

Extraction of Semantically Coherent Rules from Interpretable Models

Parisa Mahya¹  and Johannes Fürnkranz^{1,2} 

¹*Institute for Application-Oriented Knowledge Processing (FAW), Johannes Kepler University, Linz, Austria*

²*LIT Artificial Intelligence Lab, Johannes Kepler University, Linz, Austria*

fi

Keywords: Human-Centered Explainable AI, Interpretable Models, Inductive Rule Learning, Semantic Coherence.

Abstract: With the emergence of various interpretability methods, the quality of the interpretable models in terms of understandability for humans is becoming dominant. In many cases, interpretability is measured by convenient surrogates, such as the complexity of the learned models. However, it has been argued that interpretability is a multi-faceted concept, with many factors contributing to the degree to which a model can be considered to be interpretable. In this paper, we focus on one particular aspect, namely semantic coherence, i.e., the idea that the semantic closeness or distance of the concepts used in an explanation will also impact its perceived interpretability. In particular, we propose a novel method, Cognitively biased Rule-based Interpretations from Explanation Ensembles (CORIFEE-Coh), which focuses on the semantic coherence of the rule-based explanations with the goal of improving the human understandability of the explanation. CORIFEE-Coh operates on a set of rule-based models and converts them into a single, highly coherent explanation. Our approach is evaluated on multiple datasets, demonstrating improved semantic coherence and reduced complexity while maintaining predictive accuracy in comparison to the given interpretable models.


1 INTRODUCTION


Machine learning systems are increasingly used in various fields and have become capable of solving complex problems and making decisions in the real world. Models that achieve a strong predictive performance also tend to become increasingly more complex, as, e.g., exemplified by random forests or deep neural networks. Therefore, we need to deal with the trade-off between the performance of a machine learning model and its interpretability. As a result, a new field, eXplainable Artificial Intelligence (XAI) (Samek and Müller, 2019; Došilović et al., 2018), emerged with a focus on interpreting and explaining the behavior of black-box models. The goal is to provide or increase trust, confidence, and transparency, especially for high-stake decisions. The importance of explaining black-box models leads to numerous studies and research on proposing interpretable models and post-hoc explanation methods, which produce explanations in different formats such as rules, feature importance weights, etc.

Interpretability is a well-known concept, but nevertheless, there are hardly any precise mathematical

definitions for it (Linardatos et al., 2021). Among various definitions for interpretability, we single out (Doshi-Velez and Kim, 2017), who define interpretability as "the ability to explain or present in understandable terms to humans", and (Miller, 2019) where it is defined as "the degree to which a human can understand the cause of a decision." These human-centered views on explanations are often neglected in the current XAI literature, where explanations are typically assessed by their complexity (simpler explanations being perceived as more comprehensible), and their fidelity to the black-box model (i.e., the degree to which the white-box surrogate explanation coincides with the black-box model). However, even if these objectives are taken for granted, there remain often multiple diverse explanations to choose from, a phenomenon also known as the Rashomon effect (Müller et al., 2023).

Despite various definitions proposed for interpretability, it is a domain-specific concept, i.e., it could be different for different user groups, or in relation to the model, domain knowledge, target model, etc. Accordingly, the interpretability of a model's explanation can be measured by taking into account human knowledge and feedback or experiments with end-users as part of the assessment. Since human

^a  <https://orcid.org/0000-0002-5709-4074>

^b  <https://orcid.org/0000-0002-1207-0159>

feedback is not always available, interpretability is often measured with the complexity of the learned concepts. While this is important, it is also not the only relevant criterion. For example, it has been argued that different cognitive biases may contribute to the perceived interpretability of concepts and should thus be evaluated and possibly optimized in rule learning (Kliegr et al., 2021; Fürnkranz et al., 2020).

In this paper, we focus on one of these aspects, namely semantic coherence, i.e., the idea that a meaningful concept definition should be composed of conditions that are semantically similar to each other. In linguistics, semantic coherence refers to the sense relationships between propositions, units, and sentences in a text. Because of the existing relations, texts are logically and semantically consistent for readers and listeners. Our goal is to develop a method that is able to improve the semantic coherence of learned rule sets without significantly sacrificing accuracy. More precisely, we propose a method that generates a more coherent interpretable model from a pool of rule models while maintaining approximately the same level of accuracy. The results of our proposed method are evaluated by comparing the accuracy and a semantic coherence score, which measures how semantically related concepts are to each other, and indirectly specifies the understandability of the generated explanations and interpretable models to humans.

This article is organized as follows. Section 2 briefly reviews related work, Section 3 describes and reviews the semantic coherence concepts, definitions, and measurements, Section 4 describes our research goals and the used terminology, Section 5 the proposed CORIFEE-Coh method, and Section 6 discusses the results.

2 RELATED WORK

Even though the fundamental goal of XAI is to improve user understanding of the models, only a few recent studies explore user experiences with explanations and reveal some pitfalls (e.g. Ehsan and Riedl, 2024; de Bruijn et al., 2022). The results show that end users often find the generated explanations hard to use and the reasoning distracting and time-consuming (Lai et al., 2023; Xie et al., 2020; Bansal et al., 2021; Wang and Yin, 2021). Thus, recent work has focused on examining the effectiveness and acceptability of explanations by taking into account users' perception (Suffian et al., 2023), as well as their background knowledge and preferences. The idea of "putting humans in the loop" refers to human-centered approaches (Ehsan et al., 2022; Lai et al., 2023).

2.1 Semantic Coherence in XAI

There is only little prior research on enhancing the interpretability of rule-based models with a particular focus on improving their semantic coherence. Kiefer (2022) presents an architecture CaSE that uses semantic interrogations to provide meaningful and coherent explanations. The architecture combines with a modified version of LIME (Ribeiro et al., 2016) and enables semantic alignment between humans and machine learning via topic modeling techniques. The proposed method is applied to the feature-based interpretability methods by providing meaningful topics for a group of features. Gabriel et al. (2014) propose a variant of a separate-and-conquer rule learning algorithm focusing on semantic coherence. The work's emphasis on semantic coherence aligns with our approach; however, it developed a rule learning algorithm while our proposed method aims at improving the semantic coherence of multiple, existing rule-based explanations. Confalonieri et al. (2021) introduce TREPAN, which extracts decision trees from black-box models using ontologies and improves the understandability and interpretability of explanations using ontologies. A group of users tests the understandability of the extracted and enhanced explanations.

2.2 Extracting Explanations from Rule-Based Models

Our approach aims at extracting a more coherent explanation than the original ones from a set of interpretable models. This is related to research that focuses on improving interpretability by reducing the size of the learned models. Approaches include approximating a single tree from a random forest (Zhou and Hooker, 2016), which is interpretable and simplifies the prediction process, or extracting the best trees from a random forest (Khan et al., 2020) based on the trees' individual performance and their Brier case. Another study by Souza et al. (2022) focuses on the interpretability of decision trees using a novel new metric called explanation size. Other works focus on rule extraction from the trees in a random forest model. SIRUS (Stable and Interpretable RULE Set) (Bénard et al., 2021) aims at generating more stable and compact rules based on the frequency of the rules in a random forest. The most frequent rules are extracted as they represent strong and robust patterns. RF+HC (Mashayekhi and Gras, 2015) targeted the large number of trees generated in a random forest, and tried to reduce the number of rules such that the comprehensibility improves.

3 SEMANTIC COHERENCE

In this section, we dive into semantic coherence concepts and definitions and provide the techniques to measure the concept.

3.1 Concept and Definitions

The word coherence is originally based on the Latin verb "cohaerer" which means to stick or to connect together. Coherence has different meanings and definitions in various applications, and numerous studies focus on conducting studies on defining and evaluating coherence in the context (Skusa, 2006; Sanders, 1997; Bolte et al., 2003; Bresó-Pla et al., 2023). In linguistics, semantic coherence is defined as what makes a text semantically meaningful. In other words, it describes the connectivity in text as semantic consistency of phrases, units, synonyms, etc. Robert De Beaugrande describes coherence as "the continuity of senses" and "mutual access and relevance within a configuration of concepts and relations" (De Beaugrande and Dressler, 1981). In general, coherence is inherently a subjective measurement, and it is the outcome of a cognitive process that is related to the background knowledge of the targeted group of people. It expresses the ability to perceive meaningful relations between concepts and the knowledge that is available in the text, as well as logical connections between units. As a result, coherent semantic text can be easily read and understood by humans (Vakulenko et al., 2018).

Table 1: Fragments of coherent and incoherent text.
(a) coherent text.

The student wakes up early in the morning.
She gets ready.
Then she goes to school.

(b) incoherent text.

The student wakes up early in the morning.
She has a younger sister.
Then she goes to school.

Table 1 shows two examples of coherent and incoherent texts, which explain the morning routine of a student. In Table 1a, the main topic is the student and her habits of going to school, and the reader can understand the flow of the text. However, in Table 1b, the text is a combination of habits and personal information of the student, and the reader cannot easily understand the connections between the sentences.

This article focuses on the interpretability of the rules and explanations. As Thagard noted the importance of coherence for explanatory power, we evaluate the interpretability of learned rules by aiming to measure the coherence of the provided explanations. We operationalize the concept of coherence to the semantic similarity in text (Gabriel et al., 2014). The semantic coherence in rules implies that the conditions in a rule are consistent with each other, and all the rules are coherent with the existing background knowledge.

Table 2: Example of highly and lowly coherent rulesets.
(a) highly coherent ruleset.

```
Salary = high :-
    marital-status = married,
    relationship = wife.
Salary = high :-
    workclass = private,
    occupation = manager.
```

(b) lowly coherent ruleset.

```
Salary = high :-
    age >= 24.
    relationship = not-in-family,
    hours-per-week >= 46.
Salary = high :-
    marital-status = married,
    age >= 44,
    education = master.
```

Table 2 shows examples of coherent and incoherent rulesets. Ruleset 2a, describes the features of a person who earns a high salary. The first rule is defined upon two features, `marital-status` and `relationship`, which are related to the person's personal environment, and rule 2 reports the two features that are related to the person's working status. Since the features in the two rules describe two general characteristics of a person, it is easy for humans to understand the ruleset; therefore, it is considered a highly coherent ruleset. In contrast, in ruleset 2b, the first rule describes a person with a high salary with three features that refer to three main concepts and characteristics of a person, i.e., age, relationship, and economy. In the second rule, the features again describe the targeted person with three different characteristics. Thus, the ruleset has a lower coherence than the first ruleset.

The assumption behind this work is that all other things being equal—in particular, there is no differ-

ence in the discriminatory power of the rules—a more coherent rule is a more preferable explanation than a less coherent rule. Our goal is to reformulate previously learned multiple interpretable models into a semantically more coherent model.

3.2 Measuring Semantic Coherence

To quantitatively evaluate semantic coherence, we estimate the semantic similarity by measuring the degree of taxonomical proximity. Typically, there are two main approaches to measuring semantic similarity: (i) corpus-based methods measure similarity between concepts based on the information obtained from corpora, and (ii) knowledge-based measures estimate similarities using ontologies and knowledge graphs. In our work, we follow a knowledge-based approach. This section reviews some of the most known techniques to measure semantic similarity based on ontologies.

Recently, many works on semantic similarity have been based on the ontological representation of knowledge (Sánchez et al., 2012; Zhu and Iglesias, 2017). An ontology has been defined as “an explicit specification of a conceptualization of a domain” (Gruber, 1995), or “a concrete and formal representation of what terms mean within the scope in which they are used” (Hogan et al., 2021). Thus, ontologies and knowledge graphs allow us to define the semantic roots of the terms in a graph and, in that way, to reason about their semantic relations. A critical review of various definitions of knowledge graph can be found in (Ehrlinger and Wöß, 2016).

Formally, a *knowledge graph* \mathcal{KG} is defined as a directed labeled graph $\mathcal{KG} = (V, E)$ where V represents a set of *nodes* or vertices and E is a set of *edges* connecting the nodes. Fig. 1 illustrates an example of an ontology in which c_{root} is the root node and c_i and c_j (shown in green) are two concepts for which we need to estimate their semantic similarity. The *least common subsumer* (LCS) represented as c_{lcs} is the closest common ancestor between the two concepts, as shown in red in Fig. 1. $P_k(c_i, c_j)$, is the k -th possible *path* of all paths between the two concepts, and we use $|P_k|$ to denote the number of nodes in this path. The *length* $l(c_i, c_j) = \min_k |P_k(c_i, c_j)|$ is the shortest path between the two concepts, which will go through c_{lcs} , as shown in blue in Fig. 1. The *depth* $d(c_i)$ of a concept is the shortest path length from the root to the concept, i.e., $d(c_i) = \min_k |P_k(c_{root}, c_i)|$, where c_{root} is the root node of the ontology.

Different techniques have been proposed to measure the semantic similarity between two concepts on the basis of a knowledge graph:

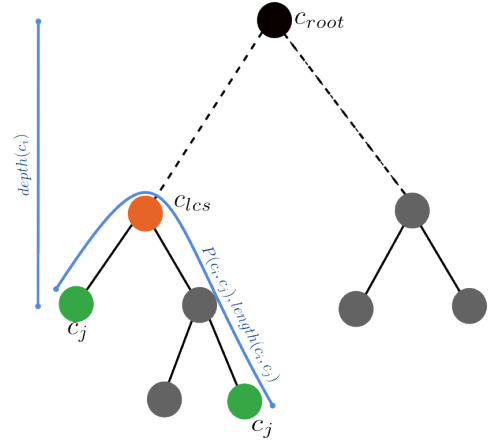


Figure 1: Illustration of ontology terminology.

- One of the simple metrics is the *path similarity* metric (1) which estimates the similarity of c_i and c_j as inversely proportional to their shortest path (Rada et al., 1989).

$$\text{sim}_{\text{path}}(c_i, c_j) = \frac{1}{1 + l(c_i, c_j)} \quad (1)$$

- The *Leacock&Chodorow* also known as *lch* semantic similarity metric (Fellbaum and Miller, 1998) is based on the shortest path between c_i and c_j and the maximum depth of taxonomy. It is calculated as

$$\text{sim}_{\text{lch}}(c_i, c_j) = -\log \frac{l(c_i, c_j)}{2 \cdot D} \quad (2)$$

where $D = \max(d(c_i), d(c_j))$ is the maximum depth of the taxonomy.

- the *Wu-Palmer* similarity metric (Wu and Palmer, 1994) is defined as

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

relating the depth of each concept c_i and c_j to the depth of their least common subsumer.

- the *Li* similarity metric (Li et al., 2003) is a parameterized method that allows to trade off the importance of the depth of the concepts and the length of their path. It is defined as

$$\text{sim}_{\text{li}}(c_i, c_j) = e^{-\alpha l(c_i, c_j)} \cdot \frac{e^{\beta d(c_{lcs})} - e^{-\beta d(c_{lcs})}}{e^{\beta d(c_{lcs})} + e^{-\beta d(c_{lcs})}} \quad (4)$$

where α and β are the two parameters that specify the contribution of the path length and the depth of the concepts.

In our experiments, we have tried various definitions, but have not found substantial differences between these for our work. In the following, we confine ourselves to reporting the results with the Wu-Palmer metric.

4 PROBLEM STATEMENT

The work in this article is in the context of a more general framework that aims to discover Cognitively biased Rule-based Interpretations from Explanation Ensembles (CORIFEE). CORIFEE is a meta-XAI method that takes multiple explanations in the form of rules as input, and generates new explanations that are more aligned to the user’s cognitive preferences. The problem context is inspired by the *Rashomon effect* (Breiman, 2001), which states that there are typically multiple models structured as rules or trees that explain the data equally well (in terms of a given performance measure), but often based their models on very different feature sets and may thus have very different semantical interpretations. It has also been argued that this phenomenon applies to multiple explanations in XAI (Müller et al., 2023).

In this work, we develop CORIFEE-Coh, an instantiation of CORIFEE, which focuses on semantic coherence and aims to generate a more coherent explanation from a pool of interpretable models. Since various interpretable models provide different explanations of the data, in our approach, we extract dominant features and rules from the pool of used features and re-assemble them to a new, more coherent explanation.

4.1 Terminology

In the remainder, we use terminologies and common terms described in the following. A *rule* is an if-then statement represented as r . It consists of a body (if-part) and a head (then-part). The body might have multiple conditions connected by conjunctions. The length of a rule is the number of conditions in the body. *Rule sets* are represented as $\mathcal{R} = \{r_1, \dots, r_k\}$. An *interpretable model* I is a model that can be inspected and its basic operation can be inspected in internalized by a human. In the context of this work, interpretable models are always rule sets, possibly generated from non-interpretable models by an explanation method. *Attributes* are the dimensions of a dataset \mathcal{D} , denoted as $\mathcal{A} = \{a_1, \dots, a_n\}$. Examples are characterized by specifying individual *values* $v_j^{(i)}$ for each of the attributes a_i . A combination of attribute and value, which can be evaluated as true or false for

any given example, is known as a *feature* f .

As the input for our method, we are given several rulesets \mathcal{R}_k , a dataset \mathcal{D} which has been used for training these rule sets, and a measure $s(\mathcal{R})$ for characterizing the semantic coherence of a ruleset. Our objective is to find a rule set \mathcal{R}' with increased semantic coherence, i.e., where $s(\mathcal{R}') > s(\mathcal{R}_k)$.

5 METHODOLOGY

This section is focused on elucidating the method’s functionality. As a general overview, the algorithm generates a semantically coherent explanation by selecting a subset of conditions from features that are clustered based on their semantic similarity. It comprises two main phases: $\text{PREPROCESS}(I)$ where we first prepare the data by cleaning, renaming, and filling the missing values, and then construct a *focused knowledge graph*, which extracts the parts that are relevant for the given dataset from a larger knowledge graph. This is then used as the basis for generating a set of semantically coherent rules from a given set of interpretable models.

5.1 Focused Knowledge Graph

For creating a focused knowledge graph that captures information about a dataset, we use WordNet and ConceptNet to extract the hypernym path of the nodes that correspond to the database attributes. We then use $\text{LABELEDNODES}()$ to create a subject-verb-object (svo) triple from the hypernym path in which the subject is the attribute and the object is the attribute’s value, both of them representing the nodes in \mathcal{KG} , and the verb specifies the relation using $\text{WEIGHTEDRELATIONS}()$ in Algorithm 1. The list of svo triples is further used to form a knowledge graph, in which a group of nodes is labeled as the attributes to form the concepts. Fig. 2 illustrates a simplified example of a knowledge graph for the *Adult* dataset in which the nodes in green color represent the attributes (concepts) in the dataset and the nodes in gray color depict the attributes’ values.

5.2 Coherent Rule Generation

Based on the focused knowledge graph, CORIFEE-Coh forms rules, initially starting from single-condition rules, where the single conditions are subsets selected from clustered features. The final rules consist of individual rules formed by adding conditions to optimize a trade-off between heuristic measurement and semantic similarity. To

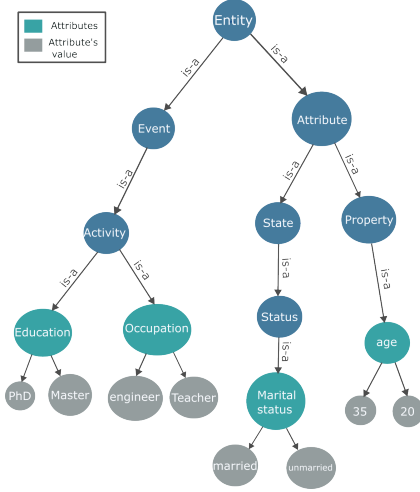


Figure 2: Example of a focused knowledge graph derived from the *Adult* dataset.

that end, it first applies the clustering method to the attribute nodes (green in Figure 2) to form clusters of attributes based on their distance in the knowledge graph. Using clusters in the explanation generation process leads to more general rule sets and ensures that the rules are based on the conceptual closeness.

The explanation generation procedure mainly consists of two steps: intracluster candidate generation and intercluster candidate generation. Utilizing clustering techniques and factoring in intercluster and intracluster aspects contributes to the semantic enrichment of the ultimate explanation.

The *intracluster candidate generation* finds the candidate conditions within each cluster for each class in the dataset. At this step, the algorithm starts by iterating into each cluster, and, for each cluster c_i , it gets all attribute nodes and their values and saves the resulting feature f_i in *clusterFeatures* as explained in Algorithm 1. It then iterates through each feature in *clusterFeatures* with an empty rule and constructs it by adding a condition that satisfies the coverage and precision criterion. The generated rule is appended to the \mathcal{R}_c .

The *intercluster candidate generation* step intends to generate highly semantic coherence rules \mathcal{R}' . It starts by iteratively getting all pairs of rules in \mathcal{R}_c and evaluates whether the merged rule is valid using the *MERGEVALIDATION* function in Algorithm 1. The function is mainly responsible for verifying the validity of the merged rules by assessing the feasibility of merged intervals for numerical features and avoiding any conflicting features in categorical features. In addition, it checks whether the merged rules belong to a class and whether H is maximized. The heuristic evaluation H forms a trade-off between accuracy and

Input: Pool of interpretable models I .
Output: A more coherent explanation.

```

    ▷ Creating a focused knowledge graph
     $I_p \leftarrow \text{PREPROCESS}(I)$ 
     $ns \leftarrow \text{LABELEDNODES}(I_p)$ 
     $rs \leftarrow \text{WEIGHTEDRELATIONS}(I_p)$ 
     $\mathcal{KG} \leftarrow \text{CREATEGRAPH}(ns,rs)$ 

    ▷ Cluster nodes using Wu-Palmer
     $C = \text{CLUSTER}(ns, \text{sim}_{\text{wup}})$ 

    ▷ Form Initial Rules

```

```

 $\mathcal{R}_c = []$ 
for  $c_i$  in  $C$  do
    clusterFeatures  $\leftarrow$  get all  $(V_j, V_k)$ 
    for  $f_i \in \text{clusterFeatures}$  do
        for  $l \in \text{CLASSES}$  do
            if  $\text{COVERAGE}(f_i, l) \geq th_c \wedge$ 
                $\text{PRECISION}(f_i, l) \geq th_p$  then
                 $r \leftarrow \{f_i \rightarrow l\}$ 
                 $\mathcal{R}_c \leftarrow \mathcal{R}_c \cup r$ 
            end
        end
    end
end
end

```

```

    ▷ Merge Rules
 $\mathcal{R}' \leftarrow []$ 
for each  $(r_i, r_j) \in \mathcal{R}_c^2$  do
    mergeValid  $\leftarrow$ 
         $\text{MERGEVALIDATION}(r_i, r_j)$ 
     $r \leftarrow \text{MERGE}(r_i, r_j)$ 
    if  $\text{mergeValid} \wedge H(r) \geq \beta$  then
         $\mathcal{R}' \leftarrow \mathcal{R}' \cup r$ 
    end
end
return  $\text{POSTPROCESS}(\mathcal{R}')$ 

```

Algorithm 1: CORIFEE-Coh.

explainability, defined as

$$H(\text{rule}) = (1 - \alpha) \cdot \text{DIS}(\text{rule}) + \alpha \cdot \text{COH}(\text{rule}) \quad (5)$$

in which the α parameter specifies the contribution of a conventional rule learning heuristic *DIS* and a measure for the semantic coherence *COH* in explanation generation. As a rule learning heuristic, we select the m-estimate (Džeroski et al., 1993), a generalization of the Laplace estimate, which has been shown to provide a tunable trade-off between precision, which tends to overfit, and weighted relative accuracy, which tends to over-generalize (Janssen and Fürnkranz, 2010). It is defined as

$$\text{DIS}(r) = \frac{p + m \cdot \frac{p}{p+n}}{p + n + m} \quad (6)$$

where p is the positive examples out of all positive examples P that are covered by the rule, n is the negative examples out of all negative examples N that are covered by the rule and m a user-settable parameter that realizes the above-mentioned trade-off.

For estimating the semantic coherence of a rule, we use the Wu-Palmer measure as defined above (3), which is a normalized similarity score, and estimates the average semantic similarity over all pairs of attributes (nodes labeled as "Attribute" in \mathcal{KG}) in the merged rules. As it considers the least common subsumer depth, it is well-aligned with how humans perceive similarity based on the shared meaning in a taxonomy, and it is computationally efficient compared to measurements such as Li semantic similarity which includes exponential factors. Consequently, we define the semantic coherence COH as

$$\text{COH}(r) = \frac{1}{L} \cdot \sum_{i=1}^{l-1} \sum_{j=i+1}^l \text{sim}_{\text{wup}}(a_i, a_j) \quad (7)$$

where a_i and a_j are two attributes in the rule of length l , and $L = \binom{l}{2}$ is the number of all pairs of attributes.

The parameter α in (5) allows to trade-off between the rule learning heuristic and the semantic coherence part. If $\alpha = 0$, the scoring method generates conditions by only considering the heuristic. As α increases, the semantic coherence gets more dominant, and the heuristic method decreases its importance. Accordingly, the conditions that satisfy a threshold for the scoring method are selected as the final conditions, which are part of the explanation.

As the final step, we perform the post-processing to generalize the rules by pruning the generated rules. This step applies to the final explanation \mathcal{R}' and is accomplished by iteratively evaluating conditions within each rule on a validation set and removing the conditions that do not worsen the error rate. By eliminating unnecessary rules, we simplify and generalize the rules and prevent overfitting.

6 RESULTS

In the following, CORIFEE-Coh is evaluated on multiple datasets, and the performance of the method is assessed in terms of accuracy, semantic similarity, and the number of found rules.

The four selected datasets are binary classification data from the UCI repository (Dua and Graff, 2017): The *Adult* dataset specifies whether a person earns more than 50K per year, the *Hepatitis* dataset contains the occurrence of hepatitis among people and determines whether they survive or die from it,

the *HeartDisease* dataset specifies whether the presence of heart disease and the *Titanic* dataset describes which passengers survived the Titanic disaster.

Since the main input to the method is a pool of interpretable models, we use random forests to generate a low number (typically 2 to 4) trees, convert each tree to a separate ruleset, and the generated rulesets constitute the pool of rules from which CORIFEE-Coh constructs a semantically coherent model.

6.1 Comparison to Random Forest

In this experiment, we train a random forest with a few trees in the dataset and use all the extracted rules from the trees to create the pool of interpretable models as CORIFEE-Coh input. The extraction of the rules is performed by iterating through each node in each decision tree in the random forest. In addition, we use the number of rules for a single rule-based model extracted from a random forest.

CORIFEE-Coh is then executed with different α parameters, and for each α parameter, the performance is evaluated through accuracy, semantic similarity, and the number of rules. The evaluation results of different α parameters are compared against the performance of random forest.

The results are shown in Fig. 3. For *Hepatitis*, as expected, the semantic similarity increases as semantic coherence contributes more to the scoring. The accuracy reaches its highest value in $\alpha = 0.8$. Because of the threshold defined for the scoring, the accuracy reported is the best value that can be achieved. The number of rules decreases and reaches its lowest value for $\alpha = 1$ where the scoring is purely based on the semantic score. By comparing the best result for $\alpha = 0.8$ and the result for the random forest, we see that the semantic similarity significantly improves, and the accuracy is nearly the same as the random forest, which indicates a good trade-off between interpretability and accuracy. CORIFEE-Coh is able to decrease the complexity substantially and, at the same time, increase the semantic coherence while maintaining a reasonable level of accuracy.

The results on *Adult* dataset reported in Fig 3 show that the semantic similarity is improved as the contribution of semantic coherence increases. The accuracy has its highest value for $\alpha = 0$. However, semantic coherence has no contribution to this α parameter, and semantic similarity has its lowest value. An acceptable balance between the accuracy and semantic similarity can be seen in $\alpha = 0.8$. By comparing the results from $\alpha = 0.8$ the random forest, we see that they both have quite comparable performance in terms of accuracy, but the semantic similar-

<i>Hepatitis</i>				
Algorithm	α	Accuracy	Semantic Similarity	Number of Rules
CoRI α EE-Coh	0	0.780	0.170	33
	0.2	0.735	0.290	28
	0.4	0.770	0.339	22
	0.6	0.729	0.231	18
	0.8	0.778	0.381	11
Random Forest		0.787	0.162	50

<i>Adult</i>				
Algorithm	α	Accuracy	Semantic Similarity	Number of Rules
CoRI α EE-Coh	0	0.815	0.208	62
	0.2	0.809	0.228	44
	0.4	0.768	0.238	39
	0.6	0.795	0.240	35
	0.8	0.803	0.383	30
1	0.752	0.405	15	
Random Forest		0.827	0.137	34553

<i>Heart Disease</i>				
Algorithm	α	Accuracy	Semantic Similarity	Number of Rules
CoRI α EE-Coh	0	0.796	0.074	35
	0.2	0.775	0.083	31
	0.4	0.788	0.097	30
	0.6	0.790	0.100	25
	0.8	0.8	0.145	14
1	0.776	0.222	9	
Random Forest		0.833	0.069	70

<i>Titanic</i>				
Algorithm	α	Accuracy	Semantic Similarity	Number of Rules
CoRI α EE-Coh	0	0.697	0.177	51
	0.2	0.708	0.193	45
	0.4	0.679	0.216	42
	0.6	0.682	0.210	40
	0.8	0.719	0.221	34
1	0.694	0.250	30	
Random Forest		0.759	0.129	272

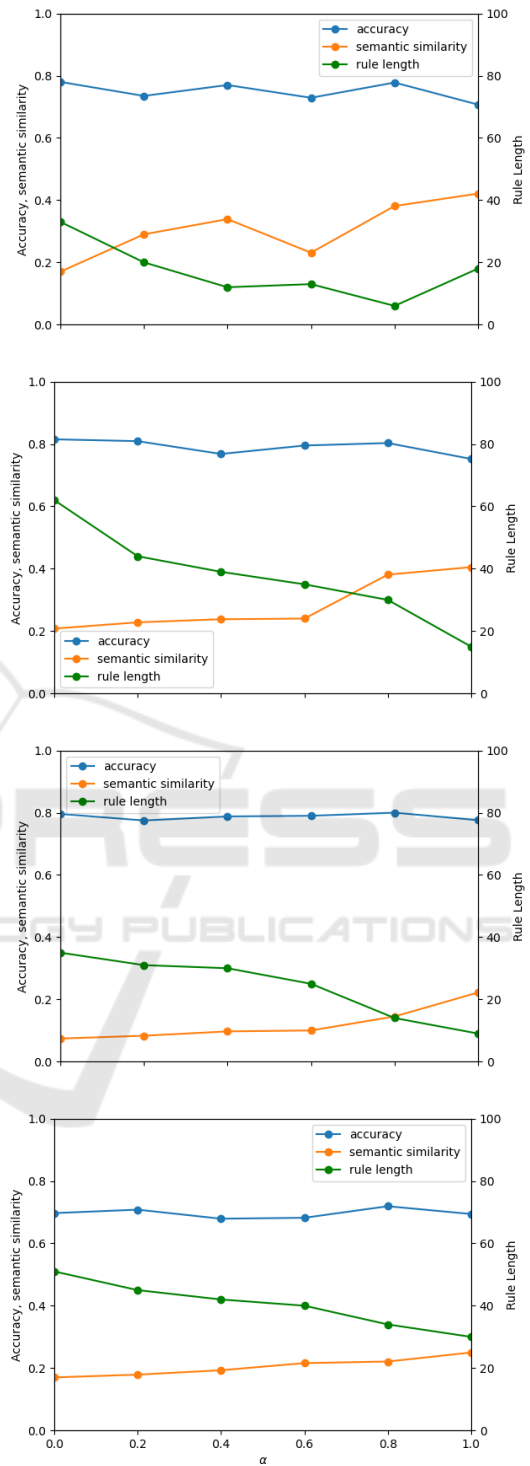


Figure 3: Results on the datasets (from top to bottom) *Hepatitis*, *Adult*, *Heart Disease*, and *Titanic*.

ity improves, and the number of rules decreases significantly.

The results on *Heart Disease* and *Titanic* datasets are similar, and follow the pattern that the semantic

coherence improves. In particular, they also confirm that the best trade-off between accuracy and semantic similarity is obtained in $\alpha = 0.8$ for both datasets.

Table 3 compares CoRI α EE-Coh with a fixed pa-

Table 3: Summary of the results on all the datasets.

Dataset	CORIFEE-Coh ($\alpha = 0.8$)			Random Forest		
	Accuracy	Semantic Similarity	Number of Rules	Accuracy	Semantic Similarity	Number of Rules
<i>Hepatitis</i>	0.872	0.381	11	0.901	0.162	5760
<i>Adult</i>	0.803	0.383	30	0.827	0.137	34553
<i>Heart Disease</i>	0.800	0.145	14	0.833	0.069	70
<i>Titanic</i>	0.719	0.221	34	0.759	0.129	272

parameter setting ($\alpha = 0.8$) to the random forest from which its rules were generated. We observe that the former gives a better trade-off between accuracy and interpretability, resulting in more coherent rulesets at the expense of reducing predictive accuracy.

6.2 Comparison to Individual Models

Since the pool of interpretable models consists of rules extracted from trees in random forest, in this section, we investigate the performance of each individual tree in the random forest model. Tables 4a and 4b report the accuracy and semantic similarity of the trees in a random forest on *Hepatitis* (3 trees) and *Heart Disease* (2 trees).

Table 4: Comparison to individual trees of a random forest on two datasets using the default value $\alpha = 0.8$.

(a) <i>Hepatitis</i>		
Tree	Accuracy	Semantic Similarity
CORIFEE-Coh	0.872	0.381
tree1	0.723	0.089
tree2	0.745	0.121
tree3	0.749	0.085

(b) <i>Heart Disease</i>		
Tree	Accuracy	Semantic Similarity
CORIFEE-Coh	0.800	0.145
tree1	0.766	0.079
tree2	0.755	0.063

The results of both datasets show that the accuracy and semantic similarity score are less than the results reported for $\alpha = 0.8$ in Fig. 3.

6.3 Sample Rules

As reported in previous sections, CORIFEE-Coh generates rule sets with considerably higher semantic similarity compared to the existing rule sets in the pool. This section will investigate the coherency of the rule sets learned by random forest and CORIFEE-Coh.

Table 5: Fragments of rule sets generated by CORIFEE-Coh and random forest on *Adult* dataset.

(a) rule sets generated by CORIFEE-Coh.	
1.	<code>income > 50k :- capital-gain >= 5119, marital-status = married-civ-spouse, relationship = husband.</code>
2.	<code>income <= 50K :- capital-gain <= 7585, capital-loss <= 2441, race = White, age <= 38.</code>

(b) rule sets learned by random forest.	
1.	<code>income > 50K :- hours-per-week >= 44.5, capital-gain >= 4307, capital-loss <= 2377, 32 <= age <= 39.5.</code>
2.	<code>income <= 50K :- marital-status = never-married, capital-gain <= 7073.5, age <= 33.5, capital-loss <= 2266.5, education = HS-grad, occupation = sales.</code>

Table 5 displays fragments of the rule sets generated by random forest and CORIFEE-Coh on *Adult* dataset. From the user's standpoint, a comparison of the results in ruleset 5a and ruleset 5b illustrates that the former is more coherent: each rule gives features related to a few related concepts, whereas the rules in the latter refer to various different concepts. For example, rule 1 in ruleset 5a describes the financial concept (capital-gain) and the marital status aspect (marital-status and relationship), while rule 2 in

ruleset 5b provides an explanation based on financial, marital status, working class, and career.

7 CONCLUSION

In this paper, we introduced CORIFEE-Coh, a novel method capable of generating coherent explanations that are more understandable for humans from a pool of interpretable models, such as the result of a random forest. The findings and results demonstrate that the method is able to generate a new explanation with considerably enhanced semantic coherence and fewer rules compared to the semantic coherence of the rulesets in the pool, while maintaining nearly the same accuracy as the underlying models in the pool.

Future studies could explore how the method can be tailored to match users' preferences, improving the understandability of the generated explanation according to the targeted users.

REFERENCES

- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). 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*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Bénard, C., Biau, G., Da Veiga, S., and Scornet, E. (2021). Interpretable random forests via rule extraction. In *International Conference on Artificial Intelligence and Statistics*, pages 937–945. PMLR.
- Bolte, A., Goschke, T., and Kuhl, J. (2003). Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. *Psychological Science*, 14(5):416–421. PMID: 12930470.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215.
- Bresó-Pla, V., Falkowski, A., González-Alonso, M., and Ivez Pozo, K. M. (2023). EFT analysis of new physics at COHERENT. *Journal of High Energy Physics*, 2023(5).
- Confalonieri, R., Weyde, T., Besold, T. R., and Moscoso del Prado Martín, F. (2021). Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471.
- De Beaugrande, R. and Dressler, W. (1981). *Introduction to Text Linguistics*. A Longman paperback. Longman.
- de Bruijn, H., Warnier, M., and Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2):101666.
- Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 210–215.
- Dua, D. and Graff, C. (2017). UCI machine learning repository.
- Džeroski, S., Cestnik, B., and Petrovski, I. (1993). Using the m -estimate in rule induction. *Journal of Computing and Information Technology*, 1:37–46.
- Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. In Martin, M., Cuquet, M., and Folmer, E., editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems (SEMANTiCS2016)*, volume 1695 of *CEUR Workshop Proceedings*, Leipzig, Germany. CEUR-WS.org.
- Ehsan, U. and Riedl, M. O. (2024). Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns*, 5(6):100971.
- Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., and Riedl, M. O. (2022). Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.
- Fellbaum, C. and Miller, G. (1998). *Combining Local Context and Wordnet Similarity for Word Sense Identification*.
- Fürnkranz, J., Kliegr, T., and Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 109(4):853–898.
- Gabriel, A., Paulheim, H., and Janssen, F. (2014). Learning semantically coherent rules. In *Proceedings of the 1st International Conference on Interactions between Data Mining and Natural Language Processing - Volume 1202, DMNLP'14*, pages 49–63, Aachen, DEU. CEUR-WS.org.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907–928.
- Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37.
- Janssen, F. and Fürnkranz, J. (2010). On the quest for optimal rule learning heuristics. *Machine Learning*, 78(3):343–379.
- Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., and Lausen, B. (2020). Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*, 14(1):97–116.

- Kiefer, S. (2022). Case: Explaining text classifications by fusion of local surrogate explanation models with contextual and semantic knowledge. *Information Fusion*, 77:184–195.
- Kliegr, T., Štěpán Bahník, and Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295:103458.
- Lai, V., Zhang, Y., Chen, C., Liao, Q. V., and Tan, C. (2023). Selective explanations: Leveraging human input to align explainable AI. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).
- Li, Y., Bandar, Z., and Mclean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1).
- Mashayekhi, M. and Gras, R. (2015). Rule extraction from random forest: the rf+hc methods. In Barbosa, D. and Milios, E., editors, *Advances in Artificial Intelligence*, pages 223–237, Cham. Springer International Publishing.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.
- Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauchhage, C., and Welke, P. (2023). An empirical evaluation of the rashomon effect in explainable machine learning. In Koutra, D., Plant, C., Rodriguez, M. G., Baralis, E., and Bonchi, F., editors, *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD): Research Track, Part III*, pages 462–478, Turin, Italy. Springer.
- Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, San Francisco, CA, USA. ACM.
- Samek, W. and Müller, K.-R. (2019). *Towards Explainable Artificial Intelligence*. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. CRC Press, Cham. ISBN 978-3-030-28954-6.
- Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, 24(1):119–147.
- Skusa, M. (2006). Semantic coherence in software engineering. In *ICEIS Doctoral Consortium, Proceedings of the 4th ICEIS Doctoral Consortium, DCEIS 2006, In conjunction with ICEIS 2006, Paphos, Cyprus, May 2006*, pages 118–129. ICEIS Press.
- Souza, V. F., Cicalese, F., Laber, E., and Molinaro, M. (2022). Decision trees with short explainable rules. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12365–12379. Curran Associates, Inc.
- Suffian, M., Stepin, I., Alonso-Moral, J. M., and Bogliolo, A. (2023). Investigating human-centered perspectives in explainable artificial intelligence. In *CEUR Workshop Proceedings*, volume 3518, pages 47–66. CEUR-WS.
- Sánchez, D., Batet, M., Isern, D., and Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9):7718–7728.
- Vakulenko, S., de Rijke, M., Cochez, M., Savenkov, V., and Polleres, A. (2018). Measuring semantic coherence of a conversation. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I*, volume 11136 of *Lecture Notes in Computer Science*, pages 634–651. Springer.
- Wang, X. and Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces, IUI '21*, page 318–328, New York, NY, USA. Association for Computing Machinery.
- Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL '94*, pages 133–138, USA. Association for Computational Linguistics.
- Xie, Y., Chen, M., Kao, D., Gao, G., and Chen, X. A. (2020). CheXplain: Enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Zhou, Y. and Hooker, G. (2016). Interpreting models via single tree approximation. *arXiv: Methodology*.
- Zhu, G. and Iglesias, C. A. (2017). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85.