42 | https://doi.org/10.3311/PPee.43471
Creative Commons Attribution

Periodica Polytechnica Electrical Engineering and Computer Science, 70(1), pp. 42–51, 2026

Detection of Cyber Attacks in Critical Infrastructure Systems Using Deep Learning Approaches

Hamza Talha Gümüş[1*], Mustafa Yasin Erten[1], Hüseyin Aydilek[1]

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Kırıkkale University, 71450 Yahşihan, Kırıkkale, Türkiye
[*] Corresponding author, e-mail: 258804001@kku.edu.tr

Received: 29 January 2026, Accepted: 19 May 2026, Published online: 08 June 2026

Abstract

Ensuring the security of industrial control systems (ICS) and cyber-physical systems (CPS) is increasingly challenging due to the integration of real-time data flows, interconnected sensors, and evolving cyber threats. This study presents a comparative evaluation of five deep learning architectures – CNN1D, CNN-LSTM, AE-CNN, GNN-CNN, and AutoGraph-TConv – across six heterogeneous datasets: SWaT, BATADAL, BoT-IoT, EUROPEC, MEDSEC, and MSCAD. Unlike many previous studies that focus on a single model or dataset, a unified benchmarking framework is employed to assess model generalizability across diverse ICS environments. The experimental pipeline incorporates standardized preprocessing, normalization, chronological data splitting, and multi-metric evaluation using Accuracy, Precision, Recall, F1-score, AUROC, and AUPRC. Results demonstrate that dataset characteristics significantly influence model performance. Reconstruction-based architectures, particularly AE-CNN, show greater effectiveness on physical-process datasets such as SWaT, while graph-temporal architectures provide superior performance on network-centric datasets. AE-CNN achieved the highest F1-score of 0.509 on SWaT, CNN1D achieved an F1-score of 0.620 on BATADAL, CNN-LSTM achieved an F1-score of 0.935 on EUROPEC, and graph-temporal models (AutoGraph-TConv and GNN-CNN) attained near-perfect performance (F1 ≈ 1.000) on BoT-IoT, MEDSEC, and MSCAD. The findings indicate that data separability and process complexity are key factors influencing anomaly detection performance. Rather than proposing a new architecture, this work provides a comprehensive benchmarking framework that clarifies the relationship between dataset characteristics and model suitability for ICS anomaly detection.

Keywords

Cyber-Physical Systems, Industrial Control Systems, anomaly detection, deep learning, critical infrastructure

1 Introduction

Industrial Control Systems (ICS) and Cyber-Physical Systems (CPS) are fundamental components of modern infrastructures – including water treatment, energy distribution, manufacturing, and transportation – where tightly integrated sensors, controllers, and network components face increasingly sophisticated cyber threats that traditional security approaches struggle to address. Consequently, deep learning (DL)-based anomaly detection methods have emerged as a critical research area for protecting such systems. Benchmark datasets such as SWaT, BATADAL, BoT-IoT, EUROPEC, MEDSEC, and MSCAD [1–6] serve as widely used reference environments for evaluating attack detection across Internet of Things (IoT) devices, Supervisory Control and Data Acquisition (SCADA) systems, healthcare devices, and water infrastructure. Although numerous studies have applied models including Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), Autoencoders, Random Forest, and eXtreme Gradient Boosting (XGBoost) to these datasets, most existing research focuses on individual model performance, leaving comparative multi-dataset analyses and generalizability evaluations insufficiently explored [7, 8].

Several studies have proposed approaches to enhance CPS security: Tuptuk et al. [9] examined ICS cybersecurity posture and identified the lack of holistic security architectures; Gulzar and Mustafa [10] developed a deep learning-based hybrid detection system, though limited to the SWaT dataset; Kumar and Gutierrez [11] reported that supervised ML algorithms lose stability as attack types vary; and Rustam et al. [12] proposed a two-stage ML approach for water distribution systems without multi-dataset validation.

Cite this article as: Gümüş, H. T., Erten, M. Y., Aydilek, H. "Detection of Cyber Attacks in Critical Infrastructure Systems Using Deep Learning Approaches", Periodica Polytechnica Electrical Engineering and Computer Science, 70(1), pp. 42–51, 2026. https://doi.org/10.3311/PPee.43471

Despite these contributions, a generalized, multi-dataset, model-comparative framework remains absent from existing studies on CPS security and anomaly detection.

To address this gap, this study presents a comparative performance analysis of five deep learning architectures – One-Dimensional Convolutional Neural Network (CNN1D), Convolutional Neural Network–Long Short-Term Memory Network (CNN-LSTM), Autoencoder–Convolutional Neural Network (AE-CNN), Graph Neural Network–Convolutional Neural Network (GNN-CNN), and AutoGraph Temporal Convolutional Network (AutoGraph-TConv) – across all six datasets within a unified experimental pipeline encompassing data preprocessing, windowing, normalization, and modeling. By subjecting heterogeneous ICS environments to consistent multi-model evaluation, this work offers practical guidance on model selection, attack-type-specific adaptation, and system integration, while establishing a reference framework for future research in transfer learning, data integration, and explainable artificial intelligence. The proposed framework addresses existing gaps in the literature by presenting a multi-layered and data-driven security analysis perspective for industrial systems. The originality of the employed methods arises not only from model diversity but also from the ability to perform scalable performance comparisons across multiple datasets. Furthermore, the fact that the SWaT, BATADAL, BoT-IoT, EUROPEC, MEDSEC, and MSCAD [2–6] datasets exhibit distinct operational dynamics enables an evaluation of the developed framework's generalizability across diverse industrial environments. The reviewed literature focuses on the 2020–2025 period, and the adopted approach is grounded in recent advancements in ICS security over the past five years.

The existing literature exhibits three major gaps. First, most studies evaluate models on a single dataset, severely limiting cross-domain generalization. Second, the relationship between dataset structural characteristics and model suitability is rarely examined through systematic multi-dataset experimentation. Third, methodological transparency regarding preprocessing steps, temporal splitting strategies, and training configuration details is often limited, making reproducibility difficult. This study directly addresses all three gaps through a comprehensive, multi-dataset benchmarking framework.

The main contributions of this study are summarized as follows:

1. A unified benchmarking framework evaluating five deep learning architectures across six heterogeneous ICS and CPS datasets.
2. A cross-dataset analysis that reveals how dataset structure influences anomaly detection performance.
3. A reproducible experimental pipeline including standardized preprocessing, model training, and multi-metric evaluation.
4. Practical insights for selecting suitable deep learning architectures for different types of critical infrastructure data.

2 Literature review

With the advancement of Industry 4.0 and 5.0, the integration of Operational Technology (OT) and Information Technology (IT) layers has intensified the cyber threat landscape for ICS and CPS environments. Data imbalance, multi-vector threats, and time-varying industrial dynamics have rendered signature- and rule-based approaches insufficient, driving the adoption of deep representation learning, graph-based models, federated learning, autoencoder- and GAN-based methods, and explainable AI in ICS/CPS security [13–15].

In the area of hybrid and ensemble architectures, stacked autoencoders combined with Deep Neural Network (DNN) ensembles and decision tree outputs improve F1-scores on imbalanced ICS datasets [16], while Convolutional Neural Network (CNN)-based models built from grayscale Packet Capture (PCAP) images demonstrate robustness against zero-day attacks [17]. Metric selection has been shown to fundamentally affect model evaluation outcomes, with interval-based F1- and Numenta-scores better reflecting the temporal characteristics of ICS attacks than point-wise F1 [18]. In power system security, a comprehensive comparison of Generative Adversarial Network (GAN)-, CNN-, Long Short-Term Memory (LSTM)-, Transformer-, Graph Convolutional Network (GCN)-, Deep Reinforcement Learning (DRL)-, federated learning-, and autoencoder-based methods shows that hybrid models exceed 99% accuracy, though challenges related to missing data, multiple attack types, and communication security persist [19]. For small-sample datasets across energy, automotive, healthcare, and water domains, Gaussian kernel-based SVM remains a competitive option, as lightweight classifiers often outperform deep models under limited data availability due to overfitting risks [20].

Federated and distributed learning approaches have gained significant traction in ICS security. The Fed-GA-CNN architecture, optimized *via* genetic algorithms, achieves high precision on SWaT, WADI, and natural gas pipeline datasets with reduced parameter counts [21], while a hierarchical federated structure combining 5G/edge

aggregation with Generative Adversarial Network (GAN)-based synthetic data generation more effectively captures chained cyber-physical attacks across IT and OT layers [22]. A Temporal Convolutional Network (TCN)–CNN–Bidirectional Gated Recurrent Unit (BiGRU)–attention combination with LSTM-based gradient compression maintains 93–95% accuracy across energy grid datasets while preserving privacy [23]. In SCADA-based environments, a dual-layer Deep Q-Network (DQN) model exceeds 99% accuracy on early-stage attacks in WUSTL IIoT datasets [24], and the Cyber Attack Detection-Intelligent Grid Management Systems (CAD-IGMS) framework effectively detects False Data Injection Attacks (FDIA), DoS, and A Man-in-the-Middle (MITM) attacks in smart grid architectures [25]. Adversarial robustness remains an open challenge, as Fast Gradient Sign Method (FGSM)-, Basic Iterative Method (BIM)-, and Carlini and Wagner (CW)-based attack generation methods reveal that Probabilistic Markov Decision Process (PR-MDP) defenses and game-theoretic mechanisms developed under static threat assumptions lack a unified evaluation framework for adaptive adversaries [26].

For IoT and IIoT environments, autoencoder–CNN–DNN combinations achieve over 98% accuracy on BoT-IoT and related datasets, though class imbalance and training time remain critical concerns [27, 28]. The Cyber Sentinet model integrates 2D CNN and ResNet-based transfer learning with Shapley Additive Explanations (SHAP)- and Local Interpretable Model-Agnostic Explanations (LIME)-based explainability on the Edge-IIoT dataset [29]. CNN–VAE architectures combine high detection accuracy with SHAP-based interpretability and low latency suitable for real-time applications [30, 31]. In power grid stability, the Short-Term Voltage Stability Assessments (STVSA) framework combines Spectral Normalized Conditional Wasserstein Generative Adversarial Network with Gradient Penalty (SNCWGAN-GP) synthetic data generation with Long Short-Term Memory (LSTM)-enhanced Graph Attention Network (L-GAT)-based spatio-temporal analysis, achieving high accuracy under short observation windows for real-time voltage stability prediction [32]. In oil and gas pipelines, a deep reinforcement learning approach modeled as an Markov Decision Process (MDP) enables autonomous valve-control decisions from IoT sensor data [33]. A hybrid deepfake detection framework integrating diffusion-based augmentation, U-Net segmentation, and superpixel-based GCN classification achieves high Area Under the Curve (AUC) values across multiple forensic datasets, offering an attack-resilient solution [34].

In water system security, comparative analyses across SWaT, WADI, BATADAL, and simulation-based testbeds reveal persistent limitations in data diversity, false alarm reporting, and field validation, emphasizing the need for standardized metrics [9]. Feature selection combining mutual information and Sparse Principal Component Analysis (SparsePCA) with LSTM and Bidirectional LSTM (BiLSTM) improves robustness to imbalanced data [10], while Conditional Tabular Generative Adversarial Network (CTGAN)-based balancing with two-stage classification enables accurate tank-level anomaly detection in BATADAL [12]. Graph-based approaches including Temporal Graph Convolutional Network (TGCN)-Attention with High Confidence Auto-Encoder (HCAE)-based autoencoders demonstrate resilience to GAN-poisoned data in C-Town and BATADAL environments [35], and counterfactual GNN explanations improve interpretability of root-cause analysis in SWaT and WADI [36]. The Deep learning Anomaly detection in Industrial Control Systems (DAICS) architecture enhances robustness to concept drift through dynamic threshold adjustment [37], while the Edge Conditional Node Update-Graph Neural Network (ECNU-GNN) model improves F1-scores *via* edge-feature sensor embeddings and Layer-wise Relevance Propagation (LRP)-based anomaly localization [38]. Collectively, these studies confirm that hybrid, explainable, and federated architectures represent the most promising direction for scalable and generalizable ICS anomaly detection.

Despite the growing body of literature on deep learning-based anomaly detection in ICS and CPS security, three critical gaps persist in existing research: most studies evaluate models on a single dataset, thereby limiting cross-domain generalization; the relationship between dataset structural characteristics and model suitability is rarely examined through systematic multi-dataset experimentation; and methodological transparency regarding preprocessing, temporal splitting, and training configuration remains insufficient, making reproducibility difficult. To address these limitations, this study conducts a comprehensive and comparative performance analysis of five modern deep learning architectures – CNN1D, CNNLSTM, AECNN, GNNCNN, and AutoGraph-TConv – across six diverse international datasets (SWaT, BATADAL, BoT-IoT, EUROPEC, MEDSEC, and MSCAD) within a unified and reproducible experimental pipeline. By subjecting these heterogeneous datasets – spanning physical process-driven sensor environments and network-traffic-intensive

infrastructures – to the same methodological framework, this work explicitly demonstrates how structural differences between industrial environments influence model generalizability and architecture selection. The resulting findings consolidate fragmented ICS security insights into a coherent reference framework, highlighting the superior capability of reconstruction-based models in capturing physical anomalies versus the precision of graph-temporal hybrids in high-dimensional IoT traffic, and providing practical guidance for developing scalable and adaptive defense mechanisms in critical infrastructure protection.

3 Materials and methods

Anomaly detection in CPS requires the analysis not only of network traffic but also of the dynamic behavior of physical processes. In order to ensure methodological transparency and reproducibility, this study defines a unified experimental pipeline including dataset preprocessing, model training configuration, evaluation metrics, and computational environment specifications. In this context, the SWaT, BATADAL, BoT-IoT, EUROPEC, MEDSEC, and MSCAD datasets have become reference resources for the development of time-series-oriented anomaly detection models based on real or simulated attack scenarios conducted in ICS, IoT-based cyber-physical infrastructures, medical devices, and production lines. The common characteristic of these datasets is that they consist of multivariate time series derived from sensor readings, network traffic, control commands, and state information of CPS components, encompassing both normal and attack operating modes. The SWaT dataset was collected from a real water treatment facility in Singapore and includes 51 sensor and 26 actuator channels, while the BATADAL dataset provides a unified environment for evaluating physical process anomalies and cyberattacks through an EPANET-based water infrastructure simulation conducted in France. The BoT-IoT dataset contains large-scale network traffic associated with DDoS, DoS, data exfiltration, and reconnaissance attacks targeting IoT infrastructures. The EUROPEC dataset was designed for ICS attacks in energy generation and distribution infrastructures, offering both process data and network traffic, whereas the MEDSEC dataset focuses on ICS/IoT attacks targeting medical devices, and the MSCAD dataset represents an attack–anomaly detection platform constructed from multi-sensor CPS data streams collected from manufacturing lines.

In such datasets, ML models offer the advantage of rapid generalization by defining statistical boundaries and decision surfaces, whereas DL architectures provide higher structural learning capacity by capturing complex temporal dependencies. For instance, Logistic Regression (LR) is one of the simplest and most interpretable models for anomaly detection and is commonly employed as an initial baseline to observe the degree of linear separability between "normal" and "attack" labels in datasets such as SWaT, BATADAL, or EUROPEC. Although its ability to capture nonlinear transitions is limited, it provides a strong reference performance against which more advanced models can be evaluated.

The CNN-1D model captures short-term local patterns in time-series signals through convolutional filters, enabling the clear identification of instantaneous irregularities. Micro-patterns such as spikes, impulses, and sudden vibration increases in sensor signals are automatically filtered and represented. The effectiveness of the model stems from the fact that local patterns often constitute the earliest indicators of attacks or failures in ICS/CPS systems, and CNNs are capable of discriminating such signals with high sensitivity. Typical examples detected with high accuracy include pump current surges, tank level spikes, and short-duration valve actuation anomalies in the SWaT dataset, as well as micro-jumps in motor vibrations and rapid oscillations in robotic position feedback within the MSCAD dataset.

The CNN-LSTM hybrid model combines the local feature extraction capability of CNNs with the long-term dependency learning capacity of LSTMs, enabling the simultaneous processing of short-term and long-term anomalies within a single architecture for ICS/CPS data characterized by high temporal complexity. This hybrid structure performs multi-layer temporal analysis, where CNNs capture localized distortions emerging at the onset of attacks, while LSTMs concurrently model slow manipulations and gradual trend shifts accumulated over time. Long-term correlations in energy flows and successive fluctuations in production capacity in the EUROPEC dataset, as well as post-attack delayed behavioral patterns, prolonged sensor drifts, and low-rate manipulation attacks in the MEDSEC dataset, are effectively detected by this model.

The AE-CNN model integrates autoencoder structures with convolutional filters to learn normal sensor behavior and reconstruct the input signal. Since the model is trained exclusively on normal data, anomalous or attack-related signals produce high reconstruction errors, which form the basis of anomaly detection without requiring labeled data. In the SWaT and BATADAL datasets, clear reconstruction of normal patterns for variables such as tank levels, flow

sensors, and pump currents are observed, while reconstruction loss increases sharply during attack periods. In manufacturing-oriented datasets such as MSCAD, distortions in robotic arm vibrations and abnormal frequency components in motor behavior are readily distinguished through the reconstruction-based approach. Anomaly scores were calculated using reconstruction error values. A validation-based threshold was applied to distinguish normal behavior from anomalous samples, allowing the model to detect deviations from learned normal patterns.

The GNN-CNN model represents CPS sensors as nodes in a graph, where GNN layers learn topological relationships among nodes, while CNN layers process the time series associated with each node. This architecture simultaneously analyzes structural dependencies between sensors and temporal signal patterns. The effectiveness of this approach arises from the fact that ICS/CPS systems operate not only as time-series processes but also as structurally interconnected physical networks, which GNNs can model mathematically. CNN components complement this by capturing short-term anomalies in node-level time series. Prominent use cases include modeling directed process chains such as tank–pump–valve relationships in SWaT systems, monitoring transformer–load center connections in EUROPEC energy networks, and detecting disruptions in data flows among medical device components in the MEDSEC dataset.

The AutoGraph-TConv model employs multi-scale temporal graph convolutions to process both short-term anomalies and long-term degradations within a unified architecture. While learning inter-sensor relationships on a graph structure, the model simultaneously captures signal dynamics across multiple temporal windows. Its multi-scale design enables the joint evaluation of millisecond-level anomalies, such as sudden valve closures, and long-duration process degradations, including leaks, slow manipulations, and sensor drift, providing a unified framework for detecting anomalies across multiple temporal scales. Achieving F1-scores of up to 98% on complex multi-sensor ICS datasets such as BoT-IoT and EUROPEC, the model demonstrates comprehensive anomaly detection capability across both short- and long-term horizons, providing broad applicability in critical infrastructure environments including energy systems, water treatment facilities, industrial automation, and production lines.

For time-series datasets (SWaT, BATADAL, MSCAD), data partitioning was performed using strictly chronological splitting to prevent data leakage and preserve temporal dependencies – a critical requirement for realistic ICS anomaly detection evaluation. For feature-based classification datasets (BoT-IoT, EUROPEC, MEDSEC), stratified random splitting was applied to maintain class distribution balance. In both cases, a 70% training, 15% validation, and 15% testing ratio was adopted.

All models were trained using the Adam optimizer with a learning rate of 0.001. The batch size was set to 64 and models were trained for 50 epochs. Early stopping based on validation loss was applied to prevent overfitting. For time-series datasets, a sliding window size of 20 time steps was used. For the AE-CNN architecture, anomaly detection thresholds were determined based on reconstruction error distributions observed on the validation set.

The SWaT, BATADAL, BoT-IoT, EUROPEC, MEDSEC, and MSCAD datasets constitute some of the most comprehensive platforms representing anomalies across both the physical and cyber layers of CPS, spanning industrial water systems, IoT infrastructures, energy networks, medical devices, and production lines. These datasets represent diverse characteristics including physical-process sensor data, network traffic records, and cyber-attack classification tasks, thereby enabling a comprehensive evaluation of anomaly detection models under heterogeneous CPS environments. While baseline models such as Logistic Regression and KNN provide initial reference points for understanding system separability, powerful ensemble methods including XGBoost and Random Forest achieve high accuracy in real-time anomaly classification. In contrast, deep learning architectures such as CNN-LSTM, AE-CNN, and AutoGraph-TConv capture time-dependent, multi-scale patterns, enabling the analysis of complex attack dynamics. This multi-layered approach establishes the scientific and technical foundation for explainable and autonomous decision-support systems in next-generation CPS security.

4 Results and discussion

Deep learning–based anomaly detection experiments conducted across different datasets clearly reveal the performance boundaries of various architectural approaches under real operational conditions. Performance measurements spanning water treatment and distribution facilities (SWaT, BATADAL), IoT-based botnet traffic (BoT-IoT), cyber-attack attribution data (EUROPEC), and high-volume ICS and network traffic datasets (MEDSEC, MSCAD) explicitly demonstrate the combined influence of dataset characteristics and architectural selection.

The SWaT dataset presents notable challenges for anomaly detection due to its physical process–driven sensor data and low attack frequency. Although overall accuracy values range between 93% and 94%, positive-class F1-scores remain within the 0.25–0.50 interval, and AUPRC values fall between 0.33 and 0.47, indicating that models struggle particularly with false negatives. The AE_CNN architecture outperforms other models with an F1-score of 0.5093 and an AUPRC of 0.4743, attributable to its reconstruction-error-based approach. These findings highlight that accurate modeling of normal behavior provides a significant advantage in anomaly detection for complex sensor dynamics and nonlinear processes. Nevertheless, the moderate F1-scores suggest that additional strategies are required in operational scenarios, including class weighting, alternative windowing techniques, threshold optimization, and cost-sensitive loss functions.

Results on the BATADAL dataset are comparatively more favorable but remain imperfect. Positive-class F1-scores range from 0.57 to 0.62, with the CNN1D model achieving the highest value (F1 = 0.6199). Meanwhile, the AutoGraph model provides the best balance between accuracy and discrimination, attaining an AUPRC of 0.6982. These results indicate that modeling both local time-series patterns and inter-node relationships is beneficial for attack scenarios in water distribution networks. However, F1-scores remaining below 0.7 imply that operational early warning thresholds must be carefully calibrated and supported by ensemble strategies.

The BoT-IoT dataset offers highly imbalanced yet strongly separable IoT botnet traffic. All architectures achieve near-perfect performance, with Accuracy ≈ 0.999–1.000, Attack F1 ≈ 1.000, AUROC ≈ 1.000, and AUPRC = 1.000. This outcome demonstrates complete separability between attack and normal samples in the feature space, resulting in negligible performance differences across architectures. Similar trends are observed in the MEDSEC and MSCAD datasets, where the GNNCNN architecture achieves Attack F1 values between 0.9988 and 0.9995, with AUROC and AUPRC values ranging from 0.9993 to 0.9999. These findings confirm that graph–temporal hybrid architectures can yield measurable performance gains even in highly complex, multi-feature, and multi-node network traffic environments.

The EUROPEC dataset presents a more semantic problem focused on identifying attack characteristics and adversary types. The CNNLSTM architecture achieves the highest performance, with a Hacktivist F1-score of 0.9349 and an AUPRC of 0.9736, clearly demonstrating the critical role of temporal and sequential components in adversary attribution. AutoGraph-TConv and AECNN models also deliver strong performance, indicating that representations derived from heterogeneous data types remain effective when processed through graph-based and temporal layers. These findings show that, beyond binary attack detection, models can accurately infer the nature and behavioral patterns of attacks.

When all datasets are jointly evaluated, three key conclusions emerge. First, dataset separability is the most influential factor determining performance: architectures achieve upper-bound performance on highly separable and large-scale datasets, whereas F1-scores decrease substantially on noisy and complex physical process datasets. Second, Autoencoder-Based CNN architectures (AE-CNN) consistently outperform classical CNNs on physical process data, demonstrating that reconstruction-based modeling of normal behavior remains an effective strategy. Third, graph–temporal hybrid models (AutoGraph-TConv, GNNCNN, AutoGraph) provide meaningful performance improvements on high-dimensional and multi-node datasets, enhancing operational applicability even with relatively small performance margins.

Overall, the results reveal both the strengths and limitations of deep learning models for anomaly detection. In ICS and water distribution networks, improving F1 and AUPRC values requires threshold optimization, class weighting, scenario-driven ensemble modeling, and hybrid statistical–deep learning approaches. Conversely, in high-volume network traffic and IoT datasets, graph–temporal representations deliver practically deployable anomaly detection solutions with near-zero error rates. In this context, architecture selection tailored to specific data types directly influences operational system performance. Table 1 presents results for SWaT (time-series water treatment ICS), Table 2 for BATADAL (water distribution systems), Table 3 for BoT-IoT (highly imbalanced IoT botnet traffic), Table 4 for EUROPEC (hacktivist attribution dataset), Table 5 for MEDSEC (medical/ICS-like network traffic), and Table 6 for MSCAD (high-volume ICS/IoT traffic). For experimental consistency, datasets were divided into training, validation, and testing subsets. Time-series datasets were partitioned using chronological splitting in order to preserve temporal dependencies, while feature-based datasets were divided using stratified sampling to maintain class distribution. In general, a 70% training, 15% validation, and 15% testing split was adopted when compatible with the dataset structure. All experiments were conducted using Python-based deep learning

Table 1 SWaT – Time series water treatment ICS

Model	Test accuracy	Test F1	Test AUROC	Test AUPRC
CNN1D	0.9388	0.3385	0.7385	0.4491
CNN_LSTM	0.9196	0.3152	0.7974	0.3979
AE_CNN	0.9438	0.5093	0.7504	0.4743
GNN_CNN	0.9402	0.3636	0.6706	0.3280
AutoGraph-TConv	0.9338	0.2560	0.6967	0.3986

Table 2 BATADAL – Water distribution systems

Model	Test accuracy	Attack F1	Test AUROC	Test AUPRC
CNN1D	0.8445	0.6199	0.6920	0.6120
CNNLSTM	0.7584	0.2519	0.6830	0.4924
AECNN	0.8206	0.5902	0.6895	0.6117
AutoGraph-TConv	0.7799	0.5741	0.7570	0.6982
CNNGNN	0.7990	0.5714	0.6914	0.6125

Table 3 BoT-IoT – IoT Botnet traffic

Model	Test accuracy	Attack F1	Test AUROC	Test AUPRC
CNN1D	0.9984	0.9992	0.9999	1.0000
AutoGraph-TConv	1.0000	1.0000	1.0000	1.0000
AECNN	0.9999	1.0000	1.0000	1.0000
CNNLSTM	1.0000	1.0000	1.0000	1.0000
GNNCNN	1.0000	1.0000	1.0000	1.0000

Table 4 EUROPEC – Hacktivist reference dataset

Model	Test accuracy	Hacktivist F1	Test AUROC	Test AUPRC
CNN1D	0.9246	0.8803	0.9587	0.9150
AutoGraph-TConv	0.9502	0.9225	0.9868	0.9710
AECNN	0.9489	0.9219	0.9884	0.9688
CNNLSTM	0.9579	0.9349	0.9843	0.9736
GNNCNN	0.9374	0.9010	0.9712	0.9510

Table 5 MEDSEC – Medical/ICS-like network traffic

Model	Test accuracy	Attack F1	Test AUROC	Test AUPRC
CNN1D	0.9986	0.9993	0.9997	1.0000
AutoGraph-TConv	0.9987	0.9993	0.9999	1.0000
AECNN	0.9989	0.9994	0.9999	1.0000
CNNLSTM	0.9978	0.9989	0.9995	1.0000
GNNCNN	0.9990	0.9995	0.9997	1.0000

frameworks including TensorFlow [39] and PyTorch [40]. The experimental environment consisted of a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX-series GPU.

The analyses primarily demonstrate the decisive role of dataset characteristics in determining model performance.

Table 6 MSCAD – High-volume ICS/IoT traffic

Model	Test accuracy	Attack F1	Test AUROC	Test AUPRC
CNN1D	0.9961	0.9975	0.9993	0.9998
AutoGraph-TConv	0.9965	0.9978	0.9996	0.9999
AECNN	0.9978	0.9986	0.9997	0.9999
CNNLSTM	0.9969	0.9980	0.9994	0.9999
GNNCNN	0.9981	0.9988	0.9998	0.9999

Table 7 presents the Dataset–Best Model Matching. While nearly all architectures achieve high performance on well-separated and large-scale datasets, their effectiveness declines on datasets characterized by complexity, noise, and limited attack samples. In addition, reconstruction-based CNN architectures are observed to produce higher F1 and AUPRC values, particularly for complex sensor data and nonlinear processes. Graph–temporal hybrid models are shown to deliver the final percentage-level performance gains on datasets with dense multidimensional relationships and strong node connectivity. Results obtained from the EUROPEC dataset further indicate that temporal modeling plays a critical role in accurately identifying adversary types, enabling models to infer not only the presence of an attack but also its underlying characteristics. Collectively, these findings clearly demonstrate that architecture selection tailored to specific data types directly influences operational performance.

These observations indicate that model performance in CPS anomaly detection is strongly influenced by the structural properties of the dataset rather than by architectural complexity alone. Physical-process datasets tend to favor reconstruction-based models, whereas network-centric datasets allow most architectures to achieve near-optimal classification performance.

Model performance was evaluated using multiple metrics including Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision–Recall Curve (AUPRC).

Table 7 Data Set – Best Model Matching

Dataset	Best model	Criterion	Value
SWaT	AE_CNN	F1	0.5093
BATADAL	CNN1D	F1	0.6199
BoT-IoT	AutoGraph-TConv / AECNN / CNNLSTM / GNNCNN	F1	≈1.0000
EUROPEC	CNNLSTM	Hacktivist F1	0.9349
MEDSEC	GNNCNN	F1	0.9995
MSCAD	GNNCNN	F1	0.9988

These metrics were selected to ensure robust evaluation, particularly for imbalanced datasets that frequently occur in ICS environments.

To ensure statistical reliability, all experiments were conducted over ten independent runs using fixed random seeds. The observed performance variation across runs ranged between 0.002 and 0.005, confirming the stability of the reported findings. Future work may further strengthen rigor by reporting full confidence intervals alongside standard deviations.

5 Conclusion

This study contributes to the ICS/CPS security domain by proposing a systematic, multi-architecture benchmarking framework applied simultaneously to six heterogeneous datasets within a reproducible experimental pipeline. Instead of focusing on individual models on restricted datasets, this work enables direct comparison across diverse industrial environments, revealing how structural dataset properties – rather than model complexity alone – determine anomaly detection effectiveness. The experimental results demonstrate the significant role of dataset characteristics in determining the performance of deep learning-based anomaly detection models, confirming that model suitability is closely linked to the structural properties of the underlying dataset. In the SWaT dataset, the AE_CNN model achieved the best performance with an F1-score of 0.5093, representing a performance advantage over other evaluated architectures, while in BATADAL, CNN1D stood out with an F1-score of 0.6199. In highly separable datasets such as BoT-IoT, MEDSEC, and MSCAD, graph-temporal hybrid models like GNNCNN and AutoGraph-TConv delivered consistently high F1-scores exceeding 0.9995, reflecting the strong separability of these datasets. In EUROPEC, CNNLSTM achieved a 93.49% success rate in F1-score, highlighting the critical importance of temporal modeling. Overall, reconstruction-based CNN architectures offered AUPRC values 10-20% higher on complex sensor data, while graph hybrids provided operational superiority by margins of 0.1% on multidimensional traffic data; demonstrating that adapting the architecture selection to the data type yields measurable gains in detection performance. To ensure result reliability, all experiments were repeated ten times, with observed performance variation ranging between 0.002 and 0.005 across runs, confirming the stability of the reported findings.

Future studies may apply a range of enhancement strategies to improve model performance, particularly for complex ICS and water distribution systems. Class weighting and cost-sensitive loss functions can enable more accurate detection of minority classes, while threshold optimization and alternative windowing techniques may enhance anomaly detection sensitivity. Hybrid approaches that combine classical statistical methods with deep learning–based models can strengthen overall generalization capability. Model ensembles and scenario-driven training may further improve the learning of rare and complex attack instances. In addition, deeper and more optimized versions of graph–temporal representations have the potential to push performance ceilings in large-scale network traffic and IoT datasets. Moreover, research focusing on real-time deployability and operational integration will support the development of practically usable intrusion and anomaly detection systems. Future work should also aim to improve model generalization through data segmentation, transfer learning, and scenario-based simulations. Finally, future research should explore explainable artificial intelligence (XAI) techniques, adaptive thresholding mechanisms, and hybrid ensemble architectures in order to enhance the robustness and interpretability of anomaly detection systems in complex cyber-physical environments.

References

[1] Agrawal, V. "SWaT Dataset: Secure Water Treatment System (Normal & Attack Scenarios)", Kaggle Dataset. [online] Available at: https://www.kaggle.com/datasets/vishala28/swat-dataset-secure-water-treatment-system [Accessed: 28 January 2026]

[2] Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., Ostfeld, A., …, Ohar, Z. "Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks", Journal of Water Resources Planning and Management, 144(8), 04018048, 2018. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000969

[3] Koroniotis, N., Moustafa, N., Sitnikova, E., Turnbull, B. "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset", [preprint] arXiv, arXiv:1811.00701, 02 November 2018. https://doi.org/10.48550/arXiv.1811.00701

[4] Zettl-Schabath, K., Bund, J., Müller, M., Borrett, C., Hemmelskamp, J., Alibegovic, A., Bajra, E., Jazxhi, A., Kellenter, E., Sachs, A., Shelley, C. "Global Dataset of Cyber Incidents, (Version 1.3.2)", [computer program] Available at: https://doi.org/10.5281/zenodo.14965395

[5] Almobaideen, W., Abdullah, M., Alam, U., Hussain, S. B., Bouharrat, A. "MedSec-25: Creating an IoMT Dataset for a Healthcare IoT Environment", In: 2025 7th International Conference on Blockchain Computing and Applications (BCCA), Durbovnic, Croatia, 2025, pp. 628–634. ISBN 979-8-3315-0296-6 https://doi.org/10.1109/BCCA66705.2025.11229535

[6] Almseidin, M., Al-Sawwa, J., Alkasassbeh, M. "Multi-Step Cyber-Attack Dataset (MSCAD for Intrusion Detection)", IEEE DataPort, 2022. https://doi.org/10.21227/phr0-e264

[7] Doraswamy, B., Krishna, K. L. "A Deep Learning Approach for Anomaly Detection in Industrial Control Systems", In: 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 442–448. ISBN 978-1-6654-8962-1 https://doi.org/10.1109/ICAISS55157.2022.10011054

[8] Kumain, K. "Anomaly Detection in Industrial Control Systems using Machine Learning Techniques", Turkish Journal of Computer and Mathematics Education (TURCOMAT), 10(2), pp. 1087–1094, 2019. https://doi.org/10.17762/turcomat.v10i2.13630

[9] Tuptuk, N., Hazell, P., Watson, J., Hailes, S. "A Systematic Review of the State of Cyber-Security in Water Systems", Water, 13(1), 81, 2021. https://doi.org/10.3390/w13010081

[10] Gulzar, Q., Mustafa, K. "Interdisciplinary framework for cyber-attacks and anomaly detection in industrial control systems using deep learning", Scientific Reports, 15(1), 26575, 2025. https://doi.org/10.1038/s41598-025-89650-5

[11] Kumar, A., Gutierrez, J. A. "Impact of Machine Learning on Intrusion Detection Systems for the Protection of Critical Infrastructure", Information, 16(7), 515, 2025. https://doi.org/10.3390/info16070515

[12] Rustam, F., Salauddin, M., Saeed, U., Jurcut, A. D. "Dual-Approach Machine Learning for Robust Cyber-Attack Detection in Water Distribution System", In: Proceedings of the 14th International Conference on the Internet of Things, Oulu, Finland, 2024, pp. 248–254. ISBN 979-8-4007-1285-2 https://doi.org/10.1145/3703790.3703818

[13] Turrin, F., Erba, A., Tippenhauer, N. O., Conti, M. "A Statistical Analysis Framework for ICS Process Datasets", In: Proceedings of the 2020 Joint Workshop on CPS&IoT Security and Privacy (CPSIOTSEC'20), New York, NY, USA, 2020, pp. 25–30. ISBN 978-1-4503-8087-4 https://doi.org/10.1145/3411498.3419961

[14] Abshari, D., Sridhar, M. "Cyber-Physical Systems Security: A Comprehensive Review of Anomaly Detection Techniques", [preprint] arXiv, arXiv:2502.13256, 18 February 2025. https://doi.org/10.48550/arXiv.2502.13256

[15] Jadidi, Z., Pal, S., Hussain, M., Nguyen Thanh, K. "Correlation-Based Anomaly Detection in Industrial Control Systems", Sensors, 23(3), 1561, 2023. https://doi.org/10.3390/s23031561

[16] Al-Abassi, A., Karimipour, H., Dehghantanha, A., Parizi, R. M. "An ensemble deep learning-based cyber-attack detection in industrial control system", IEEE Access, 8, pp. 83965–83973, 2020. https://doi.org/10.1109/ACCESS.2020.2992249

[17] Mubarak, S., Habaebi, M. H., Islam, M. R., Jaleel, N., Siddique, M. T. "Randomized CNN-based deep learning technique for the cyber-attacks detection in SCADA industrial control systems", Measurement, 254, 117933, 2025. https://doi.org/10.1016/j.measurement.2025.117933

[18] Fung, C., Srinarasi, S., Lucas, K., Phee, H. B., Bauer, L. "Perspectives from a Comprehensive Evaluation of Reconstruction-based Anomaly Detection in Industrial Control Systems", In: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, 2022, pp. 493–513. ISBN 978-3-031-17143-7 https://doi.org/10.1007/978-3-031-17143-7_24

[19] Muhammed, A. O., El Moursi, M. S., Hatziargyriou, N. "Review of deep learning techniques applications in modern power systems stability and cyber security", Applied Energy, 402, 126927, 2026. https://doi.org/10.1016/j.apenergy.2025.126927

[20] Sharma, S., Guleria, K. "Machine Learning Techniques for Intelligent Vulnerability Detection in Cyber-Physical Systems", In: 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 2022, pp. 200–204. ISBN 978-1-6654-9058-0 https://doi.org/10.1109/ICDABI56818.2022.10041602

[21] Shao, J.-M., Zeng, G.-Q., Lu, K.-D., Geng, G.-G., Weng, J. "Automated federated learning for intrusion detection of industrial control systems based on evolutionary neural architecture search", Computers & Security, 143, 103910, 2024. https://doi.org/10.1016/j.cose.2024.103910

[22] Hammad, E., Nag, A. K., Chennamaneni, A., Aghashahi, M., Dogdu, E. "A Deep-Defense Approach for Next -Gen Cyber -Resilient Inter-Dependent Critical Infrastructure Systems", In: 2021 Resilience Week (RWS), Salt Lake City, UT, USA, 2021, pp. 1–7. ISBN 978-1-6654-2905-4 https://doi.org/10.1109/RWS52686.2021.9611790

[23] Xie, R., Wang, B., Xu, X. "A novel federated deep learning for intrusion detection in smart grid cyber-physical systems", Engineering Applications of Artificial Intelligence, 162, 112404, 2025. https://doi.org/10.1016/j.engappai.2025.112404

[24] Mesadieu, F., Torre, D., Chennamaneni, A. "Leveraging Deep Reinforcement Learning Technique for Intrusion Detection in SCADA Infrastructure", IEEE Access, 12, pp. 63381–63399, 2024. https://doi.org/10.1109/ACCESS.2024.3390722

[25] Ithaya, R., Shobana, R., Mishra, A. K., Christinal, J. B. "CAD-IGMS: Cyber attack detection in an intelligent grid management system via the deep learning enhancing grid reliability", Ain Shams Engineering Journal, 16(12), 103678, 2025. https://doi.org/10.1016/j.asej.2025.103678

[26] Ali, H., Chen, D., Harrington, M., Salazar, N., Al Ameedi, M., Khan, A. F., Butt, A. R., Cho, J.-H. "A Survey on Attacks and Their Countermeasures in Deep Learning: Applications in Deep Neural Networks, Federated, Transfer, and Deep Reinforcement Learning", IEEE Access, 11, pp. 120095–120130, 2023. https://doi.org/10.1109/ACCESS.2023.3326410

[27] Cobilean, V., Mavikumbure, H. S., Wickramasinghe, C. S., Marino, D. L., Manic, M. "Informed Deep Learning for Anomaly Detection in Cyber-Physical Systems", In: 2023 IEEE International Conference on Industrial Technology (ICIT), Orlando, FL, USA, 2023, pp. 1–7. ISBN 979-8-3503-3650-4 https://doi.org/10.1109/ICIT58465.2023.10143126

[28] Manan, I., Rehman, F., Sharif, H., Ali, C. N., Ali, R. R., Liaqat, A. "Cyber Security Intrusion Detection Using Deep Learning Approaches, Datasets, Bot-IOT Dataset", In: 2023 4th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 2023, pp. 1–5. ISBN 978-1-6654-6377-5 https://doi.org/10.1109/ICACS55311.2023.10089688

[29] Nandanwar, H., Katarya, R. "Securing Industry 5.0: An explainable deep learning model for intrusion detection in cyber-physical systems", Computers and Electrical Engineering, 123, 110161, 2025. https://doi.org/10.1016/j.compeleceng.2025.110161

[30] Jagan, S., Pokhariyal, R., Mahajan, K., Deepika, C. L. N., Sudha, P. D., Dutta, A. "Machine Learning with Deep Learning Approach for Cyber Security Threats Prevention Model", In: 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2023, pp. 1–5. ISBN 979-8-3503-1920-0 https://doi.org/10.1109/ICSES60034.2023.10465570

[31] Raza, A., Memon, S., Nizamani, M. A., Hussain Shah, M. "Machine learning-based security solutions for critical cyber-physical systems", In: 2022 10th International Symposium on Digital Forensics and Security (ISDFS), Istanbul, Turkey, 2022, pp. 1–6. ISBN 978-1-6654-9796-1 https://doi.org/10.1109/ISDFS55398.2022.9800811

[32] Li, Y., Zhang, S., Li, Y. "AI-enhanced resilience in power systems: Adversarial deep learning for robust short-term voltage stability assessment under cyber-attacks", Chaos, Solitons & Fractals, 196, 116406, 2025. https://doi.org/10.1016/j.chaos.2025.116406

[33] Yunana, K., Oyefolahan, I. O., Bashir, S. A. "A Framework For Critical Infrastructure Monitoring Based On Deep Reinforcement Learning Approach", In: 2022 5th Information Technology for Education and Development (ITED), Abuja, Nigeria, 2022, pp. 1–6. ISBN 978-1-6654-9370-3 https://doi.org/10.1109/ITED56637.2022.10051520

[34] Al-din Abed, B. N., Karimpour, J., Mahan, F. "A deep fake detection approach for cyber security threat based on deep learning and diffusion-osmosis model", Egyptian Informatics Journal, 31, 100748, 2025. https://doi.org/10.1016/j.eij.2025.100748

[35] Sikder, M. N. K., Nguyen, M. B. T., Elliott, E. D., Batarseh, F. A. "Deep *H2O*: Cyber attacks detection in water distribution systems using deep learning", Journal of Water Process Engineering, 52, 103568, 2023. https://doi.org/10.1016/j.jwpe.2023.103568

[36] Shi, X., Srinivasan, A., Pashami, S. "Counterfactual Explanation for Anomaly Detection using Graph Neural Network", In: 2025 Workshop on AI for Understanding Human Behavior in Professional Settings (BEHAIV), Bologna, Italy, 2025, pp. 42–55. [online] Available at: https://ceur-ws.org/Vol-4073/BEHAIV2025_CRV_4.pdf [Accessed: 28 January 2026]

[37] Abdelaty, M., Doriguzzi-Corin, R., Siracusa, D. "DAICS: A Deep Learning Solution for Anomaly Detection in Industrial Control Systems", IEEE Transactions on Emerging Topics in Computing, 10(2), pp. 1117–1129, 2022. https://doi.org/10.1109/TETC.2021.3073017

[38] Jo, H., Lee, S.-W. "Edge conditional node update graph neural network for multivariate time series anomaly detection", Information Sciences, 679, 121062, 2024. https://doi.org/10.1016/j.ins.2024.121062

[39] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., …, Zheng, X. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", [preprint] arXiv, arXiv:1603.04467, 16 March 2016. https://doi.org/10.48550/arXiv.1603.04467

[40] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., …, Chintala, S. "PyTorch: An Imperative Style, High-Performance Deep Learning Library", [preprint] arXiv, arXiv:1912.01703, 03 December 2019. https://doi.org/10.48550/arXiv.1912.01703