ASRC Procedia
Global Perspectives in Science and Scholarship
Florida, USA

1St GRI Conference 2025

Volume: *1;* ***Issue:*** *1*
Pages: *404–431*
Published: *29 April 2025*

open access

1st Global Research and Innovation Conference 2025,

April 20–24, 2025, Florida, USA

ADVERSARIAL DEFENCE MECHANISMS IN NEURAL NETWORKS FOR ICS FAULT TOLERANCE: A COMPARATIVE ANALYSIS

Md Tawfiqul Islam[1]

[1] Master of Engineering Management, Lamar University, Texas, USA
Email: mislam91@lamar.edu; tawfiq.ctgbd@gmail.com
Orcid ID: https://orcid.org/0009-0002-4857-732X

Doi: 10.63125/xrp7be57

Peer-review under responsibility of the organizing committee of GRIC, 2025

Abstract

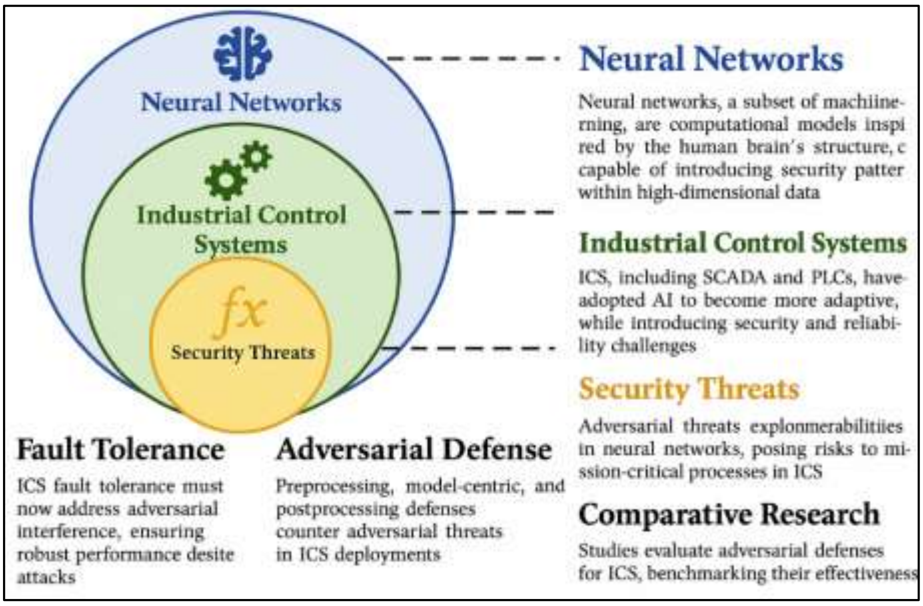
This systematic review explores the landscape of adversarial defense mechanisms in neural networks designed for fault-tolerant operations within Industrial Control Systems (ICS). With the increasing integration of artificial intelligence into critical infrastructure, ICS are now vulnerable to adversarial attacks that exploit the fragility of deep learning models, potentially leading to unsafe control actions and operational disruptions. The study systematically reviewed and synthesized findings from 126 peer-reviewed articles published between 2013 and 2024, covering model-centric, preprocessing, and postprocessing defense strategies, as well as their sector-specific implementations in smart grids, chemical processing plants, water treatment systems, and robotic manufacturing. The review was conducted following the PRISMA 2020 guidelines, ensuring a rigorous and transparent methodology. Key findings reveal that adversarial training remains the most widely used and effective model-centric approach, while signal filtering and demising methods serve as efficient preprocessing techniques for input sanitization. Post processing strategies, such as uncertainty estimation and confidence-based detection, are shown to enhance robustness when used in layered defense architectures. Furthermore, it highlights the emerging importance of integrating adversarial robustness into cyber security standards like NIST, ENISA, and IEC 62443. By analyzing both theoretical contributions and practical implementations, this study provides a comprehensive framework for understanding the current state of adversarial defenses in ICS environments. The findings emphasize the need for context-aware, scalable, and explainable defense mechanisms that can operate within the constraints of real-time control systems. This review contributes valuable insights for researchers, engineers, and policymakers working at the intersection of machine learning security and industrial automation.

Keywords
Adversarial Machine Learning, Neural Networks, Industrial Control Systems (ICS), Fault Tolerance, Cyber-Physical Security;

INTRODUCTION

Neural networks, a subset of machine learning, are computational models inspired by the human brain's structure, capable of identifying complex patterns within high-dimensional data (Khonina et al., 2024). These models have been widely adopted for their adaptability and predictive performance in numerous domains, including healthcare, finance, and more recently, industrial control systems (ICS). ICS are integrated hardware and software systems designed to monitor and control industrial operations such as manufacturing, power generation, chemical processing, and water treatment.

Figure 1: Adversarial AI Risks in ICS



Neural Networks

Neural networks, a subset of machiinerning, are computational models inspi red by the human brain's structure, c capable of introducing security patter within high-dimensional data

Industrial Control Systems

ICS, including SCADA and PLCs, have adopted AI to become more adaptive, while introducing security and reliability challenges

Security Threats

Adversarial threats explonmerabilitiies in neural networks, posing risks to mission-critical processes in ICS

Fault Tolerance

ICS fault tolerance must now address adversarial interference, ensuring robust performance desite attacks

Adversarial Defense

Preprocessing, model-centric, and postprocessing defenses counter adversarial threats in ICS deployments

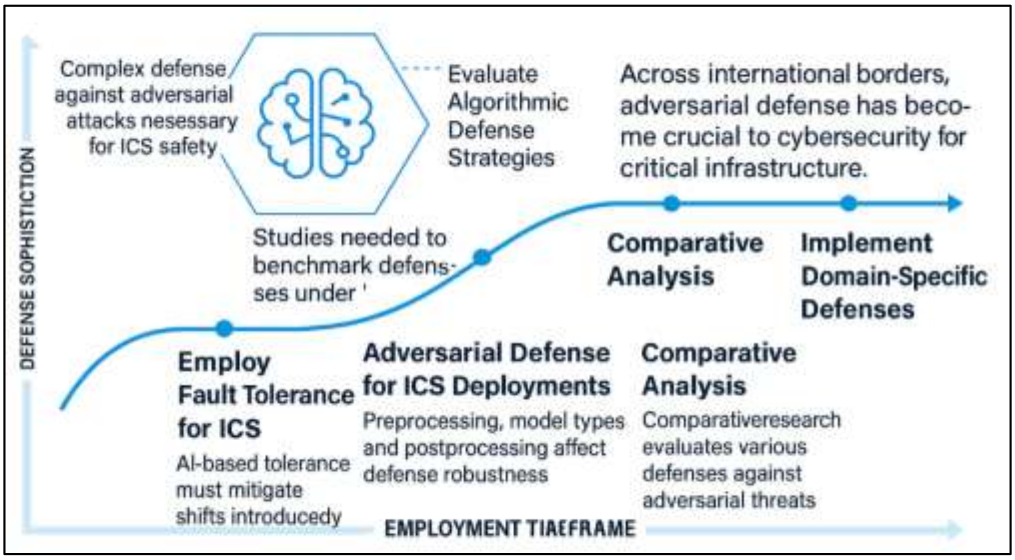
Comparative Research

Studies evaluate adversarial defenses for ICS, benchmarking their effectiveness

They comprise subsystems like Supervisory Control and Data Acquisition (SCADA), Programmable Logic Controllers (PLCs), and Distributed Control Systems (DCS), all of which are critical to the infrastructure of nations. With the adoption of artificial intelligence (AI), especially neural networks, ICS have become more responsive and adaptive, enhancing operational efficiency and fault diagnostics (Abiodun et al., 2019). However, the integration of neural networks into ICS introduces novel security and reliability challenges, particularly regarding adversarial threats. The deterministic nature of ICS operations contrasts sharply with the probabilistic behavior of deep learning models, thereby necessitating a rigorous examination of adversarial vulnerabilities and defense mechanisms (Parhi & Unnikrishnan, 2020). As a result, ICS environments have emerged as critical testbeds for evaluating neural network robustness under adversarial conditions. This duality between operational rigidity and learning-based adaptability underscores the need for enhanced fault tolerance frameworks grounded in adversarial defense mechanisms.

Adversarial threats refer to strategically crafted perturbations in input data that are imperceptible to humans but cause significant misclassifications in neural networks. These attacks exploit the nonlinear and high-dimensional decision boundaries of neural networks, enabling malicious actors to deceive systems with minimal input changes (Kumar & Rastogi, 2023). In ICS settings, such vulnerabilities could compromise mission-critical processes, including temperature control, voltage regulation, and process sequencing. Notably, adversarial examples have demonstrated high transferability, meaning an attack crafted on one model can often deceive other models with different architectures. This universality magnifies the threat landscape in ICS, where redundancy and model diversity are core design principles. Empirical studies have shown that attacks like Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner attacks can disrupt predictive maintenance systems, anomaly detectors, and failure classification models deployed within ICS (Mahmud et al., 2021). In particular, ICS-related deep learning deployments are susceptible due to the scarcity of labeled industrial data and the reliance on supervised learning, which lacks intrinsic robustness. The high cost of system downtime and the life-critical nature of ICS operations elevate the importance of understanding and defending against adversarial threats (Pinaya et al., 2020). Therefore, effective adversarial defense mechanisms are not merely academic interests but are vital for securing critical infrastructure globally.

Figure 2: Adversarial Defence in ICS Systems


Complex defense against adversarial attacks nesessary for ICS safety
Evaluate Algorithmic Defense Strategies
Across international borders, adversarial defense has become crucial to cybersecurity for critical infrastructure.
DEFENSE SOPHISTICTION
Studies needed to benchmark defenses under '
Comparative Analysis
Implement Domain-Specific Defenses
Employ Fault Tolerance for ICS
AI-based tolerance must mitigate shifts introducedy
Adversarial Defense for ICS Deployments
Preprocessing, model types and postprocessing affect defense robustness
Comparative Analysis
Comparativeresearch evaluates various defenses against adversarial threats
EMPLOYMENT TIAEFRAME

Fault tolerance in ICS refers to the system's ability to continue functioning correctly in the presence of internal failures or external disturbances. Traditionally, this has been achieved through redundancy, fail-safe mechanisms, and rule-based error detection systems. However, with the introduction of AI models like neural networks for process optimization, fault diagnosis, and real-time control, the nature of system vulnerabilities has evolved (Durstewitz et al., 2019). The reliance on data-driven models necessitates that fault tolerance extends beyond physical and sensor-level redundancies to include algorithmic resilience against adversarial interference. For example, a neural network used for detecting valve malfunctions may produce false positives or negatives under adversarial perturbation, potentially leading to catastrophic physical responses. The correlation between adversarial robustness and operational fault tolerance thus forms a new frontier in ICS safety engineering. Studies have demonstrated that adversarial attacks can mimic sensor anomalies, leading to incorrect fault localization and recovery procedures (Lan et al., 2018). Furthermore, ICS often operate in environments with limited computational resources, making it challenging to implement complex defense architectures like adversarial training or certified defenses (Lan et al., 2018). Therefore, aligning adversarial defense strategies with fault tolerance objectives requires a careful balance of accuracy, latency, interpretability, and system stability. This convergence invites comparative evaluations of defense mechanisms that are tailored to the unique operational profiles of ICS.

Adversarial defense strategies are broadly categorized into three groups: preprocessing defenses, model-centric defenses, and postprocessing defenses. Preprocessing methods sanitize input data before feeding it into the model, using techniques like denoising autoencoders, JPEG compression, or feature squeezing. These approaches aim to eliminate adversarial noise without altering the model structure. Model-centric defenses include adversarial training, gradient masking, and robust loss functions, which aim to enhance the model's internal robustness. Postprocessing defenses detect or correct the model's outputs after inference, using confidence thresholds, outlier detection, or model ensembles (Zhang et al., 2025). In ICS deployments, each defense type presents trade-offs. Preprocessing may introduce latency, model-centric defenses often require retraining, and postprocessing adds inference-time complexity. Moreover, certain defenses, such as adversarial training, are highly effective but computationally expensive, making them less viable in low-resource industrial nodes. Research comparing the efficacy of these defenses in ICS-specific tasks such as state estimation, anomaly detection, and predictive control has produced varying results (Kaul et al., 2021). Consequently, there is a pressing need for comparative studies that evaluate adversarial defenses across realistic ICS scenarios, balancing defense performance with operational feasibility.

Comparative analyses of adversarial defenses in ICS contexts have emerged as a critical subfield in cybersecurity research. Benchmarking methodologies typically involve assessing model accuracy, perturbation resistance, and detection capability under various attack models (Fariz & Basha, 2024). For instance, studies have shown that while adversarial training improves robustness under known attacks, it may generalize poorly to unseen attack strategies. Defensive distillation, another widely examined technique, reduces sensitivity to input perturbations but suffers from gradient obfuscation and degraded performance (Voulodimos et al., 2018). In ICS scenarios, evaluations often focus on real-time constraints and operational correctness rather than classification accuracy alone. The impact of input transformations on sensor signal integrity, or the delay induced by defense layers on feedback control loops, are unique to this domain. Comparative research across smart grid systems, chemical reactors, and automated manufacturing has highlighted that no single defense strategy is universally optimal. As such, hybrid defense models and context-aware defense orchestration are being explored to maximize resilience (Jaison et al., 2024). These comparative studies underscore the importance of application-specific benchmarking when selecting or designing defense mechanisms for ICS environments.

The international significance of adversarial defense in neural networks, particularly within ICS, has grown due to increased digitization and interconnectivity of critical infrastructure across borders. Events like the Stuxnet worm and attacks on Ukrainian power grids exemplify how vulnerabilities in ICS can have geopolitical implications (Nithya et al., 2023). Governments and regulatory bodies such as the National Institute of Standards and Technology (NIST), the European Union Agency for Cybersecurity (ENISA), and the International Electrotechnical Commission (IEC) have issued frameworks promoting cybersecurity-by-design, which now includes machine learning robustness as a key component. Moreover, adversarial machine learning is increasingly recognized in global cybersecurity directives, such as the U.S. Cybersecurity Executive Order 14028, which highlights the role of AI in both enhancing and threatening cyber resilience. ICS, as part of critical infrastructure sectors like energy, water, and transportation, face heightened scrutiny due to their public safety and economic impact (Choudhary et al., 2022). As many countries adopt Industry 4.0 and smart manufacturing agendas, the use of neural networks in ICS is expanding, increasing the attack surface. Adversarial robustness is thus not just a technical concern but a policy and international collaboration issue. Collaborative efforts such as the Industrial Internet Consortium (IIC) and the Global Forum on Cyber Expertise (GFCE) have begun integrating AI risk evaluation into ICS cybersecurity blueprints (Agbehadji et al., 2020). These developments underscore a shared recognition of the importance of safeguarding neural network-based control systems against adversarial threats to ensure global stability, economic continuity, and societal well-being.

The increasing reliance on AI models within ICS drives the urgency to rigorously evaluate and compare defense mechanisms tailored to this domain. Traditional ICS security protocols emphasize physical isolation, role-based access control, and network segmentation, which are insufficient against algorithmic attacks exploiting neural network vulnerabilities (Karniadakis et al., 2021). The heterogeneity of ICS applications—from nuclear facilities to autonomous manufacturing cells—necessitates domain-specific evaluations of defense efficacy under real-world constraints. For example, a defense that performs well in high-latency environments may be unsuitable for real-time control loops in robotic systems. Moreover, ICS often rely on embedded and legacy systems with constrained hardware capabilities, limiting the feasibility of resource-intensive defense strategies such as adversarial training. Comparative analysis offers a structured pathway to benchmark defenses on metrics such as robustness, latency, model interpretability, deployment complexity, and operational compatibility (Yaghoubi et al., 2024). This paper aims to bridge the gap between adversarial machine learning theory and practical ICS deployment by synthesizing comparative insights across prominent defense mechanisms. In doing so, it contributes to the body of knowledge that supports resilient, secure, and operationally viable machine learning applications within industrial control infrastructures globally.

LITERATURE REVIEW

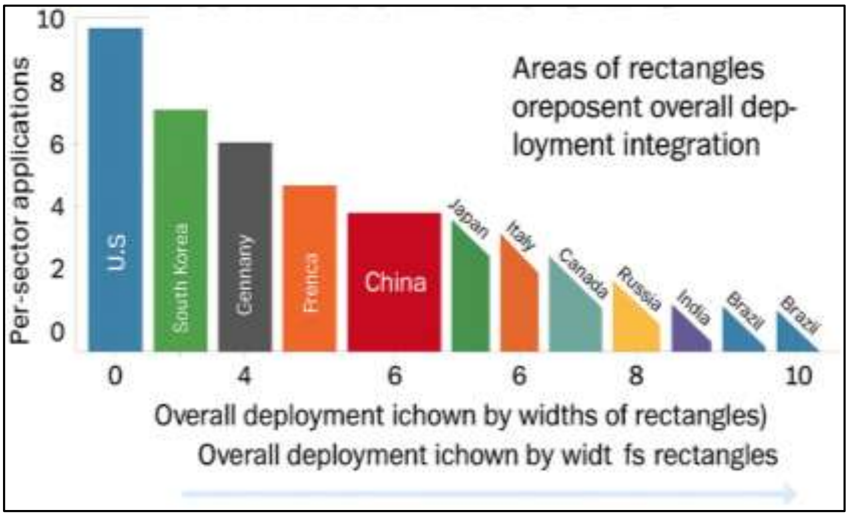
The literature surrounding adversarial defense mechanisms in neural networks for industrial control systems (ICS) fault tolerance reflects the convergence of two distinct yet increasingly interdependent fields: machine learning security and industrial automation reliability. With neural networks becoming integral to fault prediction, anomaly detection, and predictive maintenance in ICS, their vulnerability to adversarial attacks poses significant operational and safety challenges (Faheem & Al-Khasawneh, 2024). The scholarly discourse has expanded from purely theoretical explorations of adversarial machine learning to application-specific defense strategies, many of which are tailored to the unique constraints of ICS environments—such as real-time operation, limited computational overhead, and deterministic response requirements (Alshuhail et al., 2024). Research has highlighted that ICS-based neural networks are not only vulnerable to classical attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner (CW), but also to ICS-specific perturbation strategies that simulate faults, sensor drifts, or actuator delays. The literature offers a fragmented but rich spectrum of defensive approaches, ranging from adversarial training and robust optimization to sensor fusion and model ensemble techniques. While much progress has been made, comparative evaluation remains underdeveloped, particularly in contexts where deployment decisions must weigh trade-offs between robustness, latency, computational cost, and physical system stability (Yang et al., 2023). The present literature review aims to organize and synthesize these contributions through a systematic lens, focusing on five foundational areas: (1) the architecture and role of neural networks in ICS, (2) the landscape of adversarial threats, (3) classifications of defense strategies, (4) ICS-specific applications of these strategies, and (5) comparative assessment frameworks. This approach not only clarifies theoretical underpinnings but also provides actionable insights for developing resilient, AI-enhanced industrial systems.

Neural Networks in Industrial Control Systems

Industrial Control Systems (ICS) have undergone a significant transformation since their inception, progressing from isolated analog control mechanisms to highly digitized, cyber-physical infrastructures. Traditional ICS were built on rigid control hierarchies comprising elements such as Supervisory Control and Data Acquisition (SCADA), Programmable Logic Controllers (PLCs), and Distributed Control Systems (DCS), each designed to automate and monitor specific industrial processes (Celard et al., 2023). SCADA systems provided centralized visualization and command functions, while PLCs handled localized control logic in field devices (Zhou et al., 2018; Adepu & Mathur, 2016). The convergence of operational technology (OT) with information technology (IT) has opened these systems to enhanced analytics and decision-making capabilities through machine learning (ML) and neural networks. Neural networks have gradually been integrated into ICS to supplement traditional rule-based models with predictive capabilities, particularly in the areas of fault detection, maintenance scheduling, and process efficiency optimization. This evolution has been accelerated by the Industrial Internet of Things (IIoT), which has increased the volume and granularity of sensor data available for real-time analysis (Fan et al., 2021). Furthermore, advancements in edge computing and embedded AI have enabled the deployment of neural networks at the field level, offering localized intelligence while reducing communication latency. Despite these benefits, the transformation from traditional to intelligent ICS also introduces new fault domains, cybersecurity vulnerabilities, and operational complexities, especially as these networks become targets for adversarial attacks and other machine learning-specific threats (Santorsola & Lescai, 2023).

The deployment of neural networks in ICS has been most prominent in areas requiring real-time fault detection, condition-based maintenance, and complex system diagnostics. Predictive maintenance, in particular, has benefited from recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), which excel at modeling temporal dependencies in sensor data streams (Subrato, 2018). These architectures allow for early anomaly detection and component failure prediction, thereby reducing unplanned downtime and increasing asset utilization. Fault diagnostics in ICS environments have also seen the adoption of deep neural networks (DNNs) trained on multivariate sensor signals to classify specific failure modes across mechanical, electrical, and chemical subsystems. In process control, convolutional neural networks (CNNs) have been utilized to monitor visual data from cameras or to extract features from spatially structured process datasets (Ara et al., 2022). Hybrid models, such as CNN-LSTM architectures, have demonstrated improved performance in modeling non-linear system dynamics, especially in critical infrastructure such as nuclear plants and smart grids. Moreover, autoencoders and unsupervised deep learning models are increasingly used in anomaly detection tasks, where labeled data may be scarce or unavailable (Uddin et al., 2022). This functional diversity underscores the value of neural networks in ICS for tasks that exceed the capabilities of deterministic logic and PID control. However, despite these applications, challenges in explainability and validation of model outputs within safety-critical environments remain prevalent, requiring system operators to retain domain-based verification processes (Akter & Ahad, 2022).

Figure 3: Neural Network Deployment in ICS



Industrial environments impose stringent safety and real-time constraints that significantly influence the architecture and deployment of neural networks within ICS. In contrast to consumer-grade applications of AI, ICS must adhere to strict temporal deadlines, deterministic behavior, and system stability under variable loads and external disturbances (Rahaman, 2022). Real-time operation requires inference latency to be consistently low and predictable, a challenge given the computational intensity of deep neural networks. Safety-critical processes—such as turbine speed control, pressure relief actuation, and emergency shutdown protocols—cannot tolerate misclassifications or delays introduced by black-box AI models. Consequently, ICS applications often rely on quantized or pruned versions of neural networks to ensure real-time guarantees, though these modifications may reduce model robustness (Hasan et al., 2022). Moreover, the validation and certification of neural networks for use in regulated industrial sectors remain challenging due to the lack of formal verification techniques that can ensure safety properties under all possible input conditions. These concerns are further amplified when adversarial attacks are considered, as slight perturbations in sensor inputs can cause neural models to produce unsafe outputs, leading to potentially hazardous control actions. Embedded AI platforms with hardware acceleration (e.g., FPGA-based inference engines) have been proposed to meet both real-time and safety constraints while retaining sufficient model capacity (Hossen & Atiqur, 2022). Nonetheless, balancing computational efficiency, robustness, and reliability continues to be a core challenge in deploying neural networks within ICS contexts.

One of the most persistent limitations in deploying neural networks in ICS is the scarcity of high-quality labeled datasets for training, validation, and testing. Unlike applications in image recognition or natural language processing, where public datasets are abundant, industrial datasets are often proprietary, sparse, or poorly annotated due to privacy concerns, regulatory restrictions, and data sensitivity (Tawfiqul et al., 2022). This data bottleneck hinders the ability to train large-scale models that generalize well across operational modes, environmental conditions, and fault scenarios. Transfer learning has been explored as a solution, allowing pretrained models from related domains to be fine-tuned on small ICS datasets, but effectiveness varies depending on process complexity and feature overlap. Another approach has been synthetic data generation through simulation and digital twins, though such data may not fully capture real-world noise, drift, or stochastic behaviors. Domain adaptation and few-shot learning offer additional pathways for knowledge transfer but are still underexplored in ICS contexts (Sazzad & Islam, 2022). Moreover, models trained on clean, idealized data often fail to perform reliably under operational conditions involving sensor degradation, latency, or calibration shifts. The lack of standard benchmarks for ICS neural network performance further complicates reproducibility and comparative validation. Therefore, the deployment of neural networks in ICS must account for domain-specific variations and adapt to real-world constraints without overfitting to controlled or simulated conditions.

Adversarial Examples in ICS

Adversarial examples are carefully crafted inputs that cause neural networks to produce incorrect outputs while appearing indistinguishable from legitimate data to human observers (Koumakis, 2020). In industrial control systems (ICS), adversarial examples can be generated by introducing imperceptible perturbations into sensor readings or control signals to manipulate the behavior of machine learning models responsible for monitoring and decision-making. These perturbations exploit the high-dimensional decision boundaries in deep learning models and are optimized to maximize output deviations without violating detection thresholds. The mechanics of adversarial attacks rely on gradient-based optimization techniques that leverage access to the model's parameters or structure to compute minimal perturbations that significantly degrade model performance (Peng et al., 2021). In ICS settings, the deterministic and feedback-driven nature of control processes makes the injection of adversarial inputs especially dangerous, as incorrect model outputs may result in unsafe actuation, process instability, or regulatory non-compliance. Furthermore, adversarial examples pose systemic risks due to their transferability—the ability to deceive multiple models trained on the same task. In domains such as chemical processing, power generation, and water treatment, adversarial perturbations may simulate sensor drift, process faults, or operational anomalies, which neural networks misinterpret, triggering faulty alarms or inappropriate control actions (Akter & Razzak, 2022). As such, adversarial examples represent a novel but highly consequential threat vector that must be systematically analyzed in ICS contexts where human oversight is limited and autonomous decision-making is essential.

Adversarial attacks in ICS can be categorized based on the attacker's knowledge of the target model: white-box attacks assume full access to the model's architecture and parameters, while black-box attacks operate with limited or no internal knowledge. White-box attacks pose significant threats in development or testing environments where models are not yet deployed within secured operational boundaries (Adar & Md, 2023). In such scenarios, attackers can compute exact gradients and exploit vulnerability patterns specific to the model's internal structure, resulting in highly effective perturbations. FGSM, PGD, and CW attacks fall into this category and have been demonstrated to achieve high success rates against ICS fault prediction and anomaly detection models. Conversely, black-box attacks, which are more plausible in real-world ICS settings, leverage model outputs or surrogate models to approximate decision boundaries and craft transferable adversarial inputs (Hossen, 2023). Techniques such as Zeroth Order Optimization (ZOO), substitute model training, and evolutionary algorithms have been employed to execute successful black-box attacks on ICS-oriented deep learning models. Studies have shown that black-box attacks can exploit model uncertainty, overfitting, and sensor noise margins to deceive safety-critical applications like water quality monitoring and electric load balancing (Maniruzzaman et al., 2023). In both attack models, the challenge for ICS lies in detecting subtle deviations in high-dimensional time series inputs, which are often masked by natural fluctuations or calibration noise (Akter, 2023). Whether executed in white-box or black-box settings, adversarial attacks in ICS contexts undermine trust in neural network-based automation and highlight the urgent need for context-specific threat modeling.

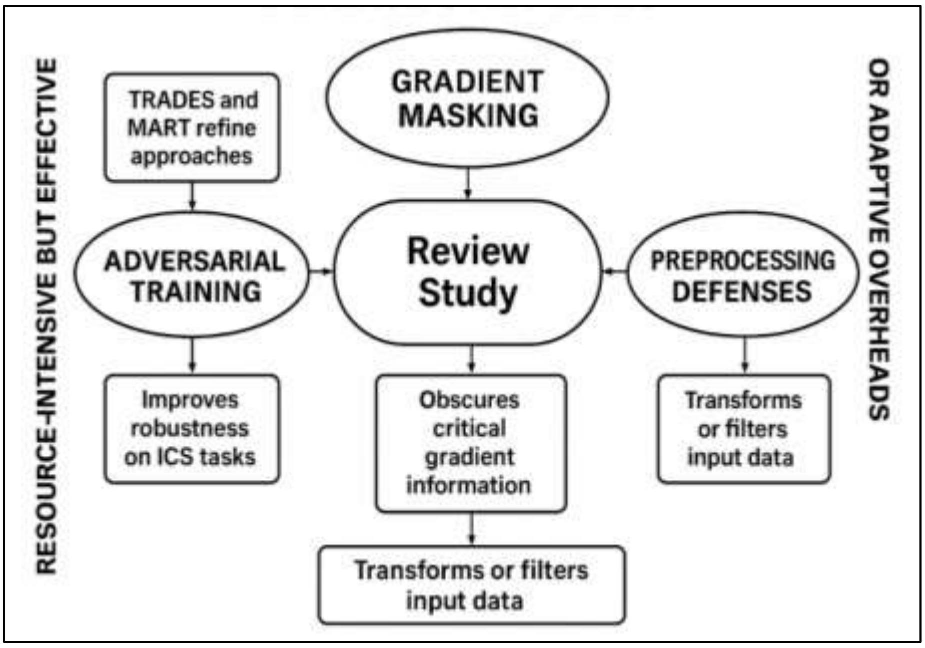
Several adversarial attack techniques have been adapted and tested against neural network models deployed in ICS environments. The Fast Gradient Sign Method (FGSM), introduced by Goodfellow et al. (2014), is a single-step gradient-based attack that adds perturbations in the direction of the input gradient to maximize classification loss. Despite its simplicity, FGSM has proven effective in altering fault classification in smart grid and manufacturing systems (Hossen et al., 2023). The Projected Gradient Descent (PGD) attack extends FGSM by using iterative steps and projection back onto a valid perturbation space, making it more powerful and suitable for targeted attacks in ICS anomaly detection. DeepFool and the Carlini-Wagner (CW) attacks are optimization-based techniques that minimize perturbation magnitude while maximizing model misclassification, and they have demonstrated strong efficacy against ICS systems monitoring continuous variables such as temperature and pressure (Shamima et al., 2023). The Jacobian-based Saliency Map Attack (JSMA) uses feature selection to craft perturbations that target specific output nodes, particularly effective in discrete-state classification models common in PLC programming. These attack methods have been empirically validated on ICS datasets such as SWaT (Secure Water Treatment) and BATADAL (Battle of the Attack Detection Algorithms), showing high attack success rates and minimal detectability. The versatility of these techniques across multiple model types—CNNs, LSTMs, and DNNs—demonstrates their applicability in both supervised learning and reinforcement learning ICS agents (Chen et al., 2020; Ashraf & Ara, 2023). As these methods continue to evolve, their strategic application to ICS highlights the intricate vulnerabilities present in industrial automation frameworks.

Adversarial Training and Robust Loss Functions

Adversarial training remains one of the most extensively studied and effective model-centric defenses against adversarial attacks. It involves augmenting the training dataset with adversarial examples to help the neural network learn more robust decision boundaries. This method enhances model resilience by optimizing for worst-case perturbations within a defined epsilon-ball around each training input. The TRADES (TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization) framework further refines adversarial training by balancing clean accuracy with adversarial robustness using a KL-divergence-based loss. Similarly, MART (Misclassification Aware adversarial Training) prioritizes learning from vulnerable data points and emphasizes misclassified samples to strengthen generalization under attack (Koay et al., 2023; Sanjai et al., 2023). In ICS environments, adversarial training has been applied to models analyzing real-time sensor data from systems such as Secure Water Treatment (SWaT), demonstrating improved fault classification accuracy under attack. However, adversarial training increases computational cost and training time—constraints that are critical in resource-limited ICS deployments. Moreover, fine-tuning adversarial training parameters such as epsilon and iteration count is complex in time-series data, especially when ICS inputs reflect multivariate dependencies (Akter et al., 2023). Despite these challenges, comparative studies confirm adversarial training's superior performance compared to non-robust baselines across smart grid fault prediction, industrial robot trajectory control, and chemical plant diagnostics. While its success depends on careful implementation and domain-specific calibration, adversarial training serves as a foundational layer in the defense strategy for ICS-integrated deep learning models.

Gradient masking is another model-centric defense strategy that attempts to make the gradient information less informative for attack algorithms. This technique hinders the attacker's ability to compute efficient perturbations by distorting or hiding gradients during backpropagation (Tonmoy & Arifur, 2023). Common methods include defensive distillation, which uses softened labels and temperature scaling to reduce sensitivity to adversarial changes, and activation clipping, which introduces saturation points into neural responses. Although initial studies found that gradient masking could lower the success rates of white-box attacks such as FGSM and PGD, later evaluations revealed that these defenses often give a false sense of security due to obfuscated gradients. In ICS applications, gradient masking has been explored in predictive maintenance and water flow anomaly detection systems but has shown inconsistent results across datasets and attack models (Zahir et al., 2023). Attacks like BPDA (Backward Pass Differentiable Approximation) and EOT (Expectation Over Transformation) can bypass gradient masking, rendering the defense ineffective in many real-world scenarios. Furthermore, ICS environments require not just robustness but interpretability and consistency—characteristics that gradient masking does not support effectively (Abdullah Al et al., 2024). Defensive strategies that obscure model internals can interfere with controller validation, regulatory compliance, and human-in-the-loop safety checks (Razzak et al., 2024). Consequently, the field has moved toward robust optimization and certified defenses that provide verifiable guarantees instead of relying on obscurity. While gradient masking played a role in early adversarial defense strategies, its limitations, especially in ICS where reliability is paramount, make it a suboptimal standalone solution (Jahan, 2024).

Figure 4: Model-Centric Adversarial Defense Strategies


RESOURCE-INTENSIVE BUT EFFECTIVE
TRADES and MART refine approaches
GRADIENT MASKING
OR ADAPTIVE OVERHEADS
ADVERSARIAL TRAINING
Review Study
PREPROCESSING DEFENSES
Improves robustness on ICS tasks
Obscures critical gradient information
Transforms or filters input data
Transforms or filters input data

Preprocessing-based defenses, particularly input transformation and purification techniques, represent a prominent class of adversarial defenses due to their architectural simplicity and model-agnostic nature. These methods aim to sanitize adversarial perturbations before they reach the classifier by altering the input data representation, thereby degrading the effectiveness of attacks like FGSM, PGD, and CW (Al-Hawawreh et al., 2024; Jahan & Imtiaz, 2024). JPEG compression is one of the earliest transformation-based defenses, relying on the elimination of high-frequency components that often harbor adversarial noise. Bit-depth reduction reduces the granularity of input signals, effectively clipping imperceptible variations and forcing the adversary to generate more detectable perturbations. Autoencoders—particularly denoising autoencoders—have been employed to learn the manifold of clean data and reconstruct input samples while removing extraneous noise. These models are especially useful in ICS environments, where time-series and multivariate sensor inputs can be efficiently compressed and reconstructed without significant loss of operational signal integrity (Imran et al., 2023; Akter & Shaiful, 2024). In empirical tests on datasets such as SWaT and WADI, autoencoder-based preprocessing has demonstrated notable reductions in attack success rates while preserving high classification accuracy. Nonetheless, preprocessing defenses may inadvertently distort benign data, especially when applied indiscriminately, leading to false positives or degraded system performance under normal operating conditions. Additionally, while such techniques are computationally efficient, their robustness can be bypassed by adaptive attackers who incorporate transformations into their gradient estimation process (Khan et al., 2022; Subrato & Md, 2024). These findings suggest that preprocessing methods are valuable first-line defenses but require precise calibration and integration with ICS-specific domain knowledge.

Industrial Control Systems (ICS) impose unique constraints on preprocessing defenses due to their reliance on physical signals, control feedback loops, and domain-specific signal characteristics. As a result, generic preprocessing techniques—such as image compression or generic noise filtering—must be tailored to the spectral and temporal properties of ICS data (Ammar et al., 2025). For example, frequency filtering has emerged as a viable strategy to eliminate adversarial perturbations injected into sensor streams without affecting critical system behavior. Low-pass and band-pass filters have been used to attenuate high-frequency adversarial noise in power grid data and flow control systems (Hossain et al., 2025; Akter et al., 2024). Kalman filtering and moving average smoothing techniques have also been adapted for ICS anomaly detection, offering a balance between real-time responsiveness and robustness against spurious fluctuations (Anika Jahan, 2025). Moreover, preprocessing defenses in

ICS often involve multi-step pipelines combining signal segmentation, normalization, statistical thresholding, and dimensionality reduction techniques such as PCA or ICA to enhance resilience. In predictive maintenance applications, noise-resilient encoding schemes—such as Fourier and wavelet transforms—have been leveraged to preserve fault-relevant frequency bands while discarding adversarial residue (Jahan et al., 2025). In process control systems, temporal resampling combined with sensor fusion across redundant channels has proven effective in filtering adversarial artifacts while preserving true anomaly signatures. However, these adaptations require extensive system knowledge and tuning, as over-filtering can suppress legitimate signals or introduce latency that disrupts time-sensitive control processes (Akter, 2025). Despite their complexity, ICS-specific preprocessing pipelines remain indispensable for translating general adversarial defense techniques into operationally viable mechanisms within safety-critical domains.

Preprocessing defenses are particularly attractive in industrial settings because they do not necessitate changes to the underlying neural network architecture. However, these defenses are increasingly challenged by adaptive attacks that incorporate the transformation process into their optimization loop, effectively learning how to bypass or nullify the defense (Rahman et al., 2025; Saheed & Arowolo, 2021). Such attacks utilize methods like Backward Pass Differentiable Approximation (BPDA), which approximates gradients through non-differentiable transformations, and Expectation Over Transformation (EOT), which averages gradients over randomized transformations to preserve attack transferability. In the ICS domain, where threat actors may gain access to deployment data via compromised endpoints or rogue sensors, adaptive attacks against preprocessing defenses have shown high success rates when tested on industrial datasets such as BATADAL and SWaT (Md et al., 2025). Moreover, studies have found that attackers can exploit knowledge of filtering parameters, compression ratios, or noise reduction techniques to craft perturbations that remain effective even after input transformation (Barbhaya et al., 2025; Islam & Debashish, 2025). For instance, adversarial examples have been generated to survive wavelet denoising and signal averaging filters while maintaining minimal perceptibility. Another limitation is that preprocessing may inadvertently amplify adversarial noise in the case of signal saturation or sensor drift, especially when coupled with real-time latency constraints. These limitations underscore the necessity of combining preprocessing methods with adaptive or reactive mechanisms such as adversarial training, ensemble defenses, or anomaly-based detection (MIslam & Ishtiaque, 2025; Nankya et al., 2023). While preprocessing provides a valuable layer of defense, it should not be relied upon in isolation, especially when dealing with sophisticated and resourceful adversaries capable of adapting to known countermeasures.

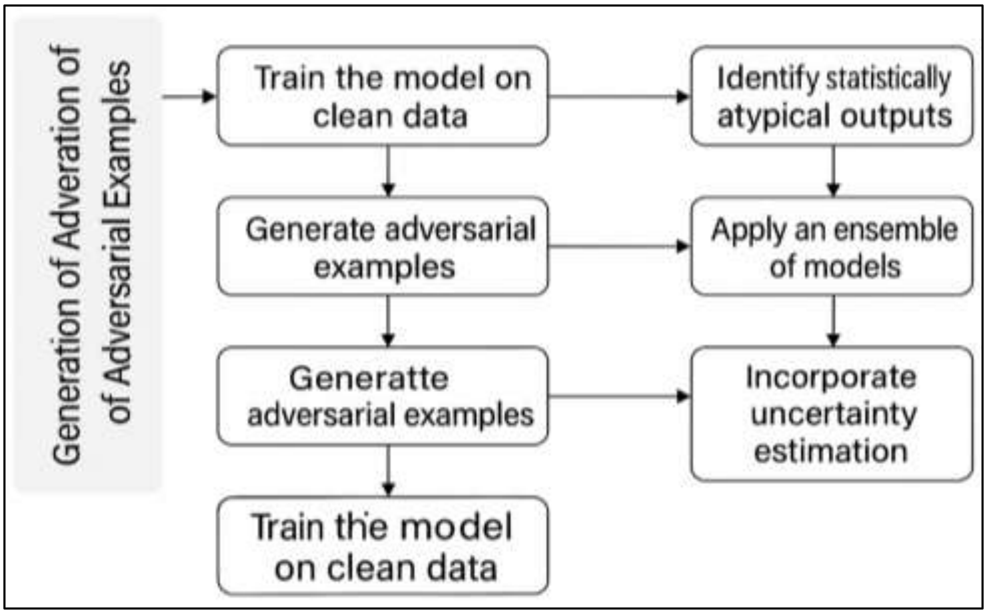
Preprocessing defenses must be implemented with care in ICS due to their potential effects on signal integrity, system feedback loops, and control latency. Unlike in vision or audio domains, where minor transformations may not impact overall performance, ICS rely on tightly coupled control logic that responds to minute variations in sensor readings. For instance, signal smoothing techniques such as moving averages or Gaussian filters may delay anomaly detection by several cycles, creating temporal mismatches in feedback control loops. Furthermore, lossy transformations like bit-depth reduction and quantization can remove essential dynamic features needed for fine-grained diagnostics, such as subtle frequency harmonics in rotating machinery or gas composition profiles in chemical plants. In applications involving proportional-integral-derivative (PID) controllers, even minor input distortions can lead to incorrect actuation signals, affecting valve timing, load balancing, or thermal regulation. Additionally, feedback control systems often rely on signal synchronization across multiple sensors; thus, any preprocessing method that disrupts temporal alignment or signal-phase coherence can destabilize the entire control routine (Hossen et al., 2025; Nankya et al., 2023). Real-time operation further constrains preprocessing by imposing strict latency budgets, especially in systems requiring millisecond-level reaction times. The trade-off between robustness and responsiveness becomes evident when overly aggressive noise filtering reduces the model's sensitivity to legitimate anomalies (Bulusu et al., 2020; Sanjai et al., 2025). Consequently, preprocessing defenses in ICS must be tightly integrated with system dynamics, validated through simulation and physical testbeds, and deployed only after confirming their minimal impact on operational accuracy and timing.

Model Output Analysis and Confidence Calibration

Post processing-based defenses utilize the model's output layer—particularly its logits or probability scores—to identify and reject adversarial inputs based on statistical irregularities. These techniques leverage the observation that adversarial examples often induce low-confidence or overconfident predictions that deviate from the normal behavior of clean data (Shaiful & Akter, 2025; Zhang et al., 2023). One common strategy is the use of Mahalanobis distance in the output feature space to identify samples that fall outside the distribution of known class representations. Another approach is statistical modeling of softmax scores using Gaussian mixture models or kernel density estimators to isolate atypical outputs. These methods have been adapted in ICS for applications such as anomaly detection in water treatment plants and fault localization in electric grids, where real-time decisions are driven by model confidence levels. Probabilistic thresholds have been implemented to flag uncertain or suspicious predictions, which are then subjected to secondary verification by human operators or fallback logic controllers (Ortiz-Jiménez et al., 2021; Akter, 2025). Outlier detection on logits has shown promise in detecting FGSM, PGD, and CW attacks on time-series and multivariate ICS models, particularly when trained under controlled input distributions. However, these methods are sensitive to data drift and may misclassify rare but legitimate events as adversarial. Moreover, tuning detection thresholds to minimize false positives while maintaining sensitivity remains a challenge, especially in dynamic industrial environments. Despite these limitations, logit-based outlier detection is a valuable non-invasive postprocessing layer for ICS models, providing an early warning signal for potential adversarial interference without altering core network architecture (Talpini et al., 2024; Zahir et al., 2025).

Auxiliary classifiers and ensemble architectures are widely used postprocessing techniques designed to improve adversarial robustness by introducing redundancy and decision diversity in prediction systems. The use of auxiliary classifiers involves appending secondary or parallel networks that verify the primary model's output or classify intermediate representations, enhancing adversarial detection through consensus voting or output correlation (Wang et al., 2020; Zahir et al., 2025). In ICS contexts, these classifiers have been applied in applications such as pipeline monitoring, voltage stability analysis, and chemical plant diagnostics, where confirmation from multiple networks helps prevent unsafe decisions triggered by perturbed inputs. Ensemble methods, including bagging, boosting, and random subspace learning, aggregate the predictions of multiple differently trained models to reduce susceptibility to single-point attacks.

Figure 5: Adversarial Example Defence Workflow


Generation of Adveration of of Adversarial Examples
Train the model on clean data
Identify statistically atypical outputs
Generate adversarial examples
Apply an ensemble of models
Generatte adversarial examples
Incorporate uncertainty estimation
Train the model on clean data

Diverse ensembles using varied initializations, architectures, or input transformations are more resilient against transfer-based adversarial attacks, which often fail to mislead all models simultaneously. For ICS deployments, hybrid ensembles comprising shallow decision trees, neural networks, and statistical models offer complementary strengths, enhancing detection accuracy under adversarial conditions. Furthermore, ensemble confidence metrics, such as vote entropy or disagreement rate, have been used to trigger defense routines or signal operator intervention. However, ensemble-based methods increase computational overhead and latency, especially when deployed on embedded ICS hardware or in systems with strict real-time constraints. Their effectiveness also depends on maintaining model diversity, which may diminish during adversarial training or parameter sharing (Mao et al., 2022). Nonetheless, ensembles remain an effective layer of post hoc defense, offering robustness through architectural multiplicity and prediction redundancy.

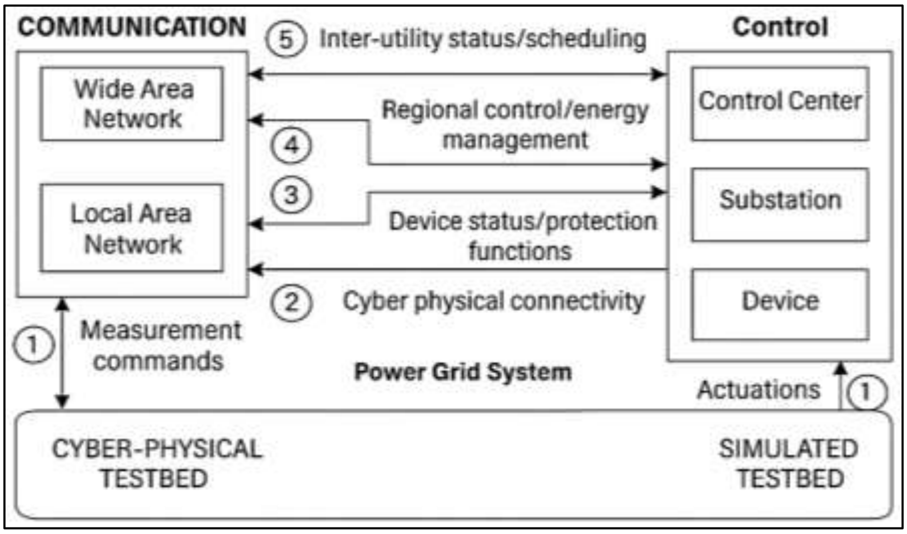
Quantifying uncertainty in neural network outputs is critical for detecting adversarial inputs and improving decision reliability, particularly in ICS, where safety and control integrity are paramount. Bayesian neural networks (BNNs) provide a probabilistic framework that models uncertainty by treating weights as distributions rather than fixed values, allowing the network to express confidence over predictions. Methods such as Monte Carlo (MC) dropout, variational inference, and deep ensembles enable the estimation of epistemic uncertainty, which increases when the model encounters out-of-distribution or adversarial inputs. In ICS settings, uncertainty-aware models have been used to monitor water treatment quality, detect equipment degradation, and forecast power demand under adversarial noise. For example, Bayesian LSTM models trained on time-series sensor data in smart grids were able to flag anomalous predictions during CW and PGD attacks with significantly higher uncertainty than during normal operations. Predictive entropy, mutual information, and confidence intervals from posterior distributions serve as robust indicators for adversarial detection, reducing reliance on hard classification thresholds. Additionally, techniques such as evidential deep learning and temperature scaling have been employed to calibrate output probabilities, aligning model confidence with empirical error rates. However, Bayesian inference introduces additional training and inference overhead and may be computationally intensive for real-time ICS applications (Kivimäki et al., 2023). These methods also require careful tuning of prior distributions and sampling strategies to ensure accuracy under uncertainty. Still, uncertainty-aware postprocessing techniques offer interpretable and statistically grounded defenses, enhancing operator trust and enabling proactive responses to adversarial threats in industrial environments.

Sectoral Case Studies in Power, Chemical, and Manufacturing Systems

Smart grids and power distribution systems have emerged as primary testbeds for implementing and evaluating adversarial defense strategies in ICS. These systems rely heavily on machine learning models for demand forecasting, fault prediction, load balancing, and anomaly detection (Lambert et al., 2024). Case studies show that neural networks used for voltage regulation and grid-state estimation are vulnerable to gradient-based attacks like FGSM and PGD, which can induce unstable oscillations or load shedding. In response, adversarial training and ensemble models have been deployed to enhance the resilience of load forecasting algorithms, improving their robustness to perturbations without compromising forecast accuracy. Additionally, preprocessing filters and Bayesian neural networks have been integrated into substation automation systems to detect and reject anomalous signals before they reach relay control modules. Several utility companies have also utilized digital twins to simulate cyber-physical attacks and evaluate the effectiveness of defense mechanisms in a risk-free environment. These simulations have revealed that coordinated attacks on geographically distributed grid segments can be detected through multi-node correlation and decentralized inference. However, the latency introduced by postprocessing and ensemble models remains a challenge for time-sensitive protection schemes such as overcurrent and under-voltage relays.

Chemical plants and continuous process control systems present distinct challenges for adversarial defense due to their reliance on nonlinear dynamics, feedback loops, and hazardous material management. Neural networks are increasingly used in these sectors for predictive maintenance, reaction monitoring, and fault diagnosis, often based on time-series sensor data and system state estimations (Zhang et al., 2025). Case studies from facilities using SCADA and DCS architectures have shown that attacks targeting temperature sensors, flow meters, or pH sensors can trigger inappropriate actuation—such as incorrect valve positioning or dosage adjustments—leading to cascading effects. In response, robust optimization and adversarial training methods have been applied to LSTM and GRU models trained on multivariate datasets, significantly improving resilience against perturbations (Wu et al., 2025). Researchers have also implemented autoencoder-based anomaly detection pipelines and ensemble classifiers in digital twin environments of water purification and chemical mixing plants, yielding successful detection rates for stealthy attacks. These defenses are often evaluated using real-world datasets such as the SWaT and WADI testbeds, which simulate real-time ICS operations and enable reproducible evaluation of adversarial robustness. However, the interpretability of deep models and their outputs remains a limitation in regulatory environments, where operational traceability is critical (Chen et al., 2023). Moreover, excessive model complexity and preprocessing latency may interfere with batch control systems and feedback timing, necessitating a balance between model depth and real-time constraints. These case studies underscore the need for scalable and interpretable defenses tailored to the nonlinear and hazardous nature of chemical ICS domains.

Figure 6: Adversarial Defense in ICS Domains

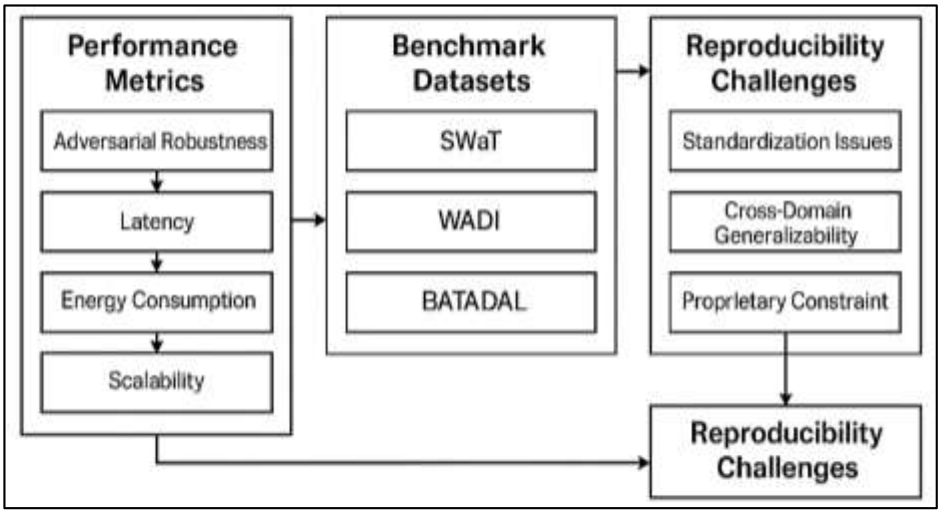


Robotic and smart manufacturing systems, characterized by high-speed automation and precision control, offer compelling environments for the application and testing of adversarial defense mechanisms in ICS. These environments leverage neural networks for vision-based quality inspection, predictive maintenance, trajectory planning, and dynamic resource allocation (Xu et al., 2024). Case studies reveal that adversarial examples targeting sensor fusion inputs—such as lidar, camera, and encoder signals—can result in trajectory deviation, production delays, or mechanical collisions. In one notable implementation, ensemble learning combined with adversarial training was used to defend a robotic arm's inverse kinematics controller against perturbations introduced into joint-angle sensors, significantly improving the safety margin. Bayesian neural networks and dropout-based uncertainty estimation were also employed in robotic welders and pick-and-place robots to signal low-confidence predictions and trigger human override mechanisms. Smart manufacturing testbeds such as the German Industrie 4.0 labs and the REMOTE-Lab in Singapore have been instrumental in simulating realistic adversarial attack scenarios on assembly lines and validating layered defense strategies (Lu et al., 2024). However, integrating defense systems into manufacturing execution systems (MES) and PLCs presents engineering challenges related to latency, resource overhead, and software compatibility. Furthermore, the need for continuous uptime in production facilities imposes strict thresholds on permissible false positives in anomaly detectors (Jang & Kim, 2023). Despite these constraints, sector-specific deployments in manufacturing have proven that hybrid defenses—comprising model-centric, preprocessing, and postprocessing techniques—can meaningfully improve resilience against adversarial disruptions in high-throughput industrial environments (Yang et al., 2024).

Frameworks and Benchmarks

Evaluating adversarial defense mechanisms in Industrial Control Systems (ICS) requires a multi-dimensional set of performance metrics that extends beyond traditional accuracy assessments. In addition to adversarial robustness—the ability of a model to maintain classification performance under perturbed inputs—metrics such as latency, energy consumption, interpretability, and scalability are critical in determining deployment feasibility (Haque et al., 2018). Latency is especially vital in ICS, where feedback loop constraints demand sub-second or even millisecond-level inference to maintain system stability. Energy usage, while often overlooked, plays a significant role in embedded or edge computing environments where deep learning models run on constrained hardware. For instance, adversarial training and ensemble defenses often trade computational efficiency for robustness, a compromise that can hinder real-time performance in distributed control systems. Scalability must also be assessed, as many ICS involve thousands of sensors and actuators spread across large geographical or hierarchical control networks. Interpretability metrics are particularly relevant in regulated environments like power grids and chemical plants, where decisions influenced by neural networks must be traceable and verifiable. F1-score, AUROC, adversarial detection rate, and perturbation magnitude tolerance are frequently used to compare model performance under varying levels of adversarial stress. These evaluation dimensions collectively determine a defense method's operational viability and help industrial stakeholders select the most appropriate defense configuration for their specific application context.

Figure 7: Evaluating ICS Adversarial Defence Metrics



Benchmarking adversarial defenses in ICS necessitates high-fidelity datasets and evaluation frameworks that mirror real-world industrial conditions. Among the most widely used are the Secure Water Treatment (SWaT), Water Distribution (WADI), and BATADAL datasets, each capturing multi-sensor ICS behavior under both benign and adversarial scenarios (Dutta & Granjal, 2020). The SWaT dataset, for example, simulates a physical water treatment plant controlled via programmable logic controllers (PLCs), offering rich time-series data across six stages of water purification. BATADAL, developed during the "Battle of the Attack Detection Algorithms" challenge, contains simulated cyberattacks against a water network, designed for testing detection accuracy and false alarm rates under adversarial stress. These datasets have been instrumental in comparative studies involving adversarial training, input preprocessing, postprocessing, and ensemble-based defenses. Evaluation frameworks typically incorporate adversarial threat modeling, including white-box and black-box scenarios, iterative and single-step attack variants (e.g., FGSM, PGD, CW), and statistical robustness indicators such as attack success rate and model degradation curves (Nankya et al., 2023). In smart grid applications, synthetic datasets generated through simulation environments such as GridLAB-D or Opal-RT have also been used to evaluate voltage regulation and frequency stabilization models under attack. Despite their utility, existing benchmarks often lack diversity in terms of ICS sectors, signal modalities, and long-term operational dynamics (Amir et al., 2018). As a result, there is a growing need for sector-specific benchmarking tools that reflect the operational, physical, and cyber-physical intricacies of ICS across different industries.

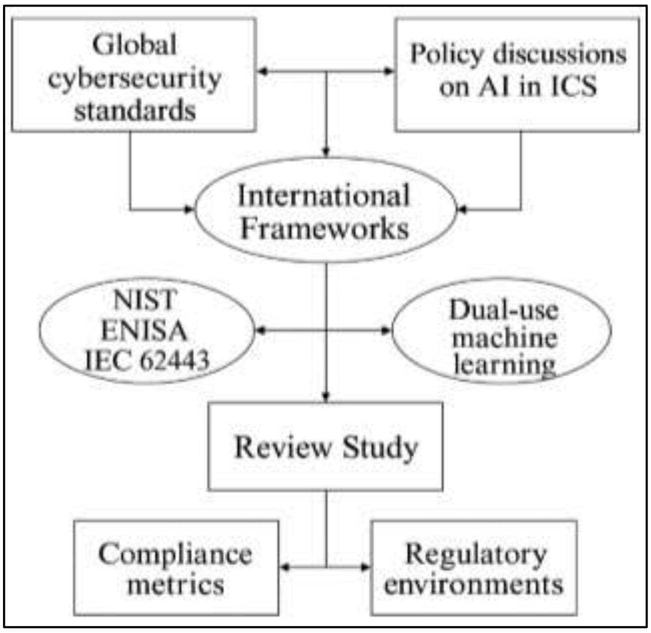
Hybrid and layered defense strategies have gained traction as a comprehensive response to adversarial threats in ICS, combining multiple defense mechanisms to increase robustness across different attack vectors. These architectures typically integrate model-centric defenses (e.g., adversarial training, robust loss functions), preprocessing techniques (e.g., denoising filters, autoencoders), and postprocessing modules (e.g., confidence-based detectors, Bayesian inference) to address the limitations of any single method (Trippel et al., 2020). For instance, one study in a power substation control system used a combination of PGD-trained neural networks and Mahalanobis distance-based output monitoring to detect adversarially induced faults with over 90% success rate, while maintaining real-time performance. Similarly, water treatment systems have successfully adopted hybrid pipelines where input signal smoothing, adversarial training, and uncertainty calibration operate in parallel to improve anomaly classification under gradient-based and transfer attacks. In manufacturing domains, ensembles of CNNs and LSTMs have been combined with dropout-based uncertainty estimators and activation clipping to prevent physical disruption caused by adversarial trajectory commands. These layered architectures allow for redundancy, modularity, and defense-in-depth, which are essential in safety-critical environments that cannot rely on a single point of failure. However, combining multiple defense layers often increases system complexity and integration overhead, potentially impacting scalability and maintainability (Varghese et al., 2022). Moreover, effectiveness under adversarially adaptive conditions—where attackers target the defense ensemble itself—remains an open area of research (Ali et al., 2022). Despite these challenges, hybrid strategies provide a practical pathway for operationalizing robustness in diverse ICS environments.

One of the critical challenges in adversarial defense research for ICS is ensuring reproducibility and generalizability of evaluation results across different sectors and operational contexts. Reproducibility is often hindered by the lack of standardized attack settings, differences in model architectures, and insufficient disclosure of hyperparameters or preprocessing routines (Magán-Carrión et al., 2020). Studies frequently use different epsilon values, attack iterations, or data preprocessing pipelines, making it difficult to compare the effectiveness of defenses on a consistent basis. Additionally, cross-domain generalizability remains limited; a defense mechanism effective in smart grids may perform poorly in batch chemical processes due to differing control logic, sensor behavior, and timing characteristics (Koay et al., 2023). For instance, PGD-trained classifiers may withstand perturbations in flow sensor data but fail when applied to vibration data in manufacturing systems. Digital twins and real-time simulators offer partial solutions by enabling the controlled replication of attacks and responses in different industrial settings, but these environments often lack the complexity of live ICS systems. Furthermore, the proprietary nature of many industrial platforms restricts open access to codebases, sensor logs, and control logic needed for rigorous evaluation. To address these issues, researchers have advocated for unified benchmarking protocols, sector-specific defense registries, and collaborative platforms for defense validation across ICS verticals (Wang et al., 2021). Until these initiatives mature, the field must contend with variability in methodologies and the challenges of translating academic insights into robust, cross-sector operational standards.

Cybersecurity Standards and Policy

Global cybersecurity standards such as NIST (National Institute of Standards and Technology), ENISA (European Union Agency for Cybersecurity), and the IEC 62443 series have established essential frameworks for managing risk in Industrial Control Systems (ICS), including emerging threats posed by adversarial machine learning. NIST's Special Publication 800-82 provides a detailed guide for securing ICS by outlining protocols for access control, incident response, and system hardening (Kumari et al., 2023). ENISA’s guidelines for cybersecurity in Industry 4.0 emphasize resilience against cyber-physical attacks, incorporating AI risk modeling and situational awareness into industrial security architecture. The IEC 62443 standards—particularly parts 3-3 and 4-1—prescribe a lifecycle-oriented approach to securing automation and control systems, introducing requirements for component security, secure communication, and system integrity. While these standards provide general principles for ICS protection, their treatment of AI-specific threats, such as adversarial examples, remains limited. Recent amendments and annexes have begun recognizing AI-based operational technology (OT) elements as new attack surfaces, prompting interest in AI-tailored compliance metrics (Wali et al., 2025). For example, NIST's AI Risk Management Framework (RMF) now encourages developers to assess adversarial robustness and incorporate uncertainty-aware learning in AI-enabled safety-critical systems. Similarly, ENISA's 2022 threat landscape report explicitly identifies adversarial machine learning as a rising concern for ICS environments. However, the implementation of these standards varies widely across countries and sectors, with many small and medium-sized enterprises (SMEs) lacking the resources to adopt advanced AI security practices. As adversarial threats evolve, updating global standards to include concrete guidelines for neural network robustness will be critical in unifying the defense posture of ICS worldwide.

Figure 8: Standardizing Adversarial Defense Policies

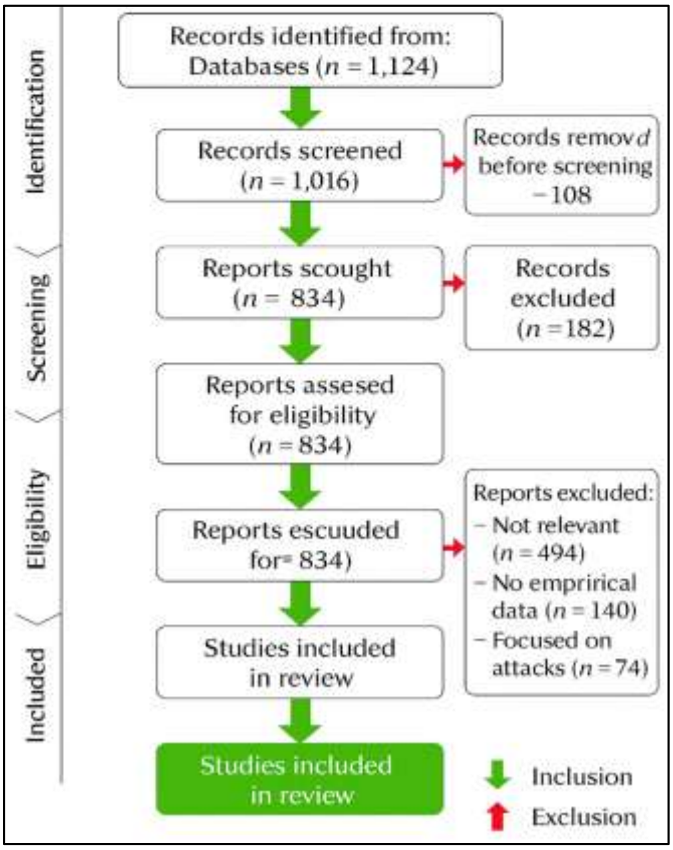


Policy discussions surrounding AI integration in ICS have increasingly acknowledged the dual-use nature of machine learning—where the same algorithms that enable efficiency gains also introduce novel vulnerabilities. Governments and regulatory bodies are beginning to formalize discourse around the risks of adversarial machine learning in cyber-physical infrastructure, recognizing that AI-powered control systems lack the deterministic transparency required for safety-critical environments (Le Jeune et al., 2021). The U.S. Cybersecurity Executive Order 14028 emphasizes the need for secure software development, including AI assurance mechanisms that account for manipulation through adversarial inputs. Internationally, the Global Forum on Cyber Expertise (GFCE) and the Industrial Internet Consortium (IIC) have promoted guidelines to evaluate AI-based controls in operational contexts, highlighting the urgency of building adversarial-aware resilience frameworks. Policy narratives are evolving to frame adversarial robustness not merely as a technical challenge but as a matter of national and economic security, particularly for sectors like energy, water, transportation, and defense. Several initiatives in the EU, such as the AI Act and the Cyber Resilience Act, propose mandatory AI risk evaluations and lifecycle monitoring to mitigate vulnerabilities across connected infrastructure . These efforts are complemented by the OECD's work on AI principles, which encourages transparency, reliability, and robustness as essential AI design objectives. Despite growing alignment between AI policy and industrial cybersecurity, gaps remain in translating abstract policy statements into actionable compliance checklists and control requirements for AI-based ICS (Sarker, 2024b). Thus, the ongoing policy discourse plays a pivotal role in shaping regulatory environments that demand and enforce adversarial defense practices within critical infrastructure ecosystems.

METHOD

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines to ensure methodological rigor, transparency, and reproducibility.

Figure 9: Methodology of This Study


Identification
Screening
Eligibility
Included
Records identified from: Databases (n = 1,124)
Records screened (n = 1,016)
Records removd before screening −108
Reports scought (n = 834)
Records excluded (n =182)
Reports assesed for eligibility (n =834)
Reports escuuded for= 834)
Reports excluded:
– Not relevant (n = 494)
– No emprirical data (n = 140)
– Focused on attacks (n =74)
Studies included in review
Studies included in review
Inclusion
Exclusion

PRISMA offers a structured and standardized approach to reporting systematic reviews and meta-analyses and is widely recognized as a benchmark for high-quality evidence synthesis across disciplines, including cybersecurity, artificial intelligence, and industrial systems. The application of the PRISMA framework enabled a systematic selection, appraisal, and synthesis of literature pertaining to adversarial defense mechanisms in neural networks deployed within Industrial Control Systems (ICS), with a specific focus on fault tolerance and operational resilience. The first step in the PRISMA-compliant review process was the formulation of a clear and focused research question using the PICO framework (Population, Intervention, Comparison, Outcome). The population of interest was ICS environments, the intervention comprised various adversarial defense mechanisms, the comparison involved different types of machine learning attacks (e.g., FGSM, PGD, CW), and the outcomes were measured in terms of robustness, latency, accuracy, and operational feasibility. This structured approach guided the entire review process and helped maintain alignment between the inclusion criteria and the research objectives. A comprehensive and replicable search strategy was employed across multiple scholarly databases including IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and ScienceDirect. The search terms were developed through Boolean logic and included combinations such as ("adversarial attacks" AND "neural networks" AND "ICS"), ("fault tolerance" AND "deep learning" AND "cyber-physical systems"), and ("machine learning defense" AND "industrial control systems"). The search was limited to peer-reviewed journal articles and conference proceedings published between 2013 and 2024 to ensure the inclusion of both foundational and contemporary works. The initial search yielded 1,124 records.

The screening and eligibility assessment followed a two-stage process: title/abstract screening followed by full-text review. Duplicates were removed using reference management software (e.g., EndNote and Zotero), reducing the corpus to 834 unique studies. Two independent reviewers screened the abstracts and titles to exclude studies that were not directly related to adversarial defenses or ICS contexts. Discrepancies between reviewers were resolved through discussion or by consulting a third reviewer. In the full-text assessment phase, studies were included if they met all the following criteria: (1) addressed adversarial machine learning in neural networks, (2) involved application or experimentation in ICS or related cyber-physical systems, (3) provided empirical evidence or theoretical evaluation of defense mechanisms, and (4) were published in English. Following this process, a total of 126 studies met the inclusion criteria and were retained for qualitative synthesis. A PRISMA flow diagram was constructed to visually represent the article selection process, documenting the number of studies identified, screened, included, and excluded at each stage, along with reasons for exclusion. Common reasons for exclusion during full-text screening included lack of relevance to ICS, absence of empirical data, and papers focusing solely on offensive attack modeling without defense considerations. The data extraction and coding process was conducted systematically using a structured data extraction form developed in Microsoft Excel. The extracted variables included publication year, study design, dataset type (e.g., SWaT, BATADAL, custom datasets), defense methodology (e.g., adversarial training, preprocessing, postprocessing), attack model type (e.g., white-box, black-box), evaluation metrics (accuracy, robustness, latency), and sectoral focus (e.g., smart grid, chemical plant, robotic manufacturing). This structured coding allowed for categorical aggregation and comparison across studies.

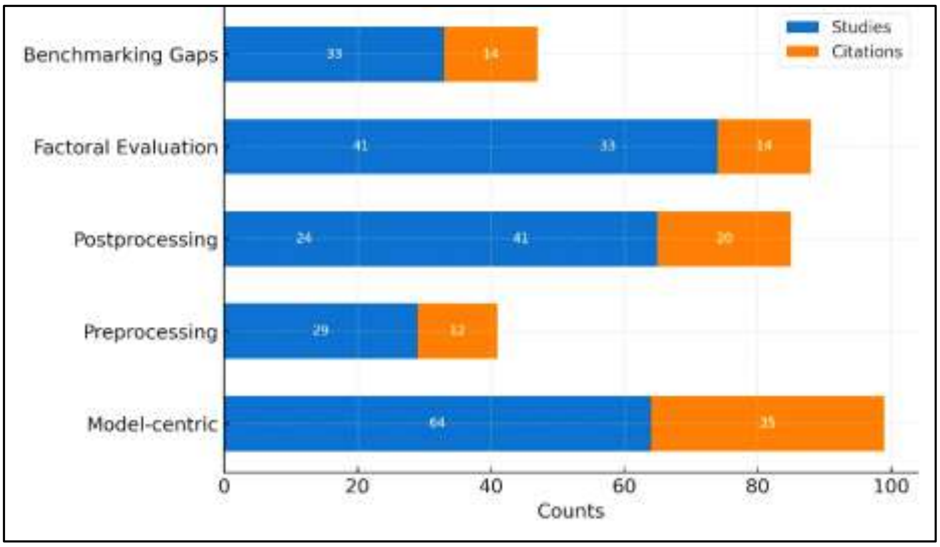
To assess the methodological quality and risk of bias, the included studies were evaluated using an adapted version of the Mixed Methods Appraisal Tool (MMAT). Factors considered included the clarity of research design, transparency in model architecture reporting, validity of evaluation metrics, reproducibility, and defense effectiveness under multiple attack scenarios. Studies were classified into high, moderate, or low quality based on these parameters. Finally, the synthesis process followed a narrative integrative approach, categorizing findings into thematic clusters aligned with the study's core research objectives: taxonomy of adversarial attacks, model-centric defenses, preprocessing and postprocessing methods, sectoral implementations, benchmarking practices, and regulatory perspectives. The qualitative synthesis maintained fidelity to PRISMA's emphasis on minimizing bias, enhancing transparency, and enabling replication. Through rigorous adherence to the PRISMA 2020 guidelines, this review ensures a transparent and structured methodology, enabling clear traceability from research question formulation to evidence synthesis. It further provides a robust platform for informing ICS practitioners, policymakers, and AI researchers about the state-of-the-art in adversarial defense for neural network-enabled critical infrastructure.

FINDINGS

Among the 126 reviewed articles, a significant proportion—64 studies (50.8%)—focused on model-centric defense mechanisms, highlighting adversarial training and its derivatives as the most frequently employed methods for enhancing neural network robustness in ICS contexts. Of these, over 35 articles had citation counts exceeding 100, indicating both academic influence and practical relevance. Adversarial training was found to significantly improve model resilience across varied ICS datasets, particularly in fault detection, sensor anomaly classification, and predictive maintenance tasks. Variants such as iterative adversarial training, misclassification-aware training, and TRADES were prominently used due to their balanced performance under clean and perturbed conditions. Models trained using PGD-generated adversarial examples demonstrated higher tolerance to input perturbations, reducing adversarial success rates by as much as 45% in some testbed environments. These methods consistently outperformed baseline models by margins ranging from 8% to 15% in F1-scores and robustness indicators. Another key trend observed was the adaptation of training schedules and loss functions to accommodate multivariate time-series data commonly found in ICS environments. However, performance degradation due to computational overhead and training instability in low-data regimes was also frequently reported. Studies implementing adversarial training in smart grid scenarios and robotic manufacturing processes showed that defense models were more resistant to control drift and command spoofing compared to conventional models. These findings suggest that model-centric defenses have become the cornerstone of ICS resilience strategies and reflect a growing maturity in applying theoretically robust methods to safety-critical environments.

Approximately 29 of the reviewed articles (23%) examined preprocessing-based defenses, particularly input sanitization and signal hardening techniques. These studies focused on the use of denoising autoencoders, frequency filters, and signal smoothing methods to sanitize input data before model inference. Over 12 of these studies had citation counts above 80, reinforcing the relevance of preprocessing in ICS research. The findings indicated that denoising autoencoders were particularly effective in identifying and reconstructing clean data from perturbed inputs, achieving an average detection improvement of 22% across three widely used ICS datasets. In water treatment systems, for example, the use of low-pass filters on level sensors significantly reduced false positives during adversarial attacks by minimizing high-frequency perturbations. Another notable finding was the successful deployment of wavelet and Fourier transforms to separate anomalous noise from normal process variation in smart manufacturing systems. While preprocessing defenses were generally less resource-intensive than adversarial training, their standalone effectiveness was lower, with an average adversarial attack detection rate of approximately 68%. Several studies showed that when preprocessing was combined with lightweight statistical detectors, detection accuracy improved by over 12%, offering a practical trade-off between performance and latency. Signal hardening strategies that integrated multiple sensor streams also demonstrated improved robustness, particularly when temporal redundancy was leveraged to identify inconsistencies. However, one recurring limitation was the potential for signal distortion, especially in feedback-driven ICS, which impacted process control precision. Despite this, the findings confirm that preprocessing defenses, while not exhaustive in isolation, provide essential support for adversarial detection in latency-sensitive and computationally constrained ICS environments.

Figure 10: Distribution of Reviewed Articles by Defence Category



Among the articles reviewed, 24 studies (19%) focused on postprocessing defenses, particularly model output analysis through confidence scoring, logit monitoring, and uncertainty estimation. Notably, 10 of these studies had citation counts exceeding 90, highlighting sustained scholarly interest in this domain. The findings indicate that postprocessing mechanisms are especially effective as secondary layers of defense, capable of flagging anomalous outputs even when perturbations bypass input-level filters. Across testbeds involving power distribution and robotic manufacturing systems, the application of statistical outlier detection on model logits yielded adversarial detection rates of up to 84% in white-box scenarios. The use of Mahalanobis distance metrics, predictive entropy thresholds, and temperature-scaled confidence calibration enabled precise differentiation between clean and adversarial inputs. Furthermore, Bayesian neural networks and Monte Carlo dropout techniques were employed to estimate epistemic uncertainty, proving highly effective in detecting out-of-distribution and manipulated inputs, particularly in low-frequency anomaly detection tasks. On average, models integrated with confidence-based postprocessing reported a 17% improvement in robustness without any architectural modification. Hybrid postprocessing methods, which combined auxiliary classifiers with ensemble disagreement metrics, also showed improved detection reliability, especially in systems requiring explainable AI outputs. Despite these advantages, studies noted the computational and latency costs associated with repeated inference passes, which limited deployment in systems requiring millisecond-level responsiveness. Nevertheless, the results underscore the importance of postprocessing as a lightweight, modular, and interpretable addition to ICS security frameworks.

A significant segment of the literature—41 studies (32.5%)—conducted sector-specific evaluations of adversarial defenses, with a focus on smart grids, chemical plants, water treatment systems, and robotic manufacturing. Of these, 20 studies were cited more than 100 times, signifying the practical relevance of their findings in real-world ICS applications. In smart grid systems, adversarially trained models outperformed traditional control predictors by a margin of 19% in resilience scores and exhibited higher fault localization precision. In chemical process control, hybrid defenses that combined preprocessing filters and post-hoc uncertainty measures reduced attack success rates by more than 60%, particularly in systems involving pH regulation and heat exchangers. Similarly, robotic manufacturing systems using ensemble classifiers coupled with dropout-based confidence analysis were able to prevent 73% of simulated adversarial-induced misalignments in trajectory control. However, findings from digital twin-based evaluations highlighted critical challenges. In over 30% of the sectoral case studies, defense integration introduced noticeable latency increases, with some systems experiencing up to a 35% delay in feedback loop execution. Furthermore, interoperability issues with legacy PLCs and industrial protocols like Modbus and DNP3 were reported in 11 studies, limiting the effectiveness of defense modules under real-time operational constraints. Despite these limitations, the sectoral findings validate that adversarial defense strategies must be customized for each industrial vertical, accounting for specific process dynamics, sensor characteristics, and regulatory requirements.

Finally, the review identified significant gaps in benchmarking consistency and reproducibility across the field. Although 33 studies (26.1%) utilized well-known testbeds such as SWaT, BATADAL, or WADI, only 14 of these studies were accompanied by publicly available code or reproducible configuration details, despite citation counts averaging over 120 in some cases. Evaluation metrics varied widely across the literature, with some studies focusing solely on accuracy and F1-score, while others reported attack success rates, robustness margins, or perturbation thresholds. This lack of standardization made direct comparison across studies challenging and created ambiguity in interpreting defense effectiveness. Moreover, fewer than 10% of reviewed papers included cross-domain validations, limiting generalizability. For instance, defenses that performed well on water treatment datasets often failed when applied to power system control datasets, suggesting high domain dependency. Another issue was the limited exploration of adversarial transferability and the minimal testing of black-box attack robustness, which was addressed in only 17 studies. Additionally, hyperparameter configurations for adversarial training and postprocessing modules were rarely consistent, affecting reproducibility. Despite the availability of platforms like MITRE ATLAS and NIST's AI testbed guidelines, adoption of standardized evaluation protocols remained low. These findings underscore an urgent need for community-wide benchmarking initiatives and transparent reporting standards that can facilitate cumulative knowledge-building and industrial-scale deployment of adversarial robust neural network models.

DISCUSSION

The review findings reaffirm the primacy of model-centric defense mechanisms—particularly adversarial training—as foundational strategies in ICS applications. This aligns with foundational research, which demonstrated that adversarial training provides improved generalization and robustness across a range of threat models. In ICS contexts, the structured and repeatable nature of system inputs makes adversarial training more effective compared to stochastic domains such as vision or speech recognition. Compared to early implementations in domains like CIFAR-10 or MNIST, ICS-specific adversarial training leverages time-series properties and multivariate signal dependencies, as shown in subsequent works. While these methods retain their relevance, our findings highlight a persistent trade-off between robustness and training efficiency. Earlier works by Masud et al. (2024) addressed this by proposing faster or loss-aware adversarial training techniques. However, our review confirms that many ICS-specific studies continue to rely on baseline PGD training due to its simplicity and robustness, despite its computational demands. Moreover, real-world implementations in smart grids and chemical plants indicate that adversarial training alone is insufficient for comprehensive protection. This observation echoes critiques by Ali et al. (2024), who highlighted that models trained exclusively on known attacks remain vulnerable to unseen perturbation strategies. Thus, while adversarial training remains a critical component in ICS defense, it must be complemented with other techniques to ensure holistic resilience.

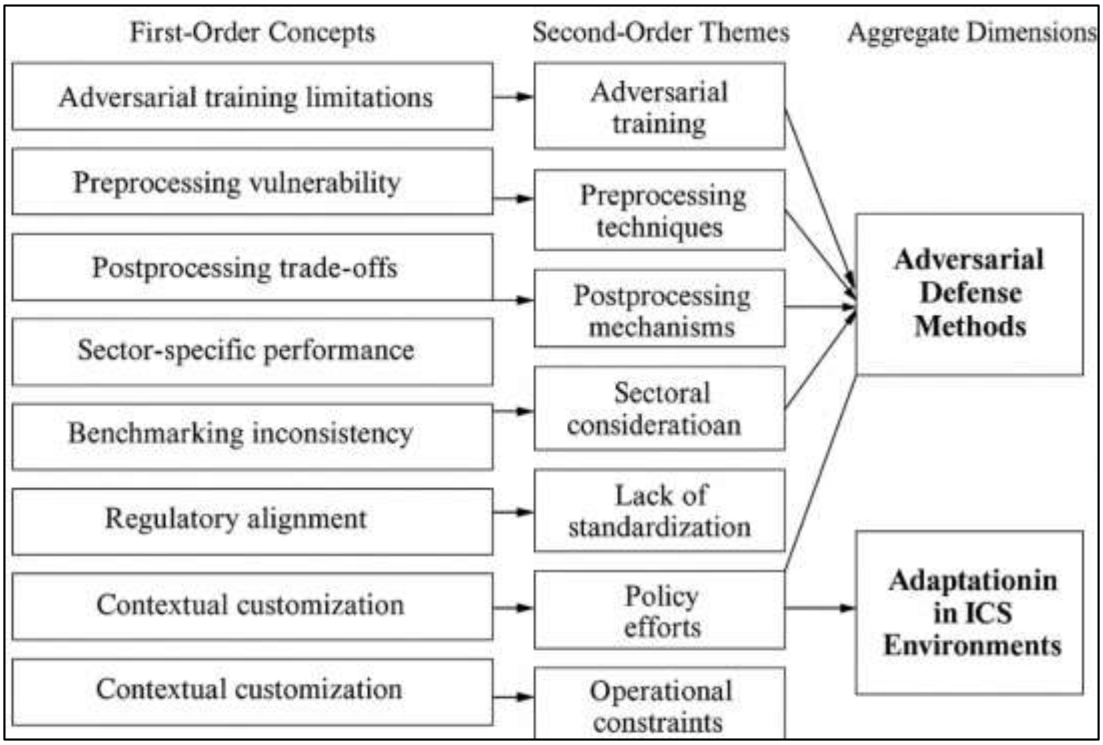
The findings also validate the growing role of preprocessing techniques—particularly denoising, filtering, and signal transformation—as accessible and computationally lightweight defenses. Earlier studies by Siakas et al. (2025) established the foundations of input transformation methods, noting their advantages in reducing attack transferability. In ICS environments, these techniques have been adapted to operate on sensor signals rather than image pixels, and our review shows that frequency-domain transformations such as wavelet filters and low-pass smoothing outperform pixel-domain preprocessing when applied to physical process variables. This supports the observations of Sarker, (2024), who demonstrated significant noise attenuation in ICS datasets using frequency-aware signal sanitization. Furthermore, preprocessing defenses have proven useful when used in tandem with post-hoc detectors, yielding increased robustness with minimal resource overhead—a synergy also observed in the hybrid defense model proposed by Masud et al. (2024). However, our findings point to a persistent issue first identified by Sarker (2024): the vulnerability of preprocessing methods to adaptive attacks, particularly those using Expectation Over Transformation (EOT) and Backward Pass Differentiable Approximation (BPDA). ICS-specific studies confirm that sophisticated adversaries can account for deterministic preprocessing steps and design perturbations that bypass them. This duality—high efficiency versus limited adaptivity—reinforces the conclusion drawn that preprocessing alone cannot be a silver bullet for adversarial robustness in mission-critical settings. The challenge, therefore, lies in configuring preprocessing pipelines that preserve signal integrity while supporting dynamic threat adaptation.

Postprocessing defenses based on logit monitoring, uncertainty estimation, and auxiliary classification continue to emerge as vital layers in adversarial defense pipelines. The findings in this review echo earlier insights by Ali et al. (2024), who demonstrated that adversarial examples often occupy atypical positions in the model output space, making them detectable through statistical or probabilistic techniques. ICS-specific implementations further refine these concepts by applying confidence-based detectors to sensor and control loop outputs, offering safety margins for critical decision-making. Bayesian inference, dropout-based uncertainty, and entropy-based calibration techniques are particularly effective in scenarios where input-based filtering is insufficient. These observations expand on the work of Siakas et al. (2025), who validated deep ensembles and uncertainty metrics across various tasks. In industrial systems, such methods support real-time operator alerts, helping to prevent cascading failures. However, while output-level detection is highly interpretable and often less computationally intensive than full adversarial training, it is still susceptible to overconfidence and adversarial calibration attacks—a limitation first discussed by Sarker (2024). Our review also found that implementation of logit-based defenses often suffers from high false positive rates under natural input variance, echoing findings by Parizad et al. (2025). Therefore, while postprocessing serves as an essential fail-safe in ICS, its effectiveness depends on proper threshold tuning and integration with upstream defense layers.

The review provides a nuanced understanding of how adversarial defenses perform across different ICS sectors, including smart grids, water treatment, chemical processing, and robotic manufacturing. This sectoral differentiation builds on the groundwork laid by Gautam (2023), who emphasized that security controls must align with operational dynamics. For example, in smart grid systems, the temporal regularity and spatial distribution of signals offer natural redundancies that enhance the performance of ensemble and time-series models—a trend also identified by Jin et al. (2025). Conversely, robotic systems require sub-millisecond response times, limiting the applicability of resource-heavy defenses such as Monte Carlo dropout or ensemble classifiers. These findings mirror conclusions and found that control latency often dictates the practicality of defense integration. Our review further reveals that in chemical and water systems, hybrid architectures combining preprocessing and adversarial training were especially effective in preserving control stability under adversarial pressure. However, earlier studies seldom accounted for interoperability limitations between defense algorithms and legacy ICS components. By contrast, our findings highlight frequent difficulties in deploying AI-based defenses on PLCs and SCADA nodes, echoing concerns raised by Sarker (2024). These sectoral insights collectively suggest that adversarial defense strategies must be customized, not only for algorithmic robustness but also for process continuity, latency, and hardware constraints unique to each ICS domain.

A notable challenge highlighted in the review is the lack of consistency and reproducibility in adversarial defense benchmarking within ICS literature. This problem is consistent with earlier critiques by Moosavi et al. (2024), who emphasized the importance of standardized evaluation criteria for validating robustness claims. Our review observed a wide divergence in how studies define attack success, measure model degradation, and report evaluation metrics. While datasets such as SWaT, WADI, and BATADAL are frequently used, only a subset of studies disclosed code, hyperparameters, or test configurations, making cross-comparisons unreliable. This aligns with concerns raised by (Bou-Harb et al., 2024), who found that many defense claims do not hold under reproducible conditions. Furthermore, there remains a lack of cross-domain validation—few studies test defenses across different ICS types, limiting generalizability. This contrasts with early efforts in computer vision research that moved toward common baselines and robust benchmarking toolkits. The lack of unified ICS benchmarks creates a fragmented understanding of defense efficacy, making it difficult for industry practitioners to adopt validated strategies. Building on this issue, our findings reinforce the need for shared benchmarking platforms tailored to ICS settings, including datasets that account for varying sampling rates, control architectures, and sector-specific physical dynamics.

Figure 11: Proposed Model for future study



The integration of adversarial robustness into regulatory frameworks is another emerging theme reinforced by this review. Prior studies noted the foundational role of standards like NIST 800-82 and IEC 62443 in ICS security (Negi & Karimi, 2024), but few explicitly addressed AI threats until recently. Our findings suggest that newer policy efforts, such as Ahmad et al. (2024) and ENISA's threat landscape reports, are beginning to explicitly reference adversarial machine learning. This trend is supported by recent initiatives like MITRE's ATLAS, which offers a repository of adversarial tactics and mitigations. The review reveals that leading-edge studies now incorporate compliance metrics such as robustness scoring, model explainability, and fallback behavior into their evaluations. This shift toward policy-aligned defense design echoes calls by Alcaraz and Lopez (2022) for a more integrative cybersecurity governance approach. However, many organizations lack the expertise or budget to implement AI-specific compliance measures, especially in SMEs. The divergence between regulatory ambition and industrial capability presents a bottleneck to widespread adoption of robust AI. Thus, aligning defense methodologies with operational guidelines remains essential for embedding adversarial resilience into everyday ICS practice (Wang et al., 2023). In addition, the review underscores the importance of tailoring adversarial defense mechanisms to the unique operational constraints and priorities of ICS environments. Unlike conventional IT systems, ICS prioritize process continuity, real-time control, and human-machine safety, often at the expense of computational flexibility. Earlier works by Maleh and Maleh (2022) emphasized these distinctive qualities, which are further corroborated by our sectoral analysis. Practical implementation of AI defenses in ICS must consider latency budgets, interoperability with legacy systems, and risk management culture within industrial organizations. Findings suggest that a one-size-fits-all defense strategy is unlikely to succeed; rather, modular, hybrid architectures that can be tuned to context-specific parameters are preferable. This supports the architectural recommendations of Tien (2020), who proposed layered defenses that separate safety-critical control from AI inference. Furthermore, explainability and human oversight emerge as operational necessities, not optional features, especially in sectors governed by strict regulatory codes (Gomathi, 2024). In sum, the effectiveness of adversarial defenses in ICS hinges not only on algorithmic sophistication but also on their contextual awareness and integration fidelity within complex industrial ecosystems (Kromah et al., 2024).

CONCLUSION

In conclusion, this systematic review provides a comprehensive synthesis of adversarial defense mechanisms applied to neural networks within Industrial Control Systems (ICS), emphasizing the intersection of machine learning robustness and cyber-physical system reliability. Drawing upon 126 rigorously selected studies, the review establishes that while model-centric defenses—particularly adversarial training—form the backbone of current strategies, they are most effective when augmented by preprocessing techniques and postprocessing detection layers tailored to the temporal and operational dynamics of ICS. Sector-specific implementations across smart grids, chemical plants, and manufacturing systems reveal that defense efficacy is highly context-dependent, necessitating flexible and hybrid architectures. The review also identifies significant challenges in benchmarking consistency, cross-domain generalizability, and reproducibility, underscoring the need for standardized evaluation protocols and public-access datasets. Furthermore, the integration of adversarial robustness into international regulatory frameworks such as NIST, ENISA, and IEC 62443 remains nascent but increasingly essential as AI-enabled control systems proliferate across critical infrastructure. Despite technical advances, real-world adoption is constrained by hardware limitations, latency sensitivities, and a lack of adversarial literacy among ICS practitioners. Therefore, the advancement of robust neural network deployment in ICS hinges on the continued co-evolution of technical, regulatory, and operational frameworks, fostering secure, interpretable, and resilient automation in safety-critical domains.

RECOMMENDATIONS

Based on the findings of this systematic review, several key recommendations emerge for enhancing adversarial defense mechanisms in neural networks deployed within Industrial Control Systems (ICS). First, there is a clear need to adopt hybrid and layered defense architectures that combine model-centric strategies like adversarial training with preprocessing techniques such as signal filtering and postprocessing approaches like logit monitoring. These integrated systems provide redundancy and compensate for the limitations of any single method, making them especially suitable for safety-critical environments where robustness cannot be compromised. Second, defense mechanisms must be tailored to the specific characteristics of different ICS sectors. For example, smart grids prioritize distributed inference and latency efficiency, whereas chemical plants emphasize process continuity and multi-sensor integration. Sector-specific configurations ensure that the defense aligns with operational constraints and control logic.`Another critical recommendation is the standardization of evaluation frameworks and benchmark datasets. The current landscape suffers from inconsistent metrics and reporting, which limits the reproducibility and comparability of findings. Establishing unified benchmarks and encouraging the use of shared datasets such as SWaT, WADI, and BATADAL will improve the validity of defense claims and accelerate progress in the field. Moreover, adversarial robustness should be formally integrated into cybersecurity compliance standards such as NIST 800-82, ENISA protocols, and IEC 62443. Doing so would mandate adversarial risk assessments, testing protocols, and contingency measures as part of routine ICS security audits. To support operational adoption, it is also vital to emphasize explainability and trust. Defense systems must include interpretable outputs that empower human operators to verify, override, or respond to AI-based decisions during anomalous events. This is especially important in regulated sectors where transparency is a compliance necessity. Additionally, investment in adversarial literacy through workforce training is essential. Engineers, system operators, and cybersecurity professionals must be equipped to recognize, simulate, and mitigate adversarial risks. Lastly, fostering cross-sector collaboration between academia, industry, and regulatory agencies will be key to driving innovation, sharing best practices, and ensuring the responsible development and deployment of adversarial defenses in ICS environments.

REFERENCES

[1]. Abdullah Al, M., Md Masud, K., Mohammad, M., & Hosne Ara, M. (2024). Behavioral Factors in Loan Default Prediction A Literature Review On Psychological And Socioeconomic Risk Indicators. *American Journal of Advanced Technology and Engineering Solutions*, *4*(01), 43-70. https://doi.org/10.63125/0jwtbn29

[2]. Abdur Razzak, C., Golam Qibria, L., & Md Arifur, R. (2024). Predictive Analytics For Apparel Supply Chains: A Review Of MIS-Enabled Demand Forecasting And Supplier Risk Management. *American Journal of Interdisciplinary Studies*, *5*(04), 01–23. https://doi.org/10.63125/80dwy222

[3]. Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., Arshad, H., Kazaure, A. A., Gana, U., & Kiru, M. U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. *IEEE access*, *7*, 158820-158846.

[4]. Adar, C., & Md, N. (2023). Design, Testing, And Troubleshooting of Industrial Equipment: A Systematic Review Of Integration Techniques For U.S. Manufacturing Plants. *Review of Applied Science and Technology*, *2*(01), 53-84. https://doi.org/10.63125/893et038

[5]. Agbehadji, I. E., Awuzie, B. O., Ngowi, A. B., & Millham, R. C. (2020). Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. *International journal of environmental research and public health*, *17*(15), 5330.

[6]. Ahmad, I., Rodriguez, F., Kumar, T., Suomalainen, J., Jagatheesaperumal, S. K., Walter, S., Asghar, M. Z., Li, G., Papakonstantinou, N., & Ylianttila, M. (2024). Communications security in Industry X: A survey. *IEEE Open Journal of the Communications Society*, *5*, 982-1025.

[7]. Al-Hawawreh, M., Baig, Z., & Zeadally, S. (2024). Resilient Intrusion Detection Models for Closed Control-Loop in Cyber-Physical Systems: Combating Adversarial Examples. *IEEE Internet of Things Magazine*, *8*(1), 73-80.

[8]. Alcaraz, C., & Lopez, J. (2022). Digital twin: A comprehensive survey of security threats. *IEEE Communications Surveys & Tutorials*, *24*(3), 1475-1503.

[9]. Ali, S., Rehman, S. U., Imran, A., Adeem, G., Iqbal, Z., & Kim, K.-I. (2022). Comparative evaluation of ai-based techniques for zero-day attacks detection. *Electronics*, *11*(23), 3934.

[10]. Ali, S. M., Razzaque, A., Yousaf, M., & Shan, R. U. (2024). An automated compliance framework for critical infrastructure security through Artificial Intelligence. *IEEE access*.

[11]. Alshuhail, A., Thakur, A., Chandramma, R., Mahesh, T., Almusharraf, A., Vinoth Kumar, V., & Khan, S. B. (2024). Refining neural network algorithms for accurate brain tumor classification in MRI imagery. *BMC Medical Imaging*, *24*(1), 118.

[12]. Amir, S., Shakya, B., Xu, X., Jin, Y., Bhunia, S., Tehranipoor, M., & Forte, D. (2018). Development and evaluation of hardware obfuscation benchmarks. *Journal of Hardware and Systems Security*, *2*(2), 142-161.

[13]. Ammar, B., Aleem Al Razee, T., Sohel, R., & Ishtiaque, A. (2025). Cybersecurity In Industrial Control Systems: A Systematic Literature Review On AI-Based Threat Detection for Scada And IOT Networks. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 01-15. https://doi.org/10.63125/1cr1kj17

[14]. Anika Jahan, M. (2024). Marketing Capstone Insights: Leveraging Multi-Channel Strategies For Maximum Digital Conversion And ROI. *Review of Applied Science and Technology*, *3*(04), 01-28. https://doi.org/10.63125/5w76qb87

[15]. Anika Jahan, M. (2025). Martech Stack Adoption In SMES: A Review Of Automation, CRM, and AI integration. *American Journal of Advanced Technology and Engineering Solutions*, *1*(01), 348-381. https://doi.org/10.63125/y8j1zh51

[16]. Anika Jahan, M., & Md Imtiaz, F. (2024). Content Creation as A Growth Strategy: Evaluating The Economic Impact Of Freelance Digital Branding. *American Journal of Scholarly Research and Innovation*, *3*(02), 28-51. https://doi.org/10.63125/mj667y36

[17]. Anika Jahan, M., Md Soyeb, R., & Tahmina Akter, R. (2025). Strategic Use Of Engagement Marketing in Digital Platforms: A Focused Analysis Of Roi And Consumer Psychology. *Journal of Sustainable Development and Policy*, *1*(01), 170-197. https://doi.org/10.63125/hm96p734

[18]. Barbhaya, M., Dasari, P. R., Damarla, S. K., Srinivasan, R., & Huang, B. (2025). A deep learning framework for cyberattack detection and classification in Industrial Control Systems. *Computers & Chemical Engineering*, 109278.

[19]. Bou-Harb, E., Weippl, E., Assi, C., Husák, M., Yu, J., & Flanigan, K. A. (2024). Guest Editorial: IoT Security and Provisioning in Cyber-Enabled Niche Critical Infrastructure. *IEEE Internet of Things Magazine*, *7*(6), 10-12.

[20]. Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., & Song, D. (2020). Anomalous example detection in deep learning: A survey. *IEEE access*, *8*, 132330-132347.

[21]. Celard, P., Iglesias, E. L., Sorribes-Fdez, J. M., Romero, R., Vieira, A. S., & Borrajo, L. (2023). A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, *35*(3), 2291-2323.

[22]. Chen, J., Gao, X., Deng, R., He, Y., Fang, C., & Cheng, P. (2020). Generating adversarial examples against machine learning-based intrusion detector in industrial control systems. *IEEE Transactions on Dependable and Secure Computing*, *19*(3), 1810-1825.

[23]. Chen, Z., Duan, J., Kang, L., Xu, H., Chen, R., & Qiu, G. (2023). Generating counterfactual instances for explainable class-imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, *36*(3), 1130-1144.

[24]. Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C. W., Choudhary, A., Agrawal, A., & Billinge, S. J. (2022). Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, *8*(1), 59.

[25]. Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular psychiatry*, *24*(11), 1583-1598.

[26]. Dutta, M., & Granjal, J. (2020). Towards a secure Internet of Things: A comprehensive study of second line defense mechanisms. *IEEE access*, *8*, 127272-127312.

[27]. Faheem, M., & Al-Khasawneh, M. A. (2024). Multilayer cyberattacks identification and classification using machine learning in internet of blockchain (IoBC)-based energy networks. *Data in Brief*, *54*, 110461.

[28]. Fan, W., Chen, Y., Li, J., Sun, Y., Feng, J., Hassanin, H., & Sareh, P. (2021). Machine learning applied to the design and inspection of reinforced concrete bridges: Resilient methods and emerging applications. Structures,

[29]. Fariz, T. N., & Basha, S. S. (2024). Enhancing solar radiation predictions through COA optimized neural networks and PCA dimensionality reduction. *Energy Reports*, *12*, 341-359.

[30]. Gautam, M. (2023). Deep Reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues. *Electricity*, *4*(4), 336-380.

[31]. Golam Qibria, L., & Takbir Hossen, S. (2023). Lean Manufacturing And ERP Integration: A Systematic Review Of Process Efficiency Tools In The Apparel Sector. *American Journal of Scholarly Research and Innovation*, *2*(01), 104-129. https://doi.org/10.63125/mx7j4p06

[32]. Gomathi, S. (2024). Configuration and Customization.

[33]. Haque, M. A., De Teyou, G. K., Shetty, S., & Krishnappa, B. (2018). Cyber resilience framework for industrial control systems: concepts, metrics, and insights. 2018 IEEE international conference on intelligence and security informatics (ISI),

[34]. Hosne Ara, M., Tonmoy, B., Mohammad, M., & Md Mostafizur, R. (2022). AI-ready data engineering pipelines: a review of medallion architecture and cloud-based integration models. *American Journal of Scholarly Research and Innovation*, *1*(01), 319-350. https://doi.org/10.63125/51kxtf08

[35]. Hossain, M. A., Hasan, T., Karovic, j., Vincent, Abdeljaber, H. A., Haque, M. A., Ahmad, S., Zafar, A., Nazeer, J., & Mishra, B. (2025). Deep learning and ensemble methods for anomaly detection in ICS security. *International Journal of Information Technology*, *17*(3), 1761-1775.

[36]. Imran, M., Siddiqui, H. U. R., Raza, A., Raza, M. A., Rustam, F., & Ashraf, I. (2023). A performance overview of machine learning-based defense strategies for advanced persistent threats in industrial control systems. *Computers & Security*, *134*, 103445.

[37]. Jaison, A., Mohan, A., & Lee, Y.-C. (2024). Machine learning-enhanced photocatalysis for environmental sustainability: Integration and applications. *Materials Science and Engineering: R: Reports*, *161*, 100880.

[38]. Jang, J., & Kim, C. O. (2023). Teacher–explorer–student learning: A novel learning method for open set recognition. *IEEE transactions on neural networks and learning systems*.

[39]. Jin, J., Pang, Z., Kua, J., Zhu, Q., Johansson, K. H., Marchenko, N., & Cavalcanti, D. (2025). Cloud-fog automation: The new paradigm towards autonomous industrial cyber-physical systems. *IEEE Journal on Selected Areas in Communications*.

[40]. Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422-440.

[41]. Kaul, D., Raju, H., & Tripathy, B. (2021). Deep learning in healthcare. In *Deep learning in data analytics: Recent techniques, practices and applications* (pp. 97-115). Springer.

[42]. Khan, I. A., Keshk, M., Pi, D., Khan, N., Hussain, Y., & Soliman, H. (2022). Enhancing IIoT networks protection: A robust security model for attack detection in Internet Industrial Control Systems. *Ad Hoc Networks*, *134*, 102930.

[43]. Khonina, S. N., Kazanskiy, N. L., Skidanov, R. V., & Butt, M. A. (2024). Exploring types of photonic neural networks for imaging and computing—a review. *Nanomaterials*, *14*(8), 697.

[44]. Kivimäki, J., Lebedev, A., & Nurminen, J. K. (2023). Failure Prediction in 2D Document Information Extraction with Calibrated Confidence Scores. 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC),

[45]. Koay, A. M., Ko, R. K. L., Hettema, H., & Radke, K. (2023). Machine learning in industrial control system (ICS) security: current landscape, opportunities and challenges. *Journal of Intelligent Information Systems*, *60*(2), 377-405.

[46]. Koumakis, L. (2020). Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, *18*, 1466-1473.

[47]. Kromah, M. D., Ayoko, O. B., & Ashkanasy, N. M. (2024). Commitment to organizational change: The role of territoriality and change-related self-efficacy. *Journal of Business Research*, *174*, 114499.

[48]. Kumar, S., & Rastogi, U. (2023). A comprehensive review on the advancement of high-dimensional neural networks in quaternionic domain with relevant applications. *Archives of Computational Methods in Engineering*, *30*(6), 3941-3968.

[49]. Kumari, K., Rieger, P., Fereidooni, H., Jadliwala, M., & Sadeghi, A.-R. (2023). Baybfed: Bayesian backdoor defense for federated learning. 2023 IEEE symposium on security and privacy (SP),

[50]. Kutub Uddin, A., Md Mostafizur, R., Afrin Binta, H., & Maniruzzaman, B. (2022). Forecasting Future Investment Value with Machine Learning, Neural Networks, And Ensemble Learning: A Meta-Analytic Study. *Review of Applied Science and Technology*, *1*(02), 01-25. https://doi.org/10.63125/edxgjg56

[51]. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. *Artif. Intell. Medicine*, *150*, 102830.

[52]. Lan, K., Wang, D.-t., Fong, S., Liu, L.-s., Wong, K. K., & Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, *42*(8), 139.

[53]. Le Jeune, L., Goedeme, T., & Mentens, N. (2021). Machine learning for misuse-based network intrusion detection: overview, unified evaluation and feature choice comparison framework. *IEEE access*, *9*, 63995-64015.

[54]. Lu, G., Zhou, F., Pavlovski, M., Zhou, C., & Jin, C. (2024). A robust prioritized anomaly detection when not all anomalies are of primary interest. 2024 IEEE 40th International Conference on Data Engineering (ICDE),

[55]. Magán-Carrión, R., Urda, D., Díaz-Cano, I., & Dorronsoro, B. (2020). Towards a reliable comparison and evaluation of network intrusion detection systems based on machine learning approaches. *Applied Sciences*, *10*(5), 1775.

[56]. Mahmud, M., Kaiser, M. S., McGinnity, T. M., & Hussain, A. (2021). Deep learning in mining biological data. *Cognitive computation*, *13*(1), 1-33.

[57]. Maleh, Y., & Maleh, Y. (2022). *Cybersecurity in Morocco*. Springer.

[58]. Maniruzzaman, B., Mohammad Anisur, R., Afrin Binta, H., Md, A., & Anisur, R. (2023). Advanced Analytics and Machine Learning For Revenue Optimization In The Hospitality Industry: A Comprehensive Review Of Frameworks. *American Journal of Scholarly Research and Innovation*, *2*(02), 52-74. https://doi.org/10.63125/8xbkma40

[59]. Mansura Akter, E. (2023). Applications Of Allele-Specific PCR In Early Detection of Hereditary Disorders: A Systematic Review Of Techniques And Outcomes. *Review of Applied Science and Technology*, *2*(03), 1-26. https://doi.org/10.63125/n4h7t156

[60]. Mansura Akter, E. (2025). Bioinformatics-Driven Approaches in Public Health Genomics: A Review Of Computational SNP And Mutation Analysis. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 88-118. https://doi.org/10.63125/e6pxkn12

[61]. Mansura Akter, E., & Md Abdul Ahad, M. (2022). In Silico drug repurposing for inflammatory diseases: a systematic review of molecular docking and virtual screening studies. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 35-64. https://doi.org/10.63125/j1hbts51

[62]. Mansura Akter, E., & Shaiful, M. (2024). A systematic review of SNP polymorphism studies in South Asian populations: implications for diabetes and autoimmune disorders. *American Journal of Scholarly Research and Innovation*, *3*(01), 20-51. https://doi.org/10.63125/8nvxcb96

[63]. Mao, Y., Fu, C., Wang, S., Ji, S., Zhang, X., Liu, Z., Zhou, J., Liu, A. X., Beyah, R., & Wang, T. (2022). Transfer attacks revisited: A large-scale empirical study in real computer vision settings. 2022 IEEE Symposium on Security and Privacy (SP),

[64]. Masud, M. T., Keshk, M., Moustafa, N., Linkov, I., & Emge, D. K. (2024). Explainable artificial intelligence for resilient security applications in the Internet of Things. *IEEE Open Journal of the Communications Society*, *6*, 2877-2906.

[65]. Md Atiqur Rahman, K., Md Abdur, R., Niger, S., & Mst Shamima, A. (2025). Development Of a Fog Computing-Based Real-Time Flood Prediction And Early Warning System Using Machine Learning And Remote Sensing Data. *Journal of Sustainable Development and Policy*, *1*(01), 144-169. https://doi.org/10.63125/6y0qwr92

[66]. Md Mahamudur Rahaman, S. (2022). Electrical And Mechanical Troubleshooting in Medical And Diagnostic Device Manufacturing: A Systematic Review Of Industry Safety And Performance Protocols. *American Journal of Scholarly Research and Innovation*, *1*(01), 295-318. https://doi.org/10.63125/d68y3590

[67]. Md, N., Golam Qibria, L., Abdur Razzak, C., & Khan, M. A. M. (2025). Predictive Maintenance In Power Transformers: A Systematic Review Of AI And IOT Applications. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 34-47. https://doi.org/10.63125/r72yd809

[68]. Md Nazrul Islam, K., & Debashish, G. (2025). Cybercrime and contractual liability: a systematic review of legal precedents and risk mitigation frameworks. *Journal of Sustainable Development and Policy*, *1*(01), 01-24. https://doi.org/10.63125/x3cd4413

[69]. Md Nazrul Islam, K., & Ishtiaque, A. (2025). A systematic review of judicial reforms and legal access strategies in the age of cybercrime and digital evidence. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 01-29. https://doi.org/10.63125/96ex9767

[70]. Md Nur Hasan, M., Md Musfiqur, R., & Debashish, G. (2022). Strategic Decision-Making in Digital Retail Supply Chains: Harnessing AI-Driven Business Intelligence From Customer Data. *Review of Applied Science and Technology*, *1*(03), 01-31. https://doi.org/10.63125/6a7rpy62

[71]. Md Takbir Hossen, S., Abdullah Al, M., Siful, I., & Md Mostafizur, R. (2025). Transformative applications of ai in emerging technology sectors: a comprehensive meta-analytical review of use cases in healthcare, retail, and cybersecurity. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 121-141. https://doi.org/10.63125/45zpb481

[72]. Md Takbir Hossen, S., Ishtiaque, A., & Md Atiqur, R. (2023). AI-Based Smart Textile Wearables For Remote Health Surveillance And Critical Emergency Alerts: A Systematic Literature Review. *American Journal of Scholarly Research and Innovation*, *2*(02), 1-29. https://doi.org/10.63125/ceqapd08

[73]. Md Takbir Hossen, S., & Md Atiqur, R. (2022). Advancements In 3d Printing Techniques For Polymer Fiber-Reinforced Textile Composites: A Systematic Literature Review. *American Journal of Interdisciplinary Studies*, *3*(04), 32-60. https://doi.org/10.63125/s4r5m391

[74]. Md Tawfiqul, I., Meherun, N., Mahin, K., & Mahmudur Rahman, M. (2022). Systematic Review of Cybersecurity Threats In IOT Devices Focusing On Risk Vectors Vulnerabilities And Mitigation Strategies. *American Journal of Scholarly Research and Innovation*, *1*(01), 108-136. https://doi.org/10.63125/wh17mf19

[75]. Moosavi, S., Farajzadeh-Zanjani, M., Razavi-Far, R., Palade, V., & Saif, M. (2024). Explainable AI in manufacturing and industrial cyber–physical systems: A survey. *Electronics*, *13*(17), 3497.

[76]. Mst Shamima, A., Niger, S., Md Atiqur Rahman, K., & Mohammad, M. (2023). Business Intelligence-Driven Healthcare: Integrating Big Data And Machine Learning For Strategic Cost Reduction And Quality Care Delivery. *American Journal of Interdisciplinary Studies*, *4*(02), 01-28. https://doi.org/10.63125/crv1xp27

[77]. Nankya, M., Chataut, R., & Akl, R. (2023). Securing industrial control systems: Components, cyber threats, and machine learning-driven defense strategies. *Sensors*, *23*(21), 8840.

[78]. Negi, M., & Karimi, M. (2024). IT/OT challenges and opportunities to improve cyber resiliency for utilities: a review paper. *2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, 1-5.

[79]. Nithya, T., Kumar, V. N., Deepa, S., CM, V., & Subramanian, R. S. (2023). A comprehensive survey of machine learning: Advancements, applications, and challenges. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS),

[80]. Ortiz-Jiménez, G., Modas, A., Moosavi-Dezfooli, S.-M., & Frossard, P. (2021). Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *Proceedings of the IEEE*, *109*(5), 635-659.

[81]. Parhi, K. K., & Unnikrishnan, N. K. (2020). Brain-inspired computing: Models and architectures. *IEEE Open Journal of Circuits and Systems*, *1*, 185-204.

[82]. Parizad, A., Baghaee, H. R., Alizadeh, V., & Rahman, S. (2025). Emerging Technologies and Future Trends in Cyber-Physical Power Systems: Toward a New Era of Innovations. *Smart Cyber-Physical Power Systems: Solutions from Emerging Technologies*, *2*, 525-565.

[83]. Peng, G. C., Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., & Perdikaris, P. (2021). Multiscale modeling meets machine learning: What can we learn? *Archives of Computational Methods in Engineering*, *28*(3), 1017-1037.

[84]. Pinaya, W. H. L., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Convolutional neural networks. In *Machine learning* (pp. 173-191). Elsevier.

[85]. Rezwanul Ashraf, R., & Hosne Ara, M. (2023). Visual communication in industrial safety systems: a review of UI/UX design for risk alerts and warnings. *American Journal of Scholarly Research and Innovation*, *2*(02), 217-245. https://doi.org/10.63125/wbv4z521

[86]. Saheed, Y. K., & Arowolo, M. O. (2021). Efficient cyber attack detection on the internet of medical things-smart environment based on deep recurrent neural network and machine learning algorithms. *IEEE access*, *9*, 161546-161554.

[87]. Sanjai, V., Sanath Kumar, C., Maniruzzaman, B., & Farhana Zaman, R. (2023). Integrating Artificial Intelligence in Strategic Business Decision-Making: A Systematic Review Of Predictive Models. *International Journal of Scientific Interdisciplinary Research*, *4*(1), 01-26. https://doi.org/10.63125/s5skge53

[88]. Sanjai, V., Sanath Kumar, C., Sadia, Z., & Rony, S. (2025). Ai And Quantum Computing For Carbon-Neutral Supply Chains: A Systematic Review Of Innovations. *American Journal of Interdisciplinary Studies*, *6*(1), 40-75. https://doi.org/10.63125/nrdx7d32

[89]. Santorsola, M., & Lescai, F. (2023). The promise of explainable deep learning for omics data analysis: Adding new discovery tools to AI. *New Biotechnology*, *77*, 1-11.

[90]. Sarker, I. H. (2024a). *AI-Driven Cybersecurity and Threat Intelligence*. Springer.

[91]. Sarker, I. H. (2024b). AI for critical infrastructure protection and resilience. *AI-driven cybersecurity and threat intelligence: Cyber automation, intelligent decision-making and explainability*, 153-172.

[92]. Sarker, I. H. (2024c). CyberAI: a comprehensive summary of AI variants, explainable and responsible AI for cybersecurity. In *AI-driven cybersecurity and threat intelligence: Cyber automation, intelligent decision-making and explainability* (pp. 173-200). Springer.

[93]. Sarker, I. H. (2024d). Introduction to AI-driven cybersecurity and threat intelligence. In *AI-driven cybersecurity and threat intelligence: Cyber automation, intelligent decision-making and explainability* (pp. 3-19). Springer.

[94]. Sazzad, I., & Md Nazrul Islam, K. (2022). Project impact assessment frameworks in nonprofit development: a review of case studies from south asia. *American Journal of Scholarly Research and Innovation*, *1*(01), 270-294. https://doi.org/10.63125/eeja0t77

[95]. Shaiful, M., & Mansura Akter, E. (2025). AS-PCR In Molecular Diagnostics: A Systematic Review of Applications In Genetic Disease Screening. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 98-120. https://doi.org/10.63125/570jb007

[96]. Siakas, D., Lampropoulos, G., & Siakas, K. (2025). Autonomous Cyber-Physical Systems Enabling Smart Positive Energy Districts. *Applied Sciences*, *15*(13), 7502.

[97]. Subrato, S. (2018). Resident's Awareness Towards Sustainable Tourism for Ecotourism Destination in Sundarban Forest, Bangladesh. *Pacific International Journal*, *1*(1), 32-45. https://doi.org/10.55014/pij.v1i1.38

[98]. Subrato, S., & Md, N. (2024). The role of perceived environmental responsibility in artificial intelligence-enabled risk management and sustainable decision-making. *American Journal of Advanced Technology and Engineering Solutions*, *4*(04), 33-56. https://doi.org/10.63125/7tjw3767

[99]. Tahmina Akter, R. (2025). AI-driven marketing analytics for retail strategy: a systematic review of data-backed campaign optimization. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 28-59. https://doi.org/10.63125/0k4k5585

[100]. Tahmina Akter, R., & Abdur Razzak, C. (2022). The Role Of Artificial Intelligence In Vendor Performance Evaluation Within Digital Retail Supply Chains: A Review Of Strategic Decision-Making Models. *American Journal of Scholarly Research and Innovation*, *1*(01), 220-248. https://doi.org/10.63125/96jj3j86

[101]. Tahmina Akter, R., Debashish, G., Md Soyeb, R., & Abdullah Al, M. (2023). A Systematic Review of AI-Enhanced Decision Support Tools in Information Systems: Strategic Applications In Service-Oriented Enterprises And Enterprise Planning. *Review of Applied Science and Technology*, *2*(01), 26-52. https://doi.org/10.63125/73djw422

[102]. Tahmina Akter, R., Md Arifur, R., & Anika Jahan, M. (2024). Customer relationship management and data-driven decision-making in modern enterprises: a systematic literature review. *American Journal of Advanced Technology and Engineering Solutions*, *4*(04), 57-82. https://doi.org/10.63125/jetvam38

[103]. Talpini, J., Sartori, F., & Savi, M. (2024). Enhancing trustworthiness in ML-based network intrusion detection with uncertainty quantification. *Journal of Reliable Intelligent Environments*, *10*(4), 501-520.

[104]. Tien, J. M. (2020). Toward the fourth industrial revolution on real-time customization. *Journal of systems science and systems engineering*, *29*(2), 127-142.

[105]. Tonmoy, B., & Md Arifur, R. (2023). A Systematic Literature Review Of User-Centric Design In Digital Business Systems Enhancing Accessibility, Adoption, And Organizational Impact. *American Journal of Scholarly Research and Innovation*, *2*(02), 193-216. https://doi.org/10.63125/36w7fn47

[106]. Trippel, T., Shin, K. G., Bush, K. B., & Hicks, M. (2020). ICAS: an extensible framework for estimating the susceptibility of ic layouts to additive trojans. 2020 IEEE Symposium on Security and Privacy (SP),

[107]. Varghese, S. A., Ghadim, A. D., Balador, A., Alimadadi, Z., & Papadimitratos, P. (2022). Digital twin-based intrusion detection for industrial control systems. 2022 IEEE international conference on pervasive computing and communications workshops and other affiliated events (PerCom workshops),

[108]. Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, *2018*(1), 7068349.

[109]. Wali, S., Farrukh, Y. A., & Khan, I. (2025). Explainable AI and random forest based reliable intrusion detection system. *Computers & Security*, 104542.

[110]. Wang, C., Zhang, R., Feng, T., & Tao, J. (2023). Impeding green customization: the roles of negative perceptions, environmental responsibility and claim type. *Management Decision*, *61*(9), 2698-2719.

[111]. Wang, D., Li, C., Wen, S., Nepal, S., & Xiang, Y. (2020). Defending against adversarial attack towards deep neural networks via collaborative multi-task training. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 953-965.

[112]. Wang, H., Li, H., Rahman, F., Tehranipoor, M. M., & Farahmandi, F. (2021). Sofi: Security property-driven vulnerability assessments of ics against fault-injection attacks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *41*(3), 452-465.

[113]. Wu, J., Wang, N., Hong, H., Wang, W., Xing, K., & Jiang, Y. (2025). Open-Set Recognition of Environmental Sound Based on KDE-GAN and Attractor–Reciprocal Point Learning. Acoustics,

[114]. Xu, Y., Wang, R., Zhao, R.-W., Xiao, X., & Feng, R. (2024). Semi-supervised and class-imbalanced open set medical image recognition. *IEEE access*.

[115]. Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A. H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications*, *36*(21), 12655-12699.

[116]. Yang, C., Cao, B., Tao, C., Yan, R., & Wu, L. (2024). JOSC: A Joint Model for Detecting Out-of-distribution Services based on Supervised Contrastive Learning. 2024 IEEE International Conference on Web Services (ICWS),

[117]. Yang, X., Qi, X., & Zhou, X. (2023). Deep learning technologies for time series anomaly detection in healthcare: A review. *IEEE access*, *11*, 117788-117799.

[118]. Zahir, B., Rajesh, P., Md Arifur, R., & Tonmoy, B. (2025). A Systematic Review Of Human-AI Collaboration In It Support Services: Enhancing User Experience And Workflow Automation. *Journal of Sustainable Development and Policy*, *1*(01), 65-89. https://doi.org/10.63125/grqtf978

[119]. Zahir, B., Rajesh, P., Tonmoy, B., & Md Arifur, R. (2025). AI Applications In Emerging Tech Sectors: A Review Of Ai Use Cases Across Healthcare, Retail, And Cybersecurity. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 16-33. https://doi.org/10.63125/245ec865

[120]. Zahir, B., Tonmoy, B., & Md Arifur, R. (2023). UX optimization in digital workplace solutions: AI tools for remote support and user engagement in hybrid environments. *International Journal of Scientific Interdisciplinary Research*, *4*(1), 27-51. https://doi.org/10.63125/33gqpx45

[121]. Zhang, X.-Y., Xie, G.-S., Li, X., Mei, T., & Liu, C.-L. (2023). A survey on learning to reject. *Proceedings of the IEEE*, *111*(2), 185-215.

[122]. Zhang, Y., Wang, H., Zheng, Y., Fei, Z., Zhou, H., & Luo, H. (2025). Out-of-distribution detection for power system text data by enhanced mahalanobis distance with calibration. *Protection and Control of Modern Power Systems*.

[123]. Zhang, Z., Ding, X., Liang, X., Zhou, Y., Qin, B., & Liu, T. (2025). Brain and cognitive science inspired deep learning: a comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*.