Exploring the accuracy-size trade-off in extended argumentative decision graphs using SHAP or LIME feature selection

Lucas Rizzo[1,2,*], Luca Longo[3,4]

[1] Technological University Dublin, School of Computer Science, Dublin, Ireland
[2] Artificial Intelligence and Cognitive Load Lab, Technological University Dublin, Dublin, Ireland
[3] School of Computer Science and Information Technology, University College Cork, Cork, Ireland
[4] Artificial Intelligence and Cognitive Load Lab, University College Cork, Cork, Ireland

Abstract

Structured argumentation frameworks provide a promising approach to enhancing the explainability of machine learning (ML) models. This study employs graph-like models learnt from data that can be used for classification tasks, also known as eXtended Argumentative Decision Graphs ($xADGs$). We leverage SHAP or LIME values to reduce inference model size while analysing the trade-off with predictive performance. By selecting the most influential features, an acceptable reduction in accuracy is expected. At the same time, a notable gain in explainability is also anticipated, as this selection can likely minimise the number of arguments and attacks related to them. Several thresholds for SHAP and LIME values were evaluated to identify the most influential features, and the corresponding impacts on the related $xADGs$ were subsequently analysed. In brief, this study further explores enhancements to $xADGs$, aiming to realise their potential to develop accurate and concise structured argumentative models.

Keywords

Computational Argumentation, Non-monotonic reasoning, Machine Learning, Explainable Artificial Intelligence, Attribution methods, Accuracy, Trade-off

1. Introduction

Most recently, computational argumentation has been used to link Machine Learning (ML) and eXplainable Artificial Intelligence (XAI) [1]. Nonetheless, despite recent advances, the development of argumentative XAI remains challenging [2]. One major challenge lies in the computational aspects of designing Argumentation Frameworks (AFs) that specify arguments and capture their dialectical relations. Moreover, extracting explanations based on AFs remains a complex task. To address these issues, the eXtended Argumentative Decision Graph ($xADG$) has been proposed [3], building on the work in [4]. Simply put, $ADGs$ are graph-like models learnt from data that can be used for classification tasks. They employ arguments with rule-based structures in the form "***if*** *feature a has value b* ***then*** *conclusion is y*". A binary attack relation is also defined, allowing arguments to attack others and be discarded in favour of their attackers. Fig. 1 illustrates an example of a Decision Tree (DT) and equivalent ADG.

The extension proposed in [3] introduced Boolean logic operators and support for multiple premises within the internal structure of arguments. By leveraging this framework, the authors presented a procedure to transform DT models into counterparts $xADGs$, which preserve the inferences of the original model. With argument-based models, users receive contrastive explanations that show why certain predictions are preferred over others, providing interpretability that DT models typically lack.

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.
lucas.rizzo@tudublin.ie (L. Rizzo); luca.longo@ucc.ie (L. Longo)
0000-0001-9805-5306 (L. Rizzo); 0000-0002-2718-5426 (L. Longo)

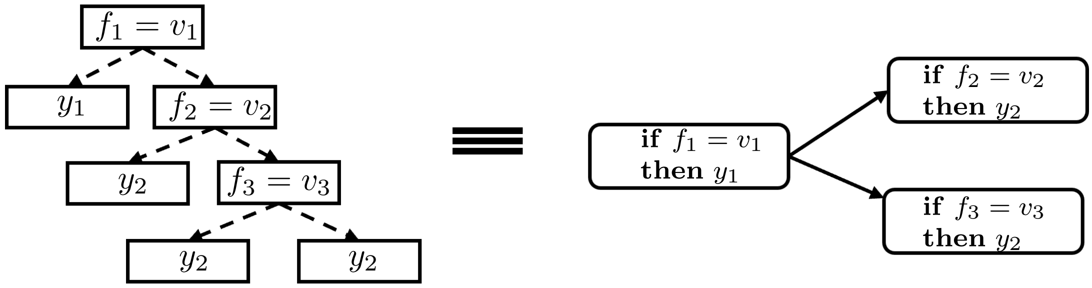

Figure 1: DT (left) and an equivalent ADG (right). Dashed arrows form paths in the DT, branching left when the condition is true and right otherwise. Solid arrows represent attacks in the ADG. In this example, if the conditions of two arguments are true and one attacks the other, the attacked argument is rejected. Given the same input values for each feature, both the DT and this equivalent ADG will produce the same outcome.

This research employs $xADGS$ and introduces and evaluates a SHAP- or LIME-based feature selection process that reduces model size by removing arguments deemed less important and their associated attacks. This approach does not require retraining the underlying DT and is expected to efficiently enhance explainability while maintaining an acceptable trade-off in accuracy. In contrast to traditional model pruning techniques, which often require retraining to accommodate new constraints, this approach enables post hoc simplification of argumentative models. Still, we do not conduct a qualitative analysis of understandability. A deeper qualitative evaluation would strengthen the validity of the results and remains a limitation in this study. However, we provide a quantitative assessment that may indicate improved perceived understandability. This assessment is conducted on classification tasks using datasets of varying scales from the UCI Machine Learning Repository [1]. The objective is to help realise the potential of $xADGs$ and advance their use in developing more accurate, concise argumentative models.

2. Literature Review

An argumentation framework consists of individual arguments (symbolic or structured) and relationships between these arguments, typically indicating support or attack. Such a framework enables the construction of argumentation graphs that can be used to evaluate which arguments can be accepted or rejected based on their conflicts with one another. Hence, it can be used in decision-making, dispute resolution, and reasoning tasks in which multiple viewpoints need to be analysed and assessed [5, 6]. These advantages have led to works focused on linking ML and argumentation as a means of explanation. For example, in [7], argumentation is employed to explain ML outcomes, providing an argumentative interpretation of both the training process and the outcomes. In [8], the authors provide a recommender system that allows the extraction of argumentative explanations in the form of bipolar argumentation frameworks. In an earlier study, [9], the authors introduce a strategy that enables experts to provide arguments as rules that link a class to feature-value pairs. These rules are then used as constraints during ML model training, thereby facilitating interactive ML between users and models. More recently, [10] uses argumentation for explainable classification, building a model that is able to justify the outcomes of a classifier in terms of assumptions that have to be accepted to achieve a predicted outcome. Still focused on explainability, some studies have explored the construction of complete or approximate mappings from black-box models to argumentation frameworks. For instance, in an earlier study, [11] introduces a neural argumentation algorithm that translates argumentation networks into standard neural networks, while [12] demonstrates that feed-forward neural networks can be represented as quantitative argumentation frameworks. In another example [13], the authors propose extracting argumentation debates from data, thereby enabling dialectical explanations of outcomes (classifications). While the approach exhibits competitive predictive performance, it does not fully

[1]UCI Machine Learning Repository, University of California, Irvine, http://archive.ics.uci.edu/ml, Accessed on: February 28, 2025

analyse the explainability of the results. In [14], arguments and attacks are derived from data-driven proxy indicators, allowing the inference of the probability that a set of arguments is accepted or rejected. Experiments with synthetic data demonstrate the feasibility of this approach. Recently, [15] focused on generating model-agnostic argumentative models from class-partitioned data, producing shorter, more interpretable rule sets. Rule-based XAI and its evaluation is usually performed via objective metrics [16]. Recently, [17, 18] empirically integrated an argumentation framework with inference rules automatically extracted via a Logic Learning Machine, showing that argumentation-based conflict resolution can enhance the inferential power of the same data-driven rules.

In summary, argumentative systems aim to produce more understandable models, though understandability lacks a unified definition and is often assumed rather than explicitly studied, due to the transparency of argumentative frameworks. Notably, once an argumentative model is generated, none of the reviewed works has explored modifications aimed only at improving understandability. This highlights a gap in the development and evaluation of techniques for restructuring argumentative models to enhance understandability while preserving predictive capacity. A key advantage of argumentative models is their natural flexibility in adjusting their knowledge base, usually in the form of arguments and relationships between them. Domain knowledge can be easily integrated, and arguments and attacks can be adjusted without limitations on the argumentation graph's topology. Most importantly, these modifications do not compromise the consistency of argument evaluation, ensuring that established approaches in the literature can still be applied. Therefore, this work exploits a structured argumentation framework previously proposed in [3], which considers both predictive accuracy and model interpretability. Initially, it is used to generate models of inference that are a complete mapping from a base DT model. This automated mapping drastically accelerates the knowledge acquisition process for defining arguments and attacks. However, in this study, the model creation pipeline is further extended. After being defined, arguments and attacks are chosen via a feature selection process, and those deemed less important are removed from the argumentation graph. The result is a model that is now an approximate mapping to the base DT, but is likely more understandable. The next subsections introduce the necessary concepts and methods for the definition of the employed framework.

3. Design and methodology

This research study implements and evaluates a pipeline for deriving $xADGs$ from decision trees (DTs). The process begins by constructing an ADG [4] that is *faithful* to the base DT, meaning it preserves the exact same inferences while providing an alternative representation grounded in arguments. This ADG then serves as the foundation for generating a faithful $xADG$, as proposed in [3], through boolean logic transformations applied within the argument internal structure, allowing modifications to the topology of the argumentation framework without modifying the inferences produced. Subsequently, a feature selection process is executed to generate a *relaxed* (unfaithful) $xADG$, where the trade-off between size (number of arguments and attacks) and accuracy becomes the main concern of the evaluation. Figure 2 illustrates the proposed pipeline, including the evaluation process to derive a relaxed $xADGs$ from a DT. The remainder of this section details the process to generate $xADGs$ followed by the feature selection process, which is the original part of this pipeline.

3.1. Generating faithful $xADGs$ from DTs

Given a DT built by any standard algorithm, such as C4.5 or CART, a faithful ADG can be derived as proposed in [4]. Intuitively, each terminal node in the DT corresponds to a predictive argument in the ADG, whereas non-terminal nodes produce non-predictive arguments (arguments with empty conclusion). Finally, arguments with different premises (input features) and conclusions, located along disjoint paths leading to distinct terminal nodes, generate attacks. When exploited by an extension-based semantics, such as grounded [19], the resulting ADG produces the same set of inferences as the DT. The same ADG can be further simplified as proposed in [3] . This simplification allows the

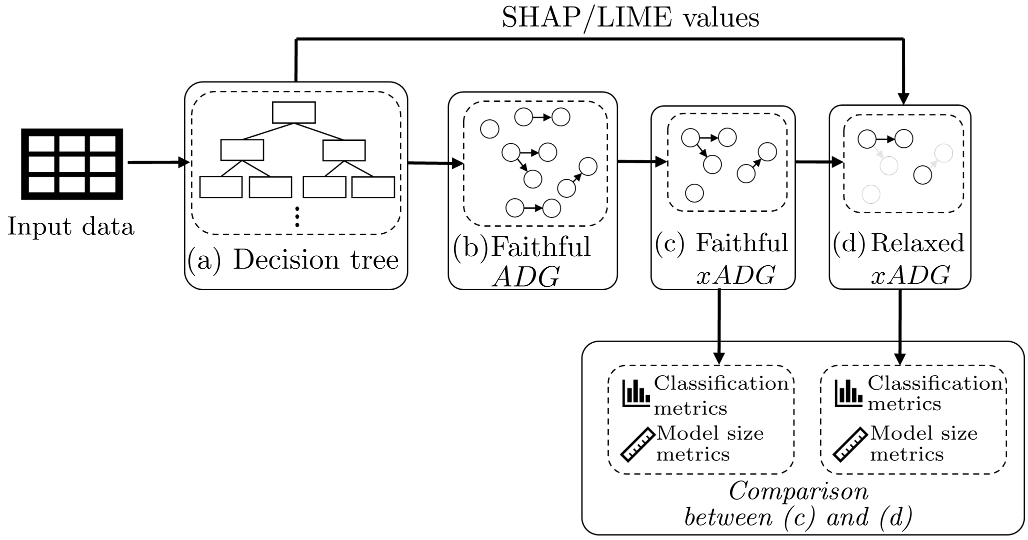

Figure 2: Design of the proposed experiment. The process begins with some input data for a binary classification problem, from which a decision tree is generated (a) and passed to the subsequent step. In step (b), a faithful ADG is constructed and used as input to generate a faithful $xADG$ in step (c). This, in turn, is carried forward to step (d), where a relaxed $xADG$ is produced by discarding certain arguments and attacks (in grey) based on a feature selection process performed with SHAP or LIME values from the DT in (a). Finally, the relaxed $xADG$ is compared to the base variant from step (c) in terms of predictive performance and size.

removal of any superfluous arguments and their related attacks that may arise from redundant siblings within the tree. Such siblings originate from the same parent node and lead to the same outcome.

Given any $(x)ADG$, a faithful $xADG$ can be derived by performing a set of modifications. Specifically, attacks and arguments can be reduced in an $xADG$ to generate a faithful variant by taking advantage of the Boolean logic employed in the internal structure of arguments. In other words, the derived $xADG$ will have a new topology and internal structure of arguments, while still producing exactly the same inferences as the input $xADG$. Two of such modifications proposed in [3] were employed:

1. Merger of two predictive arguments with the same conclusion and same set of targets and attackers (including empty sets) into a single argument. In this case, the supports of the merged arguments are concatenated by the OR operator and the conclusion is kept the same. For instance: if argument “A_1: *if feature* f_1 *has value* v_1 *then conclusion is* y” and argument “A_2: *if feature* f_2 *has value* v_2 *then conclusion is* y” have the same set of targets and attackers, they can be merged into a single argument “A_3: *if* f_1 *has value* v_1 *OR feature* f_2 *has value* v_2 *then conclusion is* y” which preserves the same targets and attackers but represents them with a single argument.

2. Removal of attacks from a non-predictive argument (empty conclusion) with addition of the negative form of its support to the attacked arguments using the AND operator. For instance: if argument “A_1: *if feature* f_1 *has value* v_1” attacks argument “A_2: *if feature* f_2 *has value* v_2 *then conclusion is* y”, then both arguments and the corresponding attack can be simplified into a single argument: “A_3: *if* f_1 *does not have value* v_1 *AND feature* f_2 *has value* v_2 *then conclusion is* y”

3.2. Feature selection and generation of relaxed $xADGs$

This research study utilizes SHAP values computed using the `TreeSHAP` algorithm from the Python SHAP package[2]. To estimate global feature importance using LIME, local explanations are averaged in multiple test instances produced with the Python LIME package[3] and its `LimeTabularExplainer` function. An aggregation over 50 samples yields a global feature importance measure comparable to

[2]https://pypi.org/project/shap/
[3]https://pypi.org/project/lime/

SHAP-style explanations. The number of samples was selected to reduce randomness from LIME and keep runtime reasonable.

Given a feature ranking obtained using SHAP or LIME importance values, a threshold parameter $\tau \in [0, 1]$ is introduced to control the elimination of features. Specifically, the lowest-ranked $\tau \times 100\%$ of features are removed, retaining only the most influential features according to the chosen explanation method. Although alternative criteria, such as a fixed threshold for feature selection, could be used, a dynamic threshold is preferred for its adaptability in assessing feature importance across diverse datasets. In other words, the dynamic threshold enables an initial evaluation of the feasibility and effectiveness of a feature selection process within $xADGs$. The goal is to balance the model's size while maintaining reasoning quality.

Given a list of selected features and a base $xADG$, any unselected features, along with their corresponding arguments and attacks, can be removed. This process does not guaranty that arguments with mutually exclusive conclusions that use different features in their supports will attack each other. Hence, this relaxation may produce an $xADG$ that is no longer faithful to the base model, potentially resulting in missing or altered inferences. However, note that this process *does not* lead to logical inconsistencies. For the purposes of this paper, the grounded semantics [20] is sufficient to decide the set of acceptable arguments (also known as grounded extension). An extension can defend its position and maintain internal coherence, even in the face of disagreement [21]. The grounded semantics is among the most commonly employed semantics, and allows membership in its extension to be decided in polynomial time [20]. In other words, this means that every inference is produced only with arguments that belong to a grounded extension and support the same conclusion. In addition, the proposed relaxation of $xADGs$ also exploits the inherent flexibility of argumentation frameworks, whose topologies can be more readily modified than those of, for example, more rigid models such as decision trees. Although outside the scope of this research article, records without inference could also be addressed by end users, such as domain experts. They could for example adapt $xADGs$ to incorporate additional knowledge or preferences more freely than in such rigid models.

3.3. Experimental setup

Three steps are proposed to implement the full pipeline depicted in Fig. 2: i) Given a DT, generate a faithful $xADG$; ii) Simplify the faithful $xADG$ by modifying the Boolean logic employed in the internal structure of arguments; iii) From the faithful $xADG$ generate a relaxed version using SHAP or LIME-based feature selection. Once the final relaxed $xADG$ is created, it can be instantiated with input data, generating a set of activated arguments and attacks on a case-by-case basis. These activated elements are evaluated using the grounded semantics to determine a unique extension. This extension can result in three possible outcomes: i) an empty extension, meaning no inference is produced; ii) an extension containing arguments with mutually exclusive conclusions that do not attack each other, also leading to no inference; and iii) an extension consisting solely of arguments supporting the same conclusion, which is then used as the output inference.

4. Results

To evaluate whether a SHAP or LIME-based feature selection can produce relaxed $xADGs$ with a good balance between explainability and predictive performance, three publicly available datasets from the UCI ML Repository were identified: Breast Cancer, Indian Liver Patient Dataset and Census Income. These were selected based on their application domains (healthcare and social sciences) and their feature and record counts. Problems in these domains typically require models that are both accurate and interpretable, while the dataset size and number of features may pose challenges to achieving these objectives. A base DT was built for each one with the optimised version of the CART algorithm provided by the `scikit-learn` package[4]. The training and testing sets used a split ratio of 80%-20%.

[4] https://scikit-learn.org/

The maximum depths of the decision trees were selected based on a preliminary analysis of balanced accuracy, stopping at the last depth that provided a meaningful improvement to avoid unnecessary complexity. Next, $xADGs$ were extracted from the base DTs as proposed in [3]. Lastly, a grid search was conducted to determine a practical feature selection threshold τ for building the relaxed $xADGs$. The threshold was varied from 0.1 to 0.25 in increments of 0.05. The goal was to ensure that a lower number of records lacked inference, maintaining the practical usefulness of the models and some level of reasoning quality of their produced inferences. These values support an initial evaluation but could be fine-tuned for specific applications.

Fig. 3 depicts the trade-off between model size and predictive power. In order to provide a more intuitive and visual analysis, the figure shows the midpoint of confidence intervals computed with 100 executions. The first column compares the number of attacks and arguments, while the second shows balanced accuracy and NA percentage. Gray lines indicate the results for the baseline faithful $xADG$. Percentage differences relative to this baseline are shown for each bar. When comparing the base $xADG$ with its relaxed counterparts, the main trade-off lies between the percentage of records with no inferences (% of NAs) and the overall model size. By varying the τ parameter, a balanced trade-off could be explored. The results demonstrate that removing selected features, along with the corresponding arguments and attacks, can substantially reduce their total count. Meanwhile, the (balanced) accuracy for the remaining records stays nearly unchanged, with a small to moderate increase in the percentage of NAs, ranging from 1.8% to 16.6%.

5. Conclusion and future work

In summary, this study proposed a neuro-symbolic approach to XAI by integrating subsymbolic post-hoc feature attribution methods, tested with SHAP and LIME, with symbolic reasoning from $xADGs$. In other words, the approach exemplifies a pragmatic hybridisation: subsymbolic models (decision trees informed by SHAP/LIME) supply empirical salience, while symbolic argumentation frameworks provide logical coherence and manipulability. A core advantage of such an approach lies in its ability to reduce model complexity and improve interoperability, employing the notions of arguments and attacks, without retraining the underlying classifier. This is particularly valuable in settings where retraining is infeasible or costly, and where more explainability of the inferential mechanism is required. Through the process of pruning arguments and attacks based on feature importance, the proposed method operationalises a transparent accuracy–size trade-off while preserving a defeasible reasoning structure grounded in formal argumentation semantics. Within the XAI landscape, this research's contribution to the body of knowledge goes beyond purely local or feature-level explanations. It offers a global, contrastive, and dialectical explanation that is closer to human reasoning practices. However, some limitations remain and are subject to future research endeavours. For example, the reliance on decision trees constrains generality. Similarly, the absence of user-centred qualitative evaluation weakens claims about understandability. However, measuring subjective explainability is known to be an open problem in XAI. Moreover, the relaxed xADGs may yield empty inferences, raising practical concerns. Also, the proposed method inherits known weaknesses of SHAP and LIME, including sensitivity to feature correlations and approximation fidelity. Nevertheless, it meaningfully contributes to both explainable AI and neuro-symbolic reasoning by showing how symbolic structures can be flexibly adapted using learned statistical insights while retaining formal guarantees of rational justification.

Future work should include additional datasets and user studies to further evaluate the understandability of large $xADGs$. The addition of domain-expert knowledge to $xADGs$ could also be tested across various contexts. In particular, this addition could help reduce the percentage of records without inference when the proposed feature selection process is applied. In terms of the structure of $xADGs$, it could be possible to explore variations that allow other types of logics to be employed by the argument's internal structure. For example, by incorporating fuzzy options that allow natural language terms to be better modelled. Another valuable line of analysis would be to assess how effective $xADGs$ are as interpretable surrogates for DTs extracted from more complex models, such as convolutional neural

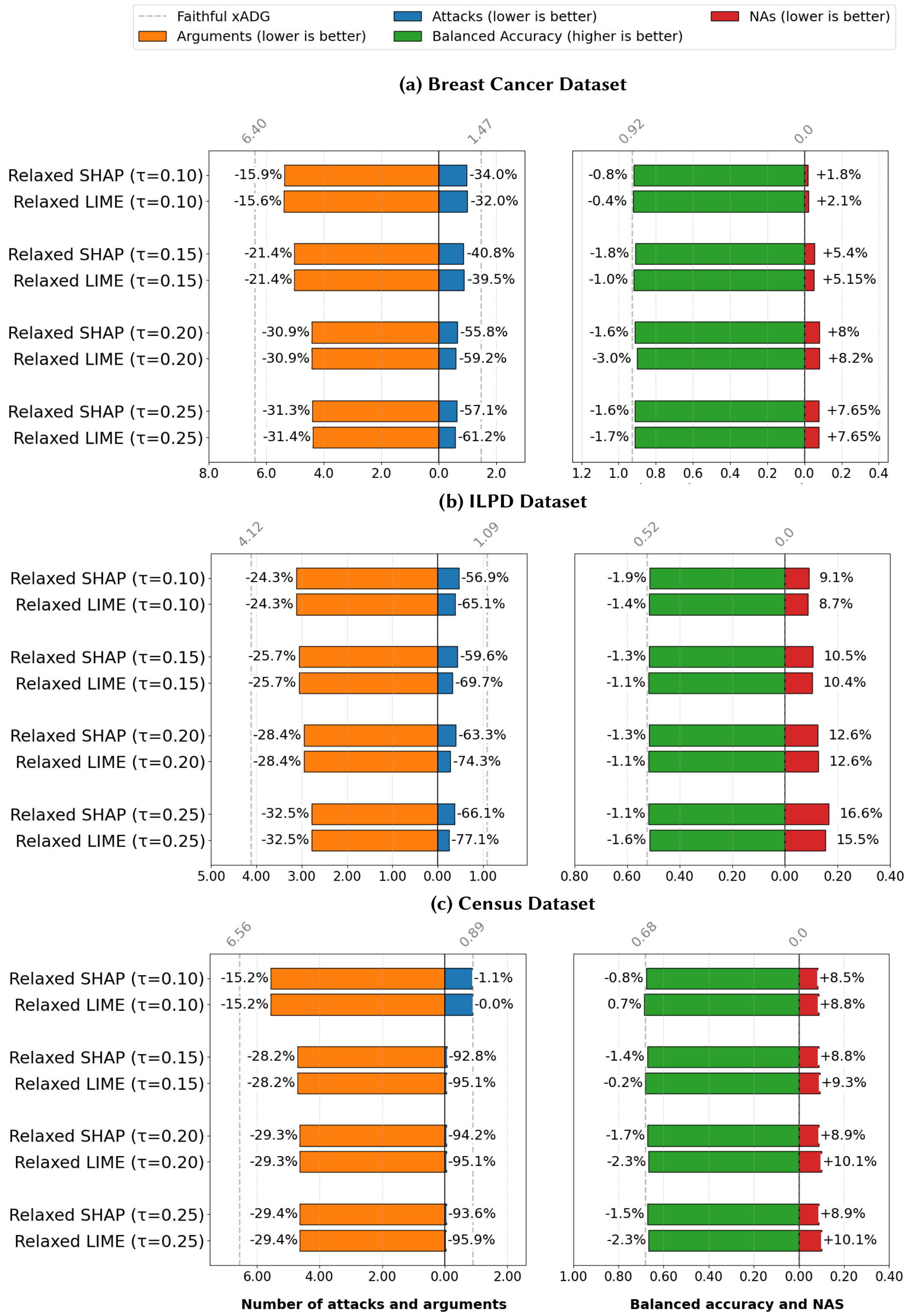

Figure 3: Comparison of faithful $xADG$ and their relaxed counterparts across three datasets.

networks or random forests. Finally, a general software solution, such as a Python package, could likely advance the use and further research on $xADGs$ despite the availability of existing public algorithms.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5. doi:10.1017/S0269888921000011.

[2] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative xai: A survey, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4392–4399.

[3] L. Rizzo, A novel structured argumentation framework for improved explainability of classification tasks, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer, Cham, 2023, pp. 399–414.

[4] P. Dondio, Towards argumentative decision graphs: Learning argumentation graphs from data, in: M. D'Agostino, F. A. D'Asaro, C. Larese (Eds.), AI^3@ AI* IA, CEUR-WS.org, 2021.

[5] L. Longo, B. Kane, L. Hederman, Argumentation theory in health care, in: 2012 25th ieee international symposium on computer-based medical systems (cbms), IEEE, 2012, pp. 1–6.

[6] O. Cocarascu, F. Toni, Argumentation for machine learning: A survey, in: Computational Models of Argument, IOS Press, 2016, pp. 219–230.

[7] S. Bistarelli, A. Mancinelli, F. Santini, C. Taticchi, An argumentative explanation of machine learning outcomes, in: Computational Models of Argument, IOS Press, 2022, pp. 347–348.

[8] A. Rago, O. Cocarascu, C. Bechlivanidis, F. Toni, Argumentation as a framework for interactive explanations for recommendations, in: Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, volume 17, 2020, pp. 805–815.

[9] M. Možina, J. Žabkar, I. Bratko, Argument based machine learning, Artificial Intelligence 171 (2007) 922–937.

[10] H. Prakken, R. Ratsma, A top-level model of case-based argumentation for explanation: formalisation and experiments, Argument & Computation 13 (2022) 159–194.

[11] A. S. D'Avila Garcez, D. M. Gabbay, L. C. Lamb, Value-based argumentation frameworks as neural-symbolic learning systems, Journal of Logic and Computation 15 (2005) 1041–1058.

[12] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 6463–6470.

[13] O. Cocarascu, A. Stylianou, K. Čyras, F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: ECAI 2020, IOS Press, 2020, pp. 2449–2456.

[14] K. Noor, A. Hunter, A bayesian probabilistic argumentation framework for learning from online reviews, in: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2020, pp. 742–747.

[15] G. Vilone, L. Longo, A global model-agnostic xai method for the automatic formation of an abstract argumentation framework and its objective evaluation., in: 1st Int. Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), volume 3209 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[16] G. Vilone, L. Rizzo, L. Longo, A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence, in: L. Longo, L. Rizzo, E. Hunter, A. Pakrashi (Eds.), Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, December 7-8, 2020, volume 2771 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 85–96.

[17] L. Rizzo, D. Verda, S. Berretta, L. Longo, A novel integration of data-driven rule generation and computational argumentation for enhanced explainable ai, Machine Learning and Knowledge Extraction 6 (2024) 2049–2073.

[18] L. Longo, S. Berretta, D. Verda, L. Rizzo, Computational argumentation and automatic rule-generation for explainable data-driven modeling, IEEE Access 13 (2025) 175565–175583.

[19] P. Baroni, M. Giacomin, Semantics of abstract argument systems, Argumentation in artificial intelligence (2009) 25–44.

[20] 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.

[21] Y. Wu, M. Caminada, M. Podlaszewski, A labelling-based justification status of arguments, Studies in Logic 3 (2010) 12–29.