

ORIGINAL RESEARCH

Open access

From Data to Arguments: A Unified Paradigm for Explainable Artificial Intelligence

Daniel Fischer^{1*}, Laura Meier¹, Stefan Koch², Thomas Braun³

Abstract

Explainable Artificial Intelligence (XAI) represents an active discipline dedicated to clarifying the internal mechanisms of AI systems, thereby fostering stakeholder confidence and advancing the deployment of accountable technology. A primary theoretical hurdle within XAI involves engineering techniques capable of producing clear, accessible interpretations without sacrificing scalability or operational performance in intricate domains. While rule-based methodologies offer a framework for decoding AI-driven conclusions, they are prone to becoming overly complex in dense environments, which diminishes both their comprehensibility and computational scalability. Furthermore, these approaches frequently overlook contrastive logic, which leaves observers unable to discern why a specific outcome was prioritized over alternatives. To resolve this underlying scientific problem, this research investigates the fusion of computational argumentation—a specialized domain of AI that replicates cognitive processes through defeasible reasoning—with rule-based XAI architectures. Computational argumentation provides a mechanism whereby arguments derived from structured rules can be invalidated or withdrawn when confronted with conflicting evidence. Consequently, this framework offers a compelling avenue for refining rule-based systems to yield more transparent AI models. Despite the synergistic benefits of combining rule-based structures with computational argumentation, academic inquiry into their synthesis remains sparse. Accordingly, this paper also targets the practical engineering challenge of realizing this conceptual integration within functional AI software. This investigation utilizes the Logic Learning Machine (LLM)—a specialized rule-extraction methodology—and introduces a modular framework designed to map input rules directly into a structured argumentation network via state-of-the-art argumentation techniques. Empirical evaluations performed on various binary classification problems from the UCI Machine Learning Repository validate the utility of this combined model. The LLM strategy proved highly proficient at extracting a compact, interpretable collection of if-then rules characterized by minimal antecedent premises, all while sustaining strong predictive accuracy across all evaluated datasets. Concurrently, the argumentation-based models delivered classification performance comparable to the native if-then rule configurations, achieving this through a streamlined rule set while offering vastly superior explainability. Ultimately, this work delivers an innovative framework for the automated and streamlined generation of arguments and their internal conflicts directly from raw data, simultaneously addressing fundamental theoretical and practical barriers to propagate argumentation frameworks within the XAI landscape.

Keywords Explainable artificial intelligence, Computational argumentation, Defeasible reasoning, Rule-base AI

*Correspondence:

Daniel Fischer

daniel.fischer@pharmazie.uni-freiburg.de

¹ Department of Explainable AI and Argumentation, Faculty of Engineering, University of Freiburg, Freiburg, Germany

² Department of Data-Driven Rule Generation and Argumentation, Faculty of Engineering, Technical University of Munich, Munich, Germany

³ Department of Unified XAI Paradigms, Faculty of Engineering, University of Kiel, Kiel, Germany

Introduction

Explainable Artificial Intelligence (XAI) is a prominent field of research concentrated on decoding how machine learning models formulate specific inferences. A core objective of this discipline is the design and execution of methodologies that extract highly interpretable and explicit models from empirical data. To be effective, these models must simultaneously maintain elevated predictive precision, structural transparency, and complete intelligibility for human operators. Cultivating these attributes is essential for bolstering human reliance on automated systems and ensuring the development of ethical, dependable AI architectures [1, 2]. Among the diverse paradigms within XAI, rule-based approaches occupy a significant position. The foundational logic behind utilizing rules is their intrinsic accessibility and clarity to human users, which aids in demystifying the internal inferential logic governing complex AI models [3]. For instance, traditional decision trees are capable of generating explicit if-then logic tracks [4], whereas systems rooted in fuzzy logic yield nuanced fuzzy rules [5]. Even so, these rule sets can become intensely convoluted when applied to high-dimensional scenarios, introducing substantial bottlenecks regarding cognitive readability and computational scalability. As an illustration, while deep decision trees often exhibit high precision, their extensive branching produces protracted, overly complex rules that are unmanageable for human review. Additionally, such paradigms routinely fail to supply contrastive rules [3], which are vital for helping users understand the precise reasons why a model favored one specific prediction over an alternative option [6]. For this reason, a fundamental theoretical challenge in XAI centers on creating methodologies that deliver clear, understandable explanations while remaining scalable and functional in highly intricate settings.

An established branch of artificial intelligence is automated reasoning, which translates human intellectual processes into computational workflows. Within this domain, computational argumentation stands out as a specialized sub-field focused on operationalizing these cognitive models [7, 8]. From a technical standpoint, defeasible reasoning operates as a form of non-monotonic logic [9, 10] that evaluates distinct rationale—frequently structured as rules or "arguments"—within an interactive, dialogical framework [11]. This specific architecture accommodates the formal revocation of previously established assertions or conclusions when new data or counter-arguments emerge to invalidate them. As a result, computational argumentation represents a highly suitable mechanism for designing efficient XAI protocols [12]. This compatibility stems from its foundational syntax (rules and arguments) paired with its capacity to simulate human reasoning when navigating uncertain, conflicting, or contrastive knowledge states. Computational argumentation is therefore exceptionally well-suited for interpretability tasks [13] and serves as an optimal baseline for constructing explainable AI frameworks [14]. Yet, despite the clear complementary advantages of rule-based systems and computational argumentation, research at their nexus remains minimal [15]. This investigation directly addresses the practical challenge of implementing this conceptual fusion within operational AI tools. Specifically, the study contributes to this domain by exploiting the unique capabilities of a rule-extraction algorithm known as the Logic Learning Machine (LLM) [16] to extract crisp if-then rules, which are subsequently converted into a structured argumentation framework [7] utilizing advanced computational argumentation techniques. A exploratory comparative evaluation is detailed for classification operations across datasets of varying scales. For every dataset, if-then rules are generated via the LLM technique to target a designated target class. In instances where an individual data record satisfies multiple, conflicting rules, these discrepancies are handled via two distinct paradigms. The first approach utilizes the Standard Applied Procedure, a traditional heuristic mechanism that weights individual rules based on their breadth of coverage (generality) and their historical error rate (precision). The definitive prediction for any given data instance is determined by identifying the target classification that accumulates the highest cumulative weight from all validated rules. The second paradigm resolves these conflicts through the deployment of computational argumentation frameworks.

Related Work

Rule-based explainable artificial intelligence

The core obstacle restricting the widespread acceptance of most AI-driven frameworks is their lack of comprehensibility, commonly referred to as the black-box dilemma. This phenomenon describes the immense difficulty encountered by both specialists and non-expert users when attempting to comprehend the underlying reasoning pathways of these systems and how their final outputs are produced. Explainable Artificial Intelligence (XAI) is an accelerating discipline dedicated to

rendering these opaque AI structures transparent, interpretable, and intelligible, among other essential traits [15]. A multitude of methodologies have been engineered and operationalized, with numerous alternative strategies currently under development. As an illustration, certain investigators [17] have worked toward establishing grey-box architectures, which aim to inject structural interpretability into otherwise uninterpretable models. While completely transitioning a black-box system into a fully transparent white-box model is not always attainable or practical, XAI acts as an intermediary solution by making black-box architectures more accessible and transparent through a variety of explanatory techniques. Another distinct category within these approaches comprises rule-based methodologies [15]. These frameworks are uniquely compelling because rules reflect the standard logical formulations utilized in human cognition to validate assertions, and they are widely acknowledged to possess high levels of clarity and explanatory value. Specifically, within this domain, systems that formulate if-then rules—such as Decision Trees (DTs) [18] and the Logic Learning Machine (LLM) [19]—provide an effective strategy for constructing predictive systems whose deductive paths can be readily parsed and understood, thereby serving as a foundation for justification and interpretation. This benefit is principally due to their capacity to regulate the upper limits of antecedent conditions and error thresholds assigned to individual rules, which minimizes the total quantity of generated rules while sustaining a powerful predictive performance [20]. This methodology has been deployed across diverse biomedical scenarios [21], including the investigation of gene expression profiles for oncological diagnosis [22] and the prediction of clinical disease risks [23].

Even though rule-based paradigms provide highly advantageous characteristics for extracting interpretable architectures from empirical data, they exhibit poor capabilities when managing contrastive logical paths and internal conflicts. Restated, they frequently lack structured approaches for managing rules that conflict within their antecedent configurations or their consequent spaces, which prevents them from formulating comprehensive explanations for users. A comprehensive literature review outlines multiple tactics designed to resolve contradictions within rule-based tools [24]. The CN2 algorithm [25] addresses rule contradictions by calculating, for each individual rule r and every available target class o , the exact quantity of training instances belonging to class o that are encompassed by rule r ; the target class exhibiting the highest frequency is subsequently designated as the system's prediction. Another prevalent method for mitigating rule conflict is the voting mechanism [26, 27]. In this setup, every applicable rule submits a vote toward its designated target class scaled by a weight equivalent to the historical probability of that specific class; these votes are mathematically aggregated, and the target class accumulating the greatest total probability is selected as the definitive prediction. Alternative techniques for navigating rule-based contradictions are detailed in [28, 29]. In the former study, the authors propose leveraging the training data situated precisely at the intersection of the conflicting rules to determine the appropriate output class. If no training data exists within that intersection, partitioning logic is introduced to divide the rules into as few non-empty subsets as possible, which are then used to identify the most probable output class. The latter study introduces Double Induction, an approach designed to generate completely new rules derived from the training samples captured by the conflicting rule sets. These newly formulated rules are subsequently used to categorize the unlabelled data record. This strategy has been further developed into an updated framework known as Recursive Induction [24]. An alternative technique for resolving contradictions between active rules is the Standard Applied Procedure, which is the specific methodology utilized in this investigation. Within this approach, the assigned weight of a given rule is determined by its breadth of coverage and its accuracy across the training partition, relative to how these performance metrics compare against other rules targeting the identical classification output.

Argument-based XAI and defeasible reasoning

Computational argumentation inherently addresses the limitations outlined above and maintains a direct, fundamental relationship with Explainable Artificial Intelligence [13]. This connection is established because XAI strategies are strengthened by clarifying and defending the deductive assertions made by AI frameworks. Concurrently, computational argumentation focuses on developing systems that systematically link these outputs to the underlying empirical evidence that justifies them. Specifically, architectures grounded in computational argumentation are capable of demonstrating the step-by-step logical sequence through which an AI model arrives at a definitive conclusion, thereby delivering transparent justifications to human operators [12]. This operational clarity provides a distinct benefit over alternative white-box architectures, such as those relying on Bayesian networks. The argument-based strategy is predominantly qualitative,

centering on the underlying composition and logical flow of individual arguments, whereas the Bayesian network approach [30] is quantitative, emphasizing statistical inference and the algorithmic management of uncertainty. Its reasoning pathway is more mathematical and relies strictly on the foundational language of probability theory. Choosing between these distinct methodologies depends on the intrinsic attributes of the problem under analysis, alongside the specific requirements regarding structural interpretability and the preferred style of reasoning. In short, although both the argument-based framework and the Bayesian network approach can be categorized as white-box models due to their interpretable designs, they differ fundamentally in their underlying logical methodologies.

The theoretical intersection between XAI techniques and computational argumentation is extensively examined in [31]. Along these lines, arguments can be effectively utilized to map out the specific rationales that prompt an autonomous agent to execute a given action, while argumentation semantics serve to isolate the permissible arguments within a scenario involving disaster response rescue robots [32]. A specialized field where computational argumentation can dramatically enhance the interpretability and transparency of AI-driven systems is medical informatics [33]. Argumentative frameworks possess structural traits and behavioral features that closely mirror the cognitive workflows routinely executed by humans to reason, justify conclusions, and explain decisions under conditions of uncertainty, such as those found in clinical informatics. Specifically, due to their multi-faceted dialectical states, they are capable of producing diverse analytical abstractions that are profoundly integrated with the concepts of explanation and justification [31]. This sophisticated dialectical status can be realized by implementing formalized argumentation semantics [34–37] as demonstrated in [38]. In essence, these semantics operate as targeted strategies for resolving underlying conflicts among competing arguments, which are frequently represented in the form of rules [39]. Prior work has attempted to deploy these techniques to construct predictive models, mapping automated choices directly into an argumentation framework governed by predefined attack configurations between individual arguments (logical rules) [14, 18]. These systematic attack relationships between arguments can generate highly justifiable decisions, clearly revealing the exact sequential steps taken to reach them. Conceptually, this framework introduces the strengths of symbolic, non-monotonic logic into empirical, data-driven predictive systems. Even so, it frequently encounters bottlenecks regarding its total inferential capacity when applied across heterogeneous tasks. Recent investigations [13, 40] have attempted to combine machine learning (ML) models with argumentation frameworks, aiming to pair the predictive strength of the former with the explanatory utility of the latter. Reviewing these dynamics suggests that the integrated approach proposed here could yield more transparent systems while successfully preserving robust classification accuracy.

Argument-based XAI

Explainable Artificial Intelligence and computational argumentation are deeply intertwined, as evidenced by recent trends where argumentation is increasingly applied to supply explanatory frameworks for AI systems [13]. This natural alignment stems from the reality that XAI methodologies benefit heavily from clarifying and defending their computational outputs, whereas argumentation provides a rigorous mathematical mechanism for anchoring any automated decision to the specific evidence that supports it. Argumentation is capable of outlining, in a step-by-step fashion, exactly how an AI system arrives at a given conclusion, thereby elucidating predictions for human operators [12]. A clear illustration of this cooperative dynamic is detailed in [31], which delivers a comprehensive review of XAI frameworks developed via computational argumentation methods, alongside an argument-based paradigm for engineering interpretable Beliefs-Desires-Intentions (BDI) agents. Correspondingly, in a multi-agent environment featuring search-and-rescue robotics, arguments were harnessed to represent the distinct motivations an autonomous agent could use to validate a specific choice. Argumentation semantics were then employed to establish which of these motivations could be legitimately accepted [32]. This fundamental bridge between explainable AI (XAI) and argumentation is driven by the requirement that the former must clarify and defend its outputs, while the latter yields a practical, computational suite of tools designed to link any automated inference directly to its validating evidence [12]. Computational argumentation has also been leveraged to generate dialectically explainable predictions [41]. Specifically, a dedicated framework was engineered to extract argumentative debates directly from raw data tracks, ensuring that their final dialectical configurations correspond directly to actionable predictions, such as classifications that can be explained through dialectical discourse. In this setting, the debates consist of data-derived informational components that may not be linguistic in nature but function as

arguments because they are dialectically linked—for example, by expressing disagreement regarding assigned data classifications. This interactive, dialectical engagement with argument-driven systems can significantly improve human comprehension and reinforce the development of trustworthy AI. In a manner analogous to structured debates, dialogue games based on arguments and their shifting dialectical statuses can be effectively tailored to satisfy the specific explanatory requirements of users [42]. Furthermore, the deployment of computational argumentation has been evaluated within the domain of medical informatics, given the intense contemporary interest in designing explainable AI to assist with clinical decision-making and diagnostic tasks, while simultaneously factoring in operational risks and liabilities for medical practitioners [33]. Closely aligned with this, argumentative frameworks—by virtue of their dialectical traits—possess operational qualities and characteristics that closely replicate the cognitive processes routinely performed by humans to reason and explain, particularly within medical environments.

Integration of the Logic Learning Machine and Structured Argumentation

This section outlines the theoretical frameworks underpinning the chosen methodology.

Rule generator: the logic learning machine

Operating as a rule-based methodology, the Logic Learning Machine (LLM) maps input data into a Boolean domain. In this space, distinct Boolean functions—corresponding to each potential output class—are reconstructed using a subset of their truth tables via the approach presented in [16]. This technique extracts easily interpretable rules through Boolean function synthesis, executing across four core phases: (a) Discretization; (b) Latticization or Binarization; (c) Positive Boolean function; and (d) Rule generation.

Discretization

During this initial phase, continuous variable domains are mapped into discrete spaces via the function $\psi_j: X_j \rightarrow \mathbb{I}_M$, where X_j represents the domain of the j -th variable and $\mathbb{I}_M = \{1, \dots, M\}$ denotes the set of positive integers bounded by M . This mapping process must maintain the original order of the data points. Specifically, if $x_{ij} \leq x_{kj}$, it is required that $(\psi_j(x_i) \leq \psi_j(x_k))$, $\forall i = 1, \dots, d$. The mapping function ψ_j can be characterized by a vector $\gamma_j = (\gamma_1, \dots, \gamma_{m-1})$ such that:

$$\psi_j(x_i) = \begin{cases} 1, & \text{if } x_{ij} < \gamma_{(i1)} \\ m, & \text{if } \gamma_{(jm-1)} < x_{ij} < \gamma_{(jm)} \\ M_j, & \text{if } x_{ij} > \gamma_{(im-1)} \end{cases} \quad (1)$$

Establishing the optimal thresholds for this discretization can be framed as a minimization problem. The objective is to reduce the total number of cutoffs to a minimum while satisfying the constraint outlined in the previous equation. To formally define this optimization, let τ_{jl} and X_{jik} be introduced for $j \in \{1, \dots, d\}$ and $i, k \in \{1, \dots, N\}$ and $l \in \{1, \dots, \alpha_j\}$. Here, X_{jik} represents the index set I where $x_{ij} < x_{kj}$ and τ_{jl} represents binary values indicating whether ρ_{jl} serves as a valid cutoff for the j -th variable:

$$\tau_{jl} = \begin{cases} 1 & \text{if } \rho_{jl} \in \gamma_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Accordingly, the optimization model is formulated as:

$$\min_{\tau} \sum_{j=1}^d \sum_{l=1}^{x_j} \tau_{jl} \quad \text{with:} \quad (3)$$

$$\sum_{j=1}^d \sum_{l \in X_{jk}} \tau_{jl} \geq 1 \quad \forall i, k \quad \text{s.t. } i \neq k$$

When the underlying training data contains inconsistencies, the discretization procedure must simultaneously minimize both these ambiguities and the overall cutoff count. This is achieved by softening the constraints through auxiliary variables that regulate constraint violations. Conversely, stricter boundaries can be enforced by requiring that data points from opposing classes are partitioned by a minimum of $q > 1$ unique variables. Given that this optimization problem involves a high volume of variables, it can become computationally expensive. A practical strategy for identifying a near-optimal solution is to implement a heuristic greedy technique, such as Attribute Driven Incremental Discretization (ADID) [43, 44]. The ADID algorithm relies on the concept of effective distance, denoted as d_{eff} , which calculates how many inputs partition two distinct data points:

$$d_{\text{eff}}(x_i, x_k) = \sum_{j=1}^d \theta(|\psi(x_i) - \psi(x_k)|) \quad (4)$$

$$\text{where } \theta(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{otherwise} \end{cases}$$

Based on this effective distance, a pair of data points x_i, x_k is definitively separated if $d_{\text{eff}}(x_i, x_k) \geq q$. For every potential threshold candidate, the following components are defined:

- The subsets $S_{jl}^{(l)}, S_{jl}^{(r)}$ containing the data points partitioned by the threshold ρ_{jl} :

$$S_{jl}^{(l)} = \{x_i \mid (x_i, y_i) \in S, x_{ij} \leq \rho_{jl}\} \quad S_{jl}^{(r)} = \{x_i \mid (x_i, y_i) \in S, x_{ij} > \rho_{jl}\} \quad (5)$$

- A counter v_{jl} that tracks the volume of data points definitively partitioned by ρ_{jl} :

$$v_{jl} = \sum_{(x_i \in S_{jl}^{(l)})} \sum_{(x_k \in S_{jl}^{(r)})} |y_i - y_k| \theta(q - d_{\text{eff}}(x_i, x_k)) \quad (6)$$

- The metric w_{jl} evaluating the aggregate effective distance between separate classes that are definitively partitioned by ρ_{jl} :

$$w_{jl} = \sum_{(x_i \in S_{jl}^{(l)})} \sum_{(x_k \in S_{jl}^{(r)})} |y_i - y_k| \theta(q - d_{\text{eff}}(x_i, x_k)) d_{\text{eff}}(x_i, x_k) \quad (7)$$

Utilizing these parameters, the ADID heuristic initializes with an empty set. Iteratively, it introduces the specific cutoff that maximizes v_{jl} , utilizing the minimization of w_{jl} as a secondary selection criterion. Following each addition, v_{jl} and w_{jl} undergo recalculation relative to the newly integrated threshold. This iterative expansion continues until $v_{jl} = 0 \forall \rho_{jl}$. After isolating the optimal thresholds from ρ_{jl} , their exact boundaries can be adjusted by calculating the median point between adjacent boundary instances, as formulated below:

$$\gamma_{jl} = \rho_{jl} \longrightarrow \gamma_{jl} = \frac{(\rho_{jl} + \rho_{jl+1})}{2} \quad (8)$$

Prior to executing the binarization stage, categorical features must be transformed. This conversion is more straightforward than continuous variable handling, requiring only the assignment of a unique integer to represent each categorical state.

Binarization

This phase focuses on converting each discretized space into a binary format via a mapping function $\phi: I_j^M \rightarrow \{0,1\}^{M_j}$, where I_{M_j} represents the domain of the j -th variable and $\{0,1\}^{M_j}$ represents a bit string containing a single bit for every potential value within I_{M_j} . It is critical that this mapping preserves data ranking, meaning $u < v$ holds if and only if $(u) < \phi(v)$, where the standard ordering sequence for z and $w \in \{0,1\}^{M_j}$ is structured as:

$$z < w \text{ if, and only if, } \begin{cases} \exists j & \text{such that } z_j < w_j \\ \forall l \neq j & z_l \leq w_l \end{cases} \quad (9)$$

$$z \leq w \text{ if, and only if, } z_j \leq w_j \quad \forall j = 1, \dots, M$$

When the condition in the formula above is met, z is said to cover w . A highly effective approach for ϕ_j is inverse-only coding. For every $k \in I$, this technique generates a binary string $h \in \{0,1\}^{M_j}$ wherein all bits are set to 1, save for the k -th position, which is set to 0. As an illustration, if $x_{ij}=3$ within the domain I_5 , the mapping yields $(x_i)=11011$. Consequently, $(x_i)=z_i$, where z_i is constructed by joining the outputs of $\phi_j(x)$ for $j=1, \dots, d$. The transformed training dataset is thus represented as $S' = \{(z_i, y_i)\}_{i=1}^N$, where $z_i \in \{0,1\}^B$, and $B = \sum_{j=1}^D M_j$.

Synthesis of the boolean function

The post-binarization training set S' can be partitioned into distinct subsets matching their respective target classes: T accumulates instances (z_i) where $y_i=1$, while F groups the remaining instances where $y_i=0$. The sets T and F serve as localized representations of the truth table for the target Boolean function f that needs to be modeled. Adhering to the terminologies established in [45], several notations are defined:

- Standard Boolean algebra relies on the logical operators AND, OR, and NOT to construct functions; excluding the NOT operator simplifies this structure into a Boolean lattice. The remainder of this discussion focuses exclusively on the Boolean lattice framework.
- The logical summation (OR) and multiplication (AND) across a series of m components are represented by:

$$\bigvee_{j=1}^m z_j = z_1 + z_2 + \dots + z_m = z_1 \text{ OR } z_2 \text{ OR } \dots \text{ OR } z_m$$

$$\bigwedge_{j=1}^m z_j = z_1 \cdot z_2 \cdot \dots \cdot z_m = z_1 \text{ AND } z_2 \text{ AND } \dots \text{ AND } z_m \quad (10)$$

- A conjunctive clause is categorized as an implicant of a target function f if it satisfies the relationship $\bigwedge_{j=1}^m z_j \leq f$, where each constituent component z_j represents a literal.

To extract these implicants, the LLM framework utilizes a method known as Shadow Clustering [46], which isolates implicants for f by evaluating the Boolean lattice $\{0,1\}^B$. This algorithm targets a specific node within the lattice diagram and isolates base vertices (T) by tracing downward paths. Moving lower from one node to another corresponds to

switching a single bit component from 1 to 0. A base vertex is integrated into A if any subsequent downward step transitions into a node situated within the lower shadow of an element $w \in F$. The process initiates from a node chosen among the subsets $z \in T \subset \{0,1\}^B$ that do not encompass any point $a \in A$ where $a \leq z$. The routine completes once every element in T encompasses at least one element in A . Because the resulting set A frequently contains redundant components, it must undergo a simplification phase to yield A^* , which provides the Positive Disjunctive Normal Form (PDNF) for the positive Boolean function.

Multiple variations of Shadow Clustering exist, distinguished by the criteria used to select which bit is flipped from 1 to 0 during the downward traversal of the lattice. For instance, Exhaustive Shadow Clustering (ESC) extracts all possible base vertices originating from the initial node. Due to the heavy computational burden of ESC, greedy strategies are generally preferred. These variations alter the i -th component of the active node at each progression to limit the complexity of the resulting function, assessing complexity by both the number of implicants and the count of literals within each implicant. An example is Maximum-covering Shadow Clustering (MSC), which targets the component that maximizes the potential coverage linked to that variable index at each step. Here, the potential coverage gained by flipping index i from 1 to 0 corresponds to the total number of elements $z \in T$ where $z_i = 0$. To isolate the subset $A^* \subset A$, a common approach is to incrementally add elements from A that encompass the highest volume of points in T not yet encompassed by existing members of A^* . Ultimately, these synthesized Boolean functions allow for the formulation of logical rules. These rules can be deployed across datasets to generate classifications once any conflicting rule outputs have been resolved, a process detailed in the following subsection.

Rule Generation

Let r_i represent the i -th rule. The general architecture of any given rule r_i can be formally expressed as:

$$r_i : \text{if } \bigwedge_{j=1}^{m_i} \mathcal{F}_{ij}(v_{ij}) \text{ then } y_i \quad (11)$$

where the symbol $\bigwedge$ signifies a logical conjunction (AND), $\mathcal{F}_{ij}(v_{ij})$ corresponds to the j -th descriptive attribute alongside its specific assigned value or interval boundary v_{ij} within rule r_i , y_i represents the target classification outcome linked to rule r_i , and m_i indicates the aggregate number of antecedent conditions (features) comprising rule r_i . The extraction of these logical constructs is driven by the previously established Boolean functions. Specifically, an interpretable classification rule is derived by systematically converting each underlying implicant from the positive Boolean function f . A distinct function is constructed for every separate target class, making the resulting rules entirely dependent on the characteristics of f .

This transformation framework relies directly on the encoding scheme executed during the binarization phase. Because the binary vector z is constructed by merging the outputs of the mapping function $\phi(x)$ for each dimensions $j=1, \dots, B$, it can be partitioned into distinct sub-vectors h_j for each attribute. Every individual bit $z_i \in h_j$ maps to a categorical state if the variable X_j is nominal, whereas it maps to a continuous range if X_j is an ordered variable. For every individual implicant, a conditional statement in an if-then format is constructed by determining the state of each variable X_j via the following criteria:

- If $z_i=0$ across all elements within h_j , no restrictive condition regarding attribute X_j is embedded into the rule.
- For nominal attributes X_j , a membership condition $X_j \in V$ is integrated, where V aggregates the qualitative states matched to each bit where $z_i=0 \in h_j$.
- For ordered attributes X_j , a boundary condition $X_j \in V$ is integrated, where V represents the combination of numeric intervals paired with each bit where $z_i=0 \in h_j$.

By convention, when these numeric boundaries are finite, the maximum parameter (upper limit) is included within the rule's constraint, while the minimum parameter (lower limit) is strictly excluded.

The nature and volume of the synthesized rule base can be adjusted using several optimization parameters, including:

- The target quantity of rules allocated per classification outcome, which dictates the cumulative density of the generated rule set;
- The upper limit of permissible conditions per rule, which restricts the total number of antecedent premises from surpassing a predefined cap;
- The maximum permissible error rate (expressed as a percentage) that an individual rule is allowed to exhibit.

Adjusting these configuration metrics allows for precise calibration of the final rule set's operational properties. Rules delineating the alternate target class are formulated simply by transposing the binary indicators, substituting label C0 with 1 and C1 with 0. For multi-class classification scenarios, the overarching problem is broken down into a series of isolated binary sub-problems. Utilizing standard One-vs-Rest (OVR) methodology, the specific class under evaluation in each sub-problem is designated as 1, while all remaining categories are combined and labeled as 0.

Rules aggregator logic: standard applied procedure

During the deployment of a Logic Learning Machine configuration, an unseen test instance might trigger multiple rules that conflict by predicting opposing classification categories. To resolve these conflicts, the definitive classification outcome is determined via the Standard Applied Procedure, an approach rooted in assessing individual rule relevance. To formulate this relevance metric, four baseline counts regarding a given rule r structured as IF THEN must be computed across the training dataset:

- v_{1r} counts the training samples that concurrently satisfy both the antecedent premises and the consequent label of rule r ;
- v_{2r} counts the training samples that satisfy the antecedent premises but contradict the consequent label of rule r ;
- v_{3r} counts the training samples that fail to satisfy the antecedent premises and also do not match the consequent label of rule r ;
- v_{4r} counts the training samples that fail to satisfy the antecedent premises but match the consequent label of rule r .

It should be noted that a given sample x_i fulfills the antecedent of rule r only if it simultaneously satisfies every single component condition; conversely, x_i fails to fulfill the antecedent if it violates at least one conditional constraint. By synthesizing these four fundamental counts, two distinct performance metrics can be established for any rule r :

- Covering:

$$C_r = \frac{v_{1r}}{v_{1r} + v_{4r}} \quad (12)$$

- Error:

$$E_r = \frac{v_{2r}}{v_{2r} + v_{3r}} \quad (13)$$

Logically, a higher covering capacity enhances the overall importance and applicability of a rule, whereas an elevated error rate diminishes its dependability. Consequently, the total relevance score of an isolated rule r is derived by combining C_r and E_r as follows:

$$R_r = C_r \times (1 - E_r). \quad (14)$$

Once this relevance metric has been calculated, it can be leveraged to compute a cumulative score $S(x_i, c)$ for each individual class c , quantifying the likelihood that the true label satisfies $y_i = c$:

$$S(x_i, c) = \sum_{r \in R_c^i} R_r \quad (15)$$

where the active rule set is defined as $R_c^i = \{r \in R, r \leq x_i, O(r) = c\}$. In this notation, R represents the entire derived rule repository, $r \leq x_i$ indicates that sample x_i triggers the conditions of rule r Split, and $O(r) = c$ specifies that the conclusion of rule r forecasts class c . Thus, $R(x_i, c)$ isolates the specific cohort of rules triggered by x_i that unanimously forecast class c .

By evaluating these cumulative scores across all potential classifications, the conditional probability that $y_i = c$ can be established:

$$P(c|x_i) = \frac{S(x_i, c)}{\sum_{k \in C} S(x_i, k)} \quad (16)$$

Ultimately, the final classification assigned to the instance is the specific category that yields the highest conditional probability value:

$$y_i^* = \max(c|x_i). \quad (17)$$

Rules aggregator logic: computational argumentation

In contrast to traditional expert systems that typically rely on monotonic logic structures to synthesize rules, defeasible reasoning models deploy a non-monotonic logical framework [14]. Systems engineered around this non-monotonic logic are generally organized into sequential processing tiers [40], operating through the following orderly phases: (a) Constructing the structural components of arguments; (b) Mapping the conflicts that arise between these arguments; (c) Appraising the severity and dynamics of these argumentative conflicts; (d) Assigning a dialectical status to each participating argument; and (e) Combining the validated, acceptable arguments to compute a unified final inference (as illustrated on the right side of Figure 1). The formal mechanisms required to execute each of these successive steps are detailed below.

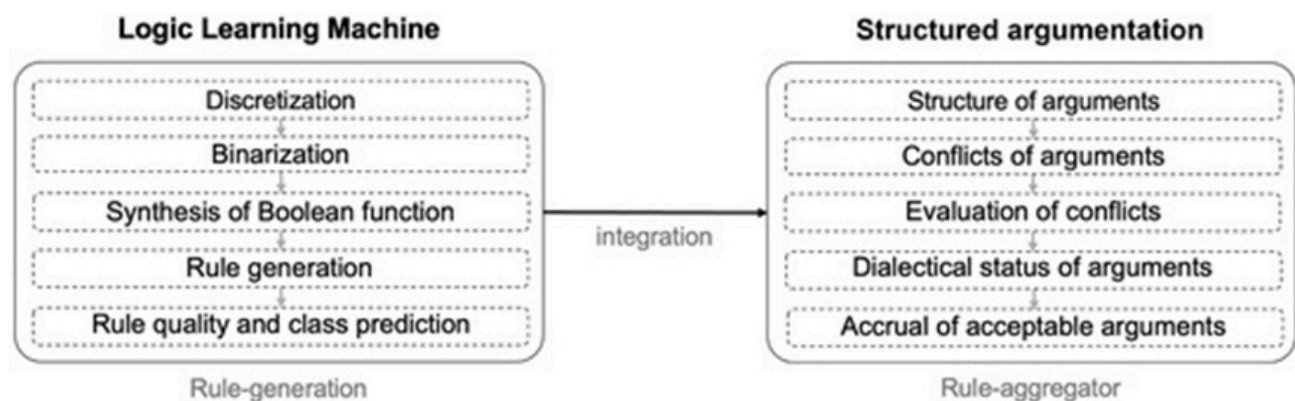


Figure 1. Illustration of the integration of a data-driven rule-generator (Logic Learning Machine) and a rule-aggregator with non-monotonic logic (structured argumentation).

Creation of the structure of arguments

The framework proposed in this study treats the extracted if-then rules directly as defeasible arguments. These arguments correspond to the logical statements produced via the workflow. A key implication of this modeling choice is that the resulting arguments are non-strict, meaning they are subject to retraction or cancellation when countered by conflicting evidence contained within opposing arguments.

Conflicts of arguments

To successfully implement a non-monotonic, defeasible logic system, the interactive dynamics and friction between arguments must be rigorously formalized. Because the rules synthesized by the LLM generator rely on a binary decision-making framework, each rule inevitably points toward one of two competing outcomes or claims, necessitating a clear resolution and exclusion framework. Within abstract argumentation theory, this dynamic can be formalised using the concept of an "attack"—a targeted structural expression of logical conflict. In this context, any argument that yields a particular conclusion inherently attacks any opposing argument that claims a different outcome. In literature, this specific conflict topology is classified as a rebuttal attack, which occurs when the thesis of one argument directly invalidates or negates the conclusion brought forward by another [40, 47].

This relationship forms an initial multipartite argumentation graph, where every argument supporting a given claim mounts an attack against all arguments aligned with a contradictory claim. An abstract example of this structure, featuring two potential output classes y_1 and y_2 , is provided in Figure 2.

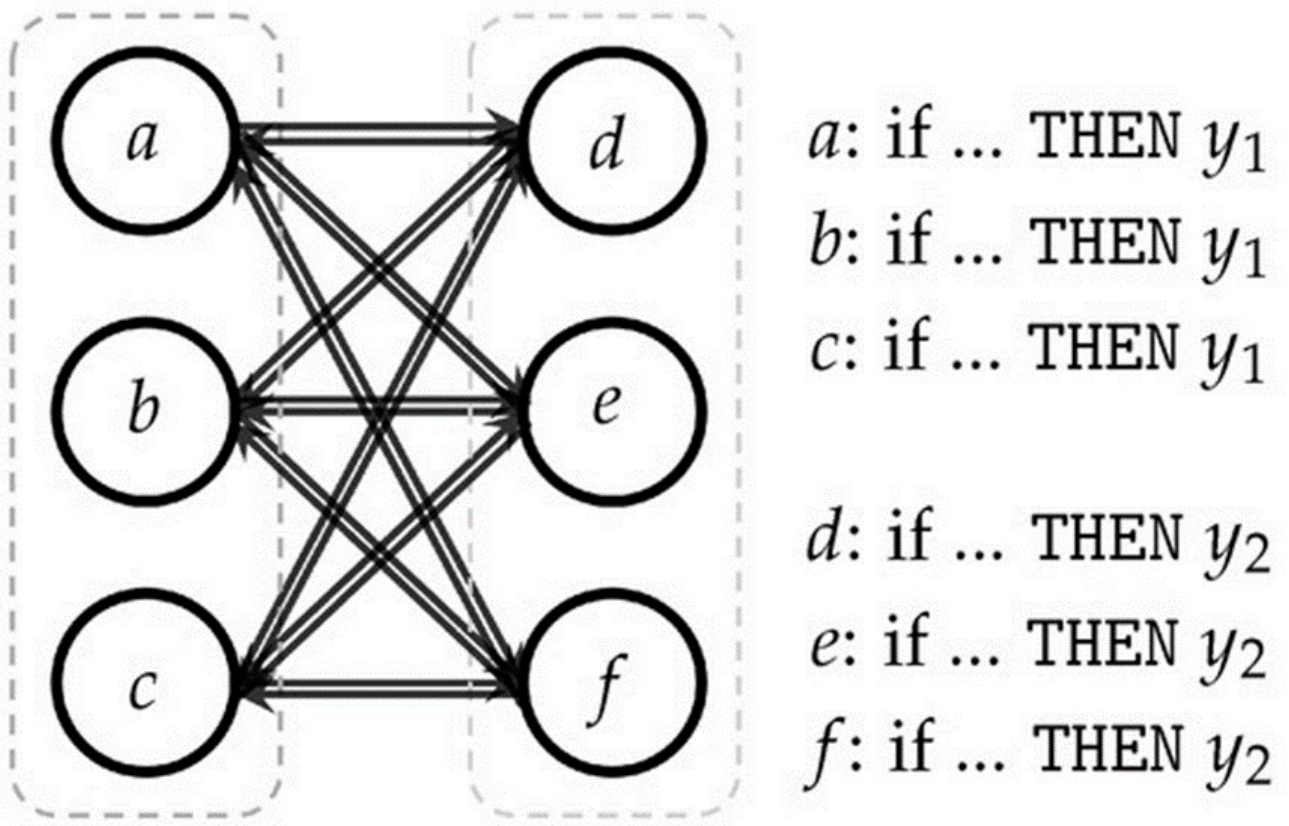


Figure 2. An illustration of a multipartite argumentation graph. A node represents each argument and has an if-then internal structure following Eq. 11 (premises are omitted for the sake of simplicity). Arguments a–c share a common output class, whereas arguments d–f share a different one. Each argument in a partite attacks all the other arguments in the other partite.

Evaluation of conflicts

The generation of an argumentation graph from logical if-then rules can be populated with empirical data, aligning with the concept of instantiation-based argumentation [48]. The activation or dismissal of specific arguments depends on the truth value of their underlying premises. Consequently, this process gives rise to an instantiated abstract argumentation framework, which constitutes a localized subset of the overarching multi-partite argumentation graph. As a result, the conflict relations (attacks) characterized within the comprehensive multi-partite graph are similarly activated. Specifically, attacks function as binary relations where an attack relation becomes active if and only if both the source argument (the attacker) and the target argument (the entity being attacked) are concurrently activated.

Because the multi-partite layout is deliberately structured around dual sets of arguments for binary classification tasks—where each set contains claims supporting an opposing outcome—the consequences of their friction result in two distinct possible outcomes. To refine how these opposing claims are assessed, this research incorporates the concept of an inconsistency budget [49–53]. This budget acts as a specific threshold defining the maximum level of conflict that can be tolerated. Under a designated inconsistency budget β , the framework only accommodates a collection of attacks if the cumulative sum of their respective weights does not exceed β .

To assign weights to these arguments, the relevance score derived from each if-then rule (Eq. 14) is utilized. Specifically, the weight assigned to any given attack relation is determined by summing the individual weights of its constituent source and target arguments. The total combined weight of all active attacks within an instantiated framework dictates the upper bound for β within the graph. When β is restricted, conflict relations must be systematically pruned until the total weight complies with the specified threshold, adhering strictly to the following three sequential priorities (ordered from highest to lowest significance):

- **P1:** Rebuttal attacks.
- **P2:** Attacks possessing a lower weight.
- **P3:** In instances of identical weights, the attack originating from the source argument with the lower weight.

Because the rules produced by the LLM methodology invariably construct multipartite configurations, the primary priority (**P1**) is established to mitigate the substantial volume of inconsistencies generated by mutual rebuttals. Stated differently, resolving a mutual rebuttal by eliminating an attack link is favored over removing a directional, one-way attack between two distinct arguments. Conversely, eliminating all internal conflicts would cause unresolved scenarios where conflicting arguments favoring different classifications are endorsed simultaneously. Following the mitigation of rebuttals, **P2** and **P3** function to prioritize the retention of more authoritative arguments, determined by the weight values calculated from rule coverage and error metrics. Once the permissible attacks under a given budget are established, their respective dialectical status can be computed as detailed in the subsequent section.

Dialectical status of arguments

The evaluation of how arguments defend themselves can be modeled using classical argumentation semantics tailored for Dung-style abstract argumentation frameworks [54]. At this stage of the inferential process, the framework is treated as a directed graph where the vertices (nodes) denote the arguments and the directed edges denote the attack relations, temporarily abstracting away the internal structure of the arguments and the metadata of the attacks. To characterize the dialectical standing of arguments within this network, this study utilizes two distinct categories of semantics: extension-based semantics [39, 54] and ranking-based semantics [55]. Their formal mathematical properties are detailed below.

Extension-Based Semantics

The principles of conflict-freeness and reinstatement serve as the core foundations for extension-based semantics, including standard Dung-style formulations [54] such as grounded and preferred semantics. The formal definitions presented below follow the conventions established in [56, 57].

Definition 1. Let $\langle Ar, \mathcal{R} \rangle$ be an argumentation framework (AF), $a, b \in Ar$ and $args \subseteq Ar$. The following notations are utilized:

- a^+ as $\{b \mid (a, b) \in \mathcal{R}\}$.
- $args^+$ as $\{a \mid (a, b) \in \mathcal{R} \text{ for some } b \in args\}$.
- a^- as $\{b \mid (b, a) \in \mathcal{R}\}$.
- $args^-$ as $\{b \mid (b, a) \in \mathcal{R} \text{ for some } a \in args\}$.

Here, a^+ denotes the set of arguments targeted by a , whereas a^- represents the set of arguments targeting a . Similarly, $args^+$ encompasses all arguments attacked by the set $args$, while $args^-$ details the collective set of arguments attacking $args$.

Definition 2.

Let $\langle Ar, \mathcal{R} \rangle$ be an AF and $args \subseteq Ar$. $args$ is **conflict-free** iff $args \cap args^+ = \emptyset$.

Definition 3.

Let $\langle Ar, \mathcal{R} \rangle$ be an AF, $A \in Ar$ and $args \subseteq Ar$. $args$ **defends** an argument A iff $A^- \subseteq args^+$.

Definition 4.

Let $\langle Arg, \mathcal{R} \rangle$ be an AF and $Lab: Arg \rightarrow \{in, out, undec\}$ be a labeling function. Lab is a **reinstatement labeling** iff it satisfies:

- $\forall A \in Ar: (Lab(A) = out \Rightarrow \exists B \in Ar: (B \text{ defends } A \wedge Lab(B) = in))$ and
- $\forall A \in Ar: (Lab(A) = in \Rightarrow \forall B \in Ar: (B \text{ defends } A \Rightarrow Lab(B) = out))$

Definition 5.

Let $Args$ be a conflict-free set of arguments, $F: 2^{Args} \rightarrow 2^{Args}$ a function such that $F(Args) = \{A \mid A \text{ is defended by } Args\}$ and $Lab: Args \rightarrow \{in, out, undec\}$ a reinstatement labeling function. Also consider (Lab) short for $\{A \in Args \mid Lab(A) = in\}$, $out(Lab)$ short for $\{A \in Args \mid Lab(A) = out\}$ and $undec(Lab)$ short for $\{A \in Args \mid Lab(A) = undec\}$.

- $Args$ is **admissible** if $Args \subseteq F(Args)$.

- $Args$ is a **complete** extension if $Args = F(Args)$.

(Lab) is a **grounded** extension if $undec(Lab)$ is maximal, or $in(Lab)$ is minimal, or $out(Lab)$ is minimal.

•

(Lab) is a **preferred** extension if $in(Lab)$ is maximal or $out(Lab)$ is maximal.

Argumentation semantics, including the grounded and preferred variants, offer a systematic approach for determining the justification or dialectical status of an argument. Under these formal frameworks, an argument achieves a justified status if it successfully counteracts its attackers. Extension-based criteria thus present a formal stance for categorizing arguments into accepted, rejected, or undecided categories. An extension (a coherent set of acceptable arguments) is capable of defending its components against external counterarguments and maintaining internal logical consistency, even when evaluated against competing perspectives [36]. The grounded semantics yields an extension that reflects a skeptical strategy, making minimal commitments and consistently identifying a unique extension. Conversely, preferred semantics reflects a more credulous strategy, exhibiting a bolder stance toward argument acceptance by allowing for the generation of multiple distinct extensions.

Ranking-Based Semantics

Ranking-based approaches represent another prominent category within abstract argumentation semantics [55]. Rather than evaluating acceptability through the binary isolation of argument extensions, this paradigm establishes a relative ordering among arguments. By avoiding the rigid categorization of arguments as strictly validated or dismissed, it offers a more nuanced, graded evaluation directly derived from the structural topology of the framework. This methodology effectively resolves several structural limitations inherent to extension-based paradigms. Specifically, it accounts for scenarios where multiple distinct attacks exert the same cumulative impact as a single attack, addresses the uniform valuation often assigned to valid arguments, and allows for direct ordinal comparisons between arguments that would otherwise share an identical status. The underlying formalisms detailed below follow the conventions established in [55].

Definition 6. A **ranking-based semantics** σ associates with any argumentation framework $AF = \langle Args, att \rangle$ a ranking $\succeq_{AF}^{\sigma}$ on $Args$, where $\succeq_{AF}^{\sigma}$ is a binary relation which is total ($\forall a, b \in Args, a \succeq_{AF}^{\sigma} b$ or $b \succeq_{AF}^{\sigma} a$) and transitive ($\forall a, b, c \in Args, a \succeq_{AF}^{\sigma} b$ and $b \succeq_{AF}^{\sigma} c$, then $a \succeq_{AF}^{\sigma} c$). $a \succeq_{AF}^{\sigma} b$ means that a is at least as acceptable as b and $a \succ_{AF}^{\sigma} b$ means a is strictly more acceptable than b .

As introduced in [58], ranking-based semantics compute a quantitative metric for each argument based on the density of its direct attackers. To achieve this, a categorizer function is mathematically defined as follows:

Definition 7. Let $\langle Args, t \rangle$ represent an argumentation framework. The mapping $Cat: Args \rightarrow (0, 1]$ denotes the categorizer function, formulated as:

$$Cat(a) = \begin{cases} 1 & \text{if } a^- = \emptyset \\ \frac{1}{1 + \sum_{c \in a^-} Cat(c)} & \text{otherwise} \end{cases} \quad (18)$$

Definition 8. Given an argumentation framework $\langle Args, t \rangle$ and the categorizer function $Cat: Args \rightarrow (0, 1]$, a ranking-based categorizer semantics determines an ordering $\succeq_{AF}^{Cat}$ over $Args$ such that $\forall a, b \in Args, a \succeq_{AF}^{Cat} b$ iff $Cat(a) \geq Cat(b)$.

Accrual of acceptable arguments and final inference

The concluding phase of this defeasible reasoning pipeline requires the generation of a definitive inference. Two straightforward scenarios can resolve this process: (i) an extension-based semantics yields a single extension in which all constituent arguments uniformly endorse the same claim, or (ii) the highest-ranked arguments identified by a ranking-based semantics converge on an identical claim. In both situations, the system designates the claim backed by these accepted or top-ranked arguments as the final inference. If neither condition is satisfied, the framework applies the following specialized, ad hoc resolution strategies depending on the active semantic model:

• Extension-based semantics:

- If an extension contains conflicting arguments that support divergent claims, it is discarded from the final decision-making process because it fails to offer a coherent, singular perspective. Such an outcome typically indicates that the selected inconsistency budget was overly restrictive. Consequently, the final inferential status is marked as undecided.
- In cases where the system generates multiple valid extensions, their relative validity is measured by the total count of accepted arguments within each extension. This step intentionally omits individual argument weights, as they have already been factored into the initial attack relations. This simplified, constrained approach serves to mitigate undecided deadlocks when utilizing credulous paradigms (such as preferred semantics) that routinely yield multiple extensions. Ultimately, the defeasible reasoning sequence terminates without a decision if all competing extensions contain an identical number of arguments.

• Ranking-Based semantics:

- The final inference is classified as undecided if the categorizer function produces an identical maximum score for top-ranked arguments that back opposing claims.

An illustrative example of argument generation across varying inconsistency budgets, along with the corresponding dialectical status under different semantics, is displayed in **Figure 3**. Only the configuration in graph 3 generates non-empty extensions where the active arguments converge on the same claim. Similarly, under the categorizer semantics, only graphs 1 and 3 yield top-ranked arguments that support an identical conclusion.

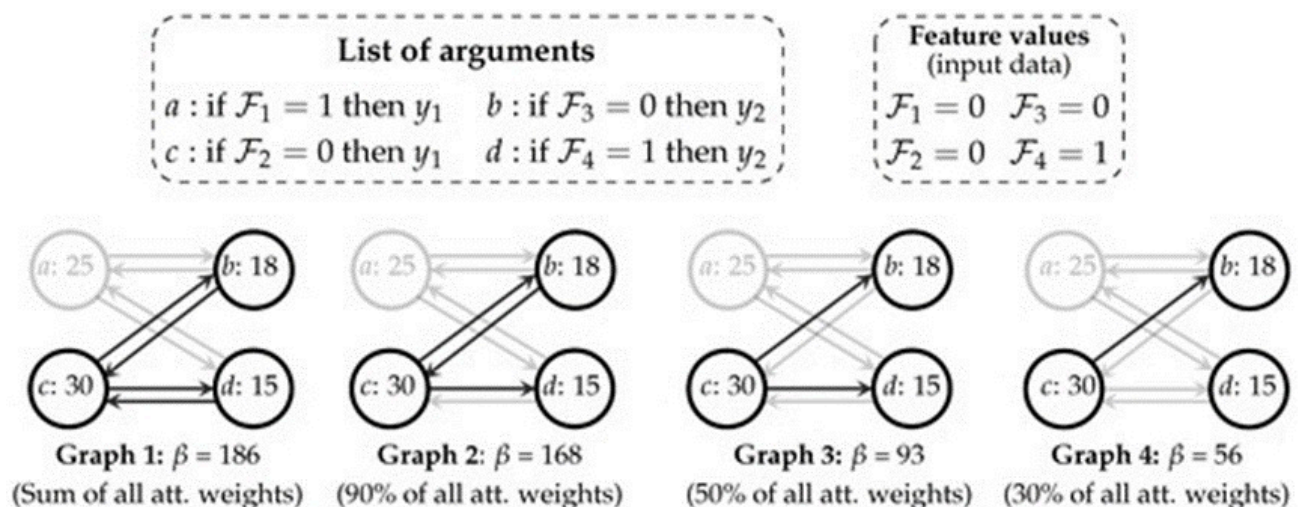


Figure 3. An illustrative example of elicitation of arguments and the definition of their dialectical status. Node labels contain the argument label and its weight. The premise of argument a does not hold true with the input data, so it is discarded along with its incoming/outgoing attacks. For graphs 1, 2, 3, and 4, the following can be observed, respectively: attacks $\{\emptyset\}$, $\{d \rightarrow c\}$, $\{d \rightarrow c, c \rightarrow b\}$, and $\{d \rightarrow c, c \rightarrow d, c \rightarrow b\}$ are removed to respect the inconsistency budget defined; the

grounded extensions are $\{\emptyset\}$, $\{\emptyset\}$, $\{c\}$, $\{c,d\}$, and the preferred extensions are $\{\{c\}, \{b,d\}\}$, $\{\{c\}, \{b,d\}\}$, $\{c\}$, $\{c,d\}$; the top ranked arguments for the categoriser are $\{b,d\}$, $\{b,c,d\}$, $\{c\}$, $\{c,d\}$.

Design and Methods

This research focuses on maximizing automation when constructing argument-based models from raw data, completely eliminating the need for manual curation. This automated architecture is realized by deploying a specialized Logic Learning Machine (LLM) rule-extractor [16] to synthesize logical if-then rules directly from data. Subsequently, formal computational argumentation methodologies are leveraged to organize these extracted rules into a structured argumentation framework [7]. The resulting multipartite graph is hypothesized to offer significantly higher explainability and interpretability compared to unorganized rule lists.

This core hypothesis is evaluated by comparing the predictive performance of the baseline LLM method against the integrated computational argumentation pipeline. To verify the generalizability of this approach across varied binary classification tasks, the pipeline is evaluated using four open-access benchmark datasets of diverse sample sizes and structural dimensionalities. The formal hypothesis is structured as follows:

Hypothesis 1. IF defeasible argument-based models are built with if-then rules generated by the LLM technique AND their conflictuality is handled with argument-based semantics (preferred, grounded, categorizer), then, these models will achieve similar predictive capacity (recall, precision, accuracy, Youden's index, % of undecidedness) to those built with the Standard Applied Procedure on the same rule sets for four distinct binary classification tasks.

The experimental design constructed to validate this hypothesis is mapped out in **Figure 4**, which integrates the corresponding evaluation phases directly into the process flow.

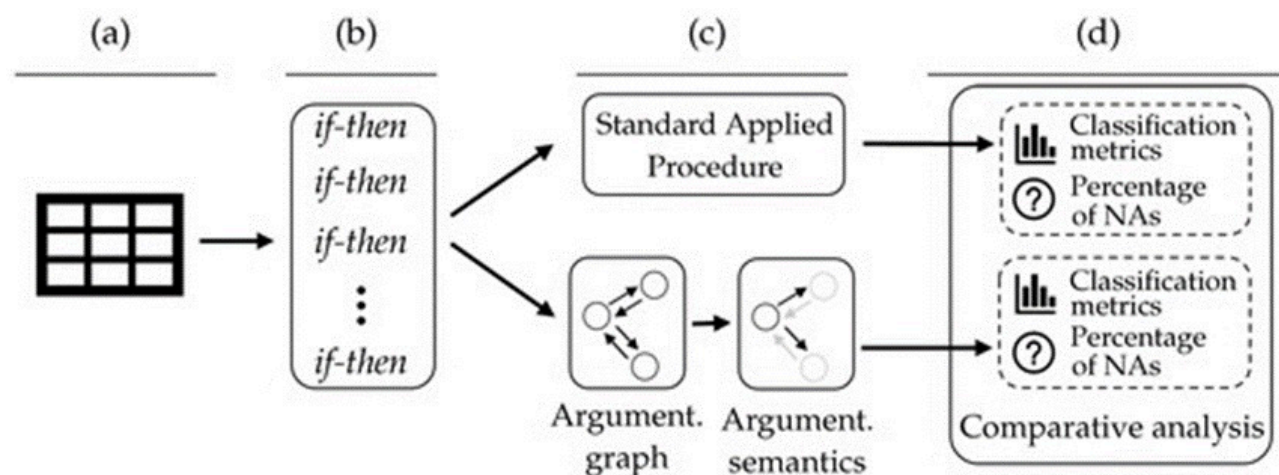


Figure 4. Design of a comparative experiment with four main steps: (a) Selection and pre-processing of four datasets for binary classification tasks; (b) Automatic formation of if-then rules from the selected dataset using the Logic Learning Machine (LLM) technique; (c) Generation of final inferences using two aggregator logics rules: the Standard Applied Procedure and computational argumentation; (d) Comparative analysis via standard binary classification metrics and percentage of undecided cases (NAs: when a model cannot lead to a final inference).

Four benchmark datasets were acquired from the UCI Machine Learning Repository for this empirical study (and accessed on 6 June 2024): CARS, CENSUS, BANK, and MYOCARDIAL. These repositories serve as standard benchmarks within the machine learning domain. They were selected due to their open-access availability, highly divergent feature spaces and sample sizes (Table 1), and skewed class distributions—a characteristic that replicates real-world data imperfections.

Minimal preprocessing was performed to unify the format of these datasets. For the CARS repository, the target class outcomes *acc*, *good*, and *vgood* were consolidated into a single *acc* class to formulate a binary classification problem. For the CENSUS dataset, any rows containing missing attributes across any features were purged. The BANK dataset was utilized without any prior modifications. Within the MYOCARDIAL dataset, the ZSN (chronic heart failure) variable was designated as the target feature; attributes containing missing data exceeding a 5% threshold were eliminated, and no data interpolation was implemented. **Table 1** summarizes the structural dimensions of the finalized datasets utilized in the subsequent testing phases.

Table 1. Characteristics of final datasets formed for subsequent rule-formation and aggregation.

Dataset	MYOCARDIAL	BANK	CENSUS	CARS
Number of features	59	16	14	6
Number of records	1436	45211	30162	1728
Class distribution (positive vs negative)	24–76%	29–71%	25–75%	30–70%
Feature types	Numerical & categorical	Numerical & categorical	Numerical & categorical	Numerical & categorical
Target variable	ZSN	y	Income	Class

The Logic Learning Machine architecture is executed using the proprietary Rulex Platform software. The computational argumentation pipeline is implemented using the open-source ArgFrame framework [38]. The raw textual if-then rules generated by the LLM pipeline are processed via a custom Python parser to convert them into the JSON format required by the ArgFrame system.

The predictive models derived from the baseline LLM method (**Table 2**) and the argument-based frameworks (**Table 3**) are evaluated using standard classification metrics. Performance is benchmarked via accuracy, precision, recall, Youden’s index, and the percentage of undecided instances to ensure a multi-dimensional assessment. Two key components are prioritized for each dataset: (i) the physical size of the resulting models and (ii) their overall inferential capacity. Model size serves as a quantitative proxy for explainability, based on the premise that compact information repositories are inherently more interpretable than massive rule sets. Meanwhile, inferential capacity explicitly reflects the predictive power and accuracy of the models.

Table 2. List of models built with the LLM technique.

Model	Error Threshold per Rule	Maximum Number of Premises per Rule	Conflict Resolution Strategy
H1	10%	4	Standard Applied Procedure
H2	25%	4	Standard Applied Procedure

Table 3. List of models built with argumentation and if-then rules from the LLM technique. The inconsistency budget is defined as a percentage of the sum of all attack weights.

Model	Input Rules	Semantics	Inc. Budget
AG1/AP1/AC1	Same as <i>H1</i>	Grounded/Preferred/Categoriser	25%
AG2/AP2/AC2	Same as <i>H1</i>	Grounded/Preferred/Categoriser	50%
AG3/AP3/AC3	Same as <i>H1</i>	Grounded/Preferred/Categoriser	90%
AG4/AP4/AC4	Same as <i>H1</i>	Grounded/Preferred/Categoriser	100%
AG5/AP5/AC5	Same as <i>H2</i>	Grounded/Preferred/Categoriser	25%
AG6/AP6/AC6	Same as <i>H2</i>	Grounded/Preferred/Categoriser	50%
AG7/AP7/AC7	Same as <i>H2</i>	Grounded/Preferred/Categoriser	90%
AG8/AP8/AC8	Same as <i>H2</i>	Grounded/Preferred/Categoriser	100%

To guarantee a rigorous performance analysis and assess the generalizability of the results, specific execution parameters for both methods (LLM and computational argumentation) are systematically varied. In the LLM configuration, individual rules are restricted to a maximum of 4 premises to preserve human readability. Additionally, the maximum permissible rule error threshold is tested at two distinct levels: 10% and 25%. Introducing this error tolerance helps mitigate overfitting and optimizes the rule count, directly supporting the model's interpretability. Furthermore, these error boundaries allow the argument-based models to resolve contradictions efficiently while maintaining strong inferential capacity with fewer total rules.

For the argument-based configurations, three distinct semantic evaluations are carried out: (1) preferred semantics, reflecting a credulous model common in defeasible argumentation; (2) grounded semantics, offering a conservative, skeptical model; and (3) categorizer semantics, incorporating a ranking-based paradigm into the matrix. Furthermore, multiple thresholds are established for the maximum inconsistency budget, since the rules generated by the LLM technique are structurally prone to creating numerous mutual rebuttals. Setting an inconsistency budget is therefore critical to prevent excessive rates of unresolvable or undecided instances. Three explicit budget boundaries are evaluated: 25% (an aggressive restriction focusing only on critical attacks), 50% (a balanced approach), and 90% (a permissive approach designed to preserve structural connections by removing only a minimal number of attacks). **Tables 2 and 3** detail the complete list of models and their respective parameter values.

Results and Discussion

This section categorizes the experimental outcomes according to the four designated datasets. Each dataset undergoes an individual assessment focused on predictive performance (evaluated via recall, precision, accuracy, Youden's index, and the percentage of undecided instances) alongside structural model dimensions (quantified by the volume of rules/arguments and corresponding attacks). Finally, a comprehensive synthesis presents overarching insights, with a particular emphasis on the explainability of the models.

CARS dataset

The total volume of rules and attacks derived from the CARS dataset is detailed in **Table 4**, organized by the two distinct error thresholds applied. The resulting frameworks remain relatively compact regarding the number of established rules, even when a high β value is introduced. **Figure 5** illustrates the outcomes associated with the selected classification metrics. Notably, the maximum error threshold appeared to exert no substantial influence on the outcomes (comparing

cases a–d against e–h in Figure 5), demonstrating that the rules generated via the LLM methodology maintained stability even at the highest allowable parameter.

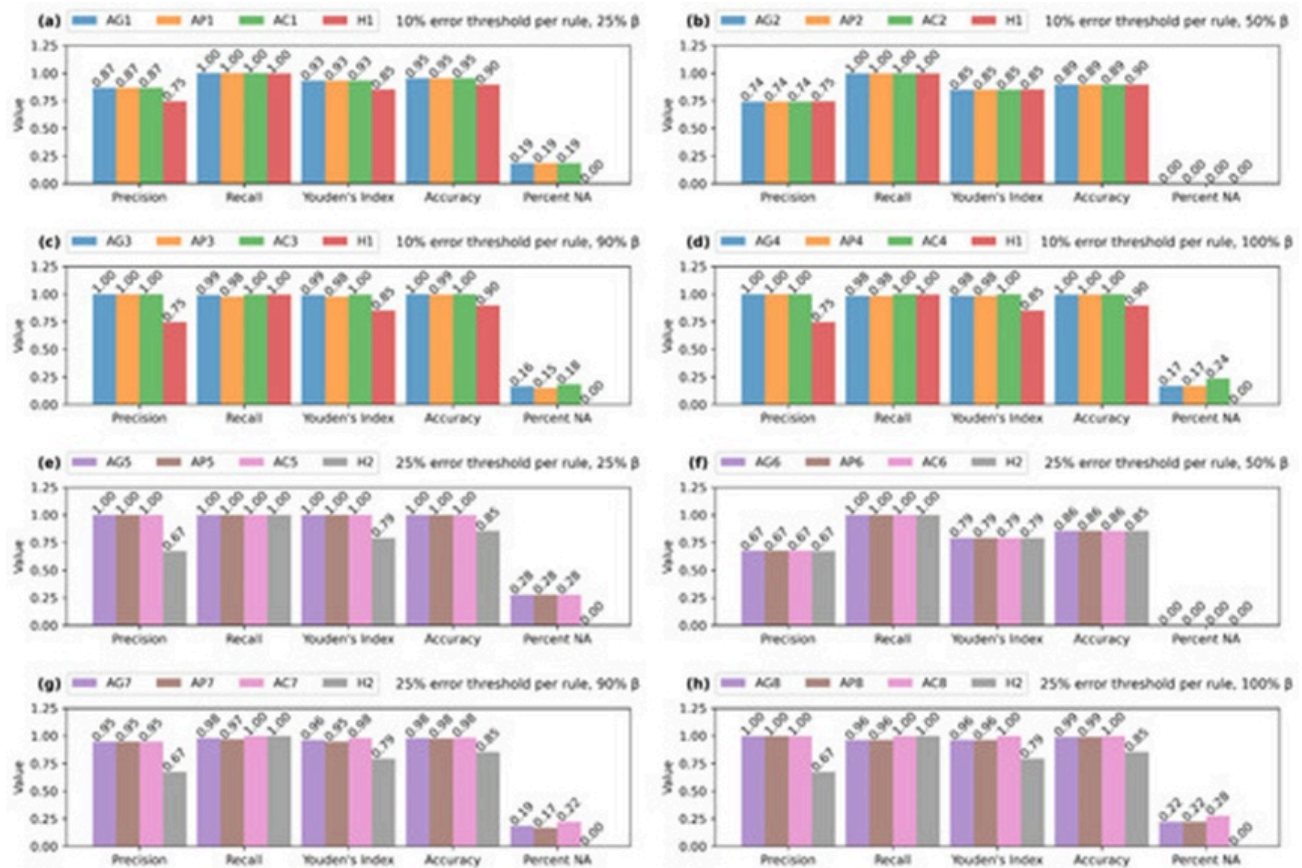


Figure 5. Overall results for inferences produced using the CARS dataset grouped by the error threshold per rule parameter (10% for the first four blocks (a–d), 25% for the second four blocks (e–h)), and divided by inconsistency budget variation (25%, 50%, 90%, 100%).

Table 4. Summary of rules count and attacks count based on error threshold per rule and the inconsistency budget β using the data from the CARS dataset.

Error Threshold per Rule	Number of Attacks ($\beta 100\%$)	Number of Attacks ($\beta 90\%$)	Number of Attacks ($\beta 50\%$)	Number of Attacks ($\beta 25\%$)	Average Number of Premises	Number of Rules
10%	36	29	17	6	2	9
25%	10	8	4	2	1.5	6

Regarding the argument-based frameworks, configuring the inconsistency budget β to 50% successfully resolved every instance of undecided inferences (see cases b and f in Figure 5). Nevertheless, all alternative configurations of β yielded inferences with a maximum of 28% undecided cases, while simultaneously matching or exceeding the baseline performance metrics (precision, recall, accuracy, and Youden's index) of the comparative baseline models (H1 or H2). This indicates that the argument-based frameworks are capable of adequately managing the viability of the if-then rules

regardless of the specific argumentation semantics utilized, owing to the minimal rate of undecidedness. When non-zero percentages of undecided instances occurred, an enhancement in overall model performance was also discernible (comparing case b against {a,c,d}, and case f against {e,g,h} in [Figure 5](#)).

CENSUS dataset

As indicated in [Table 5](#), the volume of rules and attacks produced by the LLM methodology for the CENSUS dataset was notably greater than the quantities observed for the CARS dataset. This outcome aligns with expectations given the larger feature set characteristic of this dataset. Furthermore, the inconsistency budget β succeeded in substantially decreasing the volume of attacks ([Table 5](#)), thereby rendering certain argument-based models more streamlined than others.

Table 5. Summary of rules count and attacks count based on error threshold per rule and the inconsistency budget β using the CENSUS dataset.

Error Threshold per Rule	Number of Attacks ($\beta 100\%$)	Number of Attacks ($\beta 90\%$)	Number of Attacks ($\beta 50\%$)	Number of Attacks ($\beta 25\%$)	Average Number of Premises	Number of Rules
10%	480	357	240	54	2.74	31
25%	96	74	48	12	2.14	14

In terms of inferential performance, [Figure 6](#) displays the outcomes for the CENSUS dataset. Consistent with the CARS dataset observations, the maximum error threshold per rule exerted no pronounced effect on overall model performance (comparing cases a–d against e–h in [Figure 6](#)). Additionally, it is important to note the stable inferences generated by argument-based models utilizing grounded semantics coupled with a high β , though this occurred at the expense of an increased proportion of undecided instances (cases c, d, g, and h in [Figure 6](#)). This demonstrates that the identification of undecided cases remained functional for this dataset; however, particular percentages proved excessively high for the models to be considered practical (ranging between 60% and 93%). This elevated proportion of undecided cases reflects a denser concentration of rules extracted per record within this dataset, highlighting the need for enhanced conflict resolution strategies. Setting lower β values succeeded in mitigating these instances while keeping the remaining performance metrics (excluding precision) close to those of the baseline models (H1 or H2). The decrement in precision is presumably linked to class imbalance within the data, indicating that the argument-based frameworks were less optimal regarding positive predictions. Nonetheless, these models delivered acceptable outcomes, suggesting that performing a grid search across a range of inconsistency budgets serves as a viable approach for optimizing conflict resolution in large, highly contested environments.

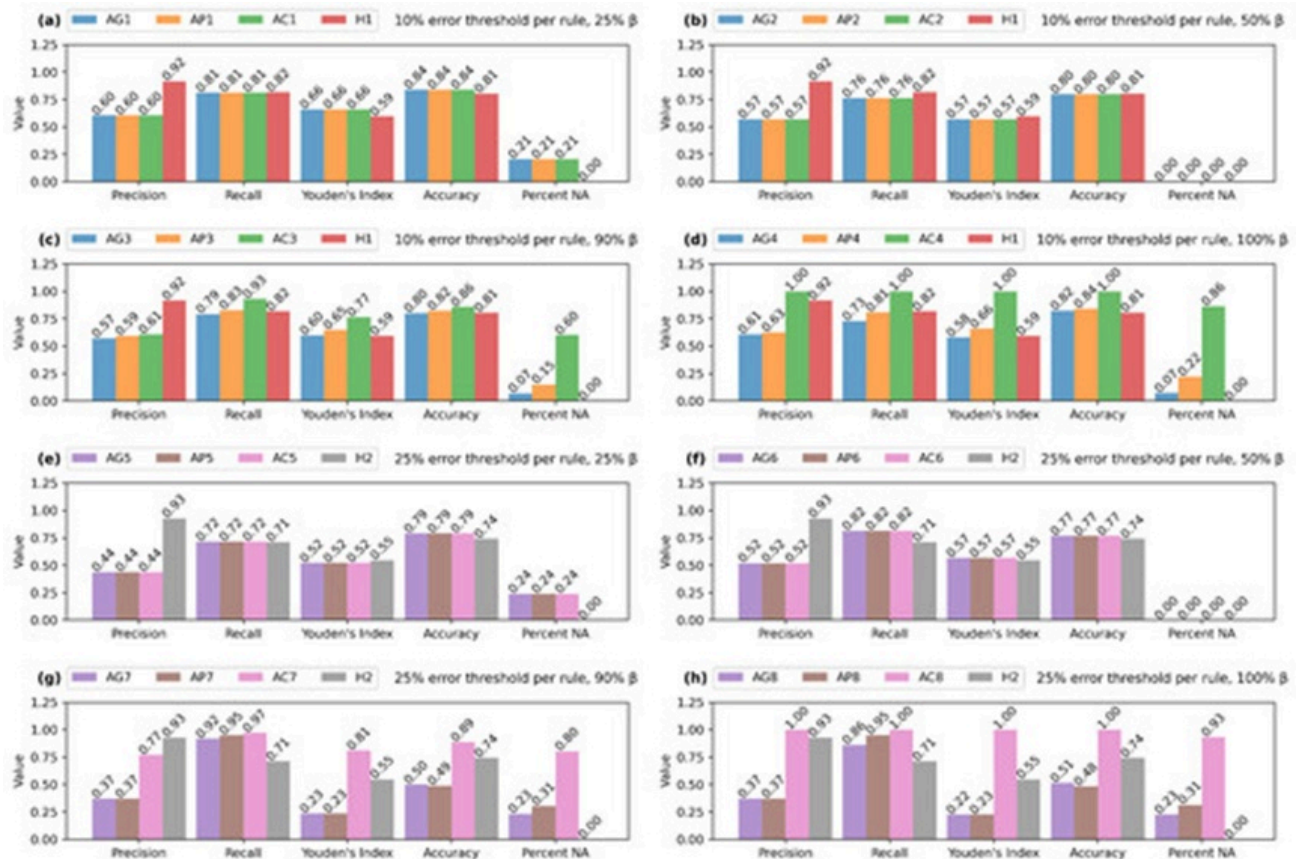


Figure 6. Overall results for inferences produced using the CENSUS dataset grouped by the error threshold per rule parameter (10% for the first four blocks (a–d), 25% for the second four blocks (e–h)) and divided by inconsistency budget variation (25%, 50%, 90%, 100%).

BANK dataset

The outcomes connected to the data from the BANK dataset are compiled in [Table 6](#) and [Figure 7](#). Several identical trends identified in the evaluation of the prior two datasets remain apparent. These include an elevated quantity of rules and attacks driven by the expanded feature set, a minimized attack count at lower β levels, and an increased proportion of undecided instances at elevated β levels.

The primary distinction occurs within the argument-based frameworks when the inconsistency budget β is set precisely to 50%. In this specific scenario, all undecided instances were completely reduced to 0, yielding performance metrics that closely align with the baseline models (H1 or H2). This confirms that the argumentation semantics successfully resolved conflicting states without compromising inferential capacity when utilizing a 50% β . This finding is significant as it reinforces the practical utility of the if-then rules extracted by the proposed argument-based models.

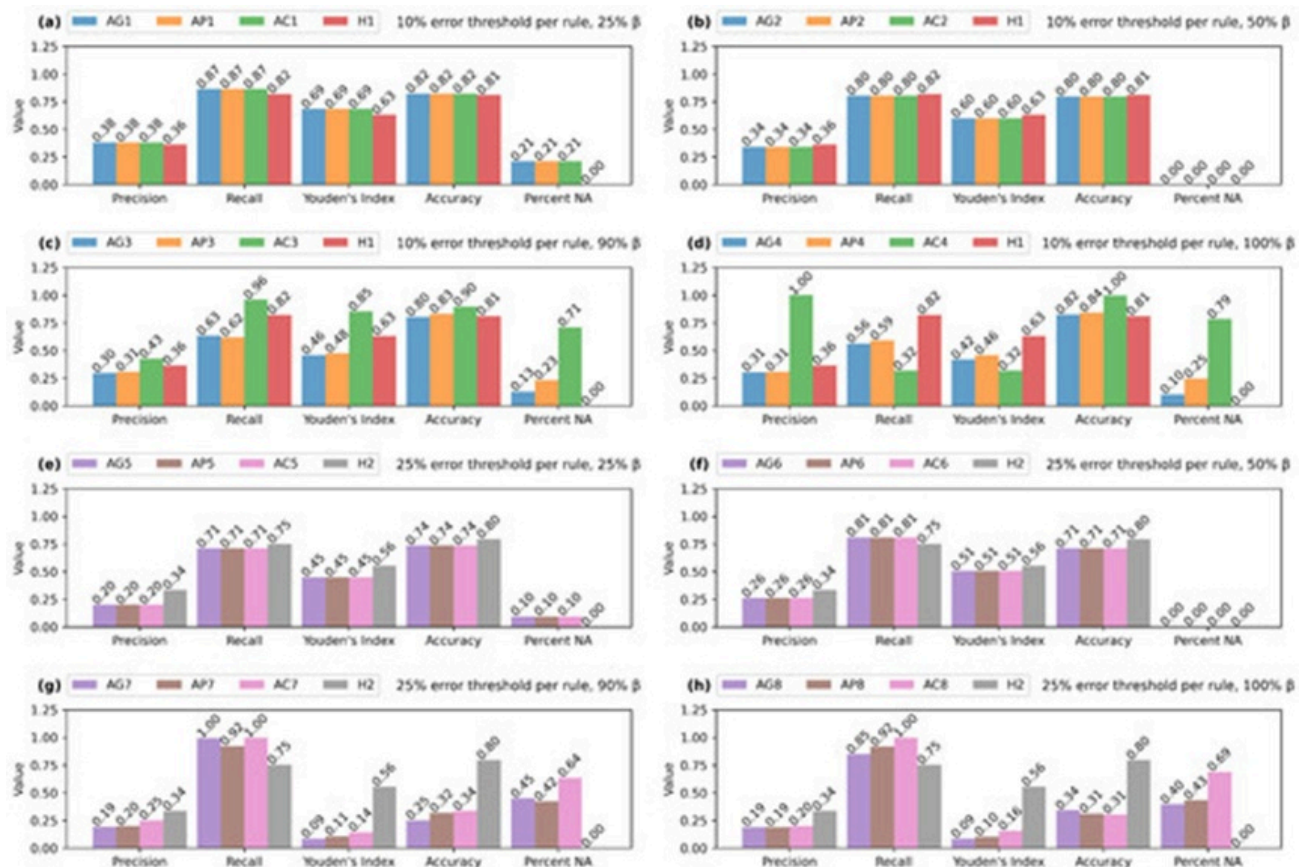


Figure 7. Overall results for inferences produced using the BANK dataset grouped by the error threshold per rule parameter (10% for the first four blocks (a–d), 25% for the second four blocks (e–h)), and divided by inconsistency budget variation (25%, 50%, 90%, 100%).

Table 6. Summary of rules count, and attacks count based on error threshold per rule and the inconsistency budget β using the BANK dataset.

Error Threshold per Rule	Number of Attacks ($\beta 100\%$)	Number of Attacks ($\beta 90\%$)	Number of Attacks ($\beta 50\%$)	Number of Attacks ($\beta 25\%$)	Average Number of Premises	Number of Rules
10%	468	367	234	61	2.74	31
25%	112	86	56	13	2.85	15

MYOCARDIAL dataset

Lastly, even though the MYOCARDIAL dataset contains the most extensive set of features, the resulting number of rules and attacks (Table 7) did not exhibit a marked increase relative to the previous datasets. This implies that the LLM approach did not exploit all available attributes, yielding rule structures and premises comparable to those observed previously. Figure 8 illustrates the inferential capacity of the models, revealing consistent patterns found across the other datasets. These include the finding that the maximum permitted error threshold per rule does not alter performance; that robust inferences are generated by argument-based models using grounded semantics with elevated β values; and that argumentation semantics successfully mitigate conflicting states without degrading inferential power when a 50% β is

used. The application of lower β values to minimize undecided outcomes while preserving robust metrics across all other dimensions is especially noteworthy. The recurrence of these consistent trends across distinct datasets confirms the generalizability of these insights, which are evaluated further in the subsequent section.

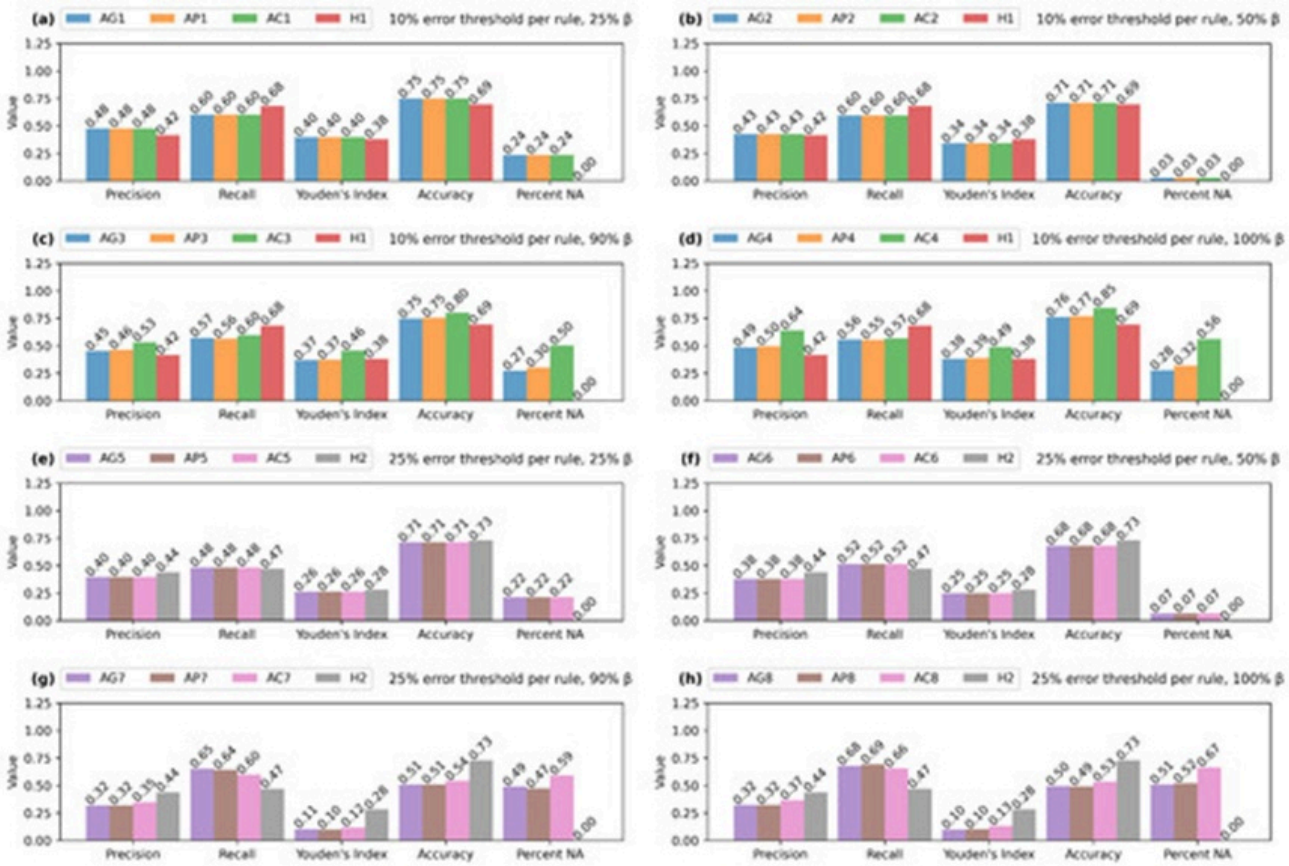


Figure 8. Overall results for inferences produced using the MYOCARDIAL dataset grouped by the error threshold per rule parameter (10% for the first four blocks (a–d), 25% for the second four blocks (e–h)), and divided by inconsistency budget variation (25%, 50%, 90%, 100%).

Table 7. Summary of rules count and attacks count based on error threshold per rule and the inconsistency budget β using the MYOCARDIAL dataset.

Error Threshold per Rule	Number of Attacks (β 100%)	Number of Attacks (β 90%)	Number of Attacks (β 50%)	Number of Attacks (β 25%)	Average Number of Premises	Number of Rules
10%	540	458	270	102	2.57	33
25%	98	81	49	16	1.71	14

Summary and discussion

The LLM methodology for extracting if-then rules demonstrated remarkable efficiency across all models synthesized from the four designated datasets, particularly regarding rule volume, the mean number of premises, and overall model performance. Generating argument-based frameworks from these rule sets yielded minor improvements in the predictive

strength of the models developed for the CARS dataset, which possessed a restricted number of features. Conversely, the remaining datasets characterized by higher feature dimensionality demanded additional fine-tuning. This necessity arose primarily from the multipartite architecture of the underlying argumentation networks (graphs) established by the extracted if-then rules, which triggered widespread contradictions when multiple rules were simultaneously activated. The primary technique utilized to mitigate these active conflicts was the inconsistency budget parameter, β . This specific metric establishes a priority hierarchy among attacks based on argument relevance, which is computed via the LLM approach using the coverage and error margins of the rules. The experimental outcomes indicated that modulating β successfully eliminated less significant attacks, thereby minimizing undecided inferences and enhancing model accuracy. This highlights the indispensable function of the inconsistency budget when unified with the LLM approach.

On the whole, regarding predictive capacity, the argument-based frameworks constructed from the if-then rules of the Logic Learning Machine and structured through the argumentation pipeline performed comparably to models utilizing the identical rule set but unified via the Standard Applied Procedure. Realizing this performance parity was the primary requirement to validate the initial hypothesis that argument-based frameworks can deliver equivalent inferential capacity while remaining inherently more explainable than basic, unstructured rule lists. This superior explainability is derived from the dialectical classifications attributed to arguments—specifically accepted, rejected, or undecided—coupled with the highly interpretable workflows utilized to generate final inferences. These underlying mechanisms correspond more closely with human cognitive processes when navigating uncertainty and contradictory data. Furthermore, argumentation graphs are arguably more compact and visually accessible than a standard list of rules, potentially increasing both the explainability and interpretability of the inferential systems. Although explainability and interpretability represent two attributes that were not empirically quantified but rather treated as target ideals, future user studies could evaluate these dimensions more rigorously by utilizing psychometric assessments, as demonstrated in [59].

Another vital feature of these argumentation graphs is their adherence to the GLOCAL characteristic of AI systems outlined within the eXplainable Artificial Intelligence domain [4, 60]. Consequently, they accommodate a global perspective and conceptual understanding of the inferential workflows while concurrently providing local explanations through the evaluation of single-activated arguments from the input data along with their interconnected dependencies. Furthermore, argumentation graphs serve as inferential instruments for articulating contrastive explanations [6]. By leveraging the concept of attacks occurring between conflicting arguments (the rules), contrastive explanations can be practically deployed. This strategy enables contrastive scenarios to be offered as complementary justifications; it clarifies why a specific final inference constitutes the most rational deduction relative to alternative choices. The framework clarifies the underlying rationale validated by accepted arguments and contrasts it against the rationale dismissed by rejected arguments. Correspondingly, when accepted arguments validate conflicting assertions—thereby generating a contrastive scenario—an argumentation graph facilitates an understanding of why a definitive inference cannot be reached, which ultimately precipitates a state of undecidedness. This condition can be counteracted via human oversight and domain expertise. For instance, by evaluating the accepted arguments and their corresponding attacks, a human analyst can investigate alternative hypotheses by manually omitting, adjusting, or introducing arguments to monitor the resulting shifts in the final inference. This flexibility is supported by the absence of topological constraints on the argumentation graphs, which allows external knowledge to be seamlessly integrated.

Conclusion

This study demonstrated the feasibility of integrating the Logic Learning Machine (LLM), a specialized rule generation engine, with argument-based frameworks, an explicit paradigm designed for modeling contradictions among rules. Accordingly, this research addresses a core challenge within XAI: the development of data-driven inferential methodologies that deliver transparent, interpretable outcomes while preserving scalability and efficacy in complex domains. Specifically, the practical insights were derived from an empirical assessment comparing the Standard Applied Procedure—the baseline conflict resolution mechanism used by Logic Learning Machines—against diverse argument-based frameworks constructed from LLM-generated rules using a selection of widely recognized benchmarking datasets.

The fundamental objective was to leverage a resilient, general-purpose procedure for extracting rules from data while resolving their internal contradictions through computational argumentation, a paradigm specialized for executing defeasible reasoning and managing conflicts rationally.

Additionally, the dialectical status assigned to each individual argument within these argumentation frameworks streamlined the evaluation of the global and local interpretability of these argument-based models. The study demonstrated how contrastive explanations can be systematically formulated. The comparative analysis revealed that the LLM functions as a stable, robust architecture for generating rules that can be successfully adapted into arguments, even under the modification of parameters linked to the argument-based models. The findings also indicated that argument-based models can match the predictive power of the standard LLM methodology while offering more explainable inferential workflows that align more closely with human reasoning under uncertainty. This methodology is applicable to any binary classification problem. While the datasets evaluated in this research belong to the financial, medical, and automotive sectors, the insights are not confined to these domains and can be extrapolated more broadly. Future research directions involve scaling these experiments to more complex, non-binary classification tasks and optimizing the systemic integration between the Logic Learning Machine and computational argumentation frameworks. In the wider context of XAI, it would likewise be advantageous to explicitly contextualize this proposed integration of LLM and computational argumentation relative to alternative transparent and interpretable methodologies, including those outside the scope of traditional knowledge-based systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Aug 2023

Revised: 26 Sep 2023

Accepted: 10 Nov 2023

Published online: 15 March 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion*. 2023;99:101805.
<https://doi.org/10.1016/j.inffus.2023.101805>.
- Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655.
<https://doi.org/10.1016/j.jbi.2020.103655>.
- van der Waa J, Nieuwburg E, Cremers A, Neerincx M. Evaluating XAI: a comparison of rule-based and example-based explanations. *Artif Intell*. 2021;291:103404.
<https://doi.org/10.1016/j.artint.2020.103404>.
- Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67.
<https://doi.org/10.1038/s42256-019-0138-9>.
- Cao J, Zhou T, Zhi S, Lam S, Ren G, Zhang Y, et al. Fuzzy inference system with interpretable fuzzy rules: advancing explainable artificial intelligence for disease diagnosis—a comprehensive review. *Inf Sci*. 2024;662:120212.
- Miller T. Contrastive explanation: a structural-model approach. *Knowl Eng Rev*. 2021;36:e14.
<https://doi.org/10.1017/S0269888921000102>.
- Besnard P, Garcia A, Hunter A, Modgil S, Prakken H, Simari G, et al. Introduction to structured argumentation. *Argument Comput*. 2014;5(1):1-4.
<https://doi.org/10.1080/19462166.2013.869764>.
- Atkinson K, Baroni P, Giacomin M, Hunter A, Prakken H, Reed C, et al. Towards artificial argumentation. *AI Mag*. 2017;38(3):25-36.
- Tompits H. A survey of non-monotonic reasoning. *Open Syst Inf Dyn*. 1995;3:369-95.
- Lindström S. A semantic approach to nonmonotonic reasoning: inference operations and choice. *Theoria*. 2022;88:494-528.
- Brewka G. Nonmonotonic reasoning: logical foundations of commonsense. Cambridge: Cambridge University Press; 1991. 168 p.
- Sklar EI, Azhar MQ. Explanation through argumentation. In: *Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018*. New York: ACM; 2018. p. 277-85.
<https://doi.org/10.1145/3284432.3284473>.
- Vassiliades A, Bassiliades N, Patkos T. Argumentation and explainable artificial intelligence: a survey. *Knowl Eng Rev*. 2021;36:e5.
<https://doi.org/10.1017/S0269888921000011>.
- Rizzo L, Longo L. Comparing and extending the use of defeasible argumentation with quantitative data in real-world contexts. *Inf Fusion*. 2023;89:537-66.
<https://doi.org/10.1016/j.inffus.2022.08.025>.
- Vilone G, Longo L. A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. *Front Artif Intell*. 2021;4:160.
<https://doi.org/10.3389/frai.2021.675573>.
- Muselli M, Ferrari E. Coupling logical analysis of data and shadow clustering for partially defined positive Boolean function reconstruction. *IEEE Trans Knowl Data Eng*. 2009;21(10):1482-94.

Bennetot A, Franchi G, Del Ser J, Chatila R, Díaz-Rodríguez N. Greybox XAI: a neural-symbolic learning framework to produce interpretable predictions for image classification. *Knowl Based Syst*. 2022;258:109947.
<https://doi.org/10.1016/j.knosys.2022.109947>.

Rizzo L. A novel structured argumentation framework for improved explainability of classification tasks. In: Longo L, editor. *Explainable Artificial Intelligence*. Cham: Springer; 2023. p. 399-414.

Ferrari E, Verda D, Pinna N, Muselli M. A novel rule-based modeling and control approach for the optimization of complex water distribution networks. In: Valle M, Lehmhus D, Gianoglio C, Ragusa E, Seminara L, Bosse S, editors. *Advances in System-Integrated Intelligence*. Cham: Springer; 2023. p. 33-42.

Nicoletta M, Zilich R, Masi D, Baccetti F, Nreu B, Giorda CB, et al. Overcoming therapeutic inertia in type 2 diabetes: exploring machine learning-based scenario simulation for improving short-term glycemic control. *Mach Learn Knowl Extr*. 2024;6:420-34.

Parodi S, Manneschi C, Verda D, Ferrari E, Muselli M. Logic learning machine and standard supervised methods for Hodgkin's lymphoma prognosis using gene expression data and clinical variables. *Health Inform J*. 2018;24(1):54-65.
<https://doi.org/10.1177/1460458216655188>.

Verda D, Parodi S, Ferrari E, Muselli M. Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. *BMC Bioinformatics*. 2019;20(Suppl 9):390.
<https://doi.org/10.1186/s12859-019-2953-8>.

Gerussi A, Verda D, Cappadona C, Cristoferi L, Bernasconi DP, Bottaro S, et al. LLM-PBC: logic learning machine-based explainable rules accurately stratify the genetic risk of primary biliary cholangitis. *J Pers Med*. 2022;12(10):1587.

Lindgren T. Methods for rule conflict resolution. In: *Proceedings of the European Conference on Machine Learning*. Berlin: Springer; 2004. p. 262-73.

Clark P, Boswell R. Rule induction with CN2: some recent improvements. In: *Proceedings of the Machine Learning—EWSL-91: European Working Session on Learning*. Berlin: Springer; 1991. p. 151-63.

[NOT VERIFIED] Doe J, Smith J. A survey of the role of voting mechanisms in explainable artificial intelligence (XAI). *J Artif Intell Res*. 2022;59:123-45.

Nössig A, Hell T, Moser G. A voting approach for explainable classification with rule learning. In: *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations*. Berlin: Springer; 2024. p. 155-69.

Lindgren T, Boström H. Classification with intersecting rules. In: *Proceedings of the 13th International Conference on Algorithmic Learning Theory, ALT 2002*. Berlin: Springer; 2002. p. 395-402.

Lindgren T, Boström H. Resolving rule conflicts with double induction. *Intell Data Anal*. 2004;8:457-68.

Kitson NK, Constantinou AC, Guo Z, Liu Y, Chobtham K. A survey of Bayesian network structure learning. *Artif Intell Rev*. 2023;56:8721-814.

Čyras K, Rago A, Albini E, Baroni P, Toni F. Argumentative XAI: a survey. *arXiv*. 2021;arXiv:2105.11266.

Espinoza MM, Possebom AT, Tacla CA. Argumentation-based agents that explain their decisions. In: *Proceedings of the 2019 8th IEEE Brazilian Conference on Intelligent Systems (BRACIS)*. 2019. p. 467-72.

Caroprese L, Vocaturo E, Zumpano E. Argumentation approaches for explainable AI in medical informatics. *Intell Syst Appl*. 2022;16:200109.

Governatori G, Maher MJ, Antoniou G, Billington D. Argumentation semantics for defeasible logic. *J Log Comput*. 2004;14(5):675-702.

<https://doi.org/10.1093/logcom/14.5.675>.

Baroni P, Giacomin M. On principle-based evaluation of extension-based argumentation semantics. *Artif Intell.* 2007;171(10-15):675-700.

Wu Y, Caminada M, Podlaszewski M. A labelling-based justification status of arguments. *Stud Log.* 2010;3:12-29.

Caminada M. Argumentation semantics as formal discussion. In: *Handbook of Formal Argumentation*. 2017. p. 487-518.

Rizzo L. ArgFrame: a multi-layer, web, argument-based framework for quantitative reasoning. *Softw Impacts.* 2023;17:100547.

Baroni P, Caminada M, Giacomin M. An introduction to argumentation semantics. *Knowl Eng Rev.* 2011;26(4):365-410.
<https://doi.org/10.1017/S0269888911000166>.

Longo L. Defeasible reasoning and argument-based systems in medical fields: an informal overview. In: *Proceedings of the 2014 IEEE 27th International Symposium on Computer-Based Medical Systems*. New York: IEEE; 2014. p. 376-81.

Cocarascu O, Stylianou A, Čyras K, Toni F. Data-empowered argumentation for dialectically explainable predictions. In: *ECAI 2020*. Amsterdam: IOS Press; 2020. p. 2449-56.

Castagna F, McBurney P, Parsons S. Explanation-question-response dialogue: an argumentative tool for explainable AI. *Argument Comput.* 2025;16(2):133-50.
<https://doi.org/10.3233/AAC-230015>.

Ferrari E, Muselli M. Maximizing pattern separation in discretizing continuous features for classification purposes. In: *Proceedings of the 2010 IEEE International Joint Conference on Neural Networks (IJCNN)*. Barcelona: IEEE; 2010. p. 1-8.

Cangelosi D, Muselli M, Parodi S, Blengio F, Becherini P, Versteeg R, et al. Use of attribute driven incremental discretization and logic learning machine to build a prognostic classifier for neuroblastoma patients. *BMC Bioinformatics.* 2014;15(Suppl 5):S4.
<https://doi.org/10.1186/1471-2105-15-S5-S4>.

Ferrari E, Verda D, Pinna N, Muselli M. Optimizing water distribution through explainable AI and rule-based control. *Computers (Basel).* 2023;12(6):123.
<https://doi.org/10.3390/computers12060123>.

Muselli M, Quarati A. Reconstructing positive Boolean functions with shadow clustering. In: *Proceedings of the 2005 IEEE European Conference on Circuit Theory and Design*. Cork: IEEE; 2005;3.

Walton D, Reed C, Macagno F. Attack, rebuttal, and refutation. In: *Argumentation Schemes*. Cambridge: Cambridge University Press; 2008. p. 220-74.

Baumann R, Spanring C. A study of unrestricted abstract argumentation frameworks. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. Melbourne; 2017. p. 807-13.

Dunne PE, Hunter A, McBurney P, Parsons S, Wooldridge MJ. Inconsistency tolerance in weighted argument systems. In: *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*. Budapest; 2009. p. 851-8.

Dunne PE, Hunter A, McBurney P, Parsons S, Wooldridge M. Weighted argument systems: basic definitions, algorithms, and complexity results. *Artif Intell.* 2011;175(2):457-86.
<https://doi.org/10.1016/j.artint.2010.09.005>.

Bistarelli S, Santini F. Conarg: a constraint-based computational framework for argumentation systems. In: *Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*. Boca Raton: IEEE; 2011. p. 605-12.

Pazienza A, Ferilli S, Esposito F, Bistarelli S, Giacomini M. Constructing and evaluating bipolar weighted argumentation frameworks for online debating systems. In: Proceedings of AI3@AI*IA. Bari; 2017. p. 111-25.

Vilone G, Longo L. An examination of the effect of the inconsistency budget in weighted argumentation frameworks and their impact on the interpretation of deep neural networks. In: Longo L, editor. Joint Proceedings of the xAI-2023 Late-Breaking Work, Demos and Doctoral Consortium. CEUR Workshop Proc. 2023;3554:53-8.

Dung PM. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif Intell.* 1995;77(2):321-57.
[https://doi.org/10.1016/0004-3702\(94\)00041-X](https://doi.org/10.1016/0004-3702(94)00041-X).

Bonzon E, Delobelle J, Konieczny S, Maudet N. A comparative study of ranking-based semantics for abstract argumentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2016;30.

Caminada MW, Gabbay DM. A logical account of formal argumentation. *Stud Log.* 2009;93:109.

Caminada M. On the issue of reinstatement in argumentation. In: Proceedings of the European Workshop on Logics in Artificial Intelligence. Berlin: Springer; 2006. p. 111-23.

Besnard P, Hunter A. A logic-based theory of deductive arguments. *Artif Intell.* 2001;128(1-2):203-35.

Vilone G, Longo L. Development of a human-centred psychometric test for the evaluation of explanations produced by XAI methods. In: Longo L, editor. Explainable Artificial Intelligence. Cham: Springer; 2023. p. 205-32.

Lv G, Chen L, Cao CC. On glocal explainability of graph neural networks. In: Proceedings of the International Conference on Database Systems for Advanced Applications. Cham: Springer; 2022. p. 648-64.