A Novel Human-Centred Evaluation Approach and an Argument-Based Method for Explainable Artificial Intelligence

Chapter · June 2022

DOI: 10.1007/978-3-031-08333-4_36

CITATION
1

READS
60

2 authors:

Giulia Vilone
Technological University Dublin - City Campus
13 PUBLICATIONS **125** CITATIONS

SEE PROFILE

Luca Longo
Technological University Dublin - City Campus
114 PUBLICATIONS **1,416** CITATIONS

SEE PROFILE

A Novel Human-Centred Evaluation Approach and an Argument-Based Method for Explainable Artificial Intelligence

Giulia Vilone(✉) and Luca Longo

The Artificial Intelligence and Cognitive Load Research Lab,
The Applied Intelligence Research Center, School of Computer Science,
Technological University Dublin, Dublin, Ireland
{giulia.vilone,luca.longo}@tudublin.ie

Abstract. One of the aim of Explainable Artificial Intelligence (XAI) is to equip data-driven, machine-learned models with a high degree of explainability for humans. Understanding and explaining the inferences of a model can be seen as a defeasible reasoning process. This process is likely to be non-monotonic: a conclusion, linked to a set of premises, can be retracted when new information becomes available. In formal logic, computational argumentation is a method, within Artificial Intelligence (AI), focused on modeling defeasible reasoning. This research study focuses on the automatic formation of an argument-based representation for a machine-learned model in order to enhance its degree of explainability, by employing principles and techniques from computational argumentation. It also contributes to the body of knowledge by introducing a novel quantitative human-centred technique to evaluate such a novel representation, and potentially other XAI methods, in the form of a questionnaire for explainability. An experiment have been conducted with two groups of human participants, one interacting with the argument-based representation, and one with a decision trees, a representation deemed naturally transparent and comprehensible. Findings demonstrate that the explainability of the original argument-based representation is statistically similar to that associated to the decision-trees, as reported by humans via the novel questionnaire.

Keywords: Explainable Artificial Intelligence · Argumentation · Human-centred evaluation · Non-monotonic reasoning · Explainability

1 Introduction

Explainable Artificial Intelligence (XAI), an emerging sub-field of Artificial Intelligence (AI), mainly aims to develop a unified approach to learning data-driven models with high prediction accuracy and a high degree of explainability. The explosion of data availability and the success of Machine Learning (ML) have led

Published by Springer Nature Switzerland AG 2022
I. Maglogiannis et al. (Eds.): AIAI 2022, IFIP AICT 646, pp. 447–460, 2022.
https://doi.org/10.1007/978-3-031-08333-4_36

to the fast development of models that can reach outstanding predictive performances. Unfortunately, most of these 'black-box' models have underlying complex structures that are difficult to comprehend and explain. Researchers have attempted to open up these black-boxes by developing numerous XAI methods that generate different formats of explanations (numerical, rules, textual, visual or mixed) [19,27]. However, the main criticism refers to the fact that these methods do not necessarily capture and describe the actual inference process of an ML model, and they merely report its outputs without attempting to verify if they are consistent with the user's domain knowledge or are instead based on spurious correlations of the data. Instead, we believe that understanding the inferential process of a model should be seen as a reasoning process that discloses the relationships between input and output. ML models are often built to discover "comprehensible, interesting knowledge (or patterns) from data" [11]. This means that a mechanism is necessary to support humans in the comprehension of the inherent learnt inferential process of a model. This mechanism should be aligned to the way human reasons. Argumentation is a multidisciplinary field within AI that focuses on how arguments can be presented, supported or discarded in a defeasible reasoning process, and it investigates formal approaches to assess the validity of the conclusions reached [4,17]. Argumentation Theory (AT) provides the basis for implementing defeasible reasoning computationally, inspired by how humans reason [17]. Our expectation is that argumentation can be a viable solution for XAI methods. This expectation was preliminarily tested via a human-centred study. This included the development of a questionnaire for explainability that was employed for comparing a traditional rule-base decision-tree and a novel argument-based explanation method.

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, Sect. 5 highlights the contribution to the existing body of knowledge and suggests future directions.

2 Related Work

Rule-based explanations represent a structured but intuitive format for conveying information to humans compactly. They can help disclose the logic of a quantitative model into a set of rules that can be read, interpreted and visualised. For this reason, scholars consider rule-based and tree-based models as naturally transparent and intelligible [7,9]. However, current methods from XAI generating rule-based explanations are usually limiting themselves to produce a list of rules mimicking the inferential process of an underlying model and not aggregated together to form a richer reasoning process [16]. Similarly, these methods do not focus on the potential inconsistencies among these rules and, should they arise, do not provide any tool to handle them [26]. Possible solutions to the above issues can be provided by Argumentation Theory (AT). This is a multidisciplinary field, inspired by how humans reason, that focuses on how

arguments can be presented, supported or discarded in a defeasible reasoning process. In formal logic, a defeasible concept consists of a set of pieces of information or reasons that can be defeated by additional information or reasons [18]. Technically speaking, AT focuses on modelling non-monotonic reasoning and it investigates formal approaches to assess the validity of the conclusions reached by arguments [4,17]. Arguments are usually designed by domain experts, forming a knowledge-base in single or multi-agent environments [24]. In a single-agent environment, arguments are often built by an autonomous reasoner, and often conflictual information tends to be minimal. In a multi-agent environment, multiple reasoners participate in argument building, and more conflicts among them are usually conceived, enabling in practice non-monotonic reasoning [20]. Defeasible argumentation can provide a sound formalisation for reasoning with uncertain and incomplete information from a defeasible knowledge-base [12]. The process of defeasible argumentation often involves the recursive analysis of conflicting arguments in a dialectical setting to determine which arguments should be accepted or discarded [10]. Abstract Argumentation Theory is the dominant paradigm, whereby arguments are abstractly considered in a dialogical structure, and formal semantics are usually adopted to partition these into conflict-free sets of arguments that can be subsequently used for supporting decision-making, explanations and justification [10,17]. Existing argument-based frameworks have some peculiar features [12,13,22]:

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

Minimal work exists on automatic argument mining from models generated by ML algorithms, and the integration between AT and ML is still a young field. [6,12,22]. A solution, based on a two-step approach, was proposed in [25]. First, rules were extracted from a given dataset with the Apriori algorithm for association rule mining. In the second step, these rules were interpreted as the input for structured argumentation approaches, such as ASPIC+ [21]. Using their argumentative inferential procedures, a new observation was classified by constructing arguments on top of these rules and determining their justification status. Alternatively, argumentative graphs were exploited to represent the structure of argument-based frameworks [23]. Arguments are treated as nodes connected by directed edges which are the attacks. The status of the arguments is given by a label (accepted or rejected), computed using argumentation semantics [2].

In summary, the literature review conducted showed that minimal work exists at the intersection of ML and AT, how models learnt can be exploited to augment their interpretation via argumentation and, in turn, improve their explainability. 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 and, successively, increase their explainability without any explicit human declarative knowledge.

3 Design

The informal research hypothesis is that the rules extracted from data-driven ML models by an XAI method support the automatic formation of an argumentation framework. This framework is expected to possess higher explainability when compared to a decision tree, another interpretable XAI method. Decision trees have been selected as a baseline since they are widely used within Computer Science because considered naturally intelligible and transparent [7,9]. The difference in the degree of explainability of the two methods was tested by considering each question of the survey separately and not by aggregating their answers. In this way, it was possible to determine which characteristics were discriminative. To test the research hypothesis, a set of phases are described in the following paragraphs and depicted in the diagram of Fig. 1.

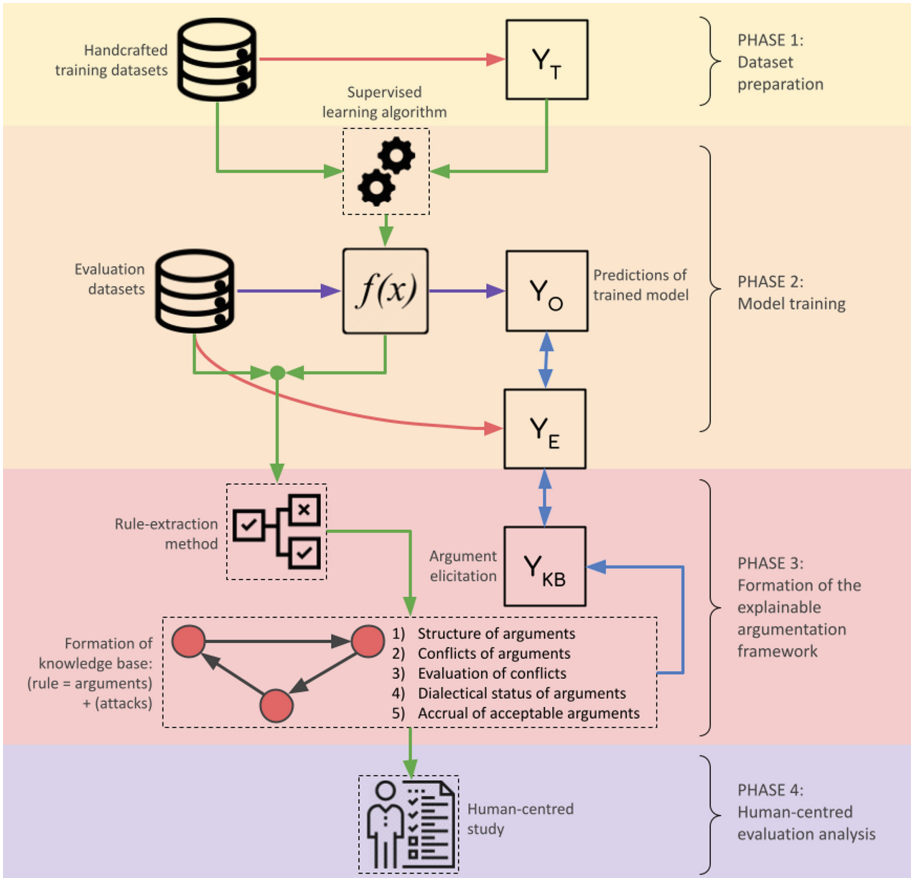


Fig. 1. High-level representation of the process to build the envisioned argument- and rule-based XAI method.

Phase 1: Dataset Preparation. The first step is to select a training dataset containing multi-dimensional data built by domain experts. The dataset must not present issues that can prevent the successful training of a model, such as the course of dimensionality or a significant portion of missing data. The block Y_T in Fig. 1 represents the labeled dependent variable. In this study, the experiment was carried out on the Airline Passenger Satisfaction dataset, publicly available from Kaggle's repository[1], which contains 129,880 records collected from an airline passenger satisfaction survey. The survey was designed to identify which factors lead to customer satisfaction for an airline, such as the quality of food and drinks, the passenger's satisfaction level with the check-in, and on-boarding services. The questionnaire contained 14 Likert-scale questions from 1 (very dissatisfied) to 5 (very satisfied), with a further option 0 meaning 'not-applicable', and four numeric questions related to the passenger's age, flight distance and departure/arrival delay in minutes. The remaining four questions were categorical and recorded the passenger's gender, the customer type (loyal/disloyal), the type of travel (personal/business) and the seating class (business/economy/economy plus). 393 records have missing data points. As they represent 0.3% of the entire dataset, these records were simply removed and not interpolated.

Phase 2: Model Training. Based on a supervised learning algorithm, a data-driven ML model is trained on the dataset to fit Y_T. The block Y_O in Fig. 1 represents the output obtained from the trained model (represented by block $f(x)$) over the evaluation dataset (test data). The block Y_E represents the original labelled dependent variables of the evaluation dataset. It is compared with Y_O to assess the model's evaluation accuracy. The architecture used in this study was a feed-forward neural network with two fully-connected hidden layers. The number of hidden nodes and the value of other hyperparameters, reported in Table 1, were determined by performing a grid search to reach the highest feasible prediction accuracy. An early stopping method was exploited to avoid overfitting during the training process by stopping it when the validation accuracy did not improve for five epochs in a row. In any case, the number of epochs was limited to 1000. The network was trained five times over five training subsets extracted from the Airline Passenger Satisfaction dataset with the five-fold cross-validation technique. The model with the highest validation accuracy was selected.

Phase 3: Formation of the Explainable Argumentation Framework. Once a model has been trained, it has to be translated into an explainable representation. In this study, an argumentation framework is formed, as described in the following five layers [17].

Layer 1: Definition of the Internal Structure of Arguments. In standard logic, an argument consists of a set premises linked to a conclusion. The selected trained model and the evaluation dataset are fed into a bespoke rule-extraction method

[1] https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction.

Table 1. Optimal hyperparameters of the neural network obtained through grid search procedure and the resulting prediction accuracy.

Model parameters	Value
Optimizer	Adam
Weight initialisation	Uniform
Activation function	Relu
Batch size	50
Hidden neurons	5
Loss function	Categorical cross-entropy
Accuracy (test set)	92.97% (92.55%)

returning a set of IF-THEN rules by using a three-step algorithm. This method prunes the not relevant input variables by recursively removing one at a time, retraining the model and checking if the prediction accuracy decreases. If this is the case, the pruned variable is removed, otherwise reinstated. Then, the evaluation dataset is split into groups according to the output class predicted by the model. This means that all the instances assigned by the model to the same class are grouped together. Finally, the Ordering Points To Identify the Clustering Structure (OPTICS) [14] algorithm was used to further divide the groups into clusters corresponding to areas of the input space with a high density of samples. Each cluster is translated into a rule by determining the two extreme samples for each relevant variable (thus, the minimum and maximum values that include all the samples in the cluster). The rule's antecedents correspond to these ranges, and the conclusion is the predicted class of the cluster's samples. A typical rule is presented below:

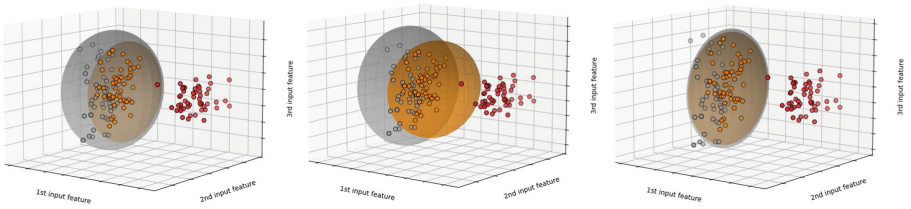
$$IF\, m_1 \leq X_1 \leq M_1\, AND \ldots\, AND\, m_N \leq X_N \leq M_N\, THEN\, Class_X \qquad (1)$$

where X_i, $i = 1, \ldots, N$ are the N independent relevant variables of the input dataset, m_i and M_i, $i = 1, \ldots, N$ are respectively the minimum and maximum values w.r.t the i-th independent variable of the instances included in the cluster. In this study, an argument coincides with an IF-THEN rule automatically generated by the rule-extraction method previously described. The premises and conclusion of an argument correspond to the rule's antecedents and conclusion.

Layer 2: Definition of the Attacks Between Arguments. Once arguments are formed, their inconsistencies are added via the notion of 'attack'. Usually, attacks are binary relations between two conflicting arguments, and they can be of different kinds. In this study, the following types were considered [17]:

- *rebutting attack* occurs when an argument negates the conclusion of another;
- *undercutting attack* occurs when an argument is attacked by arguing that there is a special case that does not allow its application.

Attacks are often specified by domain experts, and their automatic extraction is still an open research problem [8]. In this study, a novel method for automatically identifying conflicting rules was developed. In this new method, attacks were detected by checking if there were pairs of *overlapping* rules reaching different conclusions. Two rules overlap if there is an intersection area between their *covers*. The cover of a rule is the set of data points whose attribute values satisfy the rule's antecedents [15]. As shown in Fig. 2, two rules can be fully overlapping, with one rule including the second one (part a), partially overlapping (part b) or covering the same portion of the input space (part c). The first case could be seen as an undercutting attack as the first rule represents a particular case of the external rule. Partly and fully overlapping rules 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

Fig. 2. 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.

Layer 3: Evaluation and Definition of Valid Attacks. Once arguments and attacks have been represented in a dialogical structure, the formalised knowledge-base, an essential aspect of argument-based systems is the ability to determine the success of an attack. Different approaches are presented in the literature to determine a successful, thus valid, attack [17]. For example, these include a) binary attacks, b) strengths of arguments, and c) strengths of attacks. In this study, a binary notion of attack is considered; thus, the attacks that were automatically produced by the method described in layer 2 are all kept in the knowledge-base. However, for each input record, not all the arguments and attacks are activated since not all the premises are applicable. 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 [10]. Given a set of arguments where some attack others, a decision must be taken to determine 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. Here, 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-based semantics [1]. This study employed the *ranking-base categoriser* semantic, introduced by [3], 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 those 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 arguments, those fired out of the entire knowledge-base, and a final conclusion should be brought forward as the most rational conclusion associable to a single input record of the dataset. The highest-ranked argument is selected as the most representative, and its conclusion is deemed the most rationale for representing an input record of the dataset. 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 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.

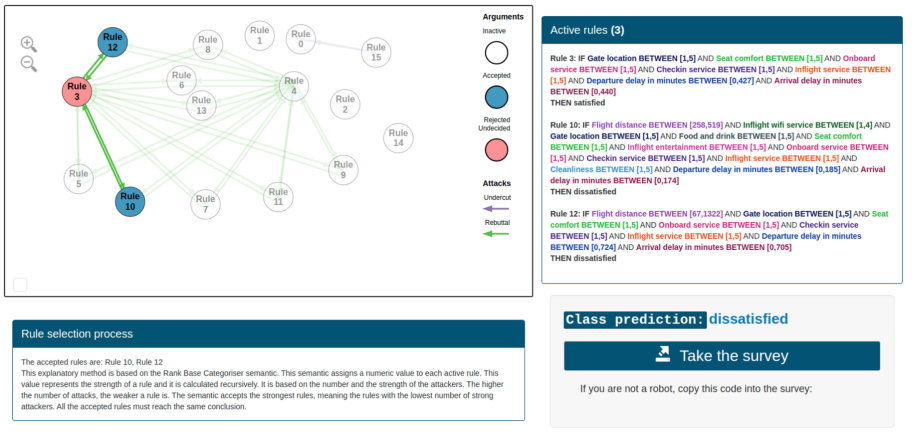
Phase 4: Human-Centred Evaluation Analysis. The degree of explainability of the proposed argument-based explanation method was evaluated involving human volunteers. Explainability is an ill-defined construct, and numerous notions underlying the construct of explanations exist [28]. In this study, a questionnaire aimed at measuring the explainability of this novel XAI method was developed by modelling a number of these notions (Table 2). The questionnaire was based on Likert-scales from 1 (strongly disagree) to 5 (strongly agree). Some of the questions were phrased negatively to minimise response and quiescence biases. The mix of positive and negative questions, presented in random order to each participant, should force the respondent to read them carefully and provide meaningful answers, thus reducing these biases [5]. The survey ended with an open-text question to collect suggestions for improving the XAI methods just used by volunteers. Five close demographic questions to collect background information about the respondents preceded the above questionnaire: their highest level of education, age, whether English is their first language, their knowledge of the airline industry and the AI technologies. Two groups of participants were randomly formed: one receiving the argument-based XAI method (Fig. 3, a, top) and another receiving a decision-tree XAI method, treated as a baseline as specified in the research hypothesis (Fig. 3, b, bottom). Both explanations methods contain an interactive

table with a subset of the data from the selected Airline Passenger Satisfaction dataset (Fig. 3, bottom). Participants could select an instance from this table to check which rules (and attacks) were fired and the final inference produced by the XAI method. Participants could spend as much time as they wished to familiarise themselves with the XAI method and the other components of the platform. Once satisfied, they could progress with the survey.

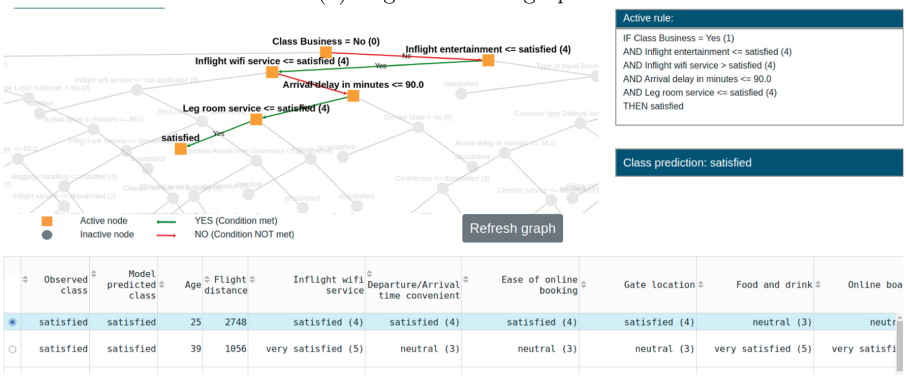
Table 2. Questions of the survey designed to assess the degree of explainability of an XAI method according to a set of notions.

#	Question	Assessed notion
1	I have learned something from the XAI method	Actionability
2	The XAI method taught me nothing new	Actionability
3	The relationship between input data & the predictions is clear	Causality
4	The relationship between input data & the predictions is vague	Causality
5	No rules in the XAI method return surprising predictions	Cognitive relief
6	The structure of the XAI method is not clear	Simplicity
7	The XAI method can be understood quickly	Explicitness
8	The XAI method takes a long time to understand	Explicitness
9	The XAI method is understandable	Understandability
10	The XAI method is incomprehensible	Comprehensibility
11	The XAI method provides useful information	Informativeness
12	The XAI method is not informative	Informativeness
13	External support was required to understand the XAI method	Intelligibility
14	The XAI method is engaging	Interestingness
15	The XAI method is not interesting	Interestingness
16	The XAI method allows me understand the ML model's logic	Mental fit
17	The ML model returns accurate predictions for all reasonable inputs	Algorithmic transparency
18	The XAI method makes me mistrust the model	Persuasiveness
19	The XAI method only includes most relevant data variables	Simplification

The final step was aimed at testing the research hypothesis by assessing, with the non-parametric Mann-Whitney U statistical test, if there are statistically significant differences in the degree of explainability of the two proposed XAI methods. The Mann-Whitney U test checks if two samples come from the same distribution. Alternatively, it tests if a cumulative distribution first-order stochastically dominates the other one, meaning that it assigns higher probabilities to the larger values. In this case, this means that the distribution of the responses related to the argumentation graph contains more positive answers ("agree" or "strongly agree") than the distribution related to the decision tree. The Mann-Whitney U test was preferred to the parametric hypothesis tests, such as the t-Student, because of the nature of the data. As they come from Likert-scale questions, it is impossible to assume that they follow the normal statistical distribution required by these parametric tests.



(a) Argumentation graph



(b) Decision tree

Fig. 3. Screenshots of the two alternative XAI methods embedded in the online platform used to carry out the evaluation survey.

4 Results and Discussion

The survey was promoted among authors' acquaintances and colleagues. It was carried out during the end of 2021 and the beginning of 2022 when Ireland was experiencing a surge of COVID-19 cases, so the only way to contact people was via online tools (emails and chats). This harmed the response rate. As a result, only 39 people completed the survey, of which 19 were presented with the decision tree and 20 with the argument graph. The majority of the participants were 25–34 and 35–44 years old with a Master's or Doctorate degree and 4+ years of experience with AI technologies, but with limited knowledge of the airline industry; 23 participants were not native English speakers (see Fig. 4).

First, the Cronbach's Alpha test was exploited to check the reliability of the explainability survey (Table 2) grouped by the two XAI methods. The alpha

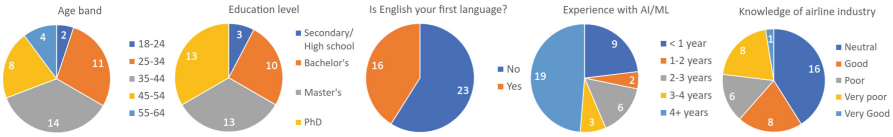


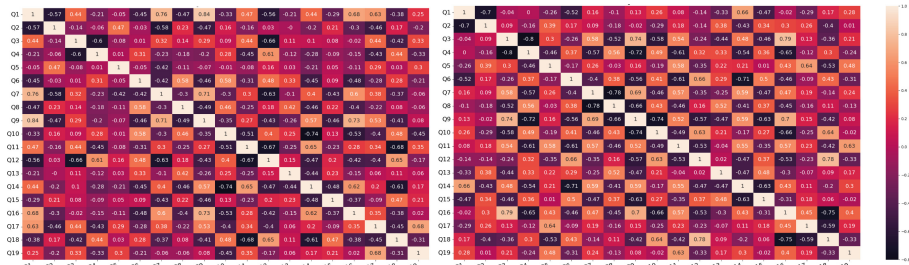
Fig. 4. Distributions of the responses given by participants to the questions related to their demographic and knowledge background.

coefficients were 0.88% for the argumentation graph and 0.9% for the decision tree, suggesting that the two surveys had high reliability. The Spearman-rank correlation coefficients were computed for each pair of the 19 Likert-scale questions to check that two notions of explainability, expected to be theoretically related, are in fact related. In particular, two questions assessing the same notion but worded differently (usually one positively and the other negatively) should be strongly correlated. For example, the first and second questions should be inversely correlated as both were designed to measure the actionability of the two XAI methods, but positively and negatively, respectively. As noticeable from the correlation matrices depicted in Fig. 5, these two questions are indeed inversely correlated in both surveys with Spearman coefficients of -0.57 (argumentation-graph) and -0.7 (decision tree). The same occurs for the other pairs: questions number 3–4, 11–12, and 14–15.

The Mann-Whitney U test returned p-values lower than the typical tolerance level of 5% for two questions: 1) 'the structure of the XAI method is not clear' (q.6) (p-value 2,19%), and 2) 'the XAI method takes a long time to understand' (q.8) (p-value 1,49%). The Mann-Whitney U test did not return any significant pieces of evidence to support the alternative hypothesis for the remaining 17 questions, meaning that there are no statically significant differences in the distributions of the responses given to these questions (see Fig. 6).

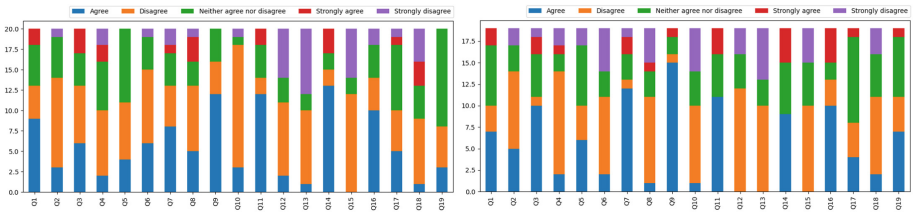
Overall, it is not possible to say that the argumentation graph was perceived neither as less nor more explainable than the decision tree, as only two questions out of 19 showed a statistically significant difference in the distribution of their answers. Furthermore, the survey was answered by a limited number of participants, and many of them have at least four years of experience with AI and ML. It is likely that they were familiar more with decision trees rather than with argumentation frameworks. However, the fact that the argumentation framework received the same scores, on average, across questions is indeed a positive outcome for its explainability. In fact, its graph-like visualisation represents the rules in a compact format as each rule is fully contained in a node and is expandable, whereas a rule in the decision tree can be represented as a long chain of nodes and edges. Decision trees always produce a set of conflict-free rules as their ML algorithms perform a series of binary splits of the input space. This is not necessarily an advantage as it returns large rulesets in terms of the number of rules. Such a case happens if there are many areas where a small perturbation of the input leads to a different prediction of the model. The decision tree needs many rules to capture all the input-output combinations.

By allowing conflicts, the argumentation framework requires, instead, a few overlapping rules with their attacks to describe the model's behaviour. The accrual of arguments determines, case by case, the most rational conclusion. In fact, the argumentation framework of this study contained 16 rules, whereas the decision tree was made by 60 rules (see Fig. 3).



(a) Argumentation graph (b) Decision tree

Fig. 5. Correlation matrices of the 19 Likert-scale questions assessing the degree of explainability of the (a) argument-based and (b) decision-tree XAI methods.



(a) Argumentation graph (b) Decision tree

Fig. 6. Distributions of the answers of the questionnaire assessing the degree of explainability of the (a) argument-based and (b) decision-tree XAI methods.

5 Conclusions

This study proposed a novel XAI method to extract IF-THEN rules automatically from ML models and treat them as arguments in the form of premises to a conclusion. Principles from AT were exploited to create an argumentation framework employing the notions of arguments and attacks among them. A novel method to automatically generate undercutting and undermining attacks among extracted arguments was devised. The hypothesis was that this argumentation framework could improve the degree of explainability of 'black-box models', namely trained neural networks with an Airline Passenger Satisfaction dataset. Two interactive interfaces were built: one for the argumentation framework and one for a decision tree which was treated as a baseline since often deemed highly

interpretable and explainable. The hypothesis was tested by designing and running a Likert-scale questionnaire of 19 questions to measure various notions related to the concept of explainability over the two interactive interfaces. Two groups of participants answered the survey after interacting with one of the two interfaces. The Mann-Whitney U test verified if the distribution of the responses related to the explainability of the argumentation framework stochastically dominated the distributions of the responses related to the explainability of the decision tree. The test did not reveal evidence to support the hypothesis of the superiority of the explainability of the argumentation-based representation over the decision tree. However, it did not also underperform the explainability of the decision trees, showing its appealing properties and characteristics for the interpretability and comprehensibility of machine-learned models. Future work will include the improvement of the internal computational mechanisms associated with the argumentation framework, particularly the definition of the valid attacks among arguments by employing gradualism, the application of other semantics for producing the dialectical status of arguments, and the strategies for their accrual in order to promote rational conclusions.

References

1. Amgoud, L., Ben-Naim, J.: Ranking-based semantics for argumentation frameworks. In: Liu, W., Subrahmanian, V.S., Wijsen, J. (eds.) SUM 2013. LNCS (LNAI), vol. 8078, pp. 134–147. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40381-1_11
2. Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. Knowl. Eng. Rev. **26**(4), 365–410 (2011)
3. Besnard, P., Hunter, A.: A logic-based theory of deductive arguments. Artif. Intell. **128**(1–2), 203–235 (2001)
4. Bryant, D., Krause, P.: A review of current defeasible reasoning implementations. Knowl. Eng. Rev. **23**(3), 227–260 (2008)
5. Choi, B.C., Pak, A.W.: Peer reviewed: a catalog of biases in questionnaires. Prevent Chronic Disease **2**(1), 1 (2005)
6. Cocarascu, O., Toni, F.: Argumentation for machine learning: a survey. In: COMMA, pp. 219–230 (2016)
7. Dam, H.K., Tran, T., Ghose, A.: Explainable software analytics. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, pp. 53–56. ACM, Gothenburg, Sweden (2018)
8. Dejl, A., et al.: Argflow: a toolkit for deep argumentative explanations for neural networks. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1761–1763 (2021)
9. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0210–0215. IEEE (2018)
10. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995)
11. Freitas, A.A.: Are we really discovering interesting knowledge from data. Expert Update (BCS-SGAI Mag.) **9**(1), 41–47 (2006)

12. Gómez, S.A., Chesnevar, C.I.: Integrating defeasible argumentation and machine learning techniques: a preliminary report. In: Proceedings of Workshop of Researchers in Computer Science, pp. 320–324 (2003)
13. Gómez, S.A., Chesnevar, C.I.: Integrating defeasible argumentation with fuzzy art neural networks for pattern classification. J. Comput. Sci. Technol. **4**(1), 45–51 (2004)
14. Kriegel, H.P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. Wiley Interdisc. Rev. Data Mining Knowl. Disc. **1**(3), 231–240 (2011)
15. Lakkaraju, H., Bach, S.H., Leskovec, J.: 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, pp. 1675–1684. ACM, San Francisco, California, USA (2016)
16. Lipton, Z.C.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2018)
17. Longo, L.: Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In: Holzinger, A. (ed.) Machine Learning for Health Informatics. LNCS (LNAI), vol. 9605, pp. 183–208. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50478-0_9
18. Longo, L.: Formalising human mental workload as a defeasible computational concept. Ph.D. thesis, Technological University Dublin (2014)
19. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: concepts, applications, research challenges and visions. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2020. LNCS, vol. 12279, pp. 1–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_1
20. Longo, L., Rizzo, L., Dondio, P.: Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning. Knowl. Based Syst. **211**, 106514 (2021)
21. Modgil, S., Prakken, H.: The aspic+ framework for structured argumentation: a tutorial. Argum. Comput. **5**(1), 31–62 (2014)
22. Modgil, S., et al.: The added value of argumentation. In: Ossowski, S. (eds) Agreement Technologies. Law, Governance and Technology Series, vol. 8, pp. 357–403. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-5583-3_21
23. Riveret, R., Governatori, G.: On learning attacks in probabilistic abstract argumentation. In: Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems, pp. 653–661 (2016)
24. Rizzo, L., Longo, L.: An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems. Expert Syst. App. **147**, 113220 (2020)
25. Thimm, M., Kersting, K.: Towards argumentation-based classification. In: Logical Foundations of Uncertainty and Machine Learning, IJCAI Workshop, vol. 17 (2017)
26. Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093 (2020)
27. Vilone, G., Longo, L.: Classification of explainable artificial intelligence methods through their output formats. Mach. Learn. Knowl. Extract. **3**(3), 615–661 (2021)
28. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021)