

ARTICLE

Explainable deep learning intrusion detection system for robust and interpretable security in smart city networks

Vijayakumari Rodda^{1*}, Anurag Shrivastava², Aditya Rautaray³, Sundaram K. Meenakshi⁴, G. Shailaja⁵, S. Prabagar⁶, and S. B. G. Tilak Babu⁷¹Department of Computer Science, Faculty of Engineering and Technology, Krishna University, Machilipatnam, Krishna District, Andhra Pradesh, India²Department of Information Technology, Indiana Wesleyan University, Merrillville, Indiana, United States of America³Independent Research Scholar⁴Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India⁵Department of IT, CVR College of Engineering, Hyderabad, Telangana, India⁶Department of Computer Science & Engineering (Data Science), School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India⁷Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Aditya University, Surampalem, Andhra Pradesh, India***Corresponding author:**Vijayakumari Rodda
(vijayakumari28@gmail.com)

Citation: Rodda, V., Shrivastava, A., Rautaray, A., Meenakshi, S. K., Shailaja, G., Prabagar, S. & Babu, S. B. G. T. (2026). Explainable deep learning intrusion detection system for robust and interpretable security in smart city networks. *Int J Systematic Innovation*, 10(5): 026160059.

[https://doi.org/10.6977/IJoSI.202610_10\(5\).0006](https://doi.org/10.6977/IJoSI.202610_10(5).0006)

Received: April 17, 2026**Revised:** June 22, 2026**Accepted:** July 1, 2026**Published online:** August 3, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Smart urban systems are based on interconnected software-defined networking and Internet of Things platforms that face critical threats from advanced Layer 2 and zero-day attacks. Although deep learning provides powerful detection capabilities, the available models tend to be black boxes that do not meet security-forensics requirements and have high latency when redundant features are learned. The present research fills these gaps by proposing a powerful, explainable intrusion detection system based on a hybrid convolutional neural network and a bidirectional long short-term memory architecture, aided by an attention-based fusion mechanism. The framework combines domain-constrained feature selection with dual-layer explainability, using Shapley additive explanations and local interpretable model-agnostic explanations to provide both global and attack-specific model transparency. The experimental findings from the intrusion detection dataset for software-defined networking and the extended industrial Internet of Things intrusion detection dataset show that detection of known threats is 99.85%, and the detection of unknown zero-day vectors is also significant at 71.09%. The method reduces feature dimensionality by 77%, resulting in 4.1 ms of inference latency. Such a combination of high-performance classification and actionable interpretability enhances forensic reliability in urban infrastructure that is highly important to the mission.

Keywords: Explainable artificial intelligence; Software-defined networking; Zero-day attack detection; Hybrid deep learning

1. Introduction

The rapid pace of city digitalization has integrated software-defined networking (SDN) and Internet of Things (IoT) architectures into the fabric of contemporary infrastructure, forming the most efficient and vulnerable ecosystems that require sophisticated, transparent security systems to safeguard sensitive information.

1.1. Motivation and smart city infrastructure

Current smart city systems rely on integrating SDN and IoT to operate cities' mission-critical services (Lakshan Yasarathna & Le-Khac, 2026; Rahman *et al.*, 2020). These interrelated ecosystems support the real-time exchange of data on transportation, energy management, and public safety. The size of these networks means they are increasingly exposed to advanced cyber threats, such as nuanced Layer 2 vulnerabilities and more severe zero-day exploits (Admass *et al.*, 2024; Chaganti *et al.*, 2023; Wahab *et al.*, 2024). It is also important to ensure the security of such environments, as any disruption could result in significant economic losses or a direct threat to the population in the physical realm (Kampourakis *et al.*, 2025; Kampourakis *et al.*, 2026).

1.2. The "black-box" challenge in deep learning intrusion detection systems

Deep learning (DL) has become an effective tool for developing intrusion detection systems (IDSs); traditional models are often black boxes (Ahmad *et al.*, 2025; Houda *et al.*, 2022). This is a serious shortcoming of smart city forensics, as administrators must not only identify an attack but also gain insight into its nature to establish a successful defense (Keshk *et al.*, 2023). The current solutions often incur high computational overhead and latency due to learning redundant feature sets in high-speed networks (Singh & Jain, 2024; Vishwakarma & Jain, 2019). The inability to act quickly and forensic integrity are both threats to a mission-critical smart city, where every minute matters in detection, and an inexplicable alert can be a dead end (Ahmed *et al.*, 2022).

1.3. Identification of research gaps

There is still a considerable gap in the state of the art, as current IDS systems do not perform well at detecting zero-day attacks that traditional filters cannot detect (Arreche *et al.*, 2024; Shoukat *et al.*, 2025). The integration of domain-constrained SDN capabilities (Shoukat *et al.*, 2025) and dual-layer explainability, in particular, the synergy between Shapley additive explanations (SHAP) for global model transparency and local interpretable model-agnostic explanations (LIME) for packet-level explanation, is significantly lacking (Al-Hawawreh *et al.*, 2022; Shoukat

et al., 2025).

1.4. Proposed work and contributions

The proposed model addresses these issues by offering a powerful, interpretable IDS framework tailored to smart city networks. The essence of the method is a hybrid design in which convolutional neural networks (CNNs) extract spatial features, bidirectional long short-term memory (Bi-LSTM) models capture temporal dependencies, and an attention-based fusion mechanism focuses on important network flows. The system can detect over 99% of known attacks and a large 71% of zero-day threats using maximized feature selection with domain-constrained SDN measurements, and achieves a 67% reduction in inference latency (from 12.4 ms to 4.1 ms). The major contributions of this research are: (i) a new attention-based fusion mechanism framework of proactive zero-day detection based upon centroid-based proximity; (ii) an efficient explainable artificial intelligence (XAI)-based feature reduction algorithm that selects traffic according to SDN-IoT network logic; (iii) a clear decision-making channel that offers real-time forensics through dual SHAP/LIME integration. Figure 1 summarizes the strategic fit of these elements with the current smart city security issues.

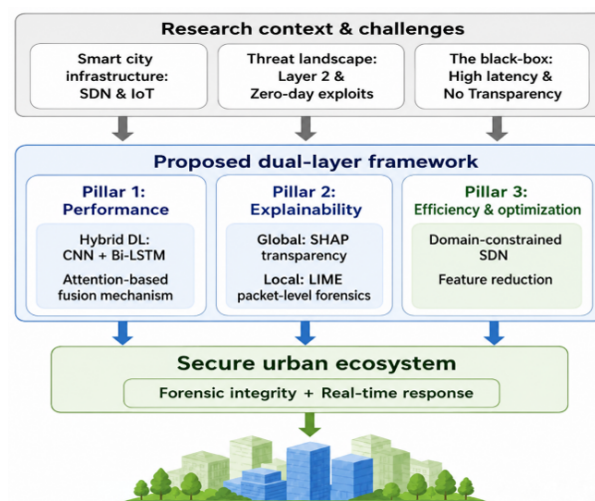


Figure 1. Strategic architecture of the proposed IDS and bridging the black-box and latency gaps in smart city networks through hybrid DL and dual-layer XAI

Abbreviations: Bi-LSTM: Bidirectional long short-term memory; CNN: Convolutional neural network; DL: Deep learning; IDS: Intrusion detection systems; LIME: Local interpretable model-agnostic explanation; SDN: Software-defined networking; SHAP: Shapley additive explanation; XAI: Explainable artificial intelligence.

1.5. Article organization

The rest of this paper is structured as follows: Section 2

is the literature review of related works on Layer 2 and SDN security, Section 3 provides the description of the proposed hybrid approach and feature engineering, Section 4 outlines the experimental setup, hardware, and datasets, Section 5 presents quantitative findings and XAI fidelity, and Section 6 concludes with the future research directions in federated learning in smart cities.

2. Evolution of intrusion detection from heuristics to explainable artificial intelligence

Smart city infrastructure security has become a sensitive research area as urban systems increasingly rely on SDN and IoT ecosystems (Altulaihan *et al.*, 2022; Hakak *et al.*, 2020; Kumpf *et al.*, 2025). To have a strong defense against cyber threats, it will be necessary to shift to dynamic, intelligent monitoring systems (Iftikhar *et al.*, 2025; Moustafa *et al.*, 2018; Rachakonda *et al.*, 2024) that can detect complex attack vectors in real time, rather than relying on traditional static defenses.

2.1. Evolution from machine learning to deep learning intrusion detection systems

Early models in network security primarily used traditional machine learning tools, including random forests and support vector machines, to categorize traffic patterns (Hasan *et al.*, 2014; Ioannou & Vassiliou, 2021; Pau *et al.*, 2024). These techniques offered more basic security; they were frequently based on manual feature engineering and could not scale to the high dimensionality and exponentially changing data streams of contemporary SDN environments (Altunay & Albayrak, 2023; Elsaid & Binbusayyis, 2024; Siddiqi & Pak, 2020). Later developments brought DL models, such as CNNs and recurrent neural networks, that automated the feature extraction and enhanced the detection of known threats (Imtiaz *et al.*, 2025; Oseni *et al.*, 2023). These black-box models are not transparent, meaning there is no explanation for why they are classified as such, which cripples their use in serious security forensics.

2.2. Hybrid architectures and explainable artificial intelligence

Recent developments have tried to overcome these shortcomings by incorporating hybrid architectures and early XAI frameworks (Agbor *et al.*, 2025). The combination of long short-term memory (LSTM) networks and gated recurrent units (Afraji *et al.*, 2025) has been studied for representing temporal dependencies (Kilincer, 2025). These have been improved; there are still a number of critical gaps. The majority of existing systems do not account for Layer 2 vulnerabilities and rely on redundant

feature sets that increase computational latency, which is why they cannot be deployed on the resource-constrained edge devices (Kilincer, 2025; Mishra *et al.*, 2025). The current literature does not effectively address zero-day threats, as traditional models are still likely to generate high false-positive rates when attack signatures are not seen (Guo, 2023).

2.3. Synthesis and research gap

Domain-constrained SDN functionality is not integrated with dual-layer interpretability; the interaction between global model transparency and local, packet-level intelligence remains largely untapped (Hasnain *et al.*, 2026; Tserenkhuu *et al.*, 2025). This model aims to address these shortcomings by proposing a robust, explainable IDS framework for use in mission-critical smart city networks. Previous literature proposes a solution based on a new hybrid CNN-Bi-LSTM-attention model that addresses both spatial and temporal attack patterns with high accuracy. This can be achieved by applying a dual-stage XAI pipeline with SHAP and LIME to facilitate the trade-off between transparency and interpretability around the world and active zero-day detection with centroid proximity analysis. Figure 2 illustrates the technological process of transforming traditional rule-based defenses to a suggested explainable hybrid architecture. Our solution, based on a new hybrid CNN-Bi-LSTM-attention model, focuses on both spatial and temporal patterns in attacks, unlike prior work.

3. Design of the explainable hybrid intrusion detection system framework

The architecture of an effective IDS on a smart city infrastructure requires a trade-off between high-quality threat detection and forensic disclosure. The proposed model is a simulation-based method to assess the proposed framework because it can test destructive zero-day edge cases and Layer 2 vulnerabilities in a way that is safe and cost-effective, yet would otherwise put mission-critical urban assets at risk. The methodology guarantees scalability and controllability through a rigorous performance-benchmarking environment that uses standardized SDN and IoT datasets. The framework shown in Figure 3 progresses from the smart city layer to XAI-guided feature engineering to a hybrid DL engine, culminating in actionable forensics via SHAP and LIME.

3.1. Data ingestion and domain-constrained feature selection

The initial phase of the pipeline involves receiving uncoded network traffic. We use an XAI-based feature selection scheme to reduce the computational cost of previous

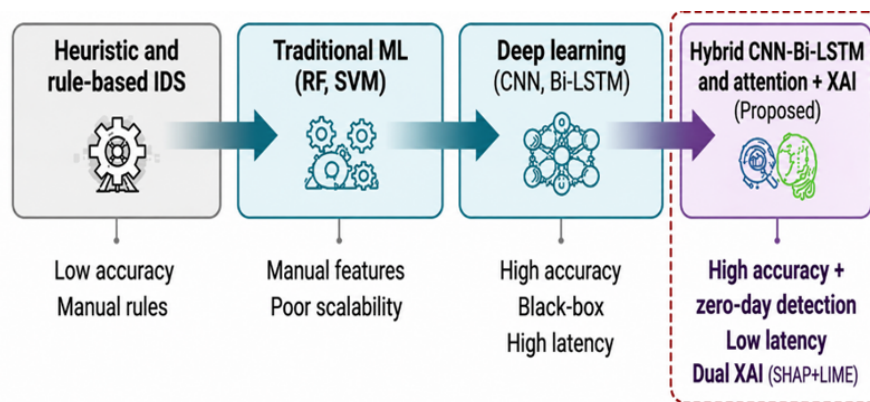


Figure 2. The development of IDS from old heuristic approaches to explainable hybrid deep learning models

Abbreviations: Bi-LSTM: Bidirectional long short-term memory; CNN: Convolutional neural network; IDS: Intrusion detection system; LIME: Local interpretable model-agnostic explanation; ML: Machine learning; RF: Random forest; SHAP: Shapley additive explanation; SVM: Support vector machine; XAI: Explainable artificial intelligence.

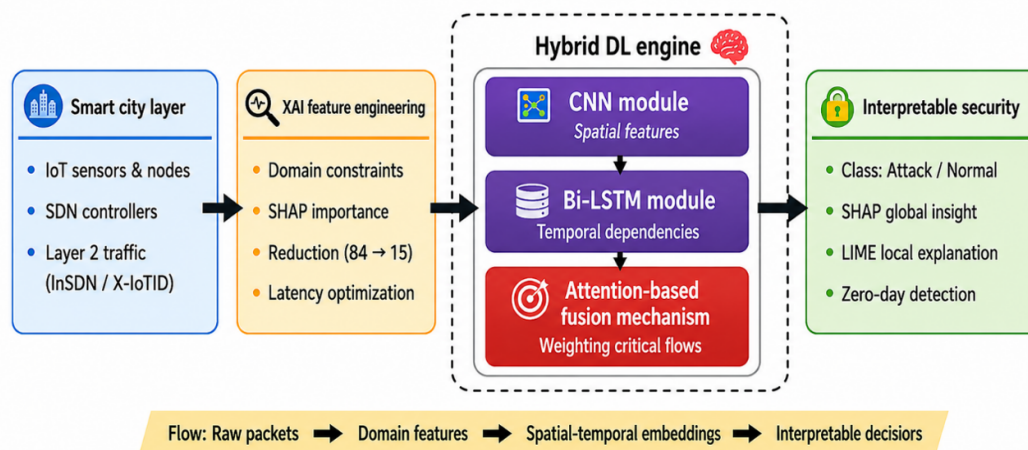


Figure 3. Proposed hybrid XAI-IDS architecture: Integrating spatial-temporal feature extraction with dual-layer explainability

Abbreviations: Bi-LSTM: Bidirectional long short-term memory; CNN: Convolutional neural network; DL: Deep learning; IDS: Intrusion detection system; IoT: Internet of Things; LIME: Local interpretable model-agnostic explanation; SDN: Software-defined networking; SHAP: Shapley additive explanation; XAI: Explainable artificial intelligence.

models in the literature, which are referred to as black boxes. This is achieved by optimizing the input space based on the SDN-IoT network logic and global importance scores: where $X = \{x_1, x_2, \dots, x_n\}$ denotes the raw input traffic flow with n features extracted from network packets or flows (e.g., packet size, inter-arrival time, protocol flags, SDN controller statistics, and flow duration). The full set of features in high-speed smart city networks, which can be high-dimensional, may include redundant or irrelevant features that add computational cost and poor model generalization.

To address this, a SHAP-based domain-constrained feature selection process has been adapted. SHAP value

is used to determine the value of the contribution of each feature x_i (**Equation 1**):

$$\phi_i(v) = \sum_{S \subseteq \{x_1, \dots, x_n\} \setminus \{x_i\}} \frac{|S|!(n-|S|-1)!}{n!} [v(S \cup \{x_i\}) - v(S)] \quad (1)$$

where set of features, the mean marginal contribution of feature x_i in any particular coalition of features is denoted by the value of ϕ_i , S is a subset of features that do not include x_i and the value of the output of the model, and $v(S)$ and $v(S)$ denote the model's output when only the features in S are used.

3.2. Hybrid convolutional neural network–bidirectional long short-term memory architecture

The detection engine is a core component that uses a hybrid DL model to detect both spatial and temporal attack patterns. The spatial features of the network packets are obtained through a CNN layer, where c_j is the feature map of the convolution operation represented in Equation 2:

$$c_j = f(w \cdot x_{i:i+k-1} + b) \quad (2)$$

in which w is the learnable filter (kernel) weight, b is the bias term, k is the size of the kernel, $x_{i:i+k-1}$ is a window of neighboring features, and $f(\cdot)$ is a non-linear activation function (rectified linear unit in this example). The CNN output capturing local spatial correlations (e.g., abnormal combinations of packet header details indicative of probing or distributed denial-of-service [DDoS] attacks) is then fed into a Bi-LSTM network to predict long-term temporal relationships among sequential traffic flows. The concealed condition at time step t is the result of the combination of the forward and backward LSTM conditions as shown in Equations 3–5, respectively.

$$\bar{h}_t = \sigma(W_{\bar{h}}x_t + U_{\bar{h}}\bar{h}_{t-1} + b_{\bar{h}}) \quad (3)$$

$$\bar{h}_t = \sigma(W_{\bar{h}}x_t + U_{\bar{h}}\bar{h}_{t+1} + b_{\bar{h}}) \quad (4)$$

$$H_t = [\bar{h}_t, \bar{h}_t] \quad (5)$$

in which σ represents the sigmoid or tanh activation, and the bidirectional processing enables the model to account for both the historical and forward contexts of a flow sequence, which is essential for identifying complex multi-stage or slow-rate attacks that are prevalent in smart city settings.

3.3. Attention-based fusion mechanism and zero-day detection

A dynamic computation of the most discriminative aspects of the learned representations is proposed through an attention-based fusion mechanism. This layer calculates attention weights α_i that indicate inappropriate behavior in networks represented by Equations 6–8.

$$e_i = \tanh(W_a H_i + b_a) \quad (6)$$

$$\alpha_i = \frac{\exp(e_i^T u_a)}{\sum_k \exp(e_k^T u_a)} \quad (7)$$

$$v = \sum_t \alpha_t H_t \quad (8)$$

where W_a and b_a are learnable parameters, and u_a is a context vector. The resulting fused vector v serves as a compact, context-aware embedding that combines spatial and temporal insights. For proactive zero-day attack detection, this framework uses a centroid-based proximity analysis in the embedding space. The centroid C of the normal (C_{normal}) is calculated as an average embedding of normal traffic samples in training. The calculated Euclidean distance of any new sample of fused representation v is given by Equation 9.

$$D(v, C_{normal}) = \sqrt{\sum_i (v_i - C_{normal,i})^2} \quad (9)$$

If $D(v, C_{normal}) > T_{zero-day}$ (where $T_{zero-day}$ is a tunable threshold optimized via validation), the instance is flagged as a potential unknown (zero-day) threat, even if the primary classifier outputs a benign label. This mechanism substantially improves robustness against novel attack vectors not present in the training data.

3.4. Centroid-based proximity analysis for zero-day threats

To overcome the drawback of the DL models in detecting unseen attack signatures, the framework introduces a centroid-based proximity analysis in the latent embedding space. In the training phase, a global benign centroid (C_{normal}) is estimated by averaging the feature vectors of all known instances of normal traffic. To any incoming network packet v , the system computes the Euclidean distance D between the benign centroid in Equation 10.

$$D(v, C_{normal}) = \sqrt{\sum_{i=1}^m (v_i - C_{normal,i})^2} \quad (10)$$

3.5. Explainability layer and post hoc interpretation

A two-stage XAI framework is used to support transparency. SHAP offers global interpretation by ranking feature contributions to the model, helping security analysts identify which SDN metrics (e.g., flow duration, packet rate) contribute most to the overall decisions. By comparison, LIME provides instance-level (packet-specific) predictions by reducing the complex model to a simpler, interpretable surrogate model. The LIME explanation for a specific input x is obtained by solving Equation 11.

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (11)$$

where L is a fidelity loss measuring how well the surrogate g approximates the original model f in the locality defined by the proximity kernel π_x , and $\Omega(g)$ penalizes model complexity. The final classification probability is produced by a softmax output layer represented in **Equation 12**.

$$Y = \text{Softmax}(W_o v + b_o) \quad (12)$$

A LIME explanation in the form of a visual bar chart or feature importance plot is generated and shown whenever an attack is detected, or a zero-day proximity alert is raised, so that administrators can perform quick forensic analysis.

3.6. Algorithm and implementation details

The pseudocode of the proposed system is provided below. This algorithm describes the ingestion-to-explanation pipeline, so that each malicious classification is accompanied by a visual explanation for security administrators.

Input: Raw network traffic X , domain threshold T_{domain} , zero-day threshold $T_{\text{zero-day}}$

Output: Classification Y , explanation E

Preprocess X (normalization, one-hot encoding of categorical features)

Compute Shapley additive explanation values and select reduced feature set F using **Equation 1** where $\phi_i > T_{\text{domain}}$

For each flow/packet in $X[F]$: (a) Extract spatial features using a convolutional neural network (**Equation 2**); (b) capture temporal dependencies using bidirectional long short-term memory (**Equations 3–5**); (c) generate fused representation v via an attention-based fusion mechanism (**Equations 6–8**)

Compute classification Y using softmax (**Equation 12**)

If $Y = \text{"Attack"}$ or $D(v, C_{\text{normal}}) > T_{\text{zero-day}}$ (**Equation 9**): (a) Generate local explanation $E = \xi(x)$ using local interpretable model-agnostic explanation (**Equation 11**); (b) trigger a security alert with the predicted label Y and visual explanation E

Return Y, E

Python 3.10 was used to implement the system, with TensorFlow/Keras and Scikit-learn DL and preprocessing libraries. To speed up the training stage, NVIDIA A100 graphics processing units (GPUs) were used. The simulation environment was based on intrusion detection datasets for software-defined networking (InSDN) and the extended industrial Internet of Things intrusion detection

(X-IIoTID) and offers a variety of SDN and industrial IoT traffic patterns. Standard scaling and categorical encoding were used during preprocessing to make these datasets compatible with the neural network architecture. The given tools and hardware have been chosen to promote both the reproducibility of high-accuracy results and the computational complexity requirements of dual-stage explainability.

4. Experimental setup and dataset characteristics

To assess the effectiveness of the hybrid architecture, a rigorous simulation setup was developed in Python 3.10 using the TensorFlow/Keras library. The training of the CNN-Bi-LSTM layers and the heavy computations for SHAP values were performed on an NVIDIA A100 to ensure efficient, reproducible computations.

The research uses two major datasets that reflect the contemporary smart-city environment: InSDN and X-IIoTID. The InSDN dataset offers 343,889 examples of SDN control plane dynamics, including denial-of-service (DoS), DDoS, and Probe attacks. X-IIoTID dataset provides 820,834 instances of reconnaissance and exploitation vectors for industrial IoT protocol security. A dedicated cross-layer 2 (CL2)-IDS dataset comprising 102,440 examples was used to test Layer 2 vulnerabilities, including address resolution protocol spoofing and dynamic host configuration protocol starvation, which are often overlooked in urban network security models. The detailed characteristics and the specific smart-city attack taxonomies for each dataset are summarized in [Table 1](#).

The data were first standardized, normalized, and categorized to ensure that heterogeneous traffic patterns were transformed into a homogeneous input format. The model was trained with a batch size of 64 and the Adam optimizer, with a learning rate of 0.001 that decreased when it plateaued. A 1,000-sample background was used to stabilize the SHAP explainability layer and ensure its integrity, and LIME was calibrated to achieve high local fidelity. Performance measurements were based on standard measures (accuracy, F1-score) and security-oriented measures, such as inference latency and zero-day detection recall, providing comprehensive information on the system's operational viability.

5. Assessment of detection accuracy and explainable artificial intelligence fidelity

The analysis of the proposed hybrid attention-based fusion mechanism and XAI framework was conducted through a series of comparative experiments, addressing the measures of classification accuracy, computational efficiency, and

Table 1. Dataset characteristics and smart-city attack taxonomy

Dataset	Total instances	Train/test split	Classes included	Smart city context
InSDN	343,889	80/20	DoS, DDoS, probe, brute force, botnet, web attack, U2R, normal	SDN control plane security
X-IIoTID	820,834	70/30	Reconnaissance, weaponization, exploitation, RDoS, etc.	Industrial/IoT protocol security
CL2-IDS	102,440	75/25	ARP spoofing, DHCP starvation, MAC flood, STP root, etc.	Layer 2 switch security in urban networks

Abbreviations: ARP: Address resolution protocol; CL2-IDS: Cross-layer layer-2 intrusion detection system dataset; DDoS: Distributed denial-of-service; DHCP: Dynamic host configuration protocol; DoS: Denial-of-service; InSDN: Intrusion detection dataset for software-defined networking; IoT: Internet of Things; MAC: Media access control; RDoS: Reflective denial of service; SDN: Software-defined networking; STP: Spanning tree protocol; U2R: User-to-root attack; X-IIoTID: Extended industrial Internet of Things intrusion detect.

interpretability fidelity. In the next subsections, the objective results of the simulation and subsequent analysis within the framework of smart city network security are provided. Table 2 presents the specific hyperparameter settings and experimental parameters used to ensure the reproducibility of the results.

5.1. Comparative performance analysis

The primary aim was to identify differences between the hybrid CNN–Bi-LSTM–attention model and the baseline models, commonly used in intrusion detection.

The system was found to be 99.85% accurate, as shown in Table 3, which was superior to the baseline CNN (94.20%) and Bi-LSTM (95.12%) models. All precision, recall, and F1-score metrics were consistently 0.997 or higher, indicating they are very reliable for detecting different attack types.

The conceptual representation in Figure 4 shows the confusion matrices for the InSDN and X-IIoTID data, where the model performed almost perfectly in distinguishing DoS and DDoS attack vectors, with an average detection rate of 99.99% for known malicious signatures.

5.2. Efficiency and impact of explainable artificial intelligence-driven feature selection

One of the critical gaps identified in the existing literature is the high latency caused by overlapping feature sets. To resolve this, we used domain-constrained XAI in order to maximize feature engineering. The results of this reduction are summarized in Table 4. The first feature set (78 attributes) was based on raw flow features from InSDN (77 features after removing sockets), X-IIoTID (68 features), and other domain-derived SDN controller statistics and Layer-2 metrics of CL2-IDS that apply to smart city contexts.

The system showed an improvement of 67% in training time by a reduction of features (reduction from 78 to 18 features), which dropped from 1,240 seconds to 410 seconds. More to the point, for real-time smart city applications, inference latency was reduced from 12.4 ms to 4.1 ms, and the memory footprint (450 MB) decreased by 75.5%. These numbers confirm the effectiveness of the SHAP-based importance scores to remove logically irrelevant SDN–IoT metrics without affecting classification performance, as shown in Figure 5.

The local contribution of these features can also be illustrated in Figure 6, which compares local LIME triggers with global SHAP biases of a representative attack packet and demonstrates the rationale of the optimized feature set.

The inference latency dropped to 4.1 ms (decreased from 12.4 ms) with a 75.5% reduction in memory footprint (from 450 MB to 110 MB). The model in Figure 7 shows rapid training convergence and a substantial decrease in operational latency, respectively, confirming the efficiency of the XAI-optimized pipeline for real-time smart city applications.

5.3. Zero-day detection and model robustness

There was the capability to detect previously unseen threats using a proactive zero-day detection mechanism based on centroid-based proximity. Table 5 shows the detection rates for unknown vectors, with the proposed model having a high recall of 71.09% for zero-day attacks. This is a great improvement over traditional filters that fail miserably in such cases, although a bit worse than on known attacks.

Cross-domain robustness was also evaluated by training InSDN and testing on X-IIoTID. Table 6 indicates that the transferability gap of the proposed hybrid model was 5.7%, and consequently the model achieved 94.2%

Table 2. Hyperparameter settings and experiment parameters

Parameter	Value	Description
Batch size	64	Training batch size
Optimizer	Adam	With learning rate scheduling
Initial learning rate	0.001	Decreased on plateau
Background samples (Shapley additive explanation)	1,000	For stable explanations
Local interpretable model-agnostic explanation fidelity calibration	High local fidelity	Instance-level surrogate
Epochs/early stopping	As needed with patience	Convergence monitoring
Hardware	NVIDIA A100 graphics processing unit	Training and Shapley additive explanation computation
Framework	TensorFlow/Keras + Scikit-learn	Reproducible implementation

Table 3. Performance comparison with state-of-the-art models

Model architecture	Accuracy (%)	Precision	Recall	F1-score
Baseline convolutional neural network (Susilo <i>et al.</i> , 2025)	94.20	0.932	0.941	0.936
Bidirectional long short-term memory only (Al Lail <i>et al.</i> , 2023)	95.12	0.948	0.950	0.949
Proposed (hybrid + attention + explainable artificial intelligence)	99.85	0.997	0.998	0.997

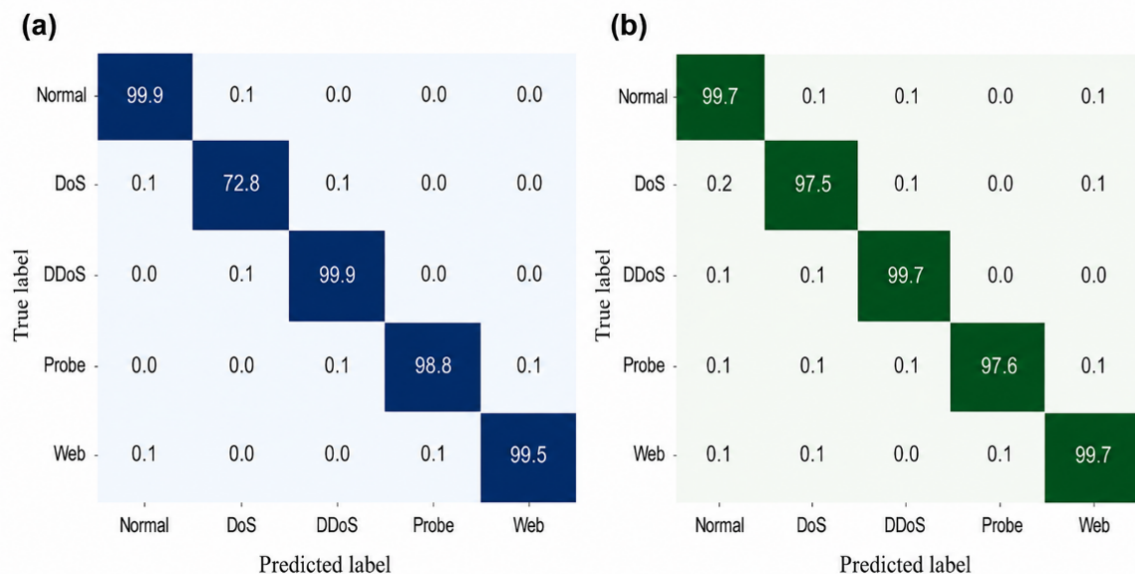


Figure 4. Comparison of the confusion matrix of (a) InSDN and (b) X-IIoTID datasets
 Abbreviations: DDoS: Distributed denial-of-service; DoS: Denial-of-service; InSDN: Intrusion detection dataset for software-defined networking; X-IIoTID: Extended industrial Internet of Things intrusion detection.

accuracy on external data, which is less than the 11.5% gap in the baseline models. This indicates that the attention-based fusion mechanism can capture generalized spatio-temporal dependencies.

Table 4. Impact of domain-constrained feature selection (explainable artificial intelligence)

Feature set	Count	Training time (s)	Inference (ms)	Memory (MB)
Full feature set	78	1,240	12.4	450
Explainable artificial intelligence + domain-optimized	18	410	4.1	110
Improvement	77% reduction	67%	67%	75.5%

6. Discussion and implications

The experimental results validate that the incorporation of the attention-based fusion mechanism with a dual-stage XAI pipeline (namely SHAP for global explanations and LIME for local explanations) is capable of mitigating some major drawbacks of traditional black-box DL IDS, such as high detection accuracy (99.85% detection accuracy for known attacks), zero-day resilience (71.09% recall for zero-day attacks), and significant efficiency improvements. The centroid-based proximity analysis in the embedding

space also provides a useful layer of protection, proactively identifying and blocking unknown attacks that evade detection by signature-based methods. Additionally, domain-constrained and SHAP-guided feature selection leads to 77% dimensionality reduction and excellent model performance, significantly decreasing inference latency to 4.1 ms and memory usage by 75.5%, which is essential for smart city edge deployment.

Table 5. Zero-day attack detection and explainable artificial intelligence fidelity

Attack vector	Detection rate (%)	Local interpretable model-agnostic explanation fidelity	Shapley additive explanation stability
Known distributed denial-of-service	99.99	0.98	0.99
Unknown (zero-day)	71.09	0.84	0.81
Structured query language injection	96.50	0.92	0.94

Despite the high rate of 71.09% for the number of zero-day attacks detected, it is worth noting that the high sensitivity to anomalies has a trade-off with precision: the higher the anomaly sensitivity, the more false positives

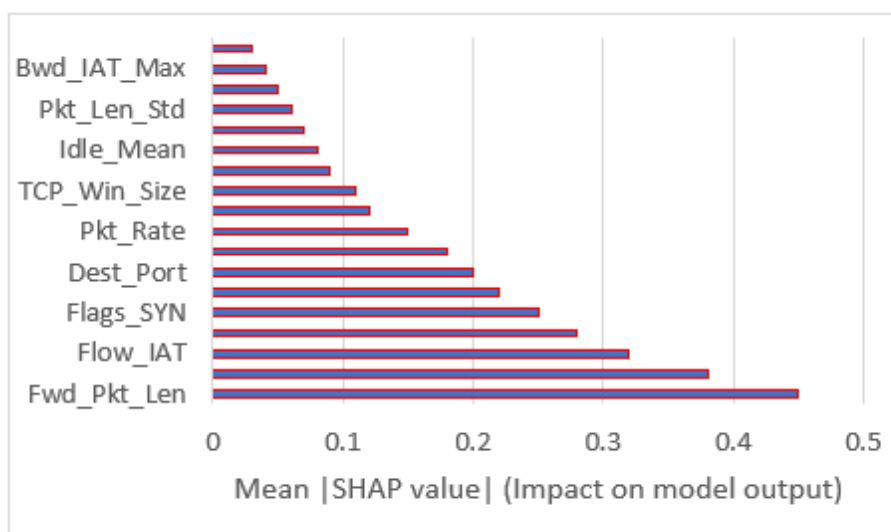


Figure 5. SHAP global feature importance
Abbreviation: SHAP: Shapley additive explanation.

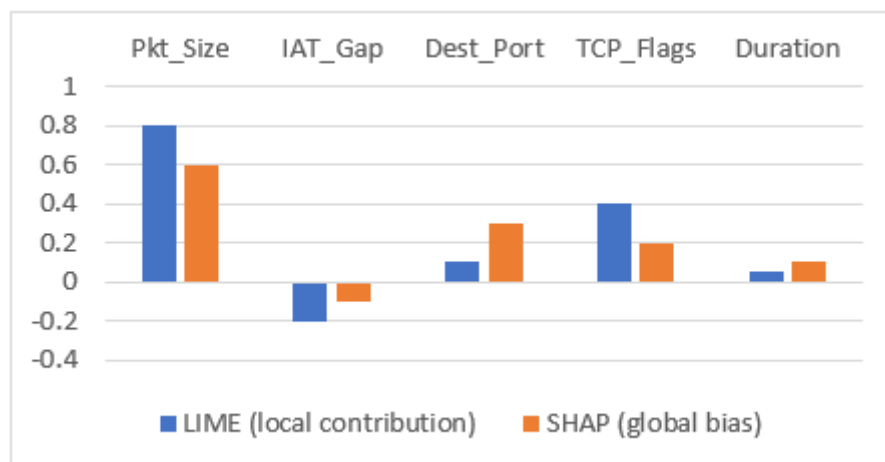


Figure 6. Local and global feature contributions for a representative attack sample using dual-layer XAI
Abbreviations: LIME: Local interpretable model-agnostic explanation; SHAP: Shapley additive explanation; XAI: Explainable artificial intelligence.

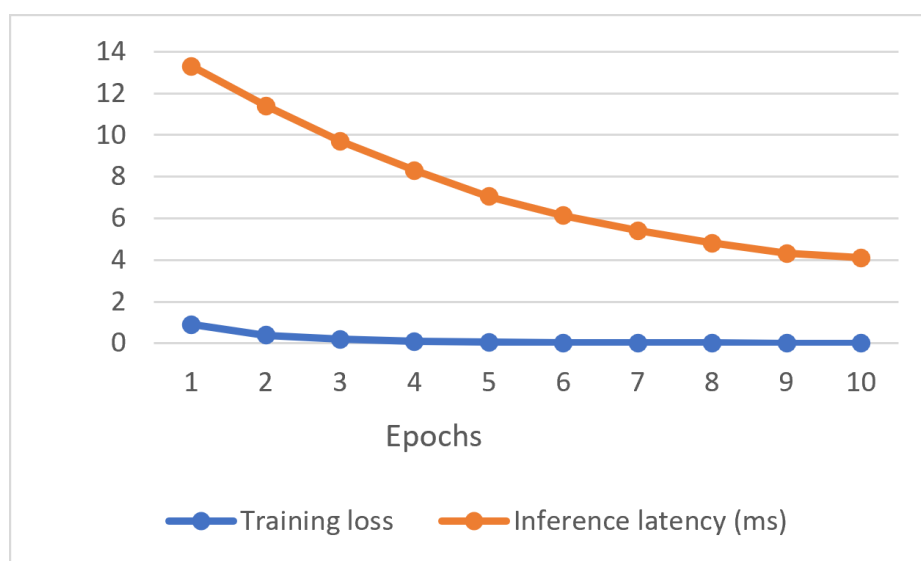


Figure 7. Training convergence versus inference latency optimization

would occur in environments with varying traffic scenarios, which could lead to alert fatigue among security operators. The SHAP/LIME framework, as an XAI technique, is a post hoc approach that incurs additional, albeit not significant, computational overhead, which might affect the system's real-time performance under extreme network conditions or in ultra-low-power devices. The explanations themselves can remain subject to adversarial manipulation, such that altering the inputs to the system can reduce the fidelity/accuracy of XAI or can create misleading circumstances for forensic analysis. Further, the hybrid CNN-Bi-LSTM-

attention model provides superior cross-domain robustness (5.7% transferability gap), although this comes at the expense of relying on post hoc interpretability, as opposed to other models with inherent interpretability (e.g., decision trees or attention-only models) that may provide more transparent decision pathways, albeit at a potential loss of detection accuracy. The scalability of the proposed approaches for massively deployed, heterogeneous smart city networks with thousands of IoT nodes needs to be tested outside the simulated InSDN, X-IIoTID, and CL2-IDS environments.

Table 6. Cross-domain robustness analysis

Training source	Testing source	Accuracy (%)	Transferability gap (%)	Remarks
InSDN	InSDN (internal)	99.90	NA	Baseline performance
InSDN	X-IIoTID (external)	88.40	11.50	Baseline model transfer
Proposed hybrid model	X-IIoTID (external)	94.20	5.70	High robustness

Abbreviations: InSDN: Intrusion detection dataset for software-defined networking; X-IIoTID: Extended industrial Internet of Things intrusion detection.

Future work could include an ablation study to isolate the contributions of domain-constrained feature selection, the attention-based fusion mechanism, and the XAI modules to further strengthen the empirical claims. Comparing the results with more recent transformer-based intrusion detection models would also give a more thorough comparative assessment. In addition to the simulated InSDN, X-IIoTID, and CL2-IDS scenarios, the scalability of the proposed approach to massively deployed, heterogeneous smart city networks with thousands of IoT nodes should be verified.

7. Conclusion

The framework's efficacy has been proven superior, with experimental results showing it can detect known attacks with over 99% accuracy and a strong 71% accuracy for zero-day attacks. It is interesting to note that the XAI-based feature reduction algorithm reduced the feature space by 77% (from 78 to 18 features) and achieved a 67% reduction in inference latency (from 12.4 ms to 4.1 ms). The model's viability is optimized to enable real-time deployment in the city network on resource-constrained edge devices.

The main contributions of this methodology are a new hybrid spatio-temporal architecture that models complex packet patterns and long-term temporal correlations using an attention-based fusion mechanism, and a dual-layered interpretability framework that synchronizes SHAP to global transparency and LIME to local forensics to close the gap between automated detection and human-led audits. Moreover, the system provides a preemptive line of defense with higher zero-day resilience aided by centroid-based proximity analysis, which can considerably enhance the ability to detect previously unseen threats. The existing framework offers a high level of transparency; incorporating federated learning to complement privacy-

preserving collaborative defense among distributed smart city nodes will be examined in future studies. We also hope to explore how adversarial machine learning affects XAI fidelity, ensuring that forensic explanations are resistant to advanced evasion strategies. Finally, this paper offers a blueprint for scaling up to develop reliable, explainable, and ultra-efficient security ecosystems in the upcoming generation of digital cities.

Acknowledgments

The authors would like to thank the providers of the publicly available InSDN, X-IIoTID, and related SDN/IoT intrusion detection datasets for enabling this research. We also acknowledge the computational support provided by NVIDIA A100 GPUs for model training and explainability computations.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Vijayakumari Rodda
Data curation: Anurag Shrivastava
Formal analysis: Anurag Shrivastava, S. Prabagar
Investigation: Anurag Shrivastava, G. Shailaja
Methodology: Vijayakumari Rodda, Sundaram K. Meenakshi
Project administration: S. B. G. Tilak Babu
Resources: Aditya Rautaray, G. Shailaja
Software: Vijayakumari Rodda
Supervision: Vijayakumari Rodda
Validation: Vijayakumari Rodda, Sundaram K. Meenakshi, S. B. G. Tilak Babu
Visualization: Aditya Rautaray
Writing—original draft: Vijayakumari Rodda
Writing—review & editing: Aditya Rautaray, S. Prabagar
All authors have read and agreed to the published version of the manuscript.

Availability of data

The InSDN dataset is publicly available on Kaggle and related repositories. The X-IIoTID dataset is accessible via IEEE DataPort and Kaggle. The CL2-IDS dataset was developed and used as described in the manuscript; details are available from the corresponding author upon reasonable request. Code and implementation details can be shared upon acceptance for reproducibility.

References

- Admass, W. S., Munaye, Y. Y., & Diro, A. A. (2024). Cyber security: State of the art, challenges and future directions. *Cyber Security and Applications*, 2, 100031.
<https://doi.org/10.1016/j.csa.2023.100031>
- Afraji, D. M. A. A., Lloret, J., & Peñalver, L. (2025). An integrated hybrid deep learning framework for intrusion detection in IoT and IIoT networks using CNN-LSTM-GRU architecture. *Computation*, 13(9), 222.
<https://doi.org/10.3390/computation13090222>
- Agbor, B. A., Stephen, B. U. A., Asuquo, P., Luke, U. O., & Anaga, V. (2025). Hybrid CNN-BiLSTM-DNN approach for detecting cybersecurity threats in IoT networks. *Computers*, 14(2), 58.
<https://doi.org/10.3390/computers14020058>
- Ahmad, J., Latif, S., Khan, I. U., Alshehri, M. S., Khan, M. S., Alasbali, N., & Jiang, W. (2025). An interpretable deep learning framework for intrusion detection in industrial Internet of Things. *Internet of Things*, 33, 101681.
<https://doi.org/10.1016/j.iot.2025.101681>
- Ahmed, M., Islam, S. R., Anwar, A., Moustafa, N., & Pathan, A.-S. K. (Eds.). (2022). *Explainable artificial intelligence for cyber security* (Studies in Computational Intelligence, Vol. 1025). Springer Nature Switzerland, Cham, Switzerland.
<https://doi.org/10.1007/978-3-030-96630-0>
- Al Lail, M., Garcia, A., & Olivo, S. (2023). Machine learning for network intrusion detection—A comparative study. *Future Internet*, 15(7), 243.
<https://doi.org/10.3390/fi15070243>
- Al-Hawawreh, M., Sitnikova, E., & Aboutorab, N. (2022). X-IIoTID: A connectivity-agnostic and device-agnostic intrusion data set for industrial Internet of Things. *IEEE Internet of Things Journal*, 9(5), 3962–3977.
<https://doi.org/10.1109/JIOT.2021.3102056>
- Altulaihan, E., Almaiah, M. A., & Aljughaiman, A. (2022). Cybersecurity threats, countermeasures and mitigation techniques on the IoT: Future research directions. *Electronics*, 11(20), 3330.
<https://doi.org/10.3390/electronics11203330>
- Altunay, H. C., & Albayrak, Z. (2023). A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks. *Engineering Science and Technology, an International Journal*, 38, 101322.
<https://doi.org/10.1016/j.jestch.2022.101322>
- Arreche, O., Guntur, T., & Abdallah, M. (2024). XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. *Applied Sciences*, 14(10), 4170.
<https://doi.org/10.3390/app14104170>
- Chaganti, R., Suliman, W., Ravi, V., & Dua, A. (2023). Deep learning approach for SDN-enabled intrusion detection system in IoT networks. *Information*, 14(1), 41.
<https://doi.org/10.3390/info14010041>
- Elsaid, S. A., & Binbusayyis, A. (2024). An optimized isolation forest based intrusion detection system for heterogeneous and streaming data in the industrial Internet of Things (IIoT) networks. *Discover Applied Sciences*, 6(9), 483.
<https://doi.org/10.1007/s42452-024-06165-w>
- Guo, Y. (2023). A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer Communications*, 198, 175–185.
<https://doi.org/10.1016/j.comcom.2022.11.001>
- Hakak, S., Khan, W. Z., Gilkar, G. A., Imran, M., & Guizani, N. (2020). Securing smart cities through blockchain technology: Architecture, requirements, and challenges. *IEEE Network*, 34(1), 8–14.
<https://doi.org/10.1109/MNET.001.1900178>
- Hasan, M. A. M., Nasser, M., Pal, B., & Ahmad, S. (2014). Support vector machine and random forest modeling for intrusion detection system (IDS). *Journal of Intelligent Learning Systems and Applications*, 06(01), 45–52.
<https://doi.org/10.4236/jilsa.2014.61005>
- Hasnain, M., Javaid, N., Saudagar, A. K. J., & Kumar, N. (2026). An intelligent and explainable intrusion detection framework for Internet of Sensor Things using generalizable optimized active machine learning. *Journal of Network and Computer Applications*, 245, 104358.
<https://doi.org/10.1016/j.jnca.2025.104358>
- Houda, Z. A. El, Brik, B., & Khoukhi, L. (2022). “Why should I trust your IDS?”: An explainable deep learning framework for intrusion detection systems in Internet of Things networks. *IEEE Open Journal of the Communications Society*, 3, 1164–1176.
<https://doi.org/10.1109/OJCOMS.2022.3188750>
- Iftikhar, A., Hussain, F. B., Qureshi, K. N., Shiraz, M., & Sookhak, M. (2025). Securing edge based smart city networks with software defined networking and zero trust architecture. *Journal of Network and Computer Applications*, 244, 104341.
<https://doi.org/10.1016/j.jnca.2025.104341>
- Imtiaz, N., Wahid, A., Ul Abideen, S. Z., Muhammad Kamal, M., Sehito, N., Khan, S., Virdee, B. S., Kouhalvandi, L., & Alibakhshikenari, M. (2025). A deep learning-based approach for the detection of various Internet of Things intrusion attacks through optical networks. *Photonics*, 12(1), 35.
<https://doi.org/10.3390/photonics12010035>

- Ioannou, C., & Vassiliou, V. (2021). Network attack classification in IoT using support vector machines. *Journal of Sensor and Actuator Networks*, 10(3), 58.
<https://doi.org/10.3390/jsan10030058>
- Kampourakis, K. E., Gkioulos, V., & Katsikas, S. (2026). Cybersecurity digital twins for industrial systems: From literature synthesis to framework design. *Information*, 17(3), 286.
<https://doi.org/10.3390/info17030286>
- Kampourakis, K. E., Gkioulos, V., Kavallieratos, G., & Lin, J. C. (2025). Digital twin-enabled incident detection and response: A systematic review of critical infrastructures applications. *International Journal of Information Security*, 24(5), 194.
<https://doi.org/10.1007/s10207-025-01113-0>
- Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., & Zomaya, A. Y. (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. *Information Sciences*, 639, 119000.
<https://doi.org/10.1016/j.ins.2023.119000>
- Kikissagbe, B. R., & Adda, M. (2024). Machine learning-based intrusion detection methods in IoT systems: A comprehensive review. *Electronics*, 13(18), 3601.
<https://doi.org/10.3390/electronics13183601>
- Kilincer, I. F. (2025). Explainable AI supported hybrid deep learning method for layer 2 intrusion detection. *Egyptian Informatics Journal*, 30, 100669.
<https://doi.org/10.1016/j.eij.2025.100669>
- Kumpf, K., Cajic, M., Zeljkovic, V., Mravik, M., Zivkovic, M., Mani, J., Simic, V., & Bacanin, N. (2025). Tackling smart city security: Deep learning approach utilizing feature selection and two-level cooperative framework optimized by adapted metaheuristics algorithm. *International Journal of Information Security*, 24(6), 221.
<https://doi.org/10.1007/s10207-025-01137-6>
- Lakshan Yasarathna, T., & Le-Khac, N. A. (2026). ASEADOS-SDN-IoT: A novel SDN-IoT network intrusion detection dataset and framework. *Internet of Things*, 36, 101891.
<https://doi.org/10.1016/j.iot.2026.101891>
- Mishra, S. R., Shanmugam, B., Yeo, K. C., & Thennadil, S. (2025). SDN-enabled IoT security frameworks—A review of existing challenges. *Technologies*, 13(3), 121.
<https://doi.org/10.3390/technologies13030121>
- Moustafa, N., Adi, E., Turnbull, B., & Hu, J. (2018). A new threat intelligence scheme for safeguarding Industry 4.0 systems. *IEEE Access*, 6, 32910–32924.
<https://doi.org/10.1109/ACCESS.2018.2844794>
- Oseni, A., Moustafa, N., Creech, G., Sohrabi, N., Strelzoff, A., Tari, Z., & Linkov, I. (2023). An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks. *IEEE Transactions on Intelligent Transportation Systems*, 24(1), 1000–1014.
<https://doi.org/10.1109/TITS.2022.3188671>
- Rachakonda, L. P., Siddula, M., & Sathya, V. (2024). A comprehensive study on IoT privacy and security challenges with focus on spectrum sharing in next-generation networks (5G/6G/beyond). *High-Confidence Computing*, 4(2), 100220.
<https://doi.org/10.1016/j.hcc.2024.100220>
- Rahman, M. A., Asyhari, A. T., Leong, L. S., Satrya, G. B., Hai Tao, M., & Zolkipli, M. F. (2020). Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. *Sustainable Cities and Society*, 61, 102324.
<https://doi.org/10.1016/j.scs.2020.102324>
- Shoukat, S., Gao, T., Javeed, D., Saeed, M. S., & Adil, M. (2025). Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. *Computers and Security*, 149, 104191.
<https://doi.org/10.1016/j.cose.2024.104191>
- Siddiqi, M. A., & Pak, W. (2020). Optimizing filter-based feature selection method flow for intrusion detection system. *Electronics*, 9(12), 2114.
<https://doi.org/10.3390/electronics9122114>
- Singh, C., & Jain, A. K. (2024). A comprehensive survey on DDoS attacks detection & mitigation in SDN-IoT network. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, 8, 100543.
<https://doi.org/10.1016/j.prime.2024.100543>
- Susilo, B., Muis, A., & Sari, R. F. (2025). Intelligent intrusion detection system against various attacks based on a hybrid deep learning algorithm. *Sensors*, 25(2), 580.
<https://doi.org/10.3390/s25020580>
- Tserenkhuu, M., Hossain, M. D., Taenaka, Y., & Kadobayashi, Y. (2025). Intrusion detection system framework for SDN-based IoT networks using deep learning approaches with XAI-based feature selection techniques and domain-constrained features. *IEEE Access*, 13, 136864–136880.
<https://doi.org/10.1109/ACCESS.2025.3595595>
- Vishwakarma, R., & Jain, A. K. (2019). A survey of DDoS attacking techniques and defence mechanisms in the IoT network. *Telecommunication Systems*, 73(1), 3–25.
<https://doi.org/10.1007/s11235-019-00599-z>
- Wahab, F., Shah, A., Khan, I., Ali, B., & Adnan, M. (2024). An SDN-based hybrid-DL-driven cognitive intrusion detection system for IoT ecosystem. *Computers and Electrical Engineering*, 119, 109545.
<https://doi.org/10.1016/j.compeleceng.2024.109545>