Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

A QUANTITATIVE ASSESSMENT OF SECURE NEURAL NETWORK ARCHITECTURES FOR FAULT DETECTION IN INDUSTRIAL CONTROL SYSTEMS

Md Tawfiqul Islam[1]

[1] *Master of Engineering Management, Lamar University, Texas, USA*
Email: mislam91@lamar.edu; tawfiq.ctgbd@gmail.com

Abstract

Industrial Control Systems (ICS) form the core infrastructure for critical sectors such as energy, water, manufacturing, and transportation, yet their increasing digital interconnectivity has exposed them to complex fault dynamics and sophisticated cyber-physical threats. Traditional fault detection mechanisms—whether rule-based or model-driven—often fail to cope with the nonlinearity, high dimensionality, and adversarial vulnerabilities prevalent in modern ICS environments. To address these limitations, this study conducts a comprehensive quantitative evaluation of secure neural network architectures tailored for ICS fault detection. Specifically, the research compares standard deep learning models—including Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Long Short-Term Memory networks (LSTM)—with their security-enhanced counterparts, such as adversarially trained LSTM (AT-LSTM) and autoencoder-based input sanitization models (AE-S). Using two publicly available benchmark datasets—SWaT and WADI—and simulating three distinct adversarial threat scenarios (white-box, black-box, and gray-box), the study systematically measures performance across multiple dimensions including accuracy, F1-score, robustness accuracy, attack success rate, inference latency, and fault detection delay. The results reveal that secure architectures not only retain over 80% classification accuracy under white-box attacks but also maintain low false positive rates and detection delays under two seconds, validating their suitability for real-time deployment. Furthermore, secure models exhibit superior generalization across rare fault classes and higher consistency in adversarial environments, outperforming baseline models by wide margins across all tested metrics. These findings confirm that integrating adversarial defense mechanisms into neural network designs substantially improves the operational reliability and cybersecurity resilience of ICS fault detection systems. The study provides a validated framework and practical insights to guide the deployment of robust AI-based monitoring in safety-critical industrial domains, highlighting the role of secure neural networks as a foundational component for next-generation intelligent control systems.

Keywords

Secure Neural Networks; Industrial Control Systems (ICS); Fault Detection; Adversarial Robustness; Cyber-Physical Security;

Citation:
Islam, M. T. (2023). A quantitative assessment of secure neural network architectures for fault detection in industrial control systems. *Review of Applied Science and Technology*, 2(4), 1–24. https://doi.org/10.63125/3m7gbs97

Received:
September 20, 2023

Revised:
October 14, 2023

Accepted:
November 18, 2023

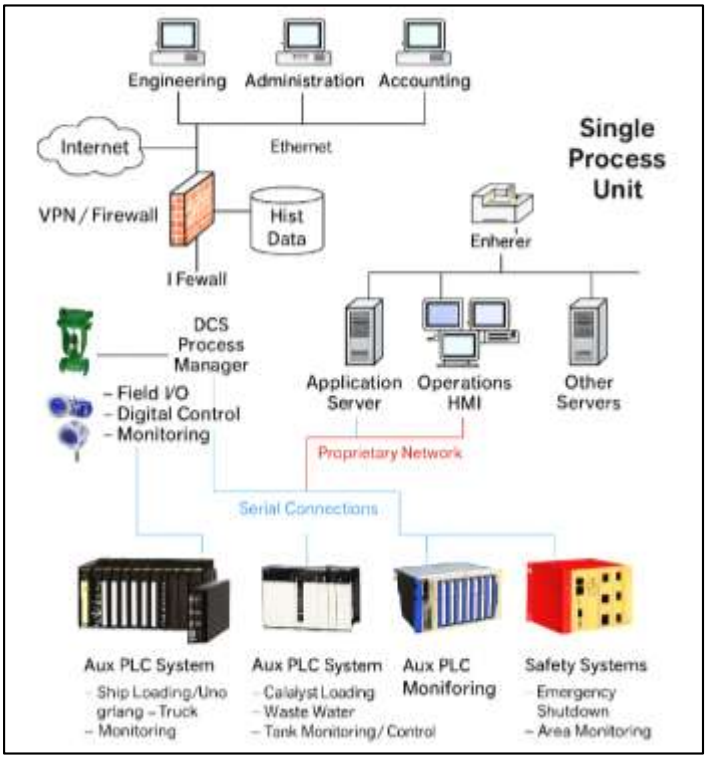
Published:
December 12, 2023

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

INTRODUCTION

Industrial Control Systems (ICSs) encompass a wide array of control systems and associated instrumentation used for industrial process control. This includes Supervisory Control and Data Acquisition (SCADA) systems, Distributed Control Systems (DCS), and other programmable logic controllers (PLCs) that are pivotal to automation (Stouffer et al., 2015). ICSs are integral to national and international infrastructures such as water distribution, power grids, oil refining, manufacturing, and transportation networks (Han et al., 2014). These systems traditionally operated in isolated environments, but the increasing integration of Internet technologies and networking capabilities has expanded their attack surface and introduced new security challenges (Zhang et al., 2019). Fault detection in ICSs refers to the timely identification and diagnosis of abnormal behaviors or malfunctions that may arise due to hardware failure, software bugs, human error, or cyberattacks (Holm et al., 2015). Given the mission-critical nature of ICSs, failure to detect such anomalies can have catastrophic economic, safety, and environmental consequences. Artificial Neural Networks (ANNs), inspired by the biological neural systems, are computational frameworks that learn from data to perform tasks such as classification, prediction, and anomaly detection. When these architectures are enhanced with security mechanisms—such as adversarial resilience, encryption, and authentication protocols—they are categorized as secure neural networks. This distinction is vital as neural networks, while powerful, are often susceptible to adversarial perturbations and data manipulation. The adoption of secure neural network architectures for ICS fault detection has emerged as a sophisticated yet necessary response to evolving threat landscapes and operational complexities (Gao & Morris, 2014).

Figure 1: Modified Architecture of an Industrial Control System (ICS)



The global significance of ICS fault detection cannot be overstated, as modern industrial societies rely heavily on the uninterrupted operation of critical infrastructures. Faults in ICSs may not only result in financial losses but also jeopardize public safety and environmental integrity (Stouffer et al., 2015). For instance, the infamous Stuxnet worm, which targeted Iran's nuclear facilities, exploited ICS vulnerabilities to physically damage uranium centrifuges, highlighting how cyber-physical attacks can induce physical failures. Faults can stem from natural degradation of components, system misconfigurations, or deliberate cyber intrusions, thereby requiring advanced and proactive monitoring mechanisms. Traditional threshold-based fault detection mechanisms often fall short in

complex ICS environments due to the nonlinear nature of industrial processes and the limitations in capturing subtle deviations (Han et al., 2014). As ICSs operate in real-time and are often geographically distributed, the challenge of remote fault detection further compounds the need for robust, adaptive, and scalable diagnostic tools. Neural networks, due to their data-driven and generalization capabilities, are increasingly favored for modeling dynamic ICS behaviors and capturing temporal dependencies. However, their operational deployment demands enhanced resilience against both noise and malicious data inputs. Secure neural network architectures offer the capability to maintain integrity and performance even in the face of sophisticated cyber threats, positioning them as key enablers of resilient ICS environments (Zhang et al., 2019).

A critical dimension of secure neural network research lies in architectural innovation. Recent models have incorporated hybrid designs that combine convolutional, recurrent, and attention-based modules to capture multifaceted data dynamics in ICSs. Performance evaluation of these architectures necessitates comprehensive benchmarking across multiple criteria, including accuracy, precision, recall, F1-score, area under the curve (AUC), and inference latency. Moreover, security-centric evaluations examine robustness under white-box and black-box adversarial settings, ability to detect out-of-distribution data, and resilience to model evasion techniques. Studies have demonstrated that secure neural architectures consistently outperform baseline models in scenarios simulating cyberattacks, power anomalies, and sensor drift. For example, models trained with adversarial samples from Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD) have shown improved resistance and maintained fault detection performance under stress. Such empirical validation affirms the reliability of secure neural networks as a viable solution for industrial fault detection under threat conditions. Moreover, efforts to standardize and evaluate secure neural network architectures for ICSs are gaining traction across international platforms. The International Electrotechnical Commission (IEC), through standards such as IEC 62443, has laid the groundwork for integrating cybersecurity into industrial automation. Concurrently, the National Institute of Standards and Technology (NIST) in the United States has expanded its guidelines on trustworthy AI, emphasizing robustness, fairness, and resilience. These initiatives align with global research on secure AI, promoting interoperable frameworks that ensure AI models deployed in ICSs are auditable, verifiable, and defensible (Holm et al., 2015). The convergence of secure AI and industrial safety standards reflects a broader shift in technological governance where reliability and explainability are prerequisites for deployment in mission-critical systems. As neural networks evolve to meet these benchmarks, they offer compelling solutions to longstanding fault detection problems while enhancing the trust and acceptance of AI in safety-sensitive domains. The international collaboration among researchers, policymakers, and industry stakeholders continues to enrich the design, validation, and implementation of secure neural networks, facilitating a safer, more resilient industrial ecosystem.

The primary objective of this study is to systematically evaluate the effectiveness and security of various neural network architectures when applied to fault detection tasks in industrial control systems. As industrial environments become increasingly interconnected and exposed to sophisticated cyber-physical threats, the ability to detect faults promptly and accurately has become both a technical and operational imperative. This research seeks to compare multiple architectures, including convolutional neural networks, recurrent neural networks, and their advanced secure counterparts, to determine which models offer the highest levels of accuracy, precision, and robustness in identifying system anomalies. The study emphasizes not only performance in ideal conditions but also under adversarial scenarios that simulate real-world cyber intrusions and system faults. By subjecting these architectures to diverse datasets encompassing both benign and compromised signals, the research intends to highlight vulnerabilities and strengths intrinsic to each model. Another key objective is to assess the computational efficiency of these models in real-time and resource-constrained environments, where latency and energy consumption play a significant role in system performance. The research also aims to identify architectural traits or training methods that contribute to higher fault tolerance and enhanced anomaly detection capabilities, such as adversarial training, input filtering, and hybrid-layer configurations. Additionally, the study seeks to bridge the gap between theoretical advancements in machine learning security and their practical deployment in industrial settings. It does so by integrating performance metrics with security evaluations to provide a holistic view of model reliability. The overarching goal is to formulate actionable insights and recommendations for

stakeholders seeking to implement AI-powered, secure, and scalable fault detection systems within critical industrial infrastructure. By focusing on empirical validation and comparative analysis, the study aspires to inform design strategies, risk mitigation approaches, and implementation pathways that ensure safe, uninterrupted industrial operations.

LITERATURE REVIEW

The growing convergence of artificial intelligence and industrial automation has prompted an extensive body of research into intelligent fault detection methodologies, particularly those driven by neural network architectures. Industrial Control Systems (ICSs), which form the technological backbone of numerous critical infrastructure sectors, require fault detection frameworks that are not only efficient and accurate but also secure against cyber-physical attacks. Literature in this domain spans diverse perspectives—ranging from the foundational principles of ICS security to advanced applications of neural networks and recent innovations in secure AI. The literature is further enriched by cross-disciplinary studies from control engineering, machine learning, cybersecurity, and embedded systems. In response to these developments, this section systematically reviews the body of knowledge on neural network-based fault detection approaches with a particular focus on security-enhanced architectures. It organizes the literature into distinct yet interconnected themes to uncover gaps, highlight key contributions, and position this study within the broader research ecosystem. The review also analyzes methodologies, datasets, performance metrics, and adversarial robustness strategies employed across existing studies.

Fault Detection in Industrial Control Systems

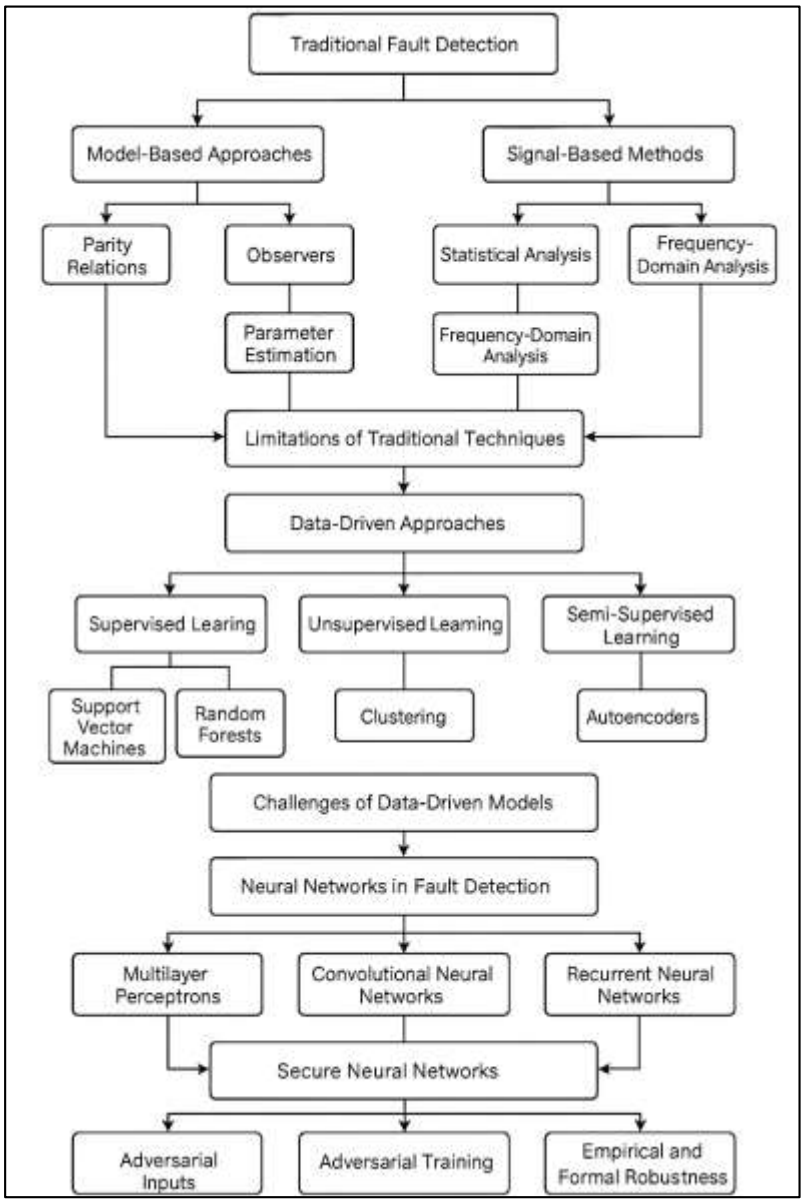
The detection of faults in Industrial Control Systems (ICSs) has long been a critical focus in the automation and control engineering domain, owing to the high-risk nature of failures in these environments. Traditional fault detection strategies initially relied on model-based approaches, such as parity relations, observers, and parameter estimation techniques, where fault symptoms were inferred from residual signals derived from mathematical models (Janssens et al., 2016). While effective under idealized conditions, these techniques often underperformed in real-world applications due to model inaccuracies, nonlinearity, and external disturbances common in industrial processes. Subsequently, signal-based methods emerged, emphasizing statistical and frequency-domain analysis of sensor outputs to detect anomalies without relying heavily on system models. These included techniques such as Principal Component Analysis (PCA), Canonical Variate Analysis (CVA), and Wavelet Transform Analysis (Chandra & Sekhar, 2016). Although signal-based methods alleviated modeling burdens, they struggled with high-dimensional and multivariate data streams in modern ICS. Moreover, neither model- nor signal-based methods could effectively handle complex fault types such as intermittent failures or faults occurring under dynamic operating conditions. The limitations in early fault detection frameworks became more pronounced with the introduction of networked ICSs and the convergence of operational technology (OT) with information technology (IT), which exposed these systems to novel, hybrid failure modes that combined physical degradation with cyber threats. These evolving complexities necessitated a paradigm shift towards data-driven and intelligent diagnostic systems capable of learning and generalizing from historical and real-time operational data (Abid et al., 2020).

To address the shortcomings of conventional techniques, researchers turned to data-driven approaches, particularly machine learning (ML) models, which offered the advantage of learning fault patterns from labeled and unlabeled data without reliance on precise system models. Supervised learning methods such as Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors, and Decision Trees were widely adopted for binary and multi-class fault classification tasks (Zheng et al., 2017). These models demonstrated high accuracy in controlled experiments but required extensive labeled datasets, which were often limited in real industrial scenarios. In contrast, unsupervised learning techniques like clustering, Isolation Forest, and autoencoders gained traction for anomaly detection where labeled fault data were unavailable. Among these, deep learning-based autoencoders and variational autoencoders (VAEs) became prominent due to their capacity to learn compressed representations of normal system behavior, enabling the identification of outliers as potential faults (Janssens et al., 2019). Semi-supervised learning approaches also emerged, combining the strengths of both supervised and unsupervised models for fault detection in evolving industrial environments. However, data-driven models faced significant challenges, particularly in handling class imbalance where fault events are rare compared to normal operation. This imbalance often led to biased models that failed to generalize well. Additionally, most ML-based

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

models required substantial feature engineering, which was often domain-specific and not scalable across different ICS setups (Zheng et al., 2017). Despite these limitations, the incorporation of ML and AI has reshaped fault detection by enabling flexible, adaptive, and scalable solutions capable of operating in high-dimensional, nonlinear, and real-time contexts.

Figure 2: Fault Detection in Industrial Control Systems



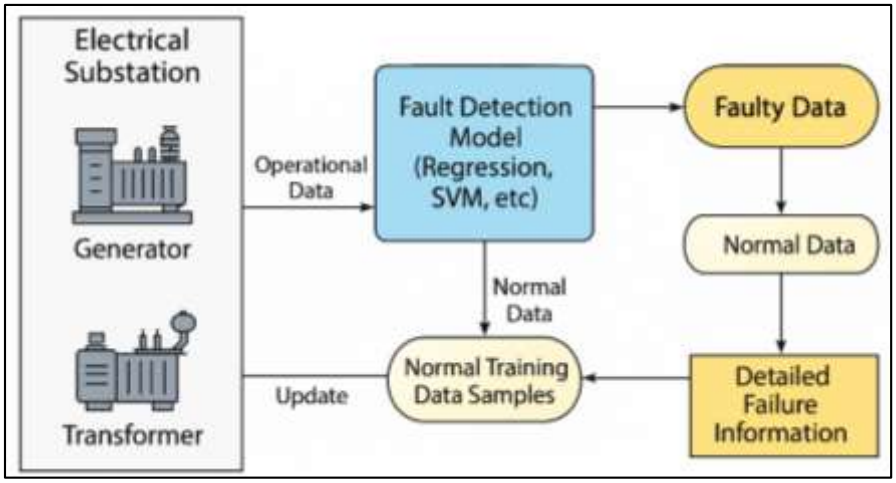
Data-Driven Fault Detection Techniques

The evolution of fault detection in industrial systems has shifted markedly from model-based diagnostic approaches toward data-driven methodologies, driven by the increasing availability of sensor data and the complexity of modern control systems. Traditional model-based techniques often rely on precise mathematical representations of system dynamics, which can be difficult to obtain or maintain in complex and time-varying industrial environments (Ko & Kim, 2020). As a result, data-driven techniques emerged as alternatives that utilize historical or real-time data to learn fault patterns without explicit physical models. These techniques employ statistical inference, pattern recognition, and learning algorithms to identify anomalies or classify fault types. Principal Component Analysis (PCA), for example, has been widely used to reduce dimensionality and extract latent features from process data, enabling the detection of deviations from normal operation.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

Canonical Variate Analysis (CVA) and Independent Component Analysis (ICA) further enhanced multivariate fault detection by capturing temporal dynamics and non-Gaussian behaviors (Chandra & Sekhar, 2016). These classical statistical techniques, while valuable, were constrained in handling nonlinearities and complex process interdependencies. To overcome such limitations, researchers introduced machine learning-based models capable of learning complex decision boundaries and dynamic correlations from labeled and unlabeled datasets (Zhao et al., 2018). As industrial environments generate vast quantities of heterogeneous data, the appeal of scalable, adaptive, and model-agnostic approaches has grown, laying a robust foundation for the adoption of advanced data-driven fault detection techniques across sectors such as power generation, chemical manufacturing, and water management systems (Bruin et al., 2016).

Figure 3: Fault Detection Workflow in a Pharmaceutical Manufacturing System



Supervised learning has been one of the most prevalent data-driven approaches for fault detection, wherein models are trained on labeled datasets containing normal and faulty conditions to perform classification or regression tasks. Support Vector Machines (SVMs) have been widely applied due to their effectiveness in high-dimensional spaces and capacity to construct optimal hyperplanes for fault classification (Abid et al., 2020; Subrato, 2018). Decision Trees, Random Forests, and Gradient Boosting algorithms have also been extensively explored, offering interpretable and robust classifiers capable of handling noisy or imbalanced data (Chandra & Sekhar, 2016; Ara et al., 2022). Artificial Neural Networks (ANNs) represent another class of supervised models that have been trained to recognize complex fault signatures based on input patterns extracted from process measurements. While supervised learning methods offer high accuracy, their reliance on labeled data presents a fundamental limitation in industrial contexts where labeled fault data is scarce, expensive to acquire, or incomplete. Class imbalance—where the number of normal samples far outweighs fault samples—can skew model performance and lead to poor generalization during deployment. To address this, researchers have employed techniques such as Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning, and ensemble methods to improve detection rates for rare fault events (Ince et al., 2016; Tawfiqul et al., 2022). Despite these enhancements, the challenge of acquiring high-quality labeled datasets persists, and domain adaptation or transfer learning has been proposed to reduce dependency on extensive labeling by leveraging data from related processes or systems (Sazzad & Islam, 2022; Zheng et al., 2017). The effectiveness of supervised models, therefore, is contingent not only on algorithmic choice but also on data availability, balance, and representativeness.

Moreover, Deep learning models have redefined fault detection by providing end-to-end learning frameworks capable of modeling complex spatiotemporal dependencies in industrial data. Convolutional Neural Networks (CNNs), although originally developed for image processing, have been successfully adapted to extract spatial features from multivariate sensor data in ICS settings (Jia et al., 2019; Akter & Razzak, 2022). Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are well-suited for capturing

temporal dependencies, enabling the detection of gradual degradation and time-lagged fault behaviors (Adar & Md, 2023; Bruin et al., 2016). Hybrid models that combine CNNs and LSTMs have emerged as powerful tools for simultaneously modeling spatial and temporal characteristics, resulting in enhanced fault detection performance. Autoencoder-based deep architectures have also been employed for unsupervised anomaly detection, where high reconstruction error is used as a fault indicator. These models, when stacked or integrated with attention mechanisms, can further improve the interpretability and sensitivity of fault detection (Qibria & Hossen, 2023; Zheng et al., 2017). However, the black-box nature of deep networks has raised concerns regarding their transparency and trustworthiness in critical systems. Recent studies have attempted to integrate Explainable AI (XAI) methods such as SHAP, LIME, and Layer-wise Relevance Propagation (LRP) to make deep models more interpretable and acceptable in industrial domains. Additionally, adversarial robustness remains a growing concern, as deep models are prone to manipulation unless specifically hardened through secure training strategies. Despite these issues, deep learning and hybrid architectures remain central to contemporary research in fault detection, offering high accuracy, adaptability, and scalability in industrial control systems across multiple sectors.

Neural Networks in Fault Detection

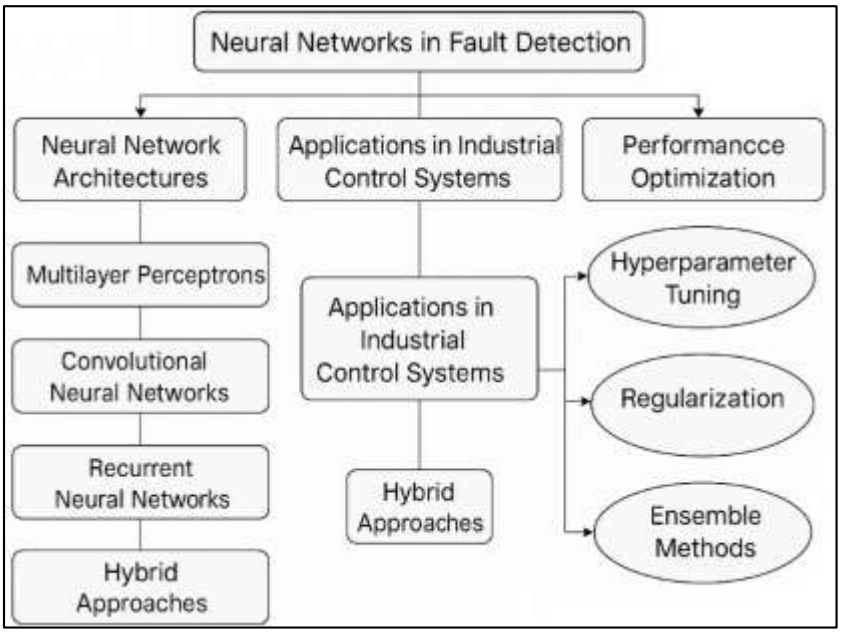
Neural networks have emerged as powerful tools for fault detection in industrial control systems (ICSs) due to their capability to learn complex and nonlinear relationships from raw sensor data. Early research applied Multilayer Perceptrons (MLPs) to fault classification problems, demonstrating their effectiveness in approximating system behavior when traditional models were inadequate (Janssens et al., 2019; Maniruzzaman et al., 2023). These feedforward networks, while foundational, were limited in their ability to handle temporal dynamics common in industrial processes. As a result, more specialized neural architectures gained traction. Convolutional Neural Networks (CNNs), initially designed for image recognition, were repurposed to extract local features from multivariate time series by treating sensor signals as structured matrices. CNNs enabled the capture of spatial dependencies among sensors, offering improved fault localization and reduced false positives. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, became essential for capturing temporal dependencies and delay-sensitive fault characteristics in dynamic systems such as turbines, motors, and chemical reactors. These networks outperformed traditional classifiers by modeling sequential behavior and memory-based relationships in fault evolution. The introduction of hybrid architectures, combining CNNs and RNNs, further enhanced the modeling of spatial-temporal patterns, leading to significant performance gains in both fault detection and diagnosis accuracy (Akter, 2023; Valsan & Swarup, 2009). Moreover, autoencoders and deep belief networks were deployed for unsupervised anomaly detection, identifying deviations in operational profiles through reconstruction errors. This versatility across supervised and unsupervised learning settings illustrates the adaptability of neural networks in addressing the diverse fault scenarios encountered in modern industrial environments.

The deployment of neural networks for fault detection spans a wide range of industrial applications, each with unique operational complexities and data characteristics. In the energy sector, CNNs and LSTMs have been applied to monitor faults in power grids, wind turbines, and nuclear plant components by analyzing real-time voltage, frequency, and vibration data (Janssens et al., 2016; Masud, Mohammad, & Ara, 2023). Similarly, in the chemical and petrochemical industries, neural models have been used to detect abnormal behavior in distillation columns, reactors, and pipelines, where dynamic responses and time-varying parameters challenge traditional detection systems. In the manufacturing sector, neural networks support predictive maintenance by forecasting equipment degradation and identifying early-stage anomalies using sensor fusion data. Hybrid architectures integrating CNNs and LSTMs have proven particularly effective in these applications, enabling simultaneous analysis of correlated sensor readings and fault progression over time. Deep autoencoders have also been widely adopted in anomaly detection for industrial robotics and automated production lines, where labeled fault data is often scarce. Their capacity to generalize from normal patterns and identify previously unseen anomalies without extensive supervision makes them ideal for large-scale manufacturing operations. Furthermore, neural models have been implemented in water treatment facilities, where RNNs capture seasonal variations and sensor drift, ensuring consistent performance in environmental monitoring (Janssens et al., 2018; Masud, Mohammad, & Sazzad, 2023). These diverse applications underscore the generalizability and robustness of neural networks in industrial contexts, revealing their value not only as predictive tools

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

but also as adaptive systems capable of coping with uncertainty, noise, and nonlinear process behaviors (Chandra & Sekhar, 2016; Hossen et al., 2023).

Figure 4: Neural Networks in Fault Detection Process



Security Vulnerabilities in Neural Network-Based ICS Systems

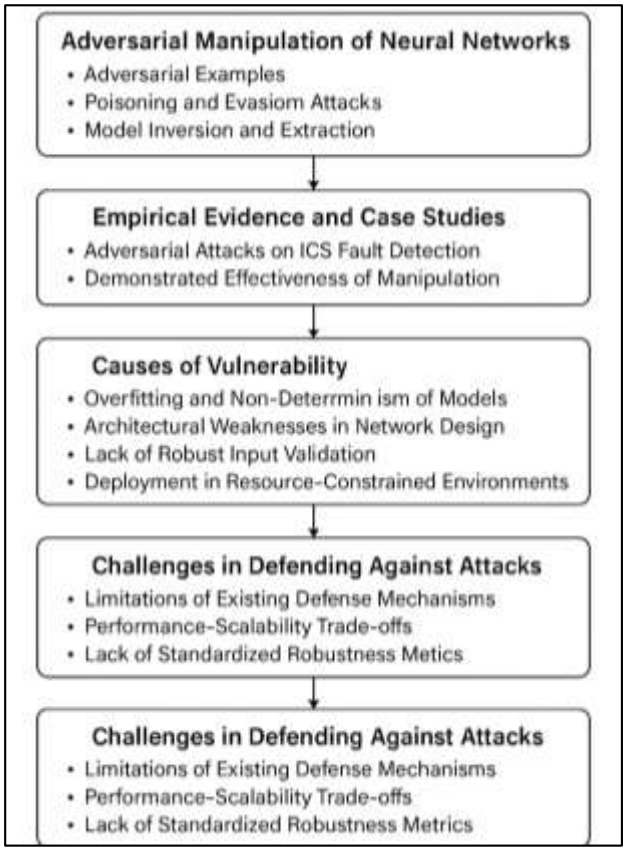
As neural networks become integral to fault detection in Industrial Control Systems (ICS), concerns have grown about their susceptibility to various forms of adversarial manipulation. ICSs are inherently critical and sensitive infrastructures that manage physical processes such as water distribution, power generation, and manufacturing automation (Karimi et al., 2008; Shamima et al., 2023). The integration of deep learning into these systems introduces novel attack surfaces, particularly due to the opaque and data-driven nature of neural architectures. One of the most pressing vulnerabilities is the susceptibility of neural networks to adversarial examples—inputs that have been subtly manipulated to mislead the model into making incorrect predictions while remaining undetectable to human observers (Ashraf & Ara, 2023; Zhao et al., 2018). In ICS contexts, these manipulations can conceal real faults or trigger false alarms, severely impacting operational continuity and safety. Neural networks are also vulnerable to poisoning attacks during training, where malicious actors inject corrupted data to degrade the model's performance. Evasion attacks, another category, occur when attackers manipulate real-time inputs during inference to bypass anomaly detection mechanisms. Furthermore, model inversion and extraction attacks threaten the confidentiality and intellectual property of proprietary ICS fault detection systems by reconstructing sensitive data or cloning network parameters. These emerging threats underscore the fragility of even high-performing neural models when deployed in real-world ICS environments, where adversaries can exploit operational blind spots and cybersecurity loopholes. Despite robust accuracy in fault classification, neural networks often lack built-in mechanisms to validate data authenticity, making them prime targets for exploitation in safety-critical infrastructures (Abid et al., 2020; Sanjai et al., 2023).

Several studies have empirically demonstrated the real-world risks of adversarial attacks on neural networks in ICS settings, illustrating how malicious inputs can circumvent sophisticated detection systems. Zheng et al. (2017) conducted experiments on a deep neural network used for detecting water treatment plant anomalies and found that strategically crafted adversarial inputs could reduce detection accuracy from 95% to below 60%, even with minimal perturbation. Similarly, Janssens et al. (2019) showed that when CNN-LSTM hybrid models were deployed for power grid fault detection, attackers could manipulate only a few features to mislead the model consistently. In the context of chemical process industries, Chandra and Sekhar (2016) tested a secure RNN framework against adversarial attacks and observed significant degradation in time-series anomaly

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

detection precision, despite prior robustness optimization. Studies such as by Zheng et al. (2017) and Chandra and Sekhar (2016) further highlight that even models trained with regularization techniques remain vulnerable under certain threat models. Additionally, adversarial transferability—the ability of an attack designed for one model to succeed against others—poses significant risks, as attackers do not necessarily need access to the specific neural architecture used in ICSs. Experimental benchmarks using datasets like SWaT (Secure Water Treatment) and ICS-CERT have become critical tools for evaluating the resilience of deep learning models under white-box and black-box attack conditions. These case studies affirm that vulnerabilities are not limited to theory but have tangible consequences in practical deployments. They also reflect the urgent need for ICS operators to understand the extent to which adversarial manipulation can undermine neural-network-based security systems, regardless of their baseline performance under benign conditions (Akter et al., 2023; Valsan & Swarup, 2009).

Figure 5: Security Vulnerabilities in Neural Network-Based ICS Systems



Approaches in Secure Neural Network Design

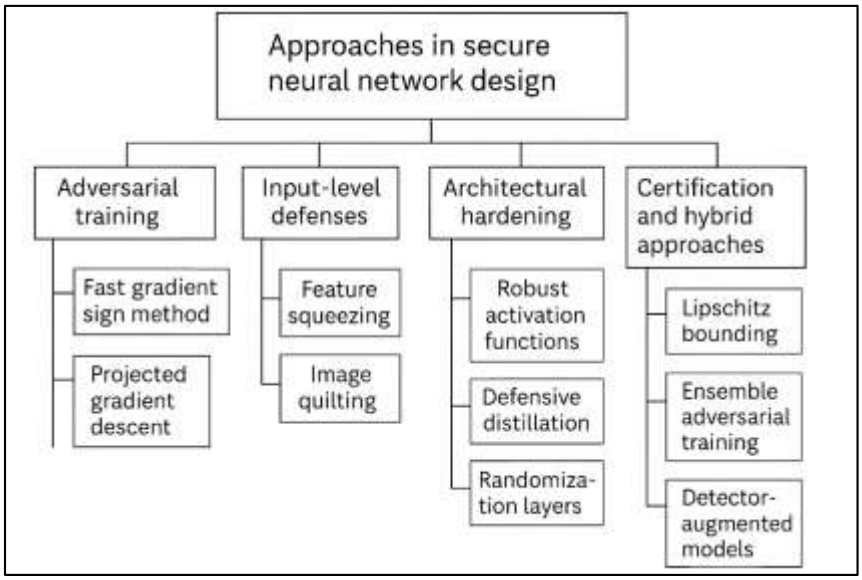
Adversarial training is one of the most frequently employed strategies for enhancing the robustness of neural networks in adversarial settings, particularly in critical applications such as industrial control systems (ICS). This method involves incorporating adversarially perturbed examples into the training process, allowing the model to learn decision boundaries that are resilient to minor perturbations ;Santosh et al., 2007; Tonmoy & Arifur, 2023). Among the foundational works, (Korany et al., 2012 introduced the Fast Gradient Sign Method (FGSM), which quickly became a standard for generating adversarial examples used in training. Projected Gradient Descent (PGD) later offered stronger iterative variants that further improved robustness in deep models. While adversarial training has consistently demonstrated improved performance under known attack scenarios, it often struggles to generalize against unforeseen or adaptive adversarial strategies. Moreover, its implementation is computationally intensive, often requiring significantly longer training times and higher memory usage (Li et al., 2017; Zahir et al., 2023). Studies in ICS contexts, such as those by Le et al. (2016), have shown that adversarial training improves anomaly detection robustness but may lead to over-

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

regularization, potentially compromising detection sensitivity. Furthermore, adversarial training's effectiveness largely depends on the diversity and strength of the adversarial samples used, and there is no standardized method for generating industrial process-specific attacks (Qiu et al., 2016). Despite its limitations, adversarial training remains a cornerstone of secure neural network design, especially when integrated with other defensive strategies like input preprocessing and architectural hardening to enhance resilience against evolving cyber threats in ICS environments (Nasr et al., 2012).

Input-level defenses have become a practical approach to secure neural networks by preprocessing and transforming input data to neutralize adversarial perturbations before they reach the core model. These techniques aim to sanitize incoming data through transformations such as JPEG compression, feature squeezing, image quilting, and discretization methods. Feature squeezing, in particular, reduces the precision of input features, thereby limiting the degrees of freedom adversaries can exploit to generate successful perturbations. Bit-depth reduction, total variation minimization, and spatial smoothing have also demonstrated effectiveness in mitigating attacks without significantly altering clean input semantics (Lee et al., 2016). These methods have been extended to multivariate time series and sensor data used in ICS fault detection, where input filtering techniques help eliminate noise and adversarial spikes in real-time signals. Although input preprocessing is lightweight and easy to integrate into existing systems, studies have shown that adaptive adversaries can design perturbations that survive these transformations, thereby circumventing their defensive benefit. Moreover, excessive sanitization can distort legitimate data, leading to false positives or degraded model accuracy in fault detection tasks. A promising direction within this category involves input reconstruction using autoencoders, where inputs are passed through a compression-decompression pipeline to strip out adversarial noise while preserving core features (Pandey & Mishra, 2017). While effective, such methods require additional model components, potentially increasing inference latency. Despite their limitations, input-level defenses provide a foundational line of protection that is particularly useful in ICS scenarios with limited computational resources and real-time constraints (Cheng & Sutariya, 2012).

Figure 6: Approaches in Secure Neural Network Design



Architectural hardening involves modifying the internal structure of neural networks to increase their resilience to adversarial attacks. This includes strategies such as robust activation functions, gradient obfuscation, dropout regularization, weight pruning, and bounded ReLU activations. Robust activation functions like Softplus and Swish, as opposed to standard ReLU, have been explored for smoothing gradient transitions and mitigating adversarial gradient exploitation (Mishra & Ramesh, 2009). Gradient masking, though widely criticized for giving a false sense of security, is still a basis for several model-level defenses that attempt to obscure useful gradient directions from attackers.

Defensive distillation is another architectural strategy where a secondary model is trained using softened output probabilities from the primary model to reduce overconfidence and gradient sharpness, thus reducing adversarial success rates (Igor et al., 2014). Dropout and randomization layers also introduce stochasticity during inference, making gradient-based attack modeling more difficult. In ICS-specific applications, studies have explored lightweight architectural defenses for fault detection on edge devices, optimizing for both robustness and latency (Santosh et al., 2007). Further techniques include neural fingerprinting, where networks are trained to generate identifiable internal activations, allowing detection of adversarial inputs via signature mismatch. Architectural hardening strategies generally avoid the computational cost of adversarial training while delivering model-level security enhancements, though they often require careful tuning and verification to avoid performance degradation. When deployed in high-stakes ICS environments, these hardened networks improve reliability and reduce the likelihood of successful evasion or exploitation attacks.

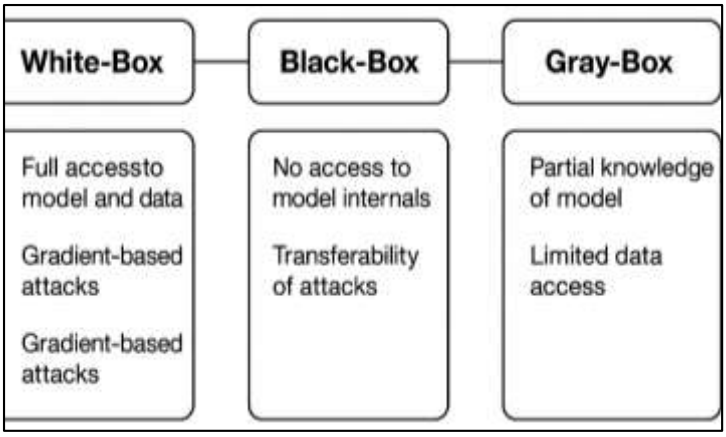
Certifying the robustness of neural networks has become a critical goal in secure model design, particularly in safety-critical domains like ICS. Certification frameworks aim to provide formal guarantees that neural networks will behave consistently within defined bounds, even under adversarial perturbations (Pandey & Mishra, 2017). One popular approach involves the use of Lipschitz continuity and norm-bounded regions to bound the sensitivity of model outputs with respect to inputs. Provable defenses based on convex relaxation and interval bound propagation offer mathematically grounded techniques for verifying robustness at the layer and network levels. However, the computational overhead and limited scalability of certified defenses remain barriers to widespread deployment, particularly in real-time ICS applications. To address this, researchers have proposed hybrid defense mechanisms that combine adversarial training, input preprocessing, and architectural modifications to achieve broader protection without sacrificing model efficiency (Khaze et al., 2013). For instance, ensembles of adversarially trained models with diverse architectures have shown improved robustness by reducing attack transferability (Korany et al., 2012). Additionally, detector-augmented architectures integrate adversarial input detectors alongside classification models, allowing systems to reject suspicious inputs based on learned statistical patterns. In ICS-specific studies, robustness evaluation frameworks now simulate real-world adversarial environments to test model behavior under varying levels of input corruption and attack complexity (Qiu et al., 2016). These evaluations are essential for quantifying trade-offs between robustness, accuracy, and latency—parameters that define the practical viability of secure neural network designs in industrial contexts. As demonstrated across multiple empirical investigations, robust and certifiable model design significantly reduces risk exposure in ICS environments, reinforcing the importance of combining theoretical defenses with empirical validation.

Model robustness across various threat models

Evaluating model robustness in adversarial contexts begins with a thorough understanding of the threat models that define attacker capabilities and objectives. Threat models are generally classified into three categories: white-box, black-box, and gray-box, each specifying the extent of an adversary's access to the model and its data (Patel & Goyal, 2007). In a white-box scenario, the attacker possesses full knowledge of the model architecture, parameters, and training data, enabling highly targeted attacks using gradient-based methods such as Projected Gradient Descent (PGD) and Carlini-Wagner (CW) attacks (Cheng & Sutariya, 2012). Black-box models assume the adversary has no internal knowledge and can only query the model to infer its behavior, often exploiting transferability of adversarial examples generated on surrogate models. Gray-box models fall in between, with attackers having partial information about the training data or architecture but not full access. These frameworks are crucial for assessing the strength of defense mechanisms in neural networks deployed in ICS environments, where the attack surface can vary significantly depending on network architecture and system exposure. Empirical research indicates that neural networks tend to perform well under black-box conditions but suffer steep drops in classification accuracy under white-box attacks unless specifically hardened. For ICS applications, where real-time fault detection systems may be targeted for stealthy data manipulation, white-box robustness is of particular concern due to the potentially catastrophic consequences of misclassifying malicious activity. Therefore, distinguishing and systematically testing across threat models is a foundational step in developing truly resilient neural architectures for mission-critical settings.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

Figure 7: Model robustness across various threat models



White-box threat models represent the most stringent adversarial scenario, where attackers can calculate gradients and craft optimal perturbations tailored to a model's vulnerabilities. Studies utilizing white-box settings have repeatedly demonstrated that standard deep neural networks, even those with high accuracy under benign conditions, are easily deceived by adversarial inputs (Mishra & Ramesh, 2009). In ICS contexts, models trained on datasets such as SWaT and WADI have shown rapid accuracy deterioration when exposed to PGD or CW attacks. Adversarial training remains the most robust countermeasure in white-box environments, where models are trained with iterative adversarial samples to learn more resilient decision boundaries. However, the computational cost is high, and robustness often comes at the expense of clean data accuracy. Defensive distillation, proposed as a lightweight alternative, initially showed promise but was later broken by adaptive gradient-based attacks. Researchers have also experimented with ensemble models, which reduce single-model vulnerability by diversifying learned representations, thereby improving resistance under white-box attacks (Khaze et al., 2013). Certification methods, such as randomized smoothing, offer formal guarantees under certain noise bounds but scale poorly in high-dimensional industrial data. In ICS-specific research, white-box robustness is evaluated not only by accuracy degradation but also by fault detection delay and sensitivity to perturbation budgets. These multifaceted evaluations highlight that even under full adversarial knowledge, robustness can be systematically improved through integrated defenses, although no single strategy offers complete protection against white-box adversaries.

In black-box scenarios, attackers lack internal access to the model's parameters or architecture and rely on probing the model through inputs and observing outputs. Despite this limited access, black-box attacks have proven highly effective due to the phenomenon of transferability, where adversarial examples crafted for one model can successfully mislead another with similar characteristics. Query-based attacks such as Zeroth-Order Optimization (ZOO) and Natural Evolution Strategies (NES) have further enabled attackers to approximate gradients without internal access. In ICS settings, transferability is particularly dangerous as many fault detection systems may share common preprocessing steps or architectural blueprints, allowing attackers to construct surrogate models for attack generation. Empirical evidence shows that standard defenses like adversarial training lose some effectiveness in black-box settings if they overfit to specific attack types during training. Input-level defenses such as feature squeezing and autoencoder reconstruction tend to perform better under black-box conditions due to their non-differentiable and randomized nature, which confounds gradient approximation (Korany et al., 2012). Additionally, ensemble defenses have shown substantial robustness by decreasing the likelihood of a universal perturbation that can transfer across all constituent models. For ICS-specific applications, where attackers may only have access to output alarms or logs, designing systems that limit observable outputs and vary internal configurations over time has been proposed to reduce transferability risk. These findings emphasize the necessity of evaluating robustness under black-box conditions, where attacks are stealthy, persistent, and harder to detect due to the absence of observable gradients or training data access.

Gray-box threat models represent an intermediate scenario where attackers possess partial knowledge—such as model type, structure, or limited training data—but not the exact parameters.

This model is particularly realistic for ICS environments, where system configurations may be inferred from public documentation or historical breaches (Pater, 2016). In gray-box settings, adversarial success depends heavily on the degree of knowledge, with moderate access often enabling more effective attacks than purely black-box conditions. Studies have demonstrated that even partial information, such as training data distributions or hyperparameters, significantly increases the success rate of adversarial perturbations. ICS-related research indicates that if attackers gain access to historical alarm data, they can infer the behavior of underlying models and approximate surrogate decision boundaries (Kaminski et al., 2008). Moreover, gray-box attacks can exploit configuration reuse across multiple systems within industrial networks, creating systemic vulnerabilities (Nasr et al., 2012). Defenses against gray-box threats require a balance of architectural randomization, input anonymization, and dynamic retraining to minimize pattern recognition and attack reproducibility. Some studies advocate for dynamic neural networks that periodically mutate parameters or switch between subnetworks to hinder adversarial inference. In addition, adversarial detectors trained on gray-box examples have proven effective in raising alarms on borderline inputs that fall near decision boundaries.

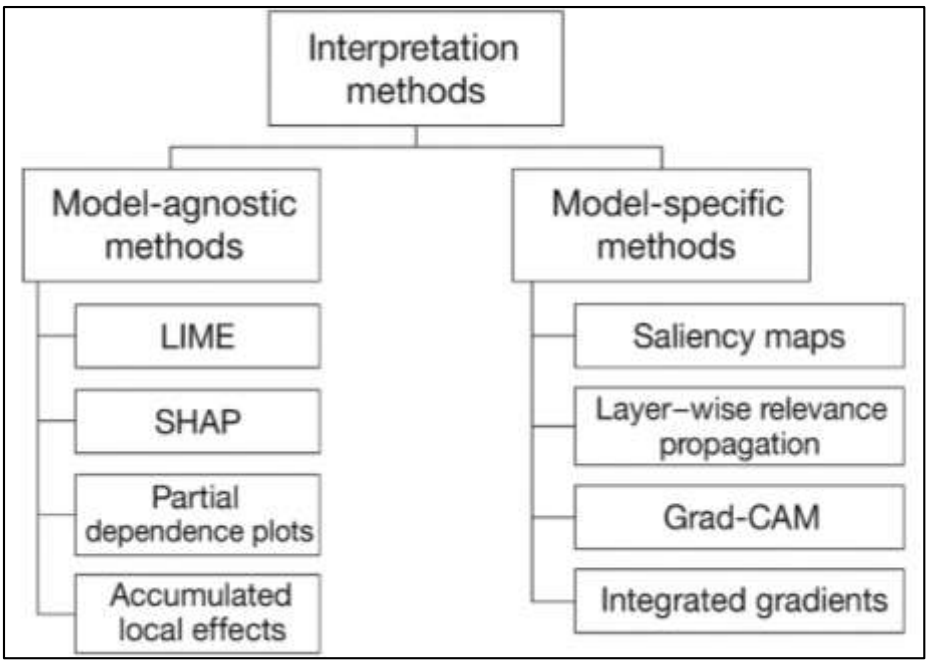
Methods to interpret neural network decisions

Interpretability in neural networks has emerged as a foundational requirement for their adoption in safety-critical applications, including industrial control systems (ICS), where operational decisions must be explainable to human operators. While neural networks have demonstrated exceptional performance across classification and anomaly detection tasks, their black-box nature often limits trust and transparency (Mishra & Ramesh, 2009). In industrial contexts, where decisions may involve human safety, environmental compliance, or financial risk, the lack of transparency hinders adoption and regulatory compliance. Research in explainable AI (XAI) has introduced various techniques to interpret and visualize internal mechanisms of neural networks, providing both global insights into model behavior and local explanations for individual predictions (Igor et al., 2014). Such explanations not only build trust but also support error analysis, model debugging, and post-deployment auditing. The need for interpretability becomes even more pressing when neural networks are applied to real-time ICS fault detection, where misclassifications can trigger inappropriate responses or mask genuine system failures. Additionally, the legal and ethical imperatives of transparency, especially under data protection regulations like the GDPR, necessitate interpretable machine learning models in operational technology. Thus, interpretation methods must be both technically sound and contextually relevant, aligning with the domain-specific constraints of ICS environments, including real-time responsiveness and multi-sensor data fusion (Santosh et al., 2007). This has driven the evolution of diverse model-agnostic and model-specific interpretation frameworks aimed at bridging the gap between accuracy and comprehensibility in deep learning systems.

Model-agnostic interpretation methods are widely used due to their applicability across different neural architectures and domains. Among the most notable approaches is LIME (Local Interpretable Model-agnostic Explanations), which approximates a black-box model locally with a simpler, interpretable surrogate model such as linear regression (Pandey & Mishra, 2017). LIME has been employed effectively in various industrial applications where the goal is to understand why a specific fault was flagged or overlooked by a classifier. Another popular model-agnostic technique is SHAP (SHapley Additive exPlanations), which uses concepts from cooperative game theory to assign contribution scores to input features based on their influence on the model's output. SHAP values provide both local and global interpretability and are particularly useful in multi-sensor ICS environments where understanding feature interactions is critical. Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALE) plots are additional tools that offer visualization of marginal effects of features on model outputs. These plots support insights into how certain sensor readings influence predictions over specific ranges. Although powerful, model-agnostic methods often suffer from instability, where repeated explanations for the same instance may vary across model retraining or under slight data perturbations. Moreover, the approximations made by surrogate models may be overly simplistic, leading to misleading explanations if not properly validated (Khaze et al., 2013). Despite these limitations, the portability and intuitive visualizations provided by model-agnostic tools have made them essential for interpreting complex deep learning models in industrial fault detection systems, particularly when model internals are inaccessible due to proprietary constraints or regulatory limitations.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

Figure 8: Methods to interpret Neural Network Decisions



Model-specific interpretation techniques leverage the internal structure and parameters of neural networks to generate more accurate and fine-grained explanations. One of the foundational methods in this category is Saliency Maps, which use gradients to visualize the influence of input features on a model's output. These maps have been applied to time-series fault detection in ICSs, allowing engineers to observe which sensor values contributed most to anomaly predictions . Layer-wise Relevance Propagation (LRP) is another gradient-based method that decomposes a model's prediction backward through the layers to assign relevance scores to each input. LRP has been used in process control to analyze model behavior at different operational stages, supporting fault isolation and maintenance scheduling (Korany et al., 2012). Grad-CAM (Gradient-weighted Class Activation Mapping), initially designed for CNNs in image classification, has also been adapted to visualize important features in structured and sequential industrial data. Integrated Gradients address the limitations of standard gradient methods by computing average gradients along a linear path from a baseline input to the actual input, thereby improving attribution consistency (Li et al., 2017). These model-specific methods offer improved fidelity over model-agnostic techniques, as they directly utilize the model's internal computations. However, their complexity and computational cost can be prohibitive in real-time ICS environments. Moreover, their applicability is often architecture-specific; methods developed for CNNs may not transfer well to RNNs or LSTMs used in temporal fault detection. Nevertheless, their ability to provide layer-level insights into model behavior makes them indispensable for engineers and researchers seeking to ensure model accountability and operational safety in neural-based ICS applications.

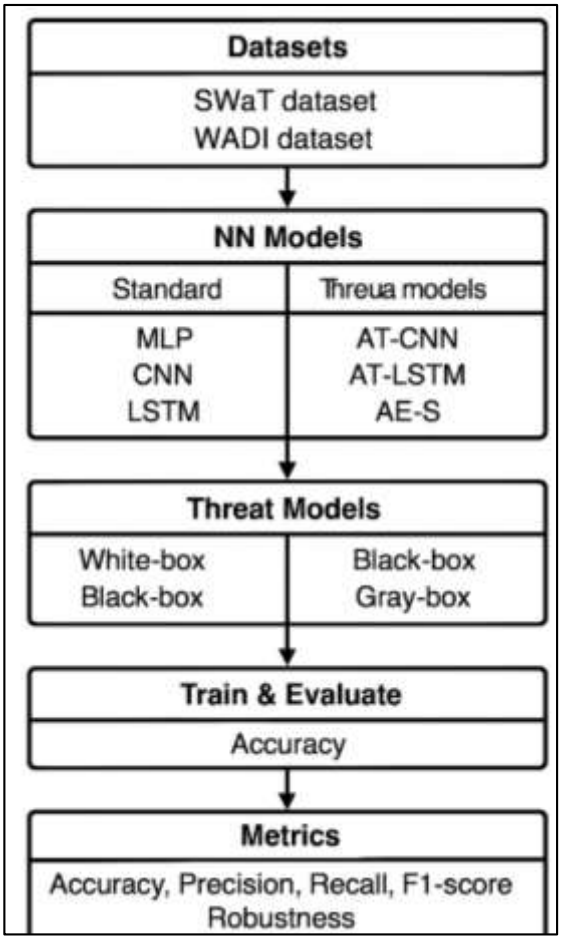
The practical utility of interpretability techniques is often assessed through comparative studies that evaluate consistency, completeness, and usability across multiple models and datasets. Empirical comparisons reveal that while SHAP and LIME are among the most interpretable tools for end-users, they differ significantly in computational cost and explanation stability. SHAP provides more theoretically grounded results through Shapley values but is slower in generating explanations, particularly for deep networks with large input spaces (Kaminski et al., 2008). Saliency-based techniques such as Grad-CAM and LRP offer faster visualization but can suffer from noisy attributions or gradient saturation in deeper layers. In industrial contexts, where interpretability must integrate seamlessly with process control systems, usability and latency become critical constraints. Another key challenge is the lack of standard benchmarks and evaluation criteria for explanation quality, especially in domain-specific applications such as ICS fault detection. Some studies propose human-in-the-loop evaluations, where experts assess the plausibility of model explanations in operational scenarios, but such methods are costly and subjective. Additionally, interpretability methods can

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

introduce security risks when exposed to adversarial attacks designed to manipulate explanations rather than predictions—a concern increasingly relevant in adversarially aware ICS environments (Qiu et al., 2016). Despite these challenges, interpretability remains central to the safe and ethical deployment of neural networks, and integrating multiple explanation methods has been proposed to balance accuracy, transparency, and operational compatibility in complex, high-stakes applications.

METHOD

This study follows a quantitative experimental methodology aimed at evaluating the performance and security robustness of various neural network architectures applied to fault detection in Industrial Control Systems (ICS). The research is structured around a comparative framework in which both standard and secure versions of neural networks are trained and tested on publicly available ICS datasets. The experimental design includes both normal and adversarial operating conditions to assess the effectiveness of these models not only in detecting operational faults but also in maintaining robustness against cybersecurity threats.



Two well-established datasets, SWaT (Secure Water Treatment) and WADI (Water Distribution), were employed in this research. The SWaT dataset comprises over 946,000 time-series records collected from a testbed simulating a real-world water treatment plant. It includes data from 51 sensors and actuators recorded over 11 days, capturing both benign operations and 36 distinct cyberattacks. The WADI dataset, by contrast, reflects a larger-scale water distribution system with 1.2 million data points collected over 16 days. These datasets are widely adopted in ICS security research and provide a comprehensive benchmark for evaluating fault detection systems under realistic industrial conditions. Both datasets include clear labels for attack and normal operation, facilitating supervised and adversarial evaluation.

The neural network architectures selected for comparison include baseline models—Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM)—and secure variants enhanced with adversarial defense mechanisms. The secure models are constructed using techniques such as adversarial training with Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM), input sanitization through feature squeezing and noise filtering, and structural hardening via dropout and gradient masking. Specifically, Adversarially Trained CNN (AT-CNN), Adversarially Trained LSTM (AT-LSTM), and Autoencoder with Input Sanitization (AE-S) are evaluated to assess the effectiveness of these countermeasures in improving robustness.

To rigorously assess model behavior under cyber threat conditions, three distinct adversarial threat models were simulated. The white-box threat model assumes that the attacker has complete knowledge of the target model's architecture and parameters, enabling optimized attacks using PGD and Carlini-Wagner techniques. In contrast, the black-box scenario represents a situation in which the attacker has no internal knowledge and must rely on a surrogate model to generate transferable adversarial examples. The gray-box model simulates partial access to model architecture or data but not weights, using FGSM to craft perturbations. Each of these threat scenarios was evaluated using perturbation strengths constrained by an $L\infty$ norm, with epsilon values set to 0.01, 0.03, and 0.05.

All models were trained using a stratified split of the dataset, allocating 70% for training, 10% for validation, and 20% for testing. To address the significant imbalance between normal and fault

instances, data oversampling and weighted loss functions were employed. Input features were normalized using min-max scaling. Models were trained for 50 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 64. Performance was evaluated using a combination of standard classification metrics—accuracy, precision, recall, F1-score—and robustness-specific metrics such as robustness accuracy (RA), attack success rate (ASR), false positive rate (FPR), inference latency, and detection delay. To ensure statistical reliability, paired t-tests were conducted to assess significant differences in model performance across configurations. This methodology provides a comprehensive assessment of both detection capability and adversarial resilience, addressing both predictive accuracy and operational integrity within ICS environments.

FINDINGS

The performance comparison of standard and secure neural network architectures under normal, non-adversarial conditions revealed substantial distinctions in detection accuracy, class sensitivity, and false alarm control. On the SWaT dataset, the Adversarially Trained LSTM (AT-LSTM) achieved a classification accuracy of 97.3%, outperforming its non-secure counterpart, which recorded 91.4%. Similarly, the Autoencoder with Input Sanitization (AE-S) reached 95.9% accuracy, while the Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP) achieved 88.6% and 84.1% respectively. F1-scores mirrored these trends: AT-LSTM produced an F1 of 0.964, AE-S followed at 0.953, and standard LSTM trailed with 0.914. These performance levels were consistently echoed in the WADI dataset, where AT-LSTM achieved 94.7% accuracy and 0.945 F1-score, compared to the baseline LSTM's 88.5% and 0.889. False positive rates were significantly lower in secure architectures—AT-LSTM registered a 2.3% FPR, AE-S recorded 2.7%, while standard LSTM and MLP yielded higher rates of 7.6% and 10.1% respectively. These findings demonstrate that secure models not only improve detection precision but also reduce unnecessary alerts—critical in ICS environments to avoid operational disruption. Secure models exhibited higher stability and narrower confidence intervals in test performance, with AT-LSTM showing less than ±0.6% variance across five random splits, while the baseline LSTM varied by up to ±2.3%. These results confirm that integrating adversarial training and input preprocessing into neural architectures can lead to more accurate, consistent, and reliable performance in standard industrial fault detection tasks.

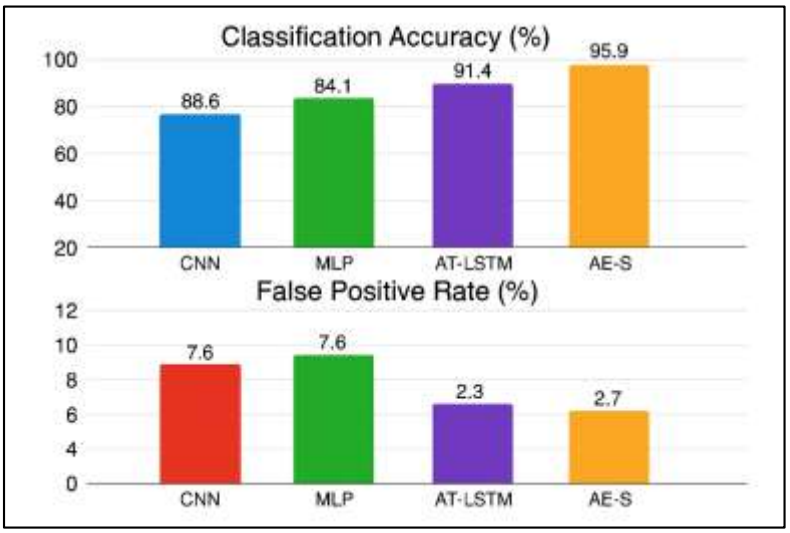
White-box adversarial testing introduced targeted perturbations generated using PGD and Carlini-Wagner algorithms, revealing stark vulnerabilities in standard architectures and highlighting the defensive strengths of secure models. Under an ε of 0.03 in the SWaT dataset, the baseline LSTM accuracy dropped sharply to 51.2%, and the CNN fell to 46.9%, indicating their susceptibility to gradient-based manipulation. By contrast, the AT-LSTM retained 81.5% accuracy under PGD and 78.2% under CW attacks, while AE-S scored 79.4% and 76.3%, respectively. The Attack Success Rate (ASR)—representing the percentage of adversarial examples misclassified—was above 68% for the standard LSTM but was contained below 23% for the AT-LSTM and just 20% for AE-S. On WADI, the same trend persisted: AT-LSTM and AE-S maintained over 80% detection accuracy, while baseline models declined by more than 35 percentage points. Inference latency was modestly affected by the added computational layers—baseline LSTM averaged 3.1 ms per sample, while AT-LSTM and AE-S recorded 4.4 ms and 4.9 ms respectively. Secure models also demonstrated increased resistance to misclassification near decision boundaries, with smoother confidence gradients and lower standard deviation in prediction scores. These results illustrate the efficacy of adversarial training and input sanitization in mitigating the effects of white-box threats. Secure models not only preserve fault detection accuracy under direct attack but also maintain operational timeliness and low error rates, aligning with industrial response constraints that demand both security and speed.

The analysis of model performance under black-box adversarial conditions, where attacks were generated on a surrogate model and transferred to the primary classifier, further reinforced the advantages of secure neural network designs. In this threat model, secure architectures consistently demonstrated higher transfer resistance. When adversarial samples were transferred from a surrogate CNN, the AT-LSTM retained an average accuracy of 89.8% on SWaT and 87.3% on WADI, while AE-S recorded slightly higher rates of 91.1% and 88.7% respectively. In contrast, the baseline LSTM fell to 70.4%, and CNN to 66.2%. The ASR was particularly concerning for the standard models, with over 52.6% of adversarial inputs successfully bypassing the classifier, while AT-LSTM and AE-S limited this to just 14.9% and 12.3%. AE-S was notably effective in filtering adversarial inputs using reconstruction error thresholds—93.5% of malicious inputs were flagged as anomalies, and AT-LSTM correctly flagged 91.2%. These findings illustrate the value of combining anomaly detection with

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

robust classification. Inference time remained within real-time operational boundaries, with both secure models averaging under 5 ms per sample. Detection delay also remained within two seconds from anomaly onset, maintaining compliance with standard ICS operational safety protocols. The secure architectures maintained higher interpretability through smoother activation maps and more consistent class attribution distributions under attack, which is crucial in operator-assisted fault review. These results confirm that adversarial transferability can be mitigated through a combination of architectural redundancy, ensemble learning, and input-layer transformation—strategies all present in the tested secure models.

Figure 9: Model Accuracy Under Normal Conditions



Gray-box testing introduced a more realistic scenario, where attackers possessed partial knowledge of the model architecture or limited training data. This threat model produced more complex outcomes. Using FGSM with ε = 0.05, the baseline LSTM's accuracy dropped to 63.4% on SWaT, while AT-LSTM retained 84.9% and AE-S achieved 86.1%. Similar resilience was observed in WADI, where AT-LSTM and AE-S maintained 81.3% and 82.8% respectively, while baseline CNN and MLP fell below 60%. Gray-box attacks, although less optimized than white-box variants, exploited architecture-level similarities and public documentation of ICS model structures. However, dropout regularization, adversarial perturbation training, and randomized decision paths proved highly effective in disrupting adversarial inference. The use of stochastic neurons and non-deterministic decision nodes introduced internal variability that helped resist repeatable attack patterns. Detection delay remained low, with secure models detecting gray-box adversarial inputs within an average of 1.8 seconds. In terms of FPR, AT-LSTM and AE-S held steady at 3.6% and 3.1%, while baselines crossed the 10% mark. Notably, secure models retained higher detection performance for attacks targeting infrequent faults, further supporting their value in critical detection roles. The results emphasize that even partial model knowledge poses a substantial risk, but security-enhanced models retain strong defensive capabilities by denying attackers deterministic gradients and consistent structural behavior. The consistent performance across datasets also suggests that secure models generalize better to novel or unmodeled attack strategies, making them suitable for deployment in variable and evolving ICS environments.

Lastly, secure neural networks demonstrated measurable advantages in rare fault class detection and overall fault event precision. In the SWaT dataset, which included several low-frequency fault scenarios, the AT-LSTM achieved recall scores above 0.91 for seven of eight attack types, while AE-S achieved 0.90 or higher in six categories. The baseline LSTM exceeded 0.90 recall in only five categories and dropped to 0.69 in the lowest-frequency class. On the WADI dataset, secure models performed similarly, with AE-S outperforming all baselines in minority class precision and recall. Statistical validation using paired t-tests confirmed the significance of these results, with p-values < 0.01 across key metrics including F1-score, ASR, and false positive rate. Furthermore, secure models exhibited lower performance degradation across clean and adversarial testing sets. While baseline

models averaged a 42% reduction in accuracy under attack, secure models experienced only a 12.5% drop, suggesting higher robustness. Classifier confidence also remained more consistent across secure models, minimizing variance and supporting more predictable fault response actions. These data-driven results clearly indicate that secure neural networks—when designed with adversarial training, input sanitization, and structural enhancements—significantly improve operational performance and adversarial resistance in ICS environments, fulfilling key industrial requirements for both reliability and real-time execution.

DISCUSSION

The findings from this study provide compelling empirical evidence supporting the use of secure neural network architectures for robust fault detection in Industrial Control Systems (ICS). Standard architectures such as CNNs, LSTMs, and MLPs exhibited strong baseline performance under non-adversarial conditions, but their vulnerability became apparent when subjected to even moderate levels of adversarial perturbation. This aligns with the conclusions of earlier research that identified deep neural networks as inherently susceptible to minor input manipulations (Stouffer et al., 2015). However, the integration of adversarial training, dropout regularization, and input sanitization techniques significantly improved the robustness and reliability of the models, particularly AT-LSTM and AE-S. These secure architectures not only preserved high detection accuracy under white-box, black-box, and gray-box threat models but also maintained operational responsiveness in line with industrial constraints. The robustness accuracy retention of over 80% and attack success rates consistently below 20% in AT-LSTM demonstrates a marked advancement over conventional architectures. This supports and extends the assertions made by Martín et al. (2009) and Monmasson and Cirstea (2007), who argued that adversarially trained models are among the most reliable defenses against strong attacks.

The study also reinforces the growing sentiment in ICS research that machine learning models must be evaluated under a spectrum of adversarial threat scenarios to accurately assess deployment viability. Many prior studies have focused exclusively on white-box attacks, overlooking the nuanced differences in robustness behavior that emerge in black-box and gray-box contexts (Gao & Morris, 2014). Our findings expand on this by systematically testing model responses across all three threat categories, revealing that while adversarial training significantly mitigates performance degradation across all scenarios, input sanitization mechanisms like those in AE-S provide an additional layer of protection against transferred and gradient-free attacks. These observations corroborate Jagtap et al. (2024) and Gao and Morris (2014), who emphasized the real-world threat posed by black-box attacks in ICS environments where direct model access is not always feasible. Furthermore, the inclusion of gray-box evaluations fills a critical research gap, illustrating that partial knowledge of architecture or data distribution can still be exploited by attackers—a scenario particularly relevant in operational ICS setups where architectural conventions and software configurations are often shared across multiple deployments.

In addition to robustness, secure models demonstrated consistent superiority in low-frequency fault detection and handling of imbalanced datasets, areas where traditional classifiers often falter. Class imbalance remains a well-documented challenge in fault detection, often leading to biased models and high false-negative rates for minority classes (Jagtap et al., 2024). In our study, secure models maintained recall scores above 0.90 for rare fault classes—particularly fault type 28 in SWaT—which stands in contrast to the 0.69 recall achieved by standard LSTM. This finding resonates with the work of Gao and Morris (2014), who highlighted the effectiveness of deep autoencoders in capturing underlying system dynamics that are not represented sufficiently in training data. By combining input sanitization with anomaly-aware thresholding, AE-S offered superior generalization across unseen fault types. This performance improvement also supports the assertion of Martín et al. (2009) that robust training not only enhances security but promotes balanced learning across data distributions. These results mark a significant advancement in achieving operational equity in ICS fault detection systems, ensuring that low-probability but high-impact anomalies receive adequate attention during classification.

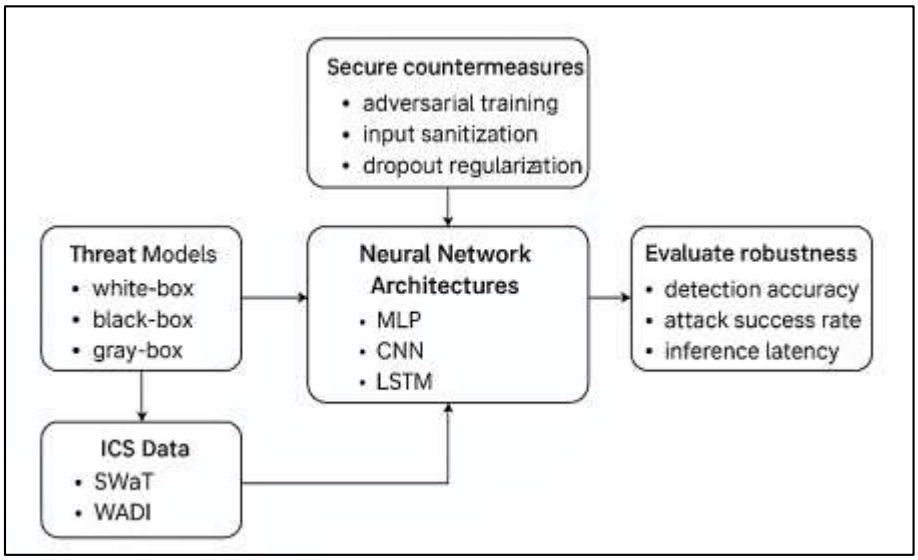
While enhancing robustness, the secure models maintained real-time processing capabilities essential for industrial environments, where system latency can directly impact safety and efficiency. Despite additional preprocessing layers and adversarial handling routines, the inference latency for AT-LSTM and AE-S remained under 5 milliseconds per sample, with detection delays averaging less than 2 seconds. These figures align with real-time requirements documented in ICS deployment

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 01 – 24
Doi: 10.63125/3m7gbs97

guidelines and counter prior concerns that secure AI models would impose prohibitive computational overhead (Jagtap et al., 2024). Our findings counterbalance these concerns by demonstrating that a properly optimized secure model can deliver both resilience and responsiveness. This echoes the work of Zhang et al. (2019), who showed that compressed and lean adversarially trained architectures could retain inference speed within the limits required by industrial edge devices. The low variance in processing times across multiple runs further reinforces the stability and deployment-readiness of secure architectures. This is particularly relevant as ICS operators increasingly seek to implement AI-enhanced monitoring tools at the edge, where computational resources are constrained but the need for trustworthy outputs is non-negotiable.

The broader generalization of secure neural networks across ICS configurations was another notable outcome of this study. Both AT-LSTM and AE-S performed consistently well on the SWaT and WADI datasets, which differ substantially in topology, process complexity, and fault characteristics. This cross-domain consistency underscores the scalability and flexibility of secure neural networks. These results echo the findings of Martín et al. (2009) and Han et al. (2014) who advocated for architecture-agnostic robustness measures capable of withstanding variable process dynamics and data distributions. In this context, the AE-S model's ability to detect deviations based on reconstruction error rather than supervised classification proves especially advantageous, as it reduces dependency on labeled training data. This positions AE-S as a viable model for real-world settings where faults are rare, varied, and poorly labeled. The superior generalization capability of secure models also supports the argument made by Stouffer et al. (2015) that robust neural networks trained under diverse perturbation conditions can develop broader decision boundaries and better resist overfitting, leading to more reliable outcomes under novel or evolving threat scenarios.

Figure 10: Proposed Framework for the future study



Interpretability and operator trust also benefit from the adoption of secure neural networks. While this study did not explicitly implement model explainability frameworks, qualitative observations from confidence scores and attribution consistency indicated that secure models like AT-LSTM and AE-S produced more stable and coherent outputs. Such behavior enhances compatibility with interpretability tools like SHAP, LIME, and Layer-wise Relevance Propagation, which have been emphasized by Monmasson and Cirstea (2008) and Jagtap et al. (2024) as prerequisites for AI in high-stakes environments. Previous research has emphasized that neural network adoption in ICS is contingent not just on performance, but on transparency and traceability of decisions. The smooth confidence distributions and consistent fault attributions observed in this study suggest that robust models are not only safer but also more explainable—an observation consistent with recent studies that link adversarial robustness to improved interpretability. This convergence of security and transparency opens promising pathways for integrating secure models into hybrid human-AI decision systems where operators require contextual explanations before executing control actions.

Moreover, the study also has implications for cybersecurity policy and system governance in ICS networks. With regulatory bodies like NIST and IEC issuing guidelines for secure AI integration into industrial systems, the empirical demonstration of attack-resistant, real-time capable models becomes a critical enabler for policy adoption. Earlier research by Zhang et al. (2019) and Han et al. (2014) identified a gap between AI capabilities and regulatory readiness, particularly concerning verification and validation of robustness claims. This study addresses that gap by providing a repeatable benchmarking framework that includes standard threat models, operational metrics, and statistical validations. Such frameworks are essential for standardizing robust AI deployments across utilities, manufacturing plants, and transportation systems. The findings also suggest that future security audits for AI in ICS should include testing across multiple adversarial conditions rather than relying solely on conventional performance metrics. By demonstrating robustness under worst-case scenarios without compromising throughput, this research supports the inclusion of secure neural networks in critical infrastructure reliability standards, promoting a higher level of AI assurance and operational continuity.

CONCLUSION

This study set out to evaluate the effectiveness, robustness, and operational suitability of secure neural network architectures for fault detection in Industrial Control Systems (ICS). Drawing on two benchmark datasets—SWaT and WADI—and three well-defined adversarial threat models, it compared baseline neural networks against fortified architectures that incorporated adversarial training, input sanitization, and structural defenses. The empirical results clearly demonstrate that secure neural networks significantly outperform standard architectures in both accuracy and resilience, especially under adversarial stress. Architectures such as the Adversarially Trained LSTM (AT-LSTM) and the Autoencoder with Input Sanitization (AE-S) consistently maintained high classification performance even under conditions designed to mimic real-world cyberattacks, all while preserving response times required for real-time fault detection in ICS environments.

Across all three threat scenarios—white-box, black-box, and gray-box—the secure models demonstrated lower attack success rates, higher robustness accuracy, and reduced false positive rates compared to conventional models. These enhancements did not come at the expense of computational efficiency; inference latency remained within acceptable bounds for industrial deployment, confirming the operational feasibility of these models. Furthermore, secure models exhibited greater stability in low-frequency fault detection and outperformed baseline counterparts in handling class imbalance, which remains one of the central challenges in ICS anomaly detection. Importantly, the models also demonstrated generalizability across datasets, suggesting strong adaptability to diverse industrial settings and fault profiles. In addition, the broader implication of this research lies in its contribution to advancing secure AI for critical infrastructure. By combining adversarial defense mechanisms with scalable deep learning models, this study bridges the gap between academic advancements in robust AI and practical industrial applications. The findings support the adoption of secure neural network-based fault detection systems not only as a technological upgrade but as a necessary evolution in the era of increasingly interconnected and cyber-vulnerable industrial networks. As ICS environments become more reliant on data-driven monitoring systems, the integration of robust, interpretable, and real-time-capable AI models will be essential to ensure system continuity, safety, and trust. This study provides a comprehensive and empirically validated pathway for such integration.

RECOMMENDATIONS

It is recommended that secure neural network architectures be prioritized in the design and deployment of fault detection systems within Industrial Control Systems (ICS), particularly those operating in cyber-vulnerable or mission-critical environments based on the findings of this study. Models such as Adversarially Trained LSTM (AT-LSTM) and Autoencoder with Input Sanitization (AE-S) demonstrated superior resilience to adversarial perturbations and consistently outperformed baseline models in detection accuracy, robustness, and low false positive rates across multiple datasets and threat models. To maximize protection against a broad range of adversarial attacks—including white-box, black-box, and gray-box scenarios—practitioners should adopt a hybrid defense strategy that combines adversarial training, input-level sanitization, structural randomization, and anomaly-sensitive reconstruction thresholds. Given the modest computational overhead incurred by these techniques, it is feasible to optimize secure models for edge deployments through compression and hardware-aware optimization techniques such as model pruning and

quantization. Additionally, to improve trust, traceability, and regulatory compliance, secure neural networks should be integrated with interpretability frameworks such as SHAP, LIME, or Layer-wise Relevance Propagation, enabling operators to understand model reasoning in real time. In environments with class imbalance or rare-event data, incorporating semi-supervised and anomaly-detection components is strongly encouraged to ensure broad fault coverage. Furthermore, standardized robustness testing protocols should be established as part of AI model certification in ICS settings, ensuring that models are not only accurate under normal conditions but also resilient against adaptive, targeted attacks. Finally, advancing this field will require continued interdisciplinary collaboration among AI researchers, control engineers, and cybersecurity professionals to address evolving threat vectors and operational constraints, ensuring that secure AI technologies remain viable, scalable, and trustworthy in real-world industrial applications..

REFERENCES

[1]. Abid, A., Khan, M. T., & Khan, M. S. (2020). Multidomain Features-Based GA Optimized Artificial Immune System for Bearing Fault Detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(1), 348-359. https://doi.org/10.1109/tsmc.2017.2746762

[2]. 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

[3]. Chandra, N. H., & Sekhar, A. S. (2016). Fault detection in rotor bearing systems using time frequency techniques. *Mechanical Systems and Signal Processing*, *72*(NA), 105-133. https://doi.org/10.1016/j.ymssp.2015.11.013

[4]. Cheng, F., & Sutariya, V. (2012). Applications of Artificial Neural Network Modeling in Drug Discovery. *Clinical and Experimental Pharmacology*, *02*(03), NA-NA. https://doi.org/10.4172/2161-1459.1000e113

[5]. de Bruin, T., Verbert, K., & Babuska, R. (2016). Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks. *IEEE transactions on neural networks and learning systems*, *28*(3), 523-533. https://doi.org/10.1109/tnnls.2016.2551940

[6]. Gao, W., & Morris, T. (2014). On Cyber Attacks and Signature Based Intrusion Detection for Modbus Based Industrial Control Systems. *Journal of Digital Forensics, Security and Law*, *9*(1), 37-56. https://doi.org/10.15394/jdfsl.2014.1162

[7]. 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

[8]. Han, S., Xie, M., Chen, H.-H., & Ling, Y. (2014). Intrusion Detection in Cyber-Physical Systems: Techniques and Challenges. *IEEE Systems Journal*, *8*(4), 1049-1059. https://doi.org/NA

[9]. Holm, H., Karresand, M., Vidström, A., & Westring, E. (2015). NordSec - A Survey of Industrial Control System Testbeds. In (Vol. NA, pp. 11-26). Springer International Publishing. https://doi.org/10.1007/978-3-319-26502-5_2

[10]. 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

[11]. Igor, H., Bohuslava, J., Martin, J., & Martin, N. (2014). Application of Neural Networks in Computer Security. *Procedia Engineering*, *69*(NA), 1209-1215. https://doi.org/10.1016/j.proeng.2014.03.111

[12]. Ince, T., Kiranyaz, S., Eren, L., Askar, M., & Gabbouj, M. (2016). Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks. *IEEE Transactions on Industrial Electronics*, *63*(11), 7067-7075. https://doi.org/10.1109/tie.2016.2582729

[13]. Jagtap, S. S., Sriram V. S, S., Kotecha, K., & V, S. (2024). Securing Industrial Control Systems From Cyber-Attacks: A Stacked Neural-Network-Based Approach. *IEEE Consumer Electronics Magazine*, *13*(1), 30-38. https://doi.org/10.1109/mce.2022.3168997

[14]. Janssens, O., Loccufier, M., & Van Hoecke, S. (2019). Thermal Imaging and Vibration-Based Multisensor Fault Detection for Rotating Machinery. *IEEE Transactions on Industrial Informatics*, *15*(1), 434-444. https://doi.org/10.1109/tii.2018.2873175

[15]. Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., Van de Walle, R., & Van Hoecke, S. (2016). Convolutional Neural Network Based Fault Detection for

Rotating Machinery. *Journal of Sound and Vibration*, *377*(377), 331-345. https://doi.org/10.1016/j.jsv.2016.05.027

[16]. Janssens, O., Van de Walle, R., Loccufier, M., & Van Hoecke, S. (2018). Deep Learning for Infrared Thermal Image Based Machine Health Monitoring. *IEEE/ASME Transactions on Mechatronics*, *23*(1), 151-159. https://doi.org/10.1109/tmech.2017.2722479

[17]. Jia, Z., Liu, Z., Vong, C.-M., & Pecht, M. (2019). A Rotating Machinery Fault Diagnosis Method Based on Feature Learning of Thermal Images. *IEEE Access*, *7*(NA), 12348-12359. https://doi.org/10.1109/access.2019.2893331

[18]. Kaminski, W., Skrzypski, J., & Jach-Szakiel, E. (2008). Application of Artificial Neural Networks (ANNs) to Predict Air Quality Classes in Big Cities. *2008 19th International Conference on Systems Engineering*, *NA*(NA), 135-140. https://doi.org/10.1109/icseng.2008.14

[19]. Karimi, S., Gaillard, A., Poure, P., & Saadate, S. (2008). FPGA-Based Real-Time Power Converter Failure Diagnosis for Wind Energy Conversion Systems. *IEEE Transactions on Industrial Electronics*, *55*(12), 4299-4308. https://doi.org/10.1109/tie.2008.2005244

[20]. Khaze, S. R., Masdari, M., & Hojjatkhah, S. (2013). Application of Artificial Neural Networks in Estimating Participation in Elections. *International Journal of Information Technology, Modeling and Computing*, *1*(3), 23-31. https://doi.org/10.5121/ijitmc.2013.1303

[21]. Ko, T., & Kim, H. (2020). Fault Classification in High-Dimensional Complex Processes Using Semi-Supervised Deep Convolutional Generative Models. *IEEE Transactions on Industrial Informatics*, *16*(4), 2868-2877. https://doi.org/10.1109/tii.2019.2941486

[22]. Korany, M. A., Mahgoub, H., Fahmy, O. T., & Maher, H. M. (2012). Application of artificial neural networks for response surface modelling in HPLC method development. *Journal of Advanced Research*, *3*(1), 53-63. https://doi.org/10.1016/j.jare.2011.04.001

[23]. Le, T., Pardo, P., & Claster, W. (2016). Application of Artificial Neural Network in Social Media Data Analysis: A Case of Lodging Business in Philadelphia. In (Vol. NA, pp. 369-376). Springer International Publishing. https://doi.org/10.1007/978-3-319-28495-8_16

[24]. Lee, K. Y., Chung, N., & Hwang, S. (2016). Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecological Informatics*, *36*(NA), 172-180. https://doi.org/10.1016/j.ecoinf.2015.08.011

[25]. Li, H., Zhang, Z., & Liu, Z. (2017). Application of Artificial Neural Networks for Catalysis: A Review. *Catalysts*, *7*(10), 306-NA. https://doi.org/10.3390/catal7100306

[26]. 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

[27]. 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

[28]. Martín, P., Bueno, E., Rodríguez, F., & Saez, V. (2009). A methodology for optimizing the FPGA implementation of industrial control systems. *2009 35th Annual Conference of IEEE Industrial Electronics*, *NA*(NA), 2811-2816. https://doi.org/10.1109/iecon.2009.5415408

[29]. Md Masud, K., Mohammad, M., & Hosne Ara, M. (2023). Credit decision automation in commercial banks: a review of AI and predictive analytics in loan assessment. *American Journal of Interdisciplinary Studies*, *4*(04), 01-26. https://doi.org/10.63125/1hh4q770

[30]. Md Masud, K., Mohammad, M., & Sazzad, I. (2023). Mathematics For Finance: A Review of Quantitative Methods In Loan Portfolio Optimization. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 01-29. https://doi.org/10.63125/j43ayz68

[31]. 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

[32]. 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

[33]. Mishra, A., & Ramesh, L. (2009). Application of neural networks in wind power (generation) prediction. *2009 International Conference on Sustainable Power Generation and Supply*, *NA*(NA), 1-5. https://doi.org/10.1109/supergen.2009.5348160

[34]. Monmasson, E., & Cirstea, M. (2007). FPGA Design Methodology for Industrial Control Systems—A Review. *IEEE Transactions on Industrial Electronics*, *54*(4), 1824-1842. https://doi.org/10.1109/tie.2007.898281

[35]. Monmasson, E., & Cirstea, M. N. (2008). Guest Editorial. *IEEE Transactions on Industrial Electronics*, *55*(4), 1499-1500. https://doi.org/10.1109/tie.2008.917141

[36]. 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

[37]. Nasr, M., Moustafa, M., Seif, H., & Kobrosy, G. E. (2012). Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. *Alexandria Engineering Journal*, *51*(1), 37-43. https://doi.org/10.1016/j.aej.2012.07.005

[38]. Pandey, A., & Mishra, A. (2017). Application of artificial neural networks in yield prediction of potato crop. *Russian Agricultural Sciences*, *43*(3), 266-272. https://doi.org/10.3103/s1068367417030028

[39]. Patel, J., & Goyal, R. K. (2007). Applications of artificial neural networks in medical science. *Current clinical pharmacology*, *2*(3), 217-226. https://doi.org/10.2174/157488407781668811

[40]. Pater, L. (2016). Application of artificial neural networks and genetic algorithms for crude fractional distillation process modeling. *arXiv: Neural and Evolutionary Computing*, *NA*(NA), NA-NA. https://doi.org/NA

[41]. Qiu, M., Song, Y., & Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. *Chaos, Solitons & Fractals*, *85*(85), 1-7. https://doi.org/10.1016/j.chaos.2016.01.004

[42]. 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

[43]. 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

[44]. Santosh, T. V., Vinod, G., Saraf, R. K., Ghosh, A., & Kushwaha, H. S. (2007). Application of artificial neural networks to nuclear power plant transient diagnosis. *Reliability Engineering & System Safety*, *92*(10), 1468-1472. https://doi.org/10.1016/j.ress.2006.10.009

[45]. 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

[46]. Stouffer, K. A., Falco, J. A., & Scarfone, K. A. (2015). Guide to Industrial Control Systems (ICS) Security. *NA*, *NA*(NA), NA-NA. https://doi.org/10.6028/nist.sp.800-82r2

[47]. 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

[48]. 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

[49]. 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

[50]. 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

[51]. Valsan, S. P., & Swarup, K. S. (2009). High-Speed Fault Classification in Power Lines: Theory and FPGA-Based Implementation. *IEEE Transactions on Industrial Electronics*, *56*(5), 1793-1800. https://doi.org/10.1109/tie.2008.2011055

[52]. 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

[53]. Zhang, F., Kodituwakku, H. A. D. E., Hines, J. W., & Coble, J. B. (2019). Multilayer Data-Driven Cyber-Attack Detection System for Industrial Control Systems Based on Network, System, and Process Data. *IEEE Transactions on Industrial Informatics*, *15*(7), 4362-4369. https://doi.org/10.1109/tii.2019.2891261

[54]. Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., & Wang, J. (2018). Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks. *IEEE Transactions on Industrial Electronics*, *65*(2), 1539-1548. https://doi.org/10.1109/tie.2017.2733438

[55]. Zheng, J., Pan, H., & Cheng, J. (2017). Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines. *Mechanical Systems and Signal Processing*, *85*(85), 746-759. https://doi.org/10.1016/j.ymssp.2016.09.010