information

MDPI

Article

Enhancing Rule-Based Explanations via Cognitive Bias Towards Semantic Relevance

Parisa Mahya [1,*] and Johannes Fürnkranz [1,2]

[1] Institute for Application-Oriented Knowledge Processing (FAW), Johannes Kepler University, 4040 Linz, Austria
[2] LIT Artificial Intelligence Lab, Johannes Kepler University, 4040 Linz, Austria
* Correspondence: parisa.mahya@faw.jku.at

Abstract

As interest in explainable artificial intelligence (XAI) continues to grow, a critical gap remains between the explanations generated by models and their interpretability by human users, particularly when cognitive biases and semantic relevance are not adequately addressed. This paper introduces CORIFEE-Rel, a novel human-centered meta-XAI method aimed at bridging this gap by producing rule-based explanations that are both interpretable and semantically aligned with the target domain concept. CORIFEE-Rel synthesizes outputs from a diverse pool of interpretable models and employs a knowledge graph-driven heuristic that combines semantic relevance with traditional rule learning metrics. This approach ensures that the resulting explanations are deeply tied to core domain concepts while maintaining the clarity and structure needed for human understanding. Empirical evaluations conducted across multiple datasets demonstrate that CORIFEE-Rel achieves higher semantic relevance than random forest and JRIP rule-based explanations without notable compromises in predictive accuracy. The results highlight the ability of CORIFEE-Rel to generate rule-based explanations that are semantically related to the target concepts while maintaining predictive performance.

Keywords: explainable artificial intelligence; cognitive bias; semantic relevance; interpretability; rule learning

1. Introduction

As machine learning models are increasingly integrated into various applications and domains such as healthcare, finance, and autonomous systems, concerns about their interpretability and trustworthiness have grown significantly. These concerns have led to the emergence of a new area, explainable AI (XAI), a field dedicated to making the decision-making process of the underlying black-box model transparent and interpretable to humans [1–3]. However, existing XAI techniques usually focus on generating interpretable explanations while maintaining predictive performance, but fail to explicitly address cognitive aspects of human understandability, which can reduce the perceived user trust and acceptance [4–6].

This limitation becomes particularly important in high-stakes domains, where explanations must be not only accurate and interpretable but also related to the underlying domain concepts. For example, in healthcare, physicians may expect a medical explanation to emphasize clinical indicators of a disease, whereas a financial analyst may expect explanations to focus on economically relevant factors. Ensuring that explanations remain semantically

Academic Editor: Yu-Cheng Wang

Received: 17 August 2026
Revised: 16 September 2026
Accepted: 17 September 2026
Published: 19 September 2026

relevant to the main concept of a domain is an important step toward generating more useful and meaningful explanations.

In response to these limitations and to address these shortcomings, the XAI field is shifting toward human-centered XAI (HCXAI) [7,8]. HCXAI draws upon insights from cognitive science, psychology, and related fields to generate explanations that are not only accurate but also tailored to users' cognitive limitations, domain expertise, and the context of use, thereby improving human interpretability. In particular, it has been argued that cognitive biases influence how explanations are perceived and interpreted [9,10].

While existing approaches have incorporated cognitive aspects into explanation generation, they primarily focus on enhancing the interpretability and usability of explanations [11]. However, limited research has addressed the semantic alignment of explanations with the core topics of a given domain. Consequently, explanations may be interpretable without strongly aligned with the domain's central concepts. This highlights an important gap between explanation generation and semantic alignment.

To address this gap, we investigate semantic relevance as a cognitively motivated dimension of explainability, with the goal of generating explanations that are more semantically aligned with the main topic of the underlying knowledge base. To achieve this, we propose a meta-XAI method that takes multiple rule-based explanations as input and synthesizes a new explanation that exhibits high semantic relevance to the target concept. The method uses a knowledge graph to quantify and measure the relevance score of the explanation.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related works and studies, Section 3 describes semantic relevance concept and definitions, Section 4 outlines our research goals and terminologies, Section 5 introduces and explains CoRIfEE-Rel method, and Section 6 discusses the results.

2. Related Works

Since the fundamental goal of XAI is to enhance the users' understanding of black-box models, numerous studies have highlighted critical challenges and pitfalls encountered during the examination of user experiences with explanations. Research shows that users often find the generated explanations difficult to understand and use, with the reasoning process both time-consuming and distracting [12–14]. In response, XAI methods have increasingly turned toward improving the interpretability and accountability of explanations by incorporating aspects of cognitive science and psychology into the explanation process [15].

Recently, Rajabi and Etminani [16] presented a comprehensive review of knowledge graph-based XAI systems and identified four principal roles of knowledge graphs in explainability: feature extraction, relationship extraction, knowledge graph construction, and reasoning. Their findings highlight the importance of semantic knowledge and structured domain representation for generating human-understandable explanations. Similarly, recent studies have explored the use of knowledge graphs to support explainability through semantic integration and structured domain representations [17].

Beyond the scope of XAI, semantic relevance has been extensively studied in natural language processing (NLP) tasks such as summarization and question answering. For example, Wang et al. [18] introduce a deep neural network model to evaluate semantic relevance between questions and candidate answers on social media platforms, leading to improved answer ranking performance. Similarly, Ruotsalo and Hyvönen [19] propose a method for ranking large answer sets by assessing their semantic relevance to a query using on-tology-based representations and RDF languages. In another study, De Boni and

Manandhar [20] present an algorithm that computes semantic similarity to evaluate answer relevance with respect to a given question.

Natural-language explanation has also been explored in XAI to present explanatory information in a more human-readable form. Costa et al. [21] propose a character-level, attention-enhanced LSTM that generates personalized natural-language explanations of recommendations directly from user reviews. Colas et al. [22] present a knowledge-graph-based approach that generates fact-grounded natural-language explanations from item features rather than user reviews, jointly learning explanation generation and recommendation scoring. Breve et al. [23] propose a hybrid prompt-learning strategy to generate natural-language justifications explaining the security and privacy risks of if–then automation rules. These approaches demonstrate the role of natural-language generation in communicating explanatory information to users. In contrast, in our work, we focus on selecting and combining rule conditions according to the semantic relevance, while natural-language rendering is used only as a complementary presentation of the generated rules.

Further works include the Hybrid Co-Attention Network (HCAN) by Rao et al. [24], which integrates convolutional and LSTM-based encoders to combine semantic matching for NLP tasks. This approach aims to bridge the methodological divide between traditional IR approaches and semantic understanding.

In the domain of text summarization, You et al. [25] proposes a BERT-based method that incorporates topic information fusion and semantic relevance. Their approach extracts topic keywords and integrates them with source documents, enhancing summary readability through improved semantic alignment. Experimental results demonstrate improved readability of the generated summaries. Additionally, Ma et al. [26] explore both text summarization and text simplification, aiming to enhance the relevance between source texts and their simplified version. To achieve this, they propose a neural model, ensuring high semantic similarity between texts and their summaries.

Semantic relevance has also been explored through ontological frameworks. Rhee et al. [27] introduce a method for semantic relevance assessment that goes beyond structural constraints, enabling the comparison of conceptually diverse resources. Building on this, Rhee et al. [28] develop a graph-based algorithm to quantify semantic relevance between resources based on a graph structure and demonstrate its application through two implementation examples.

Rule-based explanations are among widely used forms of interpretable machine learning because of their transparency and readability. However, conventional rule learning algorithms are typically optimized for predictive performance rather than human understanding. As a result, generated rules may include features that are not semantically related, reducing their perceived plausibility and interpretability for end users [9,10]. These observations suggest that explanations can be evaluated with respect to cognitively aligned properties such as semantic relevance.

Several approaches aim to improve interpretability while preserving predictive behavior. GLocalX aggregates rule-based explanations into a compact global explanation [29], and Born-Again Tree Ensembles construct a single interpretable decision tree reproducing the behavior of an ensembles [30]. CoRIfEE-Rel differs in its objective: rather than primarily optimizing fidelity to a fixed source model, it uses a pool of rule-based explanations and explicitly incorporates knowledge-graph-based semantic relevance into explanation construction.

While recent XAI research has begun to consider semantic context, continuity, and similarity to produce human-understandable explanations [31–33], and knowledge graph-based approaches have been shown to enhance interpretability across multiple domains [16], the explicit integration of semantic relevance into explanation generation as an optimization

criterion remains underexplored. Our work addresses this gap by integrating knowledge graph-driven semantic relevance as a core criterion for constructing explanations in XAI.

3. Semantic Relevance

3.1. Semantic Relevance Definition

Semantic relevance refers to the degree to which information and concepts are meaningfully related to a given cognitive context [34]. It originates in cognitive psychology as a core mechanism for prioritizing information, where the human brain evaluates the conceptual importance of information through semantic memory. Early research into human memory and decision-making has shown that individuals do not treat all pieces of information equally; instead, the human cognitive system assesses contextual plausibility, approving information that aligns with existing mental frameworks or prior knowledge while filtering out irrelevant data [35,36].

According to Relevance theory, introduced by Sperber and Wilson [37], human cognition tends to maximize relevance with the least effort, where relevance is defined as the balance between cognitive effects and processing efforts. From this perspective, information is semantically relevant if it has a significant contribution to the user's mental model.

3.2. Semantic Relevance in Machine Learning

As artificial intelligence and machine learning applications and systems increasingly focus on human-like processing and decision-making, the notion of semantic relevance has also been adapted in these fields, particularly in NLP and information retrieval. Semantic relevance has been used in the form of model attention, feature selection, and similarity metrics. For instance, transformer-based architectures [38], as an attention mechanism, explicitly model the parts of the input that are most relevant for generating an output. Furthermore, feature-based methods such as SHAP [39] or saliency maps [40] provide explanations of the model's internal decision-making by specifying the inputs that mostly contributed to the model's prediction process.

In information retrieval, semantic relevance is utilized to rank items or documents according to their contextual alignment with the user query, which impacts usability and performance [41]. However, semantic relevance in machine learning can be data-dependent, i.e., models may associate inputs with outputs due to dataset biases or correlations that mimic semantic logic. This can cause problems in real-world applications because of the lack of alignment between statistical and human-understandable relevance.

3.3. Semantic Relevance in XAI

Semantic relevance in XAI also introduces a new challenge. People often interpret AI explanations through their cognitive expectations, rather than evaluating whether the explanations reflect the model's internal reasoning. This can be referred to as one aspect of cognitive biases [15,42].

This bias arises because humans tend to naturally favor explanations that are more aligned with familiar concepts or contexts. For example, in medical imaging, an explanation that highlights a visible tumor tends to be more acceptable for physicians because it uses terms relevant to the disease and domain knowledge, even though the model makes the decision based on irrelevant data, such as image borders [43]. Therefore, even accurate and valid explanations may be rejected by the targeted users if they fail to align with users' cognitive expectations [44].

Most current XAI techniques aim to provide explanations without fully considering human needs, contexts, or cognitive limitations [5]. These approaches and methods often prioritize performance and fidelity to the underlying black-box models.

To address these limitations, human-centered XAI emphasizes the importance of generating explanations that incorporate human-in-the-loop approaches, in which targeted users and their needs are considered during the explanation generation process [6]. This aims to support cognitively plausible explanations.

In this work, we focus on the semantic relevance dimension of human-centered XAI and introduce a novel approach to generate explanations that are aligned with the context without significantly sacrificing accuracy and transparency.

4. Problem Statement

We present our new method within the broader framework, Cognitively biased Rule-based Interpretations from Explanation Ensembles (CoRIfEE). CoRIfEE is a meta-XAI approach that takes a set of rule-based explanations as input and merges them into a single explanation that better matches the user's cognitive preferences. The strategy of taking multiple rule-based explanations is inspired by the *Rashomon effect* [45], which highlights that multiple models can perform equally well on the same data while relying on different feature sets, leading to distinct explanations. This phenomenon has also been observed in the context of interpretable explanations, showing that the existence of multiple equally valid explanations provides a principled foundation for generating personalized or cognitively aligned explanations [46].

Because different interpretable models yield different aspects of the data, our method identifies and extracts features and rules from a pool of diverse explanations that meet specific criteria. They are then combined into new explanations based on heuristic measurements beyond traditional factors like syntactic simplicity or fidelity to the original model. In this article, we introduce an instantiation of the CoRIfEE platform, named CoRIfEE-Rel that focuses on semantic relevance.

4.1. CoRIfEE-*Coh and* CoRIfEE-*Pref*

In addition to CoRIfEE-Rel, the CoRIfEE framework includes other instantiations that address distinct dimen-sions of human cognitive biases, each designed to generate more cognitively aligned explanations. Two notable variants are CoRIfEE-Coh, which focuses on semantic coherence, and CoRIfEE-Pref, which incorporates user preferences.

CoRIfEE-Coh is designed to generate semantically coherent explanations, i.e., rules in which all components are conceptually and semantically related and consistent. Like CoRIfEE-Rel, this variant operates on a pool of rule-based explanations and uses a defined heuristic that balances traditional rule learning with a coherence score derived from the knowledge graph. The resulting explanations are more coherent than the explanations from the interpretable models [47].

CoRIfEE-Pref introduces personalization into explanation generation by modeling user preferences. This method assumes that different users, such as domain experts in medicine versus patients, may prioritize different types of concepts and features when interpreting explanations. CoRIfEE-Pref models such preferences as weights over nodes in the knowledge graph and incorporates them into the heuristic used during rule generation. The generated explanations are more aligned with the user's background knowledge and preferences [48].

4.2. CoRIfEE-*Rel*

In this section, we introduce CoRIfEE-Rel, another instantiation of the CoRIfEE platform, developed to generate explanations that are highly relevant to the main concept of the given domain and dataset. This method operates on a pool of interpretable models and constructs new explanations by selecting and recombining subsets of conditions in the

interpretable model. The selection process prioritizes conditions that exhibit high semantic similarity to the domain's main topic, ensuring that the generated explanations maintain semantic relevance while maintaining predictive performance to the original models.

To illustrate the concept of semantic relevance in text, we present two examples in Table 1, both centered on the main concept of financial stability. Table 1a demonstrates high semantic relevance, as it directly involves financial actions and terminology, such as "saving money" and "managing debt", which are main topics within the financial domain. This framing maintains a clear and strong connection to the main topic. In contrast, Table 1b exhibits low semantic relevance. Although it also addresses financial stability, it employs terms like "education" and "family background", which are not directly related to finance. These terms are more commonly associated with the domains of social science and personal development.

Table 1. Examples of high and low semantic relevance to the main concept of financial stability.

(**a**) High semantic relevance.	(**b**) Low semantic relevance.
If a person regularly saves money and manages their debt effectively, then they are likely to achieve financial stability.	If a person has a higher level of education and comes from a supportive family background, then they are more likely to achieve financial stability.

In our work, CORIFEE-Rel aims to adopt a meta-XAI design so that semantic relevance can be applied to any rule-based explanations from different models without specifying and modifying each underlying algorithm. This allows the method to be applicable to diverse explanations produced by different interpretable models.

4.3. Terms and Notations

In this article, we use the notations as described in the following:

- A **rule r** is an if–then statement, consisting of a body (if-part) and a head (then-part). The body includes one or multiple conditions connected by conjunctions.
- **length** of a rule, denoted by $length(\mathbf{r})$, is the number of conditions in the body of the rule $\mathbf{r}$.
- **Rule sets**, presented as $\mathcal{R} = \{\mathbf{r_1}, \cdots, \mathbf{r_k}\}$, is a collection of rules.
- **Dataset** $\mathcal{D}$ consists of examples, each represented by a vector of attributes $\mathbf{A} = \{\mathbf{a_1}, \cdots, \mathbf{a_n}\}$. For each attribute $\mathbf{a_i}$ (where $i = 1, 2, \cdots, n$), the value associated with the j-th example in the dataset $\mathcal{D}$ is denoted as $\mathbf{v}_i^{(j)}$.
- A **Feature** f is defined as a Boolean condition, represented by a combination of an attribute $f.attr$ and its corresponding value $f.value$.

The method also utilizes a knowledge graph, formally defined as a directed labeled graph $\mathcal{KG} = (V, E)$, where V denotes the set of vertices (or nodes), and E represents the set of labeled edges connecting these vertices. In this work, the vertices V include nodes representing concepts, attributes, and attribute values.

Figure 1 shows an example of a knowledge graph in which the nodes in blue illustrate the attributes and the nodes in green represent attributes' values in the dataset.

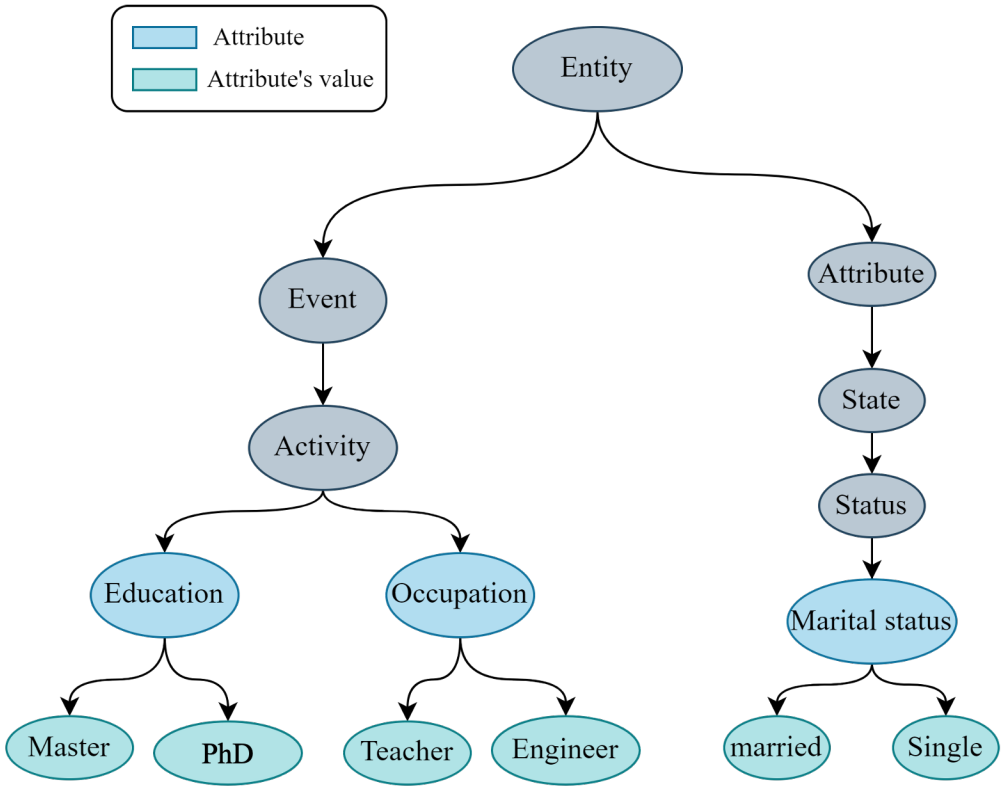


Figure 1. Fragment of a domain-specific knowledge graph constructed from the *Adult* dataset, where blue nodes represent attributes and green nodes denote attributes' values.

The method takes as inputs a pool of interpretable models, denoted by $\mathcal{I} = \{\mathcal{R}_i\}$, where each $\mathcal{R}_i$ is a ruleset, along with a dataset $\mathcal{D}$, and a focused knowledge graph $\mathcal{KG}$. Additionally, a scoring function $s(\mathcal{R})$ is defined to quantify the semantic relevance of a given ruleset $\mathcal{R}$. Our objective is to construct a new ruleset $\mathcal{R}'$ from the initial pool of interpretable models such that its semantic relevance is improved, i.e., $s(\mathcal{R}') > s(\mathcal{R}_i)$ for all $\mathcal{R}_i \in \mathcal{I}$.

5. Methodology

As discussed in previous sections, CoRIfEE-Rel focuses on semantic relevance by generating rule-based explanations that are highly relevant to the main topic or concept of the domain.

This section provides a detailed, step-by-step description of the main processes in the method.

To enhance understandability and transparency, we present an illustrative example in Figure 2 that outlines the main steps of the CoRIfEE-Rel method. The example begins with a dataset $\mathcal{D}$, which includes the attributes $\mathbf{A} = \{$"capital-gain","education","gender", "marital-status"$\}$. The interpretable model $\mathcal{I}$ contains two rules, $\mathbf{r_1}$ and $\mathbf{r_2}$, where one of the features is defined as $f =$ ("education","master"). The main steps of CoRIfEE-Rel, as illustrated in the figure, are described in detail in the subsequent sections.

5.1. CoRIfEE-*Rel Inputs*

The CoRIfEE-Rel method takes three main inputs:

- **Pool of interpretable models ($\mathcal{I}$)**: A collection of rule-based explanations generated from different interpretable machine learning models, such as random forest or JRip. The explanations are in the if–then format, making them more understandable to humans.

- **Dataset ($\mathcal{D}$)**: The dataset on which the method operates, providing the attributes and values used for rule generation and evaluation.
- **Knowledge Graph ($\mathcal{KG}$)**: A directed, labeled knowledge graph that represents domain knowledge and captures semantic relationships among attributes and their values.

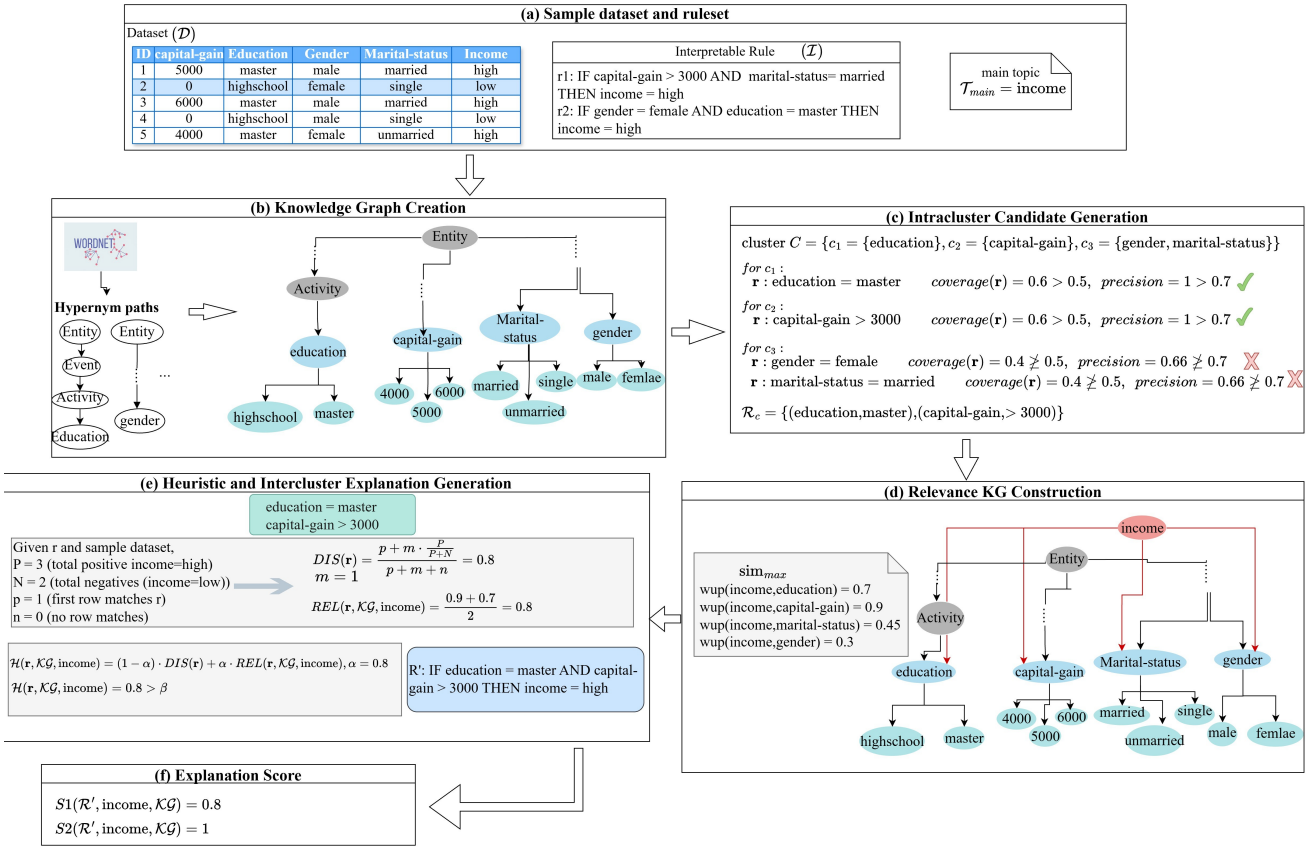


Figure 2. Illustrative example of CoRIFEE-Rel methodology. (**a**) A sample dataset, an interpretable model, and the main topic. (**b**) Construction of the domain-specific knowledge graph $\mathcal{KG}$. (**c**) Generation of intracluster candidate rules. (**d**) Modification of the knowledge graph to incorporate and connect the main concept to the existing nodes. (**e**) Computation of heuristic and generation of intercluster explanation. (**f**) Final calculation of scores S1 and S2 for the generated rule.

5.2. Dataset

We select four binary classification datasets from the UCI repository [49]. The datasets are selected because they represent diverse domains and contain identifiable attributes that can be associated with external sources of background knowledge. These datasets have been widely used in machine learning and XAI research [50,51]:

- *Hepatitis* dataset contains medical records of patients diagnosed with hepatitis, aiming to predict whether a patient will survive or die.
- *Adult* dataset determines whether a person's yearly earnings are greater than 50 K.
- *Bank Marketing* dataset provides information about customer interactions with marketing campaigns, serving to predict campaign success.
- *Water Quality* dataset contains data on multiple water quality indicators used to assess water cleanliness.

Before rule generation, each dataset is preprocessed as described in Section 5.5. The resulting preprocessed data is used as the input to the rule generation and semantic-relevance procedures.

5.3. Knowledge Graph Construction

Knowledge graphs are constructed from WordNet [52] by extracting hypernym paths for dataset attributes and representing them as subject–verb–object (SVO) triples, where the subject and object correspond to nodes along the hypernym path, and the verb denotes a labeled edge connecting them. These triples are then used to build the knowledge graph. For example, consider the hypernym path extracted for "education" from WordNet in part (b), Figure 2. The subject–verb–object triple from the example is defined as ("Entity","manifested as", "Event"), ("Event", "constitutes", "Activity"), and ("Activity", "encompasses", "Education"). Each triple is subsequently represented as a directed, labeled edge in the knowledge graph.

Rather than using the entire WordNet, CoRIfEE-Rel operates on a focused domain-specific subgraph containing only the concepts associated with the attributes and their semantic relationships. This graph, denoted by $\mathcal{KG}_0$, is constructed independently of a particular main topic and serves as the basis for the subsequent topic-specific graph enrichment and semantic relevance computation.

Before querying WordNet, dataset attribute names are normalized by replacing separators such as hyphens and underscores with spaces and, when necessary, expanding dataset-specific abbreviations to their corresponding terms. For each normalized attribute, candidate noun synsets are retrieved from WordNet. When multiple candidate synsets are available, the synset with closest semantic meaning to the attribute is selected. For compound or ambiguous attributes, the selection is guided by the dataset context. If no match is found, the closest semantically related concept is selected and used as a substitute for the attribute.

5.4. Semantic Relevance Measurement

The primary objective of CoRIfEE-Rel is to generate explanations that are highly semantically relevant to a given main concept or topic. To quantify this semantic relevance, we introduce a relevance heuristic $REL(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG})$, which measures how a given rule $\mathbf{r}$ is semantically relevant to a given main topic $\mathcal{T}_{main}$ over a domain-specific knowledge graph.

To compute the heuristic, the initial knowledge graph $\mathcal{KG}_0$ is further enriched with the main topic node to obtain a topic-conditioned knowledge graph $\mathcal{KG}$. The purpose of this enrichment is to explicitly represent the semantic relationship between the main topic and the concepts associated with the dataset attributes.

We begin by creating a node corresponding to $\mathcal{T}_{main}$. For each node labeled as "attribute", $label(v,$ "attribute") in $\mathcal{KG}_0$, we consider the concepts along its hypernym path, denoted by V_{attr}. Wu–Palmer similarity is calculated between $\mathcal{T}_{main}$ and each concept in V_{attr}. The concept with the highest similarity is selected as the most semantically related concept for that attribute. This node, denoted as v_{max}, is then connected to the main topic node via a new edge:

$$v_{max} = \{v \in V_{attr} \mid \text{wup}(\mathcal{T}_{main}, v) = \text{sim}_{max}\} \tag{1}$$

$$E = E \cup \{(\mathcal{T}_{main}, \text{sim}_{max}, v_{max})\} \tag{2}$$

As in (2), for semantic similarity, we use *Wu–Palmer* similarity (wup) [53] defined as:

$$\text{wup}(c_i, c_j) = \frac{2 \cdot d(c_{lcs})}{d(c_i) + d(c_j)} \tag{3}$$

In this equation, $d(c_i)$ and $d(c_j)$ denote the depth of concept c_i and c_j, and $d(c_{lcs})$ represents the depth of the least common subsumer.

This process is repeated for each node in V_{attr}, resulting in the topic-conditioned knowledge graph $\mathcal{KG}$. $\mathcal{KG}_0$ is therefore independent of a particular main topic, whereas the topic-conditioned graph $\mathcal{KG}$ incorporates the semantic connections established for the selected main topic.

As shown in Figure 2, the attributes "education", "capital-gain", "marital-status", and "gender" have the highest similarity scores to the main topic and are therefore selected as v_{max}. For simplicity, we assume that the highest similarity scores correspond to attribute-level nodes; however, in our experiments on datasets, the most similar nodes could also lie higher in the hierarchy (e.g., at the concept or category level). New edges are then added between these nodes and the newly introduced node "income" in the knowledge graph, shown in red in the figure.

The semantic relevance measure is designed to capture two complementary aspects of the relationship between an attribute and the main topic: semantic similarity and structural proximity. Wu–Palmer similarity captures the conceptual similarity in the semantic hierarchy, while the shortest path distance in the topic-conditioned knowledge graph captures their structural proximity. A condition should therefore receive a higher relevance contribution when its attribute is conceptually more similar to the main topic and structurally closer to it.

For each condition f in a rule, we define its relevance contribution as:

$$\frac{2 \cdot \text{wup}(f.attr, \mathcal{T}_{main})}{1 + \text{dist}(f.attr, \mathcal{T}_{main})} \tag{4}$$

where $\text{wup}(f.attr, \mathcal{T}_{main})$ denotes the Wu–Palmer similarity between the attribute and the main topic, and $\text{dist}(f.attr, \mathcal{T}_{main})$ denotes their shortest path distance in the $\mathcal{KG}$. The addition of 1 in the denominator prevents division by zero when the attribute and the main topic coincide. The factor 2 ensures that, for the minimum nonzero graph distance of one edge, the contribution reduces to the Wu–Palmer similarity.

In this formulation, semantic similarity contributes positively to the relevance, while structural distance discounts that contribution as the attribute is farther from the main topic in the graph.

The relevance of an entire rule is then defined as the average relevance contribution of its conditions:

$$\text{REL}(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG}) = \frac{1}{|\mathbf{r}|} \cdot \sum_{f \in \mathbf{r}} \frac{2 \cdot \text{wup}(f.attr, \mathcal{T}_{main})}{1 + \text{dist}(f.attr, \mathcal{T}_{main})} \tag{5}$$

Here, $\mathbf{r}$ denotes a rule consisting of multiple attribute-value pairs, where each $f \in \mathbf{r}$ represents an individual condition and $f.attr$ denotes its corresponding attribute. The term $\text{wup}(f.attr, \mathcal{T}_{main})$ measures the semantic similarity between the attribute, and $\mathcal{KG}$, $\text{dist}(f.attr, \mathcal{T}_{main})$ gets the length of the shortest path between the main topic and the attribute in the graph.

The normalization $|r|$ converts the accumulated score into an average contribution across the conditions of a rule. As a result, the score is not inherently increased by using fewer conditions or decreased by using more conditions. Therefore, rule length does not directly affect REL, instead, the score reflects the average semantic alignment of the conditions with the main topic. In this approach, longer rule can obtain a high REL score when its conditions are strongly aligned with the target concept, while a shorter rule can obtain a low REL score when its conditions are weakly related to the main topic.

Accordingly, the introduced *REL* should be interpreted as a graph-based semantic measure rather than as a measure of predictive feature importance. It quantifies the semantic alignment of the attributes appearing in a rule with the selected main topic according to the knowledge representation. It does not directly measure the relevance, and usefulness of an explanation as perceived by human users.

5.5. Semantic Relevance Explanation Generation

This section elaborates on the main processes in CoRIfEE-Rel to generate semantically relevant explanations. The method consists of a sequence of dataset preprocessing, knowledge-graph construction and topic-conditioned enrichment, semantic clustering, and intracluster and intercluster rule generation to general final explanation. Algorithm 1 summarizes the operations and intermediate outputs passed between different steps.

The method first preprocesses the dataset by renaming attributes, discretizing numerical features, and imputing missing values. The knowledge graph $\mathcal{KG}$, constructed as described in Section 5.3 and enriched according to the procedure in Section 5.4, is subsequently used to construct a weighted graph over the dataset. Louvain community detection is applied to this graph to obtain the set of clusters $C = \{c_1, \cdots, c_m\}$. Louvain is selected as it can identify clusters in graph-structured semantic data without requiring the number of clusters to be specified. The clustering step ensures that the resulting rules are semantically and conceptually coherent, while also enhancing computational efficiency.

The method utilizes clusters to generate an explanation through a two-level approach: intracluster and intercluster rule generation.

Rule generation is performed separately for each target class. clusterFeatures contains the feature-value conditions associated with the current cluster and class.

The *intracluster* rule generation gets a set of candidate rules $\mathcal{R}_c$ within each cluster by combining the conditions that meet specified thresholds for coverage and precision. For each cluster c_i in clusters C, each condition is individually assessed based on the thresholds. Conditions that fulfill the criteria are then added to the rule **r**. This process continues until all clusters have been visited and processed, and no additional conditions can be appended to **r** without violating the threshold constraints as described in Alogrithm 1. The resulting rules are then added to $\mathcal{R}_c$. Typically, the default values for the coverage and precision thresholds, denoted as th_c and th_p, are set to 0.5 and 0.7, respectively. The thresholds are based on a combination of theoretical foundations and empirical validation. A coverage threshold of 0.5 ensures that a rule must apply to at least half of the relevant instances, while a precision threshold of 0.7 ensures that the rule makes correct predictions in at least 70% of the cases it covers. These thresholds provide a balance between generality and accuracy, supporting the rule generation that effectively captures significant patterns in the data. Furthermore, as the rule generation starts with single-condition rules that represent concepts within a cluster, the approach naturally limits the number of generated rules and reduces the search space by filtering out redundant or low-quality conditions.

To clearly illustrate the intracluster candidate generation, we refer to the example in Figure 2. The attributes are first clustered based on their semantic similarity in the knowledge graph, resulting in three clusters: c_1, c_2, and c_3, as shown in step (c). The intracluster candidate generation phase then focuses on identifying valid rules within each cluster. For clusters c_1 and c_2, single-condition rules "if education=master then income=high" and "if capital-gain>3000 then income=high" are formulated, both of which satisfy the minimum coverage and precision thresholds. Cluster c_3 does not yield any rules that meet the required criteria. The two valid rules from clusters c_1 and c_2 are stored in $\mathcal{R}_c$ and passed to the intercluster rule generation step.

The *intercluster* rule generation aims to generate the final explanation. It iterates through pairs of rules in $\mathcal{R}_c$, evaluates the merge validation of the two rules, and, in that case, it appends them to $\mathcal{R}'$. Each pair of candidate rules $\mathbf{r}_i$ and $\mathbf{r}_j$ is evaluated against two defined functions. The candidate rules are first evaluated using the *merge_validation* function which checks whether their antecedent conditions can be combined without contradiction. For numerical attributes, the conditions are considered compatible when their resulting interval is non-empty and logically consistent. For categorical attributes, a merge is rejected when the two rules specify conflicting values for the same attribute. Conditions involving different attributes are combined. If the validation succeeds, the two rules are merged to form a new candidate rule, $\mathbf{r}_{merge}$, consisting of the compatible antecedent conditions of $\mathbf{r}_i$ and $\mathbf{r}_j$.

After successful completion of the merge validation, $\mathbf{r}_{merge}$ is evaluated using a heuristic evaluation function, denoted by $\mathcal{H}$, that balances predictive performance and semantic relevance and is defined as:

$$\mathcal{H}(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG}) = (1 - \alpha) \cdot DIS(\mathbf{r}) + \alpha \cdot REL(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG}) \tag{6}$$

where the α parameter controls the relative influence of a traditional rule learning heuristic and the semantic relevance measurement introduced in (5). The rule learning heuristic employs the m-estimate, a metric that offers a configurable trade-off between precision and weighted related accuracy, formally defined as:

$$DIS(\mathbf{r}) = \frac{p + m \cdot \frac{P}{P+N}}{p + n + m} \tag{7}$$

In this equation, p represents the positive examples out of all positive examples P that are covered by the rule $\mathbf{r}$, n denotes the negative examples of all negative examples N that are covered by the rule, and parameter m serves as a tunable factor that specifies the mentioned trade-off.

The parameter α in (6) is between 0 and 1 and specifies the trade-off between the rule learning heuristic and relevance heuristic. With the increase in the α parameter, the contribution of the semantic relevance in the heuristic gets more dominant, as the rule learning heuristic loses its importance. Subsequently, the conditions that satisfy the threshold β are appended to the final explanation $\mathcal{R}'$ as explained in Algorithm 1.

The threshold β serves as a minimum acceptance criterion that filters out rules with low $\mathcal{H}$ and ensures that only rules with sufficient semantic relevance and predictive quality are included in the final explanation.

To clarify the proposed heuristic and intercluster rule generation process, we again refer to the example illustrated in Figure 2. The candidate merged rule comprises two conditions: “education=master” and “capital-gain>3000”. To compute the overall heuristic $\mathcal{H}$, both the semantic relevance *REL* and the rule learning heuristic *DIS* must be evaluated. Based on the assumed Wu–Palmer similarity values between the main topic “income” and the attributes in the merged rule, and considering a shortest path length of 1 in both cases, the semantic relevance score for the rule is calculated using Equation (5) as $REL(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG}) = 0.8$. For the rule learning heuristic, according to Equation (7), the rule covers one positive instance and no negative instances, with total positive and negative counts $P = 3$ and $N = 2$, respectively, resulting in a rule learning score of $DIS = \frac{1.6}{2} = 0.8$. Finally, Equation (6) combines the two scores to balance predictive accuracy with semantic relevance. With a trade-off parameter set to $\alpha = 0.8$, the heuristic is computed as $\mathcal{H} = 0.8$.

Algorithm 1 Pseudocode on CoRIfEE-Rel.

Input: Pool of interpretable models $\mathcal{I}$.
Input: Knowledge graph $\mathcal{KG}_0$
Input: Dataset $\mathcal{D}$
Input: Main topic or core concept $\mathcal{T}_{main}$.
Output: Explanation $\mathcal{R}'$ highly relevant to the main topic.

Preprocess $\mathcal{D}$: rename attributes, discretize numerical attributes, and impute missing values.

$\mathcal{KG} \leftarrow$ Update $\mathcal{KG}_0$ by computing Wu–Palmer similarity between $\mathcal{T}_{main}$ and the nodes and adding the corresponding semantic connections.
cluster $C = \{c_1, ..., c_m\} \leftarrow$ cluster features based on their semantic similarity using the Louvain community detection method.

$\mathcal{R}_c = []$
for c_i in C **do**
    clusterFeatures $\leftarrow$ get all feature-value conditions $(\mathbf{V}_j, \mathbf{V}_k)$ for one target class from c_i.
    $\mathbf{r} \leftarrow \emptyset$
    for f_i in clusterFeatures **do**
        if $coverage(\mathbf{r} \cup f_i) \geq th_c$ and $precision(\mathbf{r} \cup f_i) \geq th_p$ **then**
            $\mathbf{r} \leftarrow \mathbf{r} \cup f_i$
        end if
    end for
    $\mathcal{R}_c \leftarrow \mathcal{R}_c \cup \mathbf{r}$
end for

$\mathcal{R}' \leftarrow []$
for each $\mathbf{r}_i$ and $\mathbf{r}_j$ in $\mathcal{R}_c$ **do**
    mergeValid, $r_{merge} \leftarrow$ merge_validation($\mathbf{r}_i, \mathbf{r}_j$)
    if mergeValid **then**
        Calculate $DIS(\mathbf{r}_{merge}) \leftarrow$ compute using (7)
        Calculate $REL(\mathbf{r}_{merge,\mathcal{T}_{main},\mathcal{KG}}) \leftarrow$ compute using (5)
        $\mathcal{H}(\mathbf{r}_{merge}, \mathcal{T}_{main}, \mathcal{KG}) = (1-\alpha) \cdot DIS(\mathbf{r}_{merge}) + \alpha \cdot REL(\mathbf{r}_{merge}, \mathcal{T}_{main}, \mathcal{KG})$

        if $\mathcal{H}(\mathbf{r}_{merge}, \mathcal{T}_{main}, \mathcal{KG}) \geq \beta$ **then**
            $\mathcal{R}' \leftarrow \mathcal{R}' \cup \mathbf{r}_{merge}$
        end if
    end if
end for
Remove duplicate rules and redundant conditions from $\mathcal{R}'$
return $\mathcal{R}'$

6. Results

6.1. Experimental Setup

We evaluate the performance of the CoRIfEE-Rel on the introduced datasets in Section 5.

CoRIfEE-Rel also requires an ensemble of interpretable models, a domain-specific knowledge graph, and the main topic, which is specified based on the dataset. In our experiment, we use a few decision trees from a random forest and extract rules from them to serve as interpretable models.

In the present evaluation, accuracy refers to predictive accuracy with respect to the ground-truth class labels and should not be interpreted as explanation fidelity to a particular source. CoRIfEE-Rel operates as a meta-XAI approach that synthesizes rules from a pool of interpretable models rather than approximating the predictions of a single fixed black-box

model. Accordingly, our evaluation examines whether incorporating graph-based semantic relevance can produce semantically aligned rule-based explanations while maintaining predictive accuracy. Evaluating agreement with a specific source model would be a different objective that requires a specific model-fidelity metric.

For each dataset, we define a main concept ($\mathcal{T}_{main}$) to guide the semantic relevance computation and explanation generation. Table 2 lists the topics.

Table 2. Main topics ($\mathcal{T}_{main}$) used for datasets in the experiments.

Dataset	Main Topic ($\mathcal{T}_{main}$)
Hepatitis	Liver disease
Adult	Income
Bank Marketing	Deposit
Water Quality	Potable

Note: Dataset names are shown in italics.

To evaluate CoRIfEE-Rel, we consider two sources of input rules: Random Forest and JRip. For Random Forest, we select 3-4 decision trees from the trained forest. Each selected decision tree is converted into rules by representing each root-to-leaf path as if–then rules. In each rule, the antecedent consists of the combined split conditions along the decision paths, while the consequence corresponds to the class predicted at the leaf node. For JRip, the learned rules are used directly as the input rule pool. CoRIfEE-Rel is then applied separately to these two sources.

The generated explanation by CoRIfEE-Rel is evaluated through the two metrics described in the following.

To evaluate the quality of the generated explanation by CoRIfEE-Rel, we introduce two complementary scoring metrics: $S1$ and $S2$. Both metrics provide different assessments of the relevance alignment of the explanations generated by CoRIfEE-Rel and are complementary rather than independent validation metrics. They are based on graph-based semantic alignment and do not, by themselves, establish human-perceived explanation relevance. $S1$ and $S2$ operate over a common set of components: $\mathcal{R}$ represents the generated explanation by CoRIfEE-Rel, consisting of a set of rules $\mathbf{r} \in \mathcal{R}$, $REL(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG})$ is the function that computes the relevance of rule $\mathbf{r}$ with respect to the main topic, given the knowledge graph $\mathcal{KG}$, as defined in (5), and $|\mathcal{R}|$ denotes the number of rules in the explanation.

The first metric, $S1$, measures the overall relevance of the explanation by averaging the individual scores of the rules:

$$S1(\mathcal{R}, \mathcal{T}_{main}, \mathcal{KG}) = \frac{\sum_{\mathbf{r} \in \mathcal{R}} REL(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG})}{|\mathcal{R}|} \tag{8}$$

This score ranges from 0 to 1, with higher values indicating that the rules in the explanation are, on average, more relevant to the main topic. Thus, $S1$ provides a holistic view of explanation alignment with the target concept or topic.

The second metric, $S2$, evaluates the quality of the explanation based on whether each rule meets a pre-defined relevance threshold θ. It calculates the proportion of rules whose relevance score meets the threshold:

$$S2(\mathcal{R}, \mathcal{T}_{main}, \mathcal{KG}, \theta) = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} I(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG}, \theta) \tag{9}$$

The indicator function $I()$ is defined as:

$$I(\mathbf{r}) = \begin{cases} 1, & \text{if } REL(\mathbf{r}, \mathcal{T}_{main}, \mathcal{KG}) \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Here, θ represents the minimum acceptable relevance score for a rule to be considered valid. As with $S1$, the $S2$ score also lies in the range [0, 1], where higher values reflect a greater proportion of high-relevant rules.

Figure 2 shows an example of the scoring evaluation. Based on the explanation $\mathcal{R}'$, and the calculated REL for ruleset with length 1 in Section 5, the score $S1$ is computed as $S1 = \frac{0.8}{1} = 0.8$. In addition, (9) measures the proportion of rules while individual relevance scores exceed a predefined threshold θ. Assuming $\theta = 0.6$, the score $S2$ is calculated as $S2 = \frac{1}{1} = 1$.

Together, the $S1$ and $S2$ metrics offer complementary assessments of a generated explanation. While $S1$ quantifies the average relevance of all rules, $S2$ highlights how many rules meet a minimum relevance standard. This dual-perspective evaluation supports a more detailed analysis of CoRIFEE-Rel performance.

From a computational perspective, the main cost of CoRIFEE-Rel arises from rule generation. By clustering semantically related attributes, the method reduces the search space and avoids exhaustive rule combination, thereby improving scalability and making the approach practical.

6.2. Quantitative Results

We evaluate the performance of CoRIFEE-Rel across different values of the α parameter, which specifies the contribution of semantic relevance in the explanation generation process. The results are compared against the random forest model and are reported in Figure 3.

To examine the effect of α, we perform a sensitivity analysis to illustrate how changing the α affects accuracy, $S1$, and $S2$.

For the *Hepatitis* dataset, both scoring metrics, $S1$ and $S2$, increase as the α parameter increases, indicating that greater emphasis on semantic relevance leads to explanations with higher semantic relevance scores. Based on the α analysis in Figure 3, we use $\alpha = 0.8$ as a configuration for the subsequent comparisons. This value assigns higher emphasis on semantic relevance while maintaining the contribution of predictive rule-learning heuristic. At this setting, the scoring metrics improve compared to random forest, while maintaining a comparable accuracy. This shows that CoRIFEE-Rel can generate explanations with higher semantic relevance with slight changes in predictive performance.

A similar trend is observed for the *Adult* dataset. At $\alpha = 0$, the heuristic relies solely on the traditional rule learning heuristic, resulting in the highest accuracy but the lowest relevance scores. As α increases, the semantic relevance of the explanations improves. At $\alpha = 0.8$, we observe a trade-off between predictive accuracy and semantic relevance, with relevance scores exceeding those of random forest while predictive accuracy changes only modestly.

The results on *Bank Marketing* and *Water Quality* follow similar patterns. As semantic relevance is given more weight, the relevance scores increase, while accuracy is slightly reduced.

The setting $\alpha = 0$ and $\alpha = 1$ can be interpreted as an ablation study of the components of the heuristic. At $\alpha = 0$, the semantic relevance component REL is removed and candidate selection is entirely based on DIS, whereas at $\alpha = 1$, DIS is removed and the rules selected are based only on REL.

Based on the sensitivity analysis, we use $\alpha = 0.8$ as a representative point for the subsequent comparative evaluation. This value assigns higher weight to semantic relevance while maintaining the contribution of the predictive rule-learning heuristic. It provides a common fixed configuration for comparing CORIFEE-Rel across datasets.

For the comparative evaluation at $\alpha = 0.8$, we use five-fold cross-validation. The complete data-dependent rule-generation process is performed independently using the training set of each fold. The source model is trained on the training fold, and CORIFEE-Rel performs candidate generation and rule selection using only the training data. The knowledge graph derived from WordNet is constructed from the attributes in the dataset and it is fixed across folds. The generated explanation is then evaluated on the test fold. This ensures that the information from the test fold is not used during preprocessing, rule extraction, candidate selection and explanation generation.

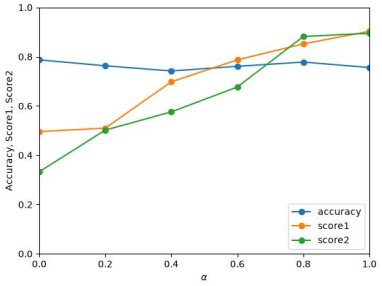
For prediction all generated rules whose conditions are satisfied by a test instance in a class vote, and the class receiving majority of votes is assigned. If no generated rule covers the test instance, the majority class in the corresponding training fold is assigned. In the case of a tie between classes, the majority class in the training fold is used to break the tie.

Table 3 summarizes the mean and standard deviation of the results obtained using five-fold cross-validation for CORIFEE-Rel at the fixed representative configuration $\alpha = 0.8$. This setting gives greater weight to semantic relevance while maintaining a contribution from the predictive rule-learning heuristic. Two separate CORIFEE-Rel experiments are reported for each dataset; one using rules from a few decision trees from random forest, and one using rules from JRIP. Random Forest and JRIP represent two commonly used approaches for generating interpretable rule-based explanations. Across all datasets, CORIFEE-Rel improves the semantic relevance $S1$ and $S2$ relative to JRIP and random forest. Overall, the results demonstrate that CORIFEE-Rel achieves a strong balance between predictive performance and semantic relevance relative to both models, consistently improving semantic relevance explanations at the expense of slightly reducing predictive performance. The small standard deviations indicate that the improvements are stable across different folds.

Table 3. Comparison of CORIFEE-Rel ($\alpha = 0.8$) with Random Forest and JRIP rule-based explanations.

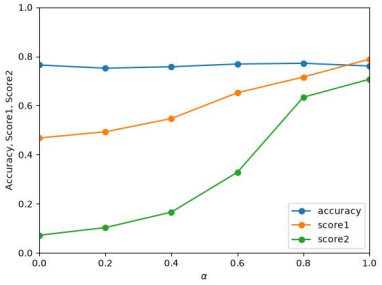
Dataset	Method	Accuracy	S1	S2
Hepatitis	Random Forest	$0.803_{\pm 0.050}$	$0.713_{\pm 0.027}$	$0.768_{\pm 0.033}$
	CORIFEE-Rel	$0.778_{\pm 0.026}$	$0.852_{\pm 0.024}$	$0.882_{\pm 0.029}$
	JRIP	$0.789_{\pm 0.040}$	$0.831_{\pm 0.028}$	$0.812_{\pm 0.021}$
	CORIFEE-Rel	$0.770_{\pm 0.028}$	$0.896_{\pm 0.023}$	$0.868_{\pm 0.019}$
Adult	Random Forest	$0.798_{\pm 0.013}$	$0.472_{\pm 0.036}$	$0.325_{\pm 0.029}$
	CORIFEE-Rel	$0.772_{\pm 0.021}$	$0.716_{\pm 0.029}$	$0.634_{\pm 0.030}$
	JRIP	$0.802_{\pm 0.018}$	$0.748_{\pm 0.027}$	$0.754_{\pm 0.025}$
	CORIFEE-Rel	$0.783_{\pm 0.023}$	$0.816_{\pm 0.019}$	$0.907_{\pm 0.013}$
Bank Marketing	Random Forest	$0.893_{\pm 0.012}$	$0.639_{\pm 0.027}$	$0.526_{\pm 0.018}$
	CORIFEE-Rel	$0.875_{\pm 0.017}$	$0.746_{\pm 0.023}$	$0.698_{\pm 0.020}$
	JRIP	$0.878_{\pm 0.019}$	$0.623_{\pm 0.022}$	$0.704_{\pm 0.016}$
	CORIFEE-Rel	$0.844_{\pm 0.015}$	$0.835_{\pm 0.017}$	$0.918_{\pm 0.013}$
Water Quality	Random Forest	$0.622_{\pm 0.010}$	$0.641_{\pm 0.021}$	$0.689_{\pm 0.027}$
	CORIFEE-Rel	$0.614_{\pm 0.018}$	$0.732_{\pm 0.019}$	$0.744_{\pm 0.014}$
	JRIP	$0.596_{\pm 0.013}$	$0.769_{\pm 0.014}$	$0.724_{\pm 0.018}$
	CORIFEE-Rel	$0.587_{\pm 0.014}$	$0.886_{\pm 0.011}$	$0.896_{\pm 0.015}$

Algorithm	α	**Accuracy**	$S1$	$S2$
CoRIfEE-Rel	0	0.787	0.496	0.373
	0.2	0.763	0.510	0.502
	0.4	0.742	0.698	0.576
	0.6	0.761	0.787	0.677
	0.8	0.778	0.852	0.882
	1	0.756	0.903	0.895
Random Forest		0.803	0.713	0.768



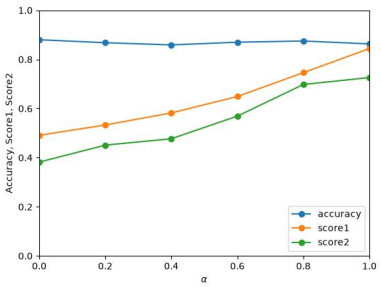
(**a**) *Hepatitis* dataset

Algorithm	α	**Accuracy**	$S1$	$S2$
CoRIfEE-Rel	0	0.765	0.468	0.072
	0.2	0.752	0.493	0.108
	0.4	0.758	0.547	0.166
	0.6	0.769	0.652	0.328
	0.8	0.772	0.716	0.634
	1	0.761	0.789	0.707
Random Forest		0.798	0.472	0.325



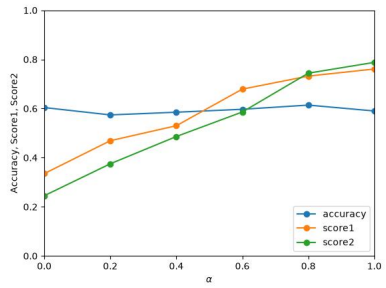
(**b**) *Adult* dataset

Algorithm	α	**Accuracy**	$S1$	$S2$
CoRIfEE-Rel	0	0.880	0.491	0.382
	0.2	0.868	0.533	0.451
	0.4	0.859	0.582	0.477
	0.6	0.870	0.649	0.569
	0.8	0.875	0.746	0.698
	1	0.863	0.844	0.726
Random Forest		0.893	0.639	0.526



(**c**) *Bank Marketing* dataset

Algorithm	α	**Accuracy**	$S1$	$S2$
CoRIfEE-Rel	0	0.604	0.335	0.245
	0.2	0.574	0.469	0.375
	0.4	0.585	0.530	0.486
	0.6	0.597	0.679	0.586
	0.8	0.614	0.732	0.744
	1	0.590	0.761	0.788
Random Forest		0.622	0.641	0.689



(**d**) *Water Quality* dataset

Figure 3. Performance of CoRIfEE-Rel and random forest across four datasets, presented both in tabular (**left side**) and graphical (**right side**) form.

Since the random forest-based CoRIfEE-Rel setting uses rules extracted from a selected set of decision trees, we additionally examine the performance of these individual trees on the *Hepatitis* dataset. Table 4 reports the accuracy, $S1$, and $S2$ values of the same selected individual decision trees used to construct the random forest pool in Table 3 and compares with the CoRIfEE-Rel explanation generated from that pool.

Table 4. Comparison to individual trees of a random forest on *Hepatitis* dataset using the representative configuration $\alpha = 0.8$.

Tree	Accuracy	S1	S2
CORIFEE-Rel	0.778	0.852	0.882
tree1	0.695	0.581	0.549
tree2	0.726	0.537	0.516
tree3	0.754	0.624	0.585

Sensitivity Analysis

To examine whether $S2$ improvements depend on the choice of the threshold θ, we conduct a sensitivity analysis over $\theta \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$. It is important to note that θ is an evaluation threshold used only in the computation of $S2$, and it does not affect the generation or selection of rules in the final explanation.

Table 5 reports the results across all four datasets for both random forest and JRIP. As expected, $S2$ decreases monotonically as the threshold becomes more restrictive. In addition, CORIFEE-Rel consistently achieves higher $S2$ than its corresponding baseline in all tested thresholds across all datasets.

Table 5. Sensitivity analysis of $S2$ with respect to the threshold θ for random forest and JRIP and the corresponding CORIFEE-Rel explanation.

Dataset	Method	0.4	0.5	0.6	0.7	0.8
Hepatitis	Random Forest	0.907	0.887	0.768	0.527	0.334
	CORIFEE-Rel	0.947	0.915	0.882	0.628	0.522
	JRIP	0.922	0.865	0.812	0.625	0.421
	CORIFEE-Rel	0.968	0.874	0.868	0.749	0.634
Adult	Random Forest	0.574	0.431	0.325	0.205	0.098
	CORIFEE-Rel	0.832	0.755	0.634	0.563	0.419
	JRIP	0.891	0.811	0.754	0.556	0.327
	CORIFEE-Rel	0.983	0.952	0.907	0.748	0.533
Bank Marketing	Random Forest	0.568	0.469	0.526	0.273	0.189
	CORIFEE-Rel	0.843	0.736	0.698	0.568	0.434
	JRIP	0.877	0.764	0.704	0.564	0.418
	CORIFEE-Rel	0.988	0.952	0.918	0.768	0.402
Water Quality	Random Forest	0.841	0.736	0.689	0.518	0.357
	CORIFEE-Rel	0.966	0.865	0.744	0.524	0.360
	JRIP	0.867	0.764	0.704	0.446	0.318
	CORIFEE-Rel	0.985	0.927	0.896	0.641	0.534

Note: Dataset names are shown in italics.

We further examine the effect of the acceptance threshold β, which determines whether a candidate merged rule should be appended to the final explanation. As reported in Table 6, we vary β from 0.4 to 0.8 while keeping the remaining parameters fixed. Across the four datasets and both random forest and JRIP, increasing β leads to higher $S1$ and $S2$ scores, while predictive accuracy remains relatively stable at lower values of β and decrease under stricter acceptance thresholds. This behavior is expected as larger β applies a stricter acceptance criterion and selects candidate rules with higher heuristic values.

Table 6. Sensitivity analysis of CoRIfEE-Rel with respect to the acceptance threshold β across all datasets and base models.

Dataset	β	Accuracy	S1	S2
Hepatitis /Random Forest	0.4	0.780	0.807	0.822
	0.5	0.782	0.816	0.843
	0.6	0.778	0.852	0.882
	0.7	0.761	0.874	0.905
	0.8	0.754	0.914	0.926
Hepatitis/JRip	0.4	0.781	0.824	0.783
	0.5	0.774	0.857	0.824
	0.6	0.770	0.896	0.868
	0.7	0.768	0.918	0.885
	0.8	0.742	0.922	0.897
Adult/Random Forest	0.4	0.784	0.642	0.548
	0.5	0.777	0.679	0.583
	0.6	0.772	0.716	0.634
	0.7	0.767	0.736	0.665
	0.8	0.761	0.752	0.688
Adult/JRip	0.4	0.789	0.759	0.827
	0.5	0.789	0.784	0.866
	0.6	0.783	0.816	0.907
	0.7	0.769	0.834	0.916
	0.8	0.758	0.862	0.933
Bank Marketing/Random Forest	0.4	0.885	0.667	0.579
	0.5	0.880	0.729	0.642
	0.6	0.875	0.746	0.698
	0.7	0.869	0.784	0.759
	0.8	0.864	0.809	0.772
Bank Marketing/JRip	0.4	0.856	0.776	0.845
	0.5	0.851	0.819	0.883
	0.6	0.844	0.835	0.918
	0.7	0.839	0.852	0.924
	0.8	0.835	0.876	0.936
Water Quality/Random Forest	0.4	0.619	0.674	0.668
	0.5	0.616	0.664	0.692
	0.6	0.614	0.732	0.744
	0.7	0.607	0.788	0.785
	0.8	0.594	0.797	0.825
Water Quality/JRip	0.4	0.590	0.836	0.853
	0.5	0.588	0.861	0.874
	0.6	0.587	0.886	0.896
	0.7	0.574	0.909	0.916
	0.8	0.570	0.934	0.937

Note: Dataset names are shown in italics.

6.3. Qualitative Results

The rules illustrate a contrast in semantic relevance with the underlying concept of "liver disease". For instance, the rules learned from random forest include conditions such as "sex = male" and "histology = no", which are not directly associated with liver pathology. In contrast, the rules generated by CoRIfEE-Rel are more closely aligned with medically meaningful indicators, such as "ascites = no", SGOT and albumin ranges, which are strongly associated with liver function and disease.

These differences reflect the capacity of CoRIfEE-Rel to generate interpretable and semantically relevant rules grounded in domain-relevant features.

To improve interpretability, we translated selected rule fragments from Table 7 into narrative form in Table 8, expressing logical conditions as natural language statements. This style conveys the model's decision logic in a format more accessible to clinicians and domain experts, highlighting how specific combinations of patient attributes relate to predicted outcomes (e.g., survival or death). By avoiding formal rule syntax, the narrative descriptions enhance semantic transparency and facilitate qualitative assessment of model reasoning.

Table 7. Fragments of rule sets generated from random forest and CoRIfEE-Rel on *Hepatitis* dataset.

(**a**) Rule sets learned by random forest		(**b**) Rule sets generated by CoRIfEE-Rel	

`class = 1 :-`	`sex = female,`
	`steroid = yes,`
	`anorexia = no.`
		`class = 1 :-`	`ascites = no,`
			`SGOT <= 40,`
			`liver firm = no,`

.		.	
.		.	
.		.	

`class = 0 :-`	`sex = male,`
	`steroid = no,`
	`histology = yes.`
		`class = 0 :-`	`albumin >= 1.9,`
			`liver big = yes,`
			`SGOT >= 150.`

.		.	
.		.	
.		.	

Table 8. Alternative representations of the generated rule sets from random forest and CoRIfEE-Rel on the *Hepatitis* dataset.

(**a**) Rule sets generated by random forest	(**b**) Rule sets generated by CoRIfEE-Rel
A patient is predicted to survive if the patient is female, is receiving steroid treatment, and does not experience anorexia.	A patient is predicted to survive if ascites is absent, the SGOT level is below 40, and the liver is not firm.
...	...
A patient is predicted to die if they are male, is not receiving steroid treatment, and has a positive histology result.	A patient is predicted to die if albumin level is above 1.9, the SGOT level is above 150, and the liver is enlarged.
...	...

7. Discussion

The results from our work show that CoRIfEE-Rel generally increases graph-based semantic relevance with slight changes in predictive accuracy. This reflects the defined goal of the heuristic $\mathcal{H}$, which balances the predictive component DIS and the semantic relevance component REL.

The sensitivity analysis further illustrates the behavior of the evaluation parameters. Increasing α gives greater weight ot semantic relevance during rule selection, while increasing β applied stricter acceptance criterion and generally increases $S1$ and $S2$, with some reduction in predictive accuracy at higher values. θ is used only as an evaluation threshold for $S1$ and $S2$; increasing θ leads to consider high relevant rules and generally decreases $S2$, without affecting the generated explanations.

CoRIfEE-Rel is also applied separately to two baseline models: Random Forest and JRip. The current evaluation, however, measures graph-based semantic relevance rather than human-perceived explanations quality, so human evaluation remains a direction for future work.

8. Conclusions

In this work, we introduced CoRIfEE-Rel, a novel instantiation of the CoRIfEE meta-XAI platform that focuses on incorporating semantic relevance into rule-based explanation generation. Motivated by the principles of human-centered XAI, CoRIfEE-Rel addresses key limitations that rule-based explanations may contain features that are not well aligned with the central concepts of the application domain. By integrating domain-specific knowledge graphs and a relevance heuristic that balances semantic relevance with predictive accuracy, CoRIfEE-Rel provides a systematic approach to generate explanations that are not only accurate but also semantically relevant to the domain concept.

Experimental results across four datasets show that CoRIfEE-Rel improves the semantic relevance of explanations (as measured by the S1 and S2 scores) while maintaining a comparable level of accuracy to underlying models. The results provide evidence that CoRIfEE-Rel can improve graph-based semantic alignment while maintaining comparable predictive accuracy.

Nonetheless, future work includes incorporating user studies to empirically validate the cognitive and practical benefits of CoRIfEE-Rel, particularly in high-stakes, domain-specific contexts such as healthcare or finance, where expert interpretation is critical. Such evaluations help to determine whether improvements in semantic relevance enhance human understanding, trust, and decision-making. Human evaluation studies constitute another direction for future work and could provide valuable insights into the role of semantically relevant explanations in supporting human–XAI collaboration.

In addition, while we use WordNet in our experiments, the framework is not restricted to a particular semantic source. Future work will explore the use of domain-specific knowledge graphs and ontologies to further enhance semantic relevance in real-world applications.

Overall, CoRIfEE-Rel offers a promising direction for advancing human-centered XAI by tailoring rule-based explanations with the semantic structures that reflect human cognition.

Author Contributions: Conceptualization, P.M. and J.F.; Investigation, P.M.; Methodology, P.M.; Software, P.M.; Supervision, J.F.; Validation, P.M. and J.F.; Visualization, P.M.; Writing—Original Draft, P.M.; Writing—Review and Editing, P.M. and J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve human participants or human-subject research requiring ethical approval.

Informed Consent Statement: Not applicable. This study did not involve human participants or human participant data.

Data Availability Statement: The datasets analyzed in this study are publicly available from the UCI Machine Learning Repository. These include the Adult dataset, https://archive.ics.uci.edu/dataset/2/adult; Hepatitis dataset, https://archive.ics.uci.edu/dataset/46/hepatitis; Bank Marketing dataset, https://archive.ics.uci.edu/dataset/222/bank+marketing; and Water Quality dataset, https://archive.ics.uci.edu/dataset/733/water+quality+prediction-1 (accessed on 1 March 2025).

Acknowledgments: Supported by Johannes Kepler University Open Access Publishing Fund.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hanif, A.; Beheshti, A.; Benatallah, B.; Zhang, X.; Habiba; Foo, E.; Shabani, N.; Shahabikargar, M. A Comprehensive Survey of Explainable Artificial Intelligence (XAI) Methods: Exploring Transparency and Interpretability. In *Proceedings of the Web Information Systems Engineering—WISE, Melbourne, VIC, Australia, 26–27 October 2023*; Zhang, F.; Wang, H.; Barhamgi, M.; Chen, L.; Zhou, R., Eds., Springer: Singapore, 2023; pp. 915–925.

2. Thunki, P.; Reddy, S.R.B.; Raparthi, M.; Maruthi, S.; Babu Dodda, S.; Ravichandran, P. Explainable AI in Data Science—Enhancing Model Interpretability and Transparency. *Afr. J. Artif. Intell. Sustain. Dev.* **2021**, *1*, 1–8.
3. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
4. Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; Riedl, M.O. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In Proceedings of the 24th International Conference on Intelligent User Interfaces, New York, NY, USA, 17–20 March 2019; IUI '19; pp. 263–274.
5. Al-Ansari, N.; Al-Thani, D.; Al-Mansoori, R.S. User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. *Hum. Behav. Emerg. Technol.* **2024**, *2024*, 4628855. https://doi.org/10.1155/2024/4628855.
6. Kim, J.; Maathuis, H.; Sent, D. Human-centered evaluation of explainable AI applications: A systematic review. *Front. Artif. Intell.* **2024**, *7*, 1456486. https://doi.org/10.3389/frai.2024.1456486.
7. Ehsan, U.; Watkins, E.A.; Wintersberger, P.; Manger, C.; Kim, S.S.Y.; Van Berkel, N.; Riener, A.; Riedl, M.O. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 11–16 May 2024; CHI EA '24.
8. Hong, S.; Park, W. Developing user-centered system design guidelines for explainable AI: A systematic literature review. *Artif. Intell. Rev.* **2025**, *58*, 386. https://doi.org/10.1007/s10462-025-11363-y.
9. Kliegr, T.; Štěpán Bahník.; Fürnkranz, J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif. Intell.* **2021**, *295*, 103458. https://doi.org/https://doi.org/10.1016/j.artint.2021.103458.
10. Fürnkranz, J.; Kliegr, T.; Paulheim, H. On cognitive preferences and the plausibility of rule-based models. *Mach. Learn.* **2020**. *109*, 853–898.
11. Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* **2021**, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473.
12. Ehsan, U.; Riedl, M.O. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* **2024**, *5*, 100971.
13. de Bruijn, H.; Warnier, M.; Janssen, M. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Gov. Inf. Q.* **2022**, *39*, 101666.
14. Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M.T.; Weld, D. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 8–13 May 2021; CHI '21.
15. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38.
16. Rajabi, E.; Etminani, K. Knowledge-graph-based explainable AI: A systematic review. *J. Inf. Sci.* **2024**, *50*, 1019–1029. https://doi.org/10.1177/01655515221112844.
17. Tiddi, I.; Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **2021**, *302*, 103627. https://doi.org/10.1016/j.artint.2021.103627.
18. Wang, B.; Wang, X.; Sun, C.; Liu, B.; Sun, L. Modeling Semantic Relevance for Question-Answer Pairs in Web Social Communities. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 1230–1238.
19. Ruotsalo, T.; Hyvönen, E. A Method for Determining Ontology-Based Semantic Relevance. In *Proceedings of the Database and Expert Systems Applications, Regensburg, Germany, 3–7 September 2007*; Wagner, R., Revell, N., Pernul, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 680–688.
20. De Boni, M.; Manandhar, S. The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering. In Proceedings of the New Directions in Question Answering, Budapest, Hungary, 24–26 March 2003; pp. 138–144.
21. Costa, F.; Ouyang, S.; Dolog, P.; Lawlor, A. Automatic Generation of Natural Language Explanations. In Proceedings of the Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces, New York, NY, USA, 7–11 March 2018; IUI '18 Companion. https://doi.org/10.1145/3180308.3180366.
22. Colas, A.; Araki, J.; Zhou, Z.; Wang, B.; Feng, Z. Knowledge-Grounded Natural Language Recommendation Explanation. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Singapore, 7 December 2023*; Belinkov, Y., Hao, S., Jumelet, J., Kim, N., McCarthy, A., Mohebbi, H., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 1–15. https://doi.org/10.18653/v1/2023.blackboxnlp-1.1.
23. Breve, B.; Cimino, G.; Deufemia, V. Hybrid Prompt Learning for Generating Justifications of Security Risks in Automation Rules. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*. https://doi.org/10.1145/3675401.

24. Rao, J.; Liu, L.; Tay, Y.; Yang, W.; Shi, P.; Lin, J. Bridging the Gap between Relevance Matching and Semantic Matching for Short Text Similarity Modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019*; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5370–5381. https://doi.org/10.18653/v1/D19-1540.
25. You, F.; Zhao, S.; Chen, J. A Topic Information Fusion and Semantic Relevance for Text Summarization. *IEEE Access* **2020**, *8*, 178946–178953. https://doi.org/10.1109/ACCESS.2020.2999665.
26. Ma, S.; Sun, X.; Xu, J.; Wang, H.; Li, W.; Su, Q. Improving Semantic Relevance for Sequence-to-Sequence Learning of Chinese Social Media Text Summarization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 2, pp. 635–640.
27. Rhee, S.K.; Lee, J.; Park, M.W. Ontology-based semantic relevance measure. In *Proceedings of the First International Conference on Semantic Web and Web 2.0 in Architectural, Product and Engineering Design—Volume 294*; RWTH Aachen: Aachen, Germany, 2007; SW+W2.0'07, pp. 63–68.
28. Rhee, S.K.; Lee, J.; Park, M.W. Semantic relevance measure between resources based on a graph structure. In Proceedings of the 2008 International Multiconference on Computer Science and Information Technology, Wisła, Poland, 20–22 October 2008; pp. 229–236. https://doi.org/10.1109/IMCSIT.2008.4747244.
29. Setzu, M.; Guidotti, R.; Monreale, A.; Turini, F.; Pedreschi, D.; Giannotti, F. GLocalX - From Local to Global Explanations of Black Box AI Models. *Artif. Intell.* **2021**, *294*, 103457. https://doi.org/https://doi.org/10.1016/j.artint.2021.103457.
30. Vidal, T.; Schiffer, M. Born-again tree ensembles. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; ICML'20.
31. Huang, Q.; Mezzi, E.; Mutlu, O.; Kofinas, M.; Prasad, V.; Khan, S.A.; Ranguelova, E.; van Stein, N., Beyond the Veil of Similarity: Quantifying Semantic Continuity in Explainable AI. In *Explainable Artificial Intelligence*; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 308–331. https://doi.org/10.1007/978-3-031-63787-2_16.
32. Pesquita, C. Towards Semantic Integration for Explainable Artificial Intelligence in the Biomedical Domain. In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021)—Volume 5: HEALTHINF, INSTICC, Online, 11–13 February 2021; pp. 747–753. https://doi.org/10.5220/0010389707470753.
33. Donadello, I.; Dragoni, M., SeXAI: A Semantic Explainable Artificial Intelligence Framework. In *AIxIA 2020 – Advances in Artificial Intelligence*; Springer: Ber-lin/Heidelberg, Germany, 2021; pp. 51–66. https://doi.org/10.1007/978-3-030-77091-4_4.
34. Sartori, G.; Lombardi, L. Semantic Relevance and Semantic Disorders. *J. Cogn. Neurosci.* **2004**, *16*, 439–452. https://doi.org/10.1162/089892904322926773.
35. Anderson, J.R.; Crawford, J. *Cognitive Psychology and its Implications*; WH Freeman: New York, NY, USA, 1995.
36. Bransford, J.D.; Johnson, M.K. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *J. Verbal Learn. Verbal Behav.* **1972**, *11*, 717–726.
37. Sperber, D.; Wilson, D. *Relevance: Communication and Cognition*; Blackwell: Oxford, UK, 1995.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2017*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
39. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; NIPS'17, pp. 4768–4777.
40. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proceedings of the Workshop at International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
41. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
42. Jacovi, A.; Goldberg, Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386.
43. Ribeiro, M.T.; 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, 13–17 August 2016; KDD '16, pp. 1135–1144.
44. Lai, V.; Tan, C. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 17 January 2019; FAT* '19, pp. 29–38.
45. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *16*, 199 – 231.

46. Rosenberger, J.; Schröppel, P.; Kruschel, S.; Kraus, M.; Zschech, P.; Förster, M. Navigating the Rashomon Effect: How Personalization Can Help Adjust Interpretable Machine Learning Models to Individual Users. *arXiv* **2025**, arXiv:2505.07100. https://doi.org/10.48550/arXiv.2505.07100.
47. Mahya, P.; Fürnkranz, J. Extraction of Semantically Coherent Rules from Interpretable Models. In Proceedings of the 17th International Conference on Agents and Artificial Intelligence—Volume 1: IAI, INSTICC, Setúbal, Portugal, 23–25 February 2025; pp. 898–908.
48. Mahya, P.; Fürnkranz, J. Biasing Rule-Based Explanations Towards User Preferences. *Information* **2025**, *16*, 535. https://doi.org/10.3390/info16070535.
49. Dua, D.; Graff, C. UCI Machine Learning Repository. 2017. Available online: https://archive.ics.uci.edu/ (accessed on 1 March 2025).
50. Peng, J.; Zou, K.; Zhou, M.; Teng, Y.; Zhu, X.; Zhang, F.; Xu, J. An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients. *J. Med. Syst.* **2021**, *45*, 61. https://doi.org/10.1007/s10916-021-01736-5.
51. Rezk, N.G.; Alshathri, S.; Sayed, A.; El-Din Hemdan, E.; El-Behery, H. XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. *Bioengineering* **2024**, *11*, 1016. https://doi.org/10.3390/bioengineering11101016.
52. Fellbaum, C. *WordNet: An Electronic Lexical Database*; The MIT Press: Cambridge, MA, USA, 1998.
53. Wu, Z.; Palmer, M. Verbs Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Las Cruces, NM, USA, 27–30 June 1994; ACL '94, pp. 133–138. https://doi.org/10.3115/981732.981751.