A Global Model-Agnostic XAI method for the Automatic Formation of an Abstract Argumentation Framework and its Objective Evaluation

Giulia Vilone[1], Luca Longo[1]

[1]*The Artificial Intelligence and Cognitive Load research lab,*
The applied Intelligence Research Center, School of Computer Science,
Technological University Dublin, Dublin, Ireland

Abstract

Explainable Artificial Intelligence (XAI) aims to train data-driven, machine learning (ML) models possessing both high predictive accuracy and a high degree of explainability for humans. Comprehending and explaining the inferences of a model can be seen as a defeasible reasoning process which is expected to be non-monotonic meaning that a conclusion, linked to a set of premises, can be withdrawn when new information becomes available. Computational argumentation, a paradigm within Artificial Intelligence (AI), focuses on modeling defeasible reasoning. This research study explored a new way for the automatic formation of an argument-based representation of the inference process of a data-driven ML model to enhance its explainability by employing principles and techniques from computational argumentation, including weighted attacks within its argumentation process. An experiment was conducted on five datasets to test, in an objective manner, if the explanations of the proposed XAI method are more comprehensible than decision trees, which are considered naturally transparent. Findings demonstrate that usually the argument-based method can represent the logic of the model with fewer rules than a decision tree, but further work is required to achieve the same performances in terms of other characteristics, such as fidelity to the model.

Keywords

Explainable artificial intelligence, Argumentation, Non-monotonic reasoning, Method evaluation, Metrics of explainability

1. Introduction

XAI, a sub-field of AI, aims to develop a unified approach to learning data-driven models that are both highly accurate in their predictions and explainable to experts and laypeople. The explosion in the quantity of available data and the success of ML, especially Deep Learning, have led to the development of new models with outstanding predictive performances. However, most of these models have complex, non-linear structures that are hard to understand and explain. Researchers have proposed numerous XAI methods generating explanations in different formats (numerical, rules, textual, visual or mixed) [1, 2]. The XAI methods returning rule-based explanations extract

1st International Workshop on Argumentation for eXplainable AI (ArgXAI, co-located with COMMA '22), September 12, 2022, Cardiff, UK
giulia.vilone@tudublin.ie (G. Vilone)
0000-0002-4401-5664 (G. Vilone); 0000-0002-2718-5426 (L. Longo)

a set of rules mimicking the inferential process of a complex ML model [3]. However, these methods do not necessarily capture and describe the actual inferential process. They just report the relationships between inputs and outputs as learned by the model without verifying if they are consistent with the background knowledge of the application field or are instead based on spurious correlations of the data. Understanding the inferential process of a model should be seen as a non-monotonic reasoning process [4]. This requires a mechanism replicating the way human reasons to support humans in the comprehension of the inherent inferential process learnt by a model. Argumentation is a multidisciplinary subfield of AI that studies how arguments can be presented, supported or discarded in a defeasible reasoning process. It also investigates formal approaches to evaluate the validity of the conclusions reached at the end of the reasoning process[5, 6]. Argumentation Theory (AT) provides the basis for implementing these processes computationally [6] and it is inspired by how humans reason. This research experiment shows that AT can be a viable solution for building novel global model-agnostic XAI methods generating argument-based explanations. The quality of these explanations was preliminarily tested via an objective study based on eight quantitative metrics that assess distinct aspects of rule-based explanations, thus providing vital insights on the inferential process of a ML model [7], and compared to another rule-extraction XAI method generating Decision Trees (DTs), which are considered as naturally transparent [8, 3].

The remainder of this manuscript is organised as follows. Section 2 summarises the strategies used by scholars to generate rule-based explanations of ML models and how to assess the quality of these explanations. Section 3 describes the design of a primary research experiment. Section 4 discusses the findings of this experiment and its limitations. Lastly, Section 5 highlights the contribution to the existing body of knowledge and suggests future directions.

2. Related work

Rule-based explanations are a structured but still intuitive format for reporting information to humans in a compact way. They represent the logic of a ML model as a ruleset that can be easily read, interpreted and visualised. Therefore, scholars consider rulesets and DTs as naturally transparent and intelligible [8, 3]. However, current rule-extraction XAI methods merely produce a rulesets mimicking the inferential process of an underlying complex model. The rules can also be in conflict with the expert domain knowledge, thus perplexing the users of such models. It must remember that such rules aim at faithfully representing the relationships captured by the model during its training process between the independent variables of the input data with its target variable. Thus, this conflict can be an essential signal of an issue occurring during training. Similarly, the XAI methods do not provide any tool to handle potential inconsistencies among the extracted rules, should they arise. Thus, these rules are not suitable to support a richer reasoning process [9]. AT provides formal approaches to model non-monotonic logic and assess the validity of the conclusions reached by a set of arguments to be considered as acceptable [5, 6]. Non-monotonic logic consists of a family of formal frameworks devised to capture and represent defeasible inferences. In formal logic, a defeasible concept consists of a set of pieces of information or arguments that can be rebutted by additional information or arguments [10]. Generally, arguments are designed by domain experts to create a knowledge-base in single or

multi-agent environments [11]. In a single-agent environment, arguments are constructed by an autonomous reasoner, thus conflictual information tends to be minimal. In a multi-agent environment, multiple reasoners participate in argument construction, so more conflicts among them usually arise, enabling in practice non-monotonic reasoning [12]. Defeasible argumentation supplies a sound formalisation for reasoning with uncertain and incomplete information from a defeasible knowledge-base [13]. The process of defeasible argumentation frequently requires the recursive analysis of conflicting arguments in a dialectical setting to determine which arguments should be accepted or discarded [14]. Abstract AT (AAT) is the dominant paradigm, whereby arguments are abstractly considered in a dialogical structure. Formal semantics are habitually adopted to identify conflict-free sets of arguments that can subsequently support decision-making, explanations and justification [14, 6]. Existing AAT-based frameworks have common features: [13, 15, 16]:

- a defeasible knowledge-base in the form of interactive *arguments*, usually formalised with a first-order logical language;
- a set of *attacks* that are modelled whenever two arguments are in conflict;
- a *semantic* which consists of mechanism for conflict resolution. It implements in practice non-monotonicity and provides a dialectical status to the arguments.

The integration between AT and ML is still a young field. Minimal work exists on automatic argument and attack mining from data-driven ML models, how the interpretation of these models can be augmented via argumentation to, in turn, improve their explainability. [17, 13, 15]. In relation to this, the first issue is the automatic extraction of rules and their conflicts from these models. The second issue is their automatic integration into an argumentation framework that can serve as a mechanism for interpreting and explaining the inferential process of such models without any explicit human declarative knowledge. A two-step approach for AT-ML integration was proposed in [18]. In the first step, rules are extracted from a given dataset with the Apriori algorithm for mining association rules. In the second step, the rules are fed into structured argumentation approaches, such as ASPIC+ [19]. Using their argumentative inferential procedures, new observations are classified by constructing arguments on top of these rules and determining their justification status. Another study exploits argumentative graphs to depict the structure of argument-based frameworks [20]. Arguments are the nodes connected by directed edges representing attacks. The status of the arguments is provided by a label (accepted or rejected) and is determined by using argumentation semantics [21].

3. Design

The informal research hypothesis of this study is that a ruleset extracted by an XAI method from data-driven ML models supports the automatic formation of an argumentation framework. The expectation is that this framework possesses a higher degree of explainability when compared to other formats of explanations considered naturally interpretable and transparent in Computer Science, like a DT. The difference in the degree of explainability of the two methods was tested in an objective and quantitative manner with eight metrics that measure different aspects of a ruleset, such as number and length of its rules. The research hypothesis was tested by carrying out a set of phases described in the following paragraphs and depicted in the diagram of Fig. 1.

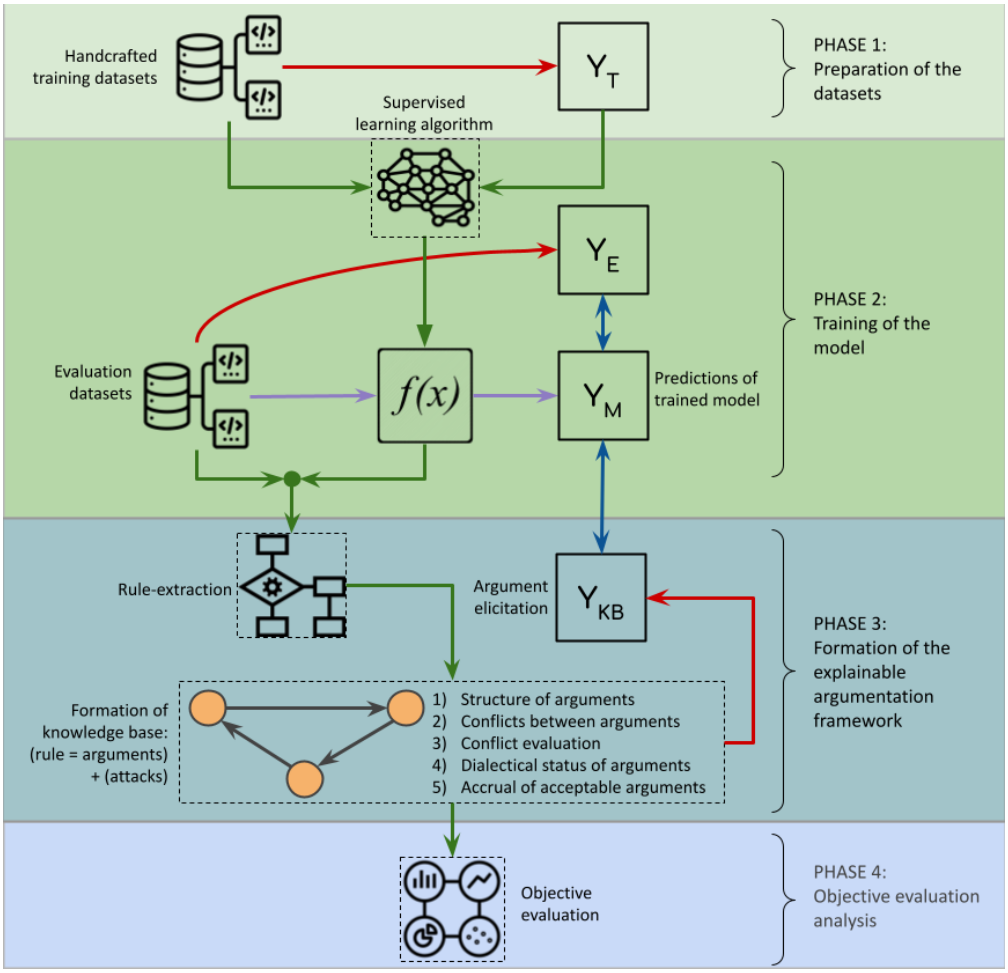


Figure 1: High-level representation of the process to build the envisioned argument-based XAI method.

3.1. Phase 1: Dataset preparation

The first step was to select a few training datasets containing multi-dimensional data built by domain experts, so they cannot contain data produced by an algorithm. The datasets must not present issues that can impede the successful training of a model, such as the course of dimensionality or a significant portion of missing data. The labeled target variable, represented by block Y_T in Fig. 1, must be categorical, ideally with more than two target classes, whereas the independent features should be a mix of continuous and categorical predictors. In this study, the experiment was carried out on five public datasets downloaded from Kaggleor the UCI Machine Learning Repository (see Tab. 1).The Adult database, based on the 1994 US Census, was designed to train ML models to predict if a person earns or not over $50K on annual basis. Avila contains data about 800 images of a Latin copy of the Bible, called the Avila Bible, manufactured during the XII century by 12 Italian and Spanish copyists who were individuated from a paleographic analysis of the manuscript. The model must associate each image with the copyist who drew

it. The Credit Card Default dataset was created to train ML models that predict if Taiwanese clients will fail to repay their credit card debts. The Hotel Bookings dataset includes booking information for a city hotel and a resort hotel such as the booking date, length of stay, and the number of adult and child guests, among other things. The target variable represents the final status of the reservation, whether it was cancelled, checked-out or the client did not show up. Online Shopper Intention records thousands of sessions on e-commerce websites. The negative target class represents customers who did not buy anything, whilst the positive class are sessions that ended with a purchase.

The datasets were preprocessed to avoid data-related issues in the model's training process. None of the selected datasets have missing data, so no action was required. However, the input features "fnlwgt" of the Adult dataset, which is the statistical weights measuring how many US citizens are represented by each subject, and the Client ID from the Credit Card Default dataset had to be discarded because they did not represent discriminative attributes. All the data in the independent features were scaled into the range [0,1]. Features with very large values might dominate over other in the training process of the model. Then, a correlation analysis was performed on each dataset to detect pairs of highly correlated features and discard one of the two to reduce the risk of multicollinearity. There is no consensus on the thresholds between strong, moderate and weak correlations. In this study, the absolute Spearman's rank correlation coefficients were grouped into three segments: values in the range (0, 0.33) were considered weak, (0.33, 0.66) moderate, and (0.66, 1) strong correlations. The best subset selection analysis was carried out to chose which variable from a strongly-correlated pair had to be discarded [22]. A linear regression model was built over each combination of the independent features excluding one from each strongly correlated pairs. These models were then sorted in descending order according to their R^2 values and the first one was selected. The best subset selection approach was chosen for its simplicity and because it requires little computational time and resources. Some of the chosen datasets are unbalanced, meaning that one specific class contains more instances than the others. This disparity can lead some learning algorithms to classify all the instances into the majority class and ignore the minority one. To avoid this, each dataset was split into a training and a validation subsets with the stratified five-fold cross-validation technique to ensure that each class was represented with the same proportion as in the original dataset. Furthermore, the Synthetic Minority Over-Sampling Technique (SMOTE) [23] was applied to the training datasets to up-sample the minority classes.

Table 1
Properties of the five datasets selected for the experiment.

Dataset	Total instances	No. of input features	No. of continuous (categorical) features	No. of classes
Adult	48,842	14	6 (8)	2
Avila	20,867	10	10 (0)	12
Credit Card Default	30,000	23	20 (3)	2
Hotel Bookings	119,385	23	16 (7)	3
Online Shopper Intention	12,330	17	14 (3)	2

3.2. Phase 2: Model training

A feed-forward neural network with two fully-connected hidden layers was trained on each datasets to fit Y_T. The block Y_M in Fig. 1 represents the predictions obtained from the trained model (represented by block $f(x)$) over the evaluation dataset (test data) whose original labelled target variable is depicted by block Y_E. Y_E is compared with Y_M to assess the model's prediction accuracy. The number of hidden nodes and the value of other model's hyperparameters, reported in Tab. 2, were determined with a grid search to reach the highest feasible prediction accuracy. To avoid overfitting, the training process was early stopped when the validation accuracy did not improve for five epochs in a row. The networks were trained five times over the five training subsets extracted from the datasets with the five-fold cross-validation technique. The models with the highest validation accuracy were chosen. Lastly, the not relevant input features were pruned by recursively removing one at a time, retraining the selected model and checking if its prediction accuracy decreased. If this was not the case, the pruned variable was removed.

Table 2
Optimal hyperparameters of neural networks obtained through grid search procedure, grouped by dataset, and their resulting accuracies.

Model parameters	Dataset list				
	Adult	Avila	Credit Card Default	Hotel Bookings	Online Shop. Intention
Optimizer	Adam	RMSprop	Adamax	SGD	SGD
Weight initialisation	Uniform	He-Unif.	Normal	Lecun-Unif.	He-Unif.
Activation function	Tanh	Relu	Softplus	Softplus	Softmax
Dropout rate	0%	0%	10%	0%	0%
Batch size	128	16	16	8	8
Hidden neurons	16	32	32	24	8
Accuracy (validation)	83% (79%)	98% (91%)	68% (79%)	65% (59%)	84% (87%)

3.3. Phase 3: Formation of the explainable argumentation framework

The trained models were translated into an explainable argument-based representation which can be easily embedded into an online interactive platform where the argumentation framework is represented as a graph (an example can be found in [4], page 10). The process of argumentation towards the achievement of a justifiable conclusion, as emerged from theoretical works of AT, can be broken down into five layers [6], as depicted in Fig. 2 and detailed in the following subsections.

Layer 1: definition of the internal structure of arguments. In standard logic, an argument consists of a set premises leading to a conclusion, or more formally:

Definition 3.1 (Argument)**.** *An argument Ar is a tentative inference* $\rightarrow$ *that links one or more premises* P_i *to a conclusion C and can be written as* $Ar : P_1, \ldots, P_n \rightarrow C$.

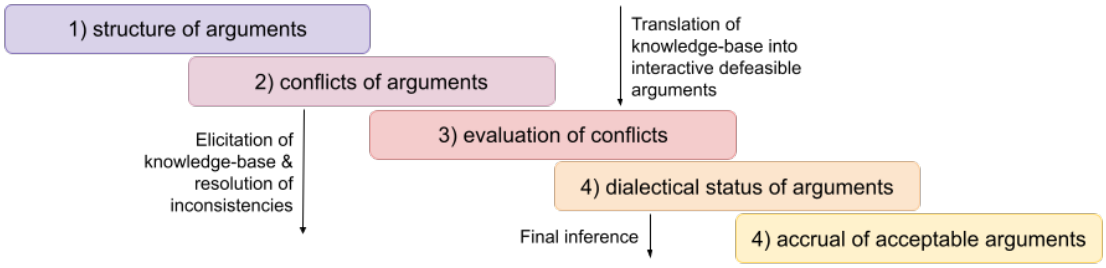


Figure 2: Five layers upon which argumentation systems are generally built, retrieved from [6].

In this study, an argument corresponds to an IF-THEN rule, thus the premises and conclusion of an argument correspond to the rule's antecedents and conclusion. The ML models and the evaluation datasets were fed into a bespoke rule-extraction method that generates a set of IF-THEN rules by using a two-step algorithm. First, each dataset was divided into groups according to the target class as predicted by the model. In other words, all the instances assigned by the model to the same class were grouped together. Second, the Ordering Points To Identify the Clustering Structure (OPTICS) [24] algorithm was exploited to further split the groups into clusters that coincide with areas of the input space having a high density of samples. Then, each cluster was translated into a rule by finding, for each relevant feature, the minimum and maximum values that include all the samples in the cluster. These ranges determine the rule's antecedents, whereas the conclusion corresponds to the predicted class of the cluster's samples. A typical rule is:

$$IF\, m_1 \leq X_1 \leq M_1\, AND \ldots AND\, m_N \leq X_N \leq M_N\, THEN\, Class_X \tag{1}$$

where X_i, $i = 1,\ldots,N$ are the N independent relevant features, m_i and M_i, $i = 1,\ldots,N$ are the minimum and maximum values w.r.t the i-th independent feature of the samples included in the cluster.

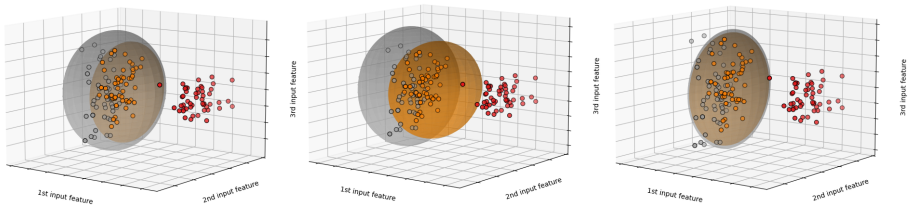
Layer 2: definition of the attacks between arguments. The inconsistencies between the formed arguments were modelled via the notion of *attack*. Generally, attacks are binary relations between two conflicting arguments. They can be of different kinds [6], but only the following two types were considered in this study.

Definition 3.2 (Rebutting attack)**.** *Given two distinct arguments $A, B \in AR$, where AR represents the set of all the arguments, with $A : P_1,\ldots,P_n \rightarrow C_1$, $B : P_1,\ldots,P_m \rightarrow C_2$, A is rebuttal of B and is denoted as (A,B) if C_1 logically contradicts C_2. A rebuttal attack is symmetrical, so it holds that iff (A,B), then $\exists(B,A)$.*

Definition 3.3 (Undercutting attack)**.** *Given an argument $A \in AR$ that challenges some or all of the premises used to construct another argument $B \in AR$, A undercuts B and is denoted as (A,B) when A claims there is a special case that does not allow the application of the inference rule $(\rightarrow)$ of argument B.*

Attacks are usually specified by domain experts, but in this study they can be automatically extracting by identifying conflicting rules. Two rules are conflictual if they are *overlapping* and reach different conclusions. Two rules overlap if their *covers* intersect. The cover of a rule

corresponds to the set of input instances whose attribute values satisfy the rule's antecedents [25]. As depicted in Fig. 3, two rules can be 1) fully overlapping, with one rule including the second one (part a), 2) partially overlapping (part b) or 3) sharing the same cover (part c). The first case could be seen as an undercutting attack because the internal rule represents an exception of the external one. The remaining two cases could be equivalent to a rebutting attack as two rules start from the same premises, at least in part, but reach different conclusions.



(a) Undercutting attack (b) Rebutting attack (c) Rebutting attack

Figure 3: Relative positions of two conflicting rules that can be a) fully overlapping, with one rule including the other, b) partially overlapping or c) covering the same area of the input space (retrieved from [4]).

Layer 3: evaluation and definition of valid attacks. Once arguments and attacks are embodied in a dialogical structure, the formalised knowledge-base, a fundamental characteristics of argument-based systems is their ability to determine the success of an attack. Different approaches can be found in the literature to decide if an attack is successful, thus valid, including a) binary attacks, b) strengths of arguments, and c) strengths of attacks [6]. In this study, a weighted notion of attack is considered; weights represent the strength of the attacks. There are various ways to compute these weights [26]. Here, they are computed as the percentage of instances belonging to the intersection of the covers of two conflictual rules that are assigned by the model to the same target class of the conclusion of the attacking rule:

$$w_{(A,B)} = \frac{|\{x \in cover(A) \cap cover(B) : f(x) = C_A\}|}{|\{x \in cover(A) \cap cover(B)\}|} \tag{2}$$

where x represents an input instance of the training dataset, C_A is the conclusion of the attacking rule (argument) A, and $|\bullet|$ is the cardinality function. For example, two conflicting rules have respectively target classes Q and S as conclusion and there are in their cover intersection 20 instances classified by the model in class Q and 30 in class S. In this case, the attack from the second rule with conclusion S is stronger than the attack from the first rule and has a weight equal to $\frac{30}{50}$. The weight of the reciprocal attack is $\frac{20}{50}$. It might happen that the difference in the number of instances per class is small, like 20 versus 21. In this case, is it fair to say that the rule with conclusion S is actually stronger than the other rule? As a consequence, the concept of *inconsistency budget* [26] was used to set a threshold on the fraction of supporting instances of the attacking rules. In this study, it was set equal to 0.55, meaning that an attack must be supported by at least 55% of the samples in the cover intersection. Future work will involve a study to fine-tune it. It is important to underline that all the arguments formed in layer one have the same importance and the notion of weight of argument is not used in this study. Not all the

arguments are activated by each training instance since not all their premises might be satisfied. The activated portion of the knowledge-base is considered for the next computations.

Layer 4: definition of the dialectal status of arguments. Dung-style acceptability semantics investigate the inconsistencies that might emerge from the interaction of arguments [14]. Given a set of arguments where some attack others, it must be decided which arguments can be accepted. In Dung's theory, the internal structure of arguments is not considered. This leads to an abstract argumentation framework (AAF) which is a finite set of arguments and attacks. In Dung's terms, usually, an argument defeats another argument if and only if it represents a reason against the second argument. It is also essential to assess whether the defeaters are defeated themselves to determine the acceptability status of an argument. This is known as *acceptability semantics*: given an AAF, it specifies zero or more conflict-free sets of acceptable arguments. However, other semantics have been proposed in the literature, not necessarily based on the notion of acceptability, such as the *ranking-base categoriser* semantic, introduced by [27] and employed in this experiment, which consists of a recursive function that rank-orders a set of arguments from the most to the list acceptable. The rank of an argument is inversely proportional to the number of its attacks and the rank of the attacking arguments. This semantic deems as acceptable the argument(s) with the lowest number of attacks and/or attacks coming from the weakest arguments.

Layer 5: Accrual of acceptable arguments. The previous layer produces a rank of activated arguments, and a final conclusion should be brought forward as the most rational conclusion associable to a single input instance. The highest-ranked argument is selected as the most representative, and its conclusion is deemed the most rationale. In the case of ties (multiple arguments with the highest rank), these are grouped into sets according to the conclusion they support. The set with the highest cardinality is deemed the most representative of an input record of the dataset, and the conclusion supported by its argument(s) is deemed the most rationale. In the case of ties with respect to cardinality, the input case is treated as undecided, as not enough information is available to associate a possible conclusion.

3.4. Phase 4: Objective evaluation analysis

The evaluation of the degree of explainability of the two XAI methods, the argument-based one developed in this study and the DT created with the C4.5 learning algorithm, followed the same process proposed in [7]. Eight metrics were selected to assess, objectively and quantitatively, the degree of explainability of their rulesets (see Tab. 3). The objectivity is achieved by excluding any human intervention in this evaluation process. Two metrics, *number of rules* and *average rule length*, measure the syntactic simplicity of the rules and should be minimised as short rulesets are deemed more interpretable [25]. *Fraction of classes* and *fraction of overlap* enhance the clarity and coherence of the extracted rules. Whilst the fraction overlap should be minimised to avoid conflicts between the rules, the fraction of classes should be maximised to guarantee that all the target classes, even the minor ones, are considered. A ruleset must also be *complete*, *correct*, *faithful* to the model's predictions, and *robust* to small perturbations of the inputs. To assure that the C4.5 algorithm returned the most compact and accurate DT, a grid search was carried out on the following hyperparameters: 1) the criterion function to measure the quality of a split (Gini,

Entropy, Log-Loss), 2) the maximum depth of the DT (from 6 to 48), and the minimum number of instances required to 3) split an internal node (from 2 to 16) and 4) be at a leaf node (from 1 to 8).

Table 3
Objective metrics to assess the explainability of rulesets.

Factor	Definition	Formula
Completeness	Ratio of input instances covered by rules (c) over total input instances (N)	$\frac{c}{N}$
Correctness	Ratio of input instances correctly classified by rules (r) over total input instances	$\frac{r}{N}$
Fidelity	Ratio of input instances on which the predictions of model and rules agree (f) over total instances	$\frac{f}{N}$
Robustness	The persistence of methods to withstand small perturbations of the input (δ) that do not change the prediction of the model ($f(x_n)$)	$\frac{\sum_{n=1}^{N} f(x_n) - f(x_n+\delta)}{N}$
Number of rules	The cardinality of the ruleset (A) generated by the two XAI methods under analysis	$\|A\|$
Average rule length	The average number of antecedents, connected with the AND operator, of the rules. a_i represents the number of antecedents of the i^{th} rule and $R = \|A\|$ the number of rules	$\frac{\sum_{i=1}^{R} a_i}{R}$
Fraction of classes	Fraction of the output class labels in the data are predicted by at least one rule in a ruleset R. A rule r is represented by a tuple (s,c) where s is the set of antecedents and c is a class label. $\|C\|$ represents the number of class labels	$\frac{1}{\|C\|}\sum_{c' \leq C} \mathbb{1}(\exists r = (s,c) \in R \| c = c')$
Fraction overlap	The extent of overlap between every pair of rules. Given two rules r_i and r_j, overlap is the set set of instances that satisfy the conditions of both rules	$\frac{2}{R(R-1)}\sum_{r_i,r_j,i\leq j} \frac{overlap(r_i,r_j)}{N}$

4. Results and discussion

The values of the metrics calculated over the two rulesets extracted with the C4.5 learning algorithm and the proposed argument-based XAI method are summarised in Tab. 4. Both methods generate complete rulesets, meaning that they cover the entire input space and all the output classes. The only exception occurs in the Avila dataset where the ruleset of the argument-based method does not consider one of the 12 output classes. This is due to the presence of several attacks, some of which have high weights, towards the rules having this class in their conclusions. Modifying the inconsistency budget might fix this issue. The C4.5 method scores higher in terms of correctness, fidelity and robustness throughout the five datasets. It can also be considered the most coherent method as its rulesets reach completeness without overlapping areas. On the other hand, it generated rulesets that contains more and longer rules than the argument-based method with only one exception represented by the Online Shopper Intention datasets where the argument-based method extracted more rules than the C4.5. However, these rules contained less antecedents, on average. In the other three datasets (Adult, Avila, and Hotel Bookings), the C4.5 returns thousands of rules whereas the argument-based method never reaches the 500

rules. Such big numbers of rules would hinder the explainability of these rulesets as struggle with reading and retaining such a big amount of information. Overall, the argument-based XAI method generates simpler ruleset that are potentially more comprehensible than the C4.5 DT, but there is the need to identify a way to fine tune the inconsistency budget to reach the optimal argumentation framework.

Table 4
Quantitative measures of the degree of explainability of the rulesets automatically generated by a novel argument-based XAI method and the C4.5 decision tree learning algorithm over five datasets.

Metric	**Adult**	**Avila**	**Credit Card Default**	**Hotel Bookings**	**Online Shopper Intention**
Argument-based XAI method					
Completeness	1.0	1.0	1.0	1.0	1.0
Correctness	0.7	0.55	0.65	0.64	0.89
Fidelity	0.81	0.52	0.8	0.52	1.0
Robustness	0.04	0.01	0.16	0.13	0.55
Number of rules	294	139	491	151	108
Average rule length	11.8	8.99	7.0	32.27	2.0
Fraction of overlap	0.9	0.64	0.87	0.99	0.15
Fraction of classes	1.0	0.92	1.0	1.0	1.0
C4.5 decision tree					
Completeness	1.0	1.0	1.0	1.0	1.0
Correctness	0.81	0.12	0.6	0.71	0.89
Fidelity	0.99	0.59	0.99	0.97	1.0
Robustness	0.99	0.60	0.95	0.96	0.98
Number of rules	4064	1614	686	6041	52
Average rule length	13.61	11.41	12.1	15.8	6.25
Fraction of overlap	0.0	0.0	0.0	0.0	0.0
Fraction of classes	1.0	1.0	1.0	1.0	1.0

5. Conclusions

This study presented a novel XAI method to form an argumentation framework with weighted attacks representing the inferential process of complex data-driven ML models. These models were trained on five datasets with handcrafted features manually engineered by humans. Eight quantitative and objective metrics were used to assess the degree of explainability of the rulesets extracted by the proposed XAI method and a DT, used as baseline. The results suggested the presence of a trade-off between completeness, number of rules and average length, measuring the syntactic simplicity of the rulesets, and the other five metrics. The C4.5 algorithm usually generate bigger rulesets, but it is more correct and faithful to the model than the argument-based method. In conclusion, the proposed XAI method returns rulesets that are complete, simpler and smaller in terms of rule cardinality and length, thus more comprehensible. However, they are not as faithful to the model, correct and robust as the C4.5 DTs. Future work will extend this research study by training deeper neural networks, employing datasets with additional types of input data, like texts and images, fine tuning the inconsistency budget between weighted attacks to obtain

the optimal set of arguments and attacks, and using semantics designed for handling weighted argumentation frameworks. The evaluation of the argumentation frameworks will include a human-centred study, as done in [4], to compare the outcome of the objective metrics with users' perception of their explainability.

References

[1] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2020, pp. 1–16.

[2] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Machine Learning and Knowledge Extraction 3 (2021) 615–661.

[3] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.

[4] G. Vilone, L. Longo, A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2022, pp. 447–460.

[5] D. Bryant, P. Krause, A review of current defeasible reasoning implementations, The Knowledge Engineering Review 23 (2008) 227–260.

[6] L. Longo, Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning, in: Machine Learning for Health Informatics, Springer, 2016, pp. 183–208.

[7] G. Vilone, L. Longo, A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods, Frontiers in artificial intelligence 4 (2021).

[8] H. K. Dam, T. Tran, A. Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, ACM, 2018, pp. 53–56.

[9] Z. C. Lipton, The mythos of model interpretability, Commun. ACM 61 (2018) 36–43.

[10] L. Longo, Formalising human mental workload as a defeasible computational concept, The University of Dublin, Trinity College, 2014.

[11] L. Rizzo, L. Longo, An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems, Expert Systems with Applications 147 (2020) 113220.

[12] L. Longo, L. Rizzo, P. Dondio, Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning, Knowledge-Based Systems 211 (2021) 106514.

[13] S. A. Gómez, C. I. Chesnevar, Integrating defeasible argumentation and machine learning techniques: A preliminary report, in: In Procs. V Workshop of Researchers in Comp. Science, 2003, pp. 320–324.

[14] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial intelligence 77 (1995) 321–357.

[15] S. Modgil, F. Toni, F. Bex, I. Bratko, C. I. Chesnevar, W. Dvořák, M. A. Falappa, X. Fan, S. A. Gaggl, A. J. García, et al., The added value of argumentation, in: Agreement technologies, Springer, 2013, pp. 357–403.

[16] S. A. Gómez, C. I. Chesnevar, Integrating defeasible argumentation with fuzzy art neural networks for pattern classification, Journal of Computer Science & Technology 4 (2004) 45–51.

[17] O. Cocarascu, F. Toni, Argumentation for machine learning: A survey., in: COMMA, 2016, pp. 219–230.

[18] M. Thimm, K. Kersting, Towards argumentation-based classification, in: Logical Foundations of Uncertainty and Machine Learning, IJCAI Workshop, volume 17, 2017.

[19] S. Modgil, H. Prakken, The aspic+ framework for structured argumentation: a tutorial, Argument & Computation 5 (2014) 31–62.

[20] R. Riveret, G. Governatori, On learning attacks in probabilistic abstract argumentation, in: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, 2016, pp. 653–661.

[21] P. Baroni, M. Caminada, M. Giacomin, An introduction to argumentation semantics, The knowledge engineering review 26 (2011) 365–410.

[22] R. R. Hocking, R. Leslie, Selection of the best subset in regression analysis, Technometrics 9 (1967) 531–540.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[24] H.-P. Kriegel, P. Kröger, J. Sander, A. Zimek, Density-based clustering, Wiley interdisciplinary reviews: data mining and knowledge discovery 1 (2011) 231–240.

[25] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2016, pp. 1675–1684.

[26] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, M. Wooldridge, Weighted argument systems: Basic definitions, algorithms, and complexity results, Artificial Intelligence 175 (2011) 457–486.

[27] P. Besnard, A. Hunter, A logic-based theory of deductive arguments, Artificial Intelligence 128 (2001) 203–235.