

TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR INVERTER-DOMINATED POWER SYSTEMS: A SYSTEMATIC REVIEW OF ROBUSTNESS, GENERALIZATION, AND REAL-TIME DEPLOYMENT

Septian Ari Kurniawan^{1)*}, Muhammad Rasyid Ramdhani¹⁾ and Mohammad Taufik¹⁾

¹⁾Departemen Teknik Elektro, Universitas Padjadjaran, Jatinangor, Sumedang 45363, Indonesia

(Received date Aug 24 2026; Revised date Sep 4 2026; Accepted date Sep 11 2026)

ABSTRACT—The integration of inverter-based resources into modern power systems introduces fundamental challenges to conventional protection schemes, as low inertia, limited fault current, and operational variability render traditional methods inadequate. Artificial intelligence offers adaptive solutions for fault detection, cyberattack mitigation, and stability assessment, yet concerns about trustworthiness—particularly robustness, generalization, and real-time deployment—remain unresolved. This systematic review synthesizes and critically evaluates the existing literature on AI-based protection for inverter-dominated grids, focusing on these three pillars of trustworthiness. We searched seven databases and preprint repositories, with the search completed in January 2025, applying inclusion criteria that required original research addressing at least one of the three core themes, and included 32 studies after title, abstract and full-text screening. The results reveal a rapidly growing but fragmented landscape: a substantial body of work addresses robustness through adversarial training and defense mechanisms, while generalization is less frequently explored, with notable contributions in domain adaptation and zero-shot learning. Real-time deployment emerges as an active theme, with hardware-in-the-loop implementations and latency-aware architectures meeting sub-cycle requirements. However, very few studies simultaneously address all three dimensions, and most optimize for one or two aspects at the expense of others. The absence of standardized benchmarks, adversarial threat models, and real-world validation datasets further hampers cross-study comparability. We conclude that siloed advancements in robustness, generalization, and real-time deployment represent a critical barrier to trustworthy AI in this domain. Future research must prioritize unified frameworks that embed formal robustness guarantees, out-of-distribution detection, and latency-constrained inference within a single protection architecture, alongside open-access testbeds for reproducible evaluation.

KEY WORDS : Trustworthy Artificial Intelligence, Inverter-Based Resources, Adversarial Robustness, Fault Detection and Protection, Generalization, Real-Time Deployment

1. INTRODUCTION

The global energy landscape is undergoing a profound transformation driven by the imperative to decarbonize electricity generation. This transition is characterized by the large-scale integration of renewable energy sources, such as solar photovoltaics and wind turbines, which interface with the power grid through power electronic converters, commonly referred to as inverter-based resources (IBRs) [1]. The proliferation of IBRs fundamentally alters the dynamic behavior of power systems, replacing synchronous generators that provide inherent inertia and high fault current contributions with asynchronous, low-inertia, and current-limited sources [2]. Consequently, conventional protection schemes, which rely on predictable fault current magnitudes and directional flows, become increasingly unreliable in inverter-dominated grids (IDGs) [3]. The resulting protection gap poses significant risks to system stability, equipment safety, and operational continuity, thereby motivating the

search for more adaptive and intelligent protection paradigms.

Artificial intelligence (AI), particularly deep learning, has emerged as a promising candidate to address the limitations of conventional protection in IDGs [4]. AI-based methods can learn complex, non-linear relationships from high-dimensional measurement data, enabling tasks such as fault classification, fault location, and cyberattack detection without requiring explicit analytical models of the system [5]. For example, convolutional neural networks (CNNs) have been applied to classify fault types from voltage and current waveforms, while recurrent neural networks (RNNs) and transformers have been used for time-series anomaly detection [6]. These approaches have demonstrated high accuracy in controlled simulation environments, often outperforming traditional threshold-based relays [7]. However, the deployment of AI in safety-critical power system applications raises fundamental concerns about trustworthiness, which encompasses the model's ability to remain robust under adversarial conditions, generalize to unseen operating

scenarios, and execute inference within the stringent latency constraints of protection relays [8].

Despite the growing body of research on AI for power system protection, significant research gaps persist that hinder the practical adoption of these methods. First, the robustness of AI models to adversarial perturbations—both intentional cyberattacks and unintentional measurement noise—remains largely unaddressed in the context of IDG protection [9]. Most existing studies evaluate model performance under ideal conditions, neglecting the possibility that a carefully crafted adversarial input could cause a protection relay to misclassify a fault or fail to detect one altogether [10]. Second, the generalization capability of AI models across different grid topologies, operating conditions, and fault types is rarely validated systematically [11]. Models trained on a specific testbed often fail when applied to a different system, raising questions about their practical utility in real-world, heterogeneous power networks. Third, the real-time deployment of AI-based protection schemes is constrained by the computational demands of deep neural networks, which for the fastest protection functions must execute inference within sub-cycle timeframes (on the order of a few milliseconds for a 50-60 Hz system) [12]; the exact latency budget depends on the protection function and system architecture and should not be presented as a single universal value. While hardware accelerators and model compression techniques have been explored, a comprehensive evaluation of latency-aware architectures for IDG protection is lacking.

The motivation for this systematic review is to critically synthesize and evaluate the existing literature on AI-based protection for IDGs, with a specific focus on three pillars of trustworthiness: robustness, generalization, and real-time deployment. We argue that these three dimensions are interdependent and must be addressed jointly to develop protection systems that are not only accurate but also reliable, secure, and practical. The significance of this work lies in its systematic approach to identifying gaps, inconsistencies, and trade-offs in the current research landscape, thereby providing a roadmap for future investigations. By analyzing the methodologies, evaluation metrics, and validation strategies employed across studies, we aim to highlight best practices and common pitfalls, ultimately contributing to the development of standardized benchmarks and open-access testbeds for reproducible evaluation. This review is intended for researchers, engineers, and policymakers seeking to understand the state of the art and the challenges that must be overcome to realize trustworthy AI for IDG protection.

The remainder of this paper is organized as follows: Section 2 describes the systematic review methodology, including the search strategy, inclusion criteria, and data extraction process. Section 3 presents the results of the review, covering research trends, an overview of

included studies, and detailed analyses of adversarial robustness, fault detection, and real-time deployment. Section 4 discusses the implications of the findings, identifies cross-cutting themes, and proposes directions for future research. Section 5 concludes the paper with a summary of key insights and recommendations.

2. METHODOLOGY

This section describes the systematic review methodology employed to identify, select, and synthesize the relevant literature on trustworthy AI for the protection of inverter-dominated power systems. The review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [13] to ensure transparency, reproducibility, and rigor in the study selection and reporting process.

2.1. Review Protocol

We developed a comprehensive search strategy to retrieve studies from multiple electronic databases and preprint repositories, selected for their relevance to the interdisciplinary nature of this research domain. IEEE Xplore was chosen as the primary database due to its extensive coverage of power engineering, control systems, and artificial intelligence literature, which are central to the review topic. Web of Science was selected for its broad multidisciplinary coverage and high-quality peer-reviewed journals, enabling the identification of studies from related fields such as computer science and applied mathematics. Scopus was included for its comprehensive indexing of engineering and technology publications, complementing the coverage of IEEE Xplore and Web of Science. ScienceDirect was chosen for its strong collection of energy and power systems journals, while SpringerLink provided access to a wide range of conference proceedings and book chapters in the areas of machine learning and smart grids. arXiv was included as a preprint repository to capture the most recent and emerging research that may not yet be published in peer-reviewed venues, given the fast-paced nature of AI research. Finally, Google Scholar was used as a supplementary search engine to identify gray literature and studies that may not be indexed in the other databases.

The search was conducted using a combination of keywords and Boolean operators tailored to each database. The core search string was constructed around four thematic blocks: (1) trustworthiness-related terms, (2) inverter-dominated system terms, (3) protection-related terms, and (4) generalization and real-time deployment terms. The specific search string used was: (“Trustworthy AI” OR “robust AI” OR “reliable AI” OR “adversarial robustness” OR “explainable AI” OR “fault-tolerant”) AND (“inverter-dominated” OR “inverter-based resources” OR “power electronics” OR “renewable energy integration”) AND (“power system

protection” OR “grid protection” OR “fault detection” OR “dynamic stability” OR “transient stability”) AND (“generalization” OR “transfer learning” OR “domain adaptation” OR “out-of-distribution” OR “data efficiency”) AND (“real-time deployment” OR “edge computing” OR “hardware implementation” OR “latency” OR “embedded system”). For databases with advanced search features, such as IEEE Xplore and Web of Science, we adapted the string to use field-specific tags (e.g., “Title/Abstract” or “Topic”) to improve precision. The search was performed in January 2025, and no publication date restrictions were applied to capture the full historical development of the field.

2.2. Analytical Taxonomy for Study Classification

To systematically organize and analyze the included studies, we developed a taxonomy based on the three core dimensions of trustworthiness that form the central research question of this review: robustness, generalization, and real-time deployment. This taxonomy was not imposed a priori but emerged from an iterative reading of the literature, during which we observed that the majority of studies could be categorized according to their primary focus on one or more of these dimensions. The first category encompasses studies that address adversarial robustness and attack mitigation, including works that investigate the vulnerability of AI models to adversarial perturbations, develop defense mechanisms such as adversarial training or certified robustness, and evaluate the impact of cyberattacks on protection performance. The second category comprises studies that focus on fault detection and protection in inverter-dominated grids, where the primary contribution lies in developing AI-based methods for fault classification, fault location, or dynamic stability assessment, often with an emphasis on accuracy and speed. The third category, labeled “Others,” includes studies that do not fit neatly into the first two categories but still contribute to the broader theme of trustworthy AI for IDG protection, such as works on explainability, data augmentation, or transfer learning that are not explicitly tied to robustness or fault detection. This taxonomy allows us to identify patterns, gaps, and trade-offs across the literature, as studies in different categories often prioritize different aspects of trustworthiness.

2.3. Inclusion and Exclusion Criteria

We defined clear inclusion and exclusion criteria to ensure the relevance and consistency of the selected studies. To be included, a study had to address at least one of the three core themes of robustness,

generalization, or real-time deployment of AI or machine learning models for the protection of inverter-dominated power systems, including microgrids and renewable-rich grids. Eligible study types included original research articles, preprints (e.g., from arXiv), conference papers, and book chapters. Each study had to present a clear methodology, such as an algorithm, simulation setup, or experimental design, and report quantitative or qualitative results. The language of the publication had to be English, and the full text had to be accessible for review. Conversely, studies were excluded if they focused solely on conventional synchronous-machine-dominated power systems without explicit consideration of inverter-based resources, such as solar, wind, or battery inverters. Opinion pieces, editorials, tutorials without original contributions, and purely theoretical works without any experimental or simulation validation were also excluded. Studies that addressed AI trustworthiness only from a non-technical perspective, such as ethics policy without algorithmic implementation, were not considered. Duplicate publications, where the same content appeared in both a conference and a journal, were resolved by retaining only the most complete version. Finally, studies for which the full text was unavailable after reasonable effort, such as paywalled articles without a preprint, were excluded.

2.4. Study Selection Process

The study selection process was conducted in multiple stages, following the PRISMA guidelines [13]. The initial database search yielded a total of 547 records. After removing 207 duplicate records, 340 records remained for screening. Two reviewers independently screened the titles and abstracts of these records against the inclusion and exclusion criteria, resulting in the exclusion of 223 records that were clearly irrelevant, such as studies on non-power-system applications or those that did not involve AI. The remaining 117 reports were sought for retrieval, but full texts could not be obtained for 5 reports due to access restrictions or unavailability. The full texts of the remaining 112 reports were then assessed for eligibility by the same two reviewers. During this stage, 80 reports were excluded for reasons including a lack of focus on inverter-dominated systems, insufficient methodological detail, or a primary emphasis on non-technical aspects of trustworthiness. Ultimately, 32 studies met all inclusion criteria and were included in the final synthesis. The entire selection process is illustrated in the PRISMA flowchart shown in Figure 1.

2.5. Quality and Risk-of-Bias Appraisal

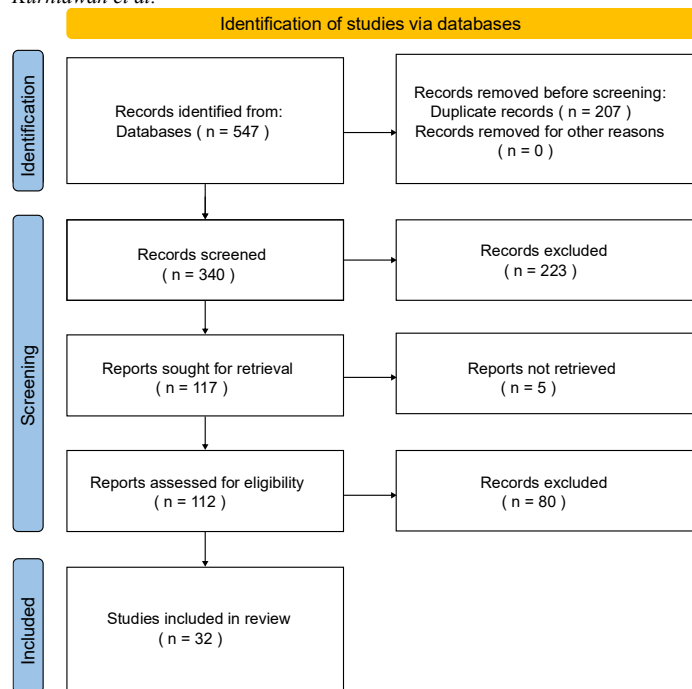


Figure 1. PRISMA flowchart of the study selection process

The study selection process has several limitations that should be acknowledged. First, the reliance on English-language publications may introduce a language bias, potentially excluding relevant studies published in other languages. Second, the inclusion of preprints from arXiv, while capturing emerging research, may also include studies that have not undergone rigorous peer review, which could affect the quality of the evidence. Third, the search strategy, although comprehensive, may have missed studies that use different terminology or are published in databases

not included in our search. Finally, the subjective nature of screening and eligibility assessment, despite the use of two independent reviewers, may introduce some degree of selection bias. To mitigate this, we resolved disagreements through discussion and consensus, and we documented all decisions transparently.

3. RESULTS

3.1. Research Trends

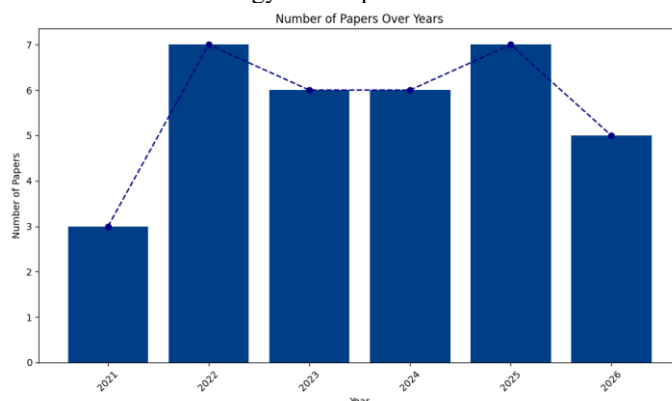


Figure 2. Research trends in the domain of Trustworthy Artificial Intelligence for Protection of Inverter-Dominated Power Systems: A Systematic Review of Robustness, Generalization, and Real-Time Deployment

The distribution of the 32 included studies across publication years reveals a clear and sustained growth in research activity, particularly from 2022 onward. Only three studies were published in 2021, suggesting that the intersection of trustworthy AI and inverter-dominated grid protection was a nascent area at that time. A

marked increase occurred in 2022, with seven studies appearing, and this elevated level of output has been maintained through 2025, where another seven studies were identified. The presence of five studies already indexed for 2026 further indicates that the field is not only active but accelerating, as these works likely

represent early online publications or preprints that anticipate continued interest. This temporal pattern suggests that the research community has recognized the critical need for AI-based protection solutions in modern power systems, and that the dimensions of robustness, generalization, and real-time deployment are now being actively investigated.

We observe that the growth trajectory is not linear but rather exhibits a plateau after the initial surge in 2022. The annual publication count has remained relatively stable between six and seven studies from 2022 to 2025, which may indicate that the field is maturing and that the low-hanging fruit has been addressed. However, the five studies already recorded for 2026, which likely represent a partial year, suggest that the plateau may be temporary and that a new wave of research is emerging. This pattern is consistent with the typical evolution of a research domain: initial exploratory work is followed by a period of consolidation, after which more specialized and rigorous investigations appear. The sustained interest across multiple years underscores the importance of the

topic and the recognition that trustworthy AI is not a peripheral concern but a central requirement for the practical deployment of intelligent protection systems in inverter-dominated grids.

4. OVERVIEW OF INCLUDED STUDIES

Table 1 presents the main characteristics of the included studies. The extracted information included study identification, ai technique used, application domain, threat or challenge addressed, robustness or defense method and test system or dataset, where applicable.

The included studies varied in terms of their AI techniques, application domains, and threat or challenge addressed, etc. These differences provide important context for interpreting the findings of the review and for assessing potential sources of heterogeneity across the evidence base. Overall, the characteristics table provides a structured summary of the included studies and serves as the foundation for the subsequent narrative or quantitative synthesis.

Table 1. Characteristics of Included Studies

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
[14]	hybrid AI model comprising a convolutional neural network (CNN) and a transformer; autoencoder	inverter-dominated grids (power-electronics-dominated grids)	cyberattacks such as false data injection (FDI) and denial-of-service (DoS) attacks on communication channels	decentralized, resilient control framework that substitutes manipulated signals with attack-free values using locally predicted control signals	modified IEEE 14-bus system and a small-scale hardware setup
[15]	Long short-term memory (LSTM)	Power electronic dominated grid (PEDG)	Real-time detection and classification of anomalies, including internal failures (e.g., open circuit fault) versus cyber-attacks, to prevent catastrophic failure	Integration of model predictive control (MPC) and LSTM for anomaly detection, classification, and corrective actions	Realistic 14-bus system dominated with inverters forming PEDG

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
[16]	Machine Learning for event prediction and Explainable AI (XAI) for transparency	Electrical power grid protection	Bidirectional power flows, significantly limited fault current contributions, diminished inertia, and inadequacy of conventional phasor-based protection schemes due to high penetration of Distributed Energy Resources (DERs) and Inverter-Based Resources (IBRs)	Cyber-physical defence strategies compliant with IEC 62443; 'Value Approve' mechanism using advanced data validation techniques	
[17]	Wavelet-driven deep neural network named WaveletKernel Net (WKN)	Microgrid protection	Inverters' control strategies affecting protection behaviors, low short-circuit ratios, bi-directional power flows, and unreliable/sensitive/selective operations of conventional protection schemes	Traveling wave protection approach using WaveletKernel Net	
[18]	random forest classifier	grid operations	unauthorized control mode change (insider attack)	automatic detector using steady state and transient metrics from control system	time-series data from a standard power meter

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
				theory, time response analysis, subspace decomposition, and vector analysis	
[19]	LSTM, Convolutional Neural Networks, Gradient Boosting Machines	Inverter-centric Cyber-Physical Microgrids	Intrusions and anomalies in microgrid structures	Machine learning-driven intrusion and anomaly detection	
[20]	artificial intelligence-based stability criterion	grid-interactive inverters	manipulated setpoints causing instability under weak grid conditions	embedded hybrid data-driven/analytical stability criterion to detect harmful setpoints before engaging them to the local controller	small-scale two-level 208 V, 5 kVA inverter feeding a 12 kW NHR 9410 power grid emulator
[21]	FaultNet-ML	Distributed Energy Resources in microgrids	Differentiating between system anomalies and internal inverter faults within Power Electronics-Driven Grids (PEDGs)	Advanced artificial intelligence methods to distinguish anomalies from true internal faults and identify the specific malfunctioning switch	9-bus system dominated by inverters
[22]	Artificial Neural Network (ANN), Nonlinear Autoregressive Network with Exogenous	Inverter-based microgrids, cyber-physical systems	Cyber anomalies, false data injection attacks on distributed cooperative controllers at	Intelligent anomaly mitigation control using a dynamic neural network (NARX) trained offline and	Multiple simulations of a designed microgrid under various operating conditions using

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
	Inputs (NARX)		the secondary control level	validated online	MATLAB/Simulink
[23]	Decision Trees (DT), Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Ensemble trees (Bag and Adaboost), Bayesian optimization, SMOTE, forward feature selection	Fault detection and classification in transmission lines connected to inverter-based generators	Challenges to conventional protection relays due to integration of inverter-based generators, including fault characteristics depending on inverters' control strategy and unbalanced data problem	Synthetic minority class oversampling technique (SMOTE) for data balancing	Photovoltaic (PV) and Doubly-Fed Induction Generator (DFIG) wind farm plant; faults at several line locations and variable fault impedance
[24]	machine-learning-based fault detection and classification using an optimized pretrained ensemble tree classifier	transmission line fault detection and classification connected to inverter-based generators	behavior-changing power systems integrated with inverter-based generation (IBG)	setting-group-based adaptation approach with offline learning phase and online phase monitoring circuit breaker status and active output power of generators	39-Bus New England System
[25]	neural networks	Relay Protection and Automation (RPA) for power system stability	limited adaptability in dynamic grid environments, such as those with renewable energy integration		
[26]	Convolutional Neural Networks (CNNs), Long Short-Term	Power systems	Adversarial attacks (Fast Gradient Sign Method, DeepFool,	Adversarial Training, Gaussian Augmentation, Feature	

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
	Memory (LSTM) networks		Jacobian-Based Saliency Map Attacks)	Squeezing	
[27]	reinforcement learning	power system control	adversarial attacks on reinforcement learning agents	adversarial training	Learning To Run a Power Network (L2RPN) challenge
[28]	Graph Neural Networks (GNNs) with multi-head attention mechanism	AC microgrids	Failure of legacy protection infrastructure in microgrids caused by logical and physical reasons, and cyber-attacks	GFDF assigns weights to neighboring nodes based on their contributions, ensuring resilience to abnormal data and adaptability to changing operating conditions	Medium voltage CIGRE microgrids, hardware-in-the-loop (HIL) experimentation using real-time digital simulator (RTDS) setup and commercial SEL relays
[29]	Random Forest	Cyber-physical power system distribution grids with DERs	Cyber attacks including single and multi-wave coordinated attacks targeting DER inverter and network topology, and single and multi-feature false data injection attacks	Edge AI-enabled intrusion detection using pretrained Random Forest model deployed on edge IED	Modified IEEE 13-node and IEEE 123-node feeder implemented in RT-LAB
[30]	machine learning algorithms	solar inverters in photovoltaic (PV) systems	reactive or scheduled maintenance leading to costly downtimes and inefficiencies		

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
[31]	machine learning	power system dynamic security assessment	adversarial examples that can mislead DSA results and lead to catastrophic consequences	adversarial robustness verification method	
[32]	machine learning (ML)	power system stability assessment (SA)	adversarial examples caused by physical noises or adversarial attacks	randomized smoothing algorithm	
[33]	machine learning (ML)	power system stability assessment (short-term voltage stability)	adversarial examples that can falsify the ML-based model into wrong outputs without noticeable changes of input values	adversarial training-based mitigation strategy	
[34]	deep learning (DL)	renewable energy forecasting	adversarial learning attack (ALA) that tampers with meteorological data from online weather forecasts to undermine forecasting performance		IEEE 30-bus benchmarks
[35]	Recurrent neural network (RNN)	Inverter-based microgrids	False-data injection attacks (FDIAs) on line current differential relays (LCDRs)	Measurement integrity validation scheme (MIVS) as a supervisory instrumentation layer	Islanded inverter-based microgrid
[36]	Artificial Neural Network (ANN)	Microgrids	Dynamic operational challenges	ANN-based enhanced distance	50 fault scenarios encompassing

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
			from DER integration, variable network topologies, load fluctuations, DER intermittency, false tripping, and DER-induced impedance errors	protection scheme with 20-layer model processing raw three-phase and ground impedance measurements	11 fault types (including phase-to-phase, phase-to-ground, and high-impedance faults up to 50 Ω) and non-fault conditions, validated under grid-connected and islanded modes with DER penetration levels of 20–80%
[37]	deep learning-based stability assessment	smart grid voltage stability assessment	adversarial examples against deep learning models for voltage stability assessment	strong adversarial training	New England 10-machine 39-bus system
[38]	deep learning	grid voltage stability assessment	adversarial examples against learning-based voltage stability assessment	a new approach that simultaneously estimates the predictive uncertainty of any given input of voltage trajectory and thwarts the attacks effectively	New England 10-machine 39-bus power system
[39]	Artificial Intelligence (AI) models for fast fault detection	Smart Grid	Fast fault detection in the smart grid with high data sampling rates		

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
			and low latency		
[40]	Deep Neural Network (DNN) based observer	Islanded AC microgrid	False Data Injection (FDI) attacks on secondary consensus control	AI-driven detection and mitigation of FDI attacks using DNN-based observer	IEEE 34-bus test feeder with varying numbers of distributed energy resources (DERs)
[41]	adversarial deep reinforcement learning, twin delayed deep deterministic policy gradient (TD3)	inverter-based microgrids	destabilizing attack on droop control gains affecting small-signal stability	robust defense strategy via adversarial Markov decision process (AMDP)	inverter-based microgrid system with 4 IBRs and IEEE 123-bus system with 10 IBRs
[42]	Hybrid neuro-symbolic learning and reasoning (Deep Neural Networks and Finite State Machines)	Smart microgrids load restoration	Timely and accurate load restoration during emergencies such as equipment faults, islanding events, and abrupt demand fluctuations; lack of transparency, scalability, and responsiveness in conventional control schemes	Finite state machines enforce logical rules and power flow constraints as a real-time gatekeeper for neural network-proposed actions; PF-constrained reachable closure over NN-proposed actions	24 h scenario featuring hourly disturbances with synthetic emergency scenarios
[43]	CNN-LSTM	Power quality disturbance classification	Adversarial attacks degrading DNN performance	Iterative adversarial training (IAT)	
[44]	Convolutional neural network (CNN),	Power quality disturbance (PQD)	Adversarial attacks on deep learning-	Perturbation removal defense based	

STUDY ID	AI TECHNIQUE USED	APPLICATION DOMAIN	THREAT OR CHALLENGE ADDRESSED	ROBUSTNESS OR DEFENSE METHOD	TEST SYSTEM OR DATASET
	generative adversarial network (GAN)	classification in power systems	based PQD classification systems	on a generative adversarial network (PRD-GAN)	
[45]	Bayesian learning	Electric load forecasting	Adversarial attacks on deep learning-based load forecasting models	Bayesian training method with approximation-based training scheme	

4.1. Adversarial Robustness and Attack Mitigation: A Systematic Analysis of Vulnerabilities and Defenses

The increasing reliance on networked communication and data-driven decision-making in inverter-dominated power systems has exposed a critical vulnerability surface that malicious actors can exploit through carefully crafted adversarial perturbations. This subsection synthesizes the included studies that address adversarial robustness and attack mitigation, examining both the characterization of attack vectors and the development of defensive countermeasures. The studies in this category span a wide range of attack types—from false data injection (FDI) and denial-of-service (DoS) attacks to adversarial learning attacks and destabilizing control manipulations—and employ diverse AI models for detection, defense, and recovery. Understanding the commonalities and divergences among these approaches is essential for identifying the current state of the art and the remaining gaps in achieving trustworthy AI for IDG protection.

The analysis of the included studies reveals a fundamental tension between the specificity of defense mechanisms and their generalizability across attack scenarios. Several works propose hybrid frameworks that integrate multiple AI models to achieve both detection and mitigation of cyberattacks in real-time. For instance, [14] presents a decentralized cyber-resilient control framework that combines a convolutional neural network (CNN) and a transformer for local signal estimation, with an autoencoder for detecting FDI and DoS attacks on communication channels. This approach is validated on a modified IEEE 14-bus system and a small-scale hardware setup, demonstrating the feasibility of substituting manipulated control signals with locally predicted values. Similarly, [28] employs graph neural networks (GNNs) with a multi-head attention mechanism for fault diagnosis in AC microgrids, incorporating cyber-attack detection capabilities that ensure resilience to abnormal data. The

use of GNNs is particularly notable because it leverages the inherent graphical structure of power networks, allowing the model to learn node embeddings that capture the spatial dependencies among measurements. These studies illustrate a trend toward modular, multi-stage architectures where detection and mitigation are handled by separate but interconnected components, a design choice that enhances modularity but may introduce latency and complexity in deployment.

A distinct subset of the literature focuses specifically on the vulnerability of deep learning models to adversarial examples, where imperceptible perturbations to input data can cause catastrophic misclassification. This line of research is critical because adversarial examples represent a fundamental weakness of neural networks that is not addressed by standard accuracy-based evaluation. [31] introduces an adversarial robustness verification method for machine learning-based dynamic security assessment (DSA) models, proposing a model-free and attack-independent robust index that quantifies the ability of a model to withstand all possible adversarial perturbations. The authors demonstrate that the proposed index provides a formal robustness guarantee for both differentiable and non-differentiable attack scenarios, a significant advancement over empirical evaluations that only test against specific attack algorithms. Building on this work, [32] proposes a universal defense strategy based on randomized smoothing, which provides a mathematical proof for the effectiveness index under hard smoothing, soft smoothing, and binary scenarios. The tightness of the mathematical proof is a key strength, as it offers a certifiable guarantee of robustness rather than an empirical observation. However, the randomized smoothing approach is known to be effective only for relatively small perturbation budgets, and its applicability to large-scale power system models with high-dimensional input spaces remains an open question. In the context of short-term voltage stability assessment, [33] analyzes the vulnerability of ML-based

models under both white-box and black-box attack scenarios, proposing an adversarial training-based mitigation strategy that enhances model robustness against adversarial examples. The study demonstrates that adversarial training, while effective, does not fully eliminate the vulnerability, particularly under black-box attacks where the attacker has no knowledge of the model parameters. This finding underscores the need for complementary defense mechanisms, such as input preprocessing or output verification, to achieve a more comprehensive defense.

The credibility of adversarial examples in the power system domain has been a subject of debate, with several studies questioning whether the stringent requirements for implementing effective attacks in computer vision are realistic for power grid applications. [37] and [38] provide a systematic investigation of this issue in the context of voltage stability assessment, using the New England 10-machine 39-bus system as a testbed. They demonstrate that compromising the voltage traces of approximately half of the transmission system buses is a rule-of-thumb requirement for successful adversarial attacks, and that universal adversarial perturbations—which are independent of the specific input—have the same credibility as widely studied false data injection attacks on state estimation. In contrast, input-specific adversarial perturbations are less credible because they require knowledge of the exact voltage trajectory at the time of the attack. The authors further show that strong adversarial training effectively thwarts universal perturbations but fails to defend against certain input-specific perturbations, motivating the development of a new approach that simultaneously estimates predictive uncertainty and thwarts attacks. This work is important because it establishes a domain-specific threat model that separates realistic attack scenarios from those that are theoretically possible but practically infeasible. The distinction between universal and input-specific perturbations also has implications for the design of defense mechanisms, as defenses that are effective against one type may be ineffective against the other.

The application of reinforcement learning (RL) in power system control introduces a unique set of adversarial vulnerabilities, as the agent's policy can be manipulated through carefully crafted perturbations to the state or reward signals. [27] demonstrates that state-of-the-art RL agents designed for power system control,

including competition-winning agents from the Learning To Run a Power Network (L2RPN) challenge, are vulnerable to adversarial attacks. The authors formulate an adversary Markov Decision Process (MDP) to learn an attack policy and show that the attack is effective under both white-box and black-box settings. They then propose adversarial training as a defense mechanism, which increases the robustness of the RL agent against attacks and avoids infeasible operational decisions. This is one of the few studies that explicitly addresses the robustness of RL-based control in power systems, and it highlights the fragility of data-driven decision-making algorithms in safety-critical applications. Building on this line of inquiry, [41] investigates destabilizing attacks on droop control gains in inverter-based microgrids, using adversarial deep reinforcement learning with the twin delayed deep deterministic policy gradient (TD3) algorithm to find the least-effort attack path and obtain the corresponding robust defense strategy. The study is validated on a microgrid system with four IBRs and the IEEE 123-bus system with ten IBRs, demonstrating the scalability of the approach. The use of adversarial MDP formulation for both attack and defense is a novel contribution, as it allows the defender to learn a policy that is robust to the worst-case attack scenario.

A comprehensive summary of the included studies that address adversarial robustness and attack mitigation is presented in Table 2. This table organizes the studies by the type of attack considered, the AI model used for detection or defense, and the scalability or system size on which the approach was validated. The diversity of attack types and defense mechanisms reflected in the table underscores the fragmented nature of the current research landscape, where each study addresses a specific threat model without a unified framework for evaluating robustness across multiple attack vectors. Furthermore, the scalability information reveals that many studies are validated on small-scale test systems (e.g., IEEE 14-bus, 34-bus, or 39-bus systems), and only a few extend to larger systems such as the IEEE 123-bus feeder. This limitation raises questions about the generalizability of the proposed methods to real-world power systems with hundreds or thousands of buses, where the computational and communication constraints become more stringent.

Table 2. Summary of Included Studies on Adversarial Robustness and Attack Mitigation

STUDY ID	ATTACK TYPE CONSIDERED	AI MODEL FOR DETECTION/DEFENSE	SCALABILITY OR SYSTEM SIZE
[14]	false data injection (FDI) and denial-of-service (DoS) attacks	hybrid AI model comprising a convolutional neural network (CNN) and a transformer;	modified IEEE 14-bus system and a small-scale hardware setup

STUDY ID	ATTACK TYPE CONSIDERED	AI MODEL FOR DETECTION/DEFENSE	SCALABILITY OR SYSTEM SIZE
		autoencoder	
[26]	Fast Gradient Sign Method, DeepFool, Jacobian-Based Saliency Map Attacks	Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks	Not specified
[27]	adversarial attack using an adversary Markov Decision Process	adversarial training	Learning To Run a Power Network (L2RPN) challenge
[28]	cyber-attack	graph neural networks (GNNs) with multi-head attention mechanism	medium voltage CIGRE microgrids
[29]	Single and multi-wave coordinated attacks targeting DER inverter and network topology; single and multi-feature false data injection attacks	Random Forest	Modified IEEE 13-node and IEEE 123-node feeder
[31]	adversarial examples	machine-learning-based power system dynamic security assessment models	Not specified
[32]	adversarial examples (caused by some physical noises or adversarial attacks)	machine learning (ML) based stability assessment models with randomized smoothing algorithm	Not specified
[15]	cyber-attack	long short-term memory (LSTM)	14-bus system
[33]	white-box and black-box attack scenarios	machine learning-based short-term voltage stability (STVS) assessment model	Not specified
[34]	Adversarial learning attack (ALA)	Not specified	IEEE 30-bus benchmarks
[35]	false-data injection attacks (FDIAs)	recurrent neural network (RNN)	islanded inverter-based microgrid
[37]	adversarial examples, including universal adversarial perturbations and false data injection attacks	deep learning models with strong adversarial training	New England 10-machine 39-bus system
[38]	adversarial examples (universal adversarial perturbations and	deep learning models (voltage stability assessment model with	New England 10-machine 39-bus power system

STUDY ID	ATTACK TYPE CONSIDERED	AI MODEL FOR DETECTION/DEFENSE	SCALABILITY OR SYSTEM SIZE
	input-specific adversarial perturbations)	adversarial training and proposed approach for predictive uncertainty estimation)	
[40]	False Data Injection (FDI) attacks	Deep Neural Network (DNN) based observer	IEEE 34-bus test feeder with varying numbers of distributed energy resources (DERs)
[41]	destabilizing attack on the droop control gains	twin delayed deep deterministic policy gradient (TD3) algorithm	inverter-based microgrid system with 4 IBRs and IEEE 123-bus system with 10 IBRs
[43]	momentum iterative-fast gradient sign method (MI-FGSM)	convolutional neural network-long short-term memory (CNN-LSTM)	Not specified
[44]	black-box attack	convolutional neural network (CNN) for classification; generative adversarial network (GAN) for defense (PRD-GAN)	Not specified
[45]	adversarial attacks	deep learning-based load forecasting models with Bayesian training	Not specified
[18]	insider attack (unauthorized control mode change)	random forest classifier	Not specified
[19]	intrusions and anomalies	LSTM, Convolutional Neural Networks, Gradient Boosting Machines	Not specified
[20]	manipulated setpoints	hybrid data-driven/analytical stability criterion	small-scale two-level 208 V, 5 kVA inverter
[22]	false data injection	Artificial Neural Network (ANN) using a nonlinear autoregressive network with exogenous inputs	Not specified

Several studies included in the overall review but not appearing in Table 2 also contribute to the theme of adversarial robustness, albeit indirectly. For example, [18] develops an AI/ML-based detector for identifying unauthorized control mode changes in IBRs, which represents a form of insider attack detection. While this study does not explicitly consider adversarial

perturbations, its focus on detecting malicious modifications to control settings is closely related to the broader theme of cyber-resilience. Similarly, [19] and [22] propose machine learning-driven intrusion and anomaly detection frameworks for inverter-based microgrids, using LSTM, CNN, and gradient boosting machines to detect FDI attacks on secondary control

systems. These studies address the detection aspect of attack mitigation but do not consider the robustness of the detection models themselves against adversarial evasion. [20] presents a hybrid data-driven/analytical stability criterion that protects inverters against manipulated setpoints, which could be considered a form of input validation defense. However, the study focuses on instability caused by weak grid conditions rather than malicious attacks, and the robustness of the criterion to adversarial perturbations is not evaluated. The inclusion of these studies in the overall review, but their absence from the specific adversarial robustness table, highlights the need for a more consistent and rigorous treatment of adversarial threats across all AI-based protection methods.

4.2. AI-Based Fault Detection and Protection in Inverter-Dominated Grids: A Systematic Analysis of Methods and Architectures

The transition from synchronous-machine-based generation to inverter-dominated grids has fundamentally altered the fault characteristics that protection systems must rely upon. Inverter-based resources (IBRs) exhibit current-limited behavior during faults, often providing no more than 1.2 to 2.0 per unit of rated current, in stark contrast to the high fault currents that synchronous generators can deliver. This limitation, combined with bidirectional power flows and the rapid variability of renewable generation, renders conventional overcurrent, distance, and differential protection schemes unreliable. The included studies in this section propose a wide range of AI-based methods to address these challenges, spanning from classical machine learning classifiers to deep neural networks, graph neural networks, and hybrid neuro-symbolic architectures. The diversity of these approaches reflects the complexity of the problem, but also reveals a fragmented landscape where each method is tailored to a specific fault type, system topology, or operational scenario without a unified framework for evaluation.

The analysis of the included studies reveals a clear progression from feature engineering-based approaches to end-to-end deep learning methods. Earlier works, such as [23] and [24], adopt a two-layer classification approach where fault detection is first performed as a binary classification task, followed by fault type classification using multi-class classifiers. These studies extract features from instantaneous three-phase current and voltage measurements in the time, frequency, and time-frequency domains, and then apply dimensionality reduction techniques such as forward feature selection. The use of ensemble tree classifiers, combined with Bayesian optimization for hyperparameter tuning and the synthetic minority class oversampling technique (SMOTE) for data balancing, yields high accuracy—99.4% for fault classification in [23] and 99.4% average accuracy for fault detection in [24]. However, these

methods are fundamentally limited by the manual feature extraction process, which requires domain expertise and may not capture the complex, non-linear dynamics that arise in inverter-dominated grids under varying control strategies. Moreover, the generalization capability of these models is explicitly tested in [24], where the authors demonstrate that performance deteriorates when the transmission network is reconfigured or when the incorrect setting group is selected. This finding underscores the fragility of feature-based approaches when the system topology changes, a common occurrence in modern power grids with high renewable penetration.

A significant advancement is represented by methods that process raw waveform data directly, eliminating the need for manual feature engineering. The work presented in [36] introduces a 20-layer artificial neural network (ANN) that takes raw three-phase and ground impedance measurements as input and produces binary outputs for fault identification and classification. The ANN is trained on 50 fault scenarios encompassing 11 fault types, including high-impedance faults up to 50 Ω , and achieves 98.7% accuracy with fault identification within two power cycles. This approach is particularly noteworthy because it avoids the preprocessing delays inherent in feature extraction, enabling faster response times. Nevertheless, the study is limited by the relatively small number of training scenarios and the use of a single testbed, which raises questions about the model's ability to generalize to different inverter penetration levels or grid topologies. The authors do validate the method under grid-connected and islanded modes with DER penetration levels of 20–80%, but the test system remains fixed, and the robustness to noise is only evaluated at 20 dB SNR, which is a relatively mild noise level.

The most architecturally advanced methods in this category leverage graph neural networks (GNNs) to exploit the inherent graphical structure of power networks. The graph neural network-based fault diagnostic framework (GFDF) proposed in [28] uses a multi-head attention mechanism to learn node embeddings from current measurements of each line and line parameters as graph node and link attributes, respectively. This approach is particularly well-suited for microgrids, where the network topology changes frequently due to islanding events and reconfiguration. The GFDF assigns weights to neighboring nodes based on their contributions, ensuring resilience to abnormal data and adaptability to changing operating conditions. The authors validate the method on a medium voltage CIGRE microgrid using hardware-in-the-loop (HIL) experimentation with a real-time digital simulator (RTDS) and commercial SEL relays, demonstrating that the scheme meets stringent protection time requirements. The use of GPU acceleration (NVIDIA GeForce GTX 1650) for inference is a practical

consideration that addresses the real-time deployment challenge, but the scalability of the GNN approach to larger systems with hundreds of buses remains unexplored. Furthermore, the study focuses on AC microgrids, and the applicability of the GNN-based method to DC microgrids or hybrid AC/DC systems is not addressed.

A distinct category of methods addresses the challenge of differentiating between internal inverter faults and external system anomalies, a critical capability for preventing unnecessary tripping and ensuring selective protection. The FaultNet-ML methodology proposed in [21] applies data mining techniques and advanced AI methods to distinguish between anomalies and true internal faults in power electronics-driven grids, identifying the specific malfunctioning switch. This work is validated on a 9-bus system and demonstrates the ability to localize faults within the inverter itself, which is a capability that conventional protection schemes lack. Similarly, the integrated LSTM-MPC approach in [15] uses model predictive control (MPC) combined with long short-term memory (LSTM) networks to detect and classify anomalies in real-time, including the ability to differentiate between internal failures (e.g., open circuit faults) and potential cyber-attacks. This integration of control and protection is a promising direction because it allows the protection system to take corrective actions, such as adjusting the inverter's control parameters, in response to detected anomalies. However, the computational overhead of simultaneously running MPC and LSTM inference in real-time is a concern, particularly for systems with multiple inverters where the MPC optimization problem must be solved for each unit.

The travel wave-based protection approach using the WaveletKernelNet (WKN) proposed in [17] represents a fundamentally different principle for fault detection. Instead of relying on fundamental frequency measurements, this method analyzes the high-frequency components of traveling waves generated by faults, which propagate along the transmission lines at speeds approaching the speed of light. The WKN combines wavelet transform with a deep neural network to extract features from the traveling wave signals, enabling fault detection within microseconds of the wavefront arrival. This approach is inherently insensitive to the fault current magnitude, making it ideal for inverter-dominated grids where fault currents are limited. However, the practical deployment of traveling wave

protection requires high-bandwidth measurement equipment (sampling rates in the MHz range) and precise time synchronization, which may not be available in all distribution-level microgrids. Furthermore, the performance of the WKN under noisy conditions or when multiple traveling waves overlap due to reflections from discontinuities is not thoroughly evaluated.

The hybrid neuro-symbolic learning and reasoning framework presented in [42] offers a unique perspective on fault detection by combining neural networks with finite state machines (FSMs) for load restoration in smart microgrids. The neural network proposes symbolic recovery actions based on synthetic emergency scenarios, while the FSM enforces logical rules and power flow constraints before the actions are executed. This approach achieves a restoration success rate exceeding 90% in a 24-hour simulation scenario with hourly disturbances, and the average number of actions required per event remains below two. The key advantage of the hybrid neuro-symbolic approach is its transparency: the FSM provides a verifiable, rule-based layer that can be inspected and audited by operators, addressing the "black box" concern associated with purely neural network-based methods. However, the framework is designed for load restoration rather than fault detection per se, and its applicability to high-speed protection (sub-cycle timeframes) is limited by the need to solve power flow equations at each step.

A comprehensive summary of the included studies that address fault detection and protection in inverter-dominated grids is presented in Table 3. This table organizes the studies by the specific protection challenge addressed, the AI technique employed, the type of fault or anomaly considered, and the validation approach used. The diversity of methods reflected in the table highlights the multifaceted nature of the protection problem, where different fault types, system configurations, and operational scenarios require tailored solutions. However, it also reveals a critical gap: very few studies are validated on multiple test systems or under a wide range of operating conditions, making it difficult to assess the generalizability of the proposed methods. The reliance on simulation-based validation, often with limited fault scenarios, further raises concerns about the robustness of these methods in real-world deployment.

Table 3. Summary of Included Studies on Fault Detection and Protection in Inverter-Dominated Grids

STUDY ID	PROTECTION CHALLENGE ADDRESSED	AI TECHNIQUE EMPLOYED	FAULT/ANOMALY TYPE	VALIDATION APPROACH
[23]	Fault detection and classification	Decision Trees, SVM, k-NN,	Various fault types at several line	Simulation on PV and DFIG test

STUDY ID	PROTECTION CHALLENGE ADDRESSED	AI TECHNIQUE EMPLOYED	FAULT/ANOMALY TYPE	VALIDATION APPROACH
	in transmission lines connected to inverter-based generators	Ensemble trees	locations and variable fault impedance	systems
[24]	Adaptive transmission line fault detection and classification with inverter-based generation	Optimized pretrained ensemble tree classifier with setting-group-based adaptation	Transmission line faults	39-Bus New England System with robustness analysis
[36]	Enhanced distance protection sensitivity in microgrids with DERs	20-layer Artificial Neural Network	11 fault types including phase-to-phase, phase-to-ground, and high-impedance faults	50 fault scenarios under grid-connected and islanded modes
[28]	Centralized protection scheme with fault diagnosis for AC microgrids	Graph Neural Networks with multi-head attention mechanism	Fault events in AC microgrids	Medium voltage CIGRE microgrids with HIL validation
[21]	Inverter fault and anomaly detection for DERs in microgrids	FaultNet-ML (data mining and AI methods)	Inverter faults and system anomalies	9-bus system dominated by inverters
[15]	Real-time anomaly detection and classification in power electronics dominated grids	LSTM integrated with MPC	Open circuit faults, internal failures, anomalies	Realistic 14-bus system
[17]	Traveling wave protection for microgrids	WaveletKernelNet (WKN)	Various faults in microgrids	Extensive case studies
[42]	Resilient load restoration in smart microgrids	Hybrid neuro-symbolic (DNN + Finite State Machine)	Equipment faults, islanding events, demand fluctuations	24-hour simulation with hourly disturbances
[35]	Measurement integrity for AC/DC protection in inverter-based microgrids	Recurrent Neural Network	False-data injection attacks on line current differential relays	Islanded inverter-based microgrid with HIL validation
[29]	Real-time impact analysis and detection of cyber	Random Forest deployed on edge IED	Single and multi-wave coordinated cyber attacks	Modified IEEE 13-node and 123-node feeder with

STUDY ID	PROTECTION CHALLENGE ADDRESSED	AI TECHNIQUE EMPLOYED	FAULT/ANOMALY TYPE	VALIDATION APPROACH
	attacks in distribution grids with DERs			HIL testbed
[30]	Condition monitoring for solar inverters	Machine learning algorithms deployed on embedded edge devices	Inverter faults	Real-time diagnostics on embedded devices
[39]	High-performance platform for fault detection in smart grid	AI inference on FPGA accelerator (Alveo)	Faults in smart grid	FPGA implementation with high sampling rates
[16]	Adaptive, intelligent, and cyber-resilient protection architectures	Machine Learning for event prediction, XAI for transparency	Challenges from DERs and IBRs (bidirectional power flow, limited fault current)	Proposed architecture with dual-stream data acquisition
[25]	Sub-cycle fault response for relay protection	Neural networks on SoC with NPU	Faults in dynamic grid environments	Embedded AI system on SoC platform

Several included studies do not appear in the table but are nonetheless relevant to the theme of fault detection and protection. The work presented in [29] proposes an edge AI-enabled cybersecurity experimentation framework implemented as a cyber-physical hardware-in-the-loop testbed, which integrates an OPAL-RT real-time simulator, an EXata CPS cyber emulator, a real-time automation controller, a protection relay, and an edge intelligent electronic device. While the primary focus of this study is on cyberattack detection and impact analysis, the testbed itself serves as a validation platform for fault detection methods, and the authors deploy a pretrained Random Forest model on the edge IED for online inference. The testbed is validated on modified IEEE 13-node and 123-node feeders, demonstrating high detection accuracy and low inference latency. Similarly, the high-performance platform proposed in [39] uses an FPGA accelerator (Alveo) to achieve low-latency AI inference for fault detection, with the ability to process up to 7 data streams simultaneously at sampling rates as high as 800,000 per second. This work addresses the real-time deployment challenge directly, but the fault detection model itself is not described in detail, and the focus is on the hardware architecture rather than the protection algorithm. The adaptive protection architecture proposed in [16] introduces a centralized, software-defined protection paradigm that uses machine learning

for event prediction and explainable AI for transparency. While this study is conceptual and does not present experimental results, it contributes a vision for the future of protection that is fully aligned with the theme of this subsection. Finally, the sub-cycle fault response approach in [25] demonstrates that neural networks deployed on System-on-Chip platforms with dedicated Neural Processing Units can achieve fault detection in real-time, avoiding cloud-dependent latency. This work is notable for its explicit focus on the latency constraint, which is a critical requirement for any protection scheme that must operate within a few milliseconds of fault inception.

5. DISCUSSION

The synthesis of findings across the included studies reveals a research landscape that is simultaneously vibrant and fragmented, where significant progress has been made in individual dimensions of trustworthiness but where a cohesive, multi-faceted approach remains conspicuously absent. Taken together, the evidence suggests that the field has matured from an initial phase of proof-of-concept demonstrations to a more rigorous phase where researchers are actively grappling with the practical challenges of adversarial robustness, generalization, and real-time deployment. However, a critical pattern emerges across the literature: the overwhelming majority of studies optimize for one or

two of these dimensions while neglecting the third, and very few works attempt to address all three simultaneously. This siloed advancement represents perhaps the most significant barrier to the practical adoption of AI-based protection in inverter-dominated grids, as a protection scheme that is robust to adversarial attacks but cannot execute within sub-cycle timeframes, or one that generalizes well across topologies but is vulnerable to simple perturbations, is unlikely to gain the trust of system operators and regulatory bodies.

The implications of this finding are profound for both the research community and the power industry. From a theoretical perspective, the lack of integrated frameworks challenges the prevailing assumption that robustness, generalization, and real-time deployment can be achieved independently and then combined post-hoc. Our synthesis suggests that these dimensions are fundamentally interdependent and that optimizing for one may come at the expense of another. For example, adversarial training, which is the most widely adopted defense mechanism across the included studies [31] [33] [37], is known to degrade model accuracy on clean data and increase inference latency due to the need for larger model capacities or ensemble architectures. Similarly, model compression techniques that enable real-time deployment on edge devices, such as quantization or pruning, can reduce the model's representational capacity and thereby compromise its robustness to adversarial perturbations. These trade-offs are not merely theoretical concerns but have practical consequences: a protection relay that is robust to attacks but too slow to trip, or one that is fast but easily fooled by a simple perturbation, is not a viable solution for safety-critical power system applications. Therefore, future research must move beyond the current paradigm of isolated optimization and instead develop unified frameworks that explicitly model and manage these trade-offs, perhaps through multi-objective optimization or Pareto-optimal design spaces.

The practical implications of our findings extend to the design of evaluation protocols and benchmarking standards. A consistent observation across the included studies is the absence of standardized testbeds, adversarial threat models, and performance metrics that would enable fair and reproducible comparisons between different approaches. For instance, studies on adversarial robustness use a wide variety of attack algorithms—from the Fast Gradient Sign Method [26] to momentum iterative methods [43] and universal adversarial perturbations [37]—with different perturbation budgets and threat models, making it impossible to determine whether one defense is inherently superior to another or whether observed differences are artifacts of the evaluation setup. Similarly, studies on fault detection use different test systems (e.g., IEEE 14-bus, 34-bus, 39-bus, 123-bus, and custom microgrids), different fault types and

locations, and different performance metrics (e.g., accuracy, F1-score, detection time, false positive rate), further complicating cross-study comparisons. This lack of standardization is a critical barrier to progress because it prevents the research community from identifying the most promising approaches and building upon them. We therefore recommend the development of open-access, community-curated benchmarks that include standardized test systems, realistic fault and attack scenarios, and well-defined evaluation protocols for each of the three trustworthiness dimensions. Such benchmarks would not only facilitate fair comparisons but also accelerate the transition from academic research to industrial deployment by providing a common reference point for validation.

A further limitation that emerges from our synthesis is the heavy reliance on simulation-based validation, often with limited fault scenarios and idealized assumptions about measurement noise, communication delays, and system dynamics. While simulation is an essential first step for developing and testing AI-based protection methods, it cannot capture the full complexity of real-world power systems, including the stochastic nature of renewable generation, the non-ideal behavior of measurement devices, the latency and packet loss in communication networks, and the unpredictable actions of human operators. Only a handful of studies in our review incorporate hardware-in-the-loop (HIL) validation [28] [29] [35], and even fewer use actual field data or physical testbeds. This gap between simulation and reality is particularly concerning for the generalization dimension, as models that perform well on simulated data may fail catastrophically when deployed in the field due to distribution shifts that are not captured in the training data. The absence of real-world validation datasets is a systemic issue that stems from the proprietary nature of power system data and the difficulty of obtaining labeled fault events from operational grids. Addressing this limitation will require collaborative efforts between academia, industry, and regulatory bodies to create anonymized, publicly available datasets that reflect the diversity of real-world operating conditions. Until such datasets become available, the claims of generalization made in the literature must be interpreted with caution, and researchers should explicitly acknowledge the limitations of their validation approaches.

Beyond the review-process constraints discussed below, the synthesis itself has limitations that temper its conclusions. The central claim that few studies jointly address robustness, generalization and real-time deployment is currently supported narratively rather than by a reported count across the 32 studies; no formal quality or risk-of-bias appraisal was applied; the analytical taxonomy does not map cleanly onto the three pillars; and several trustworthiness dimensions raised in the reviewed literature - explainability, uncertainty

quantification, distribution-shift detection, safety assurance, cybersecurity, and hardware/communication constraints - are treated only in passing.

The methodological constraints of our review process also warrant careful consideration. Our search strategy, while comprehensive, was limited to English-language publications and a predefined set of databases, which may have introduced a language bias and excluded relevant studies published in other languages or in non-indexed venues. The inclusion of preprints from arXiv, while capturing the most recent research, also introduces the risk of including studies that have not undergone rigorous peer review, potentially affecting the quality of the evidence base. Furthermore, the subjective nature of the screening and eligibility assessment, despite the use of two independent reviewers, may have introduced some degree of selection bias, particularly in the interpretation of whether a study adequately addresses the three core themes of robustness, generalization, and real-time deployment. We attempted to mitigate this by documenting all decisions transparently and resolving disagreements through discussion, but we acknowledge that a different set of reviewers might have arrived at a different set of included studies. Finally, the relatively small number of included studies (32) limits the statistical power of any quantitative synthesis and means that our findings are primarily qualitative in nature. Future reviews with a broader scope or a more targeted focus on a specific dimension of trustworthiness may yield additional insights.

Based on the gaps and inconsistencies uncovered in our review, we identify several promising directions for future research. First, there is a pressing need for unified frameworks that embed formal robustness guarantees, out-of-distribution detection, and latency-constrained inference within a single protection architecture. Such frameworks should be designed from the ground up to address the interdependencies between these dimensions, rather than attempting to combine separately developed components. For example, a promising approach might involve the use of certified robust training methods, such as randomized smoothing [32] or Lipschitz-constrained networks, which provide provable guarantees on the model's robustness to bounded perturbations, combined with lightweight out-of-distribution detectors that can flag inputs that fall outside the training distribution and trigger a fallback to a conventional protection scheme. The inference latency of such a combined system could then be optimized through hardware-aware neural architecture search or knowledge distillation, ensuring that the entire pipeline meets the sub-cycle timing requirements of protection relays. Second, future research should explore the use of transfer learning and domain adaptation techniques to improve the generalization of AI models across different grid topologies, operating conditions, and fault

types. The included studies that address generalization [24] [36] are limited in scope and typically test only a small number of system configurations. There is a need for systematic investigations of how model performance degrades under distribution shifts and what types of data augmentation or regularization strategies are most effective at mitigating this degradation. Third, the understudied area of real-time deployment on resource-constrained edge devices deserves greater attention. While several studies have demonstrated the feasibility of FPGA-based [39] or SoC-based [25] implementations, these works often use relatively simple models and do not consider the additional computational burden of adversarial defense mechanisms or out-of-distribution detection. Future research should explore the design of specialized hardware accelerators for AI-based protection, perhaps using approximate computing or in-memory computing techniques, to achieve the required latency and energy efficiency without sacrificing model accuracy or robustness. Finally, we recommend that future studies adopt a more rigorous approach to threat modeling, distinguishing between realistic attack scenarios that are feasible in the power system domain and those that are theoretically possible but practically infeasible. The work of [37] and [38] provides a valuable template for this type of analysis, and we encourage other researchers to follow suit by explicitly stating the assumptions and constraints of their threat models and by evaluating their defenses under a range of attack budgets and attacker capabilities.

6. CONCLUSION

This systematic review synthesized the existing literature on trustworthy artificial intelligence for the protection of inverter-dominated power systems, focusing specifically on the three interdependent pillars of robustness, generalization, and real-time deployment. Our analysis of 32 included studies reveals a rapidly growing but fragmented research landscape in which significant progress has been made within individual dimensions, yet only a minority of studies address more than one trustworthiness pillar and a unified framework that simultaneously addresses all three remains conspicuously absent. The core finding of this review is that the siloed advancement of these trustworthiness dimensions represents a critical barrier to practical adoption, as protection schemes optimized for adversarial robustness often neglect real-time latency constraints, while those designed for fast inference may lack the generalization capability required for diverse grid topologies and operating conditions.

The practical and theoretical implications of our findings are substantial. From a practical standpoint, the absence of standardized benchmarks, adversarial threat models, and real-world validation datasets prevents fair cross-study comparisons and hinders the transition from

academic research to industrial deployment. Theoretically, our synthesis challenges the prevailing assumption that robustness, generalization, and real-time deployment can be achieved independently and then combined post-hoc, revealing instead that these dimensions are fundamentally interdependent and subject to inherent trade-offs. Future research must therefore prioritize the development of unified protection architectures that embed formal robustness guarantees, out-of-distribution detection, and latency-constrained inference within a single, co-designed framework. We further recommend the creation of open-access, community-curated testbeds and datasets to enable reproducible evaluation and accelerate progress toward trustworthy AI for the protection of modern, inverter-dominated power systems.

REFERENCES

- [1] S. Xu, Y. Xue, and L. Chang, "Review of power system support functions for inverter-based distributed energy resources-standards, control algorithms, and trends," *IEEE open journal of Power electronics*, 2021.
- [2] U. Muenz, S. Bhela, N. Xue, A. Banerjee, M. Reno, et al., "Protection of 100," researchgate.net, 2024.
- [3] D. Gadanayak, "Protection algorithms of microgrids with inverter interfaced distributed generation units-a review," *Electric Power Systems Research*, 2021.
- [4] G. Porawagamage, K. Dharmapala, J. Chaves, et al., "A review of machine learning applications in power system protection and emergency control: Opportunities, challenges, and future directions," *Frontiers in Smart Grids*, 2024.
- [5] P. Ilius, M. Almuhami, M. Javaid, et al., "A machine learning-based approach for fault detection in power systems," *Engineering, Technology & Applied Science Research*, 2023.
- [6] P. Rai, N. Londhe, and R. Raj, "Fault classification in power system distribution network integrated with distributed generators using CNN," *Electric Power Systems Research*, 2021.
- [7] S. Khaleefah, S. Mostafa, S. Gunasekaran, et al., "A deep learning-based fault detection and classification in smart electrical power transmission system," *JOIV : International Journal on Informatics Visualization*, 2024.
- [8] D. Kaur, S. Uslu, K. Rittichier, and A. Duresi, "Trustworthy artificial intelligence: A review," *ACM computing surveys*, 2022.
- [9] B. Manoj, M. Sadeghi, et al., "Adversarial attacks on deep learning based power allocation in a massive MIMO network," *ICC*, 2021.
- [10] C. Zhang, X. Costa-Perez, et al., "Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms," *IEEE/ACM Transactions on Networking*, 2022.
- [11] M. Kouraichi, M. Mansouri, M. Trabelsi, A. M'Halla, et al., "Deep learning for fault diagnosis in power transmission lines: Current trends, limitations, and future directions," *IEEE Access*, 2025.
- [12] I. Alhamrouni, N. A. Kahar, M. Salem, M. Swadi, et al., "A comprehensive review on the role of artificial intelligence in power system stability, control, and protection: Insights and future directions," *Applied Sciences*, 2024.
- [13] M. Page, J. McKenzie, P. Bossuyt, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
- [14] U. Asif, H. V. R. Modugu, H. AL-Hallaj, and M. B. Shadmand, "Decentralized AI-driven cyber-resilient control for real-time intrusion detection and mitigation in inverter-dominated grids," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, pp. 1–12, 2026.
- [15] M. Baker, A. Y. Fard, H. Althuwaini, and M. Shadmand, "Real-time AI-based anomaly detection and classification in power electronics dominated grids," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 4, pp. 549–559, 2023.
- [16] M. Bobček, J. Bencsik, Z. Varga, V. Szomosi, and Z. Čonka, "The future of electrical power grid protection: Adaptive, intelligent, and cyber-resilient architectures for dynamic systems," *EPJ Web of Conferences*, 2026.
- [17] D. Etingov, P. Zhang, Z. Tang, and Y. Zhou, "AI-enabled traveling wave protection for microgrids," *Electric Power Systems Research*, 2022.
- [18] A. Fiset, Z. Qu, and K. McDonald, "Control mode detection of inverter based resources in grid operations," *IECON 2025 - 51st Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–6, 2025.
- [19] K. Gokulraj and C. Venkatramanan, "Advanced machine learning-driven security and anomaly identification in inverter-based cyber-physical microgrids," *Electric Power Components and Systems*, 2024.
- [20] T. Hossen, F. Sadeque, F. Fateh, and B. Mirafzal, "Projection of safe operation for inverters using artificial intelligence-based stability criterion," *2023 IEEE Applied Power Electronics Conference and Exposition (APEC)*. pp. 3083–3088, 2023.
- [21] S. R. Kasimalla, K. Park, J. Hong, Y.-J. Kim, and H. Lee, "AI-enhanced inverter fault and anomaly detection system for distributed energy resources in microgrids," *ArXiv*, vol. abs/2411.08761, 2024.
- [22] A. Khan, O. Beg, Y.-F. Jin, and S. Ahmed, "An explainable intelligent framework for anomaly mitigation in cyber-physical inverter-based systems," *IEEE Access*, vol. 11, pp. 65382–65394, 2022.
- [23] K. A. Kharusi, A. E. Haffar, and M. Mesbah, "Fault detection and classification in transmission lines connected to inverter-based generators using machine learning," *Energies*, 2022.
- [24] K. A. Kharusi, A. E. Haffar, and M. Mesbah, "Adaptive machine-learning-based transmission line fault detection and classification connected to inverter-based generators," *Energies*, 2023.
- [25] A. Kovalenko, A. Evdakov, G. Filatova, et al., "Autonomous AI-driven grid protection: Sub-cycle fault response via NPU-optimized neural networks," in *Proceedings of the 23rd ACM conference on embedded networked sensor systems*, 2025.

- [26] D. Nicolas, K. Orozco, S. Mathew, Y. Wang, W. M. Elmannai, and G. C. Giakos, "Trustworthiness of deep learning under adversarial attacks in power systems," *Energies*, 2025.
- [27] A. Pan, Y. Lee, H. Zhang, Y. Chen, and Y. Shi, "Improving robustness of reinforcement learning for power system control with adversarial training," *ArXiv*, vol. abs/2110.08956, 2021.
- [28] U. Prasad, S. R. Mohanty, S. P. Singh, and A. Jagan, "Robust centralized protection scheme with AI-based fault diagnosis capabilities for graph-structured AC microgrids," *IEEE Transactions on Smart Grid*, vol. 16, pp. 1975–1992, 2025.
- [29] M. Rahman, A. Rahman, and W. Sun, "Edge-AI-enabled cyber-physical HIL testbed for real-time impact analysis and detection of cyber attacks in distribution grids with DERs," *IEEE Transactions on Industry Applications*, 2026.
- [30] F. Rayhan, "AI-powered condition monitoring for solar inverters using embedded edge devices," *Preprints*, 2025.
- [31] C. Ren and Y. Xu, "Robustness verification for machine-learning-based power system dynamic security assessment models under adversarial examples," *IEEE Transactions on Control of Network Systems*, vol. 9, pp. 1645–1654, 2022.
- [32] C. Ren and Y. Xu, "A universal defense strategy for data-driven power system stability assessment models under adversarial examples," *IEEE Internet of Things Journal*, vol. 10, pp. 7568–7576, 2023.
- [33] C. Ren, X. Du, Y. Xu, Q. Song, Y. Liu, et al., "Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial ...," *IEEE Transactions on Power Systems*, 2021.
- [34] J. Ruan et al., "On vulnerability of renewable energy forecasting: Adversarial learning attacks," *IEEE Transactions on Industrial Informatics*, vol. 20, pp. 3650–3663, 2024.
- [35] A. M. Saber et al., "An AI-based supervisory measurement integrity validation layer for cyber-resilient AC/DC protection in inverter-based microgrids," *IEEE Transactions on Instrumentation and Measurement*, vol. 75, pp. 2511113–2511113, 2026.
- [36] A. N. Sheta, A. A. Eladl, B. E. Sedhom, M. I. El-Afifi, P. Sanjeevikumar, and M. Zaki, "Artificial neural network-based enhanced distance protection sensitivity in microgrids," *Renewable energy focus*, vol. 54, p. 100710, 2025.
- [37] Q. Song, R. Tan, C. Ren, and Y. Xu, *Understanding credibility of adversarial examples against smart grid: A case study for voltage stability assessment*. 2021.
- [38] Q. Song et al., "On credibility of adversarial examples against learning-based grid voltage stability assessment," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, pp. 585–599, 2024.
- [39] L. Sun, L. Muguira, J. Jiménez, J. Lázaro, and W. Yong, "High performance platform to detect faults in the smart grid by artificial intelligence inference," *IEEE Transactions on Smart Grid*, vol. 15, pp. 504–512, 2024.
- [40] Md. A. Taher, M. Tariq, and A. I. Sarwat, "Enhancing security in islanded AC microgrid: Detecting and mitigating cyber attacks in secondary control through AI-based method," *IEEE Transactions on Industry Applications*, vol. 61, pp. 2124–2134, 2025.
- [41] Y. Wang and B. Pal, "Destabilizing attack and robust defense for inverter-based microgrids by adversarial deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 14, pp. 4839–4850, 2023.
- [42] A. Younesi, P. Siano, A. Moradpour, and A. Mehrizi-Sani, "Hybrid neuro-symbolic learning and reasoning for resilient load restoration in smart microgrids," *Renewable Energy*, vol. 256, p. 124401, 2025.
- [43] L. Zhang, C. Jiang, Z. Chai, and Y. He, "Adversarial attack and training for deep neural network based power quality disturbance classification," *Engineering Applications of Artificial Intelligence*, 2024.
- [44] X. Zhang, C. Jiang, M. Yu, X. Wen, J. Zhang, et al., "Adversarial black-box attack and defense for convolutional neural network-based power quality disturbance classification," *Engineering Applications of Artificial Intelligence*, 2025.
- [45] Y. Zhou, Z. Ding, Q. Wen, and Y. Wang, "Robust load forecasting towards adversarial attacks via bayesian learning," *IEEE Transactions on Power Systems*, vol. 38, pp. 1445–1459, 2023.