ARGO: An Argumentative Framework for Guided Medical Outcomes

Ciro Listone[1], Salvatore Romano[1], Nikolaos Spanoudakis[2] and Aniello Murano[1]

[1]University of Naples Federico II, Naples, Italy
[2]Hellenic Mediterranean University, Chania, Greece

Abstract

The adoption of advanced Machine Learning in critical healthcare domains is fundamentally limited by the opacity of "black-box" models and their tendency to force singular predictions even when evidence is ambiguous. To address this, we propose a symbolic learning framework for medical diagnosis rooted in Computational Argumentation. Our system learns formal, explainable diagnostic theories directly from clinical data, bypassing the need for prior statistical feature selection. To accommodate the high complexity and overlapping symptomatologies inherent in medical datasets, we introduce tunable initialization hyperparameters that optimally adapt the baseline theory to the specific domain. A defining feature of our approach is the explicit management of diagnostic uncertainty through the concept of confident accuracy. Rather than risking a statistical conjecture, the framework recognizes conflicting evidence and formulates an argumentative dilemma, thereby mimicking the human physician's differential diagnosis. Furthermore, to bridge the gap between formal logic and user accessibility, we integrate a prompt-constrained Large Language Model (LLM) layer that translates the technical execution traces into clear, human-readable clinical narratives. By matching foundational accuracy baselines on the standard Iris dataset and demonstrating remarkable efficiency across complex medical data, the system proved its capacity to identify diagnostic dilemmas and provide clinical support.

Keywords

Computational Argumentation, Explainable Artificial Intelligence, Healthcare

1. Introduction

The integration of data-driven diagnostic tools into modern medicine offers unprecedented opportunities to support clinical decision-making [1, 2]. However, the reliance on "black-box" models, such as Deep Learning architectures or complex ensembles, presents a significant barrier to their adoption in critical healthcare settings [3]. Relying on an opaque prediction, where the reasons for a potential false positive or negative remain inaccessible, is both ethically and practically unsustainable. This highlights a fundamental limitation in the standard Machine Learning (ML) paradigm, which often prioritizes maximizing global accuracy over explanatory depth. Conventional classifiers tend to force a unique prediction even when the statistical evidence distinguishing two pathologies is marginal [4]. To prevent inappropriate treatments resulting from overconfident diagnoses, modern diagnostic AI systems must encompass the structural capacity to generate verifiable justifications and explicitly recognize when clinical evidence is conflicting or insufficient [5]. To address these requirements, current research is increasingly adopting explainable, logic-based approaches. Computational Argumentation has established itself as a central paradigm within AI for modeling non-monotonic reasoning, effectively handling scenarios characterized by inconsistent or uncertain information [6, 7]. Unlike classical mathematical logic, where a proof leads undeniably to a conclusion, argumentation models the cognitive process of defeasible reasoning: conclusions are tentative and can be withdrawn in light of new evidence. Reasoning becomes a dynamic process of constructing arguments, identifying conflicts (*attacks*), and determining which claims can be considered *acceptable*. This framework is particularly valuable in

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 July 2026, Bremen, Germany

ciro.listone@unina.it (C. Listone); salvatore.romano2@unina.it (S. Romano); nspan@hmu.gr (N. Spanoudakis); aniello.murano@unina.it (A. Murano)
0009-0003-5532-9188 (C. Listone); 0009-0007-4640-8705 (S. Romano); 0000-0002-4957-9194 (N. Spanoudakis); 0000-0003-4876-3448 (A. Murano)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

CEUR-WS.org/Vol-4339/paper_10.pdf

medical diagnostics. In clinical practice, distinct symptoms often point towards competing diagnoses. Argumentation allows the system to explicitly model this conflict, determining the validity of a diagnosis based on its ability to *survive* a dialectical process of attack and defense, thereby mimicking the human physician's differential diagnosis.

Our Contribution. In this work, we present a symbolic learning framework rooted in the principles of argumentation and applied to the domain of medical diagnosis. Inspired by the foundational methodology in [8], our system applies an argumentative machine learning process directly to clinical data. The primary distinction of our research is that we deliberately isolate the argumentation phase to evaluate its intrinsic capacity to handle raw clinical features, completely bypassing the prior statistical feature selection step utilized in [8]. A further methodological difference lies in the complexity of the targeted datasets. While earlier studies often relied on simpler benchmarks like the Iris dataset [9], against which we provide a comparative baseline, the medical data utilized in our research presents far greater challenges, exhibiting high dimensionality and overlapping symptomatologies. To effectively navigate this complexity, we introduce specific hyperparameters designed to optimally initialize the baseline theory. Furthermore, our framework is structurally designed to prioritize confident accuracy: when features cannot safely discriminate between competing hypotheses, rather than risking a statistical conjecture, the model outputs an *argumentative dilemma*. This mechanism explicitly exposes the diagnostic conflict, safely guiding the clinician towards further investigation. To bridge the gap between formal logic and user accessibility, the framework incorporates a prompt-constrained LLM layer within a browser-based application, translating technical execution traces into clear clinical narratives.

Related Work. The development of our argumentation-based framework builds upon a progressive evolution in Explainable Artificial Intelligence (XAI) and its application to medical diagnostics [10, 11]. Initially, to bridge the trust gap caused by opaque ML models, research focused on feature attribution methods such as SHAP and LIME [12, 13]. While effective at quantifying the influence of individual symptoms, their numerical outputs often fail to capture the dialectical nature of human medical reasoning, which inherently requires weighing conflicting evidence. To overcome the limitations of purely statistical explanations, research has increasingly turned to argumentation-based XAI. By modeling the diagnostic process as a structured debate, where symptoms act as arguments supporting or attacking a hypothesis, these frameworks provide interactive, conflict-resolving dialogues with high semantic interpretability [14, 15, 16, 17]. At the foundational level of our methodology are practical frameworks that successfully apply logic-based argumentation to real-world decision-making. Systems like Gorgias [18, 19, 20] have demonstrated how application-level explanations can be made inherently comprehensible to domain experts. Previous methodologies [8] have shown the viability of directly learning argumentative theories from data. Our work leverages this progression of tools, from the recognition of numerical XAI limitations to the adoption of formal argumentation, to propose a system where learning and explanation are unified, addressing the management of clinical dilemmas.

Outline. The paper is structured as follows. Section 2 details the proposed methodology, encompassing the dataset characteristics, the underlying reasoning framework, and the core symbolic learning algorithm. Section 3 presents the experimental setup and provides a comprehensive analysis of the evaluation results. Section 4 concludes the paper with a discussion of our findings and outlines directions for future research.

2. Methodology

In this section, we detail the methodology underpinning our symbolic learning framework. First, we introduce the datasets utilized for this study and the specific preprocessing steps required to align the data with our symbolic approach. We then present the formal argumentation system that serves as the core reasoning engine. Finally, we provide an in-depth formulation of the proposed algorithm,

elucidating how formal argumentative theories are dynamically extracted and constructed from the underlying clinical data.

2.1. The Gorgias Cloud Reasoning Framework

To apply abstract argumentation to concrete data, we adopt Logic Programming with Priorities (LPP), which forms the backbone of the Gorgias system [18]. Gorgias belongs to the family of formal argumentation systems designed for non-monotonic reasoning, a form of reasoning where conclusions may be revised or withdrawn as new information becomes available. It assumes a logical language L capable of representing positive statements (assertions that a fact holds) and negative statements (assertions denying a fact or expressing its absence). Within this language, a notion of conflict is defined, at a minimum capturing the opposition between a statement and its negation. Based on this language, an argumentation theory is constructed from argument rules and priority relations. Each rule has an assigned label and establishes a connection between a set of premises (or conditions) and a conclusion (an option to support or a preference over rule labels). The framework distinguishes two main types of arguments: object-level argument rules, which derive claims belonging to the domain language L, and priority rules, which express a preference relation between argument rules to determine which should prevail in a conflict. These rules instantiate individual arguments that can combine into composite arguments. A composite argument consists of object-level arguments supporting domain conclusions, together with priority arguments regulating their relative strength. An *attack relation* is defined whenever arguments support conflicting claims. Attacks arise in two ways: arguments mutually attack when their decisions conflict, and a preference argument attacks the rule it ranks weaker along with any reverse preference. To illustrate these dynamics:

$$\texttt{r1} : \texttt{actinic_keratosis} \leftarrow \texttt{c21}(\texttt{P21}), \texttt{P21} = 1$$
$$\texttt{r3} : \texttt{eczema} \leftarrow \texttt{c19}(\texttt{P19}), \texttt{P19} = 1$$
$$\texttt{pr2} : \texttt{r1} > \texttt{r3} \leftarrow \texttt{c15}(\texttt{P15}), \texttt{P15} = 1$$

Here, rule r1 supports *actinic keratosis* when feature c21 is present, while r3 supports *eczema* given c19. Supporting different diseases, r1 and r3 mutually attack each other when their premises are true. Although the one is not the negation of the other, their conflict can be defined in Gorgias using the complement predicate [19]. The preference rule pr2 makes r1 stronger than r3 when c15 is present. Joining r1 with pr2 forms the composite argument {r1, pr2}, which attacks r3. r3 attacks back r1 but not pr2 and, thus, it is rendered inadmissible. Reasoning in Gorgias is grounded in *admissibility*: an argument set Δ is admissible if it is conflict-free and defends itself by attacking every argument that attacks one of its members [21]. In the example above, r1, pr2 constitutes the maximal admissible set, in a scenario where all *c15*, *c19* and *c21* hold. Constructing such sets may require recursive introduction of higher-level priority arguments. Gorgias evaluates inferential strength through these admissible sets, identifying claims as plausible when backed by evidence and definite when free from unresolved conflicts. This directly informs the system's explanatory capabilities. Rather than merely outputting a result, it generates a dual narrative: an attributive explanation outlining the supporting reasons, and a contrastive explanation clarifying why this conclusion is preferred over competing alternatives. This bridges formal reasoning and human-understandable justification, providing the transparency required for complex decision-making.

2.2. Dataset

The experimental analysis utilizes a public dataset from Kaggle [22], mapping 773 unique diseases to 377 distinct symptoms across approximately 246,000 entries. The data structure reflects the variability inherent in medical diagnosis, as a single disease may be linked to several distinct symptom groups across different samples. It should be noted that this dataset is not medically validated; authentic clinical data may differ significantly in distribution, completeness, and variability. Furthermore, the dataset's binary representation of symptoms necessitates the adoption of a Closed World Assumption, treating

unrecorded symptoms as definitively absent. While real-world clinical environments often require a ternary logic (where the absence of evidence does not equate to evidence of absence), the chosen dataset serves as an adequate testbed for our computational benchmarking. To mitigate computational overhead while preserving the statistical significance of our results, we focused on three specific subsets. These were derived through rigorous preprocessing to achieve an optimal trade-off between real-world prevalence and overall dataset frequency. The first is a mixed subset comprising 9 distinct diseases from various categories. Furthermore, to evaluate the framework's classification capabilities in more ambiguous and challenging scenarios, we generated two targeted dermatological subsets: a restricted subset focusing exclusively on the 5 most prevalent skin diseases, and an expanded subset broadening this scope to 10 skin diseases.

2.3. Implemented Approach

Experiments were conducted on a workstation equipped with an AMD Ryzen 7 9700X Processor and 32 GB RAM. The reasoning tasks were offloaded to the Gorgias Cloud system via dedicated REST APIs. This Argumentation-as-a-Service (AaaS) platform handled the execution of Prolog rules and query resolution, separating the symbolic reasoning engine from the local Python-based learning loop.

Data Abstraction and Logic Representation. The framework processes the clinical dataset $\mathscr{D}$ where each patient is represented by a binary vector $\mathbf{S}$ of symptoms and a target pathology $d \in \mathscr{C}$. These records are directly translated into a declarative logic theory $\mathscr{T}$ using two construct types:

- **Diagnostic Arguments:** Standard logic implications (Horn clauses) that infer a disease d based on a conjunction of symptoms

$$\mathtt{r}_k : \mathtt{diagnosis}(P, d) \leftarrow \mathtt{symptom}_i(P, k) \wedge \cdots \wedge \mathtt{symptom}_j(P, k)$$

  where $k \in \{0, 1\}$.
- **Dynamic Priorities:** New rules are generated on-the-fly during the learning process solely to resolve specific dilemmas and errors between diagnostic arguments.

Rule Selection Metrics. To guarantee medical relevance and filter noise, every candidate rule r generated by the algorithms is evaluated against three rigorous quality metrics also used in [8] (in parenthesis). These criteria function as local inductive thresholds that a candidate argument must satisfy to be integrated into the knowledge base:

1. **Precision (π):** Measures the local reliability of the diagnostic sign. It is defined as the ratio of correct predictions TP_r to the total activations of the rule, i.e. true (TP_r) plus false (FP_r) predictions:

$$\pi(r) = \frac{TP_r}{TP_r + FP_r}$$

2. **Recall (Coverage, κ):** Measures the clinical scope of the rule, defined as the proportion of the total patient population with disease d that is correctly identified by r. This ensures the theory explains a significant portion of the dataset:

$$\kappa(r) = \frac{TP_r}{|\{P \in \mathscr{D} \mid \text{class}(P) = d\}|}$$

3. **Specificity (Exclusivity, ϵ):** A critical metric for differential diagnosis. It quantifies how uniquely a symptom set identifies d compared to competing pathologies. Unlike precision, which validates the rule against all data, specificity specifically penalizes rules that utilize generic symptoms (e.g., *fatigue*) present across multiple disease classes, favoring distinct phenotypic markers that isolate the target pathology:

$$\epsilon(r) = \frac{TN_r}{FP_r + TN_r}$$

where FP_r denotes the number of false positives (instances satisfying rule r but belonging to other disease classes) and TN_r represents the true negatives (instances neither satisfying r nor belonging to d). Consequently, the denominator corresponds to the total number of negative examples for the target pathology.

The Symbolic Learning Algorithm. The learning strategy, detailed in Algorithm 1, is an iterative process driven by a global *Learning Assessment* (LA) score that is computed as a weighted balance that maximizes correct definite predictions while penalizing definite errors, dynamically tolerating ambiguity when evidence is insufficient to resolve a conflict. The ambiguity weight w_a is inversely proportional to the number of classes. The procedure evolves through two operational phases, each targeting a specific deficiency in the theory:

Phase 1: Confident Initialization. Rather than starting from a random state or an empty set, the algorithm initializes the initial theory $\mathscr{T}_0$ by extracting "Golden Rules" governed by our tunable hyperparameters. This extraction is performed by the FindRules function, a data-driven heuristic that generates candidate object-level arguments (Horn clauses) by systematically identifying combinations of symptoms that correlate with a specific diagnosis. These generated arguments must then strictly satisfy the imposed Precision, Recall, and Specificity thresholds. Logically, this step establishes the "common" presentation of each disease. By filtering for high specificity, we ensure that the initial theory anchors the diagnosis on unambiguous symptoms, deliberately leaving complex or atypical cases for the subsequent refinement phases.

Phase 2: Refinement Loop. In this phase, the algorithm iteratively refines the diagnostic theory by addressing both clear misclassifications and unresolved conflicts between competing diagnoses.

First, it focuses on *definite error reduction*, isolating the subset of patients that are currently misclassified. For a patient affected by disease D_{true} but incorrectly classified as D_{false}, the system searches for a **defeater argument**. Concretely, this consists in learning a new, more specific object-level rule for D_{true} that holds specifically within the error subset. Because this rule is grounded in additional symptoms that were overlooked by the more general theory, it effectively captures an exception to the standard clinical presentation. To ensure this new argument overrides the incorrect prediction, the algorithm simultaneously generates a priority rule (as defined in Section 2.1) that explicitly prefers this specific object-level rule over the more general one, but only for the relevant patients. Second, the refinement loop tackles *diagnostic dilemmas*, namely situations in which the theory derives valid but conflicting arguments for two diseases (e.g., r_A for Disease A and r_B for Disease B) with no clear preference. In these cases, the algorithm analyzes the overlap between the supporting patient sets to identify a **discriminator**: a feature f that systematically appears in patients with Disease A and is consistently absent in those with Disease B. When such a feature is found, a priority rule is generated: "If f is present, prefer r_A over r_B". If no discriminating feature exists in the data, the dilemma is retained, marking the limit of the diagnostic resolution.

System Performance Metrics. While the rule selection metrics guide the local extraction of arguments, the global performance of the framework, and the iterative progression of Algorithm 1, is evaluated using the following system-level metrics:

- **Ambiguity ($\mathscr{P}_{amb}$):** The proportion of instances where the framework safely abstains from a singular prediction by outputting an unresolved argumentative dilemma (i.e., multiple conflicting diagnoses).
- **Definite Error ($\mathscr{P}_{err}$):** The ratio of misclassified instances where the system forces a single, incorrect prediction without recognizing any underlying dilemma.
- **Confident Accuracy:** The accuracy calculated strictly on the subset of non-ambiguous predictions. It evaluates the system's correctness specifically when it claims to be certain, isolating the performance from cases of safe abstention.



Figure 1: User interface of the ARGO system demonstrating a clinical diagnostic inference. In the explanation given on the right part of the screen note the two paragraphs. The first explains why the option (Pyogenic Skin Infection) was chosen, while the second provides the reasons why it was preferred over other possible diseases given the context on the left hand side of the screen. Context in upper case letters (FOOT OR TOE PAIN) was tested and found absent.

Clinical Explainability. Although the Gorgias reasoning engine inherently provides explainability through structured system logs, these outputs remain highly technical. Specifically, the generated execution trace outlines the evaluated variables, the specific argumentation rules triggered during the inference process, and a baseline application-level string indicating which diagnostic hypotheses are logically supported by the evidence. In scenarios involving an argumentative dilemma, the log explicitly captures this conflict by returning an array of multiple supported diagnoses, each lacking a stronger counter-argument. However, while these structured JSON outputs are invaluable for system debugging and formally tracing the non-monotonic logic, they are not readily interpretable by end-users, such as clinicians or patients. To bridge this interpretability gap and provide meaningful explainability to non-expert users, we introduced a supplementary natural language processing layer to our architecture. Through a dedicated API integration, the system leverages the open-source Large Language Model Llama 3.3 [1] (specifically, the `llama-3.3-70b-versatile` model) hosted via the Groq platform[2] to translate the technical argumentation logs into comprehensive, human-readable clinical narratives. To guarantee consistency and prevent hallucinations, the LLM's output is strictly constrained by a formatted prompt. It acts as a semantic boundary: it permits minor linguistic smoothing (e.g., translating raw variables like `FOOT_OR_TOE_PAIN` into natural phrasing such as "pain in the feet or toes") to ensure readability, while strictly prohibiting the injection of external medical knowledge. When the argumentation engine encounters an unresolved diagnostic dilemma, the prompt instructs the LLM to process each conflicting hypothesis sequentially. It generates the dual "Arguments in favor" and "Arguments against" paragraphs for *every* plausible disease identified in the admissible set, transparently exposing the clinical ambiguity to the user. For each potential diagnosis identified by the reasoning engine, the model must generate exactly two distinct paragraphs: the first must strictly begin with "*Arguments in favor of [Disease Name]:*" and elucidate which symptoms present in the trace support the hypothesis; the second must strictly begin with "*Arguments against [Disease Name]:*" and clarify which expected symptoms are notably absent according to the data. Ultimately, this integration elevates the formally grounded argumentation framework into a highly transparent, user-centric decision-support system.

[1] https://ollama.com/library/llama3.3
[2] https://groq.com

Algorithm 1 Argumentation-Based Clinical Learning

Require: Dataset $\mathscr{D}$, Thresholds $\{\tau_\pi, \tau_\kappa, \tau_\epsilon\}$
Ensure: Logic Theory $\mathscr{T}$

Phase 1: Confident Initialization

$\mathscr{T} \leftarrow \emptyset$
for each diagnosis $d \in \mathscr{C}$ **do** ▷ Initial Theory Costruction
  $\mathscr{R}_d \leftarrow \textsc{FindRules}(\mathscr{D}, d)$
  $\mathscr{T} \leftarrow \mathscr{T} \cup \{r \in \mathscr{R}_d \mid \pi(r) \geq \tau_\pi \wedge \kappa(r) \geq \tau_\kappa \wedge \epsilon(r) \geq \tau_\epsilon\}$
end for
$\mathscr{M} \leftarrow \textsc{Evaluate}(\mathscr{T}, \mathscr{D})$

Phase 2: Refinement Loop

while $\mathscr{M}.errors > E_Thold$ ***and*** $\mathscr{M}.dilemmas > D_Thold$ **do**
  $\mathscr{T}_{next} \leftarrow \mathscr{T}$
  if $\mathscr{M}.errors > E_Thold$ **then**
    $\mathscr{P}_{err} \leftarrow \textsc{GetMisclassifiedSubset}(\mathscr{D}, \mathscr{T})$
    $\mathscr{R}_{def} \leftarrow \textsc{ReduceErrors}(\mathscr{P}_{err}, \tau_\pi, \tau_\kappa, \tau_\epsilon)$ ▷ Learn on error subset
    $\mathscr{T}_{def} \leftarrow \mathscr{T}_{next} \cup \mathscr{R}_{def}$
  end if
  if $\mathscr{M}.ambiguity > D_Thold$ **then**
    $\mathscr{P}_{amb} \leftarrow \textsc{GetAmbiguousSubset}(\mathscr{D}, \mathscr{T})$
    $\mathscr{P}_{prio} \leftarrow \textsc{ResolveDilemmas}(\mathscr{P}_{amb}, \tau_\pi, \tau_\kappa, \tau_\epsilon)$ ▷ Find discriminators
    $\mathscr{T}_{prio} \leftarrow \mathscr{T}_{next} \cup \mathscr{P}_{prio}$
  end if
  $\mathscr{M}_{next}, \mathscr{T}_{next} \leftarrow \textsc{EvaluateTheories}(\mathscr{T}_{def}, \mathscr{T}_{prio}, \mathscr{D})$
  if $\mathscr{M}_{next}.LA > \mathscr{M}.LA$ **then**
    $\mathscr{T} \leftarrow \mathscr{T}_{next}$
    $\mathscr{M} \leftarrow \mathscr{M}_{next}$
  else
    break
  end if
end while
return $\mathscr{T}$

3. Results

To evaluate our framework, we conducted a 5-fold cross-validation (80-20% split) across all datasets. This 5-fold configuration, aligned with foundational literature [8], was specifically chosen to prevent the fragmentation of minority disease classes and maintain training stability. Relying exclusively on symbolic computational argumentation, candidate rules are evaluated against strict induction thresholds for Precision (τ_π), Recall (τ_κ), and Specificity (τ_ϵ). Crucially, these values act as fixed, local inductive constraints during the rule generation phase, rather than global post-hoc evaluation metrics. As a baseline, we tested the framework on a discrete-categorical encoding of the Iris dataset [9] for a direct methodological comparison with [8]. The transition from this simple 4-feature baseline to the clinical datasets, which scale up to 52 binary columns across 9 classes, introduces significant structural and semantic complexity. The outcomes across these differing domains are summarized in Table 1.

The initialization thresholds $\tau_\pi, \tau_\kappa, \tau_\epsilon$ were determined via empirical grid search to maximize the LA score on the training folds. The multi-domain Top 9 Subgroups required extremely low thresholds (e.g., $\tau_\pi = 0.04$), structurally justified since in highly dimensional datasets, specific diseases are often underrepresented or characterized by atypical symptoms. Lowering these thresholds allows the system to capture rare, highly specific phenotypic markers during Phase 1, which are essential for generating targeted defeater arguments in Phase 2 without overfitting. The Iris evaluation establishes a strong

baseline (0.9089 accuracy, 0.02 ambiguity), proving our methodology is agnostic to binary versus categorical feature representations. In the dermatological domain, expanding from five to ten conditions maintains acceptable accuracy (0.8420) but substantially increases ambiguity (0.5620). This highlights a core strength of the framework: in highly overlapping clinical spaces, rather than forcing a risky prediction, the system safely isolates the conflict as an argumentative dilemma. The Top 9 Subgroups performance (0.8358 accuracy, 0.0500 ambiguity) demonstrates remarkable sample efficiency, extracting robust diagnostic theories from complex spaces despite relying on minimal initial evidence.

Table 1
Performance metrics of the symbolic argumentation framework across diverse datasets.

Dataset	Total Samples	τ_κ	τ_π	τ_ϵ	Avg Accuracy	Avg Ambiguity
Iris	150	0.33	0.30	0.60	0.9089	0.0200
Skin (Top 5)	250	0.10	0.30	0.40	0.9025	0.0300
Skin (Top 10)	500	0.07	0.13	0.50	0.8420	0.5620
Top 9 Subgroups	270	0.04	0.04	0.40	0.8358	0.0500

4. Conclusion

In this paper, we introduced a symbolic learning framework rooted in Computational Argumentation to address the critical need for transparent and reliable medical diagnostic AI. Departing from opaque "black-box" models, our approach prioritizes *confident accuracy* over forced predictions. To effectively navigate the intricacies of diverse clinical data, our methodology introduces tunable initialization hyper-parameters, allowing the baseline argumentative theory to optimally adapt to the specific complexities of the chosen dataset. By mirroring the human process of differential diagnosis, the framework identifies conflicting evidence and translates ambiguous cases into *argumentative dilemmas*. This safe abstention from risky statistical conjectures actively guides clinicians toward further investigation. Furthermore, to bridge the gap between formal logic and practical utility, we integrated a prompt-constrained LLM layer. This addition provides a twofold level of explainability: it transparently exposes the underlying non-monotonic reasoning of the AI, while simultaneously translating complex execution logs into clear, human-readable clinical narratives accessible to both medical professionals and patients.

Future Work. Future development will focus on integrating the remaining features of the foundational methodology [8]. Furthermore, we aim to extend the framework into a multi-agent architecture where specialized AI modules interact and debate over a patient's symptoms. Since argumentative dialectics can be formally modeled as turn-based dialogue games, future iterations will explore the use of temporal and strategic logics, such as Computation Tree Logic [23, 24], to rigorously verify the agents' strategies and the formal correctness of the diagnostic resolution.

Acknowledgments

This work has been partially supported by the following projects: COSMI ALGORITMICI (E63C26000840001), GAMEL (F53C25001420001), SHINE-QC (B69J25000260005).

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, NPJ digital medicine 3 (2020) 17.

[2] C. Listone, V. Malvone, A. Murano, Toward explainable diagnosis: A neurosymbolic approach, in: Proc. of the 18th International Conf. on Agents and AI - Volume 4: ICAART, INSTICC, SciTePress, 2026, pp. 3190–3197.

[3] J. Allgaier, L. Mulansky, R. L. Draelos, R. Pryss, How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare, Artificial Intelligence in Medicine 143 (2023).

[4] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, CoRR abs/1706.04599 (2017).

[5] E. ŞAHiN, N. N. Arslan, D. Özdemir, Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning, Neural computing and applications 37 (2025).

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

[7] T. Bench-Capon, P. E. Dunne, Argumentation in ai, AI 171 (2007) 619–641.

[8] N. Prentzas, C. Pattichis, A. Kakas, Explainable machine learning via argumentation, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 371–398.

[9] R. A. Fisher, Iris, UCI Machine Learning Repository, 1936.

[10] 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.

[11] L. Longo, L. Hederman, Argumentation theory for decision support in health-care: A comparison with machine learning, in: Brain and Health Informatics, Springer International Publishing, 2013.

[12] P. Gohel, P. Singh, M. Mohanty, Explainable AI: current status and future directions, arXiv preprint arXiv:2107.07045 (2021).

[13] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, G. Menegaz, A perspective on explainable ai methods: SHAP and LIME, Advanced Intelligent Systems 7 (2025) 2400304.

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

[15] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021).

[16] A. Rago, H. Li, F. Toni, Interactive explanations by conflict resolution via argumentative exchanges, arXiv preprint arXiv:2303.15022 (2023).

[17] S. L. Vasileiou, A. Kumar, W. Yeoh, T. C. Son, F. Toni, Dialectical reconciliation via structured argumentative dialogues, arXiv preprint arXiv:2306.14694 (2023).

[18] A. C. Kakas, P. Moraitis, N. I. Spanoudakis, GORGIAS: applying argumentation, Argument & Computation 10 (2018) 55–81.

[19] N. I. Spanoudakis, G. Gligoris, A. Kakas, A. Koumi, Gorgias Cloud: On-line explainable argumentation, in: Comma, 2022.

[20] N. I. Spanoudakis, A. Kakas, A. Koumi, Application level explanations for argumentation-based decision making, in: ArgXAI, 2022. URL: https://ceur-ws.org/Vol-3209/1871.pdf.

[21] A. Kakas, P. Moraitis, Argumentation based decision making for autonomous agents, in: Proc. of the second int. joint conference on Autonomous agents and multiagent systems, 2003.

[22] D. R. K, Disease-Symptom Dataset, Kaggle dataset, 2023. URL: https://www.kaggle.com/datasets/dhivyeshrk/diseases-and-symptoms-dataset, accessed: 2026-04-28.

[23] E. M. Clarke, E. A. Emerson, A. P. Sistla, Automatic verification of finite-state concurrent systems using temporal logic specifications, ACM TOPLAS 8 (1986) 244–263.

[24] V. Malvone, A. Murano, V. P. Palma, S. Romano, Findme reforged: Temporal logic AI for richer videogame scenarios, in: Proc. of ICAART'26, Vol. 3, SciTePress, 2026, pp. 2413–2420.