Power System Protection and Control
ISSN:1674-3415

BENCHMARKING AI-DRIVEN ANOMALY DETECTION: FROM TREE-BASED ENSEMBLES TO GRAPH NEURAL NETWORKS IN IOT AND TRADITIONAL SYSTEMS

H. Talha Gümüş[1], A. Gökhan Alarslan[2], Hüseyin Aydilek[3]*, Mustafa Y. Erten[4]

[1,2,3,4] Department of Electrical and Electronics Engineering, Kırıkkale University, Kırıkkale, Türkiye
talhagms06@gmail.com; alarslangokhan@gmail.com; huseyinaydilek@kku.edu.tr; mustafaerten@kku.edu.tr
[1] https://orcid.org/0009-0006-8959-0204; [2] https://orcid.org/0009-0001-0803-3072;
[3] https://orcid.org/0000-0003-3051-4259; [4] https://orcid.org/0000-0002-5140-1213
*Corresponding Author: huseyinaydilek@kku.edu.tr

Received: 19/01/2026 **Revised**: 22/02/2026 **Accepted**: 19/03/2026

***ABSTRACT*:**

The rapid expansion of cyber-physical systems and network-based digital infrastructures necessitates the evolution of intrusion detection mechanisms into more flexible, intelligent, and scalable architectures. Traditional signature-based systems prove inadequate against zero-day attacks, high-volume data flows, and the dynamic nature of IoT-based traffic, as they are limited to reacting solely to known threats. Machine learning and deep learning methods transcend these limitations by strengthening anomaly-based detection processes and identifying unknown attacks through the extraction of statistical models of normal behaviour. Predicated on technical trends from the 2020–2025 period, this study presents a comparative analysis of classical and deep learning approaches and introduces a hybrid methodology addressing contemporary challenges such as data imbalance, explainability, real-time performance, and resource efficiency. Models based on LSTM, CNN, Autoencoder, GAN, XGBoost, and Random Forest were evaluated on extensive datasets, and their generalization capacities were tested across IoT, IIoT, and traditional network scenarios. The results indicate that while no single algorithm maintains dominance across every scenario, multi-layered hybrid models demonstrate high accuracy, low false-positive rates, and strong generalization performance. These findings suggest that AI-supported intrusion detection systems will become a fundamental component of future defence architectures, highlighting the critical importance of developing holistic, multi-stage solutions in this domain.

***Keywords**: Anomaly Detection, Deep Learning, Intrusion Detection Systems, Machine Learning, Network Traffic Analysis.*

INTRODUCTION

Intrusion Detection Systems (IDS), as one of the most critical components of cybersecurity, occupy a central role in safeguarding the integrity and continuity of digital infrastructures. Contemporary network traffic is becoming increasingly complex, and the rapid proliferation of IoT- and IIoT-enabled devices is continuously expanding data volume, connection density, and the overall attack surface. This evolution constrains the effectiveness of conventional signature-based systems that can only recognize known threats, thereby creating substantial vulnerabilities against emerging attack types. In this context, anomaly-based approaches have gained prominence by offering mechanisms that can learn deviations from normal behaviour, adapt to changing conditions, and remain effective even against previously unseen attack patterns. The primary objective of this study is to enhance the security architecture by integrating machine learning and deep learning methods into intrusion detection systems, making them more intelligent, autonomous, and scalable. Accordingly, the study aims to enable the automatic identification of abnormal behaviours within high-volume data streams and to dynamically strengthen network security.

In today's context, cyberattack detection is not merely a domain of technical advancement but also a strategic necessity. Cyberattacks are increasingly orchestrated by organized, persistent, and goal-oriented actors rather than isolated individuals; in critical sectors such as finance, energy, transportation, and healthcare, cyber disruptions can extend beyond operational risk to pose threats to national security. Consequently, redesigning intrusion detection systems with artificial intelligence support contributes to accelerating defensive mechanisms, reducing reliance on human intervention, and fostering more proactive security approaches. Machine learning- and deep learning-based systems surpass traditional security solutions through their capabilities to learn from historical

DOI: 10.46121/pspc.54.1.55

incidents, derive behavioural models, and develop predictive defences against novel attack types. This shift elevates the topic beyond an academic research agenda and positions it as a foundational component of future cyber defence architectures.

Nevertheless, although the literature published between 2020 and 2025 reflects considerable progress, several gaps remain. A substantial portion of existing research conducts model development on specific datasets with limited scope, while cross-dataset generalization across heterogeneous network environments is often overlooked. Even when high accuracy is reported, models frequently exhibit notable limitations in terms of real-time performance, explainability, resource efficiency, and robustness to class imbalance. Moreover, the scarcity of comprehensive, holistic, and multi-layered AI-based approaches to intrusion detection in the Turkish-language literature significantly constrains knowledge production and its transfer to institutional practice. In particular, the limited number of studies that address machine learning and deep learning methods in a broad, systematic, and comparative manner constitutes one of the most salient research gaps targeted by this work.

In this study, a range of algorithms employed in AI-driven intrusion detection are compared under diverse data conditions, and a multi-layered framework is constructed in which machine learning and deep learning are jointly leveraged for imbalanced and high-dimensional settings. Accordingly, alongside classical techniques, convolutional neural networks, long short-term memory mechanisms, autoencoders, federated learning paradigms, and generative adversarial network-based generators are examined within a unified evaluation scheme. Beyond accuracy, additional criteria—such as false alarm rates, processing time, explainability, and robustness to class imbalance—are explicitly considered. The goal is not solely to achieve high accuracy, but to develop a more flexible and sustainable defense structure that can adapt to evolving network conditions.

Considering machine learning and deep learning in tandem enables meaningful comparisons across different network types and datasets, and, in line with the prominent trends of the 2020–2025 period, the class-imbalance problem is addressed through ML–DL combinations. In the selection process, not only accuracy but also factors such as processing time, resource consumption, explainability, and sustainability are of critical importance. Evaluating IoT, IIoT, and conventional networks together provides a broader perspective grounded in cross-dataset generalization and yields a holistic and actionable direction for the future of AI-based intrusion detection.

I. RELATED WORKS

Machine Learning (ML) and Deep Learning (DL) have become core enablers for anomaly and intrusion detection, particularly under practical constraints such as class imbalance, noise, and incomplete labeling. Classical learners and representation-based DL models are repeatedly reported to achieve strong detection performance when combined with appropriate sampling and feature engineering. Recent trends suggest that temporal and contextual modeling—via RNN/GNN/GAN families—can further improve real-time detection by capturing sequential dependencies and relational structure [1].

Hybrid architectures that combine ML and DL are emphasized for balancing accuracy and operational cost. For instance, LSTM-based autoencoder representations coupled with tree-based ensembles can yield high accuracy with reduced false positives, while stream-processing stacks and feature reduction/selection reduce computational overhead; such pipelines have reported ~92.5% accuracy for zero-day scenarios [2]. However, deep NIDS still face persistent bottlenecks—training latency, false alarms, and imbalance sensitivity—across variants such as CNN+LSTM, BiLSTM-DCNN, LSTM-AE, and GAN-assisted designs [3]. Explainability is increasingly addressed through post-hoc methods: GAN-supported minority synthesis with CNN/LSTM feature learning and LIME-based interpretation improves transparency, with very high class-specific accuracies reported for certain attacks [4]. Beyond detection, closing the loop via automated response actions is proposed as a pathway to holistic defense [5]. For large-scale DDoS, multi-task DL can improve generalization through cross-task knowledge sharing, achieving near-perfect accuracies on contemporary benchmark variants [6].

Benchmark-driven comparisons underline that supervised methods generally outperform unsupervised alternatives on CICIDS2017, though severe class imbalance degrades rare-class detection and makes time cost a decisive deployment factor [7]. Feature selection can preserve accuracy with smaller subsets [8] and yields competitive results on CICIDS2017 for ensembles and tree learners [9-10]. For zero-day emphasis, one-class and density-based detectors trained on normal traffic (e.g., LOF, OCSVM) can be advantageous, albeit with false-alarm trade-offs [11-12]. Cross-dataset pipelines in IoT/IIoT highlight that gradient-boosted trees often provide strong accuracy with low false positives, whereas deep tabular models may be more variable and costly to train [13]. On UNSW-NB15, Random Forest frequently leads among standard baselines [14].

Resource-constrained IoT settings further motivate lightweight, privacy-preserving, and transferable solutions. Stacking ensembles with CatBoost/Extra Trees/XGBoost achieve high F1 scores on TON_IoT, with MCC noted as a robust imbalance-aware indicator [15]. Federated learning combined with deep autoencoders supports scalable on-device learning, and sequential/stacked ensemble components can reduce overfitting while maintaining high accuracy [16]. Cross-dataset generalization above 94% has been reported with hybrid pipelines

DOI: 10.46121/pspc.54.1.55

that integrate feature pruning and imbalance handling [17], while LightGBM is often preferred for efficiency despite heavier ensembles sometimes scoring higher [18]. Oversampling can substantially improve challenging device categories [19], k-NN has shown strong performance in vehicular TON-IoT settings [20], and fast boosting variants such as AdaBoost can deliver very high accuracy with low training time for IoT DDoS detection [21].

MATERIAL AND METHOD

In the modern cybersecurity ecosystem, the data generated across network traffic, IoT devices, and system components is rapidly increasing in both volume and diversity, while the nature of attacks is becoming progressively more complex. In this landscape, high-quality open datasets such as NFCSECICIDS2018 [22], CICIDS2017 [23], ToN-IoT [24], and UNSW-NB15 [25] constitute foundational resources for developing ML and DL–based anomaly detection models. Because these datasets were produced across different periods, attack taxonomies, and network architectures, they are critical for benchmarking model performance, designing hybrid approaches, and evaluating real-time detection strategies.

The CICIDS001 dataset is a small-scale reference corpus created by the Canadian Institute for Cybersecurity and includes contemporary forms of network-based attacks, such as DDoS, PortScan, Botnet, and Brute Force. Owing to its relatively limited class imbalance, CICIDS001 provides a suitable starting point for classical ML models. In contrast, NFCSECICIDS2018 was generated under a similar infrastructure but spans broader time intervals, includes diverse service protocols (e.g., HTTP, HTTPS, FTP, SSH), and offers a richer feature space with nearly 80 attributes. In particular, its flow-based features enable time-window–based statistical analyses. The CICIDS2017 dataset is comparatively more balanced and is released in a format segmented by attack days in addition to normal traffic, making it well-suited for assessing temporal generalization.

The ToN-IoT dataset contains network, system, sensor, and telemetry data collected from IoT (Internet of Things) devices and reflects threats emerging in both industrial and civilian IoT infrastructures (e.g., smart homes, smart cities, Industry 4.0). It facilitates modelling threats at the cyber–physical interface, such as sensor-driven anomalies (e.g., temperature deviations) or unexpected surges in network packet rates. The UNSW-NB15 dataset can be viewed as a modernized counterpart of KDD Cup 99; it includes 49 features encompassing contemporary protocols (e.g., IPv6, SSH, HTTPS) and combines low-level attributes with higher-level application-oriented features.

Collectively, these datasets represent a spectrum of anomaly detection methodologies, ranging from classical ML models with linear decision boundaries to more expressive DL architectures. Logistic Regression (LR) [26] serves as one of the simplest yet most interpretable baselines. On datasets with balanced labeling structures such as CICIDS2017, LR estimates the probability distribution between attack and normal classes via the sigmoid function. In addition, it supports feature-importance analysis by explicitly quantifying the influence of variables (e.g., “Flow Bytes/s” or “Active Mean”) on attack propensity.

Random Forest (RF) [27], as an ensemble method, offers strong generalization capability in high-dimensional and complex network traffic. On ToN-IoT and UNSW-NB15, interactions across protocol layers (e.g., IoT sensor traffic combined with network protocol behavior) can be effectively captured by RF’s decision-tree structure. Moreover, RF-derived variable importance scores can assist system administrators in identifying which sensor or protocol fields exhibit the largest shifts during attacks, thereby contributing to explainable AI (XAI) perspectives. Extreme Gradient Boosting (XGBoost) [28] is among the most effective models for high-variance and noisy network traffic. Datasets such as CICIDS2018 and UNSW-NB15 include thousands of samples and numerous attack categories. Through gradient-based error correction, XGBoost mitigates overfitting risk and can capture fine-grained attack signatures. It is particularly advantageous for detecting low signal-to-noise attacks such as slow-rate DDoS and botnet beaconing, where classical methods often underperform.

K-Nearest Neighbors (kNN) [29] estimates attack likelihood by analyzing neighbourhood relations in high-dimensional spaces. In ToN-IoT sensor traffic, devices operating within the same physical environment typically yield samples clustered in close proximity; deviations from these clusters during attacks constitute an anomaly signal for KNN. However, the computational cost of kNN is substantial, and for large-scale datasets such as CICIDS2018 with 500,000+ flows, KD-tree or Ball-tree optimizations become necessary.

DOI: 10.46121/pspc.54.1.55

Support Vector Machines (SVM) [30] model non-linear separations using kernels such as the radial basis function (RBF), maximizing the margin in high-dimensional feature spaces. This approach is particularly effective in detecting low-frequency attacks—such as "Infiltration" in CICIDS2017—where sample scarcity is a challenge. A key advantage of SVM is its ability to maintain a robust decision boundary even with relatively limited samples. Nonetheless, as dataset size increases (e.g., in ToN-IoT sensor traffic), scalability can become a limiting factor.

Isolation Forest (IF) [31], as an unsupervised anomaly detection method, isolates unusual observations through short partitioning paths. In unlabelled IoT data contexts such as ToN-IoT or NFCSECICIDS2018, IF can detect previously unknown attack types. This capability is particularly important for zero-day attacks or unpredictable threats such as data injection. A major strength of IF is that it does not require labelled data and can be deployed efficiently in an end-to-end manner.

On the deep learning side, 1D-CNN architectures are used to learn time-series patterns in network traffic. For example, in CICIDS2017, abrupt spikes in TCP packet counts or variations in "Flow Duration" can be captured as local patterns through convolutional filters. By detecting low-level irregularities directly from the data, 1D-CNNs reduce reliance on manual feature engineering.

Hybrid CNN–LSTM models capture both the structural and temporal dimensions of network traffic, enabling the learning of evolving manifestations of attacks such as DDoS or brute force. In CICIDS2018, traffic dynamics across multiple stages of an attack (e.g., reconnaissance → exploitation → exfiltration) can be tracked over time. LSTM cells preserve long-range dependencies and encode behavioural traces of multi-step attack sequences.

Autoencoder CNN (AE-CNN) reconstructs network traffic, representing normal flows with low reconstruction error and anomalous flows with higher reconstruction error. Given the label imbalance in datasets such as ToN-IoT and UNSW-NB15, AE-CNN can be employed in unsupervised or self-supervised settings. By monitoring reconstruction loss, the model flags instances exceeding a predefined threshold as attacks.

Among more advanced architectures, AquaGraph-TConv (Multi-scale Temporal Graph Convolution) is particularly relevant for ToN-IoT scenarios. Due to the inherent characteristics of IoT sensor networks, signals exhibit distinct behaviors across multiple temporal scales (milliseconds–seconds–minutes). AquaGraph-TConv applies graph-based convolutions across multi-scale temporal windows, enabling it to learn both short-term anomalies (e.g., sudden packet bursts) and long-term anomalies (e.g., sensor drift or data leakage). Its autoencoder-like structure calibrates itself by minimizing reconstruction error.

In summary, CICIDS001, NFCSECICIDS2018, CICIDS2017, ToN-IoT, and UNSW-NB15 constitute complementary data resources that collectively represent multiple security layers spanning network traffic, IoT sensing, and system-level telemetry. While simple models such as Logistic Regression provide interpretable baselines for anomaly detection, ensemble methods such as Random Forest and XGBoost can deliver high accuracy in large-scale network settings. In contrast, deep models such as CNN–LSTM and AquaGraph-TConv capture both temporal dependencies and system–topology relationships, offering stronger suitability for today's multi-dimensional cyber threat landscape. Consequently, the integration of ML and DL forms the foundation of modern cyber defense systems, enabling adaptive and predictive security architectures.

RESULTS AND DISCUSSION

In our study, in which both classical machine learning and deep learning models were evaluated on four cyberattack datasets—CICIDS2017, CICIDS2018, UNSW-NB15, and ToN-IoT—the results reveal a consistent performance ranking across all datasets. Overall, tree-based methods (Random Forest, XGBoost) and convolutional neural network–based architectures (CNN–1D, GNN–CNN) achieved the highest accuracy, whereas the semantic-encoding–oriented Autoencoder–CNN model exhibited comparatively weak performance. On the CICIDS2017 dataset, XGBoost and Random Forest achieved accuracies of 100% and 99.99%, respectively. The fact that both AUROC and AUPRC values reached 1.0000 indicates an almost error-free separation between attack and normal traffic. Deep learning models such as CNN–1D and GNN–CNN also delivered accuracies above 99.9%. Although linear methods such as Logistic Regression and SVM achieved strong performance at approximately 99% accuracy, their capacity to capture complex traffic dynamics remained limited. By contrast, Isolation Forest achieved only 73.5% accuracy, suggesting that unsupervised approaches did not attain sufficient representational power for this dataset.

DOI: 10.46121/pspc.54.1.55

The results on CICIDS2018 follow a pattern similar to the 2017 version. Random Forest, XGBoost, CNN–1D, CNN–LSTM, AquaGraph–TConv, and GNN–CNN attained perfect discrimination with 100% accuracy. Logistic Regression and SVM maintained robust performance with 99.99% accuracy. In contrast, Isolation Forest reached only around 60% accuracy. This gap indicates that unsupervised anomaly detection algorithms tend to be less effective than supervised approaches in modern traffic datasets.

A comparable model hierarchy was observed on UNSW-NB15. XGBoost, Random Forest, and CNN-based models exceeded 99.9% accuracy. Logistic Regression and SVM achieved approximately 99.9%, while KNN yielded around 99.8%. However, the Autoencoder–CNN architecture remained at 95% accuracy. This finding suggests that classical machine learning models, when supported by appropriate feature extraction and preprocessing, can still be competitive with deep learning. In particular, XGBoost's high AUROC and AUPRC values demonstrate that the model can produce stable decision boundaries even under class imbalance.

On the ToN-IoT dataset, which reflects IoT-oriented traffic, the results were more heterogeneous. Random Forest and XGBoost again exceeded 99% accuracy, whereas KNN and CNN–1D remained in the 97–98% range. CNN–LSTM and GNN–CNN achieved strong overall performance with 96–97% accuracy. Logistic Regression was limited to approximately 90%, while Isolation Forest achieved about 73%, and Autoencoder–CNN dropped below 50%. These outcomes suggest that, in complex IoT traffic, deep convolutional layers and hybrid graph-based structures (e.g., AquaGraph–TConv, GNN–CNN) provide superior generalization compared with simpler baselines.

Considering all datasets jointly, XGBoost emerges as the overall leader in terms of both speed and accuracy. Its near-perfect classification performance, together with a low training time (e.g., only 4 seconds on ToN-IoT, makes it particularly suitable for real-time network monitoring systems. Random Forest likewise demonstrated high stability. Among deep learning approaches, CNN–1D, GNN–CNN, and AquaGraph–TConv were especially effective at modeling complex temporal dependencies.

In summary, XGBoost and Random Forest consistently yielded the highest accuracy, AUROC, and F1 scores. CNN–1D, CNN–LSTM, AquaGraph–TConv, and GNN–CNN provided the most reliable results among deep learning models. By contrast, Isolation Forest and Autoencoder–CNN exhibited lower stability, particularly on imbalanced datasets and high-dimensional IoT data. These observations also highlight that preprocessing, normalization, and feature engineering directly influence the performance of classical models.

Table I presents the results for CICIDS2017, Table 2 for CICIDS2018, Table 3 for UNSW-NB15, and Table 4 for ToN-IoT. Overall, the findings suggest that hybrid approaches may play an increasingly important role in AI-driven network traffic analysis, and that real-time, low-latency threat detection infrastructures—especially for IoT systems—can be effectively designed using such methods.

TABLE I
COMPARATIVE ANOMALY DETECTION PERFORMANCE OF ML AND DL MODELS ON THE CICIDS2017 DATASET

Model	AUROC	AUPRC	F1	TN	FP	FN	TP	Accuracy
LR	0.9998	0.9997	0.9988	14645	13	34	19170	99.93
kNN	1.0000	1.0000	0.9996	14652	6	8	19196	99.96
SVM	0.9994	0.9994	0.9990	14648	10	29	19175	99.88
RF	1.0000	1.0000	0.9999	14657	1	1	19203	99.99
IF	0.7689	0.7743	0.7169	12228	2430	7117	12087	73.5
XGBoost	1.0000	1.0000	1.0000	14658	0	0	19204	100.00
1D-CNN	0.9999	0.9997	0.9998	14652	6	3	19201	99.98
AquaGraph–TConv	0.9998	0.9998	0.9948	14587	71	127	19077	99.38
AE–CNN	0.2317	0.4168	0.0000	14656	2	19204	0	43.2
CNN–LSTM	0.9948	0.9959	0.9828	14428	230	428	18776	98.6
GNN–CNN	0.9999	0.9999	0.9995	14650	8	10	19194	99.95

DOI: 10.46121/pspc.54.1.55

Power System Protection and Control
ISSN:1674-3415

TABLE III
COMPARATIVE ANOMALY DETECTION PERFORMANCE OF ML AND DL MODELS ON THE CICIDS2018 DATASET

Model	AUROC	AUPRC	F1	TN	FP	FN	TP	Accuracy
LR	1.0000	1.0000	1.0000	22500	0	2	22498	0.9999
kNN	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000
SVM	1.0000	1.0000	1.0000	22500	0	2	22498	0.9999
RF	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000
IF	0.6397	0.5339	0.8356	13648	8852	0	22500	0.6062
XGBoost	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000
1D-CNN	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000
AquaGraph–TConv	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000
AE–CNN	0.0014	0.3070	0.0000	22496	4	22500	0	0.4999
CNN–LSTM	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000
GNN–CNN	1.0000	1.0000	1.0000	22500	0	0	22500	1.0000

TABLE IIIII
COMPARATIVE ANOMALY DETECTION PERFORMANCE OF ML AND DL MODELS ON THE UNSWNB15 DATASET

Model	AUROC	AUPRC	F1	TN	FP	FN	TP	Accuracy
LR	0.9996	0.9947	0.9911	335349	311	34	19120	0.9993
kNN	0.9997	0.9984	0.9843	113483	197	5	6315	0.9982
SVM	0.9999	0.9969	0.9925	335382	278	11	19143	0.9999
RF	1.0000	1.0000	0.9997	335651	9	4	19150	0.99996
IF	0.8680	0.1725	0.3011	260976	74684	2525	16629	0.8680
XGBoost	1.0000	1.0000	0.9998	335652	8	0	19154	1.0000
1D-CNN	1.0000	0.9998	0.9990	335629	31	8	19146	0.9999
AquaGraph–TConv	1.0000	0.9994	0.9860	335142	518	27	19127	0.9984
AE–CNN	0.9984	0.9505	0.7009	638650	32670	15	38293	0.9532
CNN–LSTM	1.0000	0.9999	0.9972	85096	24	3	4877	0.9997
GNN–CNN	1.0000	0.9997	0.9986	335616	44	8	19146	0.9999

TABLE IVV
COMPARATIVE ANOMALY DETECTION PERFORMANCE OF ML AND DL MODELS ON THE Ton-IoT DATASET

Model	AUROC	AUPRC	F1	TN	FP	FN	TP	Accuracy
LR	0.9628	0.9374	0.9303	21573	927	2126	20374	0.9045
kNN	0.9945	0.9949	0.9717	22277	223	1029	21471	0.9816
SVM	0.9785	0.9841	0.9503	21849	651	1543	20957	0.9599
RF	0.9993	0.9994	0.9901	22344	156	288	22212	0.9921
IF	0.6214	0.5167	0.7730	11889	10611	1640	20860	0.7348
XGBoost	0.9992	0.9993	0.9880	22311	189	347	22153	0.9886
1D-CNN	0.9934	0.9947	0.9706	22124	376	930	21570	0.9768
AquaGraph–TConv	0.9928	0.9942	0.9695	22307	193	1150	21350	0.9773
AE–CNN	0.3814	0.3999	0.0000	22495	5	22500	0	0.4999
CNN–LSTM	0.9912	0.9931	0.9641	22105	395	1194	21306	0.9647
GNN–CNN	0.9925	0.9940	0.9690	22176	324	1050	21450	0.9695

DOI: 10.46121/pspc.54.1.55

All experimental findings clearly indicate that machine learning and deep learning approaches can play complementary roles in network-based cyberattack detection. Although classical statistical models can achieve high performance on certain datasets, the results demonstrate that deep learning–based models exhibit superior generalization capability in complex and high-dimensional traffic scenarios. In particular, the ability of architectures such as CNN–1D and GNN–CNN to jointly capture spatial and temporal dependencies in time-series network traffic constitutes a key factor explaining the observed performance gap.

Across all experiments, XGBoost consistently delivered the highest accuracy and AUROC values on all datasets. This success can be attributed to the gradient-boosting optimization of decision trees, which provides strong discriminative power with low training time and high stability. Owing to its favorable computational profile, XGBoost offers a practical solution for direct deployment in real-time IDS and Security Information and Event Management infrastructures. Moreover, its relatively low computational cost suggests that it can be effectively utilized in embedded platforms and resource-constrained IoT devices.

In contrast, Isolation Forest and Autoencoder–CNN generally exhibited lower performance. These outcomes highlight the challenges of applying unsupervised learning approaches directly to labeled cybersecurity datasets. The complete failure of Autoencoder–CNN on certain datasets further suggests that reconstruction-based models may lose generalization capability when hyperparameters are not appropriately tuned. Future work may achieve more robust and balanced results by augmenting such unsupervised architectures with transfer learning or semi-supervised learning strategies.

Graph-based deep learning architectures, including AquaGraph–TConv and GNN–CNN, achieved high accuracy particularly on the ToN-IoT dataset. This finding indicates that models aligned with the topological nature of IoT networks can provide a critical advantage in intrusion detection. While irregular data streams and heterogeneous sensor topologies in IoT environments can limit the learning capacity of conventional CNN- or LSTM-based models, graph convolutional networks can represent and exploit this structural complexity in a more natural and effective manner.

TABLE V
BINARY CLASSIFICATION PERFORMANCE OF ML/DL AND ANOMALY-BASED APPROACHES ON THE CICIDS2017 DATASET

Model	Type	Accuracy	F1 (Normal – 0)	F1 (Anomali – 1)	ROC-AUC
Logistic Regression	ML	0.9986	0.9983	0.9988	0.9998
KNN	ML	0.9997	0.9997	0.9997	1.0000
SVM-RBF	ML	0.9984	0.9982	0.9986	0.9998
Random Forest	ML	1.0000	0.9999	1.0000	1.0000
XGBoost	ML	1.0000	1.0000	1.0000	1.0000
Isolation Forest	Anomaly	0.4223	0.5938	0.0000	–
CNN–1D	DL	0.9991	0.9989	0.9992	1.0000
AE–CNN	DL	0.9933	0.9922	0.9941	0.9997
CNN–LSTM	DL	0.9991	0.9989	0.9992	0.9999
GNN–CNN	DL	0.9993	0.9992	0.9994	1.0000
AquaGraph–TConv	DL	0.9334	0.9168	0.9445	0.9992

Experiments conducted on the CICIDS2017 dataset demonstrate that supervised machine learning and deep learning–based approaches achieve near-perfect performance for anomaly (attack) detection. Among the models trained on labeled data (0 = normal, 1 = attack), Random Forest, XGBoost, GNN–CNN, CNN–LSTM, and CNN–1D stood out by attaining F1-scores above 99% for the attack class and ROC–AUC values in the 0.999–1.000 range, indicating very high recall and extremely low false-negative rates.

In contrast, the unsupervised Isolation Forest—despite correctly identifying the normal class to a large extent (recall ≈ 0.98)—was almost unable to detect the attack class (F1 ≈ 0). This outcome highlights that, in attack-dominant datasets such as CICIDS2017, unsupervised methods require careful calibration, particularly of the contamination parameter. Although the AquaGraph–TConv model exhibited strong separability in terms of ROC–

DOI: 10.46121/pspc.54.1.55

AUC (≈ 0.999), it lagged behind other deep learning models with an accuracy of approximately 93% due to the fixed 0.5 decision threshold and class imbalance. This suggests that its performance could be improved through threshold optimization and more advanced or longer training strategies. Table 5 presents the anomaly detection results for CICIDS2017.

TABLE VI
SUMMARY OF ANOMALY DETECTION PERFORMANCE ON THE UNSW-NB15 DATASET

Model	Accuracy	F1-Score	ROC-AUC
Logistic Regression	0.9996	0.9959	0.9997
K-NN	0.9998	0.9983	0.9999
SVM-RBF	0.9998	0.9986	0.9997
Random Forest	1.0000	0.9999	1.0000
XGBoost	1.0000	0.9999	1.0000
Isolation Forest	0.9368	0.0102	0.498
CNN-1D	0.9998	0.9986	0.9998
AE-CNN	0.9998	0.9985	1.0000
CNN-LSTM	0.9999	0.9989	1.0000
GNN-CNN	0.9998	0.9985	0.9999
AquaGraph–TConv	0.9904	0.9183	1.0000

The analysis conducted on the UNSW-NB15 dataset indicates that supervised machine learning and deep learning approaches deliver near-perfect performance in anomaly detection. All supervised models—including Logistic Regression, K-NN, SVM-RBF, Random Forest, and XGBoost—distinguished the attack class with high precision, achieving accuracy and F1-score values above 99%. In particular, Random Forest and XGBoost demonstrated flawless discrimination with a ROC-AUC of 1.0. Deep learning architectures such as CNN-1D, AE-CNN, CNN-LSTM, and GNN-CNN similarly exhibited strong generalization capability, attaining accuracy and F1-scores in the 99.8%–99.9% range.

In contrast, the unsupervised Isolation Forest correctly identified normal traffic but failed to capture attack instances, indicating that it is inadequate when used in isolation on imbalanced datasets such as UNSW-NB15. Although the AquaGraph–TConv model provided perfect separability in terms of ROC-AUC, its accuracy decreased slightly (from above 99% to approximately 99%) due to the fixed 0.5 decision threshold and class imbalance. Overall, the results demonstrate that supervised approaches achieve the highest levels of accuracy and generalization on UNSW-NB15, whereas unsupervised methods exhibit limited performance in anomaly detection without careful parametric optimization. Table 6 presents the anomaly detection results for UNSW-NB15.

TABLE VII
SUMMARY OF ANOMALY DETECTION PERFORMANCE ON THE TON-IOT DATASET

Model	Accuracy	F1-Score	ROC-AUC
Logistic Regression	0.9273	0.9264	0.9626
K-NN	0.9707	0.9707	0.9903
SVM-RBF	0.9528	0.9528	0.9802
Random Forest	0.9819	0.9819	0.9979
XGBoost	0.9813	0.9813	0.9975
Isolation Forest	0.4929	0.0061	0.333
CNN-1D	0.9445	0.9453	0.9884
AE-CNN	0.9548	0.9548	0.9887
CNN-LSTM	0.9602	0.9602	0.9892
GNN-CNN	0.9568	0.9567	0.9877
AquaGraph–TConv	0.9084	0.9115	0.9779

DOI: 10.46121/pspc.54.1.55

The analysis conducted on the ToN-IoT dataset indicates that complex anomalies in IoT traffic observed in Cyber–Physical Systems (CPS) are more challenging to separate than those in conventional network datasets. Overall, supervised machine learning models achieved strong performance; in particular, Random Forest and XGBoost reliably discriminated attack patterns in network traffic with accuracies above 98% and near-perfect ROC-AUC values (≈0.998). Among classical algorithms, K-NN—arguably the most sensitive to IoT-driven behavioral patterns—produced an F1-score of approximately 97%, whereas Logistic Regression and SVM, with comparatively lower accuracy (92%–95%), captured complex IoT behaviors only to a limited extent.

Among deep learning architectures, CNN-LSTM and GNN-CNN successfully learned time-dependent IoT traffic, achieving accuracies around 96% and ROC-AUC scores close to 0.99; AE-CNN also yielded stable outcomes in feature representation. In contrast, the unsupervised Isolation Forest failed to correctly distinguish the anomaly class in IoT traffic, exhibiting high recall for the normal class (0.98) but nearly zero recall for attack instances (0.003). Although AquaGraph–TConv demonstrated strong separability in terms of ROC-AUC, its accuracy remained around 90% due to the fixed decision threshold and class imbalance. Overall, the results suggest that supervised models deliver superior performance on ToN-IoT, particularly for complex IoT anomalies, whereas unsupervised methods remain weak in high-dimensional and heterogeneous sensor data. Table 7 presents the anomaly detection results for ToN-IoT.

Integrating strong feature selectors such as Random Forest or XGBoost with CNN–LSTM or GNN–CNN architectures may further improve both learning efficiency and real-time detection speed. In particular, designs that combine XGBoost's decision boundaries with CNN-based feature extraction appear promising for low-latency threat detection systems. Furthermore, in IoT settings, transferring data to centralized servers may introduce privacy risks. In this regard, federated learning—where model parameters are trained on edge devices—can offer a strategic solution by supporting KVKK/GDPR compliance while simultaneously reducing the attack surface.

In critical domains such as intrusion detection, human interpretability of the model's decision mechanisms is essential. Future work may therefore analyze the best-performing models in this study—especially XGBoost and CNN–1D—using explanation techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), thereby contributing to the development of trustworthy and auditable IDS infrastructures.

The findings further indicate that comprehensive comparative experiments across multiple datasets demonstrate that both classical machine learning approaches and deep learning–based methods can provide high accuracy and practical deployability in network security. Nevertheless, future research should emphasize integration into real-time environments, support via hardware acceleration options such as FPGA or edge GPUs, and systematic treatment of explainability. These dimensions are expected to be among the key determinants shaping the future standards of AI-driven cybersecurity monitoring and auditing mechanisms.

CONCLUSIONS

This study presented a comprehensive benchmarking analysis of machine learning and deep learning approaches for anomaly-based intrusion detection across multiple cybersecurity datasets, including CICIDS2017, CICIDS2018, UNSW-NB15, and ToN-IoT. The experimental results demonstrate that both classical machine learning models and modern deep learning architectures can achieve high detection accuracy when applied to large-scale network traffic datasets. In particular, tree-based ensemble models such as XGBoost and Random Forest consistently delivered the highest performance across all datasets, achieving near-perfect accuracy, AUROC, and F1-score values while maintaining relatively low computational cost. The findings also highlight that deep learning architectures—including CNN–1D, CNN–LSTM, and GNN–CNN—are particularly effective in capturing complex temporal and structural dependencies in network traffic. These models demonstrated strong generalization capabilities, especially in high-dimensional and heterogeneous environments such as IoT-based networks. However, unsupervised approaches such as Isolation Forest and reconstruction-based architectures such as AE–CNN exhibited significantly lower performance in several scenarios, indicating that labeled and supervised learning strategies currently remain more reliable for practical intrusion detection applications. Another important outcome of this study is the observation that no single model provides optimal performance under all conditions. While classical machine learning models showed strong stability and efficiency, deep learning models offered better representation learning for complex traffic patterns. This suggests that hybrid architectures combining

DOI: 10.46121/pspc.54.1.55

feature-selection capabilities of tree-based methods with the representation power of deep neural networks may offer the most promising direction for future intrusion detection systems. Future research should focus on improving cross-dataset generalization, integrating explainable AI techniques such as SHAP or LIME for model transparency, and exploring privacy-preserving learning frameworks such as federated learning for distributed IoT environments. Additionally, hardware acceleration through edge computing, GPUs, or FPGA-based implementations may further enable the deployment of AI-driven intrusion detection systems in real-time operational environments.

REFERENCES

1. S. Kumar, S. Dua, and S. Rastogi, “Anomaly detection: A machine learning and deep learning perspective,” in Proceedings of the 2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3), pp. 1-6, 2023.
2. S. Banerjee, B. Basak, S. Mandal, and A. Das, “Real-time anomaly detection in network traffic: A hybrid ML-DL approach,” in Proceedings of the 2025 International Conference on Engineering Innovations and Technologies (ICoEIT), pp. 536-539, 2025.
3. G. Nidhishree and K. Vidyalakshmi, “Comprehensive analysis on anomaly detection in cybersecurity dataset using deep learning algorithms,” in Proceedings of the 2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC), pp. 1-4, 2024.
4. M. J. Hossain, K. Alam, M. F. Monir, M. M. Hoque, and T. Ahmed, “Explainable AI meets synthetic data: A deep learning framework for detecting network intrusion in NextG network infrastructure,” IEEE Access, vol. 13, pp. 114979-115001, 2025.
5. S. Ramya and S. A. E. Selvi, “Anomaly detection in network traffic data using deep learning,” in Proceedings of the 2024 2nd International Conference on Computing and Data Analytics (ICCDA), pp. 1-5, 2024.
6. M. F. Abyandani, P. Sukarno, and A. A. Wardana, “Distributed denial-of-service (DDoS) detection using multitask learning based on deep learning,” in Proceedings of the 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 383-388, 2025.
7. Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, and C. F. M. Foozy, “Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset,” IEEE Access, vol. 9, pp. 22351-22370, 2021.
8. T. Janarthanan and S. Zargari, “Feature selection in UNSW-NB15 and KDDCUP99 datasets,” in Proceedings of the 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), pp. 1881-1886, 2017.
9. K. Kurniabudi, D. Stiawan, D. Darmawijoyo, M. Y. B. Idris, A. M. Bamhdi, and R. Budiarto, “CICIDS-2017 dataset feature analysis with information gain for anomaly detection,” IEEE Access, vol. 8, pp. 132911-132921, 2020.
10. M. S. Raza, H. Arif, S. M. Errapotu, and V. Gonzalez, “Comprehensive analysis of MLP and ensemble learning approaches for intrusion detection using CICIDS2017 dataset,” in Proceedings of the 2024 IEEE Future Networks World Forum (FNWF), pp. 476-482, 2024.
11. T. Elmasri, N. Samir, M. Mashaly, and Y. Atef, “Evaluation of CICIDS2017 with qualitative comparison of machine learning algorithm,” in Proceedings of the 2020 IEEE Cloud Summit, pp. 46-51, 2020.

DOI: 10.46121/pspc.54.1.55

12. Z. Xu and Y. Liu, “Robust anomaly detection in network traffic: Evaluating machine learning models on CICIDS2017,” in Proceedings of the 2025 10th International Conference on Electronic Technology and Information Science (ICETIS), pp. 475-482, 2025.
13. H.-C.-T. Nguyen, X.-H. Nguyen, and K.-H. Le, “An automated benchmarking framework for anomaly-based intrusion detection systems,” in Proceedings of the 2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pp. 1-6, 2024.
14. U. C. Akuthota and L. Bhargava, “Evaluation of machine learning models for intrusion detection with the UNSW-NB15 dataset,” in Proceedings of the 2023 IEEE Silchar Subsection Conference (SILCON), pp. 1-5, 2023.
15. G. Guo, X. Pan, H. Liu, F. Li, L. Pei, and K. Hu, “An IoT intrusion detection system based on TON IoT network dataset,” in Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), pp. 333-338, 2023.
16. F. Naeem, A. W. Malik, S. A. Khan, and F. Jabeen, “Enhancing intrusion detection: Leveraging federated learning and hybrid machine learning algorithms on ToN_IoT dataset,” in Proceedings of the 2023 International Conference on Frontiers of Information Technology (FIT), pp. 73-78, 2023.
17. S. M. M. Ahmed, S. K. M. S. Islam, and M. S. Ullah, “Hybrid machine learning and deep learning approaches for anomaly detection using KD99 and TON_IoT datasets,” in Proceedings of the 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-6, 2025.
18. S. Ismail, S. Dandan, and A. Qushou, “Intrusion detection in IoT and IIoT: Comparing lightweight machine learning techniques using TON_IoT, WUSTL-IIOT-2021, and EdgeIIoTset datasets,” IEEE Access, vol. 13, pp. 73468-73485, 2025.
19. E. Kalifati, Ş. Şolpan, K. K. Eren, and K. Küçük, “Performance analysis of SMOTE-enhanced machine learning approaches on IoT-based TON IoT dataset,” in Proceedings of the 2025 33rd Signal Processing and Communications Applications Conference (SIU), pp. 1-4, 2025.
20. H. Sharma, H. Babbar, and A. Sharma, “TON-IoT: Detection of attacks on Internet of Things in vehicular networks,” in Proceedings of the 2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 539-545, 2022.
21. F. Chishti and G. Rathee, “ToN-IoT set: Classification and prediction for DDoS attacks using AdaBoost and RUSBoost,” in Proceedings of the 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 2842-2847, 2023.
22. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "CSE-CIC-IDS2018 on AWS: Intrusion Detection Dataset," Canadian Institute for Cybersecurity (CIC), 2018. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html
23. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Intrusion Detection Evaluation Dataset (CIC-IDS2017)," Canadian Institute for Cybersecurity (CIC), 2017. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html
24. A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "ToN_IoT Telemetry Dataset," UNSW Canberra Cyber, 2020. [Online]. Available: https://research.unsw.edu.au/projects/toniot-datasets

DOI: 10.46121/pspc.54.1.55

25. N. Moustafa and J. Slay, "UNSW-NB15 Dataset," UNSW Canberra Cyber, 2015. [Online]. Available: https://research.unsw.edu.au/projects/unsw-nb15-dataset
26. Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215-232.
27. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
28. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
29. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
30. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
31. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 413-422.

DOI: 10.46121/pspc.54.1.55