Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI
A. Salatino et al. (Eds.)

doi:10.3233/SSW240012

Semantically Describing Predictive Models for Interpretable Insights into Lung Cancer Relapse

Yashrajsinh CHUDASAMA [a,b,1], Disha PUROHIT [a,b] and Philipp D. ROHDE [a,b] and Enrique IGLESIAS [b,c] and Maria TORRENTE [d] and Maria-Esther VIDAL [a,b,c]

[a] *TIB-Leibniz Information Centre for Science and Technology, Hannover, Germany*
[b] *Leibniz University Hannover, Germany*
[c] *L3S Research Center, Hannover, Germany*
[d] *Hospital Universitario Puerta de Hierro-Majadahonda,Spain*

ORCiD ID: Yashrajsinh Chudasama https://orcid.org/0000-0003-3422-366X, Disha Purohit https://orcid.org/0000-0002-1442-335X, Philipp D. Rohde https://orcid.org/0000-0002-9835-4354, Enrique Iglesias https://orcid.org/0000-0002-8734-3123, Maria-Esther Vidal https://orcid.org/0000-0003-1160-8727

Abstract. Machine learning (ML) is becoming increasingly important in healthcare decision-making, requiring highly interpretable insights from predictive models. Although integrating ML models with knowledge graphs (KGs) holds promise, conveying model outcomes to domain experts remains challenging, hindering usability despite accuracy. We propose semantically describing predictive model insights to overcome communication barriers. Our pipeline predicts lung cancer relapse likelihood, providing oncologists with patient-centric explanations based on input characteristics. Consequently, domain experts gain insights into both the characteristics of classified lung cancer patients and their relevant population. These insights, along with model decisions, are semantically described in natural language to enhance understanding, particularly for interpretable models like LIME and SHAP. Our approach, *SemDesLC*, documents ML model pipelines into KGs, and fulfills the needs of three types of users: KG builders, analysts, and consumers. Experts' opinions indicate that semantic descriptions are effective for elucidating relapse determinants. *SemDesLC* is openly accessible on GitHub, promoting transparency and collaboration in leveraging ML for healthcare decision support.

Keywords. Knowledge Graphs, Machine Learning, Interpretability

1. Introduction

Lung cancer (LC) is Europe's leading cause of cancer death. LC is the most expensive disease in Europe, costing nearly 3 billion euros annually to care for patients [1]. Although expensive, lung cancer treatments can be more effective, and the chances of re-

[1] Corresponding Author: Yashrajsinh Chudasama, yashrajsinh.chudasama@tib.eu.

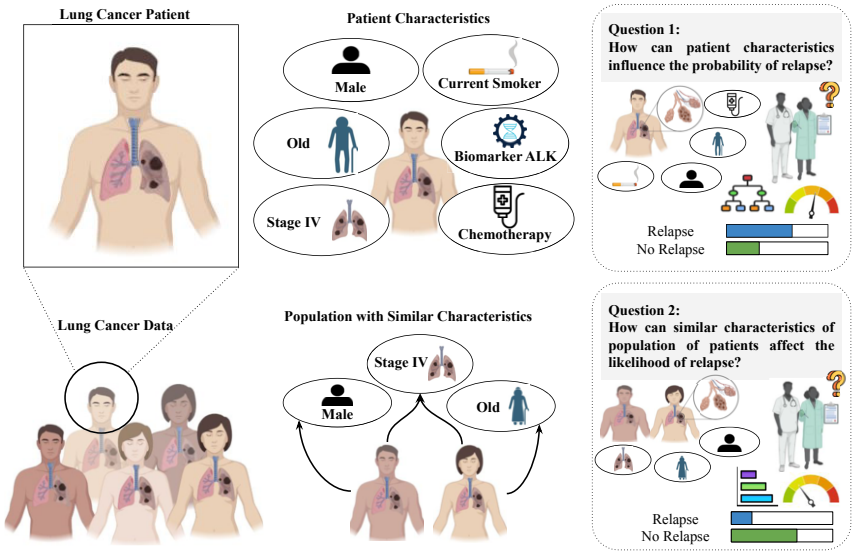


Figure 1. Motivating Example. An illustration of how the characteristics of lung cancer patients or sub-populations determine the state of *Relapse*. It also demonstrates the different questions that healthcare professionals take into account when deciding on therapy or medication for lung cancer patients.

sponding are better if discovered at an early stage. Biomedical data have grown exponentially in the previous decade. They include significant information that can be used for accurate illness diagnosis and personalized medical care, demonstrated in [2]. Central to this paradigm shift is the understanding that tailoring treatments for LC necessitates a thorough examination of patient characteristics. One critical aspect that demands attention is the forecast of *relapse*, where the resurgence of previously treated cancer poses a challenge, mainly when asymptomatic for long periods. However, within this complexity lies an opportunity: timely prediction of relapse holds the key to enhancing survival rates. Early detection and intervention may provide more treatment options and better outcomes for patients. Managing cancer relapse in lung cancer patients is crucial for extending life expectancy and enhancing life quality.

Implementing predictive models based on machine learning algorithms can greatly benefit healthcare professionals like doctors, researchers, and oncologists. These models offer promising opportunities to enhance the ability to predict the likelihood of cancer relapse in lung cancer patients [3]. By integrating diverse healthcare data, including patient demographics, medical history, and genetic markers, these models have the potential to provide actionable insights for timely interventions and personalized care plans. This convergence of technology and medical research promises to transform preventative healthcare, leading to proactive knowledge management and better outcomes for patients facing lung cancer relapse. Figure 1 highlights an example of integrating the *relapse* prediction problem with ML models for a patient and sub-populations.

Motivation. The motivation for our work arises from the fact that there is a noticeable gap that hinders oncologists from requiring automated support and interpretability of predictive model decisions. Figure 1 shows the lung cancer use case presented in the current study. While not all patients will experience relapse, effective treatments are critical for those who do. Domain experts often deal with questions like *How can patient*

characteristics influence the probability of relapse? or *How can similar characteristics of the population of patients affect the likelihood of relapse?* to determine the treatments suitable for the patients. In our lung cancer use case, patients are described by medical characteristics such as patient identifier, smoking habits, cancer stages, and treatments, along with additional features like family history, surgery, mutation, relapse, and performance status. Figure 1 illustrates the need for automated help and interpretability of prediction model decisions to bridge the communication gap with various KG users. The main objective of *SemDesLC* is to provide a pipeline that documents ML model pipelines into KGs. This enables domain experts to analyze treatment impacts, check if the patient satisfies medical protocols, and assess the likelihood of *relapse*.

State-of-the-art data-driven interpretable frameworks such as decision trees, SHAP [4], and LIME [5] generate a visualization, highlighting the important relevant features, and prediction probabilities. However, decision trees are prone to existing bias in the data, especially when the target class is imbalanced, resulting in decision trees with complex relationships less reliable in predicting the target class.

The domain experts indicated that the visualizations produced by existing interpretable frameworks (i.e., LIME and SHAP) were perceived as confusing, especially when there was a lack of explanation for the information presented in the plot. Consequently, the goal of this work is to implement a KG-based framework understandable by their users to maximize *usability*. Our aim is to achieve the following research objectives: **RO1)** Define a KG framework capable of efficiently improving the performance of predictive models. **RO2)** Utilize ML models to predict lung cancer relapse. **RO3)** Enhance understanding of ML models and the traceability of their decisions. Further, to meet our research objectives we aim to answer the following research questions: **RQ1)** Are techniques offered for KG creation efficient compared to the state-of-the-art methods? **RQ2)** How efficient are data quality assessments? **RQ3)** How efficient is federated query processing? **RQ4)** How do predictive models perform in terms of precision, recall, and F1-score? **RQ5)** How significant are the results of interpretability of *SemDesLC*?

Challenges. Various approaches have explored the challenges KG users face and their implications for *usability*. For example, Li et al. [6] report the results of a survey, which enabled the identification of three types of KG users and the main problems they faced when working with KGs: **a) Data Quality:** Users frequently identify data quality issues such as inadequate or missing data, incorrect data, or data redundancy. Furthermore, partial or erroneous data in the healthcare domain are frequently ambiguous and can lead to incorrect conclusions, which must be avoided. **b) Querying KGs:** Non-technical users have difficulty writing SPARQL queries on KGs to gain insights into predictive models, reducing their usability. **c) Lack of understanding of end user's needs:** Presenting predictive model conclusions without understanding the needs of domain experts yields results that are insignificant to domain experts. **d) Non-standarized nomenclature:** Another commonly faced challenge is the absence of defined terminology. Different groups may use the same word to represent multiple concepts or use different terms to describe the same concept. **e) Current KG Visualization Designs:** KGs visualization designs also fall short of domain specialists' expectations. As a result, many end users either do not understand the results or find it difficult to examine the insights provided. **Our Approach.** *SemDesLC*, is a computational framework that relies on KG technologies to trace and explain predictive models. *SemDesLC* creates an LC KG that integrates data collected from medical data sources and another KG that represents traces describing

the decisions made by the predictive models integrated in *SemDesLC*. The KG creation process is declaratively defined as a data integration system [7] using mapping assertions expressed in RML [8]. In addition, a federated query engine allows queries to be executed across the LC and *SemDesLC* KGs. Finally, SHACL provides the basis for validating integrity constraints expressed as SHACL shapes. *SemDesLC* is demonstrated in the lung cancer use case and evaluated in terms of the efficiency and effectiveness of the techniques integrated into *SemDesLC*. Following the categorization of users identified by Li et al. [6], the *SemDesLC* evaluation focuses on analyzing how well the needs of the *SemDesLC* users are met. Thus, we empirically evaluate the efficiency of *SemDesLC* to meet the needs of KG builders and analyze the accuracy of the predictive models to meet the needs of KG analysts. Finally, to improve communication with KG consumers, *SemDesLC* provides understandable visualizations and natural language (NL) explanations of predictive model predictions; the interpretability of these NL descriptions has been evaluated with the KG consumers of our LC use case.

Our contributions. 1) Analysis of the lung cancer use case to characterize the needs, tasks, and requirements of its users, as well as integrity constraints. **2)** *SemDesLC*, a KG-based framework that integrates state-of-the-art KG technologies to meet the needs of three types of users, i.e., KG builders, analysts, and consumers. **3)** Performance evaluation of tools used for KG creation, SHACL validation, and federated query processing. **4)** Predictive model performance analysis, including accuracy, precision, and recall. **5)** Interpretable descriptions of the predictive model results, enhanced and validated with feedback from experts in the lung cancer field. The remainder of the paper is as follows: Section 2 presents the LC use case, including requirements, roles, and tasks for KG actors. Section 3 introduces the terminologies and the proposed computational framework. The experimental evaluation of *SemDesLC* approach is reported in Section 4. Section 5 summarizes the state-of-the-art approaches. Section 6 reports the observed outcomes of *SemDesLC* and lastly in Section 7 outlined our conclusions and future works.

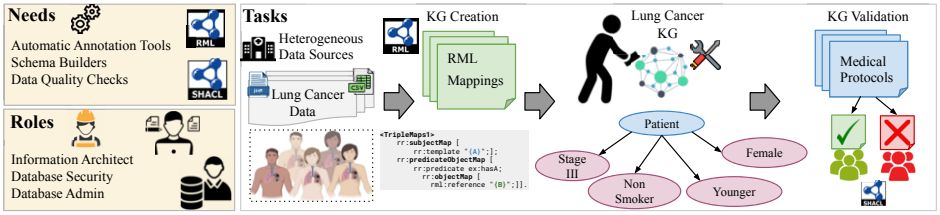
2. Use Case: Lung Cancer

In the LC use case, we deal with the classification challenge of categorizing patients as having the condition *Relapse* or *No Relapse*, depending on patients' or subpopulation characteristics. The early detection of relapse may permit the implementation of more effective treatment options, thereby improving the clinical results for patients. This can be achieved through the use of automated tools like *SemDesLC*. Figure 2 depicts the roles, needs, and tasks of the KG users in the LC use case described in this study.

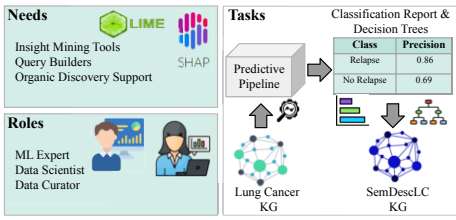
2.1. Main Actors in the LC Use Case

The following section describes in detail the different KG users in the LC use case, with an emphasis on the importance of each user in the process of the proposed framework.

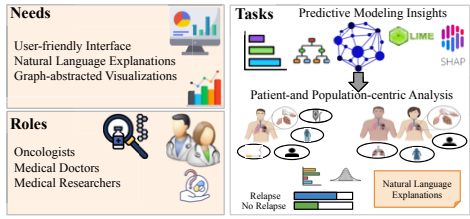
KG Builders as shown in Figure 2a encompasses users responsible for managing the data received, selecting the most appropriate database or representation mechanism to store it. Furthermore, the requirements of this user type include tools for generating KGs and performing data quality checks. The creation of KGs corresponds to the execution of the RML mapping assertions over the instances of the data sources. Additionally, KG



(a) **KG Builders** are typically experts in database systems, data management, or data modeling.



(b) **KG Analysts** are often experts in data science.



(c) **KG Consumer** are end users or domain experts.

Figure 2. Knowledge Graph Actors. The figure depicts three distinct KG actors: KG Builder (Figure 2a), KG Analyst (Figure 2b), and KG Consumer (Figure 2c) with varied skill levels in utilizing KGs. Each KG actor plays a vital role and has specific needs to accomplish the tasks required for that role.

validation is used to assess data quality using SHACL constraints. The constraints indicate whether an entity validates or invalidates the SHACL constraints, i.e., the medical protocols defined by the domain experts in the LC use case. The evaluation of constraints determines whether a specific entity violates the integrity constraints. The result is either true, indicating that the entity is valid, or false, indicating that the entity is invalid.

As illustrated in Figure 2b the role of **KG Analysts** is to utilize ML expertise and data science skills to generate insights. This is achieved through the use of interpretable tools that provide both local and global explanations. Predictive models are employed to generate classification reports and decision trees, which are used to draw conclusions based on the classification task. Furthermore, KG Analysts can evaluate data quality by reviewing generated decision trees with SHACL restrictions, allowing analysts to gain insights into data quality based on visualizations and the constraint validation report.

KG Consumers are domain experts, i.e., oncologists, medical doctors, or medical researchers as shown in Figure 2c. KG Consumers require insights that can be easily understood, as this assists them in interpreting the outcomes of predictive models. Users are more likely to interact with the data and accept the results when the data is presented in a way that hides the underlying graph structure. Furthermore, supplementary comprehensible insights would enhance their understanding, thereby increasing trust and usability of the outcomes generated by ML models for decision-making processes through knowledge-driven frameworks like *SemDesLC*. Subsection 2.2 describes how KG Consumers' requirements are collected to deliver insights that meet their demands.

2.2. Requirement Analysis

Requirement analysis is the process of identifying, documenting, and comprehending the needs and expectations of domain experts, in this case, medical doctors and researchers. This understanding is essential for the design, implementation, and manage-

ment of healthcare systems, applications, or services that align with the specific needs of the healthcare domain. By conducting a comprehensive analysis of the requirements in the healthcare domain, organizations can develop solutions that effectively address the needs of healthcare providers. The KG consumers have highlighted as relevant two types of analyses: **Patient-centric Analysis** refers to identifying characteristics of individual patients resulting in the condition of *Relapse*. As shown in Figure 1, KG Consumers are interested in identifying *"How can patient characteristics influence the probability of relapse?"*. The outcomes generated by predictive models often lack self-explanation. For example, when a predictive model indicates a young female patient is experiencing a relapse, it typically does not provide insight into the specific characteristics contributing to the occurrence of the relapse. Consequently, domain experts must investigate further into the patient's medical history before making a decision. This ultimately leads to a reduction in the use of predictive model outcomes. Therefore, patient-centric analysis requires the interpretation of a model's predicted outcome for an individual patient. It is important to present the reasons behind a particular patient being diagnosed with a *Relapse*; local interpretable tools like LIME [5], can be employed to generate local explanations.
Population-centric Analysis is a method of evaluating predictive model outcomes based on the sub-population of patients in the KG. As seen in Figure 1, KG Consumers are interested in determining *"How can similar characteristics of the population of patients affect the likelihood of relapse?"*. The predictive models may effectively analyze a range of patient characteristics, and suffer in explaining the reasoning behind their conclusions. This can pose challenges for domain experts who rely on a clear understanding of why certain predictions are made to make informed decisions. Thus, population-centric analysis allows us to identify patients with comparable features who fall into the same category, as well as evaluate predictive model decisions based on existing characteristics. For instance, a subpopulation of young female patients diagnosed with cancer at stage IV and exhibiting biomarker ALK are typically classified as belonging to the *Relapse* class. SHAP [4] generates global explanations, including feature importance, but fails to provide an overview of the model's behavior over the subpopulation.

3. SemDesLC

SemDesLC addresses the research objectives and challenges outlined in section 1 by aiding diverse KG users. This section presents the framework of *SemDesLC*, demonstrating the importance of semantically describing predictive models for interpretable insights.

3.1. Preliminaries

Data Integration System (DIS). The creation of a G is defined in terms of a data integration system $DIS_G = \langle O, S, M \rangle$ where O is a set of classes and properties of a unified ontology, S is a set of data sources, and M corresponds to mapping rules or assertions defining concepts in O as conjunctive queries over sources in S. The execution of the M rules over data from sources in S generates the instances of G.
Knowledge Graph (KG). Given a set *Con* of countable infinite constants. A *knowledge graph (KG)* is a directed *edge-labeled graph* $KG = (V, E, L)$, where $V \subseteq Con$ is a set of nodes, $L \subseteq Con$ is a set of edge labels, and $E \subseteq V \times L \times V$ is a set of edges.

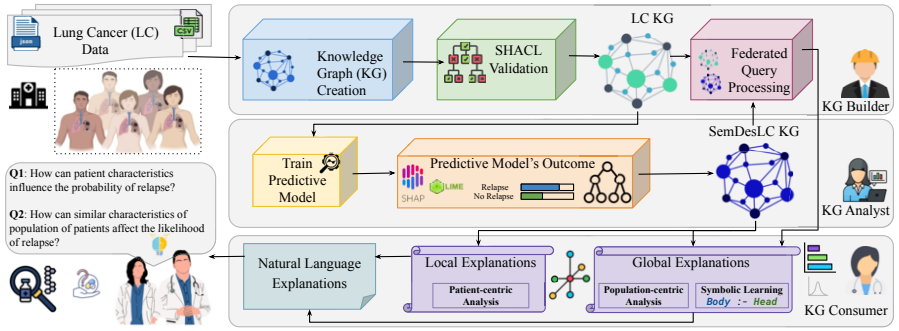


Figure 3. SemDesLC framework in a healthcare decision-support system, starting with creating KGs from several data sources about lung cancer patients, followed by ensuring data quality. Secondly, KG analysts obtain predictive models to classify target classes with probabilities. Lastly, our pipeline offers local and global explanations to determine the likelihood of relapse with visualizations and their natural language explanation.

Shapes Constraint Language (SHACL). The Shapes Constraint Language (SHACL) [9] is a language to define constraints over KGs. Constraints that are imposed over the same set of entities in the KG are comprised in a *shape*. A collection of different shapes that are validated over the same KG is called *SHACL shape schema*. The shapes in a SHACL shape schema can be connected. A constraint linking two shapes is referred to as *inter-shape constraint*. Hence, the remaining constraints are named *intra-shape constraints*.
Horn Rule. A Horn rule is defined as follows: $Body \Rightarrow Head$. The body of the rule is comprised of predicate facts. The head is a predicate fact of a single atom. All the variables in the *Head* are terms of at least one predicate fact in the *Body*. Every two predicate facts in Body share at least one variable. We say a rule $R : B_1 \wedge B_2 \wedge B_3 \wedge ..... \wedge B_n \implies r(x,y)$ where head $r(x,y)$ and body $B_1 \wedge B_2 \wedge B_3 \wedge ... \wedge B_n$.

3.2. *The* SemDesLC *Framework*

SemDesLC framework comprises a series of interconnected components designed to facilitate efficient KG creation and performing predictive analysis. *SemDesLC* receives as input heterogeneous data sources and offers an interpretable insight into predictive model decisions with natural language explanations. The interconnected *SemDesLC* framework components are as follows: **KG Creation** component receives a data integration system DIS_G as input, and generates KG. The creation of KG in biomedicine necessitates the integration of diverse data types, including drugs, genes, and clinical records, among others. Furthermore, the intricate process of generating KGs from heterogeneous data sources is addressed in the *SemDesLC* pipeline, which serves to facilitate users with varying interests. The component uses RML mapping engines [10,11,12] for the creation of LC KG that corresponds to the execution of the RML mapping assertions M over the instances of the data sources S. The generated LC KG is then delivered as input to the **SHACL Validation** component, which examines the quality of the data. The SHACL shapes represent constraints for validating the KG and for uncovering the impact of validation on models' decisions. The validation engines [13,14,15] are employed to test data quality utilizing SHACL constraints over the LC KG nodes. A validation report is generated for each constraint. The result of the validation report states *true*, indicating that the entity is valid, or *false*, indicating that the entity is not valid. The validation

component includes 15 constraints, with 10 *inter-shape constraints* and 5 *intra-shape constraints*. For instance, *"inter-shape constraint"* like *"A lung cancer patient receiving at least one treatment"* and *"intra-shape constraint* like *"A lung cancer patient should have exactly one gender"*. These constraints may be employed as medical protocols, indicating whether the patient validates or invalidates the protocols. This information may prove useful to both *KG Analysts* and *KG Consumers*.

Predictive Analysis over KGs: The generated LC KG is provided as input to the predictive model component where KG Analysts play a role in training predictive models over KGs. KG analysts examine the classification report produced by predictive models about its accuracy, precision, and recall. Predictive models with low precision or recall are more likely to produce inaccurate results. Consequently, analyzing the accuracy of predictive models is crucial before examining their outcomes. Moreover, KG Analysts can assess the quality of the data, as the approach generates decision trees with SHACL constraints, thereby enabling analysts to derive insights based on visualizations. The utilization of post-hoc interpretable tools, such as SHAP [4] and LIME [5], enables the comprehension of the rationale behind prediction model outcomes. Nonetheless, the component generates the SemDesLC KG which reflects the traced contextual knowledge about model traits and rationales, emphasizing the significance of interpretable insights.

Local and Global Explanations: Interpretable tools use prediction model outcomes to provide explanations that can be classified as local and global. As described in subsection 2.2, the requirements of the KG Consumers are collected to satisfy their needs. However, state-of-the-art interpretable tools are not sufficient, as they frequently produce visualizations that are difficult to comprehend by KG Consumers. The presentation of data in a manner that shields users from the underlying graph structure increases their willingness to interact with the data and accept predictive model outcomes.

Considering patient-centric analysis, *SemDesLC* generated a detailed representation of the outcome of LIME, incorporating information deemed essential by domain experts for evaluating a patient's characteristics or understanding why a specific patient was classified as having relapsed. *SemDesLC* provides the input characteristics of a relapsed patient by tracing into the input KG, which improves interpretability. Furthermore, domain experts receive a natural language explanation of the features that contributed to a patient's classification as *Relapse*. To illustrate, the predictive model identifies a patient as a relapse, as shown in Figure 1, *SemDesLC* provides additional characteristics that the patient is old and in stage IV, assisting oncologists to improve their decision-making processes by understanding the classification results based on the input characteristics.

In population-centric analysis *SemDesLC* outputs the features that contributed to the prediction of the model's outcome for the sub-population, in conjunction with the associated weights, to highlight the importance of the features' contributions to the model's outcome. Furthermore, the number of patients in the subpopulation that exhibit each of these features is shown. This analysis also allows domain experts to identify patients with comparable features who fall into the same category, as well as evaluate predictive model decisions based on existing characteristics. For instance, a subpopulation of young female patients diagnosed with cancer at stage IV and exhibiting biomarker ALK are typically classified as belonging to the *Relapse* class.

SemDesLC offers *Symbolic Learning*, which enables the capture of explicit patterns from the KGs and the generation of Horn Rules to derive insights from the KGs. For instance, a Horn Rule: `lc:stage(IIIC, X)` $\Rightarrow$ `lc:hasBio(PDL1, X)` states that if a patient

Table 1. Benchmark Statistics. #triples – Number of RDF triples in the KG, #entities – Number of distinct entities in the KG, #predicates – Number of distinct predicates in the KG. On the right, the Table shows lung cancer patient counts based on age, relapse, smoking habit, and gender.

Lung Cancer KG	Counts
#entities	15785
#predicates	27
#triples	71903
SemDesLC KG	**Counts**
#entities	39200
#predicates	158
#triples	228155

Features	Values	Patients Count
Age Category	Young	897
	Old	344
Relapse	Relapse	754
	No Relapse	526
Smoking Habit	Current Smoker	612
	Former Smoker	132
	Non Smoker	491
Gender	Female	870
	Male	372

is in stage *IIIC*, then it is most likely that the patient is positive for a biomarker *PDL1*. This implies that the head atom can be deduced if all body atoms are in KG. The current version of *SemDesLC* utilizes *AMIE* [16] as a rule-mining approach. *SemDesLc* offers natural language explanations of Horn rules to facilitate a more comprehensive understanding of the association observed in the KG by the KG Consumers. Furthermore, providing additional accessible insights would improve their understanding, enhancing trust and usability of knowledge-based approaches.

4. Evaluation

This section evaluates the performance of *SemDesLC* framework focusing on its efficacy and accuracy with state-of-the-art methodologies. Subsequently, an experimental investigation of patient- and population-centric analysis is conducted to determine the interpretability and understandability of prediction model outcomes from oncologists' perspectives. The purpose of this analysis is to evaluate the impact and effectiveness of integrating Semantic Web technologies with ML models, as well as to provide insights into their applicability and prospective benefits in the field of healthcare decision-making process. *SemDesLC* implementation is accessible on GitHub[2] for reproducibility.

Benchmarks. In the scope of experiments, we employ an anonymized synthetic lung cancer benchmark that comprises clinical data extracted from heterogeneous sources such as publications, clinical trials, and clinical records representing patients diagnosed with lung cancer. The benchmark includes 1242 patients with different medical characteristics. Table 1 represent the statistics of the LC KG and a subset of patient counts with features and their values documented in benchmark KG. Each patient is uniquely identified (*a.k.a, EHR*) with characteristics like a smoking habit (*e.g., Current Smoker*), demographic information, cancer mutation (*e.g., EGFR*), cancer stage (*e.g., IVB*). The benchmark also includes knowledge about drug-treatment assessment with certain START and END dates. Additionally, a score (*a.k.a, performance status*) is assigned to each patient, describing the patient's ability to perform daily physical activities. The benchmark comprises primarily young aged lung cancer patients, with 754 experiencing a

[2] https://github.com/SDM-TIB/SemDesLC

relapse and 526 completing follow-up without relapse; experts recommended the features to train predictive models for analyzing relapse. Further, the team of KG builders, analysts, and consumers evaluated together the medical guidelines for LC treatments and defined a simplified SHACL schema with 6 shapes representing 15 constraints.

Baselines. We created baselines to evaluate pipeline findings, such as KG creation, data quality checks, federated query processing, predictive models, and interpretability. We employ Morph-KGC [10], RMLMapper [12], and SDM-RDFizer [11] to create KGs. SHACL2SPARQL [17,15], Shaclex [14], and TravSHACL [13] are utilized to validate data on specific medical protocols expressed as a SHACL schema. Random Forest and Decision Trees models are trained and assessed for prediction tasks. Additionally, SHAP [4] and LIME [5] frameworks offer both local and global explanations to understand ML model decisions. ANAPSID [18], DeTrusty [19], and FedX [20] are evaluated for their ability to retrieve knowledge from a federation of KGs.

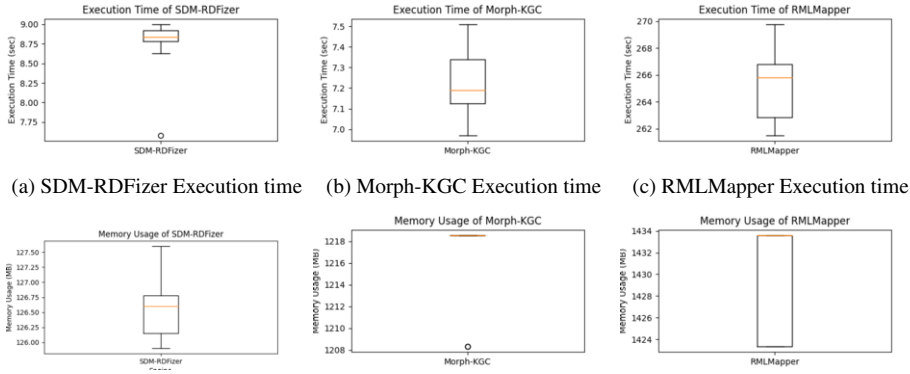
Experimental Environment. The experiments are performed in a dockerized environment, i.e., all the engines and data sources are executed in their respective Docker containers. The experiments are executed on an Ubuntu 16.04.6 LTS 64-bit machine with two Intel® Xeon® Platinum 8160 2.10 GHz CPUs, and 755 GiB DDR4 RAM. KGs are served using *Virtuoso 7.20.3238*. Each instance of Virtuoso is set up to use up to 16 GiB of memory. *MySQL 8.0.19* is utilized as a relational database to store synthetic data.

Settings for KG Creation. SDM-RDFizer v4.7.3.4, Morph-KGC v2.7.0, and RMLMapper v6.0.0 are evaluated over synthetic data. The experiments are run ten times. The execution time and memory usage are reported and compared in box plots.

Settings for KG Validation. SHACL2SPARQL, shaclex, and Trav-SHACL are evaluated over the defined SHACL schema. The experiments are run ten times. We report the average execution time and standard deviation per validation engine.

Settings for Federated Query Evaluation. The efficacy of ANAPSID, DeTrusty, and FedX is studied based on ten queries. Each query is executed ten times with each federated query engine. The average execution time and standard deviation are reported.

Settings for Predictive Models. We utilize an ensemble learning classifier, i.e., Random Forest (RF), to train our predictive models over the benchmark with optimized hyperparameters (e.g., maximum depth of the tree is 6) obtained from AutoML[3]. In *SemDesLC*, the predictive model uses *5-fold* cross-validation (CV) technique [21], and after each fold; over 20 relevant features were traced from the trained predictive model. Further, these relevant features are utilized to train a decision tree classification model, to improve the interpretability of the trained random forest model decisions. *SemDesLC* divides the training and test data sets into 70% training and 30% test sets, a common approach in ML. We assess the performance of the predictive models in terms of evaluation metrics such as Recall, Precision, F1-score, and Support. Recall depicts the ratio of counts of correctly predicted patients in *Relapse* class to total patient count with the target class *Relapse* in the benchmark. Precision is the ratio of accurately predicted patients in *Relapse* class to those projected to have class *Relapse*. The same evaluation parameters are used to categorize lung cancer patients as having *No Relapse*.



(a) SDM-RDFizer Execution time (b) Morph-KGC Execution time (c) RMLMapper Execution time

(d) SDM-RDFizer Memory Usage (e) Morph-KGC Memory Usage (f) RMLMapper Memory Usage

Figure 4. Results of KG Creation Engines. SDM-RDFizer and Morph-KGC can generate the KG in less than ten seconds. Morph-KGC has the lowest execution time. SDM-RDFizer has the lowest memory usage, with a difference in order of magnitude compared to the other engines. RMLMapper presents the highest execution time and memory usage, with a difference in execution time of an order of magnitude.

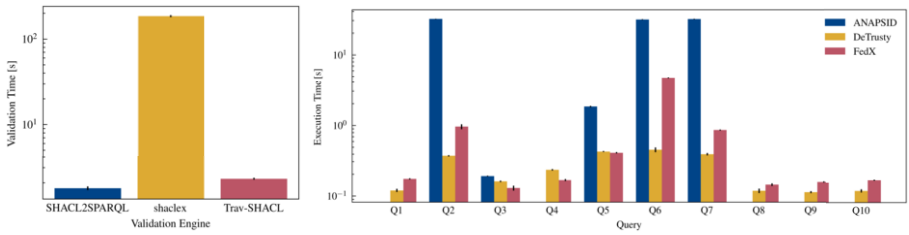
4.1. Results.

Effectiveness in KG Creation. The results of measuring the execution time and memory usage of KG creation engines SDM-RDFizer, Morph-KGC, and RMLMapper can be seen in Figure 4. SDM-RDFizer and Morph-KGC outperform RMLMapper by one order of magnitude regarding execution time. SDM-RDFizer and Morph-KGC can generate the KG in less than ten seconds, with Morph-KGC being slightly faster than SDM-RDFizer. Reason why RMLMapper takes longer to generate the KG is because it does not have an efficient method of removing duplicates and executing join. SDM-RDFizer outperforms Morph-KGC and RMLMapper by one order of magnitude regarding memory usage. For Morph-KGC, this high memory usage can be attributed to the fact that Morph-KGC uses the Python library `Pandas` for loading and preprocessing data. Unfortunately, it is well-known that this library consumes a great deal of memory. Since, SDM-RDFizer presents the lowest memory usage and a competitive execution time with Morph-KGC, SDM-RDFizer is chosen as the KG creation engine in *SemDesLC*.

Efficacy in KG Validation. The results of studying the performance of the SHACL validators SHACL2SPARQL, shaclex, and Trav-SHACL are reported in Figure 5a. SHACL2SPARQL and Trav-SHACL outperform shaclex by two orders of magnitude. Due to performance differences in Java and Python, SHACL2SPARQL is slightly faster than Trav-SHACL. Figuera et al. [13] demonstrate this fact with a Python implementation of the SHACL2SPARQL approach. Since the *SemDesLC* framework is implemented in Python Trav-SHACL has a richer feature set compared with SHACL2SPARQL. It was determined that Trav-SHACL is the optimal choice for the role of SHACL validator.

Efficacy in Federated Query Evaluation. Figure 5b shows the execution times for the three federated query engines ANAPSID, DeTrusty, and FedX for the ten queries from the benchmark. DeTrusty and FedX are capable of executing all queries and also outperform ANAPSID. Due to limited feature support, ANAPSID fails to execute some of the

[3] https://www.automl.org/



(a) KG Validation (b) Federated Query Evaluation

Figure 5. Result of KG Validation and Federated Query Evaluation. In Figure 5a, SHACL2SPARQL and Trav-SHACL perform similarly, with SHACL2SPARQL being slightly faster; both outperform shaclex by two orders of magnitude. In Figure 5b, DeTrusty and FedX are capable of executing all queries. DeTrusty outperforms all engines in seven out of ten queries. FedX is slightly faster than DeTrusty in three of ten queries.

queries, i.e., the queries using the VALUES clause or aggregates. DeTrusty outperforms FedX in seven of the ten queries. For the other three queries, FedX is slightly faster than DeTrusty. For query Q5, FedX is just a little faster than DeTrusty. However, FedX produces too many results for that query, not correctly applying the DISTINCT modifier. The answer provided by FedX includes each answer thrice. Due to the good overall performance and rich feature set, DeTrusty is integrated into *SemDesLC*.

Effectiveness in Predictive Model Evaluation and Interpretable Insights. We semantically documented and traced the trained predictive model as described in section 3 to generate the SemDesLC KG. The model behavior, local and global explanations are represented as RDF factual statements in the SemDesLC KG. Table 1 depicts the statistics of the SemDesLC KG. Executing SPARQL queries on the SemDesLC KG reveals the results for a specific patient or sub-population, additionally, predictive model characteristics like input features, classification reports, CV folds, hyperparameters, and prediction probabilities for each target class. In our predictive task of *Relapse* occurrence, the RF model showcased good performance across *5-fold* cross-validation, demonstrating the robustness and generalization of the relapse classification problem. Table 2 shows minimal variance in Precision, Recall, and F1-scores across multiple folds, indicating the model's stability and reliability across subsets of data. Moreover, the precision score ranges from 0.83 to 0.87, exhibiting the random forest model's ability to reduce false positives and accurately classify patients into target classes. Similarly, the recall metric showcases values ranging from 0.71 to 0.75, indicating the model's reliability in capturing the most relevant instances across multiple folds and reducing false negatives. Nevertheless, the *SemDesLC* pipeline traced the relevant features from the RF model. Further, the list of features is utilized to train the Decision Tree model to classify a lung cancer patient in classes - *Relapse* and *No Relapse*. Table 2 represents the classification report, demonstrating the strong performance across both *Relapse* and *No Relapse* classes, including additional metrics such as Support, Macro, and Weighted average. However, the Decision Tree model is trained using a balanced class distribution, with 519 lung cancer patients per target class. Thus, ensures balanced performance and prevents over-fitting. For the *Relapse* class, the precision score is 0.86, revealing that 86% of lung cancer patients experience relapse. The recall score for *Relapse* is 0.60, suggesting that the model correctly predicted 60% among all patients having a relapse in the data. F1-score for *Relapse* and *No Relapse* is 0.71 and 0.78 respectively, reflecting a harmonious balance between precision and recall. *SemDesLC* pipeline utilizes LIME and SHAP to provide a

Table 2. Evaluation Results. Relapse prediction task for patients with lung cancer. The table on the left shows 5-fold cross-validation (CV) results for a random forest model, including Precision, Recall, and F1-score. The table on the right displays a decision tree-generated classification report for *Relapse* and *No Relapse*, including Precision, Recall, F1-score, and Support. Support indicates the number of true instances for each target class.

Fold no.	Precision	Recall	F1-score
1	0.83	0.75	0.79
2	0.86	0.75	0.80
3	0.86	0.73	0.79
4	0.85	0.71	0.77
5	0.87	0.72	0.79

Decision Tree	Precision	Recall	F1-score	Support
Relapse	0.86	0.60	0.71	519
No Relapse	0.69	0.89	0.78	519
macro avg	0.78	0.75	0.74	1038
weighted avg	0.78	0.74	0.74	1038

global perspective by quantifying the influence of each feature across the subsets of data. LIME generates valuable insights by interpreting model predictions on a local level. Understanding the granularity of insights is crucial, especially in scenarios where specific predictions have significant implications, such as *"What is the likelihood of relapse for a young female in cancer stage IIIB receiving intravenous chemotherapy?"*. KG analysts revealed that features, such as *smoking habit* and *treatment type* exerted considerable influence on the model's outcomes. Moreover, leveraging SHAP values allows KG consumers to prioritize feature selection and KG analysts can optimize model performance for better interpretability on a global level. Additionally, symbolic learning generates Horn rules, i.e., patterns that are used for statistical analysis, guiding KG consumers with interpretability and natural language explanations. *SemDesLC* was able to mine 557 horn rules over the synthetic data. Additionally, the pipeline generates decision trees, feature importance, and SHACL validation decision tree plots, assisting and providing more comprehensive interpretability of the models' outcomes for KG consumers. Supplementary material includes *SemDesLC* implementation, statistical queries, decision tree plots, and generated KGs.

5. Related Work

In healthcare, the fusion of predictive models with KGs has gained tremendous attention, offering a promising research direction for various applications such as clinical decision-making systems, drug-drug interaction, and patient diagnosis with personalized treatment. Thus, the relevant works to our research falls under two categories:

KGs and Predictive Models in Medical domain. Predictive models [3,22] and Knowledge extraction [2,23] methodologies have been widely used to solve the prediction problem in the healthcare domain. Yang et al. [24] investigate the use of machine learning techniques such as decision trees and deep neural networks to analyze how clinical status and demographics influence the survivability of a patient with early-stage cancer. One of the most similar studies to ours investigates the use of predictive models such as random forests for tailored healthcare applications. By modeling patient-level and patient-episode health records [25], the authors create ensemble-based predictive models for diagnosing dementia and offering individualized therapies based on LIME [5] interpretations. They explain how their approach might help clinicians with dementia detection and therapy recommendations. Chandak et al. [26] propose the notion of PrimeKG, which integrates structured clinical concepts from heterogeneous data sources, detailing 17,080 diseases and their relationships reflecting biological processes, experimental

medications, and protein perturbations. They present a methodology that uses PrimeKG to predict patient drug-disease outcomes and provides treatment suggestions with textual descriptions. The authors [27] offer a methodology for creating and improving healthcare knowledge graphs using rule-based systems, entity-linking techniques, and machine learning algorithms. Despite advances in understanding multiple prognostic features, there is no clear consensus on how and which of these features should be integrated for relapse prediction. Various approaches [6,28] propose several guidelines and challenges from the perspective of end consumers. The aforementioned work illustrates the gaps and limitations of comprehending and evaluating predictive model forecasts. For instance, consumers such as oncologists, continue to lack effective frameworks for estimating a patient's likelihood of relapse in the early stage of treatment and translating back to the original characteristics of a lung cancer patient.

Semantic Web Technologies in Medical domain. In the context of Semantic Web (SW) technologies, Ristoski et al. [29] surveyed the potential for linking and integrating data from multiple sources, enabling effective data mining and knowledge discovery. Additionally, the deductive system (DS) proposed in [30] identifies drug-drug interactions caused by combining multiple drugs. In their work, the approach focuses on the application of KG-based machine learning methods for drug discovery. The authors use graph neural networks to predict molecular properties and discover possible therapeutic candidates, achieving breakthrough performance. Moreover, technologies such as ontologies play a crucial role in defining background knowledge and metadata related to the application domain. A semantic-based approach such as Knowledge4COVID-19 [27] analyzes drug-drug interactions via extracting entities and relations related to COVID-19 from Drugbank. Later, the Knowledge4COVID-19 KG is utilized to perform downstream tasks such as predicting interactions, treatment recommendations for curing the COVID-19 virus, and services to visualize the impact of a treatment drug. In [2], authors propose De4LungCancer, a health data ecosystem that utilizes controlled vocabularies and ontologies for knowledge management and analytics to describe the medical history of lung cancer patients. Furthermore, the data ecosystem utilizes the RML mapping engine [11] to build KGs from heterogeneous data sources with mapping assertions.

SHACL technologies can be used to validate data over KGs for quality assessment. An efficient SHACL validation engine [13] shows the best performance in planning and executing SHACL shape schema to determine whether entities (i.e., patients) from KGs comply with specific medical protocols. Moreover, neuro-symbolic approaches have shown significant achievement with enhanced performance and explainability of predictive models in the biomedical domain. Rivas et al. [31] demonstrate the fusion of numerical and symbolic learning for the prediction problem of treatment effectiveness over lung cancer KG. Further the relevant works [32,33] close to ours, investigate the impact of documenting and tracing predictive models on lung cancer prediction to address the issues of interpretability of decision-support systems. The authors use a federated query engine, DeTrusty [19], to extract the medical characteristics of a lung cancer patient and their insights into ML model predictions. The proposed interpretable framework facilitates oncologists to improve their understanding of the model's outcomes and recommends patients with early treatment procedures. Thus, the interpretability of the ML model is essential for consumer's trust in technologies. Similarly, *SemDesLC* resorts to SW technologies by semantically describing predictive models to ensure trust and allow consumers to interpret the model decisions for the prediction problem of LC relapse.

6. Discussion

SemDesLC shows a diverse range of analyses for various KG users. Furthermore, it was also demonstrated that identifying KG users' requirements improved the performance of *SemDesLC*. Figure 4 and Figure 5a show that efficiently creating and validating KGs improves predictive model performance and enables us to answer the **RQ1)** and **RQ2)**. Figure 5b displays the influence of the federated KGs, offering insights into how the instance was specified in the input KG, as well as the prediction models allowing to answer **RQ3)**. The predictive model evaluation answers **RQ4)**, highlighting the model's effectiveness in accurately categorizing lung cancer patients into target classes. Responding to **RQ5)**, KG consumers indicated high satisfaction with the interpretability of the LC use case and found *SemDesLC* to be informative in encouraging trust and traceability of the ML model decisions. Understanding the most important features provides actionable insights for knowledge engineering and model refinement. Further, extending the use of these predictions beyond classification problems offers new opportunities. Survival analysis can predict patients' long-term prognosis, enabling personalized treatment options. Moreover, leveraging the model's predictive capabilities for link prediction problems may entail investigating complex relationships between mutations, and patient medical history, ultimately improving our understanding of relapse response. These efforts show great potential for broadening the discipline of healthcare.

7. Conclusions and Future Work

In contrast to the state-of-the-art approaches presented in this work, *SemDesLC* proposes an independent methodology. As previously stated, it is crucial to differentiate between distinct KG users and to consider the evaluations in which these different users are interested. The current work presents a series of empirical evaluations for each KG user. *SemDesLC* has been made available to different KG users for evaluation and feedback. In particular, two KG Builders participated in the process intending to assess the performance of the tools and software required by the KG Builder. It was demonstrated that it is critical to determine the efficacy of KG creation and validation of the input data to ensure the optimal functioning of the approach, which includes the use of scalable technologies. Furthermore, two KG Analysts evaluated the predictive models' performance on the synthetic data to identify interpretable insights. Finally, three KG consumers provided feedback on their understanding of the predictive model results and the natural language explanations offered by SemDesLC for patient- and population-centric analyses that meet the requirements of KG Consumers. However, in future work, additional surveys will be distributed to various KG users. The goal is to obtain more insight into the proposed methodology and assess its efficacy and performance in different domains.

Acknowledgements

This work has been partially supported by TrustKG- Transforming Data in Trustable Insights with grant P99/2020 and the EraMed project P4-LUCAT (GA No. 53000015).

References

[1] Wood R, Taylor-Stokes G. Cost burden associated with advanced non-small cell lung cancer in Europe and influence of disease stage. BMC Cancer. 2019 03;19.

[2] Aisopos F, Jozashoori S, Niazmand E, Purohit D, Rivas A, Sakor A, et al. Knowledge graphs for enhancing transparency in health data ecosystems. Semantic Web. 2023;14(5):943-76.

[3] Janik A, Torrente M, Costabello L, Calvo V, Walsh B, Camps C, et al. Machine Learning–Assisted Recurrence Prediction for Patients With Early-Stage Non–Small-Cell Lung Cancer. JCO Clinical Cancer Informatics. 2023;(7):e2200062. Available from: https://doi.org/10.1200/CCI.22.00062.

[4] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. .

[5] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. .

[6] Li HX, Appleby G, Brumar CD, Chang R, Suh A. Knowledge Graphs in Practice: Characterizing their Users, Challenges, and Visualization Opportunities. IEEE Trans Vis Comput Graph. 2024;30(1):584-94.

[7] Lenzerini M. Managing Data through the Lens of an Ontology. AI Mag. 2018;39(2):65-74. Available from: https://doi.org/10.1609/aimag.v39i2.2802.

[8] Dimou A, Sande MV, Colpaert P, Verborgh R, Mannens E, de Walle RV. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: Bizer C, Heath T, Auer S, Berners-Lee T, editors. Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014. vol. 1184 of CEUR Workshop Proceedings. CEUR-WS.org; 2014. .

[9] Knublauch H, Kontokostas D. Shapes Constraint Language (SHACL); 2017. W3C Recommendation. Available from: https://www.w3.org/TR/2017/REC-shacl-20170720/.

[10] Arenas-Guerrero J, Chaves-Fraga D, Toledo J, Pérez MS, Corcho O. Morph-KGC: Scalable knowledge graph materialization with mapping partitions. Semantic Web. 2022.

[11] Iglesias E, Jozashoori S, Chaves-Fraga D, Collarana D, Vidal ME. SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs. In: CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York, USA: ACM; 2020. .

[12] Dimou A, De Nies T, Verborgh R, Mannens E, Van de Walle R. Automated Metadata Generation for Linked Data Generation and Publishing Workflows. In: Workshop on Linked Data on the Web; 2016. .

[13] Figuera M, Rohde PD, Vidal ME. Trav-SHACL: Efficiently Validating Networks of SHACL Constraints. In: The Web Conference. New York, NY, USA: ACM; 2021. .

[14] Labra Gayo JE, Prud'hommeaux E, Roman B, Cebrían T, Berezovskyi A. Shaclex v0.2.2; 2022. GitHub. Available from: https://github.com/weso/shaclex.

[15] Corman J, Florenzano F, Reutter JL, Savković O. SHACL2SPARQL: Validating a SPARQL Endpoint against Recursive SHACL Constraints. In: Proceedings of the ISWC 2019 Satellite Tracks. Aachen, Germany: CEUR-WS.org; 2019. p. 165-8. Available from: https://ceur-ws.org/Vol-2456/paper43.pdf.

[16] Lajus J, Galárraga L, Suchanek F. Fast and Exact Rule Mining with AMIE 3. In: The Semantic Web; 2020. .

[17] Corman J, Florenzano F, Reutter JL, Savković O. Validating SHACL Constraints over a SPARQL Endpoint. In: The Semantic Web – ISWC 2019. Cham: Springer; 2019. p. 145-63.

[18] Acosta M, Vidal ME, Lampo T, Castillo J, Ruckhaus E. ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints. In: The Semantic Web – ISWC 2011. Berlin, Heidelberg: Springer; 2011. p. 18-34.

[19] Rohde PD, Bechara M, Avellino. DeTrusty v0.15.6; 2024.

[20] Schwarte A, Haase P, Hose K, Schenkel R, Schmidt M. FedX: Optimization Techniques for Federated Query Processing on Linked Data. In: The Semantic Web – ISWC 2011. Berlin, Heidelberg: Springer; 2011. p. 601-16.

[21] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010 Jan;4(none). Available from: http://dx.doi.org/10.1214/09-SS054.

[22] Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, et al. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Scientific Reports. 2017;7. Available from: https://api.semanticscholar.org/CorpusID:26489344.

[23] Vidal ME, Niazmand E, Rohde PD, Iglesias E, Sakor A. In: Challenges for Healthcare Data Analytics Over Knowledge Graphs; 2023. .

[24] Yang Y, Xu L, Sun L, Zhang P, Farid SS. Machine learning application in personalised lung cancer recurrence and survivability prediction. Computational and Structural Biotechnology Journal. 2022;20:1811-20.

[25] Vyas A, Aisopos F, Vidal ME, Garrard P, Paliouras G. Identifying the presence and severity of dementia by applying interpretable machine learning techniques on structured clinical records. BMC Medical Informatics Decis Mak. 2022;22(1):271. Available from: https://doi.org/10.1186/s12911-022-02004-3.

[26] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. bioRxiv. 2022. Available from: https://www.biorxiv.org/content/early/2022/05/01/2022.05.01.489928.

[27] Sakor A, Jozashoori S, Niazmand E, Rivas A, Bougiatiotis K, Aisopos F, et al. Knowledge4COVID-19: A semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analyzing treatments' toxicities. Journal of Web Semantics. 2023;75:100760.

[28] Suh A, Appleby G, Anderson EW, Finelli L, Chang R, Cashman D. Are Metrics Enough? Guidelines for Communicating and Visualizing Predictive Models to Subject Matter Experts. IEEE Transactions on Visualization and Computer Graphics. 2023:1-16.

[29] Ristoski P, Paulheim H. Semantic Web in data mining and knowledge discovery: A comprehensive survey. J Web Semant. 2016.

[30] Rivas A, Vidal ME. Capturing Knowledge about Drug-Drug Interactions to Enhance Treatment Effectiveness. In: Proceedings of the 11th Knowledge Capture Conference. K-CAP '21. New York, NY, USA: Association for Computing Machinery; 2021. .

[31] Rivas A, Collarana D, Torrente M, Vidal ME. A neuro-symbolic system over knowledge graphs for link prediction. Semantic Web Journal Special Issue on Neuro-Symbolic Artificial Intelligence and the Semantic Web. 2023:1-25.

[32] Chudasama Y, Purohit D, Rohde PD, Gercke J, Vidal ME. InterpretME: A Tool for Interpretations of Machine Learning Models Over Knowledge Graphs. Semantic Web Journal Special Issue on Tools & Systems. 2024.

[33] Chudasama Y, Purohit D, Rohde PD, Vidal ME. Enhancing Interpretability of Machine Learning Models over Knowledge Graphs. In: Keshan N, Neumaier S, Gentile AL, Vahdati S, editors. Proceedings of the Posters and Demo Track of the 19th International Conference on Semantic Systems co-located with 19th International Conference on Semantic Systems (SEMANTiCS 2023), Leipzig, Germany, September 20 to 22, 2023. vol. 3526 of CEUR Workshop Proceedings. CEUR-WS.org; 2023. Available from: https://ceur-ws.org/Vol-3526/paper-05.pdf.