Toward Explainable Diagnosis: a Neurosymbolic Approach

Ciro Listone[1], Vadim Malvone[2] and Aniello Murano[1]
[1]University of Naples Federico II, Naples, Italy
[2]LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
{ciro.listone, aniello.murano}@unina.it, vadim.malvone@telecom-paris.fr

Keywords: Autoencoders, Obstruction Logic, Neurosymbolic, Healthcare

Abstract: The growing adoption of Artificial Intelligence (AI) in medicine highlights the limitations of neural models, which often lack transparency and verifiability. This work presents a theoretical formalization of a neurosymbolic game model that integrates an autoencoder with Obstruction Logic (OL) to model the diagnostic process as structured reasoning guided by knowledge and formal constraints. The autoencoder learns relationships among symptoms from data, producing continuous representations that enrich the clinical picture and capture latent structure. These representations serve as input to a logical module based on OL, a dynamic and tractable formalism that progressively eliminates diagnostic hypotheses inconsistent with established medical knowledge. The combined system merges the inferential power of neural networks with the rigor and reliability of symbolic reasoning, offering interpretable, consistent outcomes. This approach aims to enable decision-support tools that maintain clinical trustworthiness while leveraging data-driven insights, providing a pathway toward more transparent and verifiable AI-assisted medical diagnostics.

1 INTRODUCTION

Artificial Intelligence (AI) in healthcare has expanded in recent years, with applications ranging from disease onset prediction to medical image analysis and decision support in clinical practice (Mattos et al., 2024; Esteva et al., 2019; Ghaffar Nia et al., 2023; Kabir and Tomforde, 2024; Lan et al., 2021). Generative AI techniques enable direct interactions between systems, patients, and physicians, proposing diagnostic hypotheses or therapeutic suggestions based on symptom descriptions. However, these approaches face limitations: they often operate as "black boxes", and the risk of hallucinations or incorrect inferences makes them misleading, especially in critical medical contexts. A promising direction to address these limitations is the combination of logics and formal reasoning with generative approaches (i.e. *neurosymbolic*). Logic allows knowledge and inferential processes to be represented in a structured, verifiable manner, simulating reasoning rather than producing statistical correlations. While symbolic approaches provide transparency and formal guarantees, they struggle to scale or capture complex and high-dimensional clinical data. Conversely, neural networks excel at learning from unstructured data and detecting intricate patterns but lack interpretability. The neurosymbolic paradigm combines the strengths of both approaches, enabling systems that are adaptable to real-world clinical data and formally grounded. Neural networks capture correlations and patterns that are difficult to encode manually, while symbolic components ensure consistency and allow the reasoning process to be verified. In the medical domain, this synergy is relevant: neural models can analyze symptom descriptions or clinical data, whereas logical methods simulate the inferential processes guiding a physician's reasoning.

Among logical formalisms, *Obstruction Logic* (OL) (Catta et al., 2023) and its extensions (Catta et al., 2024; Catta et al., 2025; Leneutre et al., 2025) are well suited to modeling decision-making under constraints via progressive elimination of alternatives. OL defines an interaction between two agents, the *Traveler* and the *Demon*, on an abstract graph of possibilities: the Traveler seeks a goal, while the Demon blocks paths. In our clinical adaptation, the Demon encodes medical knowledge, pruning paths incompatible with observed symptoms or known correlations, while the Traveler explores the remaining diagnostic routes. The outcome is a set of hypotheses consistent with patient data and clinical knowledge, together with a formally verifiable reasoning trace.

In this work, we propose a theoretical formalization of a neurosymbolic game model combining an *autoencoder* (Michelucci, 2022; Mienye and Swart, 2025) with Obstruction Logic. The neural component learns structured representations from clinical data and scales to large knowledge bases, while the logical component constrains and validates each inference step, ensuring transparency, verifiability, and step-by-step explainability.

Outline. Section 2 reviews related work. Section 3 describes the dataset for future implementations. Section 4 presents the high-level architecture, and Section 5 details the core components of the proposed system. Section 6 illustrates an example, and Section 7 concludes with future work.

2 STATE OF ART

Research in the field of AI applied to medicine encompasses several directions, including neural models for clinical data analysis and prediction, logical and formal methods for structured medical knowledge representation, and more recent *neurosymbolic* approaches that aim to unify learning and reasoning within a single framework.

Deep neural networks underpin most automated healthcare analytics (Paul et al., 2024). They perform well in diagnostic classification, medical image segmentation, and outcome prediction (Díaz-Pernas et al., 2021; Pandit and Garg, 2021; Morid et al., 2020). Autoencoders, in particular, have emerged as powerful tools for unsupervised learning of latent representations capable of capturing structures and correlations within high-dimensional clinical data (Pratella et al., 2021; Miotto et al., 2016). They have been used for imputing missing physiological signals, detecting anomalies, and reducing dimensionality in symptom-based or genomic datasets (Kim and Chung, 2020; Badhoutiya et al., 2023). However, they still offer limited interpretability and weak support for incorporating explicit medical knowledge into inference.

The representation of medical knowledge through logical formalisms allows causal relations, diagnostic constraints, and clinical guidelines to be modeled in a structured and verifiable manner. Description logics, rule-based systems, and formal verification methods have been employed to ensure the consistency of therapeutic protocols and to check the correctness of support systems (Bonfanti et al., 2018; Ten Teije et al., 2006). These approaches offer transparency and traceability of reasoning but struggle to efficiently manage the uncertainty and variability inherent in real-world data, making them less adaptive in dynamic or incomplete clinical contexts.

To bridge statistical learning and formal reasoning, *neurosymbolic* methods integrate neural and logical components (Acharya and Song, 2025; Marra et al., 2024; Hossain and Chen, 2025). In medicine, they have constrained neural networks with ontologies, explained decisions via symbolic rules, and checked the logical coherence of generated recommendations (Samwald et al., 2015; Lu et al., 2025; Seneviratne et al., 2023). Broader results support their potential: neurosymbolic reasoning has improved diagnosis of acute abdominal pain (Sundar et al., 2021) and enriched electronic health record (EHR) analysis by combining learned representations with symbolic layers that enforce explainability and actionable knowledge (Kang et al., 2021). Other lines focus on transparent neural architectures, such as discretized interpretable multilayer perceptron models (MLPs) enabling rule extraction and biomarker identification (Bologna, 2003). Overall, neurosymbolic systems seek to combine neural generalization with symbolic verifiability, improving accuracy, transparency, and clinical reliability. Despite these advances, most architectures emphasize prediction performance, knowledge representation, or post-hoc explanations. What remains less explored is the inferential process itself: the iterative, constraint-driven, exclusion-based reasoning that characterizes medical diagnosis.

3 DATASET

The dataset designed for this work is available on *Kaggle* (Patil and Rathod, 2020) and is organized into four `.csv` files, each providing complementary information useful for modeling the problem:

- **dataset.csv**: contains the main representation of clinical data in binary format. Each row corresponds to an instance of a disease characterized by a specific combination of symptoms. The columns represent the symptoms, encoded as binary values (1 = presence, 0 = absence). Since the same disease can manifest in different forms, the same disease label may appear multiple times with partially different symptom configurations. This format reflects the real variability of clinical presentations observed in medical practice.
- **symptom_severity.csv**: provides additional quantitative information. Each symptom is associated with a severity weight measuring its clinical importance. This information allows distinguishing mild symptoms from highly discriminative or critical ones, enriching the simple binary encoding with a finer semantic dimension.

- **symptom_description.csv**: contains a short textual description for each symptom. Although not directly required for training the autoencoder, this file can be useful during the validation and interpretation phases of the model, as it provides a linguistic context that makes the symptoms more readable and understandable.
- **symptom_precaution.csv**: associates each symptom with a list of precautions to adopt when it occurs (for instance, preventive measures or recommended behaviors). This file is also not central for training but represents a valuable resource for possible application extensions of the model, as it links the diagnostic phase to practical and behavioral recommendations.

The combination of these four files thus constitutes an integrated data source that not only enables the construction of models capable of identifying symptom–disease correlations, but also provides a clinical and descriptive context useful for enhancing the interpretability and usability of the system.

This dataset can be considered an ideal starting point for the formalization of medical reasoning, as it combines structured data, quantitative information, and textual descriptions. Despite its simplicity, it allows exploration of both the neural aspects of learning symptom correlations and the symbolic aspects of logical representation and reasoning, laying the groundwork for developing more complex and realistic systems for clinical decision support. It is important to note that the dataset is not medically validated and does not necessarily reflect real clinical practice. Authentic clinical data collected in healthcare settings may differ significantly in distribution, completeness, and variability. However, for demonstrative and experimental purposes, the chosen dataset is adequate and serves as a useful testbed for assessing the validity of the proposed approach.

4 HIGH-LEVEL ARCHITECTURE

The proposed system has a *pipeline* structure composed of two main blocks (Figure 1), conceptually described in this section. In the final form of the system, the input will consist of the symptoms reported by the patient, and the output the diagnostic proposal.

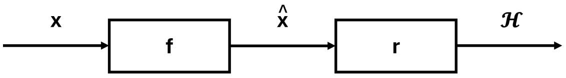
Let m be the total number of symptoms considered. We define the input vector

$$\mathbf{x} \in \{0,1\}^m$$

where, in accordance with the dataset used, $x_i = 1$ indicates the declared presence of symptom i and $x_i = 0$ its absence. The vector $\mathbf{x}$ is provided to the first block of the pipeline, and the system proceeds as follows:

1. **Symptom expansion module f:** transforms the binary vector $\mathbf{x}$ into a continuous scoring vector
$$\hat{\mathbf{x}} = f(\mathbf{x}) \in [0,1]^m$$
where $\hat{x}_j$ represents a *score* expressing the correlation or probability (depending on calibration) that symptom j is relevant, given the subset of symptoms observed in $\mathbf{x}$. The vector $\hat{\mathbf{x}}$ has the same dimensionality as the input but with real-valued components in the range $[0,1]$; so, the first block does not provide a rigid labeling but rather a continuous assessment that can be used to identify symptoms correlated with those declared by the patient, thus extending the initial symptomatic picture.
2. **Formal reasoning module r:** the second block receives as input the vector $\hat{\mathbf{x}} \in [0,1]^m$ produced by the symptom inference module f, which represents the probability or relevance of symptoms associated with the patient. This vector constitutes the initial information for formal clinical reasoning. Let n be the total number of diseases considered in the dataset, and denote
$$\mathcal{M} = \{M_1, M_2, \dots, M_n\}$$
as the set of possible diagnoses. The goal of the second block is to construct, through logic, a list of possible diseases that the patient might have.

Figure 1: Black-box architecture.



4.1 Desired Properties of the First Block

Conceptually, the module f must satisfy several properties useful for the subsequent reasoning phase:

- **Preservation of observed symptoms:** for every index i such that $x_i = 1$, the corresponding $\hat{x}_i$ is expected to be high (close to 1), ensuring that the symptoms reported by the patient remain dominant in the representation.
- **Generalization capability:** the module must assign nonzero values also to symptoms not directly observed but statistically correlated with those present in $\mathbf{x}$, enabling the expansion of the clinical picture.
- **Partial interpretability:** the continuous output $\hat{\mathbf{x}}$ provides a degree of confidence for each possible symptom, useful for presentation to clinicians and for subsequent use in symbolic reasoning.

The module f is trained on the cases present in the dataset:

$$\mathcal{D} = \{\mathbf{d}^{(1)}, \dots, \mathbf{d}^{(N)}\}$$

where each vector $\mathbf{d}^{(k)} \in \{0,1\}^m$ represents a symptomatic configuration observed for a disease instance. At a high level, the training objective can be expressed as the minimization of a loss function:

$$\min_{\theta} \sum_{k=1}^{N} \mathcal{L}\big(f_{\theta}(\tilde{\mathbf{d}}^{(k)}), \mathbf{d}^{(k)}\big)$$

where $\tilde{\mathbf{d}}^{(k)}$ is a partially observed version of $\mathbf{d}^{(k)}$ (for example, obtained by randomly masking some components), and $\mathcal{L}$ is a suitable loss function for binary/continuous vectors. This choice is motivated by the fact that the clinical task of inferring undeclared or not-yet-manifested symptoms is conceptually analogous to reconstructing missing parts of a vector. In this way, the model learns not only to reproduce observed data faithfully but also to infer correlations among symptoms, acquiring a distributed understanding of the internal relationships within the pathological domain.

In summary, the first block of the pipeline functions as a *symptom correlation module*: given a partial set of symptoms, it produces a continuous vector that encodes the degree of statistical association between the observed symptoms and all other possible ones. Rather than providing a diagnostic output, this stage yields a correlation profile that reflects how likely each symptom is to co-occur with the given input pattern. This vector then serves as the input for the subsequent logical reasoning block, which operates at a symbolic level to derive consistent diagnostic hypotheses.

4.2 Desired Properties of the Second Block

Conceptually, the module r aims to transform the continuous output $\hat{\mathbf{x}}$ from the symptom inference module into diagnostic hypotheses compatible with the clinical domain.

The module r must satisfy several key properties to ensure coherent reasoning:

- **Logical consistency:** the diagnoses produced must respect formal constraints and internal relations within the clinical domain.
- **Symbolic generalization capability:** the module must combine partially observed symptoms with predefined knowledge, generating plausible diagnoses even when information is incomplete.
- **Integration with the output of the first block:** the module must be able to interpret the continuous values of $\hat{\mathbf{x}}$, representing degrees of association with each symptom, and possibly transform them (e.g., via weighting or thresholding) to assess symptom relevance.

Let $\mathcal{K}$ denote the set of clinical rules formalized using a graph-based representation. The module r uses this information to identify which diagnoses are compatible with the observed symptoms.

Given the continuous output $\hat{\mathbf{x}} \in [0,1]^m$ of module f, module r produces a set of diagnostic hypotheses $\mathcal{H}$ according to the rule:

$$r(\hat{\mathbf{x}}, \mathcal{K}) = \mathcal{H}$$

The conceptual process unfolds in two main stages:

1. **Selection of relevant symptoms:** the vector $\hat{\mathbf{x}}$ is processed to generate a candidate set of symptoms $\mathbf{y}$ to be used in logical reasoning.
2. **Application of formal logic:** using the relations and constraints defined in $\mathcal{K}$, module r determines which diagnoses are compatible with $\mathbf{y}$, excluding symptom–disease combinations that are inconsistent.

In this way, the system does not merely suggest diagnoses based on statistical correlations but applies logical constraints to simulate a formal medical reasoning process.

The overall pipeline flow can therefore be represented as:

$$\mathbf{x} \xrightarrow{f} \hat{\mathbf{x}} \xrightarrow{r} \mathcal{H}$$

where $\mathbf{x}$ is the vector of observed symptoms, $\hat{\mathbf{x}}$ is the continuous vector of symptom correlations, and $\mathcal{H}$ is the set of diagnostic hypotheses compatible with formal clinical knowledge.

5 LOW-LEVEL ARCHITECTURE

This section presents a detailed overview of the architecture designed to tackle the proposed problem, with a specific focus on the two core components of the solution, which were introduced conceptually in the previous section.

5.1 First Block: Autoencoder

The module f, introduced as the *symptom inference module*, can be implemented through an **autoencoder**, a neural architecture widely used for learning distributed representations and reconstructing partially observed data.

An autoencoder consists of two main components:

- **Encoder:** maps the input $\mathbf{x} \in \{0,1\}^m$ into a latent representation $\mathbf{z}$. The encoder aims to compress information by capturing the most relevant correlations among the observed symptoms.
- **Decoder:** reconstructs a vector $\hat{\mathbf{x}} \in [0,1]^m$ from the latent representation $\mathbf{z}$. In this stage, the decoder produces continuous values that represent the probability or degree of correlation of each symptom, including those not present in the original input.

Formally, the autoencoder can be expressed as:

$$\mathbf{z} = \text{Encoder}(\mathbf{x}; \theta_e)$$

$$\hat{\mathbf{x}} = \text{Decoder}(\mathbf{z}; \theta_d)$$

where θ_e and θ_d represent the parameters (weights and biases) of the encoder and decoder, respectively, and $\hat{\mathbf{x}}$ corresponds to the correlations vector conceptually defined within module f.

The use of an autoencoder is particularly suitable for the goal of inferring missing or undeclared symptoms for several reasons:

- **Learning symptom correlations:** the encoder compresses information, forcing the network to capture co-occurrence patterns among symptoms, which can then be exploited by the decoder to estimate unobserved ones.
- **Partial reconstruction capability:** by training the autoencoder on masked vectors (with some symptoms hidden), the module learns to predict missing components, which in our context correspond to unreported but plausibly present symptoms.
- **Interpretable continuous output:** the decoder produces values in $[0,1]$, which can be interpreted as correlation or probability scores, providing useful information for the symbolic reasoning module.
- **Generalization:** the latent representation $\mathbf{z}$ enables generalization beyond the cases observed in the dataset, identifying symptomatic combinations not explicitly present in the training data.

During training, the autoencoder is optimized to minimize a loss function measuring the discrepancy between the observed inputs and the reconstructed outputs:

$$\min_{\theta_e, \theta_d} \sum_{k=1}^{N} \mathcal{L}\left(\hat{\mathbf{x}}^{(k)}, \mathbf{x}^{(k)}\right)$$

where $\mathbf{x}^{(k)}$ is a vector of observed symptoms, $\hat{\mathbf{x}}^{(k)}$ is the reconstruction produced by the autoencoder, and $\mathcal{L}$ can be, for instance, the binary cross-entropy (Creswell et al., 2017) for each symptom, or the Mean Squared Error (Hinton and Salakhutdinov, 2006). To encourage the ability to infer missing symptoms, a masked version of the input $\tilde{\mathbf{x}}^{(k)}$ can be used during training, so that the network learns to reconstruct hidden components. This approach is common in so-called denoising autoencoders (Vincent et al., 2010; Rubin-Falcone et al., 2023).

In summary, the internal architecture of module f leverages the properties of autoencoders to:

- capture correlations among observed symptoms,
- estimate unreported or latent symptoms,
- produce a continuous output interpretable as a relevance score.

This architectural choice makes module f particularly suitable for the task of *symptom expansion*.

5.2 Second Block: Obstruction Logic

In clinical practice, except for a few obvious cases, diagnosis rarely occurs in a direct manner but rather through exclusion: starting from an initial set of hypotheses consistent with the symptoms reported by the patient, the clinician progressively eliminates those incompatible with clinical knowledge and known symptom correlations. This strategy is incremental, constrained and selective; rather than aiming for a single definitive diagnosis, it seeks to identify a coherent set of plausible clinical possibilities.

A formal paradigm suitable for modeling this dynamic within module r is the *Obstruction Logic*, which is based on an interaction between two agents, the *Traveler* and the *Demon*, and an abstract graph representing the space of possibilities.

Let $\mathcal{G} = (V, E)$ be an abstract graph, where V is the set of nodes and E the set of directed edges. No specific assumption is made about the nature or topology of this graph: it generically represents the *space of possibilities*.

- The **Traveler** is the agent exploring the space represented by $\mathcal{G}$. Its goal is to move along allowed edges to identify one or more final states consistent with the ongoing reasoning process.
- The **Demon** is the agent that, at each step, can selectively remove a subset of edges from E, thereby constraining the paths available to the Traveler. These removals are performed according to constraints, knowledge, or strategies defined externally to the Traveler's dynamics.

The process is iterative, and each step is structured as follows:

1. **Demon's move.** At the beginning of step t, the Demon selects a subset $R_t \subseteq E$ of edges to remove.
2. **Traveler's move.** The Traveler can then move along one of the remaining edges,
$$(v_t, v_{t+1}) \in E \setminus R_t.$$
3. **Restoration.** At the end of the step, the removed edges are restored, and the process can continue.

In the classical formalism, Obstruction Logic assumes that the Demon seeks to prevent the Traveler from reaching a predefined goal state, while the Traveler attempts to reach it despite the removals.

In our scenario, Obstruction Logic is not used to model a *competitive game*, but rather a *constrained collaborative process*.

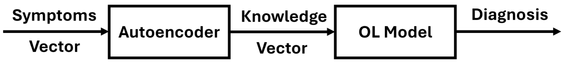
- The **Demon**, which thus becomes a **Navigator**, represents clinical knowledge: at each step, it removes paths that are inconsistent with the symptoms reported by the patient and with known symptom correlations.
- The **Traveler** represents the exploratory decision-making process that traverses the paths deemed coherent by the Navigator, progressively narrowing the space of possibilities.

In this adaptation, no single *goal node* is fixed. The final objective of the process is not to make one agent "win," but rather to determine the residual set of reachable states

$$G^{\text{eff}} \subseteq V,$$

which corresponds to the set of diagnostic hypotheses compatible with the observed symptoms and with clinical knowledge applied iteratively (Figure 2).

Figure 2: Final architecture



This formalism allows for:

- transparently and verifiably representing the process of *diagnostic exclusion*,
- naturally handling the presence of multiple plausible diagnoses, as the goal is not to select a single node but to identify a coherent set of candidates.

6 Illustrative Example

To show how the two modules interact, we present a simple example (Figure 3). Consider a symptom space consisting of three items: [Cough, Common cold, Abdominal pain]. The patient reports only *Common cold*, encoded as

$$\mathbf{x} = [0, 1, 0].$$

First block. Given its training on symptom co-occurrence statistics, the autoencoder produces a continuous reconstruction

$$\hat{\mathbf{x}} = [0.7, 0.9, 0.3],$$

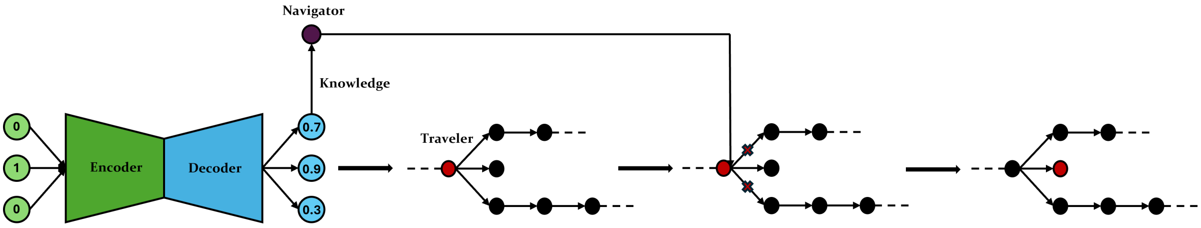
which reflects the learned correlations: strong association between Common cold and Cough, weak association with Abdominal pain. This output serves as a soft, enriched representation of the clinical picture.

Second block. Module r receives $\hat{\mathbf{x}}$ and uses it as evidence for constraining the space of possible diagnostic paths. Symptoms with low reconstructed support (here Abdominal pain) guide the Navigator to eliminate paths dependent on them, while the Traveler continues exploring paths coherent with the retained features (here respiratory-related ones). The resulting reachable set G^{eff} thus contains only diagnoses compatible with the reconstructed symptom profile, capturing the exclusion-based refinement characteristic of clinical reasoning.

7 CONCLUSIONS

We have proposed a neurosymbolic architecture that integrates an autoencoder with Obstruction Logic, with the goal of modeling and reproducing the clinical reasoning process adopted by physicians in formulating diagnostic hypotheses. The neural component, based on a denoising autoencoder, is intended to support the scalable management of large volumes of clinical information and to enable the learning of latent correlations among symptoms, thereby facilitating the inference of potentially unobserved or imperfect clinical data. At the same time, the logical component, grounded in OL, is conceived to counterbalance the limited explainability of neural models by providing an explicit and formally grounded reasoning process, whose inference steps can be inspected and verified step by step, ensuring inferential coherence and the exclusion of alternatives incompatible with established medical knowledge. The synergy between these two modules represents a step toward diagnostic systems capable of combining data-driven learning with reasoning, ensuring both adaptability and transparency. From a theoretical standpoint, the proposed formalization represents a novel contribution to the definition of a neurosymbolic framework

Figure 3: Simplified representation of the proposed pipeline applied to the illustrative example described in Section 6.



grounded in verifiable logical principles. It lays the foundation for the development of computational architectures that integrate predictive capabilities with formal control, opening the way to structural verification of the inferences produced by neural models.

Future work. We plan to develop and empirically evaluate the proposed architecture, assessing the autoencoder's ability to reconstruct realistic symptom patterns and analyzing OL in dynamic diagnostic scenarios. Next, we will extend the framework with temporal representations of symptom evolution and integrate validated medical knowledge bases. This may require more sophisticated logics and frameworks to explore (Kupferman et al., 2001; Löding and Rohde, 2003; Chen et al., 2013; Mogavero et al., 2014; Jamroga and Murano, 2014; Jamroga and Murano, 2015; Jamroga et al., 2024; Leneutre et al., 2024; Borghoff et al., 2025). We also plan to explore probabilistic and fuzzy extensions to handle uncertainty and partial evidence in clinical data; early steps in this direction include (Bulling and Jamroga, 2009; Almagor et al., 2016; Ferrando et al., 2024; Bouyer et al., 2023).

ACKNOWLEDGEMENTS

This work has been partially supported by the PNRR projects INFANT, APLAND, GAMEL and FAIR, as well as the PRIN 2020 project RIPER.

REFERENCES

Acharya, K. and Song, H. (2025). A comprehensive review of neuro-symbolic ai for robustness, uncertainty quantification, and intervenability. *Arabian Journal for Science and Engineering*, pages 1–33.

Almagor, S., Boker, U., and Kupferman, O. (2016). Formally reasoning about quality. *J. ACM*, 63(3):24:1–24:56.

Badhoutiya, A., Singh, D. P., Raj, J. R. F., Srivastava, A. P., Chari, S. L., and Khan, A. K. (2023). Anomaly detection in healthcare: A deep learning approach with autoencoders. In *ICAIIHI 2023*. Vol. 1, pages 1-6, IEEE.

Bologna, G. (2003). A model for single and multiple knowledge based networks. *Artificial Intelligence in Medicine*, 28(2):141–163.

Bonfanti, S., Gargantini, A., and Mashkoor, A. (2018). A systematic literature review of the use of formal methods in medical software systems. *Journal of Software: Evolution and Process*, 30(5):e1943.

Borghoff, U. M., Bottoni, P., and Pareschi, R. (2025). An organizational theory for multi-agent interactions integrating human agents, LLMs, and specialized AI. *Discov. Comput.*, 28(1):138.

Bouyer, P., Kupferman, O., Markey, N., Maubert, B., Murano, A., and Perelli, G. (2023). Reasoning about quality and fuzziness of strategic behaviors. *ACM Trans. Comput. Log.*, 24(3):21:1–21:38.

Bulling, N. and Jamroga, W. (2009). What agents can probably enforce. *Fundam. Informaticae*, 93(1-3):81–96.

Catta, D., Leneutre, J., and Malvone, V. (2023). Obstruction logic: A strategic temporal logic to reason about dynamic game models. In *ECAI 2023*, pages 365–372.

Catta, D., Leneutre, J., Malvone, V., and Murano, A. (2024). Obstruction alternating-time temporal logic: A strategic logic to reason about dynamic models. In *AAMAS 2024*, pages 271–280. IFAAMAS / ACM.

Catta, D., Leneutre, J., Malvone, V., and Ortiz, J. (2025). Coalition obstruction temporal logic: A new obstruction logic to reason about demon coalitions. In *IJCAI 2025*, pages 21–28. ijcai.org.

Chen, T., Forejt, V., Kwiatkowska, M. Z., Parker, D., and Simaitis, A. (2013). Automatic verification of competitive stochastic systems. *Formal Methods Syst. Des.*, 43(1):61–92.

Creswell, A., Arulkumaran, K., and Bharath, A. A. (2017). On denoising autoencoders trained to minimise binary cross-entropy. *arXiv preprint arXiv:1708.08487*.

Díaz-Pernas, F. J., Martínez-Zarzuela, M., Antón-Rodríguez, M., and González-Ortega, D. (2021). A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. In *Healthcare*, volume 9, page 153. MDPI.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.

Ferrando, A., Luongo, G., Malvone, V., and Murano, A. (2024). Theory and practice of quantitative ATL. In *PRIMA 2024*, LNCS 15395, pages 231–247. Springer.

Ghaffar Nia, N., Kaplanoglu, E., and Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3(1):5.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hossain, D. and Chen, J. Y. (2025). A study on neuro-symbolic artificial intelligence: Healthcare perspectives. *arXiv preprint arXiv:2503.18213*.

Jamroga, W., Mittelmann, M., Murano, A., and Perelli, G. (2024). Playing quantitative games against an authority: On the module checking problem. In *AAMAS 2024*, pages 926–934. IFAAMAS/ACM.

Jamroga, W. and Murano, A. (2014). On module checking and strategies. In *AAMAS 2014*, pages 701–708. IFAAMAS/ACM.

Jamroga, W. and Murano, A. (2015). Module checking of strategic ability. In *AAMAS 2015*, pages 227–235. IFAAMAS / ACM.

Kabir, M. F. and Tomforde, S. (2024). A deep analysis for medical emergency missing value imputation. In *ICAART (3)*, pages 1229–1236.

Kang, T., Turfah, A., Kim, J., Perotte, A., and Weng, C. (2021). A neuro-symbolic method for understanding free-text medical evidence. *Journal of the American Medical Informatics Association*, 28(8):1703–1711.

Kim, J.-C. and Chung, K. (2020). Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE access*, 8:104933–104943.

Kupferman, O., Vardi, M. Y., and Wolper, P. (2001). Module checking. *Inf. Comput.*, 164(2):322–344.

Lan, S., Fan, W., Yang, S., Pardalos, P. M., and Mladenovic, N. (2021). A survey on the applications of variable neighborhood search algorithm in healthcare management. *Ann. Math. Artif. Intell*, 89(8):741–775.

Leneutre, J., Malvone, V., and Ortiz, J. (2024). Reasoning about real-time and probability on obstruction logic. In *AI4CC-IPS-RCRA-SPIRIT 2024*, CEUR 3883.

Leneutre, J., Malvone, V., and Ortiz, J. (2025). Timed obstruction logic: A timed approach to dynamic game reasoning. In *AAMAS 2025*, pages 1272–1281. IFAAMAS / ACM.

Löding, C. and Rohde, P. (2003). Model checking and satisfiability for sabotage modal logic. In *FSTTCS*, pages 302–313. Springer.

Lu, Q., Li, R., Sagheb, E., Wen, A., Wang, J., Wang, L., Fan, J. W., and Liu, H. (2025). Explainable diagnosis prediction through neuro-symbolic integration. *AMIA Summits on Translational Science Proc.*, 2025:332.

Marra, G., Dumančić, S., Manhaeve, R., and De Raedt, L. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, 328:104062.

Mattos, M. S., Siqueira, S. W., and Garcia, A. C. B. (2024). Fair and equitable machine learning algorithms in healthcare: A systematic mapping. *ICAART (3)*, pages 815–822.

Michelucci, U. (2022). An introduction to autoencoders. *arXiv preprint arXiv:2201.03898*.

Mienye, I. D. and Swart, T. G. (2025). Deep autoencoder neural networks: A comprehensive review and new perspectives. *Arch Computat Methods Eng.*, 32:3981–4000.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):26094.

Mogavero, F., Murano, A., Perelli, G., and Vardi, M. Y. (2014). Reasoning about strategies: On the model-checking problem. *ACM Trans. Comput. Log.*, 15(4):34:1–34:47.

Morid, M. A., Sheng, O. R. L., Kawamoto, K., and Abdelrahman, S. (2020). Learning hidden patterns from patient multivariate time series data using convolutional neural networks: A case study of healthcare cost prediction. *J. of Biomedical Informatics*, 111:103565.

Pandit, A. and Garg, A. (2021). Artificial neural networks in healthcare: A systematic review. In *Confluence*, pages 1–6. IEEE.

Patil, P. and Rathod, P. (2020). Disease Symptom Prediction. Kaggle dataset. Accessed: 2025-11-02.

Paul, S. G., Saha, A., Hasan, M. Z., Noori, S. R. H., and Moustafa, A. (2024). A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions. *IEEE Access*, 12:15145–15170.

Pratella, D., Ait-El-Mkadem Saadi, S., Bannwarth, S., Paquis-Fluckinger, V., and Bottini, S. (2021). A survey of autoencoder algorithms to pave the diagnosis of rare diseases. *Int. J. Mol. Sci.*, 22(19):10891.

Rubin-Falcone, H., Lee, J. M., and Wiens, J. (2023). Denoising autoencoders for learning from noisy patient-reported data. In *AHLI CHIL*, pages 393–409. PMLR.

Samwald, M., Miñarro Giménez, J. A., Boyce, R. D., Freimuth, R. R., Adlassnig, K.-P., and Dumontier, M. (2015). Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. *BMC medical informatics and decision making*, 15(1):12.

Seneviratne, O., Das, A. K., Chari, S., Agu, N. N., Rashid, S. M., McCusker, J., Franklin, J. S., Qi, M., Bennett, K. P., Chen, C.-H., et al. (2023). Semantically enabling clinical decision support recommendations. *Journal of Biomedical Semantics*, 14(1):8.

Sundar, L. K. S., Muzik, O., Buvat, I., Bidaut, L., and Beyer, T. (2021). Potentials and caveats of ai in hybrid imaging. *Methods*, 188:4–19.

Ten Teije, A., Marcos, M., Balser, M., van Croonenborg, J., Duelli, C., van Harmelen, F., Lucas, P., Miksch, S., Reif, W., Rosenbrand, K., et al. (2006). Improving medical protocols by formal methods. *Artificial intelligence in medicine*, 36(3):193–209.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).