will now focus on transforming these enriched structures into a vectorized format suitable for ML, and on the optimal strategies for augmenting the input data, as illustrated in the lower part of Figure 2.

Step 4: Knowledge Graph Embedding Generation With the populated KG with enriched data structures, the subsequent step is to prepare the KG for ML model training. This requires transforming the KG into a vector space representation suitable for ML models, using knowledge graph embedding (KGE) algorithms. Having a knowledge graph $KG = (E, R)$, the goal of the embedding algorithm is to map entities E and relations R into a continuous vector space. Formally, this can be represented as function: $\phi : E \cup R \rightarrow \mathbb{R}^d$, where ϕ is the embedding function that maps each entity and relation in the KG to a d -dimensional real-valued vector in the vector space $\mathbb{R}^d$. This transformation allows the KG to be represented in a way that preserves its semantic information while being computationally efficient. In the next steps 5 & 6 we will see how these embeddings are used as such or to compute features that are added to augment the dataset for a better ML performance.

Step 5 & 6: Tabular Data Enrichment and ML Model Training After computing KGEs, our objective is to explore the integration of these embeddings to enhance the performance of ML models. We experimented with different approaches for augmenting the training set using KGEs. First, we established a baseline that trains ML models using only tabular data T, following the traditional approach, shown in Figure 1, where no KG information is being added. Then we experimented with different ways for enhancing the dataset with KGEs and training the ML models, which are shown in details in the following section.

5. Proposed Approaches for Tabular Data Enrichment and ML Model Training

In this section, we outline the eight distinct approaches we explored for integrating KG embeddings into the training dataset, each designed to evaluate the impact of enriched semantic information on model performance. [2]

5.1. Embeddings as ML Model Inputs (EmbedOnly)

We begin by our initial objective to investigate whether training a model on the vector representations generated from these KGs, using various embedding algorithms, could reveal underlying patterns and relationships within the data. Therefore, we define our first sub-hypothesis as follows.

> ***H1.1**: Using the embeddings alone, without any additional tabular data, could provide meaningful insights and capture latent relationships that enhance the model's predictive capability.*

To explore this, we first explored the EmbedOnly approach, focusing solely on the embeddings to assess their standalone effectiveness in capturing meaningful insights as shown in Figure 3.

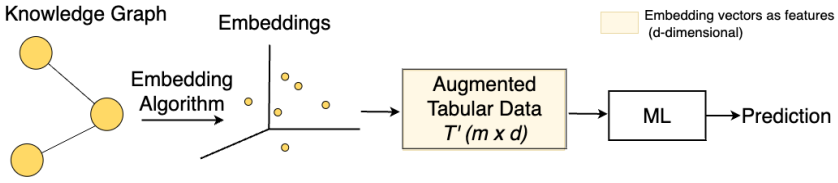

Fig. 3. Embedding vectors, highlighted in yellow, serve as inputs to the ML model.

For each instance p_i in the tabular data T, we have them represented as a subset $P \subseteq E$ of KG, where P represents the set of entities corresponding to instances in the tabular data. The embedding function: $\phi : P \cup R \rightarrow \mathbb{R}^d$ is used to map each instance entity $p_i \in P$ to a d-dimensional vector space. Consequently, during both the training and testing phases only the embeddings $\{\phi(p) \mid p \in P\} \subset \mathbb{R}^d$ derived from the instance entities are used, as outlined in Algorithm 1. This ensures that the model is trained and evaluated only on the vector representations, capturing the semantic relationships within the KG relevant to the instances.

[2] A visual representation of these approaches, following the Boxology notation [50], is available as supplementary material at: https://semsys.ai.wu.ac.at/data-augmentation/home.html.

Algorithm 1 EmbedOnly

Require: Knowledge Graph $KG = (E, R)$, Tabular Data T, Embedding Function $\phi : E \cup R \rightarrow \mathbb{R}^d$, ML Model M
Ensure: Trained ML model using only embeddings for n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
for each instance p_i in the training set T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Map entities to embeddings
    Append v_i to X_{train}
    Append label y_i corresponding to p_i to Y_{train}
end for
for each instance p_i in the test set T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Map entities (without y_i) to embeddings
    Append v_i to X_{test}
    Append label y_i corresponding to p_i to Y_{test}
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.2. Combining Embeddings with Tabular Data Features (EmbedAugTab)

Building upon EmbedOnly approach, we define our second sub-hypothesis as follows.

> ***H1.2**: Combining the KG-derived embeddings with traditional tabular data might enhance model performance by introducing additional relational information from the KG* structure.

This led us to design approaches that integrate both embeddings and tabular features, aiming to see if the KG information could complement and enrich the existing dataset. Thus, we investigated EmbedAugTab and other subsequent approaches that leverage embeddings for data augmentation based on this intuition.

EmbedAugTab approach involves training ML algorithms on datasets that integrate the original tabular data with additional columns derived from embeddings, as illustrated in Figure 4 and presented in Algorithm 2. For each instance p in the tabular dataset T, we augment T by appending the embedding vector $\phi(p)$, corresponding to the instance $p \in P$. The embedding vector $\phi(p)$ is generated using the embedding function $\phi : P \cup R \rightarrow \mathbb{R}^d$. This process yields an augmented tabular matrix T' with dimensions $m \times (n + d)$, where each row i contains the original features from T concatenated with the d-dimensional embedding vector $\phi(p)$. The resulting augmented matrix T' is then utilized to train ML models, leveraging both the original tabular features and the vector representations of the instances. In the healthcare domain, each instance p represents a patient, and the embedding vectors are added for each patient, in order to improve the models' ability to predict the presence or absence of specific diseases, such as heart disease or chronic kidney disease.

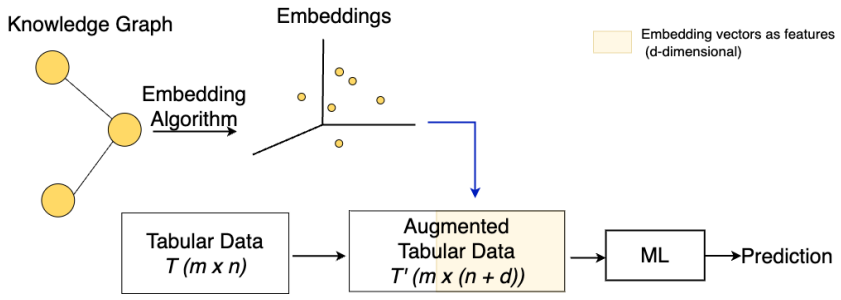

Fig. 4. Tabular dataset enrichment with embedding vectors, highlighted in yellow, used as inputs for the ML model.

Algorithm 2 EmbedAugTab

Require: Knowledge Graph $KG = (E, R)$, Tabular Data T, Embedding Function $\phi : E \cup R \to \mathbb{R}^d$, ML Model M
Ensure: Trained ML model M for each T_{train}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $X_{train} \leftarrow X_{train} \cup \text{Concatenate}(p_i, v_i)$ ▷ Append original features and embedding
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $X_{test} \leftarrow X_{test} \cup \text{Concatenate}(p_i, v_i)$ ▷ Append original features and embedding
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.3. Tabular Dataset Enrichment with Distance Measures from Knowledge Graphs (DistAugTab)

Utilizing embedding vectors directly to augment the tabular data may introduce noise. Thus, our sub-hypothesis is defined as follows.

> ***H1.3**: Extracting specific structural information from the embedding space, such as distance matrices or cluster characteristics, might enhance model performance by providing more interpretable features for distance-based models.*

This led us to introduce the DistAugTab and ClustAugTab approaches, which aim to selectively extract meaningful information from the embeddings to improve the learning process.

In DistAugTab approach, we enhance the tabular dataset T by incorporating additional features derived from embedding-based distance calculations, as illustrated in Figure 5 and presented in Algorithm 3. For each instance p_i in the dataset T, we compute its embedding vector $\vec{v}_i$ using the embedding function ϕ. To further enrich the representation of each instance, we introduce $|C|$ additional columns, where C denotes the set of target classes.

The new columns are calculated by determining the Euclidean distance between the embedding vector $\vec{v}_i$ of instance p_i and the centroid $\vec{c}_{C_j}$ of each target class $C_j \in C$. The centroid $\vec{c}_{C_j}$ is calculated as the mean of the embedding vectors $\vec{v}_i$ for all instances p_i belonging to the target class C_j. These distance-based features are added to the augmented dataset T', resulting in an expanded dataset with dimensions $m \times (n + |C|)$, where m is the number of instances and n is the original number of features.

By including these distance features, we aim to capture how closely each instance's embedding aligns with the class centroids, thereby potentially improving the model's ability to differentiate between target classes. For

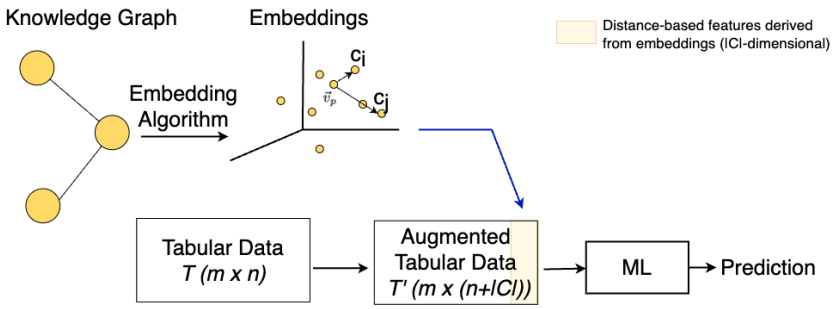

Fig. 5. Tabular dataset enrichment with distance measures from the KG (highlighted in yellow), used as inputs for the ML model.

example, in the healthcare domain, the target classes could represent the presence or absence of a disease $C = \{\text{disease}, \text{noDisease}\}$, where the distance features's aim is to help refine the model's predictions based on proximity to the centroids of the disease and noDisease classes.

Algorithm 3 DistAugTab

Require: Knowledge Graph $KG = (E, R)$, Tabular Data T, Target Classes C, Embedding Function $\phi : E \cup R \to \mathbb{R}^d$, ML Model M
Ensure: Trained ML model M for each T_{test}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    for each class $C_j \in C$ **do**
        $d_{i,j} \leftarrow \|\vec{v}_i - \vec{c}_{C_j}\|_2$ ▷ Compute Euclidean distance between p_i and class centroid C_j
    end for
    $X_{\text{train}} \leftarrow X_{\text{train}} \cup \text{Concatenate}(p_i, d_{i,1}, \ldots, d_{i,|C|})$ ▷ Append original features and distances
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    for each class $C_j \in C$ **do**
        $d_{i,j} \leftarrow \|\vec{v}_i - \vec{c}_{C_j}\|_2$ ▷ Compute Euclidean distance between p_i and class centroid C_j
    end for
    $X_{\text{test}} \leftarrow X_{\text{test}} \cup \text{Concatenate}(p_i, d_{i,1}, \ldots, d_{i,|C|})$ ▷ Append original features and distances
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.4. Embedding and Distance Features Augmented Tabular Data (EmbedDistTabAug)

This approach augments the tabular dataset by incorporating both embedding vectors and distance-based features, as depicted in Figure 6 and presented in Algorithm 4. For each instance p_i, the augmented dataset T' is expanded by adding $d + |C|$ new columns, where d represents the embedding dimension and $|C|$ denotes the number of target classes. This results in an enhanced dataset with dimensions $m \times (n + d + |C|)$, combining the original features, embedding vectors, and distances to class centroids.

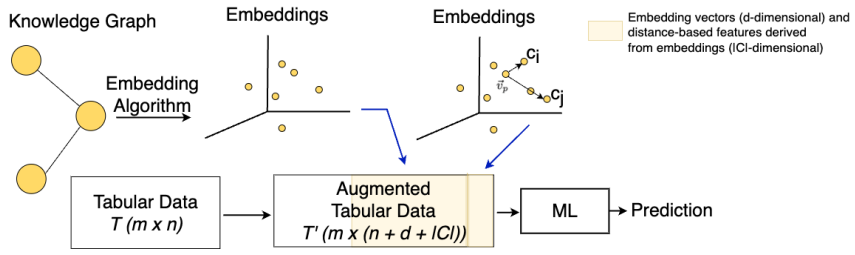

Fig. 6. Tabular dataset enrichment with distance measures from the KG and vector embeddings, highlighted in yellow, used as inputs for the ML model.

Algorithm 4 EmbedDistTabAug

Require: Knowledge Graph $KG = (E, R)$, Tabular Data T, Target Classes C, Embedding Function $\phi : E \cup R \rightarrow \mathbb{R}^d$, ML Model M

Ensure: Trained ML model M for each T_{test}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    for each class $C_j \in C$ **do**
        $d_{i,j} \leftarrow \|\vec{v}_i - \vec{c}_{C_j}\|_2$ ▷ Compute Euclidean distance between p_i and class centroid C_j
    end for
    $X_{\text{train}} \leftarrow X_{\text{train}} \cup \text{Concatenate}(p_i, v_i, d_{i,1}, \ldots, d_{i,|C|})$ ▷ Append original features, embedding, and distances
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    for each class $C_j \in C$ **do**
        $d_{i,j} \leftarrow \|\vec{v}_i - \vec{c}_{C_j}\|_2$ ▷ Compute Euclidean distance between p_i and class centroid C_j
    end for
    $X_{\text{test}} \leftarrow X_{\text{test}} \cup \text{Concatenate}(p_i, v_i, d_{i,1}, \ldots, d_{i,|C|})$ ▷ Append original features, embedding, and distances
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.5. Tabular Dataset Enrichment with Embedding Clusters' membership (ClusterAugTab)

In this approach, referred to as ClusterAugTab, we augment the tabular dataset by first computing embeddings for the data $E_{\text{train}} = \{\phi(p_i) | p_i \in T_{train}\}$, where $\phi : E \cup R \rightarrow \mathbb{R}^d$, and then clustering these embeddings into n clusters using the K-means algorithm, as shown in Figure 7 and presented in Algorithm 5. Each instance $p_i \in T$ is assigned a cluster membership based on its embedding, which is added as an additional feature to the dataset. The augmented dataset T' now has dimensions $m \times (n+1)$, where the original n features are extended by one column representing the cluster membership derived from the embeddings. This enhanced dataset is then used to train the ML model, with the added cluster-level information facilitating the grouping of similar instances. By capturing these underlying patterns in the embeddings, the model can achieve improved predictive performance.

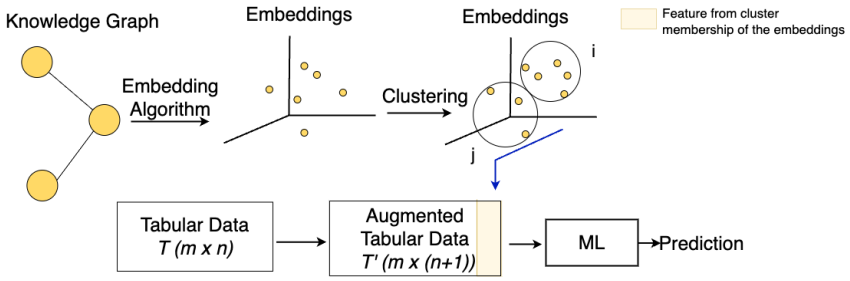

Fig. 7. Tabular dataset enrichment with embedding clusters' membership (highlighted in yellow), used as ML model inputs.

Algorithm 5 ClusterAugTab

Require: Tabular Data T, Number of Clusters n, K-means Clustering Algorithm, ML Model M
Ensure: Trained ML model M for each T_{test}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
Compute embeddings for T_{train}: $E_{\text{train}} = \{\phi(p_i) | p_i \in T_{train}\}$
Initialize K-means with n clusters
Fit K-means on E_{train} to obtain cluster memberships
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $c_i \leftarrow$ K-means cluster for v_i ▷ Assign cluster membership based on embedding v_i
    $X_{\text{train}} \leftarrow X_{\text{train}} \cup \text{Concatenate}(p_i, c_i)$ ▷ Append original features and cluster membership
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $c_i \leftarrow$ K-means cluster for v_i ▷ Assign cluster membership based on embedding v_i
    $X_{\text{test}} \leftarrow X_{\text{test}} \cup \text{Concatenate}(p_i, c_i)$ ▷ Append original features and cluster membership
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.6. Tabular Dataset Enrichment with Embeddings and Embedding Clusters' membership (EmbedClusterAugTab)

This approach, the tabular dataset is augmented by integrating both embedding vectors and cluster memberships, as shown in Figure 8 and detailed in Algorithm 6. For each instance p_i, the augmented dataset T' is expanded by appending both the d-dimensional embedding vector and the corresponding cluster membership, where d represents the embedding dimension. The resulting dataset has dimensions $m \times (n + d + 1)$, combining the original features, the learned embeddings, and the cluster assignments derived from the embeddings. This enriched representation enables the model to leverage both latent structure and group similarity for improved predictive performance.

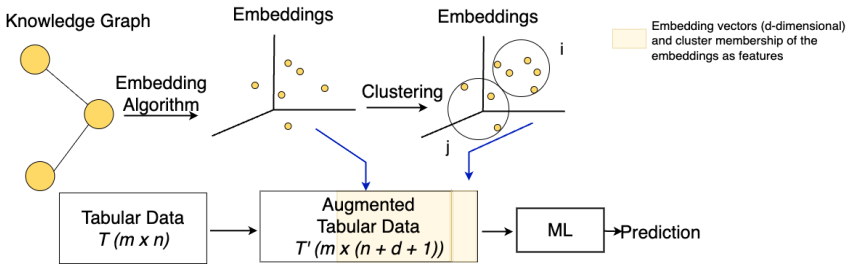

Fig. 8. Tabular dataset enrichment with embedding clusters' membership and vector embeddings, highlighted in yellow, used as ML model inputs.

Algorithm 6 EmbedClusterAugTab

Require: Tabular Data T, Number of Clusters n, K-means Clustering Algorithm, ML Model M
Ensure: Trained ML model M for each T_{test}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
Compute embeddings for T_{train}: $E_{\text{train}} = \{\phi(p_i) | p_i \in T_{train}\}$
Initialize K-means with n clusters
Fit K-means on E_{train} to obtain cluster memberships
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $c_i \leftarrow$ K-means cluster for v_i ▷ Assign cluster membership based on embedding v_i
    $X_{\text{train}} \leftarrow X_{\text{train}} \cup \text{Concatenate}(p_i, v_i, c_i)$ ▷ Append original features, embeddings and cluster membership
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $c_i \leftarrow$ K-means cluster for v_i ▷ Assign cluster membership based on embedding v_i
    $X_{\text{test}} \leftarrow X_{\text{test}} \cup \text{Concatenate}(p_i, v_i, c_i)$ ▷ Append original features, embeddings and cluster membership
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.7. Tabular Dataset Enrichment with Feature Interaction (InteraAugTab)

To further optimize the integration of KG information, we hypothesized that interactions between embeddings and existing features could reveal complex patterns. We define the sub-hypothesis as follows.

> ***H1.4**: Some classes may only be distinguishable through the combined effects of KG embeddings and tabular data.*

By developing approaches that compute these interaction terms, we aimed to enrich the feature space, enabling the model to capture dependencies arising from the integration of KG-derived and tabular features. This approach, implemented in the InteraAugTab approach, offers a multi-dimensional perspective that aims to improve accuracy and F2 score.

InteraAugTab approach augments the tabular dataset by incorporating interaction terms derived from the original features, as illustrated in Figure 9 and presented in Algorithm 7. For each instance p_i, the embedding vector v_i is computed using an embedding function ϕ. Interaction terms are then generated by element-wise multiplying each feature in p_i with each component of the embedding vector v_i. The augmented dataset T' thus contains the original

features and the interaction terms. This results in an enhanced dataset with dimensions $m \times (n + (n \times d))$, where n is the number of original features and d is the embedding dimension. The interaction terms enable the model to capture complex relationships between the original features and the latent information in the embeddings, potentially leading to improved predictive performance.

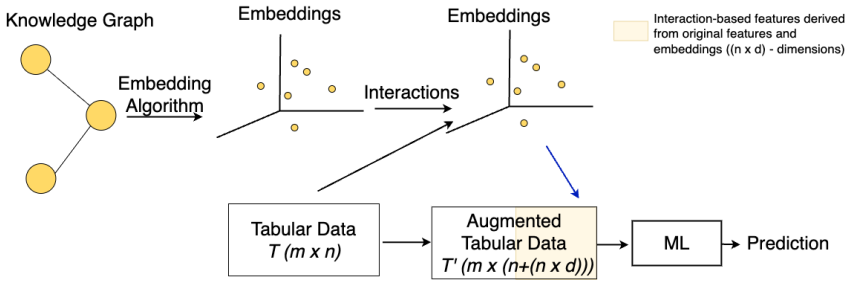

Fig. 9. Tabular dataset enrichment with feature interaction (highlighted in yellow), used as ML model inputs.

Algorithm 7 Feature Interaction Augmented Tabular Data (InteraAugTab)

Require: Tabular Data T, ML Model M
Ensure: Trained ML model M for each T_{test}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $X_{int} \leftarrow$ Compute interaction terms between original features and embedding v_i
    $X_{\text{train}} \leftarrow X_{\text{train}} \cup \text{Concatenate}(p_i, X_{int})$ ▷ Append original features, and interaction terms
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $X_{int} \leftarrow$ Compute interaction terms between original features and embedding v_i
    $X_{\text{test}} \leftarrow X_{\text{test}} \cup \text{Concatenate}(p_i, X_{int})$ ▷ Append original features and interaction terms
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

5.8. Tabular Dataset Enrichment with Embedding and Feature Interaction (EmbedInteraAugTab)

In this approach, referred to as EmbedInteractionAugTab, we augment the tabular dataset by incorporating both the embedding vectors and the interaction terms between the original features and the embedding vectors, as shown in Figure 10 and presented in Algorithm 8. Similar to InteraAugTab approach, the embeddings are computed and the interaction terms. The augmented dataset T' thus contains the original features, the embedding vectors, and the interaction terms resulting in dimensions $m \times (n + d + (n \times d))$, where n is the number of original features and d is the embedding dimension.

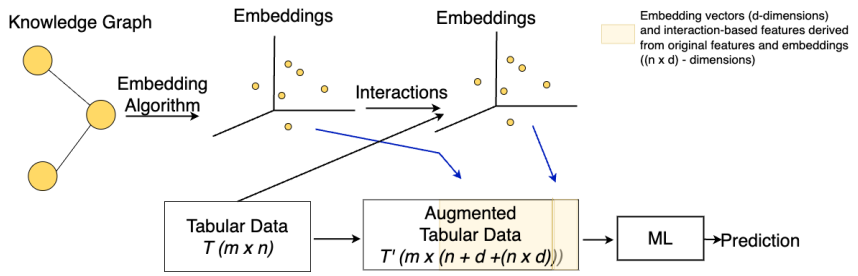

Fig. 10. Tabular dataset enrichment with feature interaction and vector embeddings, highlighted in yellow, used as ML model inputs.

Algorithm 8 EmbedInteractionAugTab

Require: Tabular Data T, Embedding Function $\phi : E \cup R \rightarrow \mathbb{R}^d$, ML Model M
Ensure: Trained ML model M for each T_{test}, T_{test} of n splits

Initialize training data $X_{\text{train}} = []$, labels $Y_{\text{train}} = []$
Initialize test data $X_{\text{test}} = []$, labels $Y_{\text{test}} = []$
for each instance p_i in T_{train} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $X_{int} \leftarrow$ Compute interaction terms between original features and embedding v_i
    $X_{\text{train}} \leftarrow X_{\text{train}} \cup \text{Concatenate}(p_i, v_i, X_{int})$ ▷ Append original features, embedding, and interaction terms
end for
for each instance p_i in T_{test} **do**
    $v_i \leftarrow \phi(p_i)$ ▷ Compute embedding vector for instance p_i
    $X_{int} \leftarrow$ Compute interaction terms between original features and embedding v_i
    $X_{\text{test}} \leftarrow X_{\text{test}} \cup \text{Concatenate}(p_i, v_i, X_{int})$ ▷ Append original features, embedding, and interaction terms
end for
Train ML model M using X_{train} and Y_{train}
Evaluate M on X_{test} and Y_{test}
return Trained model M

To address the risk of high dimensionality, which can adversely affect the performance of certain models, we implemented a dimensionality reduction step using the PCA algorithm [1]. This reduction was specifically applied to approaches integrating embeddings, namely EmbedOnlyRed, EmbedAugTabRed, and EmbedDistAugTabRed. We define our sub-hypothesis as follows:

> ***H1.5****: Reducing the dimensionality of the embedding-augmented datasets will improve model performance by eliminating redundant or noisy features, thereby retaining only the most informative ones.*

6. Experimental Analysis

In this section, we discuss the experimental goals that guide our investigation in Section 6.1 and in section 6.2 we discuss the experimental setup and materials used to achieve these goals.

6.1. Experimental Goals

The goal of our experimental evaluation is to investigate the use of KGs through knowledge graph embeddings to enhance the predictive performance of ML methods. We leverage the semantic structure of the ontologies, to represent the instances with more semantics and then through our proposed approaches use these to augment the tabular dataset for a better ML performance. The specific goals of our experiments are as follows:

Table 2
Details of the ontologies for heart and kidney disease domain.

Domain	Ontologies	Classes	Object prop.	Data prop.
Heart	Small	29	6	10
	Extended	1664	6	10
	Snomed	80	24	10
Kidney	Snomed	113	27	21

Optimal Integration of KGs into ML Pipelines (RQ1): We examine effective methods for incorporating KGs into ML pipelines to improve model performance, with a particular emphasis on accuracy and F2 score. This entails analyzing the integration strategies that can enhance the predictive power of ML models.

Influence of KG Embedding Techniques (RQ2): We seek to understand how different KG embedding algorithms affect performance outcomes in ML models when utilized to enrich tabular data. This exploration focuses on identifying which embedding techniques yield the best enhancements in model accuracy and F2 score.

Comparative Analysis of ML Algorithms with KG-Enhanced Data (RQ3): We assess the relative performance of various ML algorithms when supplemented with KG-derived information. This analysis will highlight how distinct algorithms exploit KG semantics to boost the predictive performance.

6.2. Experiment Setup

Datasets. In our experiments, we used two publicly available datasets from Kaggle: the *Heart Disease*[3] and *Chronic Kidney Disease*[4] datasets. Both datasets are used for binary classification tasks, where the goal is to predict the presence (*disease*) or absence (*no disease*) of the disease.

- Heart disease dataset consists of 303 instances, with 14 features capturing various patient health indicators relevant to diagnosing heart disease such as heart rate and cholesterol.
- Chronic kidney disease contains 400 instances and 25 features, capturing various health metrics related to chronic kidney disease such as blood pressure and albumin levels.

Both datasets contain a mix of categorical and numerical attributes, making them suitable for testing the integration of KGE with tabular data. Additional details about the datasets' features can be found in Appendix A.

Ontologies. For the heart disease, we used three different ontologies:

- The *Small* ontology, denoted as $O = (C, R, H^C)$, is a handcrafted model derived from Trepan Reloaded [10] that encapsulates the features found in the Heart Disease dataset.
- The *Extended* ontology, represented as $O = (C', R', H^C)$, is an extension of an existing ontology [5] $O = (C, R, H^C)$ which incorporates additional features from the dataset.
- The *Snomed* ontology is derived as sub-ontology from the SNOMED-CT ontology [6]. This ontology was constructed using the methodology proposed by Chen et al. [7], which focuses on extracting relevant ontological structures from SNOMED-CT based on a predefined set of seed concepts required in the output. Initially, we selected the relevant concepts in the SNOMED-CT browser [7] that align with the dataset's features. These concepts served as seed concepts in the extraction process, ensuring the resulting ontology included them.

For chronic kidney disease, we only used the third approach, extracting a sub-ontology from SNOMED-CT, due to the lack of ontologies specific to this domain. An overview of the ontologies used for both domains, including the count of classes and properties, is presented in Table 2.

[3] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
[4] https://www.kaggle.com/datasets/mansoordaku/ckdisease
[5] https://bioportal.bioontology.org/ontologies/HFO
[6] https://www.snomed.org
[7] https://termbrowser.nhs.ukmar

Table 3
Parameters for different KGE methods for different KGs.

Domain	KG	Node2Vec Param.				RDF2Vec Param.			TransH & DistMult
		dimens.	walk length	walks	window	depth	walks/node	window	params
Heart	Small	[64,128,100]	40	200	5	4	100	5	default
	Extended	[64,128,100]	60	200	10	6	150	10	default
	Snomed	[64,128,100]	50	200	7	5	100	7	default
Kidney	Snomed	[64,128,100]	50	200	7	10	100	7	default

Table 4
Parameter grid for ML methods.

Method	Parameter	(Grid) Values
KNN	n_neighbors	[20, 25, 30, 35, 40]
SVM	C; kernel; probability	[0.9, 1.0, 1.1, 1.2]; rbf; True
XGB	learning_rate	[0.08, 0.09, 0.1, 0.11]
NN	layers; activation; loss; optimizer	[32, 16, 1]; [relu, relu, sigmoid]; binary crossentropy; adam

KG embedding methods. We used four embedding methods: Node2Vec, RDF2vec, DistMult and TransH. The first two methods were selected as random-walk based models in the embedding landscape, while DistMult and TransH were chosen based on the findings in the Sem@K paper [25], which identified them as outperforming models from the semantic matching and geometric model families, respectively. An overview of these models is provided in Section 2.

In Table 3, we illustrate the parameters used for the embedding methods, tailored to the specific characteristics of the KGs. The embedding dimensions ([64, 128, 100]) were selected to provide a range of vector sizes that are large enough to capture meaningful patterns but small enough to maintain computational efficiency. For Node2Vec and RDF2Vec, the walk length and the number of walks per node were adapted to the size and complexity of each ontology. For smaller ontologies, shorter walks and fewer iterations, while larger or more complex ontologies required slightly longer walks. We averaged performance across three embedding dimensions to provide a more robust evaluation of each method and also computed the standard deviation to capture variability across runs.

ML models. In our experiments, we used four models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and a simple feedforward Neural Network (NN). KNN and SVM were chosen because they are distance-based, aligning with our hypothesis that KGEs, which are also distance-based, would enhance their performance. Whereas, XGB and NN were included to test the effect of KGEs on more complex, non-distance-based models.

To ensure robust evaluation, we used stratified 5-fold cross-validation, maintaining the same class distribution in each fold. For reproducibility, a fixed random seed was applied throughout the experiments. We initially experimented with a wide range of hyperparameters and, to reduce computational cost, we narrowed the range to focus on the best-performing configurations, as shown in Table 4. Results were averaged to ensure consistency across different configurations.

Evaluation metrics. In our experiments, we computed both accuracy and F2 score to assess model performance. We selected the F2 score as a key metric due to its relevance in disease prediction tasks, where maximizing true positive cases is critical for effectively identifying patients with the disease.

7. Results

In this section are shown the experiment results based on the experiment setup that we discussed in Section 6.2, starting with heart disease prediction, followed by kidney disease prediction. We show the concluding results for

each research question.

7.1. Heart Disease Prediction

Table 5 shows the average accuracy and F2 scores, along with the standard deviation across different vector sizes of the embeddings, for four different ML models (KNN, NN, SVM, XGB). The results include the models' baseline performance on tabular data alone, compared with their performance when the data is augmented using embeddings generated by four KG embedding algorithms (Node2Vec, RDF2Vec, DistMult, TransH). Additional results, including average recall with standard deviation across vector sizes, evaluated using different knowledge graphs, models, approaches and embedding methods, are provided in Table 11 in Appendix B. In the following, the results are analyzed based on the research questions.

Investigating the impact of various methods for data augmentation through KGE The different methods of augmenting tabular data with KG embeddings yield mixed results across models. Approaches such as EmbedAugTab, DistAugTab and EmbedDistAugTab often provide the most performance improvements, especially for models such as XGBoost and NN. For example, DistAugTab when Node2Vec is being used to generate the embeddings significantly improved the F2 score of XGB from 75.19 (baseline) to 90.85, highlighting the ability of XGB to effectively use the additional distance features from KG embeddings.

Conversely, SVM and KNN tend to struggle with complex augmentation methods, showing lower gains and even losses in some cases, as they are less suited to high-dimensional data and the resulting feature complexity.

In the following we consider the effectiveness of different approaches based on the sub-hypotheses H1.1 to H1.5.

H1.1 Analysis: Our initial hypothesis (H1.1) proposed that using KG-derived embeddings alone (EmbedOnly) could provide meaningful insights by capturing latent relationships within the data. However, the results across all models and embedding algorithms contradict this hypothesis. The EmbedOnly approach consistently underperformed the baseline for each ML model, regardless of the KG embedding method used. For example, with Node2Vec, the F2 score of SVM dropped significantly from 77.18 (baseline) to 48.31, and similar trends were observed for other models and embeddings. Even when dimensionality reduction (EmbedOnlyRed) was applied to the embeddings, the results remained poor. This suggests that the standalone embeddings lack the richness of information provided by the original tabular features, which include more direct indicators of patient characteristics and clinical factors. Additionally, using embeddings alone may introduce complexity without clear connections to the target variable, making it difficult for the models to extract useful patterns.

H1.2 Analysis: Our second hypothesis (H1.2) proposed that combining the KG embeddings with traditional tabular data (EmbedAugTab) would enhance predictive performance by adding relational information from the KG structure. This approach showed mixed results. In some cases, it led to modest improvements, such as with NN using RDF2Vec to generate the embeddings (F2 score improved from 77.44 to 78.64) or KNN with Node2Vec and DistMult. For SVM, there were F2 score gains when using any embedding algorithm and EmbedAugTabRed compared to the baseline, suggesting that the additional KG information could help refine decision boundaries for SVM's kernel-based approach. However, XGBoost often underperformed when embeddings were added (EmbedAugTab), with scores generally below the baseline. This could be due to XGBoost's preference for a simpler feature space where tabular data alone provides more direct information, making the additional, less structured KG-derived features more of a drawback than a help.

H1.3 Analysis: To address potential noise from directly using embeddings, H1.3 suggested that extracting specific structural features, such as distances from class centroids (DistAugTab) or clustering characteristics (ClustAugTab), would yield better results. The performance of DistAugTab, especially using Node2Vec to generate embeddings, supports this hypothesis, showing significant improvements over the baseline across NN, SVM, and XGBoost models. For instance, using Node2Vec to generate embeddings and then using DistAugTab approach for data augmentation boosted the F2 score of NN from 77.44% to 78.78% and XGBoost from 75.19% to 90.85%, indicating that distance-based features may help capture nuanced relationships between instances and classes that are relevant for classification. NN and SVM also performend well using DistMult and TransH for embedding generation and DistAugTab approach, likely because these models can benefit from the distance measures, making it easier to distinguish between similar instances.

Table 5

Average accuracy and F2 scores (with standard deviation across different vector sizes), across different knowledge graphs, for various models, approaches, and embedding methods in heart disease prediction.

	KNN		NN		SVM		XGBoost	
Methods	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2
Baseline	81.02	71.33	81.77	77.44	79.75	77.18	79.32	75.19
Node2Vec								
EmbedOnly	49.36 ± 2.55	47.10 ± 3.97	49.03 ± 3.26	50.71 ± 3.27	48.31 ± 4.33	48.79 ± 2.42	48.61 ± 1.52	47.50 ± 3.72
EmbedOnlyRed	49.36 ± 2.55	47.10 ± 3.97	49.00 ± 3.59	50.48 ± 2.29	48.19 ± 4.09	48.52 ± 2.21	48.93 ± 2.38	50.90 ± 4.82
EmbedAugTab	81.27 ± 0.21	71.43 ± 0.48	78.15 ± 1.27	76.08 ± 1.58	80.59 ± 0.27	78.07 ± 0.54	64.94 ± 1.29	64.56 ± 4.35
EmbedAugTabRed	80.60 ± 0.13	70.63 ± 0.27	81.02 ± 0.83	76.93 ± 1.13	79.34 ± 0.11	77.53 ± 0.18	79.72 ± 0.49	75.24 ± 0.81
DistAugTab	81.17 ± 0.12	71.54 ± 0.18	**82.17** ± 0.50	**78.78** ± 1.09	**81.81** ± 0.09	**78.36** ± 0.02	**92.51** ± 2.14	**90.85** ± 3.96
EmbedDistAugTab	**81.43** ± 0.28	**71.70** ± 0.55	77.71 ± 1.53	76.10 ± 1.83	**81.67** ± 0.32	**78.57** ± 0.25	**91.82** ± 3.11	**89.27** ± 5.73
EmbedDistAugTabRed	80.66 ± 0.20	70.76 ± 0.33	80.06 ± 0.55	75.74 ± 1.33	79.46 ± 0.20	77.71 ± 0.30	79.32 ± 0.35	75.15 ± 0.59
EmbedClustAugTab	81.17 ± 0.72	70.97 ± 1.49	72.43 ± 0.89	72.96 ± 3.61	76.10 ± 0.56	75.01 ± 2.00	55.32 ± 2.25	57.39 ± 4.94
EmbedInteraAugTab	79.07 ± 0.82	65.56 ± 1.82	75.95 ± 1.28	75.70 ± 1.90	80.12 ± 0.28	76.95 ± 1.15	69.22 ± 1.89	68.40 ± 3.69
ClustAugTab	81.21 ± 0.64	71.16 ± 1.43	78.01 ± 0.10	76.03 ± 1.14	77.58 ± 0.41	76.02 ± 0.69	62.93 ± 0.73	65.05 ± 3.99
InteraAugTab	79.06 ± 0.68	65.55 ± 1.49	78.81 ± 1.34	76.77 ± 0.41	80.11 ± 0.70	77.00 ± 1.10	74.18 ± 1.71	72.90 ± 2.57
RDF2Vec								
EmbedOnly	52.04 ± 0.69	31.42 ± 4.68	53.04 ± 1.28	22.06 ± 9.23	51.27 ± 0.92	40.14 ± 3.48	50.65 ± 1.72	43.41 ± 2.18
EmbedOnlyRed	52.04 ± 0.69	31.42 ± 4.68	53.34 ± 1.10	21.84 ± 9.30	51.27 ± 0.92	40.16 ± 3.44	51.00 ± 1.21	43.18 ± 1.71
EmbedAugTab	81.02 ± 0.00	71.33 ± 0.00	**82.07** ± 0.29	**78.64** ± 0.42	79.75 ± 0.00	77.18 ± 0.00	78.56 ± 0.54	75.30 ± 1.34
EmbedAugTabRed	79.95 ± 0.00	69.59 ± 0.00	80.32 ± 0.58	76.06 ± 0.90	79.32 ± 0.00	77.63 ± 0.00	78.77 ± 0.19	75.25 ± 0.16
DistAugTab	81.02 ± 0.00	71.33 ± 0.00	81.96 ± 0.58	78.57 ± 0.98	79.75 ± 0.00	77.18 ± 0.00	**84.38** ± 1.59	**81.62** ± 2.24
EmbedDistAugTab	81.02 ± 0.00	71.33 ± 0.00	81.85 ± 0.10	78.46 ± 0.58	79.75 ± 0.00	77.18 ± 0.00	80.60 ± 0.82	77.20 ± 0.82
EmbedDistAugTabRed	79.95 ± 0.00	69.59 ± 0.00	80.49 ± 0.77	76.45 ± 0.90	79.32 ± 0.00	77.63 ± 0.00	78.77 ± 0.19	75.25 ± 0.16
EmbedClustAugTab	**81.18** ± 0.11	71.12 ± 0.17	81.44 ± 0.63	77.48 ± 0.67	**80.16** ± 0.08	**77.36** ± 0.28	78.64 ± 0.40	75.34 ± 0.80
EmbedInteraAugTab	81.02 ± 0.00	71.33 ± 0.00	81.81 ± 0.28	78.43 ± 0.79	79.76 ± 0.02	77.18 ± 0.01	79.33 ± 1.04	76.03 ± 0.75
ClustAugTab	**81.18** ± 0.11	71.12 ± 0.17	81.81 ± 0.70	78.38 ± 1.11	**80.16** ± 0.08	**77.36** ± 0.28	79.10 ± 0.35	75.22 ± 0.42
InteraAugTab	81.02 ± 0.00	71.33 ± 0.00	**82.25** ± 0.39	**78.59** ± 0.51	79.75 ± 0.00	77.18 ± 0.00	79.23 ± 0.45	76.15 ± 0.72
DistMult								
EmbedOnly	48.04 ± 0.68	62.88 ± 7.55	46.43 ± 2.21	64.43 ± 5.32	47.14 ± 1.42	68.57 ± 2.13	47.78 ± 2.26	54.97 ± 4.72
EmbedOnlyRed	48.04 ± 0.68	62.88 ± 7.55	49.35 ± 3.91	69.01 ± 3.24	47.46 ± 0.92	64.98 ± 4.09	47.35 ± 1.69	59.07 ± 4.39
EmbedAugTab	81.07 ± 0.10	**71.40** ± 0.19	80.35 ± 0.99	78.30 ± 0.71	80.03 ± 0.31	**77.68** ± 0.40	49.60 ± 3.09	55.53 ± 1.33
EmbedAugTabRed	80.16 ± 0.15	70.03 ± 0.12	80.79 ± 0.36	76.45 ± 0.49	79.33 ± 0.08	77.71 ± 0.11	78.27 ± 0.23	74.25 ± 0.15
DistAugTab	80.88 ± 0.08	71.04 ± 0.14	**82.18** ± 0.77	**78.90** ± 0.81	**80.27** ± 0.02	77.39 ± 0.09	53.42 ± 4.61	54.84 ± 2.15
EmbedDistAugTab	80.94 ± 0.15	71.14 ± 0.21	80.57 ± 1.28	78.55 ± 0.73	80.11 ± 0.25	77.56 ± 0.29	50.49 ± 1.92	61.31 ± 1.66
EmbedDistAugTabRed	80.16 ± 0.15	70.02 ± 0.19	81.30 ± 0.11	77.69 ± 0.77	79.34 ± 0.08	77.71 ± 0.11	78.16 ± 0.25	73.92 ± 0.24
EmbedClustAugTab	**81.39** ± 0.23	71.32 ± 0.11	72.17 ± 2.71	72.09 ± 2.44	75.31 ± 1.20	73.72 ± 1.25	50.12 ± 3.03	55.90 ± 1.53
EmbedInteraAugTab	80.80 ± 0.21	69.92 ± 0.65	70.67 ± 1.42	70.85 ± 0.59	80.20 ± 0.27	77.46 ± 0.21	47.54 ± 1.95	52.40 ± 5.09
ClustAugTab	**81.43** ± 0.16	**71.45** ± 0.06	76.78 ± 0.69	75.76 ± 1.84	76.17 ± 0.76	74.26 ± 1.77	59.35 ± 2.53	62.11 ± 8.32
InteraAugTab	80.84 ± 0.13	69.93 ± 0.66	76.42 ± 1.61	74.98 ± 2.28	80.18 ± 0.35	77.49 ± 0.44	47.29 ± 0.74	48.89 ± 5.33
TransH								
EmbedOnly	48.22 ± 1.36	53.12 ± 5.31	47.88 ± 2.07	59.80 ± 8.61	48.55 ± 1.95	59.32 ± 2.70	47.57 ± 0.21	50.25 ± 11.95
EmbedOnlyRed	48.22 ± 1.36	53.12 ± 5.31	48.17 ± 1.24	57.54 ± 14.81	48.65 ± 0.36	53.07 ± 5.43	51.85 ± 2.31	54.12 ± 8.21
EmbedAugTab	**81.00** ± 0.06	71.24 ± 0.14	80.68 ± 0.31	77.51 ± 1.17	79.89 ± 0.22	77.32 ± 0.39	48.59 ± 0.85	50.51 ± 14.01
EmbedAugTabRed	80.00 ± 0.05	69.74 ± 0.13	79.95 ± 0.66	75.62 ± 0.92	79.33 ± 0.08	77.69 ± 0.03	78.16 ± 0.38	74.17 ± 0.20
DistAugTab	80.98 ± 0.02	71.27 ± 0.03	**81.99** ± 0.28	**78.55** ± 0.85	**80.10** ± 0.15	77.30 ± 0.05	75.34 ± 6.56	73.77 ± 4.42
EmbedDistAugTab	80.95 ± 0.06	71.19 ± 0.10	80.89 ± 1.80	77.97 ± 0.67	**80.11** ± 0.04	77.50 ± 0.33	55.48 ± 4.80	57.76 ± 4.51
EmbedDistAugTabRed	80.08 ± 0.00	69.95 ± 0.06	80.72 ± 0.44	76.61 ± 0.28	79.38 ± 0.09	**77.78** ± 0.05	78.20 ± 0.27	74.17 ± 0.30
EmbedClustAugTab	80.76 ± 0.43	70.42 ± 1.30	76.68 ± 2.10	74.95 ± 2.07	78.32 ± 0.91	74.42 ± 2.54	48.52 ± 1.59	50.38 ± 13.27
EmbedInteraAugTab	80.39 ± 0.22	69.04 ± 0.57	74.31 ± 4.18	72.08 ± 2.19	80.07 ± 0.39	77.14 ± 0.16	49.50 ± 1.57	48.30 ± 6.44
ClustAugTab	80.82 ± 0.38	70.55 ± 1.22	80.24 ± 0.82	75.94 ± 2.34	78.52 ± 0.93	74.52 ± 2.34	60.35 ± 1.96	56.58 ± 7.13
InteraAugTab	80.43 ± 0.25	69.10 ± 0.53	78.55 ± 2.05	75.16 ± 0.36	79.94 ± 0.35	76.94 ± 0.40	49.40 ± 0.77	41.28 ± 6.69

The ClustAugTab approach, on the other hand also showed improvements, particularly with KNN. For example, using RDF2Vec with ClustAugTab led to better clustering of instances, resulting in improved accuracy for KNN and SVM from 81.02% to 81.18% and from 79.75% to 80.16%, respectively. KNN benefited from this approach as it relies on distance metrics to identify neighbors, and having meaningful clusters aligned with its decision-making process. Similarly, SVM showed better results using RDF2Vec for embedding generations and ClustAugTab approach for data augmentation, possibly because the cluster memberships served as a valuable feature that helped define clearer support vectors for class separation.

H1.4 Analysis: Sub-hypothesis (H1.4) suggests that certain classes in the data may be more effectively distinguished when interactions between KG embeddings and traditional tabular data are considered. This hypothesis assumes that there are complex dependencies between the relational information captured by the embeddings and the raw features of the tabular data. The results of the InteraAugTab and EmbedInteraAugTab approaches provide some support for this hypothesis. For example, SVM shows minor improvement in accuracy using interaction terms as additional features, from 79.75% (baseline) to 80.11%, 80.18% and 79.94% when Node2Vec, DistMult and TransH algorithms respectively are being used to generate the embeddings. Comparing between different embedding algorithms, from the table it is shown that RDF2Vec generated the most suitable embeddings to be used for IntraAugTab approach.

H1.5 Analysis: Sub-hypothesis (H1.5) suggests that reducing dimensionality would help models by removing irrelevant or noisy features, thus focusing learning on the most informative aspects of the data. The results show mixed outcomes: while dimensionality reduction sometimes improved performance by simplifying the input space, it often failed to match the effectiveness of methods that used the full set of features without PCA.

For instance, while EmbedAugTabRed and EmbedDistAugTabRed helped reduce overfitting for XGBoost when using DistMult to generate the embeddings by eliminating redundant features, it did not outperform the EmbedAugTab and EmbedDistAugTab methods for KNN. This suggests that, while PCA can be useful for some models it might remove valuable information that more sophisticated models can use, highlighting a trade-off between feature simplification and richness.

Investigating the impact of embedding algorithm The choice of KG embedding algorithm has a significant impact on model performance across the various approaches. Each embedding method captures different aspects of the knowledge graph structure, influencing how well the derived embeddings integrate with the original tabular data and the model's ability to leverage this information.

From the results shown in Table 5 and the F2 score differences illustrated in Figure 11[8], it is evident that Node2Vec and RDF2Vec generally lead to more consistent performance improvements compared to DistMult and TransH, particularly when combined with approaches such as EmbedAugTab and DistAugTab. For example, Node2Vec embeddings with the EmbedDistAugTab approach provided the most notable gains across models, including SVM, and XGBoost. This improvement suggests that Node2Vec's random walk-based approach is effective in preserving local neighborhood information and graph structure, which seems to translate well into the feature space used by the models. The relational patterns it captures may align better with the tabular features, providing additional context that aids in classification.

RDF2Vec also showed good performance, particularly with EmbedAugTab and ClustAugTab approaches. Its ability to leverage RDF graph structures and preserve semantic relationships appears to be beneficial, especially for models such as NN and SVM.

In contrast, DistMult and TransH showed more variable results. While these methods performed well in specific scenarios—such as DistAugTab with DistMult or TransH, particularly for SVM and XGBoost—they were less consistent across different approaches. For example, while DistMult's tensor factorization approach allows it to capture specific types of relational patterns, this does not always translate into performance gains when used for approaches such as ClustAugTab, IntraAugTab or EmbedAugTab.

Moreover, the figures show that XGBoost's performance is particularly sensitive to the choice of embedding algorithm. While XGBoost generally excelled for DistAugTab or EmbedDistAugTab approach using Node2Vec,

[8]Note that EmbedOnly and EmbedOnlyRed are omitted from the figure, due to their consistently poor performance, which skewed the scale for the other approaches

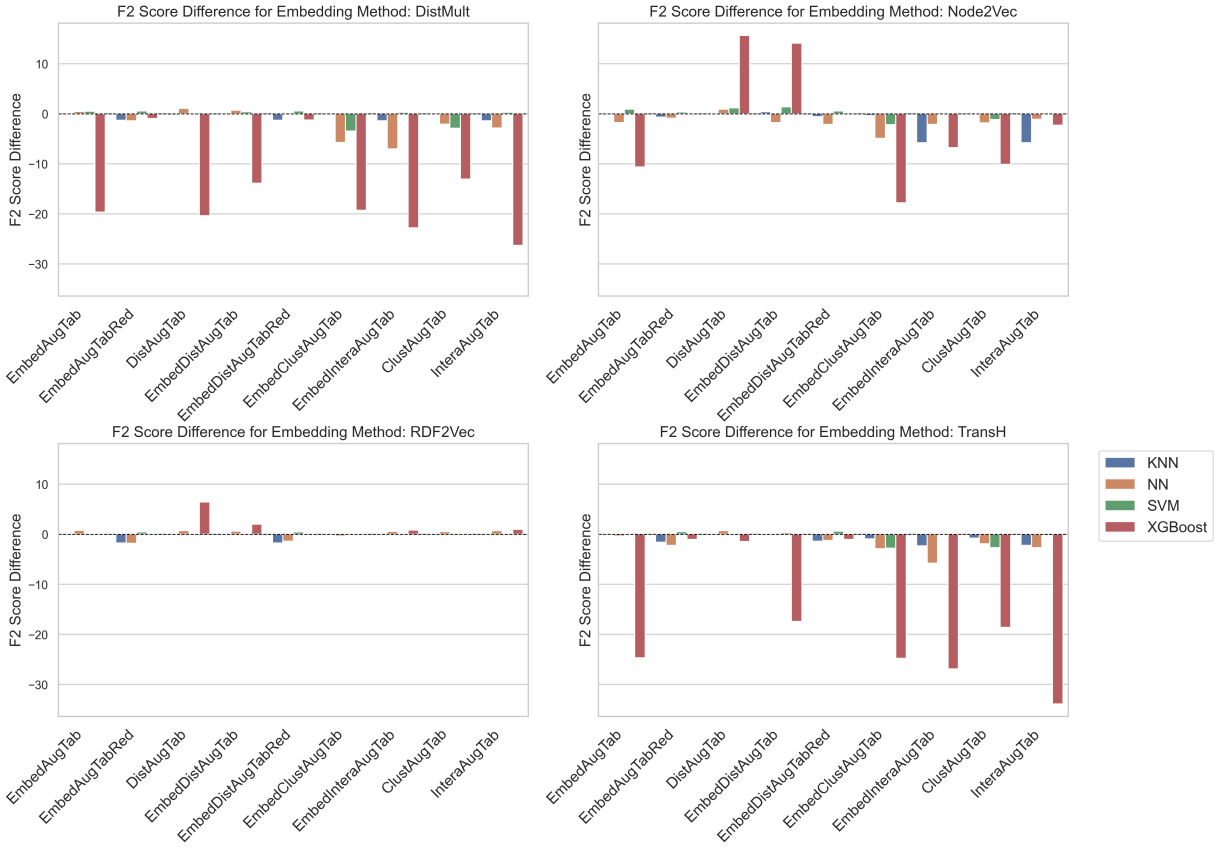

Fig. 11. F2 score differences relative to baseline across models and embedding methods, showing gain/loss for each approach for heart disease prediction

it underperformed with simpler methods such as EmbedAugTab when combined with TransH or DistMult. This suggests that XGBoost requires embeddings that add clear, structured relational information rather than purely dense vector representations. Thus, Node2Vec and RDF2Vec's ability to provide richer, more interpretable representations likely aligns better with XGBoost's learning mechanism.

In conclusion, the choice of the embedding algorithm plays a crucial role in determining the success of different data augmentation approaches. RDF2Vec consistently provides more valuable representations for enhancing model performance across a range of methods, likely due to their strength in capturing both local and global graph structures. DistMult and TransH, while potentially effective in capturing specific relational patterns, exhibit more variability and require carefully chosen augmentation methods to translate their structural information into improved model performance. These findings emphasize that selecting the right embedding algorithm is critical, as it can significantly influence how well the additional relational data is integrated into the learning process.

Investigating the impact of KGs choice Figure 12 shows the average accuracy and F2 score across all evaluated approaches implemented with each ontology. It shows that the choice of ontology (Small, Extended, or Snomed) slightly affects model performance. Using Snomed ontology generally provides the highest accuracy and F2 scores, due to its clinically structured information from medical experts, highlighting its ability to enrich predictions. The Small KG yields the poorest results among the three ontologies, arguably due to its handcrafted nature by non-medical experts, which limits its depth and relevance to complex medical relationships.

Investigating the performance of different ML models across various approaches and embedding algorithms Across the evaluated models, XGBoost and NN showed the most significant improvements when incorporating various KG augmentation methods. Specifically, XGBoost's performance saw the largest gains using the DistAugTab

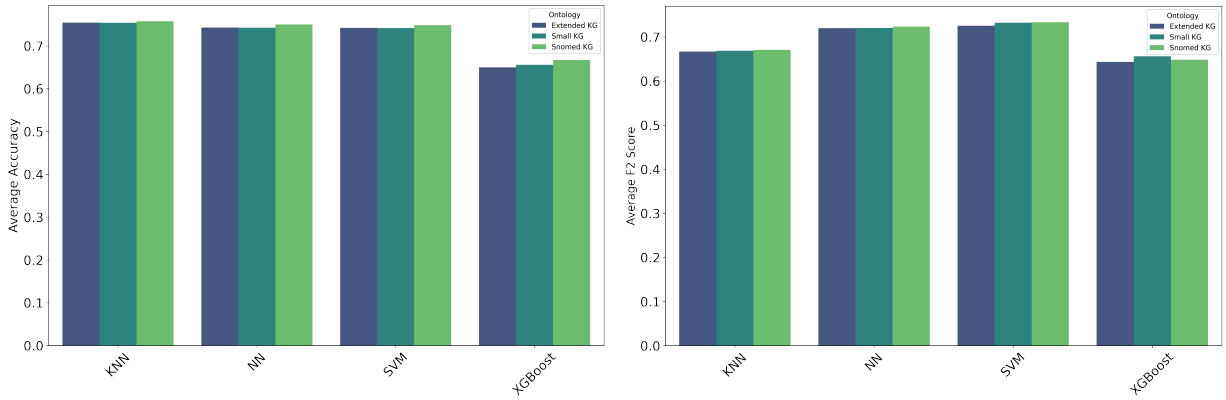

Fig. 12. Comparison of accuracy (left) and F2 (right) scores for different ML models across various KGs.

and EmbedDistAugTab approaches. For example, with Node2Vec embeddings combined using EmbedDistAugTab approach, XGBoost's F2 score increased from a baseline of 75.19% to 89.27%. This can be attributed to XGBoost's ability to effectively handle high-dimensional feature spaces, allowing it to extract valuable patterns from the distance-based features derived from the embeddings. However, XGBoost showed a lot of underperformance when the datasets were augmented with various approaches, especially when embeddings were generated with TransH and DistMult. This reduced performance may be due to the relational complexity in TransH and DistMult embeddings, which introduces interdependent features that XGBoost struggles to interpret independently.

On the other hand, KNN showed only slight performance gains when augmented with embeddings but maintained stable results across different approaches and embedding algorithms.

Looking at the average F2 scores in Table 6 when different embedding algorithms are used to generate the embeddings, we can observe that for four approaches SVM gained slightly better performance compared to the baseline, making it in general more suitable model that gains performance when additional data from KGs is being added, especially when computing the distances to the target classes, or when the vectors are added as such to augment the tabular data.

Table 6
Averages of F2 scores across embedding algorithms for different ML models and approaches in heart disease prediction

Model	KNN	NN	SVM	XGB
Baseline	71.33	77.44	77.18	75.19
EmbedOnly	48.63	48.84	54.21	49.03
EmbedOnlyRed	48.63	49.31	51.68	51.82
EmbedAugTab	**71.35**	77.22	**77.56**	61.47
EmbedAugTabRed	70.00	75.86	**77.64**	74.73
DistAugTab	71.30	**78.29**	**77.56**	**75.27**
EmbedDistAugTab	**71.34**	77.36	**77.70**	71.39
EmbedDistAugTabRed	70.08	76.21	**77.71**	74.62
EmbedClustAugTab	70.96	73.96	75.13	59.75
EmbedInteraAugTab	68.96	73.86	77.18	61.28
ClustAugTab	71.07	76.12	75.54	64.74
InteraAugTab	68.98	75.97	77.15	59.81

7.2. Kidney Disease Prediction

Table 7 shows the average accuracy and F2 scores, along with the standard deviation across different vector sizes of the embeddings, for different models, approaches and embedding methods in kidney disease prediction.

Additional results, including average recall with standard deviation across vector sizes are provided in Table 12 in Appendix B. In the following paragraphs, we will discuss the results based on the research questions, considering also the sub-hypothesis H1.1 - H1.5 from Section 5.

Investigating the impact of various methods for data augmentation through KGE From Table 7, we observe that adding distance-based features to tabular data improves ML model performance, especially for KNN and NN. Although the baseline is already high, enhancements such as distance-to-class, cluster membership features, and embedding vectors still boost performance. For example, KNN accuracy increases from 97% to 99.08% and 99.12% when the data is augmented with vector embeddings (EmbedAugTabRed) and with embeddings plus Euclidean distances to classes (EmbedDistAugTabRed), using TransH to generate embeddings. In the following we will discuss the hypothesis H1.1 - H1.5 based on the results.

H1.1 Analysis: As with heart disease prediction, using only embeddings (EmbedOnly) consistently underperforms compared to the baseline, regardless of the ML model or embedding method, contradicting the hypothesis. This suggests that embeddings alone provide less insight than tabular data for capturing relationships. Performance is particularly poor with RDF2vec embeddings, likely because RDF2vec focuses on structural patterns rather than the detailed, feature-specific information captured in tabular data.

H1.2 Analysis: In line with this hypothesis, the table shows that augmenting tabular data with embedding vectors (EmbedAugTab) generally results in similar or slightly improved accuracy and F2 scores compared to the baseline. For instance, for KNN, adding Node2Vec embeddings increases accuracy from 97% to 97.19% and the F2 score from 98.43% to 98.53%. Likewise, for NN, augmenting with RDF2Vec embeddings raises both accuracy and F2 scores from 99.92% and 99.96% to 100%.

H1.3 Analysis: This hypothesis suggests that adding structural features, such as distances to class centroids (DistAugTab) or clustering membership (ClustAugTab), also adding the embedding vectors (e.g., EmbedDistAugTab), should enhance performance. Structural features help capture relationships in the data by adding context about group distances, which is especially useful for proximity-based models such as KNN, SVM and NN. The results support this, showing performance generally remains consistent with or slightly better than the baseline. For example, for NN using Node2Vec embeddings, the DistAugTab approach increases accuracy and F2 score from 99.92% and 99.96% to 100%, as Node2Vec effectively captures neighborhood structures that align with these models' reliance on similarity. However, with TransH embeddings, which emphasize hierarchical relationships, ClustAugTab and EmbedClustAugTab slightly decrease accuracy and F2 scores for KNN, NN, and SVM, as these embeddings may introduce noise rather than meaningful distance-based information.

H1.4 Analysis: Similarly to the heart disease prediction results, Table 7 shows some results supporting the hypothesis that complex interactions between embeddings and raw tabular features can further improve model performance. SVM generally maintained its 100% performance across different embedding algorithms. RDF2Vec showed to be the most compatible embedding algorithm to be used with our approaches, as it did not lead to performance drops with any approach. KNN on the other hand showed performance improvements when embeddings are used in those two approaches generated from DistMult, with accuracy improvement from 97.00% with only tabular data to 97.29% when interaction terms were included via the InterAugTab approach.

H1.5 Analysis: In the line with this hypothesis, suggesting that tabular dimensionality reduction can help eliminate noisy features, the results for kidney prediction show mixed outcomes. For KNN, applying the PCA algorithm consistently increases both accuracy and F2 scores compared to using the full feature set, particularly for the approaches EmbedDistAugTab and EmbedAugTabRed. This improvement is expected given KNN's struggles with high-dimensional data; it relies heavily on distance calculations, and PCA effectively enhances the quality of these calculations by reducing noise and emphasizing the most informative dimensions. For instance, when using embeddings generated with Node2Vec, dimensionality reduction in EmbedAugTab and EmbedDistAugTab boosts the accuracy from 97.19% and 97.21% to 99.10% in both cases.

Conversely, the performance of NN, SVM, and XGBoost generally worsened after the dimensionality reduction step. This could be attributed to the already high baseline accuracy (ranging from 99.75% to 100%); further reducing the dimensionality might eliminate features that, while not highly significant, still contribute to the model's performance. Additionally, these models can inherently manage high-dimensional data and may not benefit as much from PCA as KNN does. As a result, the reduced feature set may lack the nuanced information that these more complex models require for optimal performance.

Table 7

Average accuracy and F2 scores (with standard deviation across different vector sizes), for various models, approaches, and embedding methods in kidney disease prediction.

	KNN		NN		SVM		XGBoost	
Methods	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2
Baseline	97.00	98.43	99.92	99.96	100.00	100.00	99.75	99.46
Node2Vec								
EmbedOnly	62.12 ± 0.90	20.45 ± 9.74	61.83 ± 1.76	25.97 ± 15.45	62.56 ± 1.33	22.70 ± 12.15	61.60 ± 1.07	26.67 ± 11.16
EmbedOnlyRed	62.12 ± 0.90	20.45 ± 9.74	61.58 ± 1.13	24.62 ± 14.89	62.92 ± 1.38	23.26 ± 12.30	61.42 ± 1.18	25.43 ± 12.01
EmbedAugTab	97.19 ± 0.11	98.53 ± 0.06	99.83 ± 0.29	99.78 ± 0.39	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedAugTabRed	**99.10** ± 0.07	**98.55** ± 0.08	99.75 ± 0.00	99.73 ± 0.23	99.83 ± 0.14	99.64 ± 0.31	99.28 ± 0.38	99.08 ± 0.44
DistAugTab	97.06 ± 0.00	98.46 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTab	97.21 ± 0.10	98.54 ± 0.05	99.92 ± 0.14	99.82 ± 0.31	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTabRed	**99.10** ± 0.07	**98.55** ± 0.08	99.75 ± 0.25	99.73 ± 0.27	99.83 ± 0.14	99.64 ± 0.31	99.47 ± 0.05	99.18 ± 0.26
EmbedClustAugTab	97.38 ± 0.25	**98.56** ± 0.24	99.92 ± 0.14	99.82 ± 0.31	99.97 ± 0.05	99.94 ± 0.10	99.75 ± 0.00	99.46 ± 0.00
EmbedInteraAugTab	96.54 ± 0.22	98.03 ± 0.32	98.83 ± 0.52	98.44 ± 1.21	99.06 ± 0.70	97.96 ± 1.53	99.75 ± 0.00	99.46 ± 0.00
ClustAugTab	97.27 ± 0.16	98.54 ± 0.13	99.75 ± 0.00	99.87 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
InteraAugTab	96.46 ± 0.22	97.99 ± 0.32	99.00 ± 0.25	98.39 ± 0.49	99.14 ± 0.63	98.14 ± 1.37	99.75 ± 0.00	99.46 ± 0.00
RDF2Vec								
EmbedOnly	59.45 ± 0.87	8.59 ± 3.09	62.50 ± 1.10	4.10 ± 1.30	56.64 ± 2.21	13.90 ± 0.58	53.19 ± 0.96	26.59 ± 3.53
EmbedOnlyRed	59.45 ± 0.87	8.59 ± 3.09	56.25 ± 0.00	7.75 ± 0.00	56.58 ± 2.29	13.89 ± 0.58	53.83 ± 1.50	27.52 ± 0.64
EmbedAugTab	97.00 ± 0.00	98.43 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedAugTabRed	**99.06** ± 0.00	**98.70** ± 0.00	99.75 ± 0.00	99.87 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.50 ± 0.00	99.33 ± 0.00
DistAugTab	97.00 ± 0.00	98.43 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTab	97.00 ± 0.00	98.43 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTabRed	**99.06** ± 0.00	**98.70** ± 0.00	99.83 ± 0.29	99.78 ± 0.39	100.00 ± 0.00	100.00 ± 0.00	99.50 ± 0.00	99.33 ± 0.00
EmbedClustAugTab	97.17 ± 0.16	98.52 ± 0.08	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedInteraAugTab	97.00 ± 0.00	98.43 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
ClustAugTab	97.17 ± 0.16	98.52 ± 0.08	99.92 ± 0.14	99.96 ± 0.08	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
InteraAugTab	97.00 ± 0.00	98.43 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
DistMult								
EmbedOnly	66.09 ± 1.99	46.62 ± 9.74	59.17 ± 1.91	44.73 ± 28.08	64.17 ± 0.00	16.95 ± 0.00	55.83 ± 6.17	26.58 ± 35.97
EmbedOnlyRed	66.09 ± 1.99	46.62 ± 9.74	51.88 ± 9.72	74.79 ± 3.58	58.75 ± 0.00	11.63 ± 0.00	63.75 ± 0.00	4.13 ± 0.00
EmbedAugTab	97.02 ± 0.04	98.44 ± 0.02	99.92 ± 0.14	99.82 ± 0.31	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedAugTabRed	**99.08** ± 0.04	**98.68** ± 0.07	99.83 ± 0.14	99.91 ± 0.08	99.75 ± 0.00	99.46 ± 0.00	99.42 ± 0.14	99.29 ± 0.08
DistAugTab	97.00 ± 0.00	98.43 ± 0.00	99.75 ± 0.25	99.87 ± 0.13	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTab	97.04 ± 0.07	98.45 ± 0.04	99.92 ± 0.14	99.96 ± 0.08	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTabRed	**99.08** ± 0.04	**98.71** ± 0.02	99.92 ± 0.14	99.96 ± 0.08	99.89 ± 0.13	99.76 ± 0.27	99.50 ± 0.00	99.33 ± 0.00
EmbedClustAugTab	97.44 ± 0.27	98.59 ± 0.26	99.92 ± 0.14	99.96 ± 0.08	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedInteraAugTab	97.27 ± 0.24	98.27 ± 0.27	98.25 ± 0.75	96.12 ± 1.74	99.72 ± 0.35	99.40 ± 0.75	95.84 ± 3.69	89.86 ± 9.52
ClustAugTab	97.38 ± 0.17	98.59 ± 0.14	**100.00** ± 0.00	**100.00** ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
InteraAugTab	97.29 ± 0.16	98.35 ± 0.19	98.00 ± 0.66	95.53 ± 1.64	99.86 ± 0.24	99.70 ± 0.52	98.33 ± 2.45	96.29 ± 5.49
TransH								
EmbedOnly	41.82 ± 3.92	67.08 ± 6.86	45.81 ± 4.30	58.45 ± 7.49	56.64 ± 13.86	33.89 ± 25.46	44.73 ± 3.76	63.87 ± 7.00
EmbedOnlyRed	41.82 ± 3.92	67.08 ± 6.86	39.79 ± 4.16	62.73 ± 9.94	67.50 ± 0.00	85.23 ± 0.00	45.03 ± 5.18	62.13 ± 12.14
EmbedAugTab	97.00 ± 0.00	98.43 ± 0.00	99.83 ± 0.14	99.91 ± 0.08	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedAugTabRed	**99.08** ± 0.04	**98.71** ± 0.02	99.83 ± 0.29	99.91 ± 0.15	99.92 ± 0.14	99.82 ± 0.31	99.50 ± 0.00	99.33 ± 0.00
DistAugTab	97.06 ± 0.00	98.46 ± 0.00	99.92 ± 0.14	99.96 ± 0.08	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.00	99.46 ± 0.00
EmbedDistAugTab	97.06 ± 0.00	98.46 ± 0.00	99.83 ± 0.14	99.91 ± 0.08	100.00 ± 0.00	100.00 ± 0.00	98.78 ± 1.68	98.86 ± 1.04
EmbedDistAugTabRed	**99.12** ± 0.00	**98.73** ± 0.00	99.83 ± 0.29	99.91 ± 0.15	99.86 ± 0.13	99.70 ± 0.27	99.50 ± 0.00	99.33 ± 0.00
EmbedClustAugTab	96.92 ± 0.18	98.32 ± 0.18	99.50 ± 0.50	99.74 ± 0.26	99.92 ± 0.14	99.96 ± 0.08	99.75 ± 0.00	99.46 ± 0.00
EmbedInteraAugTab	96.94 ± 0.06	98.36 ± 0.03	99.25 ± 0.43	98.64 ± 1.19	99.83 ± 0.29	99.64 ± 0.62	86.85 ± 4.36	66.82 ± 11.84
ClustAugTab	96.90 ± 0.20	98.31 ± 0.18	99.83 ± 0.14	99.91 ± 0.08	99.92 ± 0.14	99.96 ± 0.08	99.75 ± 0.00	99.46 ± 0.00
InteraAugTab	97.00 ± 0.00	98.40 ± 0.06	99.00 ± 0.66	98.11 ± 1.36	100.00 ± 0.00	100.00 ± 0.00	87.93 ± 4.93	69.47 ± 13.95

Investigating the impact of embedding algorithm Figure 13 shows the F2 score differences relative to the baseline across models and embedding methods, highlighting gains and losses for each approach in kidney disease prediction. The approaches EmbedOnly, EmbedOnlyRed, and EmbedInteraAugTab, as well as InteraAugTab, were excluded from analysis due to their skewed performance, particularly when using DistMult and TransH to generate the embeddings, which demonstrated low performance.

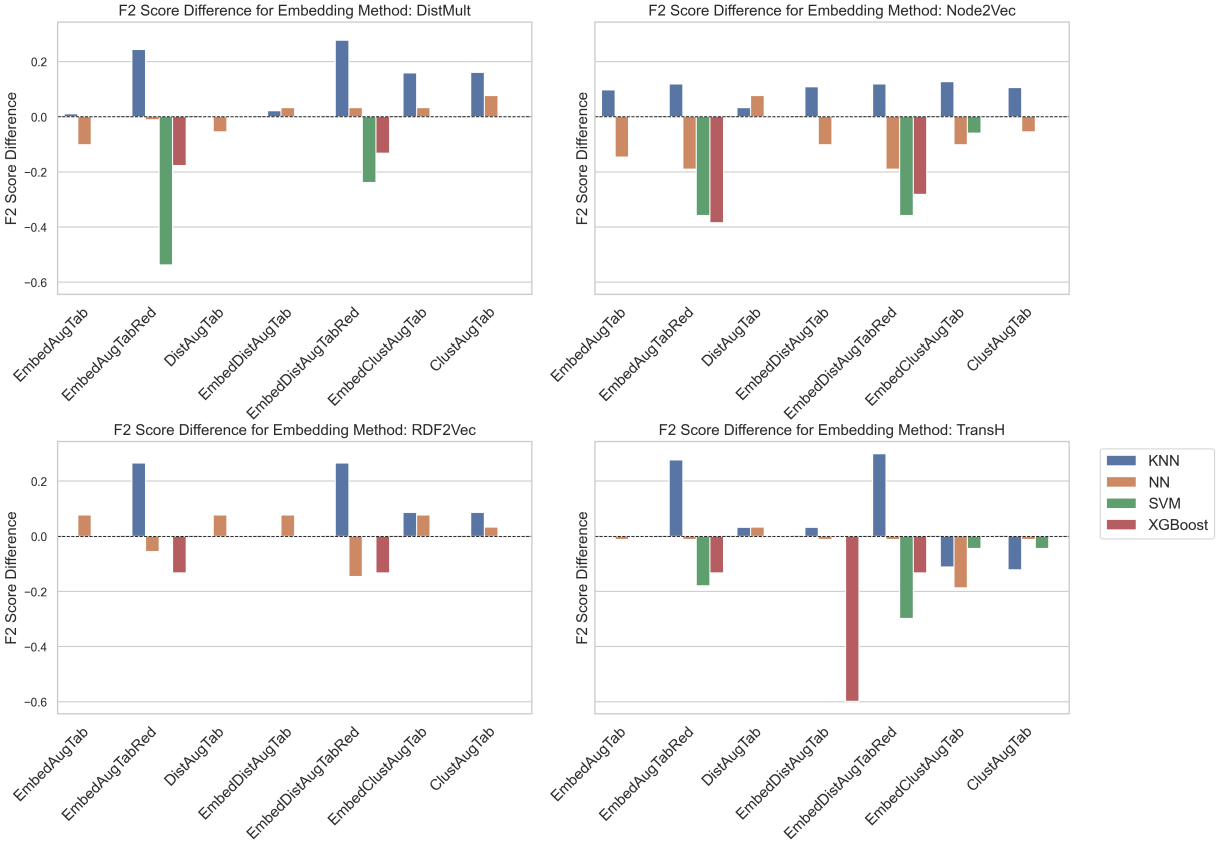

Fig. 13. F2 score differences relative to baseline across models and embedding methods, showing gain/loss for each approach for kidney disease prediction.

From the figure we see that RDF2VEC was shown to be the best suited algorithm among the four embedding algorithms for our approaches. It generates effective embeddings particularly for KNN where the F2 score is increased for some of the approaches and stayed the same for the others. Moreover for the SVM model it maintained a perfect F2 score of 100%, in comparison to other embedding algorithms where the performance dropped. This could indicate that RDF2VEC caputes relevant features that enhance the SVM's ability to establish clear decision boundaries, ultimately resulting in higher predictive performance.

In contrast, Node2Vec generated effective embeddings for the KNN model, where it slightly improved performance. However, its utility decreased for SVM, XGBoost, and NN, often leading to slight performance drops. This suggests that while Node2Vec captures local structural information well, it may introduce noise or irrelevant features for models such as SVM and XGBoost.

Using DistMult to generate the embeddings achieved performance gains with KNN but resulted in decreased performance for SVM. This inconsistency indicates that while DistMult enhances KNN's ability to capture relationships, it introduces noise for SVM's decision-making process.

Similarly, TransH performed best with KNN, especially in the EmbedAugTabRed and EmbedDistAugTabRed approaches, yet showed weaker results for other models, particularly XGBoost. This discrepancy highlights that

TransH may capture specific relational aspects beneficial for KNN but lacks the broader applicability needed for more complex models such as XGBoost.

Overall, our findings suggest that RDF2Vec algorithm is the optimal choice for embedding generation for augmenting data across various models due to its ability to enhance relevant feature representation. Conversely, Node2Vec is particularly advantageous for KNN, emphasizing the need to carefully select embedding algorithms based on the specific model and approach used to ensure the most effective performance enhancement.

Investigating the performance of different ML models across various approaches Table 8 presents the average F2 scores for various ML models in kidney disease prediction, showing the impact of different approaches for combining tabular data with embeddings, averaged across multiple embedding methods. KNN showed the most notable improvements when augmented with KG embeddings across different approaches, likely due to its weaker baseline performance compared to the rest and the suitability of distance-based metrics and dimensionality reduction for this model. Excluding cases where only embeddings were used for training, NN generally maintained its performance with only slight drops in some approaches, suggesting that NN's ability to learn complex patterns is somewhat robust to variations in feature augmentation. SVM, which achieved a perfect F2 score (100%) with only tabular data, retained this performance in EmbedAugTab, DistAugTab, and EmbedDistAugTab. Similarly, XGBoost preserved its performance with the four embedding-augmented configurations, though it experienced slight declines in the remaining cases.

Table 8

Averages of F2 scores across embedding algorithms for different ML models and approaches in kidney disease prediction

Model	KNN	NN	SVM	XGBoost
Baseline	98.43	99.96	100	99.46
EmbedOnly	35.69	33.31	21.86	35.93
EmbedOnlyRed	35.69	42.47	33.50	29.80
EmbedAugTab	**98.46**	99.88	**100.00**	**99.46**
EmbedAugTabRed	**98.66**	99.86	99.73	99.26
DistAugTab	**98.45**	**99.96**	**100.00**	**99.46**
EmbedDistAugTab	**98.47**	99.92	**100.00**	99.31
EmbedDistAugTabRed	**98.67**	99.84	99.78	99.29
EmbedClustAugTab	**98.50**	99.88	99.97	**99.46**
EmbedInteraAugTab	98.27	98.30	99.25	88.90
ClustAugTab	**98.49**	99.93	99.99	**99.46**
InteraAugTab	98.29	98.01	99.46	91.17

8. Conclusions and Future Work

In this paper, we proposed several innovative approaches to augment tabular data with semantic information by leveraging ontologies to capture domain semantics as shown in Figure 14. We utilized these ontologies to construct KGs, thereby enriching the datasets with structured ontological information. To make the knowledge graphs suitable for ML models, we employed knowledge graph embeddings to transform the graphs into a vector space representation. This process enhances the data used to train ML models by integrating domain-specific semantics, allowing the models to leverage contextual and relational information. Based on our experiment setup, we conducted experiments for heart and kidney disease prediction.

For RQ1, our experiments demonstrated that incorporating KG embeddings, particularly by augmenting tabular data with distance-based features to target classes, improves model performance in most of the cases. This enhancement is particularly evident in challenging domains such as chronic kidney disease, where accuracy and F2 scores improved despite limited room for improvement, underscoring the value of KG information for refining ML predictions, especially in data-sparse environments.

For RQ2, our findings indicate that RDF2Vec is the most effective embedding algorithm across models for both heart and kidney disease prediction, given its ability to capture relevant feature representations without performance

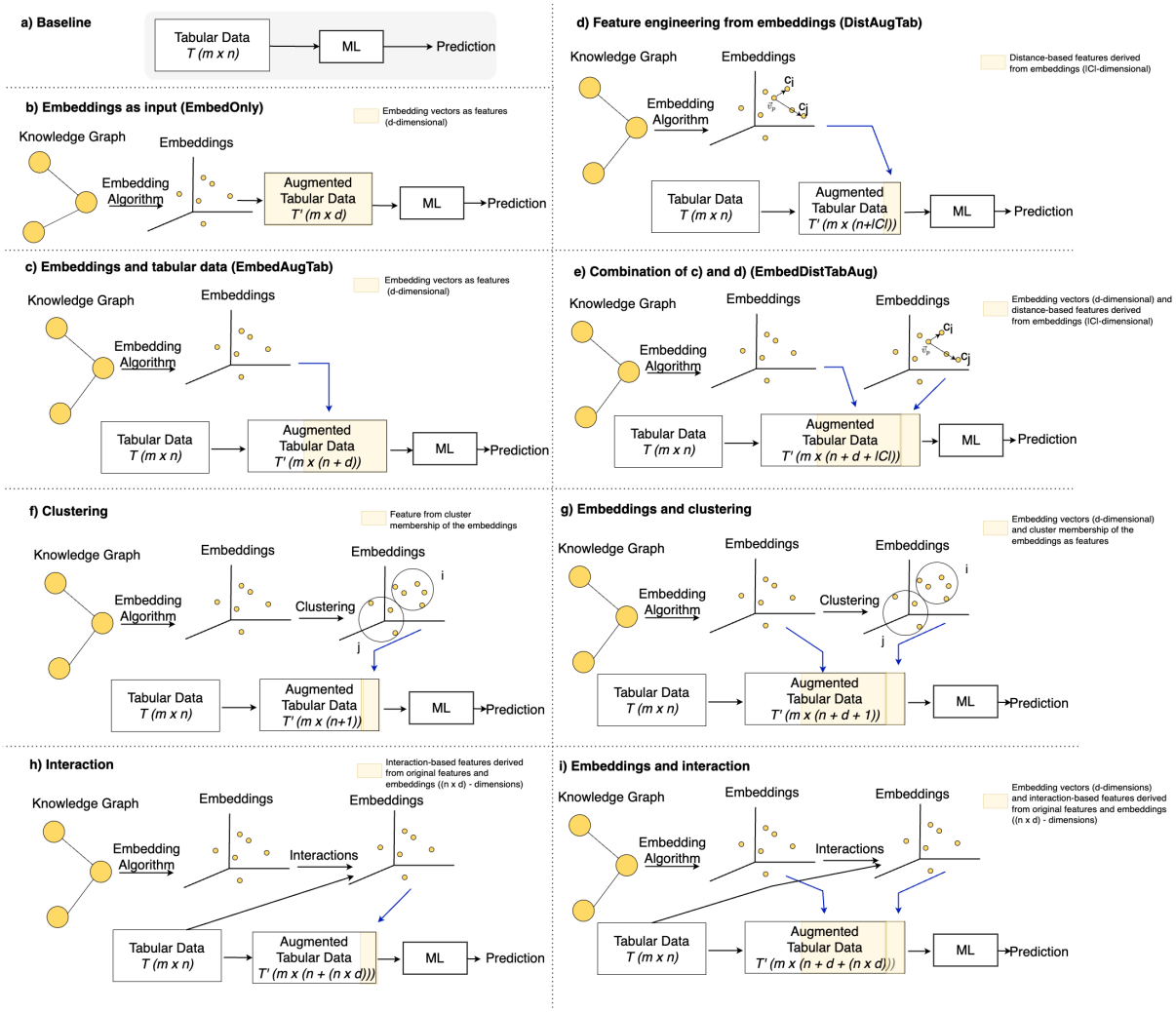

Fig. 14. Different ways of infusing the KG as input into ML pipeline.

drops. Node2Vec proved particularly beneficial for KNN in kidney disease prediction, while in heart disease prediction, Node2Vec enhanced XGBoost the most. However, XGBoost exhibited instability across approaches and embedding algorithms in both cases, suggesting the need for careful pairing of embedding methods and models.

For RQ3, in one hand for heart disease prediction overall on average SVM showed the most F2 score improvement across multiple approaches. Whereas on the other hand for kidney disease prediction, KNN showed the largest performance gains when enhanced with KG embeddings across various approaches, likely due to its weaker baseline performance and the suitability of distance-based metrics and dimensionality reduction, which complement KNN's neighbor-based approach.

Future work will explore the effectiveness of KGs across diverse domains, particularly those with limited data, by augmenting sparse datasets to address the data dependency issues in ML models. Additionally, we plan to assess the scalability of our methods based on data size and structure and experiment with more complex ML models to further optimize the integration of KG embeddings. Furthermore, we aim to explore alternative embedding models and investigate methods for mapping literals into the embedding space to evaluate their impact on model performance.

9. Acknowledgments

This work was supported by the FWF HOnEst project (V 745-N), FFG SENSE project (894802) and FAIR-AI project (904624).

References

[1] H. Abdi and L.J. Williams, Principal component analysis, *Wiley interdisciplinary reviews: computational statistics* **2**(4) (2010), 433–459.
[2] L. Ali, A. Niamat, J.A. Khan, N.A. Golilarz, X. Xingzhong, A. Noor, R. Nour and S.A.C. Bukhari, An optimized stacked support vector machines based expert system for the effective prediction of heart failure, *IEEE Access* **7** (2019), 54007–54014.
[3] K. Annervaz, S.B.R. Chowdhury and A. Dukkipati, Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing, *arXiv preprint arXiv:1802.05930* (2018).
[4] S. Bhatt, A. Sheth, V. Shalin and J. Zhao, Knowledge graph semantic enhancement of input data for improving AI, *IEEE Internet Computing* **24**(2) (2020), 66–72.
[5] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Advances in neural information processing systems* **26** (2013).
[6] A. Breit, L. Waltersdorfer, F.J. Ekaputra, M. Sabou, A. Ekelhart, A. Iana, H. Paulheim, J. Portisch, A. Revenko, A.t. Teije et al., Combining machine learning and semantic web: A systematic mapping study, *ACM Computing Surveys* **55**(14s) (2023), 1–41.
[7] J. Chen, G. Alghamdi, R.A. Schmidt, D. Walther and Y. Gao, Ontology extraction for large ontologies via modularity and forgetting, in: *Proceedings of the 10th International Conference on Knowledge Capture*, 2019, pp. 45–52.
[8] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasiński, Ł. Jasiński, R. Gono, E. Jasińska et al., Prediction of chronic kidney disease-a machine learning perspective, *IEEE access* **9** (2021), 17312–17334.
[9] C.G. Chute and C. Çelik, Overview of ICD-11 architecture and structure, *BMC medical informatics and decision making* **21**(6) (2021), 1–7.
[10] R. Confalonieri, T. Weyde, T.R. Besold and F.M. del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, *Artificial Intelligence* **296** (2021), 103471.
[11] M. Correll, E. Bertini and S. Franconeri, Truncating the y-axis: Threat or menace?, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
[12] T. Dash, S. Chitlangia, A. Ahuja and A. Srinivasan, A review of some techniques for inclusion of domain-knowledge into deep neural networks, *Scientific Reports* **12**(1) (2022), 1040.
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
[14] S. El-Sappagh, F. Franda, F. Ali and K.-S. Kwak, SNOMED CT standard ontology based on the ontology for general medical science, *BMC medical informatics and decision making* **18** (2018), 1–19.
[15] S.L. Franconeri, L.M. Padilla, P. Shah, J.M. Zacks and J. Hullman, The science of visual data communication: What works, *Psychological Science in the public interest* **22**(3) (2021), 110–161.
[16] A.d. Garcez and L.C. Lamb, Neurosymbolic AI: The 3 rd wave, *Artificial Intelligence Review* (2023), 1–20.
[17] M. Gaur, U. Kursuncu, A. Alambo, A. Sheth, R. Daniulaityte, K. Thirunarayan and J. Pathak, " Let me tell you about your mental health!" Contextualized classification of reddit posts to DSM-5 for web-based intervention, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 753–762.
[18] R. Gazzotti, C. Faron-Zucker, F. Gandon, V. Lacroix-Hugues and D. Darmon, Injecting domain knowledge in electronic medical records to improve hospitalization prediction, in: *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, Springer, 2019, pp. 116–130.
[19] A. Grover and J. Leskovec, node2vec: Scalable feature learning for networks, in: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
[20] T.R. Gruber, A translation approach to portable ontology specifications, *Knowledge acquisition* **5**(2) (1993), 199–220.
[21] A.P. Hassler, E. Menasalvas, F.J. García-García, L. Rodríguez-Mañas and A. Holzinger, Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome, *BMC medical informatics and decision making* **19** (2019), 1–17.
[22] D. Herron, E. Jiménez-Ruiz and T. Weyde, On the Benefits of OWL-based Knowledge Graphs for Neural-Symbolic Systems, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning*, Vol. 3432, CEUR Workshop Proceedings, 2023, pp. 327–335.
[23] P. Hitzler, A. Eberhart, M. Ebrahimi, M.K. Sarker and L. Zhou, Neuro-symbolic approaches in artificial intelligence, *National Science Review* **9**(6) (2022), nwac035.
[24] Y.-X. Huang, Z. Sun, G. Li, X. Tian, W.-Z. Dai, W. Hu, Y. Jiang and Z.-H. Zhou, Enabling abductive learning to exploit knowledge graph, in: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 3839–3847.
[25] N. Hubert, P. Monnin, A. Brun and D. Monticolo, Sem@K: Is my knowledge graph embedding model semantic-aware?, *Semantic Web* **14**(6) (2023), 1273–1309. doi:10.3233/SW-233508.
[26] M. Ivanović and Z. Budimac, An overview of ontologies and data resources in medical domains, *Expert Systems with Applications* **41**(11) (2014), 5158–5166.

[27] D. Jarrett, E. Stride, K. Vallis and M.J. Gooding, Applications and limitations of machine learning in radiation oncology, *The British journal of radiology* **92**(1100) (2019), 20190001.
[28] A. Jovic, M. Prcela and D. Gamberger, Ontologies in medical knowledge representation, in: *2007 29th International Conference on Information Technology Interfaces*, IEEE, 2007, pp. 535–540.
[29] R. Katarya and S.K. Meena, Machine learning techniques for heart disease prediction: a comparative study and analysis, *Health and Technology* **11** (2021), 87–97.
[30] H. Kautz, The third AI summer: Aaai robert s. engelmore memorial lecture, *AI Magazine* **43**(1) (2022), 105–125.
[31] J.D.M.-W.C. Kenton and L.K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of naacL-HLT*, Vol. 1, Minneapolis, Minnesota, 2019, p. 2.
[32] C. Kraišniković, R. Harb, M. Plass, W. Al Zoughbi, A. Holzinger and H. Müller, Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making, *Engineering Applications of Artificial Intelligence* **139** (2025), 109561.
[33] U. Kursuncu, M. Gaur and A. Sheth, Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning, *arXiv preprint arXiv:1912.00512* (2019).
[34] Y. Lin, Z. Liu, M. Sun, Y. Liu and X. Zhu, Learning entity and relation embeddings for knowledge graph completion, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29, 2015.
[35] M. Llugiqi, F.J. Ekaputra and M. Sabou, Enhancing Machine Learning Predictions Through Knowledge Graph Embeddings, in: *International Conference on Neural-Symbolic Learning and Reasoning*, Springer, 2024, pp. 279–295.
[36] S. Mohan, C. Thirumalai and G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, *IEEE access* **7** (2019), 81542–81554.
[37] D. Moussallem, M. Arčan, A.-C.N. Ngomo and P. Buitelaar, Augmenting neural machine translation with knowledge graphs, *arXiv preprint arXiv:1902.08816* (2019).
[38] D.M. Pisanelli, *Ontologies in medicine*, Vol. 102, IOS press, 2004.
[39] K. Poulinakis, D. Drikakis, I.W. Kokkinakis and S.M. Spottswood, Machine-learning methods on noisy and sparse data, *Mathematics* **11**(1) (2023), 236.
[40] E.-H.A. Rady and A.S. Anwar, Prediction of kidney disease stages using data mining algorithms, *Informatics in Medicine Unlocked* **15** (2019), 100178.
[41] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, Hierarchical text-conditional image generation with clip latents, *arXiv preprint arXiv:2204.06125* **1**(2) (2022), 3.
[42] P. Rani, R. Kumar, N.M.S. Ahmed and A. Jain, A decision support system for heart disease prediction based upon machine learning, *Journal of Reliable Intelligent Environments* **7**(3) (2021), 263–275.
[43] P. Ristoski and H. Paulheim, Rdf2vec: Rdf graph embeddings for data mining, in: *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15*, Springer, 2016, pp. 498–514.
[44] C. Ruiz, H. Ren, K. Huang and J. Leskovec, High dimensional, tabular deep learning with an auxiliary knowledge graph, *Advances in Neural Information Processing Systems* **36** (2024).
[45] M.K. Sarker, L. Zhou, A. Eberhart and P. Hitzler, Neuro-symbolic artificial intelligence, *AI Communications* **34**(3) (2021), 197–209.
[46] D. Shah, S. Patel and S.K. Bharti, Heart disease prediction using machine learning techniques, *SN Computer Science* **1** (2020), 1–6.
[47] A. Sheth, M. Gaur, U. Kursuncu and R. Wickramarachchi, Shades of knowledge-infused learning for enhancing deep learning, *IEEE Internet Computing* **23**(6) (2019), 54–63.
[48] A. Singhal et al., Introducing the knowledge graph: things, not strings, *Official google blog* **5**(16) (2012), 3.
[49] I. Szilagyi and P. Wira, An intelligent system for smart buildings using machine learning and semantic technologies: A hybrid data-knowledge approach, in: *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*, IEEE, 2018, pp. 20–25.
[50] M. van Bekkum, M. de Boer, F. van Harmelen, A. Meyer-Vitali and A. ten Teije, Modular design patterns for hybrid learning and reasoning systems, *Appl. Intell.* **51**(9) (2021), 6528–6546. doi:10.1007/S10489-021-02394-3. https://doi.org/10.1007/s10489-021-02394-3.
[51] F. Van Harmelen and A. Ten Teije, A boxology of design patterns for hybrid learning and reasoning systems, *Journal of Web Engineering* **18**(1–3) (2019), 97–123.
[52] S. Vijayarani, S. Dhayanand and M. Phil, Kidney disease prediction using SVM and ANN algorithms, *International Journal of Computing and Business Research (IJCBR)* **6**(2) (2015), 1–12.
[53] M. Wang, L. Qiu and X. Wang, A survey on knowledge graph embeddings for link prediction, *Symmetry* **13**(3) (2021), 485.
[54] Z. Wang, J. Zhang, J. Feng and Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 28, 2014.
[55] M. Wortsman, G. Ilharco, S.Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A.S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith et al., Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: *International conference on machine learning*, PMLR, 2022, pp. 23965–23998.
[56] A.L. Yadav, K. Soni and S. Khare, Heart Diseases Prediction using Machine Learning, in: *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2023, pp. 1–7.
[57] B. Yang, W.-t. Yih, X. He, J. Gao and L. Deng, Embedding entities and relations for learning and inference in knowledge bases, *arXiv preprint arXiv:1412.6575* (2014).
[58] P. Yildirim, Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction, in: *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2, 2017, pp. 193–198. doi:10.1109/COMPSAC.2017.84.

[59] C. Yin, R. Zhao, B. Qian, X. Lv and P. Zhang, Domain knowledge guided deep learning with electronic health records, in: *2019 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2019, pp. 738–747.

[60] K. Ziegler, O. Caelen, M. Garchery, M. Granitzer, L. He-Guelton, J. Jurgovsky, P.-E. Portier and S. Zwicklbauer, Injecting semantic background knowledge into neural networks using graph embeddings, in: *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, IEEE, 2017, pp. 200–205.

Appendix A. Additional Experimental Analysis

A.1. Dataset details

Tables 9 and 10 provide an overview of the features in the heart disease and chronic kidney disease datasets, respectively. Each table includes the feature names, their data types, and corresponding descriptions.

Table 9
Heart Disease Dataset Features

Feature Name	**Type**	**Description**
Age	Integer	Age of the patient in years
Sex	Categorical	Gender of the patient (male, female)
CP	Categorical	Chest pain type (typical angina, atypical angina, non-anginal pain, asymptomatic)
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
FBS	Categorical	Fasting blood sugar > 120 mg/dl (true, false)
Restecg	Categorical	Resting electrocardiographic results (normal, STWaveAnormality, LeftVentricularHypertrophy)
Thalach	Continuous	Maximum heart rate achieved
Exang	Categorical	Exercise-induced angina (yes, no)
Oldpeak	Continuous	ST depression induced by exercise relative to rest
Slope	Categorical	Slope of the peak exercise ST segment (0: upsloping, 1: flat, 2: downsloping)
CA	Integer	Number of major vessels (0-3) colored by fluoroscopy
Thal	Categorical	Thalassemia (normal, fixed defect, reversible defect)
Target	Categorical	Presence of heart disease (1 = yes; 0 = no)

Table 10
Chronic Kidney Disease Dataset Features

Feature Name	**Type**	**Description**
Age	Integer	Age of the patient in years
BP	Continuous	Blood pressure in mm/Hg
SG	Continuous	Specific gravity
AL	Integer	Albumin level
SU	Integer	Sugar level
RBC	Categorical	Red blood cells (normal, abnormal)
PC	Categorical	Pus cell (normal, abnormal)
PCC	Categorical	Pus cell clumps (present, not present)
BA	Categorical	Bacteria (present, not present)
BGR	Continuous	Blood glucose random in mg/dl
BU	Continuous	Blood urea in mg/dl
SC	Continuous	Serum creatinine in mg/dl
Sod	Continuous	Sodium level in mEq/L
Pot	Continuous	Potassium level in mEq/L
Hemo	Continuous	Hemoglobin level in gms
PCV	Integer	Packed cell volume
WC	Integer	White blood cell count
RC Count	Continuous	Red blood cell count in millions/cmm
HTN	Categorical	Hypertension (yes, no)
DM	Categorical	Diabetes mellitus (yes, no)
CAD	Categorical	Coronary artery disease (yes, no)
Appet	Categorical	Appetite (good, poor)
PE	Categorical	Pedal edema (yes, no)
Ane	Categorical	Anemia (yes, no)
Label	Categorical	Classification of the disease (ckd, notckd)

Appendix B. Additional Results

Tables 11 and 12 present the average recall (with standard deviation across different vector sizes) for various models, approaches, and embedding methods in heart disease and kidney disease prediction, respectively.

Table 11

Average recall (with standard deviation across different vector sizes), across different knowledge graphs, for various models, approaches, and embedding methods in heart disease prediction.

Methods	KNN	NN	SVM	XGBoost
Baseline	68.37	76.80	76.96	74.31
Node2Vec				
EmbedOnly	48.85 ± 5.39	53.39 ± 4.89	51.06 ± 3.52	49.21 ± 4.82
EmbedOnlyRed	48.85 ± 5.39	53.22 ± 3.71	50.73 ± 3.46	53.70 ± 6.81
EmbedAugTab	**68.38** ± 0.55	76.26 ± 2.77	**77.86** ± 0.71	66.18 ± 5.45
EmbedAugTabRed	67.69 ± 0.31	76.05 ± 1.20	**77.88** ± 0.20	**74.40** ± 0.89
DistAugTab	**68.58** ± 0.18	**78.07** ± 1.36	**77.67** ± 0.08	**90.62** ± 4.51
EmbedDistAugTab	**68.68** ± 0.61	76.57 ± 3.02	**78.03** ± 0.41	**88.73** ± 6.37
EmbedDistAugTabRed	67.83 ± 0.36	74.85 ± 1.55	**78.07** ± 0.32	**74.47** ± 0.71
EmbedClustAugTab	67.85 ± 1.72	74.53 ± 5.17	75.79 ± 2.64	60.05 ± 6.78
EmbedInteraAugTab	61.74 ± 1.95	76.74 ± 2.90	76.51 ± 1.60	69.50 ± 4.41
ClustAugTab	68.05 ± 1.67	76.27 ± 1.69	76.51 ± 1.01	67.49 ± 5.57
InteraAugTab	61.74 ± 1.60	**76.90** ± 0.97	76.56 ± 1.25	73.59 ± 3.10
RDF2Vec				
EmbedOnly	29.59 ± 4.88	20.11 ± 9.02	39.13 ± 3.76	43.12 ± 2.16
EmbedOnlyRed	29.59 ± 4.88	19.90 ± 9.07	39.15 ± 3.71	42.84 ± 1.94
EmbedAugTab	**68.37** ± 0.00	**77.92** ± 0.48	76.96 ± 0.00	**74.77** ± 1.70
EmbedAugTabRed	66.60 ± 0.00	75.17 ± 1.00	**78.02** ± 0.00	**74.79** ± 0.14
DistAugTab	**68.37** ± 0.00	**77.92** ± 1.11	76.96 ± 0.00	**81.05** ± 2.47
EmbedDistAugTab	**68.37** ± 0.00	**77.77** ± 0.84	76.96 ± 0.00	**76.54** ± 0.99
EmbedDistAugTabRed	66.60 ± 0.00	75.64 ± 0.84	**78.02** ± 0.00	**74.79** ± 0.14
EmbedClustAugTab	68.00 ± 0.18	76.57 ± 0.68	**77.02** ± 0.37	**74.79** ± 1.12
EmbedInteraAugTab	**68.37** ± 0.00	**77.75** ± 0.99	76.96 ± 0.00	**75.46** ± 0.92
ClustAugTab	68.00 ± 0.18	**77.67** ± 1.27	**77.02** ± 0.37	**74.45** ± 0.43
InteraAugTab	**68.37** ± 0.00	**77.75** ± 0.55	76.96 ± 0.00	**75.67** ± 0.96
DistMult				
EmbedOnly	**74.45** ± 9.62	**77.09** ± 8.04	**82.55** ± 2.44	63.16 ± 5.48
EmbedOnlyRed	**74.45** ± 9.62	**81.10** ± 3.72	**78.55** ± 4.97	69.72 ± 5.02
EmbedAugTab	**68.42** ± 0.22	**78.42** ± 0.64	**77.55** ± 0.45	63.31 ± 2.99
EmbedAugTabRed	67.10 ± 0.12	75.49 ± 0.58	**78.12** ± 0.12	73.67 ± 0.14
DistAugTab	68.02 ± 0.15	**78.25** ± 0.97	**77.01** ± 0.12	60.12 ± 1.30
EmbedDistAugTab	68.14 ± 0.23	**78.65** ± 0.41	**77.33** ± 0.33	70.35 ± 4.26
EmbedDistAugTabRed	67.08 ± 0.21	**77.00** ± 1.09	**78.12** ± 0.12	73.27 ± 0.21
EmbedClustAugTab	68.26 ± 0.02	74.60 ± 2.51	75.06 ± 2.40	64.17 ± 2.75
EmbedInteraAugTab	66.63 ± 0.73	73.27 ± 0.83	**77.18** ± 0.23	59.68 ± 7.25
ClustAugTab	**68.40** ± 0.11	**76.91** ± 2.39	75.38 ± 2.87	67.78 ± 9.88
InteraAugTab	66.63 ± 0.78	75.72 ± 2.37	**77.21** ± 0.50	54.14 ± 7.90
TransH				
EmbedOnly	62.07 ± 7.94	69.66 ± 11.46	69.67 ± 2.93	57.37 ± 14.74
EmbedOnlyRed	62.07 ± 7.94	66.98 ± 19.81	63.09 ± 7.23	61.92 ± 11.18
EmbedAugTab	68.24 ± 0.18	**77.13** ± 1.81	**77.09** ± 0.52	56.87 ± 17.56
EmbedAugTabRed	66.78 ± 0.12	74.67 ± 1.13	**78.10** ± 0.00	73.65 ± 0.13
DistAugTab	68.28 ± 0.04	**77.85** ± 1.14	76.96 ± 0.00	**74.60** ± 2.90
EmbedDistAugTab	68.18 ± 0.11	**77.66** ± 0.46	**77.25** ± 0.49	62.49 ± 7.83
EmbedDistAugTabRed	67.02 ± 0.06	75.72 ± 0.37	**78.20** ± 0.05	73.58 ± 0.29
EmbedClustAugTab	67.28 ± 1.61	75.76 ± 4.10	74.32 ± 3.95	56.56 ± 16.46
EmbedInteraAugTab	65.62 ± 0.78	72.87 ± 2.48	76.80 ± 0.07	52.80 ± 8.82
ClustAugTab	67.42 ± 1.51	75.28 ± 3.49	74.32 ± 3.58	60.34 ± 9.30
InteraAugTab	65.68 ± 0.70	74.81 ± 0.74	76.56 ± 0.42	42.69 ± 8.43

Table 12

Average recall (with standard deviation across different vector sizes) for various models, approaches, and embedding methods in kidney disease prediction.

Methods	KNN	NN	SVM	XGBoost
Baseline	100.00	100.00	100.00	99.33
Node2Vec				
EmbedOnly	18.44 ± 9.27	24.22 ± 15.15	20.67 ± 11.57	24.74 ± 11.09
EmbedOnlyRed	18.44 ± 9.27	22.89 ± 14.49	21.19 ± 11.74	23.56 ± 12.00
EmbedAugTab	100.00 ± 0.00	99.78 ± 0.38	100.00 ± 0.00	99.33 ± 0.00
EmbedAugTabRed	98.39 ± 0.19	99.78 ± 0.38	99.56 ± 0.38	99.11 ± 0.38
DistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTab	100.00 ± 0.00	99.78 ± 0.38	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTabRed	98.39 ± 0.19	99.78 ± 0.38	99.56 ± 0.38	99.11 ± 0.38
EmbedClustAugTab	99.89 ± 0.19	99.78 ± 0.38	99.93 ± 0.13	99.33 ± 0.00
EmbedInteraAugTab	99.72 ± 0.35	98.44 ± 1.54	97.48 ± 1.86	99.33 ± 0.00
ClustAugTab	99.94 ± 0.10	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
InteraAugTab	99.72 ± 0.35	98.22 ± 0.77	97.70 ± 1.68	99.33 ± 0.00
RDF2Vec				
EmbedOnly	7.33 ± 2.73	3.33 ± 0.00	12.30 ± 0.46	25.33 ± 3.64
EmbedOnlyRed	7.33 ± 2.73	6.45 ± 0.00	12.30 ± 0.46	26.22 ± 0.67
EmbedAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedAugTabRed	98.67 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
DistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTabRed	98.67 ± 0.00	99.78 ± 0.38	100.00 ± 0.00	99.33 ± 0.00
EmbedClustAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedInteraAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
ClustAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
InteraAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
DistMult				
EmbedOnly	47.92 ± 15.91	48.33 ± 36.55	14.44 ± 0.00	28.89 ± 41.41
EmbedOnlyRed	47.92 ± 15.91	91.67 ± 11.79	9.68 ± 0.00	3.33 ± 0.00
EmbedAugTab	100.00 ± 0.00	99.78 ± 0.38	100.00 ± 0.00	99.33 ± 0.00
EmbedAugTabRed	98.61 ± 0.10	100.00 ± 0.00	99.33 ± 0.00	99.33 ± 0.00
DistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTabRed	98.67 ± 0.00	100.00 ± 0.00	99.70 ± 0.34	99.33 ± 0.00
EmbedClustAugTab	99.89 ± 0.19	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedInteraAugTab	99.50 ± 0.29	95.33 ± 2.00	99.26 ± 0.93	89.28 ± 9.60
ClustAugTab	99.94 ± 0.10	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
InteraAugTab	99.61 ± 0.25	94.67 ± 1.76	99.63 ± 0.64	95.78 ± 6.16
TransH				
EmbedOnly	86.04 ± 12.19	72.94 ± 11.05	37.78 ± 34.01	78.80 ± 11.81
EmbedOnlyRed	86.04 ± 12.19	82.31 ± 13.45	100.00 ± 0.00	77.69 ± 18.37
EmbedAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedAugTabRed	98.67 ± 0.00	100.00 ± 0.00	99.78 ± 0.38	99.33 ± 0.00
DistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedDistAugTab	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.11 ± 0.38
EmbedDistAugTabRed	98.67 ± 0.00	100.00 ± 0.00	99.63 ± 0.34	99.33 ± 0.00
EmbedClustAugTab	99.89 ± 0.19	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
EmbedInteraAugTab	99.94 ± 0.10	98.44 ± 1.54	99.56 ± 0.77	65.44 ± 11.90
ClustAugTab	99.89 ± 0.19	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 0.00
InteraAugTab	99.94 ± 0.10	97.78 ± 1.68	100.00 ± 0.00	68.30 ± 13.58

Appendix. Cover letter

Dear Editors,

Thank you for giving us the opportunity to resubmit our manuscript! We appreciate the insightful comments provided by the reviewers, as it has helped us improve the quality of our paper. In this revised version, we have made the following main changes:

- In the Related Work section (Section 3), we have added a new summary table (Table 1) to visually compare existing NeSy-related studies with ours. We also provide an overview of the key differences between this extended version and our previous NeSy 2024 paper, and have incorporated the paper recommended by the reviewer to enrich the related work section.
- We reported standard deviations over multiple runs in our result tables. Additionally, we provide more results in the Appendix, with a particular focus on recall metrics.
- We added legends and improved the caption for Figures 3-10 to make them more understandable.
- In the Appendix we added an overview of the dataset features.
- We added supplementary material [9], where all approaches are illustrated using the boxology notation from [50]
- We corrected all the typos.

We hope these revisions address all the reviewers' comments, improve the manuscript's clarity and make it suitable for publication. Below, we have included a detailed response letter where we address each of the reviewers' comments individually.

The Author Team

Appendix. Response letter

Dear Reviewers,

We sincerely appreciate the time and effort you have invested in reviewing our manuscript. Your thorough and insightful feedback has been invaluable in refining our work, strengthening its clarity, and improving its overall quality.

In this response letter, we have reproduced the reviewers' comments as received, explicitly numbering each comment raised. We then provide a detailed response to each comment and highlight the corresponding revisions made in the manuscript where applicable.

We hope that the revisions and improvements we have incorporated make our submission suitable for acceptance.

The Author Team

Review 1:

The paper explores the enhancement of machine learning (ML) predictions in data-scarce environments through semantic-based data augmentation leveraging knowledge graphs (KGs). It enriches tabular datasets with various KG-derived embeddings and evaluates their impact on the predictive performance of ML models (e.g., KNN, SVM, XGBoost, Neural Networks) across different embedding techniques and augmentation strategies. The methodology is applied to binary classification tasks for heart disease and chronic kidney disease using public datasets. The findings demonstrate notable improvements, particularly when distance-based KG features are incorporated, with XGBoost and Neural Networks showing the most significant gains.

[9] https://semsys.ai.wu.ac.at/data-augmentation/home.html.

The paper fits well within the scope of the Neuro AI journal. It is well-written and clear. Therefore, I recommend accepting the paper with minor revisions. Below are some suggested improvements.

Presentation:

R1C1: Include a summary of the main changes in this extended version compared to the NeSy 2024 paper.
***Answer**: We have included a summary of the main changes of this paper compared to our NeSy 2024 paper in the last subsection of the Related Work section. This summary highlights the additional approaches we propose, the formalization of the approaches, the expanded methodology with two more KG embedding algorithms, and the more comprehensive evaluation applied to heart and chronic kidney disease prediction.*

R1C2: Add a summary table in the related works section to visualize the relationship between the current work and NeSy-related studies.
***Answer**: We have included Table 1 in the related works section, which provides an overview of NeSy-related studies. This table outlines the types of semantic knowledge used, the domains or tasks covered, the ML models applied, the incorporation of KGEs, and the integration methods employed. This addition helps to visually summarize the relationship between our work and existing NeSy-related studies, making it easier to compare different approaches.*

R1C3: Position tables summarizing the results in the relevant sections of the main text.
***Answer**: As recommended, we have moved the tables summarizing the results to the relevant sections of the main text to improve readability.*

R1C4: Move all algorithms to an appendix for better readability.
***Answer**: We considered moving all algorithms to an appendix but found that it compromised readability due to frequent back-and-forth references.*

R1C5: Utilize Figure 14 in the main text to reference all approaches instead of having separate figures.
***Answer**: Similar to our response to R1C4, we considered referencing all approaches in Figure 14 instead of using separate figures but found that it compromised readability.*

Experiments and Results:

R1C6: Section 6.2: While the paper reports an averaged performance across three embedding dimensions to ensure robustness, it is common practice to also average the results of multiple runs for each experiment and report the standard deviation.
***Answer**: In response to the reviewer's suggestion, we have included the standard deviations to capture variability between runs for Tables 5 and 7 in the main text, as well as for Tables 11 and 12 in the appendix, based on experiments conducted across three embedding dimensions.*

R1C7: Table 2: Provide details on how the hyperparameters for each embedding method were selected.
***Answer**: As suggested by the reviewer, we have added details on the selection of hyperparameters in Table 3 (previously Table 2). Specifically, we explain that the embedding dimensions ([64, 128, 100]) were selected to provide a range of vector sizes that are large enough to capture meaningful patterns but small enough to maintain computational efficiency. For Node2Vec and RDF2Vec, the walk length and the number of walks per node were adapted to the size and complexity of each ontology. For smaller ontologies, shorter walks and fewer iterations, while larger or more complex ontologies required slightly longer walks.*

R1C8: Section 7: Clarify how the impact of KGs was computed. Since the ontologies are used to build the KGs and subsequently implement the described approaches, specify whether the reported value represents an average across all these approaches.

***Answer**: We have clarified that the reported results represent the average accuracy and F2 score across all evaluated approaches implemented with each ontology.*

Minor comments:

R1C9:

- p10, line 44: approachess -> approaches
- p10, line 51: c_j for each target class is computed. How is the centroid computed?
- p11, line 19: and no classes -> and noDisease classes.
- p11 in Alg3, $\vec{v_i}$ is not defined
- p12, line 43: In this approach, referred to as EmbedClusterAugTab -> In this approach, referred to as ClusterAugTab
- p12, line 45: Algorithm 6 -> Algorithm 5
- p17, line 33: we only used on the third approach -> we only used the third approach
- p17, line 39: Detailed descriptions -> I would rather say 'An overview of these models is provided in Section 2.'
- p19, line 34: in the other hand -> on the other hand

***Answer**: We have addressed all minor comments by correcting the reported typos, clarifying how the centroids for the target classes are computed, and defining $\vec{v_i}$ in Algorithm 3.*

Review 2:

[Paper Summary]

The paper explores the integration of Knowledge Graphs into Machine Learning pipelines to address challenges in data-scarce domains. The authors hypothesize that enriching training datasets with semantic information derived from KGs can improve ML models' predictive capabilities. The paper explores three primary research objectives: identifying optimal methods for integrating KG-derived features into ML pipelines; analyzing the impact of various KG embedding techniques on model performance; and comparing the effectiveness of ML algorithms when augmented with KG information. To address the objective the authors tested five sub-hypotheses across eight approaches and conducted experiments on binary classification tasks focused on predicting heart and chronic kidney diseases. The results indicate substantial improvements when models are augmented with distance features derived from KG embeddings.

[Review]

I reviewed this paper for NeSy 2024, and I am pleased to note that it has significantly improved since then. The authors have addressed almost all my previous comments, resulting in a much stronger and more refined submission. The methodology and results are now presented more clearly, and the contributions are well-articulated. Additionally, the paper has been expanded considerably with new content. The paper in its current form is well-prepared and demonstrates substantial progress. However, I still have a few minor comments that could further enhance the quality and clarity of the work:

R2C1: The paper briefly mentions the types of features used in the datasets but does not provide detailed specifications. It would be beneficial to clarify whether the features are categorical, continuous, or integers. Including a table in the appendix that lists each feature along with its type and any relevant characteristics would enhance clarity and reproducibility.

***Answer**: As suggested by the reviewer, we have added Tables 9 and 10 in the Appendix A to provide more details for the features in the used datasets.*

R2C2: It is unclear whether literals are considered part of the entity set E. If literals are not included in E, the graph should be formally represented KG =(E,L,R',Tr) where L denotes the set of literals. Additionally, the paper should explain how literals are mapped into the embedding space, as their representation may influence the model's effectiveness.

***Answer**: We appreciate the reviewer's insightful comment. In our approach, literals are **not** explicitly included in the entity set E, therefore, we have updated the formal representation of the KG to $KG = (E, L, R', Tr)$, where L represents literals, which are currently not included in the embedding space. This is because the embedding models we use do not handle literals unless they are explicitly converted into entities or nodes. We acknowledge that the representation of literals may influence the model's effectiveness, and as part of future work, we plan to explore methods such as transforming literals into entity nodes or using other models that support literals (such as TransEA, LiteralE).*

R2C3: The notation for the embedding function requires more consistency. Initially, it is defined as $\phi : E \cup R \rightarrow \mathbb{R}^d$ mapping entities and relations into a d-dimensional space. Later, the notation shifts to $\phi : P \rightarrow \mathbb{R}^d$, where $P \subseteq E$.

***Answer**: We have ensured consistency in the notation for the embedding function by explicitly using $\phi : P \cup R \rightarrow \mathbb{R}^d$ throughout the manuscript where applicable*

R2C4: The meaning of the yellow sections in the "augmented tabular data" block is not explained in Figures 3–10. Please provide a legend or annotation within the figures.

***Answer**: We have added legends to all figures (Figures 3–10) and Figure 14 to clarify the meaning of the yellow sections in the "augmented tabular data" block. Additionally, we have updated the captions of the figures to explicitly describe the meaning of the yellow section.*

[Minor things]

R2C5: While the paper primarily focuses on accuracy and F2 scores, incorporating recall as an evaluation metric could offer valuable insights.

***Answer**: As suggested by the reviewer, we have incorporated recall as an additional evaluation metric. In Appendix B, we have added tables that present the average recall along with standard deviations for different vector sizes, for different ML models and different embedding methods, comparing baseline tabular performance with KG-augmented data.*

R2C6: In the abstract and introduction, you mention a focus on accuracy and F2 score without explaining why. The explanation is provided later in section 4.2. Include a brief sentence in the introduction explaining this focus or reference where the explanation can be found.

***Answer**: We have addressed this comment by adding a reference in the introduction to where the reason for choosing the selected metrics is explained.*

R2C7: Rescale the y-axis in your figures for clarity. For example, Fig 12 should have a y-axis range of 0.6 to 0.8. This will better illustrate performance differences.

***Answer**: We considered rescaling the y-axis but decided to maintain the current scale for consistency with our previous publications and because altering the y-axis range can sometimes exaggerate differences and potentially mislead readers about the actual variability in performance. To avoid unintended misinterpretations, as highlighted by [11, 15], we have chosen to keep the original y-axis scale.*

R2C8: Typos:

- Page 3 line 42: R capital letter (R subset of CxC)
- H1.2 Analysis [line 14 page 19] missing capital letter

***Answer**: We have fixed the typos.*

Review 3:

This paper presents an impressive and innovative contribution to the field of machine learning, addressing the critical challenge of improving model performance in data-scarce or sensitive scenarios. The authors' hypothesis that semantic enrichment through knowledge graph (KG) integration can enhance predictive power is both compelling and highly relevant. The introduction of novel neuro-symbolic approaches and the systematic exploration of KG embedding techniques highlight the authors' dedication to advancing the state of the art. Their rigorous evaluation across multiple ML algorithms and KG embedding methods showcases the robustness of their approach. The focus on real-world applications, such as heart disease and chronic disease prediction, further emphasizes the practical significance of their work. The results are particularly noteworthy, demonstrating remarkable improvements in F2 scores, such as a dramatic boost in XGBoost performance for heart disease prediction. These findings convincingly illustrate the potential of KG-based augmentation to transform ML performance, especially in binary classification tasks. The clear, data-driven methodology and the emphasis on accuracy and F2 scores provide valuable insights for both researchers and practitioners. This reviewer emphasises that this paper is a significant step forward in the integration of symbolic reasoning with ML techniques, paving the way for more context-aware, robust, and effective predictive models. It is a must-read for anyone interested in enhancing ML performance through innovative data augmentation strategies. Summarizing, the work presented is interesting, relevant, important, well presented and also well written and fits well into this journal. For all these reasons, this reviewer argues for acceptance of this work and provides in the following just one minor suggestion for improvement to further enhance its usefulness to the potential reader:

R3C1: Page 5, last paragraph- In addition to the excellent work of Bhatt et al. (2020), a very new paper should be mentioned here, a related work that is very interesting for the reader: Kraisnikovic, C. 2025. Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making. Engineering Applications of Artificial Intelligence, 139, 109561, doi:10.1016/j.engappai.2024.109561.

***Answer**: Thank you for your suggestion. We have incorporated the referenced paper into the Machine Learning Models in Disease Prediction paragraph of the Related Work section, as it aligns well with our discussion on leveraging embeddings for medical decision-making.*