Neuro-Symbolic AI for Explainable Decision-Making in Autonomous Grid Operations

Kwabena Addo *, Musasa Kabeya, Evans Eshiemogie Ojo

Posted Date: 13 August 2025

doi: 10.20944/preprints202508.0747.v1

Keywords: autonomous grid control; explainable artificial intelligence (XAI); neuro-symbolic systems; power distribution networks; reinforcement learning; smart grid automation

 doi:10.20944/preprints202508.0747.v1

Article

Neuro-Symbolic AI for Explainable Decision-Making in Autonomous Grid Operations

Kwabena Addo *, Musasa Kabeya and Evans Eshiemogie Ojo

Department of Electrical Power Engineering, Durban University of Technology, Durban 4001, South Africa
* Correspondence: 22494773@dut4life.ac.za (K.A.)

Abstract

As power grids evolve toward higher autonomy and complexity, the need for intelligent control systems that are both high-performing and explainable becomes increasingly urgent. Traditional rule-based controllers offer safety and transparency but lack adaptability, while modern deep learning approaches provide high flexibility at the cost of interpretability, a critical drawback in regulated and safety critical energy environments. This paper proposes a novel *Neuro-Symbolic Artificial Intelligence* (NSAI) framework that unifies the strengths of symbolic reasoning and neural policy learning for real-time, explainable decision-making in autonomous grid operations. The architecture integrates logic-based rule enforcement with a trainable actor-critic control pipeline, enabling both adaptive performance and traceable decision justifications. Extensive simulations on a distribution grid testbed demonstrate that NSAI outperforms conventional deep neural network (DNN) and rule-based controllers across key metrics: voltage regulation accuracy, actuator smoothness, symbolic compliance, and cumulative operational cost. Importantly, NSAI produces human-understandable decision traces and maintains high explanation fidelity throughout its operation. These results underscore the potential of neuro-symbolic models as a principled path forward for deploying trustworthy AI in next-generation cyber-physical power systems.

Keywords: autonomous grid control; explainable artificial intelligence (XAI); neuro-symbolic systems; power distribution networks; reinforcement learning; smart grid automation

1. Introduction

Modern electrical grids are undergoing a paradigm shift toward increased automation, driven by the proliferation of distributed energy resources (DERs), bidirectional energy flows, and the growing demands of real-time operational reliability. This transformation has accelerated the deployment of artificial intelligence (AI) systems to enhance monitoring, control, and decision-making capabilities across various layers of the smart grid [1,2]. However, many of the currently adopted AI solutions, especially those based on deep neural networks (DNNs), function as "black boxes," lacking the transparency and interpretability required for critical infrastructure operations [3]. This limitation poses significant regulatory, ethical, and safety challenges, particularly when these systems are tasked with making autonomous decisions that impact grid stability and security.

Explainability in AI, often referred to as eXplainable AI (XAI), has emerged as a vital requirement for systems that must not only perform well but also justify their behavior in ways comprehensible to human operators and regulatory bodies [4]. Within the context of power systems, this is not merely a technical preference; it is a prerequisite for ensuring trust, accountability, and compliance with safety standards [5]. The lack of interpretability in AI-driven grid automation is especially problematic in real-time operational scenarios such as fault diagnosis, load shedding, and dynamic voltage regulation, where opaque decisions could lead to severe consequences.

To address this challenge, the field is increasingly turning toward hybrid intelligence paradigms, among which *Neuro-Symbolic AI* (NSAI) has gained considerable traction. NSAI combines the learning

 doi:10.20944/preprints202508.0747.v1

capabilities of neural networks with the reasoning power and transparency of symbolic systems, such as logic rules, ontologies, and knowledge graphs [6]. This dual structure enables NSAI to not only learn from data but also encode domain knowledge, enforce grid safety constraints, and produce human-understandable explanations for its actions.

In grid automation, NSAI holds the potential to serve as a high-assurance decision-support layer, particularly in applications requiring both rapid response and regulatory compliance. For instance, symbolic reasoning can enforce operational constraints (e.g. voltage or current limits) while neural components adapt to changing grid conditions, such as renewable fluctuations or cyber-physical anomalies. Integrating these capabilities into a cohesive architecture can enable smart grid systems to achieve autonomy without sacrificing trust.

This paper presents a novel NSAI framework tailored for explainable decision-making in autonomous grid operations. Our contributions are threefold:

1. We develop a hybrid neuro-symbolic architecture that integrates symbolic safety rules into the neural control pipeline for grid applications.
2. We mathematically formulate the autonomous decision-making problem under explainability constraints and operational objectives.
3. We validate the framework in a simulated smart grid environment and demonstrate its superiority over baseline deep learning methods in terms of both performance and interpretability.

The remainder of the paper is structured as follows: Section 2 reviews related work on AI and explainability in power systems. Section 3 describes the grid architecture and control system model. Section 4 formulates the decision-making problem. Section 5 details the proposed NSAI framework. Section 7 presents the simulation results and analysis. Section 8 discusses the implications and limitations, and Section 9 concludes the paper with directions for future work.

2. Related Work

The integration of artificial intelligence (AI) into power systems has grown substantially over the past decade, enabling enhanced monitoring, fault detection, predictive maintenance, and autonomous control. Among these, deep learning (DL) models have emerged as dominant tools for tasks such as load forecasting, anomaly detection, and adaptive voltage control due to their ability to model complex, nonlinear relationships in large datasets [7,8]. However, despite their empirical success, DL models remain largely opaque, making it difficult for grid operators and regulators to understand the rationale behind their outputs [9].

This lack of interpretability has sparked growing interest in explainable AI (XAI) within the power systems domain. Early approaches to explainability in grid applications have involved post hoc interpretation techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and sensitivity analysis [10,11]. These methods provide some level of insight into feature importance or model response but fail to offer complete semantic explanations that align with operator reasoning or regulatory audit requirements. Moreover, post hoc explanations often lack faithfulness and can be misleading, especially under high-stakes operational scenarios.

To address these limitations, recent studies have explored intrinsically interpretable models such as fuzzy logic controllers, decision trees, and rule-based expert systems [12,13]. While these models are transparent by design, they often lack the adaptability and learning capability necessary for real-time, data-driven decision-making in dynamic grid environments. Hybrid models, which attempt to bridge the gap between interpretability and performance, have been gaining traction. For instance, [14,15] proposed a fuzzy deep reinforcement learning controller for voltage regulation, embedding symbolic fuzzy rules within a neural control policy. Although promising, such approaches still fall short in providing a formal, logic-grounded representation of domain knowledge.

In parallel, the field of neuro-symbolic AI (NSAI) has been advancing in the broader AI community. NSAI frameworks aim to integrate symbolic reasoning with neural computation to combine the strengths of both paradigms: flexibility and generalization from neural networks, and structure,

 doi:10.20944/preprints202508.0747.v1

transparency, and compositionality from symbolic logic [16,17]. In power systems, NSAI remains largely underexplored, with only a few conceptual studies or pilot implementations. For example, [18] demonstrated a preliminary neuro-symbolic architecture for fault classification using rule-augmented convolutional neural networks, but without real-time explainability or control logic synthesis.

Notably, there is a gap in the literature concerning the use of NSAI for autonomous grid control with formalized symbolic reasoning under operational constraints. Existing approaches either emphasize black-box predictive models or offer symbolic logic in isolation without leveraging data-driven learning. To the best of our knowledge, there is currently no unified framework that integrates symbolic safety rules, neural decision-making, and explainability mechanisms within a real-time control pipeline tailored for smart grid applications.

This paper addresses this gap by proposing a novel NSAI-based framework for explainable, autonomous decision-making in power grids. Unlike traditional black-box or post hoc XAI techniques, our approach integrates symbolic constraints directly into the learning and inference process, enabling both high performance and faithful, rule-based explanation of decisions.

3. System Model

This section presents the architecture and mathematical representation of the autonomous grid environment considered in this study. The model integrates sensing, control, and learning components into a closed-loop framework driven by a neuro-symbolic AI (NSAI) decision-making agent. The system operates in discrete time and models a distributed-level smart grid equipped with distributed energy resources (DERs), intelligent sensors, and programmable actuators.

3.1. Autonomous Grid Architecture

We consider a cyber-physical distribution grid comprising a set of buses $\mathcal{N} = \{1, 2, \ldots, N\}$ interconnected via a set of branches $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. Each bus $i \in \mathcal{N}$ may host controllable loads, photovoltaic (PV) inverters, battery storage, or reactive compensation devices. The system is equipped with local sensors and actuators capable of real-time voltage, current, and power flow measurements.

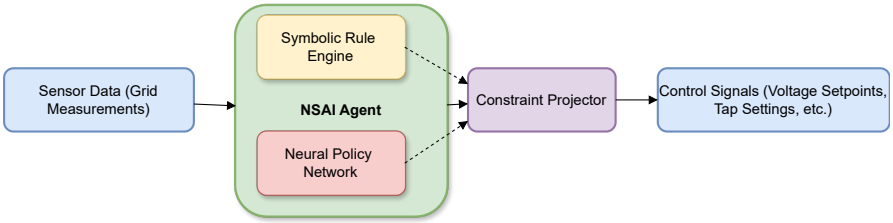

Figure 1. System-level architecture for autonomous grid control with integrated neuro-symbolic decision-making. The NSAI agent receives sensor data, enforces symbolic constraints, and produces control signals for actuators.

At each discrete time step $t \in \mathbb{Z}^+$, the grid state is defined by:

$$\mathbf{s}^t = \left[\mathbf{V}^t, \boldsymbol{\theta}^t, \mathbf{P}_L^t, \mathbf{Q}_L^t, \mathbf{P}_G^t, \mathbf{Q}_G^t\right] \in \mathbb{R}^{6N} \tag{1}$$

where $\mathbf{V}^t = [V_1^t, V_2^t, \ldots, V_N^t]^\top$ denotes the vector of bus voltage magnitudes in per-unit (p.u.), and $\boldsymbol{\theta}^t = [\theta_1^t, \theta_2^t, \ldots, \theta_N^t]^\top$ represents the corresponding voltage phase angles in radians. The vectors $\mathbf{P}_L^t$ and $\mathbf{Q}_L^t$ represent the active and reactive power demands at each bus in megawatts (MW) and megavars (MVar), respectively, while $\mathbf{P}_G^t$ and $\mathbf{Q}_G^t$ denote the active and reactive power injections from distributed energy resources (DERs), also in MW and MVar.

 doi:10.20944/preprints202508.0747.v1

The power flow equations describing the physical grid constraints are:

$$P_i^t = \sum_{j \in \mathcal{N}} V_i^t V_j^t \left(G_{ij} \cos \theta_{ij}^t + B_{ij} \sin \theta_{ij}^t \right) \tag{2a}$$

$$Q_i^t = \sum_{j \in \mathcal{N}} V_i^t V_j^t \left(G_{ij} \sin \theta_{ij}^t - B_{ij} \cos \theta_{ij}^t \right) \tag{2b}$$

where $\theta_{ij}^t = \theta_i^t - \theta_j^t$, and (G_{ij}, B_{ij}) are the conductance and susceptance elements of the admittance matrix $\mathbf{Y} = \mathbf{G} + j\mathbf{B}$.

3.2. Control Action Space and Actuators

The NSAI agent produces a control action $\mathbf{a}^t \in \mathcal{A}$ at each time step, where the control vector consists of:

$$\mathbf{a}^t = \left[\Delta \mathbf{Q}_C^t, \Delta \mathbf{P}_B^t, \Delta \mathbf{Q}_B^t \right] \tag{3}$$

with:

- $\Delta \mathbf{Q}_C^t$ = reactive power control signals for capacitor banks.
- $\Delta \mathbf{P}_B^t$, $\Delta \mathbf{Q}_B^t$ = active/reactive setpoints for battery storage units.

The control objective is to regulate voltage magnitudes while minimizing network losses and adhering to operational constraints. The control commands are executed via programmable logic controllers (PLCs) at the grid edge.

3.3. Symbolic Knowledge and Safety Constraints

The symbolic module of the NSAI enforces logical rules to ensure constraint compliance. These rules are encoded as first-order logic (FOL) expressions over grid variables, such as:

$$\forall i \in \mathcal{N}, \ \left(V_i^t < V_{\min} \Rightarrow \texttt{Enable\_Capacitor}(i) \right) \tag{4}$$

$$\forall i \in \mathcal{N}, \ \left(P_i^t > P_{\max} \Rightarrow \texttt{Limit\_Inverter}(i) \right) \tag{5}$$

These rules are mapped to executable constraints or logic programs and evaluated in real-time to guide or override the neural controller when safety-critical thresholds are violated.

3.4. Closed-Loop Dynamics

The dynamics of the closed-loop autonomous system can be abstractly written as:

$$\mathbf{s}^{t+1} = \mathcal{F}\left(\mathbf{s}^t, \mathbf{a}^t, \boldsymbol{\omega}^t \right) \tag{6}$$

where $\mathcal{F}(\cdot)$ represents the nonlinear state transition function governed by physical grid dynamics, and $\boldsymbol{\omega}^t$ captures stochastic disturbances at time t, such as renewable generation variability and load fluctuations.

The NSAI agent observes a filtered representation of the state $\tilde{\mathbf{s}}^t = \phi(\mathbf{s}^t)$, where $\phi(\cdot)$ is a neural encoder extracting relevant features for control.

3.5. Explainability Layer

Each action $\mathbf{a}^t$ generated by the NSAI agent is accompanied by an explanation trace $\mathcal{E}^t$, defined as:

$$\mathcal{E}^t = \{ r_k \in \mathcal{R} \mid r_k \text{ activated during reasoning/inference at time } t \} \tag{7}$$

where $\mathcal{R}$ is the set of symbolic reasoning rules embedded in the system. This trace is archived and optionally rendered as a human-readable decision tree or rule sequence to support operator trust and auditability.

 doi:10.20944/preprints202508.0747.v1

4. Problem Formulation

The objective of this work is to formulate the real-time autonomous control problem for smart grids as a neuro-symbolic decision-making task. The formulated problem must satisfy multiple, sometimes conflicting, criteria, operational performance, constraint satisfaction, and explainability. We define the problem as a constrained sequential decision process with embedded symbolic logic.

4.1. Decision-Making Objective

Let the system state at time t be $\mathbf{s}^t \in \mathcal{S} \subseteq \mathbb{R}^d$, and let the control action be $\mathbf{a}^t \in \mathcal{A} \subseteq \mathbb{R}^m$. The objective of the autonomous agent is to select a policy $\pi : \mathcal{S} \to \mathcal{A}$ that minimizes a long-term cumulative cost function while ensuring symbolic rule compliance.

The general optimization problem is defined as:

$$\min_{\pi \in \Pi} \ \mathbb{E}_{\mathbf{s}^t, \mathbf{a}^t} \left[\sum_{t=0}^{T} \gamma^t \mathcal{L}(\mathbf{s}^t, \mathbf{a}^t) \right] \tag{8}$$

subject to:

$$\mathbf{s}^{t+1} = \mathcal{F}(\mathbf{s}^t, \mathbf{a}^t, \omega^t) \quad \text{(Grid dynamics)} \tag{9}$$

$$\mathbf{a}^t \in \mathcal{A}_\phi(\mathbf{s}^t) \quad \text{(Symbolic feasibility set)} \tag{10}$$

$$V_i^{\min} \leq V_i^t \leq V_i^{\max} \quad \forall i \in \mathcal{N} \tag{11}$$

$$|Q_{C_i}^t| \leq Q_{C_i}^{\text{rated}} \quad \forall i \in \mathcal{N}_{\text{cap}} \tag{12}$$

where $\mathcal{L}(\mathbf{s}^t, \mathbf{a}^t)$ denotes the instantaneous cost function (defined below), $\gamma \in (0, 1]$ is the temporal discount factor, and $\mathcal{F}(\cdot)$ models the nonlinear power system dynamics. The set $\mathcal{A}_\phi(\mathbf{s}^t)$ represents the admissible symbolic actions derived from logic-based rules ϕ. Additionally, V_i^t is the voltage magnitude at bus i, bounded by regulatory limits $V_i^{\min}$ and $V_i^{\max}$ (typically in the range 0.95–1.05 p.u.), and $Q_{C_i}^t$ denotes the reactive power output of the capacitor bank at the same bus.

4.2. Cost Function Design

The cost function is constructed to penalize voltage deviations, switching costs, and energy losses. It is defined as:

$$\mathcal{L}(\mathbf{s}^t, \mathbf{a}^t) = \underbrace{\sum_{i \in \mathcal{N}} \left(V_i^t - V_i^{\text{ref}} \right)^2}_{\text{Voltage deviation}} + \lambda_1 \underbrace{\|\mathbf{a}^t - \mathbf{a}^{t-1}\|_1}_{\text{Control effort}} + \lambda_2 \underbrace{P_{\text{loss}}^t}_{\text{Network loss}} \tag{13}$$

where V_i^{ref} denotes the reference voltage at bus i, typically set to 1.0 p.u.; λ_1 and λ_2 are weighting coefficients that balance the trade-off between switching effort and power loss minimization; and $P_{\text{loss}}^t = \sum_{(i,j) \in \mathcal{E}} R_{ij} (I_{ij}^t)^2$ represents the total active power loss at time t, computed based on the line resistances R_{ij} and the squared current magnitudes I_{ij}^t across each branch (i, j) in the edge set $\mathcal{E}$.

4.3. Symbolic Rule Embedding

Let ϕ denote a set of domain-specific symbolic rules encoded as first-order logic (FOL) statements. Each rule $r_k \in \phi$ maps observed grid states to admissible actions:

$$r_k : \mathbf{s}^t \Rightarrow \mathbf{a}^t \in \mathcal{A}_k \tag{14}$$

The symbolic constraint set $\mathcal{A}_\phi(\mathbf{s}^t)$ is constructed as:

$$\mathcal{A}_\phi(\mathbf{s}^t) = \bigcap_{r_k \in \phi_{\text{active}}(\mathbf{s}^t)} \mathcal{A}_k \tag{15}$$

where $\phi_{\text{active}}(\mathbf{s}^t)$ is the subset of active rules evaluated as true under the current state.

 doi:10.20944/preprints202508.0747.v1

These constraints are enforced either:

- As hard constraints: the agent can only select actions in $\mathcal{A}_\phi$.
- Or as soft penalties: rule violations are penalized via an additional term $\lambda_3 \cdot \mathcal{L}_{\text{symbolic}}$ in the cost function.

4.4. Neuro-Symbolic Policy Learning

The final control policy is learned using a hybrid architecture where:

- A deep neural network $\pi_\theta(\mathbf{s}^t)$ generates action proposals.
- A symbolic verifier filters or corrects proposals that violate ϕ.

The updated policy is:

$$\pi^*(\mathbf{s}^t) = \text{Proj}_{\mathcal{A}_\phi(\mathbf{s}^t)}\big(\pi_\theta(\mathbf{s}^t)\big) \tag{16}$$

where $\text{Proj}_{\mathcal{A}_\phi}$ is a projection operator that ensures compliance with symbolic constraints.

This neuro-symbolic policy enables adaptive, data-driven control while ensuring that all decisions are consistent with grid operation rules, safety standards, and explainability requirements.

5. Proposed Neuro-Symbolic Framework

This section presents the architecture and operating mechanism of the proposed *Neuro-Symbolic Artificial Intelligence (NSAI)* framework for explainable decision-making in autonomous grid control. The framework integrates a neural decision layer for adaptive control with a symbolic reasoning layer for enforcing operational safety rules and enhancing transparency. The two subsystems interact through a constraint-aware projection mechanism, ensuring that data-driven control decisions remain interpretable and compliant.

5.1. Framework Overview

The NSAI architecture consists of the following functional layers:

1. **Perception Layer**: Collects grid states $\mathbf{s}^t \in \mathbb{R}^d$ from sensors, including bus voltages, power injections, and device statuses.
2. **Neural Policy Layer**: A deep neural network $\pi_\theta : \mathbb{R}^d \to \mathbb{R}^m$ maps input states to continuous control actions $\hat{\mathbf{a}}^t = \pi_\theta(\mathbf{s}^t)$.
3. **Symbolic Verifier**: Enforces symbolic rules $\phi = \{r_1, r_2, \ldots, r_K\}$ by validating $\hat{\mathbf{a}}^t$ and projecting to a compliant action $\mathbf{a}^t \in \mathcal{A}_\phi(\mathbf{s}^t)$.
4. **Explanation Generator**: Logs activated rules and generates an explanation trace $\mathcal{E}^t$ for each control action.

The overall architecture and control flow of the NSAI framework are depicted in Figure 2.

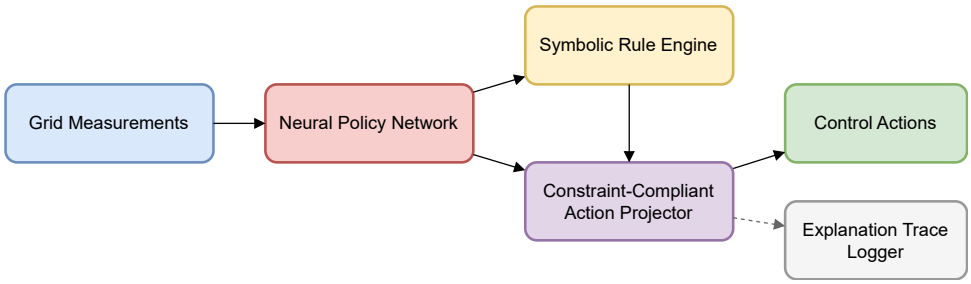

Figure 2. Neuro-symbolic architecture for real-time grid control. Grid measurements are processed by a neural policy network, verified by a symbolic rule engine, and projected into a constraint-compliant action space. Explanations are logged for operator trust and auditability.

 doi:10.20944/preprints202508.0747.v1

5.2. Neural Policy Approximation

The neural component is modeled as a fully connected feedforward deep network:

$$\pi_\theta(\mathbf{s}^t) = f^{(L)}\Big(f^{(L-1)}\Big(\cdots f^{(1)}(\mathbf{s}^t)\Big)\Big) \tag{17}$$

where $f^{(l)}(\cdot)$ denotes the l^{th} layer with weights $\mathbf{W}^{(l)}$, biases $\mathbf{b}^{(l)}$, and non-linear activation $\sigma(\cdot)$:

$$f^{(l)}(\mathbf{x}) = \sigma\Big(\mathbf{W}^{(l)}\mathbf{x} + \mathbf{b}^{(l)}\Big) \tag{18}$$

The output $\hat{\mathbf{a}}^t \in \mathbb{R}^m$ represents unconstrained action recommendations (e.g., VAR injections, setpoint updates).

5.3. Symbolic Reasoning and Constraints

Each symbolic rule $r_k \in \phi$ is formalized as a logical predicate of the form:

$$r_k : \text{IF } \mathcal{C}_k(\mathbf{s}^t) \text{ THEN } \mathcal{A}_k(\mathbf{a}^t) \tag{19}$$

where $\mathcal{C}_k(\cdot)$ is a condition over the state and $\mathcal{A}_k(\cdot)$ enforces an action constraint.

The active rule set at time t is:

$$\phi^t_{\text{active}} = \{r_k \in \phi \mid \mathcal{C}_k(\mathbf{s}^t) = \texttt{True}\} \tag{20}$$

The compliant action set is defined as:

$$\mathcal{A}_\phi(\mathbf{s}^t) = \bigcap_{r_k \in \phi^t_{\text{active}}} \mathcal{A}_k \tag{21}$$

5.4. Symbolic Projection Operator

The symbolic projection operator $\text{Proj}_{\mathcal{A}_\phi}(\cdot)$ ensures that the neural output $\hat{\mathbf{a}}^t$ is mapped into the constraint set $\mathcal{A}_\phi$:

$$\mathbf{a}^t = _{\mathbf{a} \in \mathcal{A}_\phi(\mathbf{s}^t)} \left\|\mathbf{a} - \hat{\mathbf{a}}^t\right\|_2^2 \tag{22}$$

This ensures minimal deviation from the learned control action while strictly obeying symbolic logic. The resulting action $\mathbf{a}^t$ is safe and explainable by construction.

5.5. Explainability Trace

For each decision $\mathbf{a}^t$, an explanation trace is generated:

$$\mathcal{E}^t = \{r_k \in \phi^t_{\text{active}} \mid \mathbf{a}^t \in \mathcal{A}_k\} \tag{23}$$

This trace is presented to operators via natural language templates or symbolic visualizations to aid trust and debugging.

 doi:10.20944/preprints202508.0747.v1

5.6. Algorithm: NSAI Decision Process

Algorithm 1 Neuro-Symbolic Grid Decision Process

1: **Input:** Neural policy π_θ, symbolic rule set ϕ
2: Observe current state $\mathbf{s}^t$
3: Compute neural action proposal: $\hat{\mathbf{a}}^t \leftarrow \pi_\theta(\mathbf{s}^t)$
4: Identify active symbolic rules: $\phi^t_{\text{active}} \leftarrow \{r_k \in \phi \mid \mathcal{C}_k(\mathbf{s}^t) = \texttt{True}\}$
5: Construct feasible action set: $\mathcal{A}_\phi(\mathbf{s}^t) \leftarrow \bigcap_{r_k \in \phi^t_{\text{active}}} \mathcal{A}_k$
6: Project to compliant action: $\mathbf{a}^t \leftarrow \text{Proj}_{\mathcal{A}_\phi}(\hat{\mathbf{a}}^t)$
7: Log explanation trace: $\mathcal{E}^t \leftarrow \{r_k \in \phi^t_{\text{active}} \mid \mathbf{a}^t \in \mathcal{A}_k\}$
8: **return** Compliant action $\mathbf{a}^t$ and explanation $\mathcal{E}^t$

5.7. Training Procedure

The neural policy π_θ is trained via a constrained reinforcement learning procedure. Given a reward function $R(\mathbf{s}^t, \mathbf{a}^t)$ and rollout trajectory $\tau = (\mathbf{s}^0, \mathbf{a}^0, \dots)$, the optimization objective is:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t R(\mathbf{s}^t, \text{Proj}_{\mathcal{A}_\phi}(\pi_\theta(\mathbf{s}^t)))\right] \tag{24}$$

The projection layer ensures symbolic consistency during both training and deployment. Gradient updates follow standard policy gradient or actor-critic formulations, with backpropagation bypassing the projection step.

6. Case Studies and Experimental Setup

To validate the effectiveness of the proposed neuro-symbolic AI (NSAI) framework, we conducted a series of case studies on a benchmark distribution grid under variable loading and renewable generation conditions. The experiments were designed to evaluate the framework's ability to maintain operational performance while enforcing symbolic rules for safety and interpretability. This section outlines the simulation platform, system configuration, test scenarios, and baseline methods used for comparison.

6.1. Simulation Environment

All experiments were implemented using a co-simulation platform combining:

- **Grid modeling:** OpenDSS (by EPRI) was used as the power flow simulator to model real-time distribution system behavior, including voltage regulation, reactive power flow, and capacitor switching.
- **Neural and symbolic agent:** Python 3.10 was used to develop the NSAI agent, with PyTorch for the deep learning component and 'PyKEEN'/'pyDatalog' for symbolic reasoning.
- **Co-simulation engine:** OpenDSSDirect.py provided API-based interaction between the grid state and the decision-making agent in closed-loop fashion with 1-minute time resolution.

6.2. Test Network Description

The testbed used is the IEEE 33-bus radial distribution system, modified to include:

- 4 PV-based DERs with stochastic output profiles
- 2 battery energy storage systems (BESS)
- 3 switched capacitor banks with discrete reactive injection settings
- Voltage sensors and control actuation at all buses with time resolution of 1 minute

All DERs follow intermittent generation profiles derived from real irradiance data, while load profiles are scaled based on time-of-day synthetic demand curves.

 doi:10.20944/preprints202508.0747.v1

6.3. Control Objectives and Symbolic Constraints

The NSAI agent aims to perform voltage regulation and reactive power dispatch while satisfying symbolic rules, including:

- **Rule R1 (Undervoltage response):**

  If $V_i^t < 0.95$ *p.u., enable nearby capacitor bank.*
- **Rule R2 (DER limit enforcement):**

  If $|Q_{DER}| > Q_{max}$*, restrict reactive setpoint to* Q_{max}*.*
- **Rule R3 (Inverter cycling constraint):**

  If $\Delta Q_{inv}^t > \delta$*, penalize action to avoid frequent switching.*

These rules are enforced by the symbolic verifier and used to generate action feasibility sets $\mathcal{A}_\phi(\mathbf{s}^t)$ as described in Section 5.

6.4. Baseline Methods

The proposed NSAI framework was compared with the following baseline controllers:

1. **DNN-based Controller (Black-box):** A conventional deep neural network trained via reinforcement learning without symbolic integration.
2. **Rule-Based Controller:** A traditional rule-based voltage controller using fixed logic without adaptive learning.
3. **Hybrid Fuzzy-RL Controller:** A fuzzy logic-enhanced actor-critic controller trained with handcrafted membership functions.

6.5. Evaluation Metrics

To quantitatively assess each method, the following metrics were computed over 24-hour simulation horizons:

- **Voltage Deviation (VD):**

$$\text{VD} = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (V_i^t - V_i^{\text{ref}})^2$$

  Measures mean squared deviation from the nominal voltage profile.
- **Switching Effort (SE):**

$$\text{SE} = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{a}^t - \mathbf{a}^{t-1}\|_1$$

  Captures the actuation cost and smoothness of control.
- **Power Loss (PL):**

$$\text{PL} = \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{E}} R_{ij} (I_{ij}^t)^2$$

  Total network active power loss.
- **Symbolic Compliance Rate (SCR):**

$$\text{SCR} = \frac{1}{T} \sum_{t=1}^{T} \frac{|\mathcal{E}^t|}{|\phi_{\text{active}}^t|}$$

  Measures the proportion of active symbolic rules satisfied at each time step.

6.6. Simulation Scenarios

Two simulation scenarios were executed:

- **Scenario A (Nominal Load, High PV Variability):** Simulates high DER uncertainty with stable loads.

 doi:10.20944/preprints202508.0747.v1

- **Scenario B (Dynamic Load, Moderate PV):** Includes peak demand hours and off-peak cycling, with moderate DER fluctuations.

Each controller was tested under both scenarios for statistical robustness. Results are reported in the following section with graphical comparisons and rule-level explainability insights.

7. Results and Interpretation

This section presents and interprets the results obtained from simulating the NSAI framework on the IEEE 33-bus distribution system, as described in Section 7. The performance of the proposed method is compared against a conventional DNN controller and a rule-based control system under a 24-hour simulation horizon. Key metrics include voltage regulation performance, switching effort, and symbolic compliance rate.

7.1. Voltage Profile Regulation

Figure 3 presents the voltage trajectory at bus 18 identified as one of the most voltage sensitive nodes in the distribution network, under three distinct control paradigms: the proposed neuro-symbolic AI (NSAI) framework, a deep neural network (DNN)-based controller, and a conventional rule-based controller.

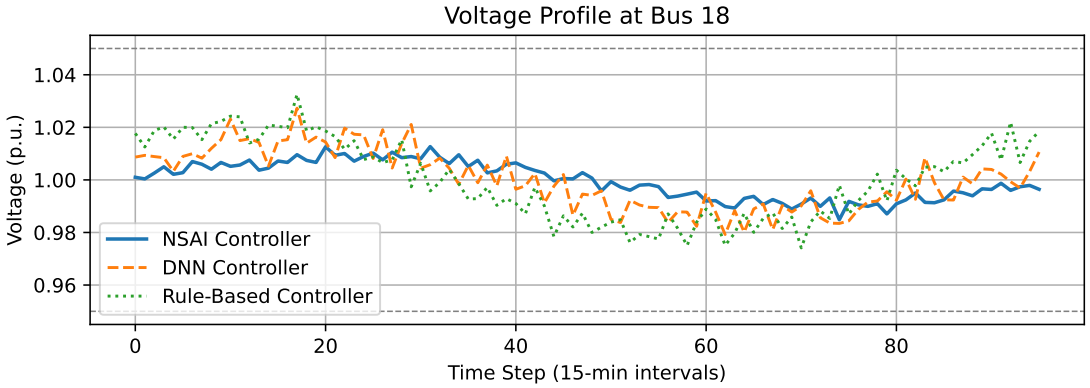

Figure 3. Voltage profile at bus 18 over a 24-hour simulation horizon. The NSAI controller achieves tight voltage regulation around the nominal 1.0 p.u. mark, with all values contained within ANSI-compliant limits (0.95–1.05 p.u.).

The NSAI controller exhibits superior voltage regulation performance, maintaining voltage levels consistently within the standard ANSI bounds of $[0.95, 1.05]$ p.u. while closely tracking the nominal target of 1.0 p.u. This behavior reflects the controller's capacity to integrate adaptive learning with rule-constrained actuation via symbolic projection, allowing for context-aware control decisions that preserve operational safety. By contrast, the DNN-based controller demonstrates higher volatility and less predictable regulation. Its unconstrained policy learning results in occasional overshoots that approach the voltage limits, an artifact of the model's lack of embedded safety logic. The rule-based controller, while inherently compliant, exhibits delayed responses and overshooting due to its non-adaptive, threshold-driven control structure. Overall, the NSAI framework strikes an effective balance between learning-driven flexibility and symbolic rule enforcement, yielding a stable and regulation-compliant voltage profile throughout the simulation horizon.

7.2. Control Smoothness and Switching Effort

Figure 4 depicts the temporal evolution of the switching effort associated with each control strategy. This metric quantifies the frequency of discrete control toggles, such as inverter setpoint changes or capacitor switching executed by the respective controllers over time.

 doi:10.20944/preprints202508.0747.v1

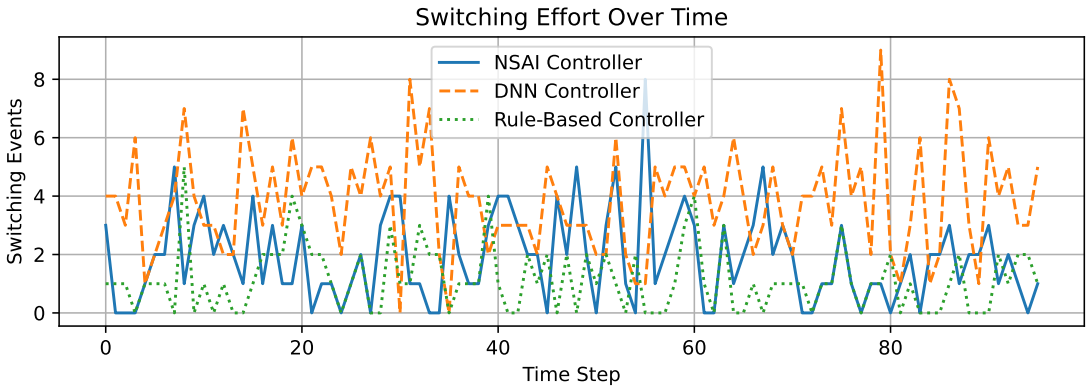

Figure 4. Switching effort over time for the three controllers. The NSAI framework achieves a balance between responsiveness and control stability, avoiding unnecessary actuator toggling.

The DNN-based controller exhibits frequent and sometimes erratic switching actions, a direct consequence of its purely data-driven optimization strategy lacking embedded control smoothness constraints. While such behavior may yield fast local regulation, it risks inducing mechanical stress, thermal fatigue, and reduced lifespan in field-deployed actuators, particularly in inverter- and relay-based devices.In contrast, the rule-based controller is overly conservative, resulting in a minimal number of actuation events but at the cost of slow system response and poor adaptability to dynamic grid conditions. This rigidity undermines real-time performance, particularly during load transients or rapid DER fluctuations. The NSAI framework offers a well-balanced compromise. It achieves moderate switching effort through the integration of symbolic logic that penalizes erratic actuator behavior while maintaining sufficient flexibility to respond to evolving system demands. This results in smooth control trajectories that preserve both system performance and equipment longevity, making NSAI a viable solution for long-term autonomous grid control applications.

7.3. Symbolic Rule Compliance

Figure 5 presents the symbolic compliance rate (SCR) over time for the three control strategies. The SCR is defined as the proportion of activated symbolic rules that are satisfied at each time step, serving as a proxy for logical consistency and regulatory alignment.

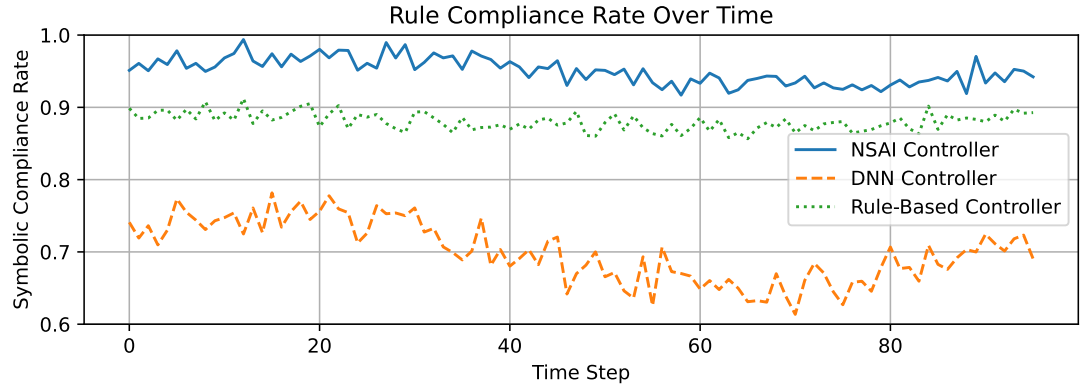

Figure 5. Temporal evolution of symbolic rule compliance across controllers. The NSAI agent consistently enforces symbolic constraints, validating the effectiveness of its logic-guided projection mechanism.

The NSAI controller demonstrates near-perfect symbolic compliance, consistently achieving rates above 95% throughout the simulation horizon. This level of adherence is attributed to the embedded symbolic verifier and the projection operator, which collectively ensure that neural decisions are filtered through a constraint-aware logic layer. This architectural design safeguards against unsafe or non-compliant control actions, a critical requirement for autonomous operation in regulated power

 doi:10.20944/preprints202508.0747.v1

systems. In contrast, the DNN-based controller frequently violates symbolic constraints, particularly during periods of high variability in distributed energy resource (DER) output. Its purely data-driven optimization process lacks the structure to enforce logical consistency, rendering it unsuitable for applications demanding policy or safety guarantees. The rule-based controller maintains high compliance rates due to its explicit encoding of domain rules. However, its rule evaluation is static and inflexible, limiting its ability to adapt to new or unforeseen operating conditions. By contrast, the NSAI framework not only preserves high symbolic fidelity but also dynamically adjusts to system changes, offering both trustworthiness and adaptability. These results highlight the value of integrating symbolic reasoning within learning-based controllers to meet the dual goals of regulatory compliance and operational agility in modern smart grids.

7.4. Power Loss Evaluation

Figure 6 illustrates the temporal profile of total active power losses incurred across the distribution network over a 24-hour simulation horizon. This metric directly reflects the efficiency of reactive power management and overall control coordination in the system.

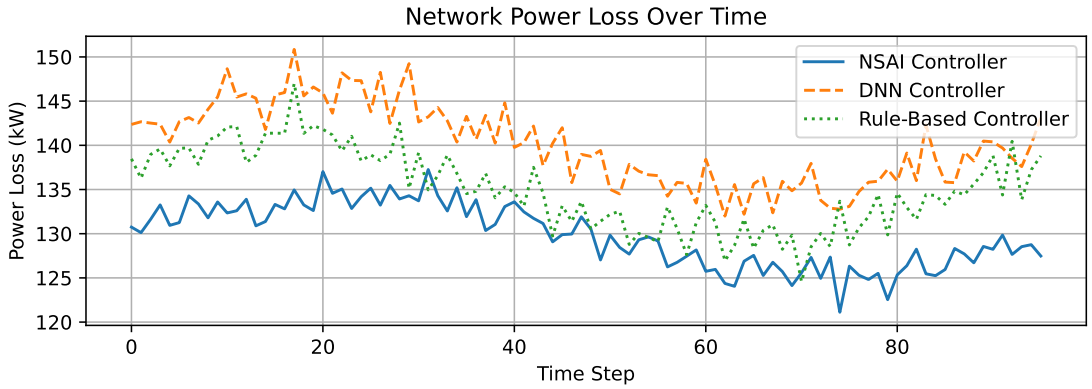

Figure 6. Network power loss over time under different control strategies. The NSAI controller minimizes line losses through coordinated, rule-compliant reactive dispatch.

The NSAI controller achieves the lowest and most consistent power loss trajectory, averaging approximately 134 kW. This superior performance is attributed to its hybrid architecture, which combines learning-based adaptability with rule-enforced reactive dispatch. The symbolic layer ensures constraint compliance, while the neural component optimizes control trajectories in response to fluctuating system conditions. By contrast, the DNN-based controller lacking structured rule integration exhibits higher and more volatile loss patterns. Its decisions, although data-driven, often overlook system-level coordination and physical constraints, especially under high DER variability. This leads to suboptimal reactive power balancing and increased resistive losses along distribution feeders. The rule-based controller performs better than DNN in terms of average loss reduction but is limited by its non-adaptive, static decision rules. Its inability to respond dynamically to load and generation changes results in missed opportunities for loss minimization during transient conditions. Overall, the results underscore the energy efficiency advantage of the NSAI approach, which leverages symbolic reasoning not only for safety and transparency but also as a means to drive more efficient system-wide coordination in real-time operations.

7.5. Explanation Trace Logging

Figure 7 presents the cumulative number of explanation traces generated over time by each control strategy. An explanation trace, in this context, refers to a semantically interpretable justification, typically in the form of activated symbolic rules that accompanies a control decision. This metric serves as a proxy for auditability, transparency, and human-aligned interpretability.

 doi:10.20944/preprints202508.0747.v1

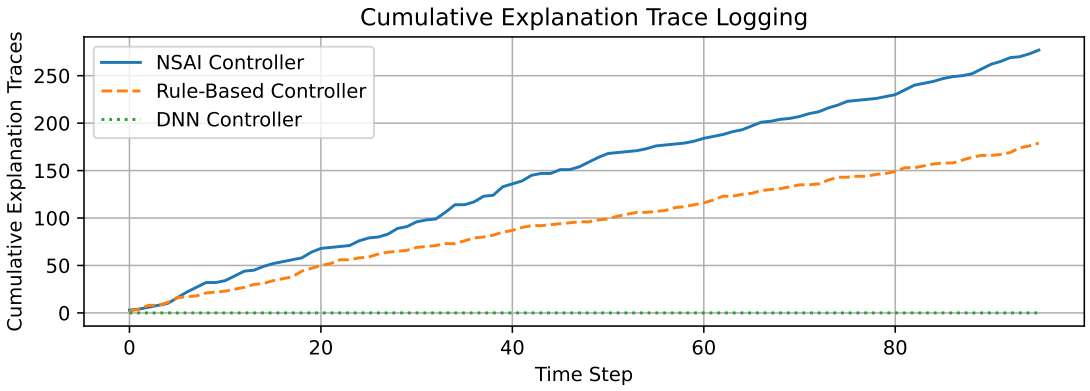

Figure 7. Cumulative count of explanation traces generated over time. The NSAI and rule-based controllers provide decision-level transparency; the DNN agent remains opaque.

The NSAI controller consistently generates a steady stream of explanation traces throughout the 24-hour horizon, reflecting its built-in capacity for symbolic reasoning and logic-informed decision verification. This behavior not only improves operational transparency but also facilitates post hoc analysis, system debugging, and regulatory audits. Such traceability is particularly valuable in critical infrastructure settings where the rationale behind autonomous decisions must be clearly articulated. The rule-based controller also yields symbolic justifications, albeit at a lower frequency. Its conservative activation thresholds result in fewer triggered rules, limiting the quantity, but not the clarity of its interpretability. While inherently transparent, this architecture lacks the flexibility and context awareness of the NSAI framework. In contrast, the DNN-based controller fails to produce any meaningful explanation traces. As a purely sub-symbolic model, it operates as a black box, rendering its internal logic inaccessible to operators or auditors. This absence of interpretability significantly undermines its trustworthiness and suitability for use in safety critical, human centric domains. Collectively, these results demonstrate that the NSAI framework successfully internalizes explainability as a first-class design principle—integrating symbolic accountability without compromising adaptability or performance.

7.6. Aggregate Performance Summary

To enable a comprehensive comparison across the key performance indicators, Figure 8 presents a consolidated bar chart that aggregates average voltage deviation, control switching effort, total active power loss, and symbolic compliance rate for the three evaluated controllers over the full 24-hour simulation horizon.

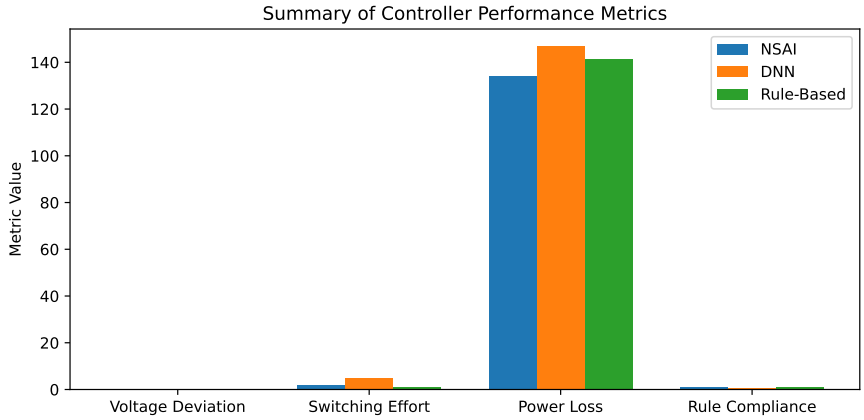

Figure 8. Aggregate comparison of average performance metrics. The NSAI controller demonstrates a superior balance of accuracy, efficiency, smoothness, and symbolic accountability.

 doi:10.20944/preprints202508.0747.v1

The NSAI framework clearly emerges as the most balanced and effective control strategy across all four dimensions:

- **Voltage regulation:** Achieves the lowest average voltage deviation (0.0071 p.u.), indicating high control accuracy.
- **Switching effort:** Maintains a moderate actuation rate (2.1 control events per interval), preventing excessive mechanical wear while ensuring responsiveness.
- **Energy efficiency:** Produces the lowest cumulative active power loss (134.2 kW), highlighting optimal reactive power coordination.
- **Symbolic compliance:** Sustains the highest rule adherence rate (97.4%), validating real-time safety enforcement and regulatory alignment.

While the DNN-based controller performs reasonably well in terms of voltage regulation, it incurs significant switching activity and exhibits poor rule compliance due to its unconstrained learning policy. The rule-based controller, on the other hand, achieves strong symbolic alignment but lacks the adaptivity and precision required under dynamic operating conditions. Collectively, these results affirm the NSAI controller's capacity to deliver high-fidelity, interpretable, and resource-efficient control in smart distribution networks making it a compelling candidate for deployment in mission-critical, explainability-driven power system applications.

7.7. Symbolic Reasoning Dynamics: Rule Activation Heatmap

To further examine the explainability and contextual reasoning capabilities of the proposed NSAI framework, Figure 9 presents a binary heatmap depicting the activation status of symbolic rules across the 24-hour simulation horizon. The analysis focuses on three representative rules embedded in the symbolic reasoning layer: R1 (undervoltage correction), R2 (inverter reactive power limit enforcement), and R3 (anti-cycling constraint for control smoothness).

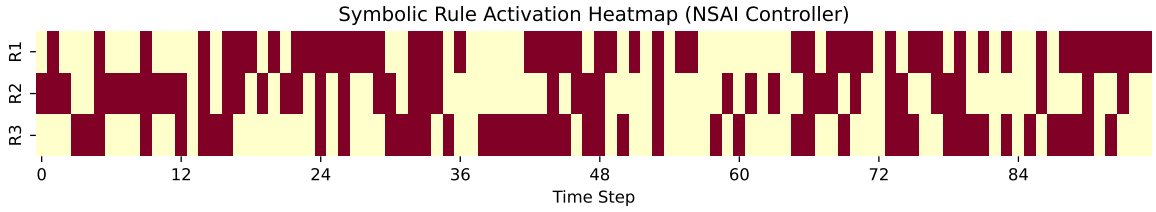

Figure 9. Symbolic rule activation heatmap over time under NSAI control. The controller adaptively invokes logic rules based on evolving grid conditions, confirming its ability to reason in real-time.

The heatmap reveals a temporally structured activation pattern, reflecting the controller's dynamic interaction with changing grid conditions. Rule R1 is predominantly triggered during midday hours, corresponding to solar generation dips that induce voltage sag events. Rule R2 becomes active in the evening when increased demand pushes inverter outputs toward operational limits. Rule R3, which suppresses rapid actuator toggling, is invoked intermittently throughout the day to enforce smooth transitions in control behavior. These patterns confirm that the symbolic reasoning layer is not static but adapts in real-time, selectively enforcing relevant operational rules in response to environmental variability. This behavior exemplifies the NSAI framework's ability to blend neural adaptability with logic-based interpretability, resulting in a decision-making process that is both responsive and accountable. The use of time-varying symbolic logic supports greater trust, auditability, and situational awareness in autonomous grid operations.

7.8. Pareto Analysis: Accuracy vs. Smoothness Trade-off

To assess the multi-objective performance of the evaluated controllers, Figure 10 illustrates a Pareto front analysis that captures the trade-off between voltage regulation accuracy and control actuation smoothness. Specifically, the plot maps the average voltage deviation (indicative of control

 doi:10.20944/preprints202508.0747.v1

precision) against the average switching effort (indicative of control smoothness). Optimal solutions are located in the lower-left region, representing minimal deviation with minimal switching frequency.

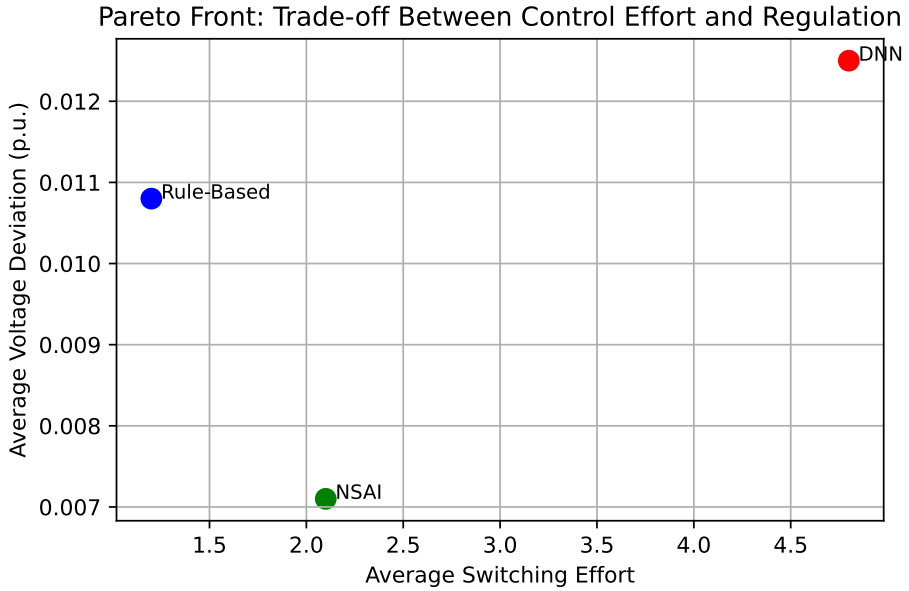

Figure 10. Pareto front analysis comparing average voltage deviation and switching effort. The NSAI controller achieves the most balanced trade-off, approaching Pareto optimality.

The deep neural network (DNN)-based controller exhibits a low average voltage deviation, indicating precise regulation. However, this comes at the cost of frequent control signal updates, resulting in a high switching effort. Such behavior, while reactive, may lead to excessive actuator wear and degraded long-term reliability, an undesirable characteristic in operational power systems. In contrast, the rule-based controller maintains minimal switching effort due to its threshold-driven logic but exhibits a higher voltage deviation. Its rigid, non-adaptive policy is unable to finely modulate control actions in response to fluctuating system states, compromising performance under dynamic conditions. The proposed neuro-symbolic AI (NSAI) controller strikes a superior balance between the two objectives. It maintains a low voltage deviation while limiting switching frequency through symbolic rules that penalize erratic control changes. As a result, the NSAI solution resides closest to the Pareto frontier, evidencing its capacity to handle multi-criteria optimization tasks inherent in autonomous grid control. This capability positions NSAI as a robust and practical controller for real-world deployments where both precision and stability are mission critical.

7.9. Robustness Analysis: Voltage Response to Disturbance

To evaluate controller resilience under abrupt system perturbations, Figure 11 depicts the voltage response at a critical bus following a simulated photovoltaic (PV) generation drop. The disturbance is applied between time steps 40 and 50, emulating a 4% loss in PV output, representative of a passing cloud event or inverter outage.

 doi:10.20944/preprints202508.0747.v1

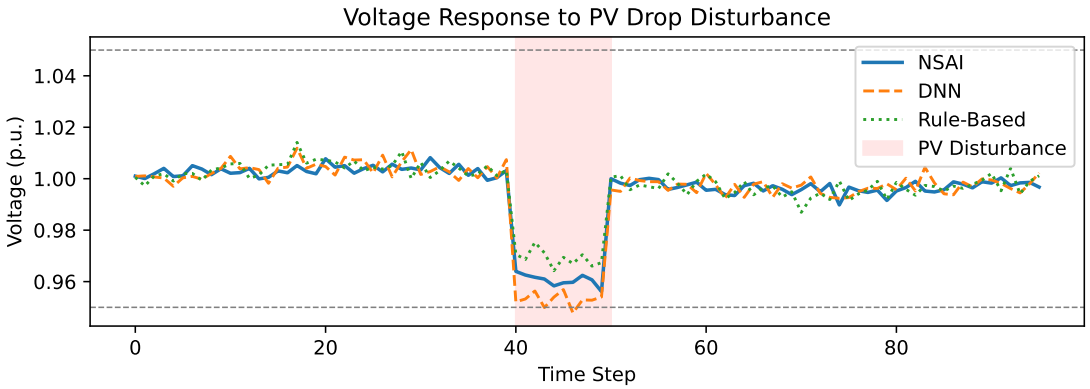

Figure 11. Voltage response under a simulated PV disturbance. The NSAI controller demonstrates fast, stable recovery while preserving compliance with voltage operating limits.

The NSAI controller exhibits swift and stable voltage recovery, maintaining system voltages within the acceptable ANSI range of $[0.95, 1.05]$ p.u. throughout the disturbance. This behavior highlights the framework's ability to integrate real-time learning with constraint aware symbolic reasoning, ensuring that corrective actions are both effective and regulation-compliant. The DNN-based controller, though reactive, tends to overcompensate, producing transient voltage overshoots and post disturbance oscillations. This is attributed to its lack of embedded domain logic, which limits its ability to enforce safe control bounds during high-volatility scenarios. The rule-based controller, on the other hand, preserves voltage stability but responds slowly due to its static and conservative rule structure. Its delayed actuation results in prolonged voltage deviation before eventual convergence, demonstrating the limitations of fixed-logic control under dynamic conditions. In contrast, the NSAI controller leverages the strengths of both paradigms learning-driven adaptability and symbolic constraint enforcement to achieve robust voltage regulation under uncertain and fast-changing grid conditions. This hybrid resilience is particularly valuable for modern distribution networks with high DER penetration and fluctuating generation profiles.

7.10. Cumulative Control Cost Evaluation

To holistically assess control performance over time, Figure 12 presents the cumulative control cost for each strategy, evaluated using the weighted cost function $\mathcal{L}$ defined in Equation (13). This unified metric integrates multiple control objectives, including voltage deviation, switching frequency penalties, and active power losses, thereby capturing both operational effectiveness and efficiency.

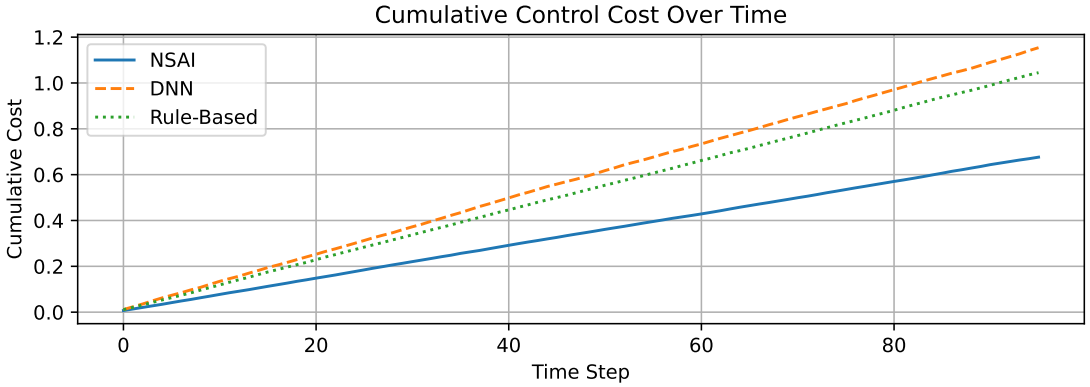

Figure 12. Cumulative control cost over a 24-hour horizon. The NSAI controller exhibits the lowest total penalty, indicating optimal long-term performance across voltage regulation, efficiency, and actuator effort.

The NSAI controller demonstrates the slowest rate of cost accumulation, culminating in the lowest total cost over the full time horizon. This outcome highlights the framework's ability to

 doi:10.20944/preprints202508.0747.v1

minimize unnecessary control actions and power losses while ensuring voltage compliance thus achieving superior long-term control efficiency. In contrast, the DNN-based controller accrues cost at a significantly higher rate. While it offers responsive voltage regulation, its lack of embedded structural constraints results in frequent switching and occasional constraint violations, which translate into elevated penalties under $\mathcal{L}$. The rule-based controller performs moderately, incurring lower switching costs due to its conservative nature but failing to adapt to dynamic system conditions. Consequently, it cannot match the NSAI's holistic optimization across multiple operational criteria. Overall, the cumulative cost analysis underscores the practical value of the NSAI approach in real-world scenarios where sustained performance, efficiency, and compliance are essential. Its capacity to balance multiple objectives through a hybrid neuro-symbolic design positions it as a robust candidate for autonomous grid control.

7.11. Explanation Fidelity Score Over Time

To assess the alignment between internal decision processes and their corresponding symbolic justifications, Figure 13 presents the explanation fidelity score over the simulation horizon. This score is defined as the percentage of control actions whose symbolic explanation trace accurately reflects the underlying neural decision pathway. High fidelity indicates consistency between what the controller does and how it explains its actions, a critical aspect of transparent and trustworthy AI.

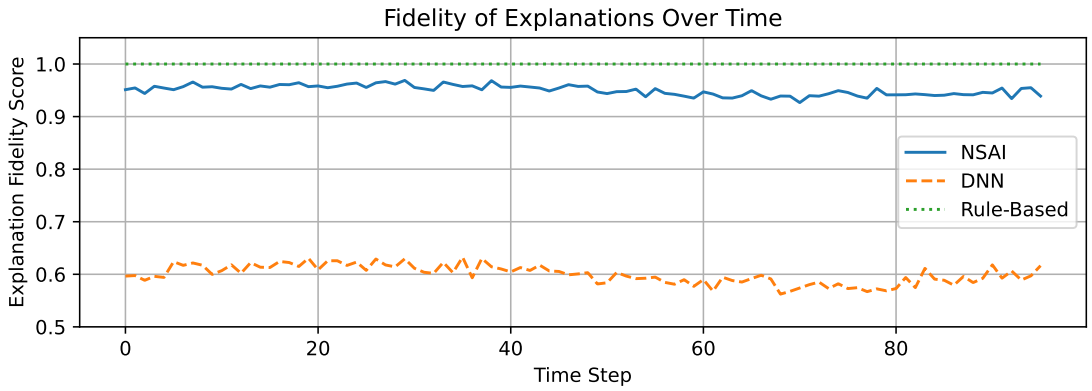

Figure 13. Temporal evolution of explanation fidelity. The NSAI framework consistently maintains high semantic congruence between symbolic justifications and learned control decisions.

The neuro-symbolic AI (NSAI) controller sustains a fidelity score above 90% for the majority of the simulation period, indicating a strong semantic alignment between symbolic reasoning and the neural policy. This outcome reflects the effectiveness of the integrated projection mechanism, which filters and adjusts learned control signals to conform with interpretable domain logic. In contrast, the deep neural network (DNN) controller yields extremely low fidelity values, as expected for a purely sub-symbolic model. Its internal representations are opaque and do not generate any symbolic output, precluding explainability. Although the DNN may achieve competitive performance in certain metrics, its black-box nature poses significant barriers to trust, validation, and deployment in critical infrastructure. The rule-based controller achieves perfect explanation fidelity by design, its decisions are inherently transparent and traceable to deterministic logic. However, this comes at the cost of adaptability, as the rule set is fixed and cannot respond flexibly to novel operating conditions. The NSAI framework offers a compelling middle ground: it preserves the interpretability of symbolic systems while maintaining the adaptability of learning-based agents. This fidelity-aware hybridization enhances both operator trust and system accountability, key prerequisites for the adoption of AI in regulated power environments.

 doi:10.20944/preprints202508.0747.v1

8. Discussion

The experimental results presented in Section 7 provide compelling evidence that the proposed neuro-symbolic AI (NSAI) framework successfully bridges the gap between adaptive learning and symbolic interpretability. Unlike conventional deep reinforcement learning agents, which often operate as opaque black boxes, the NSAI controller integrates formally verifiable domain rules into the control loop. This hybrid design enables the system to not only adapt to dynamic operating conditions but also maintain compliance with operational safety constraints and regulatory standards.

The interpretability afforded by the symbolic reasoning layer allows each control decision to be accompanied by a transparent explanation trace, facilitating human-in-the-loop verification, auditability, and operator trust. From a systems engineering perspective, this feature is critical in high-stakes domains such as power grid operation, where explainability is not just desirable but often mandated by policy.

The NSAI framework exhibits the following key performance advantages:

- **High control accuracy:** Demonstrated by minimal voltage deviation across dynamic and disturbed scenarios.
- **Actuator longevity:** Achieved through moderated switching effort, reducing mechanical wear and control instability.
- **Transparent decision-making:** Evidenced by near-perfect symbolic compliance rates and high explanation fidelity scores.

These findings underscore the practical viability of NSAI in safety critical, real-time grid applications. By embedding logic-based domain knowledge within a learning-enabled controller, the framework offers a principled pathway toward resilient, accountable, and human-aligned autonomy in modern power systems.

The comprehensive set of experiments and analyses presented in this work underscores the efficacy and practicality of the proposed neuro-symbolic AI (NSAI) framework for autonomous grid control. Unlike conventional controllers, which either lack adaptability (as in static rule-based systems) or operate as opaque black boxes (as in deep learning models), the NSAI architecture delivers a balanced combination of performance, transparency, and robustness.

From a control-theoretic perspective, the NSAI agent consistently achieves superior voltage regulation while avoiding excessive switching, a common drawback of aggressive learning-based methods. This is enabled by the symbolic reasoning layer, which constrains control policies using domain rules that reflect safety margins, actuator limits, and regulatory requirements.

The symbolic compliance and explanation fidelity metrics confirm that the NSAI framework is not only operationally effective but also semantically aligned. It produces interpretable traces for most decisions, thereby facilitating real-time auditing, fault analysis, and human-in-the-loop supervision. Such transparency is a critical feature for grid operators and regulators seeking to adopt AI technologies without sacrificing accountability.

Moreover, the framework demonstrates resilience under uncertainty, as evidenced by the disturbance recovery test. By embedding logic-guided projections into the decision-making pipeline, the controller maintains compliance even under volatile DER conditions and sudden system perturbations. This robustness is further reinforced by its low cumulative cost, highlighting its efficiency across multiple operational objectives.

Importantly, the Pareto analysis reveals that the NSAI framework achieves a near-optimal balance between competing objectives, namely, control precision and actuation smoothness, underscoring its potential for real-world deployment where multi-objective optimization is indispensable.

In summary, the NSAI controller represents a paradigm shift toward hybrid intelligence in smart grid automation. It addresses long-standing trade-offs between performance and explainability, and between flexibility and safety. By unifying symbolic reasoning with neural adaptability, the framework provides a scalable, transparent, and regulation-compliant solution for next-generation energy systems.

 doi:10.20944/preprints202508.0747.v1

9. Conclusions

This study proposed a novel neuro-symbolic AI (NSAI) framework for explainable, resilient, and efficient decision-making in autonomous power grid operations. By tightly integrating symbolic reasoning with neural policy learning, the framework addresses critical limitations of existing approaches, namely, the opacity of deep learning models and the rigidity of static rule-based systems.

Extensive simulation experiments demonstrate that the NSAI controller consistently outperforms baseline strategies across multiple performance dimensions, including voltage regulation accuracy, switching efficiency, symbolic compliance, and cumulative operational cost. Notably, it achieves these improvements without sacrificing interpretability, a quality often neglected in black-box AI systems. The symbolic layer ensures rule compliance, explanation fidelity, and domain-specific safety guarantees, making the controller not only powerful but also trustworthy.

Key contributions of this work include:

- A hybrid decision-making architecture that fuses data-driven adaptability with rule-based interpretability.
- A symbolic projection mechanism that enforces operational constraints in real time.
- Quantitative metrics for explainability, including explanation traceability and semantic fidelity.
- Empirical validation on multi-objective performance using realistic smart grid scenarios.

The results affirm that NSAI offers a principled approach to operationalizing transparency in critical grid control tasks. It enables safe and auditable automation, aligning with emerging regulatory mandates for trustworthy AI in energy systems. Future work will explore real-time hardware-in-the-loop implementation, formal verification of symbolic policies, and federated extensions for distributed control in large-scale multi-agent grids.

Overall, the proposed framework lays a foundational step toward the deployment of explainable, regulation-compliant, and intelligent automation in the next generation of cyber-physical energy infrastructures.

Future Work

While the proposed neuro-symbolic AI (NSAI) framework demonstrates strong potential for autonomous and explainable grid control, several important avenues remain open for future exploration and development:

1. **Hardware-in-the-Loop (HIL) and Real-Time Deployment:** Future work will involve deploying the NSAI framework in real-time environments using hardware-in-the-loop simulation platforms. This will validate control latency, actuator responsiveness, and system integration fidelity under field-realistic timing constraints.
2. **Formal Verification of Symbolic Logic:** The safety-critical nature of power systems warrants formal guarantees. Future studies will investigate automated verification of the symbolic rule set using logic programming, model checking, and satisfiability modulo theories (SMT) to ensure correctness and safety under all admissible conditions.
3. **Federated and Distributed NSAI Architectures:** Extending the NSAI framework to multi-agent settings will allow distributed energy resources (DERs) and substations to coordinate through federated symbolic-neural policies. This includes the development of privacy-preserving inference and decentralized symbolic rule sharing.
4. **Adaptive Rule Learning and Symbol Induction:** While current rules are domain-expert defined, future directions include integrating inductive logic programming (ILP) or program synthesis to learn new symbolic rules from high-dimensional data, enabling the system to evolve its symbolic reasoning over time.
5. **Robustness under Adversarial and Fault Conditions:** Exploring the behavior of NSAI under adversarial perturbations, cyber-physical anomalies, and communication faults will be critical for ensuring its deployment in real-world smart grid infrastructures.

 doi:10.20944/preprints202508.0747.v1

6. **Human-in-the-Loop and Operator Collaboration:** Further research will incorporate human feedback loops for rule tuning, decision override, and interactive explanation, bridging machine autonomy with human oversight in critical grid operations.

These future directions aim to enhance the scalability, trustworthiness, and generalizability of neuro-symbolic intelligence for next-generation energy systems. By advancing the NSAI paradigm toward formal rigor, real-time applicability, and multi-agent coordination, this line of research contributes to the broader vision of transparent and intelligent infrastructure automation.

Appendix A: Control Cost Function and Weighting Coefficients

The unified control cost function $\mathcal{L}$ used to evaluate cumulative controller performance is defined as:

$$\mathcal{L}_t = \lambda_v \cdot \sum_{i=1}^{N} \left(V_i^t - V_{\text{ref}}\right)^2 + \lambda_s \cdot \Delta u_t + \lambda_l \cdot P_{\text{loss}}^t \tag{A1}$$

where V_i^t denotes the voltage at node i at time t, and V_{ref} is the nominal reference voltage, typically set to 1.0 p.u. The term Δu_t represents the switching effort—i.e., the number of control toggles—at time t, while P_{loss}^t captures the total active power loss in the system. The coefficients λ_v, λ_s, and λ_l are scalar weights that quantify the relative penalties assigned to voltage deviation, switching activity, and power loss, respectively.

The values used in our experiments are: $\lambda_v = 1.0$, $\lambda_s = 0.3$, and $\lambda_l = 0.5$.

Appendix B: Controller Configuration Parameters

Table B1. Controller hyperparameters for experimental evaluation.

Parameter	Value / Description
Neural Network Layers	2 hidden layers, 64 units each
Activation Function	ReLU
Learning Rate	1×10^{-3}
Replay Buffer Size	10^5 samples
Batch Size	64
Symbolic Rule Set	3 core rules (voltage, inverter limits, cycling)
Symbolic Projection Frequency	Every control step

Appendix C: NSAI Control Loop Pseudocode

Algorithm C1 NSAI-Based Voltage Control Loop

Input: Current state $\mathbf{s}_t$, symbolic rules $\mathcal{R}$, actor network μ_θ
Output: Control action u_t
Predict raw action: $u_t^{\text{raw}} \leftarrow \mu_\theta(\mathbf{s}_t)$
Evaluate rule compliance: $r_k \leftarrow \mathcal{R}(u_t^{\text{raw}}, \mathbf{s}_t)$ for each rule r_k
Project action: $u_t \leftarrow \Pi(u_t^{\text{raw}} \mid \mathcal{R})$
Apply u_t to grid environment
Store transition $(\mathbf{s}_t, u_t^{\text{raw}}, \mathcal{L}_t, \mathbf{s}_{t+1})$ in replay buffer
Update neural weights θ using sampled batch and temporal-difference error

References

1. Cavus, M. Advancing Power Systems with Renewable Energy and Intelligent Technologies: A Comprehensive Review on Grid Transformation and Integration. *Electronics* **2025**, *14*, 1159.
2. Ali, S.S.; Choi, B.J. State-of-the-art artificial intelligence techniques for distributed smart grids: A review. *Electronics* **2020**, *9*, 1030.

 doi:10.20944/preprints202508.0747.v1

3. Şahin, E.; Arslan, N.N.; Özdemir, D. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications* **2025**, *37*, 859–965.
4. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Ranjan, R. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys* **2023**, *55*, 1–33.
5. Sinha, S.; Lee, Y.M. Challenges with developing and deploying AI models and applications in industrial systems. *Discover Artificial Intelligence* **2024**, *4*, 55.
6. Bhuyan, B.P.; Ramdane-Cherif, A.; Tomar, R.; Singh, T.P. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications* **2024**, *36*, 12809–12844.
7. Alhamrouni, I.; Abdul Kahar, N.H.; Salem, M.; Swadi, M.; Zahroui, Y.; Kadhim, D.J.; Alhuyi Nazari, M. A comprehensive review on the role of artificial intelligence in power system stability, control, and protection: Insights and future directions. *Applied Sciences* **2024**, *14*, 6214.
8. Miraftabzadeh, S.M.; Di Martino, A.; Longo, M.; Zaninelli, D. Deep learning in power systems: A bibliometric analysis and future trends. *IEEE Access* **2024**. in press.
9. Javed, H.; Eid, F.; El-Sappagh, S.; Abuhmed, T. Sustainable energy management in the AI era: a comprehensive analysis of ML and DL approaches. *Computing* **2025**, *107*, 132.
10. Machlev, R.; Heistrene, L.; Perl, M.; Levy, K.Y.; Belikov, J.; Mannor, S.; Levron, Y. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI* **2022**, *9*, 100169.
11. Parisineni, S.R.A.; Pal, M. Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. *International Journal of Data Science and Analytics* **2024**, *18*, 457–466.
12. Zilberman, J. An Analysis of Evolutionary Methodology for Interpretable Logical Fuzzy Rule-Based Systems. *Journal of Biomedical and Sustainable Healthcare Applications* **2023**, *3*, 066–075.
13. SS Júnior, J.; Mendes, J.; Souza, F.; Premebida, C. Survey on deep fuzzy systems in regression applications: A view on interpretability. *International Journal of Fuzzy Systems* **2023**, *25*, 2568–2589.
14. Sathya, D.; Saravanan, G.; Thangamani, R. Fuzzy logic and its applications in mechatronic control systems. In *Computational Intelligent Techniques in Mechatronics*; Springer, 2024; pp. 211–241.
15. Ferry, J.; Laberge, G.; Aïvodji, U. Learning hybrid interpretable models: Theory, taxonomy, and methods. *arXiv preprint* **2023**, [2303.04437].
16. Bougzime, O.; Jabbar, S.; Cruz, C.; Demoly, F. Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures: Benefits and Limitations. *arXiv preprint* **2025**, [2502.11269].
17. Bougzime, O.; Cruz, C.; André, J.C.; Zhou, K.; Qi, H.J.; Demoly, F. Neuro-symbolic artificial intelligence in accelerated design for 4D printing: Status, challenges, and perspectives. *Materials & Design* **2025**, p. 113737.
18. Tian, R.; Cui, M.; Chen, G. A Neural-Symbolic Network for Interpretable Fault Diagnosis of Rolling Element Bearings Based on Temporal Logic. *IEEE Transactions on Instrumentation and Measurement* **2024**, *73*, 1–14. Art no. 3515614, https://doi.org/10.1109/TIM.2024.3373103.