

Quantifying Domain-Application Knowledge Mismatch in Ontology-Guided Machine Learning

Pawel Bielski^a, Lena Witterauf, Sönke Jendral^b, Ralf Mikut^c and Jakob Bach^d

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

{pawel.bielski, ralf.mikut, jakob.bach}@kit.edu, lena.emma77@gmail.com, jendral@kth.se

Keywords: Ontology Quality Evaluation, Knowledge-Guided Machine Learning, Application Ontology.

Abstract: In this work, we study the critical issue of knowledge mismatch in ontology-guided machine learning (OGML), specifically between domain ontologies and application ontologies. Such mismatches may arise when OGML uses ontological knowledge that was originally created for different purposes. Even if ontological knowledge improves the overall OGML performance, mismatches can lead to reduced performance on specific data subsets compared to machine-learning models without ontological knowledge. We propose a framework to quantify this mismatch and identify the specific parts of the ontology that contribute to it. To demonstrate the framework's effectiveness, we apply it to two common OGML application areas: image classification and patient health prediction. Our findings reveal that domain-application mismatches are widespread across various OGML approaches, machine-learning model architectures, datasets, and prediction tasks, and can impact up to 40% of unique domain concepts in the datasets. We also explore the potential root causes of these mismatches and discuss strategies to address them.


1 INTRODUCTION


Motivation. Ontologies formally represent domain knowledge in a structured way. They use a set of concepts and their relationships that is understandable by both humans and machines (Min et al., 2017; Lourdasamy and John, 2018; Wilson et al., 2022). They are increasingly important for intelligent, ontology-informed applications in fields such as knowledge management, data integration, decision support, reasoning, and machine learning (McDaniel and Storey, 2020; Min et al., 2017).


One application area that is gaining interest in the machine-learning community is ontology-guided machine learning (OGML). OGML is a subfield of knowledge-guided machine learning (KGML) (von Rueden et al., 2023; Willard et al., 2023) that systematically incorporates ontological domain knowledge into machine-learning models. OGML aims to improve prediction performance, especially for rarely represented data objects, reduce training data requirements, and generate more interpretable results.


OGML methods have shown significant success in fields like computer vision (image classification, segmentation, and retrieval) and medical data processing (including text classification and patient health prediction (Choi et al., 2017; Ma et al., 2018; Yin et al., 2019)), where rich ontological background knowledge is abundant (Min and Wojtusiak, 2012).

OGML methods generally outperform ontology-uninformed machine-learning methods on average (Dhall et al., 2020; Karthik et al., 2021; Silla and Freitas, 2011). However, the underlying ontological domain knowledge may not always have the optimal structure for a particular machine-learning task, which may negatively impact particular subsets of the data. In other words, OGML methods can suffer from a mismatch between domain and application-specific knowledge, which typically arises because ontological domain knowledge is created for different purposes than the specific OGML task. Existing literature on OGML methods often ignores this type of low-quality domain knowledge and assumes that ontologies only positively impact predictions. It is crucial to understand how such mismatches manifest, how big their impact is, and how to address them. The first step in tackling this challenge is to develop a method for identifying and quantifying this mismatch in OGML approaches.

^a  <https://orcid.org/0009-0005-3242-9113>

^b  <https://orcid.org/0009-0000-0070-9595>

^c  <https://orcid.org/0000-0001-9100-5496>

^d  <https://orcid.org/0000-0003-0301-2798>

Approach. In this work, we study the important issue of knowledge mismatch between domain-specific and application-specific ontologies in OGML, which has been mostly overlooked in the literature. As a result, both the existing theory-based and empirical methods to evaluate ontology quality (Hlomani and Stacey, 2014; Wilson et al., 2022) are inadequate for detecting this mismatch. To address this gap, we propose a new OGML-aware evaluation framework based on the task-based framework (Porzel and Malaka, 2004). Because the original framework was not designed for OGML, we adapt it to account for OGML-specific aspects, such as separation of the ontology and data, different task and ground truth definitions, and the stochastic nature of machine-learning algorithms. We argue that domain-application knowledge mismatch manifests as *harmful* domain knowledge, negatively impacting the prediction performance of OGML methods. Our framework identifies such harmful parts of the ontology for a specific task by comparing the performance of the OGML method with an ontology-uninformed method.

To demonstrate the effectiveness of our framework, we apply it to two common OGML application areas: image classification and patient health prediction. For image classification, we quantify the mismatch across three biological image datasets, using the Hierarchical Semantic Embedding OGML approach by (Chen et al., 2018). For patient health prediction, we quantify the mismatch across three prediction tasks within one medical dataset, using the GRAM (Choi et al., 2017) OGML approach. Our findings reveal that such mismatches are widespread across various OGML approaches, machine-learning model architectures, datasets, and prediction tasks. We also explore the potential root causes of these mismatches based on the harmful parts of the ontology identified by our framework. Furthermore, we discuss strategies to address these issues, demonstrating that our methodology shows promise as a generalizable approach for ontology quality assessment, enabling the identification of various ontological issues.

Contributions. To summarize, our contributions are as follows:

1. We study the important but relatively overlooked problem of domain-application knowledge mismatch in ontology-guided machine learning (OGML).
2. We introduce a quality evaluation framework to quantify this mismatch and identify ontology parts that negatively impact the task performance.
3. We apply our framework in two common OGML application areas to demonstrate how to detect, interpret, and address such mismatches.
4. We provide the code and experimental results¹.

Paper Outline. Section 2 discusses background and related work. Section 3 introduces our approach. Section 4 reports on experiments from two OGML case studies. Section 5 concludes.

2 RELATED WORK

In this section, we review related work regarding approaches for ontology quality evaluation in Section 2.1, OGML in general in Section 2.2, and the issue of low-quality domain knowledge in the context of machine learning in Section 2.3.

2.1 Ontology Quality Evaluation

Creating and maintaining ontologies is a highly subjective, labor-intensive process. This process is prone to errors, as there is no standard method for creating ontologies (Capellades, 1999; Brewster, 2002; Duque-Ramos et al., 2011). Additionally, ontologies are only approximations of domain knowledge, and multiple valid ontologies can exist to represent the same knowledge (Hlomani and Stacey, 2014; McDaniel and Storey, 2020). Thus, evaluating the quality of ontologies is essential for the broader adoption of ontology-informed applications (Mc Gurk et al., 2017). This process ensures that developed ontologies are useful for specific tasks or domains and helps select the most suitable ontology for the given application (Duque-Ramos et al., 2011). Evaluation methods can significantly reduce the human effort needed to create and maintain ontologies. In particular, they can guide the construction process and enable the reuse of existing ontologies instead of building them from scratch (Capellades, 1999; Beydoun et al., 2011; McDaniel and Storey, 2020). Despite many proposed approaches to ontology quality evaluation, no universal solution exists as they address different quality aspects. (McDaniel and Storey, 2020).

Existing methods can be grouped into two broad categories: deductive (metrics-based) and inductive (empirical) (Burton-Jones et al., 2005; Hlomani and Stacey, 2014).

Deductive evaluation methods to evaluate ontology quality are theory-based metrics that quantify whether an ontology is correct according to structural

¹<https://doi.org/10.35097/zv8zqqd6ezm02vk>

properties and description-logic axioms (Hlomani and Stacey, 2014; Wilson et al., 2022). Often inspired by software-engineering research on software quality, these methods use heuristic quality criteria to identify syntactic, semantic, and structural problems that are independent of the application (McDaniel and Storey, 2020). However, because these deductive methods rely on various subjective interpretations of ontology quality, none of them has become standard (Brewster et al., 2004). Additionally, verifying whether an ontology meets specific formal criteria does not guarantee optimal performance for a particular purpose (Gómez-Pérez, 1999; McDaniel and Storey, 2020).

Inductive evaluation methods assess ontology quality by empirically testing its fitness (i.e., usefulness for a specific application) rather than its syntax, semantics, or structure (Burton-Jones et al., 2005; Wilson et al., 2022). Fitness can be quantified in terms of application fitness, which evaluates performance on a specific task, or domain fitness, which assesses performance across multiple tasks within a domain. Ontology fitness is typically quantified for the entire ontology (Porzel and Malaka, 2004; Clarke et al., 2013), but it can also be quantified for specific parts of the ontology, which can help identify improvement potentials. This process requires linking specific parts of the ontology to application performance, which is not trivial and thus often skipped in practice (Pittet and Barthélémy, 2015).

(Porzel and Malaka, 2004), (Brank et al., 2005), (Burton-Jones et al., 2005) and (Ohta et al., 2011) argue that inductive evaluation, particularly task-based evaluation, offers an objective measure of ontology quality by directly evaluating the ontology's ability to solve practical problems. Despite this, research in this area is limited. Apart from the original paper introducing task-based ontology quality evaluation (Porzel and Malaka, 2004) and a few adaptations (Clarke et al., 2013; Pittet and Barthélémy, 2015), there is little research on assessing ontology quality based on its utility for specific applications. Both (Ohta et al., 2011) and (Wilson et al., 2022) have highlighted the need for more research in this area. Specifically, evaluating ontology quality for OGML, which we address in this work, has not been previously explored.

Recent research in confident learning and data-centric AI (Wang et al., 2018; Northcutt et al., 2021; Rigoni et al., 2023) shows that analyzing predictions from traditional machine-learning methods can uncover and address ontological issues in image label hierarchies, enhancing data quality and prediction performance. Our work follows a similar direction but focuses specifically on ontology-guided machine learning.

2.2 Ontology-Guided Machine Learning

Ontology-guided machine learning (OGML) is a sub-field of knowledge-guided machine learning (KGML) that leverages structured ontological domain knowledge to enhance machine-learning models. This is usually accomplished with custom loss functions (Zeng et al., 2017; Ju et al., 2024), ontology-aware embeddings (Vendrov et al., 2016; Nickel and Kiela, 2017; Chen et al., 2018; Dhall et al., 2020; Bertinetto et al., 2020), or adapted model architectures (Brust and Denzler, 2019a). OGML methods have demonstrated significant success in domains rich in ontological background knowledge, such as medical data processing or computer vision.

In healthcare, abundant medical domain knowledge has accumulated through years of medical research, hospital administration, billing, and documentation of medical procedures. This knowledge is often organized into ontologies that group medical codes into semantically meaningful categories using parent-child relationships, e.g., the ICD-9 hierarchy of symptoms and diseases (see Section 4.2). OGML approaches leverage these ontologies for various automated medical data processing tasks, such as patient health prediction (Choi et al., 2017; Yin et al., 2019; Ma et al., 2019) or medical text classification (Arbabi et al., 2019). These methods have been shown to improve prediction performance, especially for rare diseases that are often insufficiently represented in data.

In computer vision, domain knowledge is often structured as taxonomies of labels, reflecting the hierarchical nature of many real-world datasets, such as those in biology (Silla and Freitas, 2011; Rezende et al., 2022). Even non-hierarchical datasets can be enriched with knowledge from literature or general domain-independent ontologies (Chen et al., 2018; Brust and Denzler, 2019a). OGML approaches in computer vision have been applied to tasks such as image classification (Deng et al., 2014; Goo et al., 2016; Marino et al., 2017; Chen et al., 2018; Brust and Denzler, 2019a; Bertinetto et al., 2020; Ju et al., 2024), image retrieval (Vendrov et al., 2016; Barz and Denzler, 2019).

OGML approaches typically use readily available generic or domain ontologies (Burton-Jones et al., 2005) rather than task-specific application ontologies. While research often reports that ontological domain knowledge improves average prediction performance compared to models without it (Dhall et al., 2020; Karthik et al., 2021; Silla and Freitas, 2011), there is limited recognition that not all data subsets may benefit equally in the context of a specific task.

2.3 Low-Quality Domain Knowledge

In the broader context of knowledge-guided machine learning (KGML), both (Mitchell, 1997) and (Yu, 2007) recognize that domain knowledge can be imperfect due to difficulties in its collection, definition, and representation. (Yu, 2007) also notes that domain knowledge is highly context-dependent, meaning its usefulness can vary across different tasks. The authors emphasize the importance of considering the negative impact of imperfect domain knowledge when applying KGML. (Mitchell, 1997) argues that even imperfect knowledge can be beneficial as long as the machine-learning algorithm tolerates some level of error. While some recent KGML publications explicitly design or evaluate their approaches with this in mind and quantify the impact of imperfect domain knowledge (Bielski et al., 2024; Brust et al., 2021; Deng et al., 2014), most existing KGML publications do not explicitly address this issue.

In the specific context of OGML, no studies similar to ours on the problem of domain-application knowledge mismatch have been conducted. However, several related observations have been made regarding the low quality of domain knowledge. For example, (Brust and Denzler, 2019b) investigated the discrepancy between visual and semantic similarity in OGML for image classification. They observed that the overall prediction performance may decrease in some situations compared to knowledge-uninformed baselines. (Choi et al., 2017) showed that fully randomized ontological domain knowledge can decrease the overall prediction performance in healthcare OGML applications. The above studies considered the overall negative effect on average prediction performance and did not analyze the prediction performance on subsets of the data. They also did not consider identifying specific parts of ontological domain knowledge that might have caused the decrease in the prediction performance. (Deng et al., 2014) and (Brust et al., 2021) investigated the related problem of maximizing the utility of imprecise ontologies in OGML but did not focus on identifying potential quality issues within the ontologies themselves.

The most similar work to ours is (Marino et al., 2017), where the authors analyzed the prediction performance of their OGML approach for image classification across different data subsets. They found that their OGML approach performed worse than the baselines on certain subsets of the data, attributing this to missing relationships in the ontology. While their study provided valuable insights, our work builds upon this by offering a more comprehensive framework that not only broadens the perspective

on the underlying issues but also systematically quantifies and addresses them.

3 APPROACH

Section 3.1 outlines our adaptation of the original task-based evaluation framework to OGML. Next, Section 3.2 introduces the concept of domain-application mismatch and explains how to quantify it.

3.1 Adapting Task-Based Ontology Quality Evaluation to OGML

The original task-based evaluation framework for ontologies, proposed by (Porzel and Malaka, 2004), assessed quality within ontology-informed applications by comparing task results against human-generated gold standards. While effective in its context, this framework requires significant adaptation to OGML.

Separation of Ontology and Data. In the original framework, the task is performed directly on the ontology since data and ontology are the same. In contrast, OGML distinguishes between ontology and data. The ontology is used to improve the prediction performance of a machine-learning task on the data.

Task Definition and Ground Truth Data. In the original framework, tasks were specifically designed to identify ontology issues, with the ground truth defined by humans, leading to potential subjectivity errors. In OGML, however, the machine-learning process defines the task, and the ground truth is derived directly from the data. This ensures that the evaluation is more objective and less prone to errors. However, it also necessitates linking the performance of the OGML task to specific parts of the ontology, which can be achieved by using refinement metrics, as described in Section 3.2.

Stochastic Nature of ML Algorithms. OGML introduces stochastic elements inherent in machine learning, including retraining machine-learning models multiple times with different seed values, varying train-test splits, or varying model sizes. These factors must be considered to ensure the objectivity of results.

3.2 Quantifying Domain-Application Knowledge Mismatch in OGML

In OGML, a *domain ontology* represents a broad field of knowledge, and an *application ontology* is tai-

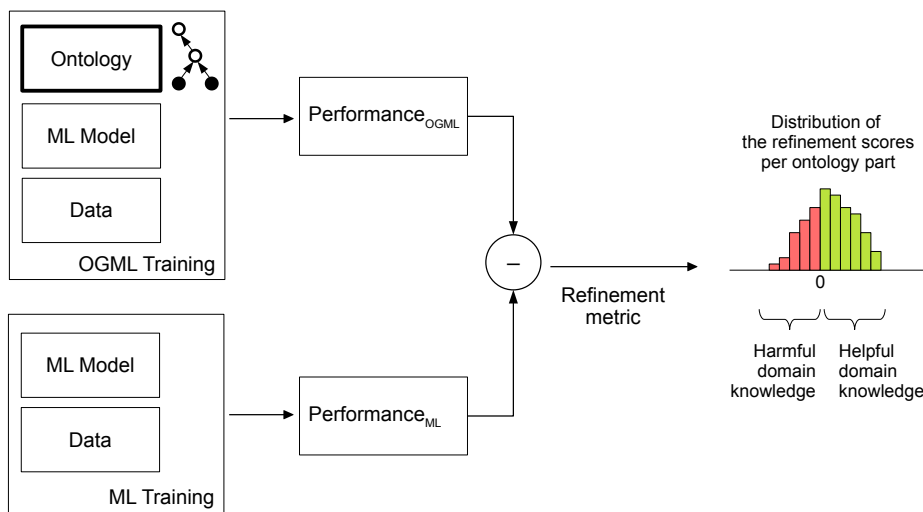


Figure 1: Proposed framework to quantify the domain-application mismatch.

lored to a specific task. A mismatch between domain knowledge and application knowledge occurs when the provided ontological knowledge for a domain is not optimally structured for the specific machine-learning task at hand. This mismatch can exist even if the domain knowledge is free of mistakes and thus has high quality from the domain perspective.

Because most OGML approaches leverage domain ontologies instead of application ontologies, it is often impossible to quantify the mismatch between domain and application knowledge directly by comparing OGML models with both types of ontologies. We argue that such a mismatch manifests itself through the existence of harmful parts of the ontological domain knowledge, which may exist independent from the fitness of the entire ontology. That is why we propose to approximate domain-application mismatch by measuring *harmful* domain knowledge.

Definition 1. A particular part of domain knowledge is *harmful (helpful)* for a particular supervised machine-learning task if it negatively (positively) affects the prediction performance compared to a knowledge-uninformed baseline. The machine-learning task comprises the datasets for training and testing, prediction target, prediction model, and evaluation metric.

Measuring harmful and helpful domain knowledge requires measuring the fitness of specific ontology parts, which can be achieved with our proposed framework (Figure 1). An ontology part is the subset of nodes and edges of the ontology that is semantically connected with a specific domain concept (e.g., unique class label) from the dataset. A *refinement metric*, which is a task-specific heuristic, links these ontology parts to application performance. We

demonstrate examples of such refinement metrics in Section 4. In general, a refinement metric assigns a score to each part of the domain knowledge: 0 indicates no impact on the prediction performance, a score greater than 0 indicates a positive impact, and a score less than 0 indicates a negative impact. The distribution of the refinement scores can be plotted, as shown in Figure 1. The parts of domain knowledge with scores below zero are harmful.

As the representation of domain knowledge may be of considerable size, e.g., for an ontology with many nodes and edges, only some parts may be harmful or helpful. Furthermore, the above definition is closely tied to *one* particular machine-learning task. In particular, some knowledge may be harmful for one task but not another. We propose to leverage our framework to quantify mismatch as follows:

Definition 2. The *knowledge mismatch* is the ratio between the number of harmful domain-knowledge parts (Definition 1) and the total number of domain-knowledge parts.

For example, if the OGML approach outperforms the ontology-uninformed baseline, but 25% of relevant domain concepts perform worse than the baseline, we consider there to be a 25% mismatch. Note that different domain concepts may occur with different frequencies in the data. For example, if 25% of the domain concepts are harmful, 10% of the data for the machine-learning task may be affected if the affected concepts are relatively infrequent or 50% of the data if they are relatively frequent.

Table 1: Comparison of the three datasets for image classification (negative refinement scores in red).

Dataset	Butterflies	Birds	VegFru
Acc. Baseline [%]	84.78	85.23	86.31
Acc. OGML [%]	85.82	88.09	88.77
Improvement [pp]	1.04	2.86	2.46
Mismatch [%]	25.50	16.50	40.21
– Data affected [%]	22.00	17.00	40.67

4 CASE STUDIES

In this section, we demonstrate how to quantify a domain-application mismatch in two common OGML application areas: image classification (Section 4.1) and patient health prediction (Section 4.2). For each area, we explain how to apply our proposed framework to quantify the mismatch and present the results. Additionally, for patient health prediction, we use our framework to identify and qualitatively describe ontological issues arising from mismatches.

4.1 Use Case 1: Ontology-Guided Image Classification

Scenario. In the first use case, we demonstrate how to quantify domain-application mismatch in computer vision. We apply our framework to the OGML approach for image classification proposed by (Chen et al., 2018). This approach incorporates structured information about parent-child relationships between image categories and subcategories (i.e., a label taxonomy) into a deep learning model. The OGML model employs a Hierarchical Semantic Embedding framework to maintain consistency in classification across different taxonomy levels.

Experimental Setup. We employ the same setup as the original paper by (Chen et al., 2018). Specifically, we use the three pre-trained machine-learning models made publicly available by the authors and apply them to the corresponding hierarchical image datasets: *Butterflies*, *Birds*, and *VegFru* (Vegetables and Fruits). These datasets contain 200 unique classes for Butterflies and Birds, and 292 classes for VegFru, each organized into a taxonomy with four levels for Butterflies and Birds, and two levels for VegFru.

Refinement Metric. To quantify the domain-application mismatch of ontological domain knowledge, we define the refinement metric as the per-class performance improvements, while the original paper assesses overall performance improvements. Our approach allows for a more detailed analysis of how the ontology impacts the model’s performance, offering insights into which specific classes benefit from the ontological knowledge and which do not. We quantify prediction performance with top-1 accuracy, as in the original paper, on the test set. We calculate the mismatch as the percentage of classes that show a decrease in prediction performance compared to the baseline. Additionally, since classes may vary in the number of examples, we also report the proportion of the test data affected by these classes.

Results. As Table 1 shows, the OGML method demonstrates overall improvements compared to the knowledge-uninformed baseline across all three hierarchical datasets. However, a substantial number of classes does not benefit from the ontological domain knowledge (highlighted in red on the distribution plots of refinement scores). Since the classes are relatively balanced in all datasets, we observe a similar percentage of data affected by the mismatch.

4.2 Use Case 2: Ontology-Guided Sequential Health Prediction

Scenario. In the second use case, we demonstrate how to quantify domain-application mismatch in medical data processing. We apply our proposed framework to an OGML approach for sequential patient health prediction, proposed by (Choi et al., 2017). This approach incorporates structured information about the hierarchical relationships of varying

Table 2: Comparison of model sizes for two variants of risk prediction (negative refinement scores in red).

Architecture	Heart Disease Prediction		Diabetes Prediction	
	Small	Large	Small	Large
Acc. Baseline [%]	71.36	79.66	70.32	85.72
Acc. OGML [%]	78.98	81.32	89.03	90.33
Improvement [pp]	7.62	1.66	18.71	4.61
Mismatch [%]	15.07	20.07	4.73	11.47
– Data affected [%]	40.79	78.76	11.46	48.90

depth between medical codes of symptoms and diseases defined by the ICD-9 classification system². It processes sequences of medical codes with a graph-based attention mechanism (GRAM) to generate semantic embeddings, considering not only individual medical codes but also their hierarchical ancestors.

Experimental Setup. We employ a similar setup as the original paper by (Choi et al., 2017), utilizing the publicly available MIMIC-III healthcare dataset (Johnson et al., 2016). Different from the computer vision use case, we train the OGML model ourselves, varying the experiments across three prediction tasks: two *risk prediction* tasks – for heart diseases and diabetes – and one *next-visit prediction* task. In the dataset, each patient visit is represented by medical codes corresponding to the diagnoses and symptoms identified during that visit. For risk prediction, the goal is to predict whether the patient’s next visit will include a diagnosis of heart disease or diabetes, based on their previous visits. For next-visit prediction, the goal is to predict all the diagnoses and symptoms recorded during the patient’s next visit, based on their previous visits.

To address the stochastic nature of machine-learning methods, we conduct five experiments for each combination of task and model size. In each experiment, we used random train-test splits – 80-20 for risk prediction and 90-10 for next-visit prediction. We then report the average performance on the test sets. Each model comprises an embedding layer, an RNN layer, and a final dense layer. The dense layer uses a sigmoid activation function for risk prediction and a softmax activation function for next-visit prediction. The larger model has an attention dimension

of 100, an RNN dimension of 200, and an embedding dimension of 300, and it is trained with a batch size of 128 for 100 epochs with early stopping. The smaller model has an attention dimension of 16, an RNN dimension of 32, and an embedding dimension of 16, and it is trained with a batch size of 32 for 50 epochs with early stopping. For further details on the experimental setup, please refer to the experimental code provided along with this paper.

Refinement Metric. To quantify domain-application mismatch in both tasks, we define the refinement metric as the per-code performance improvements based on all input-output pairs where the input sequences (patient visits) include that particular medical code. For risk prediction, the improvement is measured using binary accuracy. For next-visit prediction, we define two variants of the refinement metric. The first one is accuracy-based, similar to the risk prediction task, but using top-20 accuracy to measure the improvement, in line with the evaluation metric from the original paper. The second one measures the average rank improvement between the baseline and OGML approach by comparing the rank differences for medical codes found in the ground-truth data for the respective patient visits. In both tasks, input sequences may be counted multiple times for different medical codes. As in the computer vision use case, we calculate the mismatch as the percentage of classes (codes) that show a performance decrease, and we also report the proportion of data affected by these classes (codes).

Given the smaller dataset sizes, varying train-test splits, and a larger number of unique classes compared to the computer vision use case, we report the mismatch on the entire dataset instead of just the test set. To handle the high number of unique medical

²<http://www.icd9data.com/2015/Volume1/default.htm>

Table 3: Comparison of model sizes and refinement metrics for next-visit prediction (negative refinement scores in red).

Architecture	Next-Visit Prediction			
	Small		Large	
Acc. Baseline [%]	55.10		66.19	
Acc. OGML [%]	73.87		71.32	
Improvement [pp]	18.77		5.13	
Refinement Metric	Accuracy-based	Ranking-based	Accuracy-based	Ranking-based
Mismatch [%]	29.60	21.47	29.71	25.56
– Data affected [%]	83.50	59.61	82.67	60.78

codes (1,823 for next-visit prediction and 2,426 for risk prediction), we filter out codes that appear fewer than three times in the dataset. This results in 38.2% of unique codes being filtered out for risk prediction and 2.1% for next-visit prediction.

Results. The results, summarized in Tables 2 and 3, reveal distinct patterns across the various models and tasks. First, domain-application knowledge mismatch is evident across all model sizes, prediction tasks, and refinement metrics within the dataset. However, the degree of mismatch varies, with the mismatch for diabetes risk prediction being two to three times smaller than that for heart disease risk prediction using the same model sizes. This variation is also reflected in the differences in the distribution of refinement scores shown in the bottom row of the table. Additionally, models used for diabetes risk prediction benefit more from domain knowledge than those used for heart disease risk prediction. Further, ontological domain knowledge tends to improve the performance of smaller models more than of larger models. For risk prediction, smaller OGML models perform nearly as well as their larger counterparts, while in next-visit prediction, smaller OGML models even outperform the larger ones. Additionally, smaller models generally exhibit less domain-application mismatch, i.e., they benefit more from domain knowledge than larger models. Lastly, we observe a much higher percentage of data affected by mismatches compared to the computer vision use case, likely due to the presence of multiple medical codes in each input sequence.

4.2.1 Identifying Ontological Issues

When examining the top ten medical categories with the lowest accuracy-based refinement scores for next-visit prediction, we found several potential ontological issues (with the ICD-9 hierarchy) that could decrease the prediction performance of OGML.

Similar Concepts in Different Ontological Paths.

This issue arises when related ontological categories are placed under different paths and lack a common semantic ancestor. As a result, the OGML approach treats these categories as semantically independent, which can confuse the machine-learning model and negatively impact the prediction performance for these categories. For example, five of the ten medical codes with the lowest refinement scores are related to drug-related symptoms or diseases. These five categories fall into three paths in the ontology, without a shared common ancestor:

- *970.8: Poisoning by other specified central nervous system stimulants*, falls under the ontological category *Injury and Poisoning 800-999*
- *E950.0: Suicide and self-inflicted poisoning by analgesics, antipyretics, and antirheumatics* and *E950.4: Suicide and self-inflicted poisoning by other specified drugs and medicinal substances* both fall under the ontological category *Supplementary Classification of External Causes of Injury and Poisoning E000-E999*.
- *304.23 Cocaine dependence, in remission*, and *304.21 Cocaine dependence, continuous* both fall under the ontological category *Mental Disorders 290-319*.

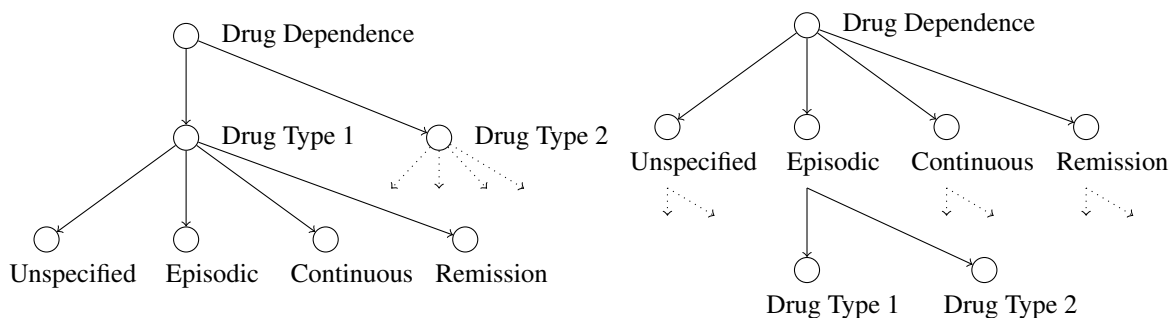


Figure 2: Suboptimal parent order (left) and potential improved one (right) for categories related to drug dependence.

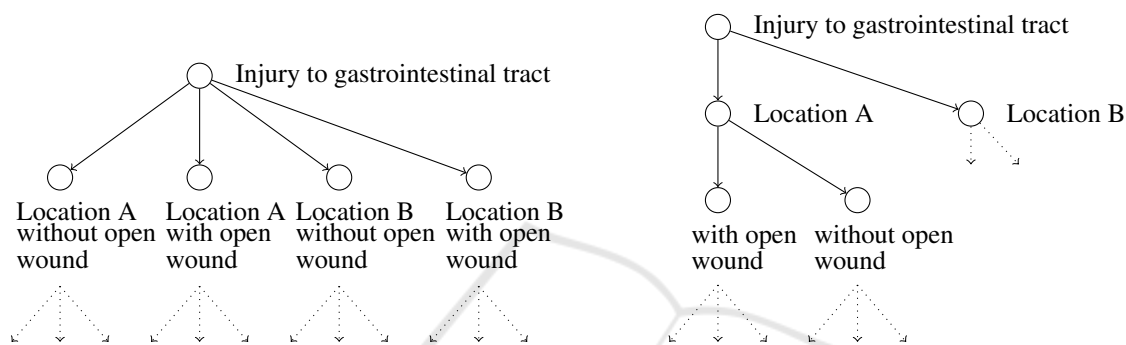


Figure 3: Suboptimal parent order (left) and potential improved one (right) for categories related to gastrointestinal tract.

Irrelevant Categorization. This issue arises when the categorization focuses on aspects that may be less relevant to the machine-learning task. For instance, consider the following codes:

- *E956 Suicide and self-inflicted injury by cutting and piercing instrument*
- *E950.0 Suicide and self-inflicted poisoning by analgesics, antipyretics, and antirheumatics*
- *E950.4 Suicide and self-inflicted poisoning by other specified drugs and medicinal substances*

While these codes categorize different types of injuries (cutting, poisoning by drugs, etc.), they are all grouped under the broader category of *Suicide and Self-Inflicted Injury (E950-E959)*. This grouping does not account for other causes of such injuries. For next-visit prediction, the focus on whether an injury is self-inflicted might be less relevant than the specific type of injury. A more effective approach could be categorizing these codes based on the type of injury (e.g., cutting, poisoning) rather than its origin, as this may be more relevant to the prediction task.

Inaccurate or Overly Broad Categories. This issue arises when a category is not specific enough or is overly broad. Categories that include terms like ‘unspecified’ or ‘other’, or have such terms in their parent

categories, are especially susceptible to this problem. For example, the code *957.1 Injury to other specified nerve(s)* is classified under the broader category *957 Injury to other and unspecified nerves*. This broad classification can include various, potentially unrelated medical codes within the same category, which may confuse the machine-learning model.

Suboptimal Ordering of Parent Categories. This issue arises when parent categories are organized to prioritize one aspect over another, which may not be optimal for the specific task. For example, the ICD-9 ontology initially classifies drug dependence by drug type and then by dependence type (Figure 2, left). This structure leads the OGML approach to treat continuous use of different drugs as unrelated and continuous versus episodic use of the same drug as more similar. Reordering to classify by dependence type first (Figure 2, right) could better capture the nuances of drug use. Similarly, Figure 3 shows that organizing wound information by location and type at the same level may not be ideal. Depending on the task, it might be more effective to classify injuries first by location and then by type, or vice versa, or even to provide both ordering paths.

5 CONCLUSIONS

In this work, we addressed the critical and often overlooked issue of domain-application knowledge mismatch in ontology-guided machine learning (OGML). We developed an OGML-aware framework to quantify these mismatches and identify harmful ontology parts, which negatively affect prediction performance. Our framework offers a practical and generalizable methodology for assessing ontology quality in OGML contexts. Thus, it improves the integration of ontological knowledge into machine-learning models, leading to more effective and reliable use of ontologies. Our case studies in image classification and patient health prediction revealed that mismatches are widespread across datasets, OGML approaches, and machine-learning architectures. This highlights the importance of aligning domain ontologies with specific application requirements in OGML contexts. Future research could refine our framework and explore its applicability across various domains and OGML methods. For example, one could apply our framework to multiple tasks in a single domain to evaluate an ontology's domain fitness. Another promising direction is automatically repairing ontologies for a given OGML task, i.e., removing harmful domain knowledge, restructuring the ontologies accordingly, and re-training the OGML model.

ACKNOWLEDGEMENTS

This research has been partially funded by the German Federal Ministry of Education and Research (BMBF) under grant 01IS17042 as part of the Software Campus project DomainML.

REFERENCES

- Arbabi, A., Adams, D. R., Fidler, S., and Brudno, M. (2019). Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. *JMIR Med. Inf.*, 7(2).
- Barz, B. and Denzler, J. (2019). Hierarchy-Based Image Embeddings for Semantic Image Retrieval. In *Proc. WACV*, pages 638–647.
- Bertinetto, L., Mueller, R., Tertikas, K., Samangoeei, S., and Lord, N. A. (2020). Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks. In *Proc. CVPR*, pages 12503–12512.
- Beydoun, G., Lopez-Lorca, A. A., García-Sánchez, F., and Martínez-Béjar, R. (2011). How do we measure and improve the quality of a hierarchical ontology? *J. Syst. Software*, 84(12):2363–2373.
- Bielski, P., Eismont, A., Bach, J., Leiser, F., Kottonau, D., and Böhm, K. (2024). Knowledge-guided learning of temporal dynamics and its application to gas turbines. In *Proc. e-Energy*, page 279–290.
- Brank, J., Grobelnik, M., and Mladenic, D. (2005). A survey of ontology evaluation techniques. In *Proc. SiKDD*, pages 166–170.
- Brewster, C. (2002). Techniques for automated taxonomy building: Towards ontologies for knowledge management. In *Proc. Annu. CLUK Res. Colloq.*
- Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data driven ontology evaluation. In *Proc. LREC*, pages 641–644.
- Brust, C.-A., Barz, B., and Denzler, J. (2021). Making every label count: Handling semantic imprecision by integrating domain knowledge. In *Proc. ICPR*, pages 6866–6873.
- Brust, C.-A. and Denzler, J. (2019a). Integrating domain knowledge: using hierarchies to improve deep classifiers. In *Proc. ACPR*, pages 3–16.
- Brust, C.-A. and Denzler, J. (2019b). Not just a matter of semantics: The relationship between visual and semantic similarity. In *Proc. DAGM GCPR*, pages 414–427.
- Burton-Jones, A., Storey, V. C., Sugumaran, V., and Ahluwalia, P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data Knowl. Eng.*, 55(1):84–102.
- Capellades, M. A. (1999). Assessment of reusability of ontologies: a practical example. In *Proc. AAAI Workshop Ontol. Manage.*, pages 74–79.
- Chen, T., Wu, W., Gao, Y., Dong, L., Luo, X., and Lin, L. (2018). Fine-Grained Representation Learning and Recognition by Exploiting Hierarchical Semantic Embedding. In *Proc. ACM MM*, pages 2023–2031.
- Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. (2017). GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *Proc. KDD*, pages 787–795.
- Clarke, E. L., Loguercio, S., Good, B. M., and Su, A. I. (2013). A task-based approach for Gene Ontology evaluation. *J. Biomed. Semant.*, 4.
- Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. (2014). Large-Scale Object Classification Using Label Relation Graphs. In *Proc. ECCV*, pages 48–64.
- Dhall, A., Makarova, A., Ganea, O., Pavllo, D., Greeff, M., and Krause, A. (2020). Hierarchical Image Classification using Entailment Cone Embeddings. In *Proc. CVPRW*, pages 3649–3658.
- Duque-Ramos, A., Fernández-Breis, J. T., Stevens, R., and Aussenac-Gilles, N. (2011). OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies. *J. Res. Pract. Inf. Technol.*, 43(2):159–176.
- Goo, W., Kim, J., Kim, G., and Hwang, S. J. (2016). Taxonomy-Regularized Semantic Deep Convolutional Neural Networks. In *Proc. ECCV*, pages 86–101.
- Gómez-Pérez, A. (1999). Evaluation of taxonomic knowledge in ontologies and knowledge bases. Technical report, University of Calgary, Alberta, Canada.

- Hlomani, H. and Stacey, D. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semant. Web J.*, 1(5):1–11.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1).
- Ju, L., Yu, Z., Wang, L., Zhao, X., Wang, X., Bonnington, P., and Ge, Z. (2024). Hierarchical Knowledge Guided Learning for Real-World Retinal Disease Recognition. *IEEE Trans. Med. Imaging*, 43(1):335–350.
- Karthik, S., Prabhu, A., Dokania, P. K., and Gandhi, V. (2021). No Cost Likelihood Manipulation at Test Time for Making Better Mistakes in Deep Networks. arXiv:2104.00795 [cs].
- Lourdusamy, R. and John, A. (2018). A review on metrics for ontology evaluation. In *Proc. ICISC*, pages 1415–1421.
- Ma, F., Wang, Y., Xiao, H., Yuan, Y., Chitta, R., Zhou, J., and Gao, J. (2019). Incorporating medical code descriptions for diagnosis prediction in healthcare. *BMC Med. Inf. Decis. Making*, 19(6).
- Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., and Gao, J. (2018). KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare. In *Proc. CIKM*, pages 743–752.
- Marino, K., Salakhutdinov, R., and Gupta, A. (2017). The More You Know: Using Knowledge Graphs for Image Classification. arXiv:1612.04844 [cs].
- Mc Gurk, S., Abela, C., and Debattista, J. (2017). Towards ontology quality assessment. In *Proc. LDQ2017*, pages 94–106.
- McDaniel, M. and Storey, V. C. (2020). Evaluating Domain Ontologies: Clarification, Classification, and Challenges. *ACM Comput. Surv.*, 52(4).
- Min, H., Mobahi, H., Irvin, K., Avramovic, S., and Wojtusiak, J. (2017). Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. *J. Biomed. Semant.*, 8(1).
- Min, H. and Wojtusiak, J. (2012). Clinical data analysis using ontology-guided rule learning. In *Proc. MIXHS*, pages 17–22.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill New York.
- Nickel, M. and Kiela, D. (2017). Poincaré Embeddings for Learning Hierarchical Representations. In *Proc. NIPS*.
- Northcutt, C., Jiang, L., and Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.*, 70:1373–1411.
- Ohta, M., Kozaki, K., and Mizoguchi, R. (2011). A Quality Assurance Framework for Ontology Construction and Refinement. In *Proc. AWIC*, pages 207–216.
- Pittet, P. and Barthélémy, J. (2015). Exploiting Users’ Feedbacks - Towards a Task-based Evaluation of Application Ontologies Throughout Their Lifecycle:. In *Proc. IC3K*, pages 263–268.
- Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In *Proc. ECAI Workshop Ontol. Learn. Popul.*
- Rezende, P. M., Xavier, J. S., Ascher, D. B., Fernandes, G. R., and Pires, D. E. V. (2022). Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings Bioinf.*, 23(4).
- Rigoni, D., Elliott, D., and Frank, S. (2023). Cleaner Categories Improve Object Detection and Visual-Textual Grounding. In *Proc. SCIA*, pages 412–442.
- Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discovery*, 22(1-2):31–72.
- Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2016). Order-Embeddings of Images and Language. arXiv:1511.06361 [cs].
- von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Gieselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., and Schuecker, J. (2023). Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Trans. Knowl. Data Eng.*, 35(1):614–633.
- Wang, H., Wang, H., and Xu, K. (2018). Categorizing concepts with basic level for vision-to-language. In *Proc. CVPR*, pages 4962–4970.
- Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. (2023). Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Comput. Surv.*, 55(4).
- Wilson, R. S. I., Goonetillake, J. S., Indika, W. A., and Ginige, A. (2022). A conceptual model for ontology quality assessment. *Semant. Web*, 14(6):1051–1097.
- Yin, C., Zhao, R., Qian, B., Lv, X., and Zhang, P. (2019). Domain Knowledge Guided Deep Learning with Electronic Health Records. In *Proc. ICDM*, pages 738–747.
- Yu, T. (2007). *Incorporating Prior Domain Knowledge into Inductive Machine Learning: its Implementation in Contemporary Capital Markets*. PhD thesis, University of Technology Sydney, Australia.
- Zeng, C., Zhou, W., Li, T., Shwartz, L., and Grabarnik, G. Y. (2017). Knowledge Guided Hierarchical Multi-Label Classification Over Ticket Data. *IEEE Trans. Netw. Serv. Manage.*, 14(2):246–260.