Exploring Knowledge Graphs for Machine Learning Enhancement

Majlinda Llugiqi

Vienna University of Economics and Business

Abstract

In recent years, machine learning (ML) methods have seen widespread adoption across various domains due to their ability to perform complex tasks with high accuracy. However, their applicability remains limited in certain critical fields, where the demand for reliable decision-making is crucial. Furthermore, the efficacy of ML models is often limited by data availability and quality, posing significant challenges especially in data-sensitive areas. Addressing these limitations, this thesis explores the integration of semantic knowledge through knowledge graphs to enhance the performance of ML models. Knowledge graphs, with their structured representation of domain-specific information, offer a way to augment ML models with domain insights, improving their performance and potentially overcoming the issues of data scarcity. This research contributes to the field of Neurosymbolic AI and Semantic Web by not only demonstrating the feasibility and benefits of combining knowledge graphs with ML but also by offering guidance on the effective construction and utilization of knowledge graphs for the purpose of ML enhancement.

Keywords

Neurosymbolic AI, Knowledge Graphs, Knowledge Infusion, Machine Learning

1. Problem Statement and Importance

The application of artificial intelligence (AI) techniques across diverse fields has significantly expanded the potential of Machine Learning (ML) models in addressing complex problems. The third wave of AI, Neuro-symbolic (NeSy) AI [1], seeks to combine the adaptability of sub-symbolic AI techniques, such as ML, with the logical structuring of symbolic AI, aiming to address the limitations of traditional AI techniques [2, 3]. Central to this approach is the incorporation of Knowledge Graphs (KGs), which organize intricate domain knowledge in a structured format that ML models can utilize effectively.

Despite the promising advancements brought by NeSy AI, several challenges remain prevalent. ML models require large datasets for training, which can be sparse, of low quality, or biased, particularly outside mainstream applications [4, 5]. This dependency on extensive and high-quality data represents a critical barrier, as acquiring adequate data is often not feasible.

Considering these challenges, an emerging area of research focuses on knowledge infusion into ML. In this thesis we explore how the infusion of KGs influence ML outcomes, especially concerning accuracy, data sparsity, and data quality. Moreover, we also investigate how the

Proceedings of the Doctoral Consortium at ISWC 2024, co-located with the 23rd International Semantic Web Conference (ISWC 2024)

majlinda.llugiqi@wu.ac.at (M. Llugiqi)
0000-0002-5008-6856 (M. Llugiqi)

quality, granularity, and relevance of KGs influence the improvement of ML models when these KG are infused. To our knowledge, these dependencies have not been extensively investigated previously. Addressing these issues entails refining the ways in which knowledge is represented and utilized, ensuring that KGs not only supplement but also synergize with ML processes to drive more accurate and reliable results.

This research focuses on the settings where a supervised ML model is used for prediction and investigates methods to enhance this through knowledge infusion. Illustrative use cases examined include the prediction of heart disease using tabular data on patients and their clinical features [6], as well as for prediction of kidney disease [7]. These studies explore how the accuracy of ML predictions can be improved by incorporating KGs, outlining the necessary steps, such as the construction of a KG and its integration into ML workflows, as outlined in Figure 1.

Problem Importance This thesis tackles a significant challenge in ML by enhancing model performance despite the constraints of insufficient or low-quality data. The main benefits of this thesis are three fold. First, by integrating semantically rich knowledge from KGs into ML pipelines, this thesis aims to reduce the reliance of ML models on large data sets, which is particularly crucial in sensitive domains where data is scarce, such as medical domain. The integration of KGs promises to improve decision-making processes, enhancing the accuracy and reliability of ML applications. Second, this thesis advances the field of NeSy AI by exploring effective methods and proposing an innovative approach to use KG insights into ML pipelines, thus enriching both the research community and practical applications. Third, the findings of this thesis are also beneficial for KG engineers, providing them with detailed insights on the characteristics that KGs should possess to optimally support ML models.

2. Related Work

The broadest area related to this research is the integration of sybmbolic and sub-symbolic AI, known as neurosymbolic AI. This integration combines reasoning with data-driven learning, significantly advancing AI capabilities. As outlined by Henry Kautz [8] there are several methods for integrating these techniques, each offering unique benefits for combining symbolic reasoning with neural networks. Sheth et al. [9] further differentiate three levels of knowledge infusion into neural models: shallow, semi-deep, and deep. Shallow infusion introduces syntactic and symbolic knowledge at the input level, semi-deep at intermediate layers, and deep infusion embeds knowledge directly within the neural network, fostering deeper interactions between knowledge and learning processes. Dash et al. [10] reviewed various techniques that include domain knowledge into deep neural networks. They emphasize the transformative impact of such knowledge integration across three main aspects of neural network: the input data, the loss function, and the network architecture (including the structure or the parameters).

Within the broader scope of NeSy described earlier, one specific focus is the integration of semantic knowledge into ML pipelines. Examples of such successful integration are as follows. Knowledge-based artificial neural networks (KBANN) [11] integrate symbolic knowledge into neural networks by encoding initial knowledge as rules into the network architecture. This

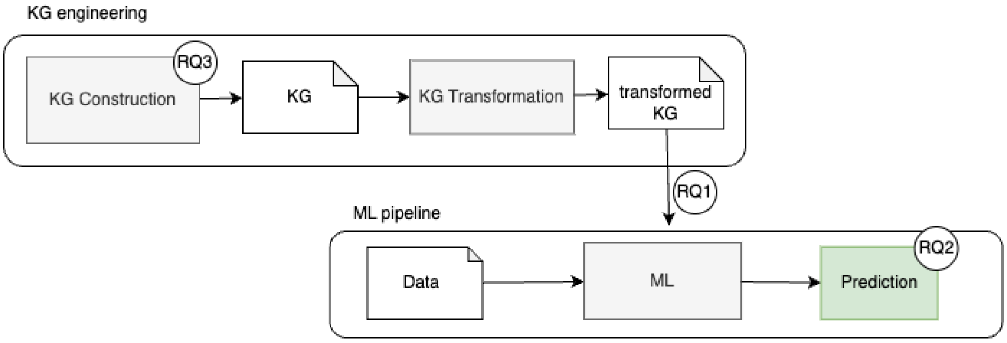

Figure 1: An overview of the methodology followed in this thesis.

approach provides a structured starting point for learning, allowing the network to refine and adapt the knowledge during training, leading to improved generalization and learning performance. In opinion mining Alfrjani et al. [12], used semantic feature to enhance data classification accuracy. Similarly, Ziegler et al. [13] enhanced neural networks for fraud detection by using graph embeddings to inject semantic knowledge, illustrating improved classification outcomes. Gazzotti et al. [14] tackled sparse data in electronic medical records by augmenting features with ontological resources, showing potential in improving predictive analytics. For image classification, [15] presents the Graph Search Neural Network, a framework that integrates large knowledge graphs into classification pipelines. This approach uses structured knowledge of object traits and relationships to boost reasoning about visual concepts with fewer examples. The findings show that knowledge graphs substantially improve multi-label classification, outperforming standard neural network baselines.

Besides the use of semantic knowledge to augment the input of the ML pipeline, there have been successful efforts to incorporate domain knowledge into the loss functions and weights of neural networks [16, 17, 18, 19]. We aim to explore further methods for leveraging knowledge from semantic structures such as knowledge graphs.

These examples underline the diverse applications and potential of semantic structures in enhancing machine learning, setting a precedent for our research. However, there are only a limited number of approaches of KG based infusion. Additionally, there is minimal empirical evidence on how KG characteristics influence this integration.

3. Research Questions and Contributions

In this thesis, we will focus on the use of the KGs to enhance the predictions of the ML models. We aim to investigate the following overall research question: **How to effectively and efficiently infuse KG into ML pipeline?** In particular, we will investigate the following sub-research questions:

RQ1: *What are the approaches to infuse KGs into ML pipelines?* Understanding the approaches to infuse KGs into ML pipelines is crucial, as KGs provide contextualized and structured information that can potentially enhance ML models. The explicit representation of knowledge in KGs enables more informed feature engineering, helps in data preprocessing, and can be used for better interpretation of model outcomes. To address this question, we aim to explore various ways to leverage KGs for enhancing the predicting performance of ML methods and introduce a novel approach. Our focus will be on identifying the most effective ways to integrate KGs into various parts of the ML pipeline, in order to improve the performance of ML models. The contribution of addressing this question is to identify and categorize various methods to leverage KGs effectively, thereby proposing an innovative approach to enhance ML workflows. This framework could serve as a guideline for practitioners and researchers to optimize the performance of their ML systems using KGs.

RQ2: *To what extent does the infusion of KGs into ML pipelines impact model performance?* Assessing the impact of KG infusion is critical to validate the theoretical benefits of KGs and understand their effectiveness in real-world applications. A key focus of our investigation will be to explore whether integrating KGs can enhance the ML performance in terms of accuracy and F-measures, compared to the baseline when no KG is used. Additionally, we aim to address the data-dependency problem inherent in ML models, enabling effective learning with fewer data by incorporating semantically rich, structured knowledge. The investigation of this question will empirically evaluate the performance changes in ML models with KG infusion across various use cases. This provides evidence-based insights into the advantages and possible limitations of KG integration, enabling more informed decisions in the development of ML models.

RQ3: *How do the characteristics of KGs influence the performance of ML models?* It is important to explore how specific KG characteristics, such as size and modeling techniques, affect ML model performance in order to guide the design of more effective KGs tailored to enhance specific ML tasks. This question is critical within the field of knowledge engineering, as it informs the development of optimized KGs tailored for integration into NeSy systems. By understanding the specific characteristics that contribute to performance improvements, we can design more effective and meaningful KGs suited for advanced ML applications. Exploring this question contributes by developing novel metrics or requirements for KGs to ensure their optimal infusion into ML models. This research could lead to a set of best practices for KG construction and refinement specifically aimed at enhancing ML applications.

4. Research Methods

In this thesis, we will adopt the Design Science methodology [20] to systematically develop and evaluate artifacts that infuse knowledge from KG into various steps of ML pipeline. In the following we show how we will address each cycles.

Relevance Cycle This cycle connects the research to the real-world by identifying the practical problems that the artifacts aim to solve and demonstrating the artifacts' impact in a real application setting. In our case, the relevance cycle involves engaging with stakeholders in the ML and semantic web communities to identify key challenges in integrating KGs with ML models, using methods such as literature study. We will collaborate with industry professionals and academic peers to ensure that our solutions are applicable and address significant needs.

Design Cycle We will design approaches that integrate KG data into ML systems, addressing our first research question (RQ1) which focuses on different ways of knowledge infusion. Our current investigation explores the infusion of KG embeddings as part of the input to the ML model, augmenting tabular data with semantics from KGs constructed using instances from the same dataset. Moreover, we will assess the impact of these integrations, and evaluate the enhancement of ML models to demonstrate how KGs can enhance ML model performance, directly tackling our second research question (RQ2). Currently we are focusing in medical domain, for heart and kidney disease prediction. Lastly, we will evaluate how different characteristics of the KGs influence ML methods' improvement, thus answering our third research question (RQ3).

To evaluate the research questions, the improvements of ML methods and the impact of KGs, we will compare the baseline when no additional knowledge is being used, with different infusion methods. We will measure accuracy and F-scores as evaluation metrics. Accuracy reflects the proportion of correct predictions, while the F-score, a combination of precision and recall, provides a balanced measure of a model's precision and its ability to identify all relevant instances. The choice for F-score is particularly crucial in fields such as medicine where the consequences of false negatives or positives are significant, making the F-score a more comprehensive indicator of model performance than accuracy alone.

Preliminary Results In our previous work [6] we introduced methods for integrating KGs into ML pipelines, specifically using KG embeddings with tabular data to improve algorithms for predicting heart disease. We conducted a comparative analysis of different methods for merging KGs with tabular datasets and used two embedding algorithms to enhance KG representation. In [7], we formalized these approaches, introduced a new variation, and extended our experiments to include the prediction of chronic kidney disease. Our results for heart disease prediction in [6] showed improvements in accuracy and F2 score, with the accuracy of the Feed-Forward Neural Network increasing from 82% to 85% and the F2 score for the K-Nearest Neighbors model improving from 71% to 80%. This is a step into the direction of answering RQ1 and RQ2. Preliminary findings also indicate that the size and modeling techniques of KGs impact ML algorithm performance, with addressing RQ3.

Rigor Cycle We will ensure the rigor cycle by incorporating knowledge from existing literature and theoretical foundations in the fields of ML and KGs. Our approach will also utilize established methodologies and frameworks from previous research to guide the design and evaluation processes. Throughout the thesis, we will maintain a rigorous scientific approach, using well-defined metrics and evaluation methods to ensure that our findings are valid, reliable, and reproducible.

The planned work for this thesis is depicted in Figure 1, outlining our structured approach to explore, implement, and evaluate the proposed enhancements.

5. Reflection and Future Work

This thesis aims to contribute in the Neuro-Symbolic AI field by exploring methods to integrate knowledge from knowledge graphs into the machine learning pipeline, aiming to enhance

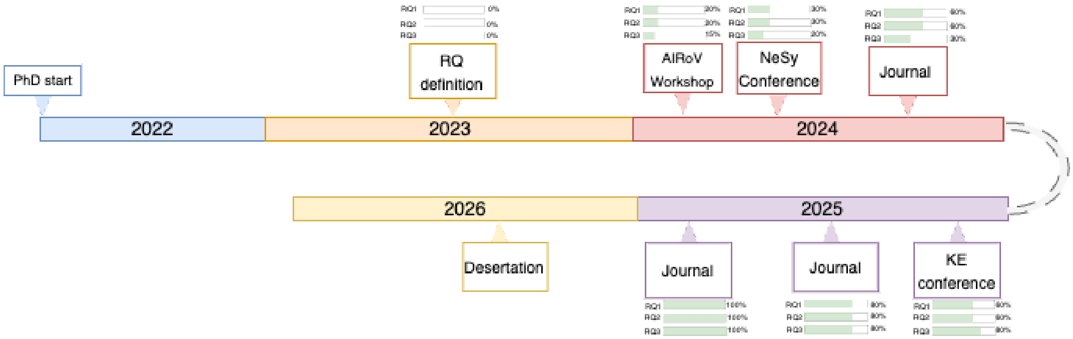

Figure 2: An overview of the timeline planned for this thesis.

their performance. Additionally, this thesis contributes in the knowledge engineering area, by identifying characteristics and modeling techniques for knowledge graphs, making them more effective for the purpose of enhancing machine learning methods in prediction.

As illustrated in Figure 2, I started my PhD in July 2022, initially focusing on familiarizing myself with the field and identifying research gaps. Until now, I have developed essential research skills, participated in PhD courses as part of my curriculum, and engaged in teaching activities. Additionally, I have co-authored several papers in the neuro-symbolic area, enhancing my understanding and expertise in this field. Moreover, I have contributed to publications and presented preliminary findings on RQ1, RQ2, and RQ3 at the Airov Workshop [6], and most recently, submitted a paper to the NeSy conference [7]. Moving forward, our plan is as follows:

- *Short-term (Now - End of 2024)*: We aim to focus on RQ1 and RQ2 by conducting extensive evaluations of the methods proposed for infusing KG embeddings to augment data. This phase will include investigating other domains beyond the medical field and exploring additional machine learning and embedding methodologies. Moreover, we plan to investigate whether the infusion of KGs can help address the data-dependency problem in machine learning. The findings from these studies are targeted for submission to a prominent journal.
- *Mid-term (Early 2025)*: We plan to address RQ3 by proposing novel evaluation metrics and characteristics of KGs that enhance ML methods, aiming for publication at a knowledge engineering conference. This work will also examine other ways to integrate semantic knowledge into the different parts of the ML pipeline.
- *Long-term (2025 - Mid 2026)*: The final phase of our research will involve further conferences and journal publications to solidify our contributions and expand the applicability of our findings. This will lead to the completion of my dissertation by mid-2026.

6. Acknowledgments

I would like to thank my supervisors, Prof. Marta Sabou and Asst. Prof. Fajar J. Ekaputra for their valuable advises, expert insights and guidance. The work presented in this paper is supported by the FWF HOnEst project (No. V 745-N).

References

[1] A. d. Garcez, L. C. Lamb, Neurosymbolic ai: The 3 rd wave, Artificial Intelligence Review 56 (2023) 12387–12406.

[2] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022) nwac035.

[3] M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence, AI Communications 34 (2021) 197–209.

[4] L. Von Rueden, S. Mayer, J. Garcke, C. Bauckhage, J. Schuecker, Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning, Learning 18 (2019) 19–20.

[5] K. Poulinakis, D. Drikakis, I. W. Kokkinakis, S. M. Spottswood, Machine-learning methods on noisy and sparse data, Mathematics 11 (2023) 236.

[6] M. Llugiqi, F. J. Ekaputra, M. Sabou, Leveraging knowledge graphs for enhancing machine learning-based heart disease prediction (2024).

[7] M. Llugiqi, F. J. Ekaputra, M. Sabou, Enhancing machine learning predictions through knowledge graph embeddings (conditionally accepted) (2024).

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

[9] A. Sheth, K. Roy, M. Gaur, Neurosymbolic Artificial Intelligence (Why, What, and How), IEEE Intelligent Systems 38 (2023) 56–62.

[10] T. Dash, S. Chitlangia, A. Ahuja, A. Srinivasan, A review of some techniques for inclusion of domain-knowledge into deep neural networks, Scientific Reports 12 (2022) 1040.

[11] G. G. Towell, J. W. Shavlik, Knowledge-based artificial neural networks, Artificial intelligence 70 (1994) 119–165.

[12] R. Alfrjani, T. Osman, G. Cosma, A hybrid semantic knowledgebase-machine learning approach for opinion mining, Data & Knowledge Engineering 121 (2019) 88–108.

[13] I. Szilagyi, 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.

[14] R. Gazzotti, C. Faron-Zucker, F. Gandon, V. Lacroix-Hugues, 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.

[15] K. Marino, R. Salakhutdinov, A. Gupta, The more you know: Using knowledge graphs for image classification, arXiv preprint arXiv:1612.04844 (2016).

[16] M. Diligenti, M. Gori, C. Sacca, Semantic-based regularization for learning and inference, Artificial Intelligence 244 (2017) 143–165.

[17] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. Broeck, A semantic loss function for deep learning with symbolic knowledge, in: International conference on machine learning, PMLR, 2018, pp. 5502–5511.

[18] H. Yu, T. Li, W. Yu, J. Li, Y. Huang, L. Wang, A. Liu, Regularized graph structure learning with semantic knowledge for multi-variates time-series forecasting, arXiv preprint arXiv:2210.06126 (2022).

[19] H. Wang, F. Zhang, M. Zhang, J. Leskovec, M. Zhao, W. Li, Z. Wang, Knowledge-aware graph neural networks with label smoothness regularization for recommender systems, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 968–977.

[20] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS quarterly (2004) 75–105.