Leveraging Knowledge Graphs for Enhancing Machine Learning-based Heart Disease Prediction

Majlinda Llugiqi
WU Vienna
majlinda.llugiqi@wu.ac.at

Fajar J. Ekaputra
WU Vienna
fajar.ekaputra@wu.ac.at

Marta Sabou
WU Vienna
marta.sabou@wu.ac.at

Abstract

Implementing machine learning in healthcare, especially for heart disease prediction, is crucial for saving lives through accurate diagnosis. Yet, the effectiveness of these models is constrained by a lack of extensive, annotated datasets, which are crucial for training robust models. Additionally, there is an under-use of existing knowledge graphs (KGs), which contain structured domain knowledge that could improve the performance of models. To address this challenge, our study introduces approaches that integrate KG embeddings with tabular data, with the goal of improving the performance of machine learning algorithms for heart disease prediction. We conduct a comparative analysis of various methodologies to merge tabular data with KGs, focusing on heart disease, and evaluate the performance of two embedding algorithms in augmenting datasets for more accurate machine learning applications. Our methodology, which involves testing embeddings from diverse KGs, has consistently shown improvements in model performance. Specifically, we increased the accuracy of the Feed-Forward Neural Network from 82% to 85% and the F2 score for the K-Nearest Neighbors model from 71% to 80%. This advancement offers a promising direction for leveraging semantic information from KGs for knowledge-enhanced machine learning in healthcare.

1 Introduction

In the rapidly evolving landscape of Artificial Intelligence (AI), the fusion of symbolic reasoning with machine learning (ML) has given rise to a novel paradigm known as Neurosymbolic AI (NeSy) [5, 9, 20]. This convergence has the potential to overcome the limitations of traditional AI by integrating the interpretability and logical rigor of symbolic AI with the adaptive learning capabilities of ML methods. Central to this integration is the use of knowledge graphs (KGs), which serve as a structured representation of domain-specific knowledge, enabling AI systems to process and interpret vast amounts of data more effectively [2, 6, 8, 25].

Despite these advancements, ML models often lead to suboptimal performance, particularly in domains where data can be sparse or of poor quality [18]. This is especially the case in the prediction of medical diagnostics, where the inconsistency and scarcity of data significantly hinder the models' performance [11].

Addressing this challenge, this paper proposes a novel approach that leverages the rich information contained within KGs to supplement the data used in ML algorithms. By integrating background knowledge from these sources, we aim to enhance the accuracy and F2 score of ML models in the medical domain. Our research is guided by the following three research questions:

- RQ1: How can we best infuse KGs as input into a ML pipeline in order to enhance its performance, in terms of accuracy and F2 score?

The First Austrian Symposium on AI, Robotics, and Vision (AIROV24).

- RQ2: In what ways do the size and structural variations of KGs affect performance gains among various ML algorithms?
- RQ3: How do different ML algorithms perform when information from KGs is infused?

In order to answer these research questions, our methodology includes the following steps: (i) enriching existing ontologies with specific instances from a dataset, then (ii) subsequently using two different embedding techniques on those KGs to generate vector embeddings. Next, (iii) we combine these embeddings with traditional tabular data, creating more comprehensive datasets that includes additional features derived from the embedding space. Finally, (iv) we use these enriched datasets to train ML models for the prediction of heart disease.

The main contributions of this research are as follows:

- We explore and compare methodologies for integrating KGs into ML pipelines, with a focus on heart disease. We introduce an innovative feature engineering approach that enriches datasets with KG insights. Our analyses evaluate the efficacy of two embedding algorithms in merging KGs with tabular data, aiming to enhance ML models' robustness. This work addresses RQ1 by demonstrating various strategies for KG infusion in ML, highlighting their potential in medical applications.
- We empirically demonstrate that incorporating KGs, with their diverse structures and sizes, generally enhances ML model performance, notably in terms of accuracy and F2 scores. We show a variation in how different algorithms benefit from KG integration; some algorithms achieve optimal performance with smaller-sized KGs, while others excel when leveraging larger graphs. This indicates that despite the size of the KG, its integration consistently improves upon baseline models. This finding responds to RQ2.
- We compare the performance of various ML algorithms, both with and without the infusion of additional knowledge derived from KGs. Our findings demonstrate the impact of this integration on algorithms' performance, effectively addressing RQ3.

The rest of the paper is structured as follows. Section 2 provides an overview of the related work, this is followed by Sections 3 and 4, where we discuss the methodology used for our approach along with the experiment setup in Section 5. Then in Section 6 we show and analyze the results of our experiments, and we conclude our findings and discuss the future work in Section 7.

2 Related Work

We start by discussing the use of ML models for heart disease prediction, then, we examine existing literature on the enhancement of ML predictions through the incorporation of semantic knowledge, and finally, we discuss a broader field where sub-symbolic and symbolic techniques are combined, known as Neurosymbolic integration.

Machine Learning for Heart Disease Prediction. The application of ML in healthcare, particularly for predicting heart disease, has been an area of significant research interest. Early efforts in this domain primarily relied on extensive tabular datasets, such as electronic health records and clinical trial data. Studies such as those by Yadav et al. [24] and Shah et al. [21] demonstrate the effectiveness of traditional ML algorithms in predicting heart disease, while Mohan et al. [16] combined two different ML techniques to enhance prediction accuracy.

The performance of ML models is often hindered by the lack of data or data of suboptimal quality. In the healthcare domain, ontologies, as discussed by [10, 17, 12], provide a structured and semantically rich layer of information, that can be used to improve the contextual understanding of ML models.

Enhancing Machine Learning Predictions with Semantic Knowledge. To address the challenges of ML methods, in terms of data-reliance, recent research has explored the integration of semantic knowledge, such as ontologies and KGs, into the ML framework, in various domains. In their study on opinion mining, Alfrjani et al. [1] merge semantic knowledge bases with ML to enhance data analysis and classification accuracy. Their methodology involves constructing a "Semantic Feature Matrix," capturing structured semantic information related to domain features in reviews. Similarly, in the field of smart building management, Szilagyi et al. [23] examine an intelligent system that

integrates ML and semantic knowledge, in the form of taxonomies, schemas, logic rules and so on. This hybrid model optimizes building management, highlighting the synergy between data-driven insights and rule-based knowledge. Taking a distinct approach, Ziegler et al.'s [26] research focuses on enhancing neural networks with semantic knowledge using graph embeddings for the credit-card fraud detection task. In their approach they used embeddings created for country nodes in DBpedia [15] as well as they augmented the dataset further with information about public holidays and showed that injecting semantic background knowledge improves the classification. Bhatt et al.[2] present four case studies showcasing the application of KGs (in different forms) to enhance ML: sentiment analysis, personalized news recommendation, machine translation and recommender systems.

The papers discussed above demonstrate that the integration of ML with semantic data yields significant benefits. Therefore, in this paper, we aim to apply this approach within the domain of heart disease prediction.

Neurosymbolic Integration The integration of symbolic and sub-symbolic AI techniques, known as neurosymbolic, significantly advances the field of AI by integrating complex reasoning with data-driven learning. There are different possibilities on integrating these two techniques. Henry Kautz, in his 2022 lecture "The Third AI Summer" [13], identified various forms of these neurosymbolic aproaches, each offering unique approaches to combining symbolic reasoning with neural networks. This classification provides a clear understanding of the diverse methodologies and potential applications of these hybrid techniques in AI. Similarly, Sheth et al. [14, 22] distinguished three levels of knowledge infusion into neural models: shallow, semi-deep, and deep. Shallow infusion involves the integration of syntactic and symbolic knowledge at the ML algorithm's input level. Semi-deep infusion incorporates this knowledge at the intermediate layers of the neural network, while deep infusion, the most integrated approach, embeds knowledge directly within the neural network, fostering a deeper interaction between knowledge and learning processes. Our research aligns with the shallow infusion approach, utilizing syntactic and symbolic knowledge as integral parts of the input to a ML algorithm. Our approach enriches the input data with semantic information, thereby enhancing the model's comprehension and processing capabilities.

Our research builds upon these foundations, using KGs as supplementary data sources to train ML models. We introduce a novel approach in the heart disease domain, incorporating embeddings from KGs into ML models, a method not previously explored, neither in medical prediction nor more broadly in NeSy inputs. Additionally, to the best of our knowledge, we are the first to examine how the size and structure of KGs, when used to train a ML model, affect its performance.

3 Problem Definition

This study aims to predict heart disease using ML techniques applied to clinical data. We focus on the challenge of accurately diagnosing heart disease based on a range of patient health indicators, highlighting the critical role of predictive analytics in improving medical outcomes.

Our primary data source is the Heart Disease dataset [1] from Kaggle, which includes information on 303 patients. For example, consider a patient profile extracted from the dataset: a 41-year-old female with atypical angina, a resting blood pressure of 130 mmHg, cholesterol levels at 204 mg/dL, a normal fasting blood sugar, left ventricular hypertrophy on ECG, a maximum heart rate of 172 bpm, no exercise-induced angina, a 1.4 mm ST depression, an upsloping ST segment, no major vessels observed via fluoroscopy, and a normal thalassemia result. The goal is to predict the presence or absence of heart disease. We refer to 'original' dataset the experiments performed with only the tabular data as shown in Figure 1a, using ML algorithms when no external knowledge is used.

4 Approach: Knowledge Graphs integration in ML models

In this section, we discuss our approach for integrating knowledge graphs (KGs) to augment the data used to train ML models, aiming to improve its performance. Our methodology consists of three steps: firstly, we generate KGs from existing ontologies; secondly, we apply KG embedding algorithms to transform these graphs; and finally, we discuss various strategies that we used to utilize these embeddings to augment tabular data, which are then used to train ML models.

[1] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

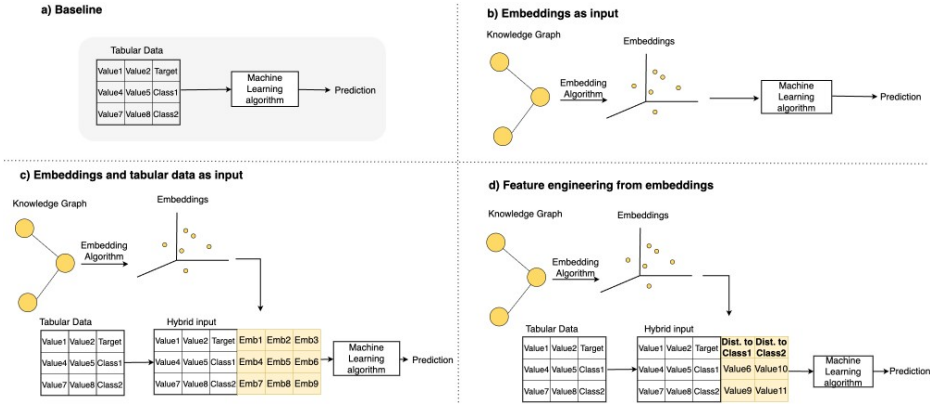

Figure 1: Approaches for medical diagnostic problem with variations of the input data: (a) Baseline method (without KG), (b) KG embeddings as direct inputs, (c) KG embeddings combined with tabular data, and (d) Feature engineering with distance metrics from KG embeddings.

Table 1: Details of the KGs for heart disease domain.

KG	Logical Axioms	Classes	Object prop.	Data prop.
Small	4637	29	6	10
Extended	6682	1664	6	10
Snomed	1963	80	24	10

Step 1. To enhance the predictive capabilities of the ML models, we have incorporated additional insights from three distinct KGs, each offering a unique perspective on the features present in the Heart Disease dataset. The KGs used we refer to as small, extended, and snomed. We obtained these KGs by populating existing ontologies with data from the original dataset discussed in the previous section. The small KG is created by populating the handcrafted ontology from Trepan Reloaded [4], which represents the features from the heart disease dataset. The extended KG, is created by populating an existing ontology [2], which we extended with the features from the dataset, and the snomed KG, is created by the population of the extracted ontology from SNOMED-CT [3], with the approach introduced by Chen et al. [3], that extracts ontologies from SNOMED-CT for specific domain based on an input set of seed-concepts that need to appear in the output ontology.

In Table 1, the details of KGs are outlined through key components: *logical axioms* that ensure the ontology's consistency by defining class and property relationships; *classes*, which are the ontology's fundamental elements e.g., 'Patient' or 'Disease'; *object properties* that link entities, such as 'hasSymptom' connecting 'Patient' to 'Symptom'; and *data properties* that detail entities with specific attributes, such as a 'Patient's' age.

Step 2. For the transformation of these KGs into a format readable by ML algorithms, we employed two embedding algorithms: RDF2Vec [19] and Node2Vec [7]. RDF2Vec is specifically designed for RDF (Resource Description Framework) graphs in the Semantic Web, using random walks to generate sequences from RDF graphs, which are then converted into embeddings via Word2Vec models, capturing both semantic and relational attributes. Node2Vec, on the other hand, is a more versatile algorithm suitable for various graph types, whether labeled or unlabeled, directed or undirected. It employs random walks with an adjustable bias parameter to explore different search strategies, effectively capturing both the local neighborhood and global structure of the graph. Thus, while both methods use random walks for generating embeddings, RDF2Vec specializes in the semantic web domain, focusing on semantic relationships, whereas Node2Vec is adaptable to diverse graph structures, emphasizing structural information.

[2] https://bioportal.bioontology.org/ontologies/HFO
[3] https://www.snomed.org

Table 2: Parameter grid for ML methods

Method	Parameter	(Grid) Values
KNN	n_neighbors	[5, 10, 15, 20, 25, 30, 35, 40, 45, 50]
SVM	C; kernel; probability	[0.1, 0.2, ..., 2.1]; linear; True
XGB	learning_rate	[0.01, 0.02, ..., 0.20]
NN	layers; activation; loss; optimizer	[32, 16, 1]; [relu, relu, sigmoid]; binary crossentropy; adam

Step 3. Leveraging the capabilities of RDF2Vec and Node2Vec, with the goal of enhancing ML algorithm performance, we have created six datasets that use the rich semantic and topological information encoded by these embeddings, either alone or in addition with the original dataset. Then we used these augmented datasets to train ML models. These approaches include:

- **Baseline**: It represents the traditional approach for training a ML algorithm (shown in Figure 1a), where as input is a tabular dataset, in our case original dataset.
- **RDF2Vec and Node2Vec**: These approaches involve training ML algorithms using datasets exclusively composed of RDF2Vec and Node2Vec embeddings, respectively. These embeddings are obtained by converting the KG into a vector space, effectively capturing the semantic and topological characteristics of the graph's resources., shown in Figure 1b.
- **Combined RDF2Vec and Combined Node2Vec**: These approaches involve training ML algorithms with datasets that are fusion of the original dataset with RDF2Vec and Node2Vec embeddings, respectively, as additional columns (shown in Figure 1c). This ensures that each instance (Patient) is represented not only by the original features but also by a corresponding vector space representation. This augmentation of original dataset is intended to provide a richer set of features to leverage both raw features and the semantic knowledge.
- **FE_RDF2Vec and FE_Node2Vec**: These approaches use datasets enhanced with RDF2Vec and Node2Vec embeddings through feature engineering for training ML algorithms (shown in Figure 1d). In addition to the original features, two extra features are incorporated to further refine the sample representations. These new features are calculated based on the Euclidean distances from the sample embeddings to predefined class centroids. Specifically, for each patient in the dataset, two additional columns are introduced. The first column represents the Euclidean distance between the patient's embedding (generated through RDF2Vec or Node2Vec) and the vectorial representation of the 'yes' class centroid, indicating the presence of heart disease. The second column, similarly, denotes the Euclidean distance to the 'no' class centroid, representing the absence of heart disease. By including these distance metrics, the datasets are enhanced to more accurately reflect the proximity of each patient's data to the known class centroids, thus potentially improving the classification performance in identifying whether a patient has or does not have heart disease.

5 Experiment Setup

The goal of our experiments is to evaluate the performance of ML models in predicting heart disease by leveraging the integration of domain knowledge from knowledge graphs with the original dataset.

For our experiments we used fours different supervised ML models, these being: (i) K-Nearest Neighbors (KNN), (ii) Support Vector Machines (SVM), (iii) eXtrem Gradient Boosting (XGB) and (iv) Feedforward Neural Network (NN). We choose the first two models because they deal with distances and since we are using KG embeddings, we expect a better performance, and the other two to investigate applicability for other more complex models. As part of the training process, we tuned the hyperparameters of the models using a grid search, for which the search spaces of the hyperparameters are shown in Table 2.

To evaluate our approach, we used two evaluation metrics: accuracy and the F2 score. Accuracy is defined as the proportion of correctly predicted instances relative to the total number of instances. This metric provides a straightforward measure of overall model performance. Alongside accuracy, we used the F2 score, which is a more advanced metric that combines recall and precision, with a greater emphasis on recall. This choice of the F2 score is particularly pertinent in the context of

heart disease prediction, where the objective is to maximize the identification of true positive cases. Such an approach is crucial in medical diagnostics, as it prioritizes the detection of all potential true cases, even at the expense of a higher false positive rate. This strategy ensures that fewer true cases are overlooked, which is vital in a healthcare setting where the cost of missing a true case can be significantly higher than the cost of a false alarm.

6 Results

In this section, we present the results from our evaluation, based on the methodology Sections 3 and 4 and setup discussed in Section 5.

6.1 Infusing Knowledge Graphs into Machine Learning Pipelines (RQ1)

Table 3 show the average accuracy and F2 scores for different models using different inputs and when information from small, extended and snomed KGs are being used respectively. We can see that using the distances of the instances to the target classes as additional features to the tabular data, to train ML models, outperforms other datasets across all the models (KNN, NN, SVM and XGB) in both accuracy and F2 score across all KGs. This indicates the effectiveness of feature engineering combined with Node2Vec embeddings in enhancing model performance irrespective of KG modeling and ML model. Notably, K-Nearest Neighbors (KNN) and Neural Network (NN) models exhibit the highest performance gains with FE-N2V. For KNN the F2 score increases from 71% to 79%, 78% and 80% with small, snomed and extended KG respectively, and for NN from 78% to 84%, 82% and 82% with small, snomed and extended KG respectively.

Additionally, these results indicate that models utilizing the FE-R2V feature engineering approach show a decrease in performance when applied to extended KGs. This suggests that the distance-to-target-class features added from RDF2Vec are less effective when dealing with larger, more complex KGs, possibly due to overfitting or noise amplification. In contrast, Node2Vec, focusing on local neighborhood structures, is likely less affected by the complexity of the KG, thus avoiding these pitfalls and maintaining more consistent performance.

On the other hand, Combined RDF2Vec, and Combined Node2Vec approaches show similar performance trends, suggesting that the addition of RDF2Vec or Node2Vec embeddings to the original dataset maintains or slightly improves model performance. Whereas, RDF2Vec and Node2Vec embeddings alone (without combination with the original dataset) generally lead to a decrease in performance. This could be due to the loss of some critical features present in the original dataset.

Moreover we can observe that, even though RDF2Vec is designed specifically for KGs, its performance relative to Node2Vec is lower for our case. Node2Vec's ability to capture diverse structural relationships and align better with our models contributes to its better performance. Additionally, the effectiveness of adding embeddings as columns to the original dataset depends on whether they provide non-redundant information. Node2Vec, focusing on preserving node neighborhoods, can offer more informative additions compared to RDF2Vec, especially when the latter overlaps significantly with the original data. Furthermore, in our feature engineering approach involving Euclidean distances to class centroids, alignment with the 'heart disease' vs. 'no heart disease' division is crucial. RDF2Vec may struggle in this regard due to the nature of the semantic relationships captured, making Node2Vec-derived distance features more predictive.

6.2 Impact of Knowledge Graph Size and Structure (RQ2)

In Figure 2, we present the average F2 scores of various models under different inputs: the original dataset with no KG and datasets enhanced with features from small, extended, or snomed KGs. The graph demonstrates that the incorporation of KG features generally enhances model performance in terms of F2 score, even if only slightly.

Notably, the neural network exhibits the best performance when trained on data augmented with the small KG, indicating its effectiveness in extracting and using the extra features introduced by this specific KG. In the case of the SVM, there is an interesting observation where both small and extended KGs yield equivalent performance improvements. This suggests that the SVM is capable of efficiently leveraging the added complexity and richness of these KGs to a similar extent. For other

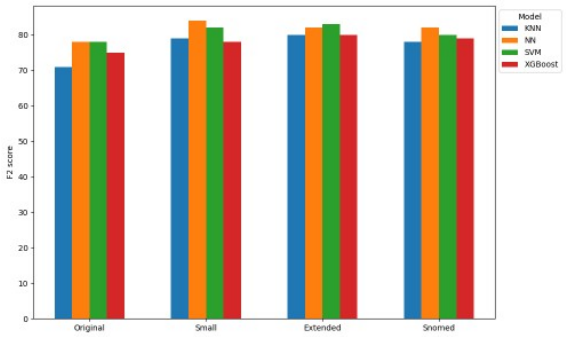

Figure 2: Average F2 Score for Different Models with no KG and Various KG-Modified Inputs.

ML models, including KNN and XGB, the extended KG emerges as the most beneficial in terms of accuracy enhancement. This may be attributed to the more comprehensive nature of the extended KG, which introduces a wider range of features and deeper semantic relationships into the dataset.

Table 3: Comparison of Accuracy and F2 Scores Across Models using various KG Inputs.

Model	Original		Rdf2Vec		Node2Vec		Comb-R2V		Comb-N2V		FE-R2V		FE-N2V	
	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2
Small KG														
KNN	0.81	0.71	0.51	0.34	0.72	0.59	0.81	0.71	0.81	0.71	0.65	0.53	**0.83**	**0.79**
NN	0.82	0.78	0.53	0.04	0.81	0.79	0.82	0.78	0.82	0.77	0.73	0.69	**0.85**	**0.84**
SVM	0.82	0.78	0.54	0.00	0.81	0.78	0.82	0.78	0.83	0.81	0.74	0.65	**0.84**	**0.82**
XGB	0.79	0.75	0.50	0.40	0.73	0.67	0.80	0.75	0.81	0.75	0.65	0.57	**0.80**	**0.78**
Snomed KG														
KNN	0.81	0.71	0.54	0.34	0.76	0.66	0.81	0.71	0.81	0.72	0.65	0.53	**0.83**	**0.78**
NN	0.82	0.78	0.57	0.29	0.79	0.75	0.82	0.78	0.82	0.77	0.70	0.65	**0.84**	**0.82**
SVM	0.82	0.78	0.54	0.00	0.80	0.75	0.82	0.78	0.82	**0.80**	0.69	0.62	**0.83**	**0.80**
XGB	0.79	0.75	0.58	0.48	0.76	0.70	**0.82**	0.77	**0.81**	0.77	0.64	0.58	**0.81**	**0.79**
Extended KG														
KNN	0.81	0.71	0.53	0.37	0.80	0.72	0.81	0.71	0.81	0.72	0.52	0.24	**0.84**	**0.80**
NN	0.82	0.78	0.54	0.05	0.80	0.78	0.82	0.79	0.83	0.80	0.55	0.26	**0.84**	**0.82**
SVM	0.82	0.78	0.54	0.00	0.79	0.76	0.82	0.78	0.83	0.80	0.56	0.22	**0.84**	**0.83**
XGB	0.79	0.75	0.53	0.43	0.77	0.73	0.82	0.77	0.81	0.77	0.54	0.41	**0.82**	**0.80**

6.3 Performance of Machine Learning Algorithms with Knowledge Graph Infusion (RQ3)

Table 3 show the consistency in the ranking of best-performing ML algorithms, regardless of the specific KG used, indicates that inherent compatibility of each algorithm with the given data type remains unchanged. NNs, being highly adaptable and capable of modeling complex non-linear relationships, consistently outperform other models. SVMs, with their robustness in high-dimensional spaces, follow closely. Lower complexity models such as KNN, despite improvements with the additional embeddings, are outperformed by more complex models. This is likely due to their inherent limitations in dealing with the augmented complexity introduced by the embeddings.

Figure 3 illustrate how F2 scores change across KNN, SVM, XGBoost, and NN models under various parameters (shown in Table 2), with and without KG inputs. In most cases, parameter tuning do not change the algorithmic performance ranking, regardless of the data input. However, KNN show a linear F2 score increase with more neighbors when using node2vec embeddings, while a performance decrease is noted with increased neighbors when using FE-R2V inputs.

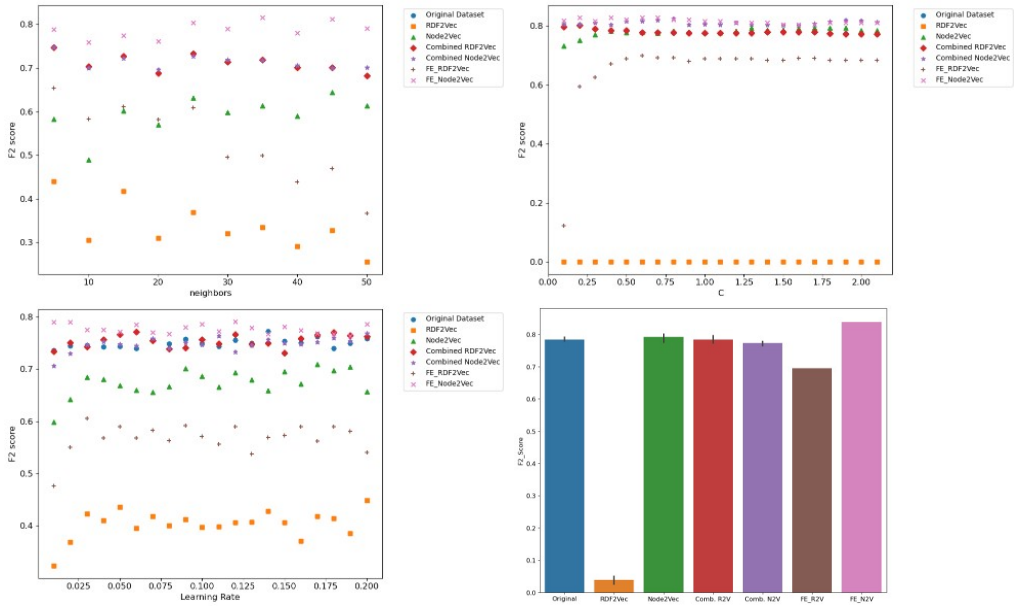

Figure 3: Average F2 scores over different parameters for (top-left) XGBoost, (top-right) NN, (bottom-left) KNN and (top-right) SVM.

7 Conclusion and Future Work

In addressing RQ1, this study demonstrates the potential of integrating knowledge graphs (KGs) with machine learning algorithms for heart disease prediction. The use of RDF2Vec and Node2Vec embeddings to incorporate semantic and topological information from KGs into the machine learning pipeline has led to enhancements in model performance. Specifically, the application of the Node2Vec embeddings as extra features in the dataset has resulted in improvements in both accuracy and F2 scores. This underscores the value of advanced feature engineering through the infusion of KG-based features, thereby providing a robust foundation for enhancing machine learning models in healthcare.

Concerning RQ2, the experiments reveal that the size and structural complexity of KGs influence the performance gains across different machine learning algorithms. While neural networks tend to favor features derived from smaller KGs, algorithms such as SVMs exhibit equal benefits from both small and extended KGs. Conversely, models such as KNN and XGB show a preference for features from extended KGs. These insights highlight the role of KG size and structure in optimizing machine learning performance, emphasizing the necessity for strategic feature selection that aligns with the algorithm's characteristics.

In response to RQ3, the comparative analysis of machine learning algorithms, when augmented with KG information, illustrates a universal improvement in performance due to the richer feature set provided by RDF2Vec or Node2Vec embeddings. However, this enhancement does not change the inherent performance ranking among the models, indicating that the optimal selection of a machine learning algorithm remains dependent on its strengths and the specific nature of the dataset. This finding points to the importance of the choice of machine learning model that complements the characteristics of the KG-infused data.

Regarding future work, we plan to include various domains beyond heart disease, exploring the wider utility of KGs in machine learning. It's also important to investigate different embedding algorithms to find more efficient ways for integrating domain knowledge, and use different advanced machine learning techniques. We also plan to focus on measuring the data dependency of machine learning algorithms and comparing this with the complementary contributions from KGs. This comparative analysis would provide valuable insights into how the integration of KG can either reduce or shift the data reliance in machine learning models, leading to a better understanding of their synergistic potential and limitations.

Acknowledgements

This work was supported by the FWF HOnEst project (V 754-N), FFG SENSE project (894802) and FAIR-AI project (904624).

References

[1] Rowida Alfrjani, Taha Osman, and Georgina Cosma. A hybrid semantic knowledgebase-machine learning approach for opinion mining. *Data & Knowledge Engineering*, 121:88–108, 2019.

[2] Shreyansh Bhatt, Amit Sheth, Valerie Shalin, and Jinjin Zhao. Knowledge graph semantic enhancement of input data for improving AI. *IEEE Internet Computing*, 24(2):66–72, 2020.

[3] Jieying Chen, Ghadah Alghamdi, Renate A Schmidt, Dirk Walther, and Yongsheng Gao. Ontology extraction for large ontologies via modularity and forgetting. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 45–52, 2019.

[4] Roberto Confalonieri, Tillman Weyde, Tarek R Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021.

[5] Artur d'Avila Garcez and Luis C Lamb. Neurosymbolic AI: The 3 rd wave. *Artificial Intelligence Review*, pages 1–20, 2023.

[6] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman 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*, pages 753–762, 2018.

[7] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[8] David Herron, Ernesto Jiménez-Ruiz, and Tillman 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*, volume 3432, pages 327–335. CEUR Workshop Proceedings, 2023.

[9] Pascal Hitzler, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Lu Zhou. Neuro-symbolic approaches in artificial intelligence. *National Science Review*, 9(6):nwac035, 2022.

[10] Mirjana Ivanović and Zoran Budimac. An overview of ontologies and data resources in medical domains. *Expert Systems with Applications*, 41(11):5158–5166, 2014.

[11] Daniel Jarrett, Eleanor Stride, Katherine Vallis, and Mark J Gooding. Applications and limitations of machine learning in radiation oncology. *The British journal of radiology*, 92(1100):20190001, 2019.

[12] Alan Jovic, Marin Prcela, and Dragan Gamberger. Ontologies in medical knowledge representation. In *2007 29th International Conference on Information Technology Interfaces*, pages 535–540. IEEE, 2007.

[13] Henry Kautz. The third AI summer: Aaai robert s. engelmore memorial lecture. *AI Magazine*, 43(1):105–125, 2022.

[14] Ugur Kursuncu, Manas Gaur, and Amit Sheth. Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning. *arXiv preprint arXiv:1912.00512*, 2019.

[15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[16] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554, 2019.

[17] Domenico M Pisanelli. *Ontologies in medicine*, volume 102. IOS press, 2004.

[18] Konstantinos Poulinakis, Dimitris Drikakis, Ioannis W Kokkinakis, and Stephen Michael Spottswood. Machine-learning methods on noisy and sparse data. *Mathematics*, 11(1):236, 2023.

[19] Petar Ristoski and Heiko 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*, pages 498–514. Springer, 2016.

[20] Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence. *AI Communications*, 34(3):197–209, 2021.

[21] Devansh Shah, Samir Patel, and Santosh Kumar Bharti. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1:1–6, 2020.

[22] Amit Sheth, Manas Gaur, Ugur Kursuncu, and Ruwan Wickramarachchi. Shades of knowledge-infused learning for enhancing deep learning. *IEEE Internet Computing*, 23(6):54–63, 2019.

[23] Ioan Szilagyi and Patrice 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)*, pages 20–25. IEEE, 2018.

[24] Anup Lal Yadav, Kamal Soni, and Shanu Khare. Heart Diseases Prediction using Machine Learning. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2023.

[25] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. Domain knowledge guided deep learning with electronic health records. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 738–747. IEEE, 2019.

[26] Konstantin Ziegler, Olivier Caelen, Mathieu Garchery, Michael Granitzer, Liyun He-Guelton, Johannes Jurgovsky, Pierre-Edouard Portier, and Stefan 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)*, pages 200–205. IEEE, 2017.