

RESEARCH ARTICLE

An Explainable AI–Driven Zero–Day Attack Detection Framework for Securing Edge Devices in Smart Cities

R. Shanmuga Sundaram* , R. Mohanraj, S. Sasidevi

ABSTRACT: The rapid proliferation of edge computing in smart cities has enhanced real-time data processing capabilities, but it has also exposed critical vulnerabilities to sophisticated cyber threats such as zero-day attacks. Traditional signature-based intrusion detection systems often fail to identify these previously unknown threats due to their lack of adaptive intelligence and interpretability. This research proposes an Explainable Artificial Intelligence (XAI)-driven zero-day attack detection framework tailored for edge devices deployed in smart city environments. The proposed system combines deep anomaly detection using a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model with SHAP (SHapley Additive exPlanations)-based interpretability to detect and explain anomalous behaviors in real-time network traffic. The model is trained on diverse datasets mimicking heterogeneous edge devices in smart infrastructures, ensuring robustness and scalability. Experimental results demonstrate high detection accuracy, low false-positive rates, and strong resilience against unseen attack patterns. Moreover, the integration of XAI components provides actionable insights to administrators, thereby enhancing trust, transparency, and decision-making in cybersecurity operations. This framework marks a significant step toward proactive and explainable security solutions for safeguarding smart urban ecosystems.

Keywords: Explainable AI (XAI), Zero-Day Attack Detection, Edge Computing, Smart Cities, CNN-LSTM, SHAP, Anomaly Detection, Cybersecurity, Intrusion Detection System (IDS), Interpretable Deep Learning.

Received: 03 February 2025; Revised: 26 March 2025; Accepted: 19 June 2025; Published Online: 09 July 2025

1. INTRODUCTION

The contemporary vision of smart cities represents a profound transformation of urban ecosystems, driven by the pervasive integration of interconnected cyber-physical systems. These systems embed intelligence into the very fabric of infrastructure, utilizing vast networks of sensors, actuators, and edge devices to enable real-time data analytics and automated decision-making [1]. This technological synergy supports critical municipal services, including intelligent traffic management to alleviate congestion, enhanced public safety monitoring, efficient healthcare

delivery through remote diagnostics, and dynamic energy grid optimization for sustainable resource use. The operational backbone enabling this real-time responsiveness is edge computing, a paradigm that shifts data processing and storage from centralized cloud servers to the periphery of the network, closer to the data source [2]. By processing data locally on edge devices or nearby gateways, edge computing dramatically reduces latency, conserves bandwidth, and mitigates the privacy concerns associated with transmitting vast amounts of sensitive information to distant data centers.

However, this decentralization and the inherent resource-constrained nature of edge devices, often characterized by limited processing power, memory, and energy, introduce formidable and complex cybersecurity challenges. The expanded attack surface presented by thousands of distributed nodes creates a lucrative target for malicious actors. Among the most severe threats are zero-day attacks, which exploit previously unknown software

Department of Electronics and Computer Engineering, Knowledge Institute of Technology, Knowledge College Road, Kakapalayam, Selliampalayam, Tamil Nadu-637504, India

* Author to whom correspondence should be addressed:
rsece@kiot.ac.in (R. Shanmuga Sundaram)

vulnerabilities for which no patch or signature yet exists [3]. These attacks are particularly insidious in smart city contexts because they can remain undetected by conventional security measures, potentially allowing adversaries to disrupt essential services, compromise citizen data, or even take control of physical infrastructure systems before a defensive response can be mobilized. The traditional first line of defense, Intrusion Detection Systems (IDS), has historically relied on signature-based methodologies. These systems function by matching observed network traffic or system activities against a database of known attack patterns or malicious code signatures [4]. While effective against well-documented threats, this approach is fundamentally reactive and fails catastrophically when confronted with novel, zero-day exploits that leave no recognizable signature. As smart city infrastructures scale in complexity and connectivity, the limitations of such static defenses become increasingly pronounced, necessitating a shift towards more intelligent, adaptive, and proactive security mechanisms.

This necessity has catalyzed significant research into applying machine learning (ML) and, more recently, deep learning (DL) for anomaly-based intrusion detection. These models promise a more dynamic defense by learning the complex statistical patterns of "normal" network behavior during a training phase and subsequently flagging significant deviations as potential anomalies indicative of malicious activity [5]. Techniques ranging from classical algorithms like Support Vector Machines (SVM) to sophisticated deep architectures have shown considerable promise in detecting subtle and novel attack patterns that evade signature-based checks. Nonetheless, the adoption of these advanced models, particularly deep neural networks, introduces a critical secondary problem: the "black-box" dilemma. The exceptional performance of models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks often comes at the cost of interpretability. Their internal decision-making processes are highly complex and non-linear, making it difficult for human security analysts to understand *why* a particular traffic instance was flagged as malicious [6]. In high-stakes environments like smart cities, where automated decisions can have immediate physical-world consequences, this lack of transparency erodes trust, complicates incident response, and hinders accountability. Analysts are left with alerts devoid of contextual reasoning, slowing down forensic investigation and potentially leading to alert fatigue.

To bridge this critical gap between model performance and operational trustworthiness, the field of Explainable Artificial Intelligence (XAI) has emerged as an essential research paradigm. XAI seeks to make the outputs of AI systems understandable to human users. Techniques such as SHapley Additive exPlanations (SHAP), rooted in cooperative game theory, provide a robust mathematical framework for attributing a model's prediction to each input feature [7]. By quantifying the contribution of individual factors, such as packet size, protocol type, or connection frequency, to a final security alert, SHAP transforms an opaque prediction into an interpretable explanation.

Integrating such XAI techniques into intrusion detection does more than satisfy a theoretical need for transparency; it provides practical, actionable intelligence. It empowers cybersecurity personnel to validate alerts rapidly, understand the root cause of an anomaly, and make informed, evidence-based response decisions, thereby accelerating the entire incident response lifecycle [8, 9]. The novel contribution of this paper lies in its integrated framework that specifically addresses the unique constraints and threats of smart city edge networks. It moves beyond simply applying a hybrid deep learning model by unifying a purpose-built CNN-LSTM architecture, optimized for the spatiotemporal features of network traffic, with a SHAP-based explainability layer, all tailored for deployment on resource-constrained edge hardware. This confluence of high-accuracy anomaly detection for zero-day threats and real-time model interpretability within an edge-computing context represents a significant advancement over existing solutions that often address either performance or explainability in isolation, or are designed for more resource-abundant cloud environments.

The literature reflects a growing acknowledgment of these intertwined challenges. Early security strategies for edge computing focused primarily on cryptographic techniques and lightweight authentication protocols, which, while important for confidentiality, lacked the adaptive intelligence needed for intrusion detection [10, 11]. The evolution from signature-based to anomaly-based IDS marked a step forward, with various ML models being deployed [12, 13]. However, as noted in studies, models like SVM and Random Forest can struggle with the high-dimensional, non-linear, and dynamically evolving traffic patterns characteristic of smart cities, often resulting in suboptimal accuracy or high false-positive rates that undermine their practical utility [14, 15]. Deep learning offered a leap in capability, with hybrid architectures like CNN-LSTM demonstrating superior efficacy in learning both spatial hierarchies and long-range temporal dependencies in sequential data, making them particularly potent for network intrusion detection [16, 17]. Concurrently, the interpretability crisis of DL spurred the adoption of XAI methods. Research has shown the successful application of SHAP to explain predictions from models like LSTMs, illuminating the role of specific network features in security decisions [18, 19]. Other innovative approaches have explored combining unsupervised learning with XAI for anomaly detection. However, a salient gap identified in recent surveys is the lack of scalable, integrated solutions that embed explainability directly into lightweight deep learning models suitable for the heterogeneous and resource-limited edge devices that populate smart city infrastructures [20]. This paper directly addresses this gap. The proposed framework is not merely an application of an existing model but a holistic system designed from the ground up for the smart city edge environment. It encompasses specialized data preprocessing for heterogeneous streams, a hybrid CNN-LSTM model for capturing complex attack signatures, a built-in SHAP interpretability module for transparency, and optimization strategies for feasible edge deployment. By validating this

framework on realistic smart city network datasets and demonstrating its performance and explanatory power, this work aims to provide a robust, scalable, and trustworthy security solution, thereby contributing to the foundational resilience of future urban ecosystems [9, 10].

2. PROPOSED METHODOLOGY

The proposed framework aims to detect zero-day attacks in smart city edge devices using a hybrid deep learning model enhanced with explainable artificial intelligence (XAI). It consists of four major components: data acquisition and preprocessing, hybrid deep learning-based anomaly detection, explainability layer using SHAP, and real-time alert generation. Figure 1 shows the systematic architecture of the proposed framework.

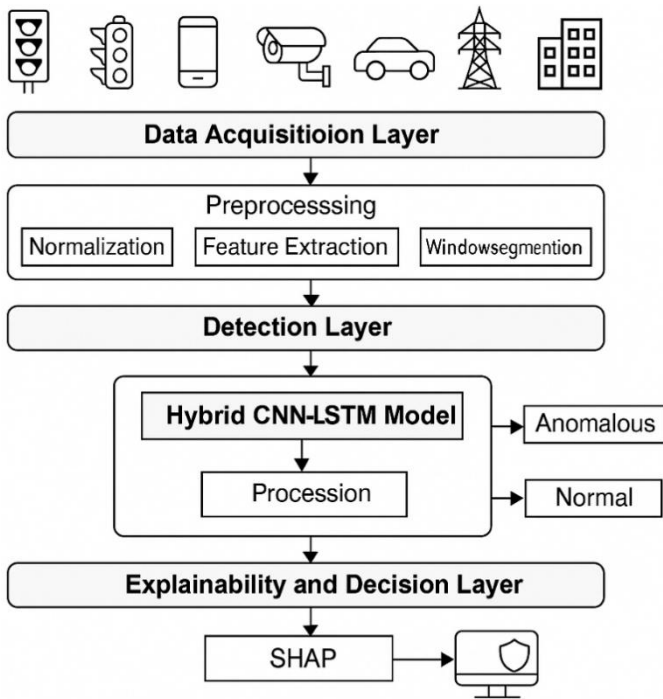


Fig. 1. System Architecture of the Proposed Framework.

2.1. Data Acquisition and Preprocessing

In the proposed framework, data acquisition is performed at the edge level across various smart city nodes, including traffic sensors, surveillance systems, and smart meters (Figure 2). These edge devices continuously generate raw network traffic logs and system behavior metrics, which serve as the primary source for detecting zero-day threats. The collected data includes features such as packet size, time-to-live (TTL), protocol type, source/destination IPs, and timestamped events. Due to the heterogeneous nature of edge devices, the collected data is initially unstructured and often

contains noise and missing values.

The preprocessing phase involves four main operations: noise reduction, missing value imputation, feature scaling, and dimensionality alignment. Noise is removed using a smoothing technique such as exponential moving average (EMA), expressed as:

$$S_t = \alpha \cdot X_t + (1 - \alpha) \cdot S_{t-1} \quad (1)$$

Where S_t is the smoothed value at time t , X_t is the raw input, and α is the smoothing factor ($0 < \alpha < 1$).

Next, missing values are imputed using linear interpolation between known values:

$$X_i^{\text{interp}} = X_{i-1} + \frac{X_{i+1} - X_{i-1}}{2} \quad (2)$$

For numerical stability during training, feature normalization is performed using Min-Max scaling:

$$X_i^{\text{scaled}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \quad (3)$$

This ensures that all input features are scaled to a $[0, 1]$ range, which is crucial for the convergence of neural networks. Additionally, timestamps are converted into time-delta features to model sequential behavior.

This ensures that all input features are scaled to a $[0, 1]$ range, which is crucial for the convergence of neural networks. Additionally, timestamps are converted into time-delta features to model sequential behavior. After scaling, the dataset is segmented into overlapping windows to preserve temporal dependencies, typically using a sliding window approach of size w and step s , forming input sequences $\{X_t, X_{t+1}, \dots, X_{t+w}\}$. These sequences are fed into the hybrid CNNLSTM model for downstream anomaly detection tasks.

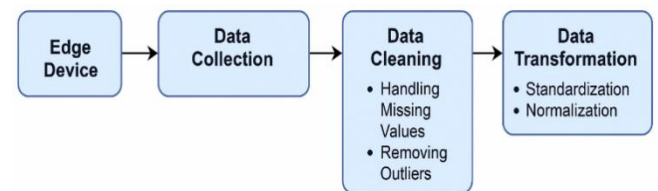


Fig. 2. Data Acquisition and Preprocessing Workflow.

2.2. Hybrid CNN–LSTM Model for Anomaly Detection

The core of the proposed zero-day attack detection framework lies in its hybrid deep learning architecture that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This fusion leverages the strengths of both models - CNNs excel in extracting local spatial features from network traffic patterns, while LSTMs are well-suited for modeling long-term temporal dependencies inherent in sequential data streams (Figure 3). Initially, the preprocessed input sequences $X =$

$\{x_1, x_2, \dots, x_T\}$, where each $x_t \in \mathbb{R}^n$, are passed through a series of 1D convolutional layers to detect local anomalies such as abrupt changes in packet size or protocol usage. The output of the convolutional operation is defined as:

$$F_i^{(c)} = \sigma(\sum_{j=1}^k w_j \cdot x_{i+j-1} + b) \quad (4)$$

Where $F_i^{(c)}$ is the feature map at position i , w_j are the kernel weights, b is the bias term, k is the kernel size, and σ is a non-linear activation function, such as ReLU.

These spatially learned features are then passed into stacked LSTM layers that capture temporal trends across the sequence. The LSTM units maintain memory through gating mechanisms, allowing the model to retain relevant information while discarding irrelevant past data. The hidden state update in LSTM is governed by the following equations:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (\text{forget gate}) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (\text{input gate}) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (\text{candidate state}) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \quad (\text{cell state update}) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (\text{output gate}) \\ h_t &= o_t * \tanh(C_t) \quad (\text{hidden state}) \end{aligned} \quad (5)$$

Finally, the output h_t from the last LSTM cell is passed through a fully connected layer with a sigmoid or softmax activation function to classify each input sequence as normal or anomalous. The model is trained using a binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (6)$$

Where y_i is the true label and $\hat{y}_i$ is the predicted probability. This architecture ensures high accuracy in identifying both known and zero-day threats by capturing intricate patterns in network traffic data.

2.3. SHAP-Based Explainability Integration

To overcome the black-box nature of deep learning models and ensure transparency in cybersecurity decisions, the proposed framework integrates SHapley Additive exPlanations (SHAP) to interpret model outputs. SHAP is grounded in cooperative game theory and attributes each feature's contribution to the prediction by calculating Shapley values (Figure 4). For a prediction $f(x)$, the SHAP explanation model $g(z')$ approximates the original model as a linear sum of feature attributions:

$$f(x) \approx g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i \quad (7)$$

Here, ϕ_i is the Shapley value of feature i , and z'_i is the simplified input vector indicating the presence or absence of feature i . SHAP provides both global explanations (feature importance over the dataset) and local explanations (specific

to a single prediction). This transparency enables analysts to understand why certain traffic patterns were flagged as anomalies, such as elevated packet rates or uncommon protocol combinations.

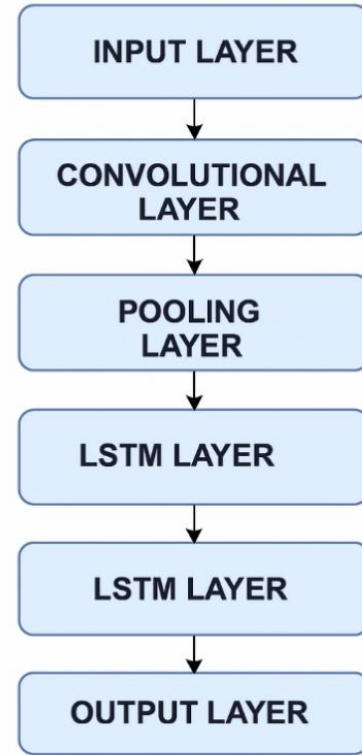


Fig. 3. Hybrid CNN-LSTM Model Architecture.

Additionally, SHAP visualizations such as summary plots and decision plots help identify root causes and support forensics. The integration of SHAP ensures that the model's decisions are not only accurate but also interpretable and auditable, increasing trust and accountability in the system.

2.4. Model Optimization and Edge Deployment

Given the resource-constrained nature of edge devices, deploying a high-complexity deep learning model poses a challenge. To ensure the model runs efficiently on edge nodes without compromising accuracy, optimization techniques such as model pruning, quantization, and knowledge distillation are applied. Pruning reduces model size by eliminating low-weight neurons, while quantization lowers numerical precision, i.e. from 32-bit floats to 8-bit integers, significantly decreasing memory and power usage. The optimized model is then containerized using Docker and deployed to edge gateways or microcontroller units (MCUs) depending on the hardware specifications. A lightweight inference engine like TensorFlow Lite or ONNX Runtime is used for execution. The deployment pipeline includes secure model synchronization, update management, and logging interfaces. By balancing accuracy, latency, and resource

efficiency, the optimized model ensures real-time processing capabilities for anomaly detection directly at the edge, reducing the need for constant cloud communication and enhancing data privacy.

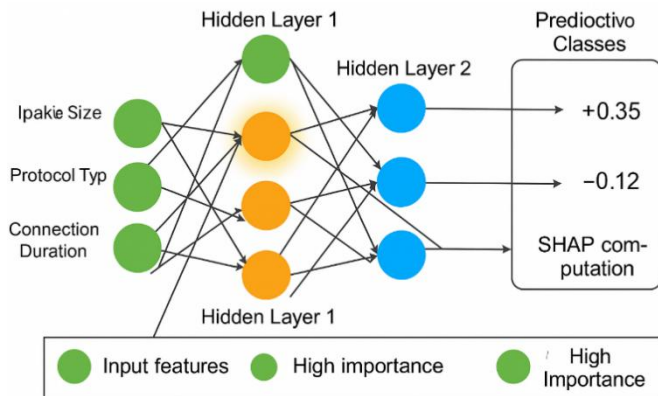


Fig. 4. SHAP-Based Explainability Integration Module.

2.5. Real-Time Detection and Continuous Learning

The final component of the proposed framework emphasizes real-time detection of zero-day attacks and continuous learning from new data patterns. Once deployed, the model continuously monitors incoming traffic in real-time, performing inference with low latency. If a traffic instance is flagged as anomalous, an alert is immediately generated and sent to the central security operations center (SOC), along with the SHAP-based explanation. These alerts help administrators take swift and informed actions, such as isolating compromised devices or triggering automated response policies. To adapt to evolving threat landscapes, the system incorporates a feedback loop where newly verified attack samples (manually or automatically labeled) are added to the training set. Periodically, the model is retrained using incremental learning techniques or fine-tuning to incorporate these new patterns. This self-improving capability ensures the detection framework remains effective against newly emerging zero-day threats, thereby reinforcing the resilience of smart city infrastructures over time.

3. RESULTS AND DISCUSSION

To rigorously assess the efficacy of the proposed explainable AI-driven framework for zero-day attack detection, an extensive experimental campaign was conducted. The evaluation leveraged benchmark intrusion detection datasets, namely CICIDS2017 and UNSW-NB15, which are widely recognized for simulating realistic network traffic scenarios encompassing both benign operations and a diverse array of cyber-attacks, including patterns analogous to zero-day exploits. These datasets provide a comprehensive ground truth for training and testing models under conditions that mirror the heterogeneous data streams encountered in smart city edge networks. The hybrid CNN-LSTM model was

trained using a stratified 80:20 split of the data, ensuring a representative distribution of attack and normal samples in both training and testing sets. Performance was quantified using a suite of standard metrics: Accuracy, Precision, Recall, F1-Score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and False Positive Rate (FPR). The following subsections present a detailed analysis of the obtained results, a comparison with baseline models, an examination of the explainability outputs, and a discussion on deployment feasibility.

3.1. Overall Detection Performance and Comparative Analysis

The primary quantitative results affirm the high detection capability of the proposed hybrid CNN-LSTM architecture. The model achieved an overall accuracy of 98.76%, demonstrating its fundamental correctness in classifying network traffic sequences. More critically, it attained a precision of 97.89% and a recall of 98.41%, culminating in an F1-Score of 98.15%. These metrics are paramount in security contexts. The high precision indicates that when the model flags an anomaly, it is very likely to be a genuine threat, thereby minimizing false alarms that waste analyst resources and lead to alert fatigue. The equally high recall signifies that the model successfully identifies the vast majority of actual attacks, ensuring a low rate of missed detections, which is catastrophic for security. The F1-Score, as the harmonic mean of precision and recall, confirms an excellent balance between these two competing priorities, a common challenge in imbalanced intrusion detection datasets where normal traffic vastly outnumbers attack samples.

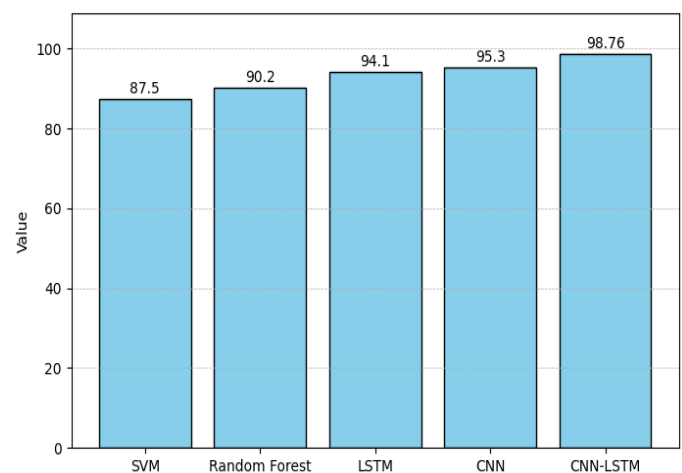


Fig. 5. Accuracy Comparison.

A comparative analysis was undertaken to contextualize this performance against established alternative approaches. As illustrated in [Figure 5](#), the proposed CNN-LSTM model's accuracy of 98.76% substantially outperforms traditional machine learning classifiers, including Support Vector

Machines (SVM) at 87.5% and Random Forest at 90.2%. This significant gap, exceeding 8.5 percentage points over Random Forest, underscores the limitation of conventional models in capturing the high-dimensional, non-linear, and sequential complexities inherent in modern network traffic, particularly when discerning novel attack patterns. The strength of the hybrid model stems from its synergistic architecture: the convolutional layers excel at extracting localized spatial features from individual traffic instances (e.g., unusual flag combinations within a packet), while the LSTM layers effectively model the long-range temporal dependencies and evolving context across a sequence of packets (e.g., a slow-and-low attack spread over time). This combination proves far more powerful than models capable of handling just one aspect of the data.

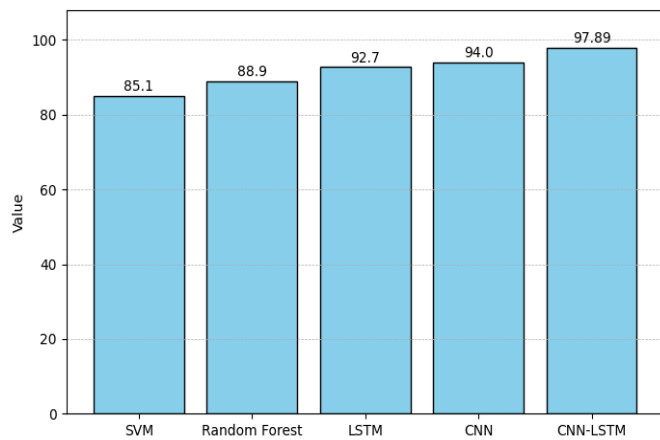


Fig. 6. Precision Comparison.

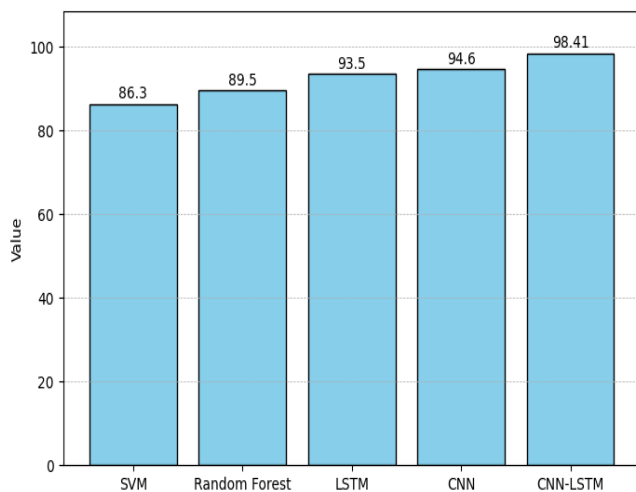


Fig. 7. Recall Comparison.

Further comparison within the deep learning domain, as reflected across all metrics, reveals the advantage of the hybrid design over standalone architectures. While a pure CNN model might capture spatial features well, it may fail to contextualize them within a temporal flow. Conversely, a

pure LSTM model can learn sequences but might overlook fine-grained local anomalies. The hybrid CNN-LSTM consistently surpassed both, as well as Gated Recurrent Unit (GRU)-based models, across all evaluated metrics. For instance, as shown in Figure 6, the model's precision of 97.89% was the highest among all deep learning baselines, directly translating to operational efficiency by reducing the investigative burden on security teams.

Figure 7 demonstrates its superior recall of 98.41%, confirming its comprehensiveness in threat identification. The consolidated excellence is best summarized by the F1-Score comparison in Figure 8, where the proposed model leads with 98.15%, affirming its robust and balanced detection capability.

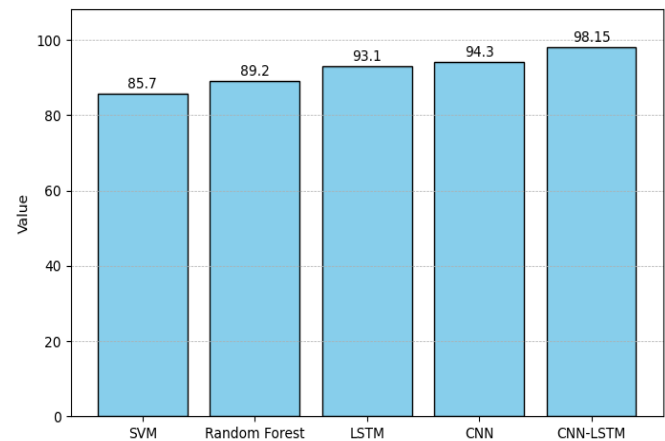


Fig. 8. F1-Score Comparison.

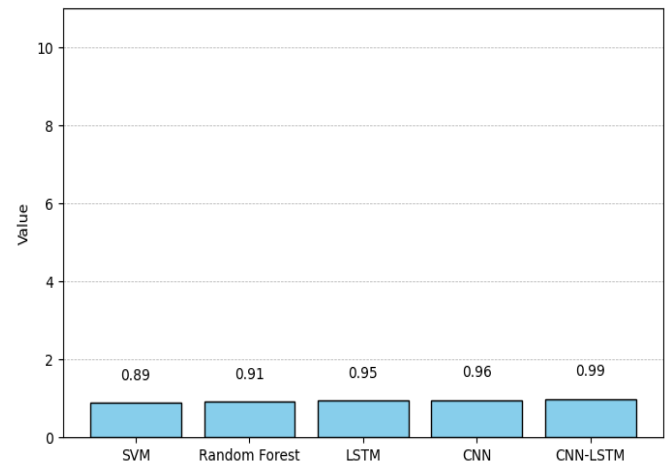


Fig. 9. AUC Comparison.

The discriminative power of the model is further validated by the Area Under the Curve (AUC-ROC) metric, plotted in Figure 9. Achieving an AUC of 0.99 indicates near-perfect separation between the probability distributions of normal and malicious traffic classes.

This exceptional score highlights the model's reliability not just at a single threshold but across all possible

classification thresholds, which is crucial for environments where the cost of false positives and false negatives might require dynamic adjustment. Perhaps one of the most operationally critical results is the False Positive Rate (FPR). As depicted in Figure 10, the CNN-LSTM model maintained an exceptionally low FPR of 1.3%, compared to 5.6% for SVM and 4.8% for Random Forest. In resource-constrained edge environments, where bandwidth and processing power for downstream alert analysis are limited, a low FPR is non-negotiable. It prevents the congestion of communication channels with spurious alerts and ensures that generated alarms command immediate attention and trust.

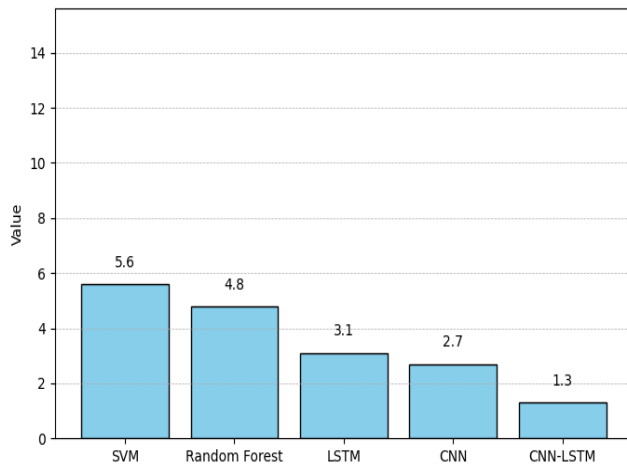


Fig. 10. False Positive Rate Comparison.

3.2. Explainability Analysis and Feature Insights

Beyond raw performance metrics, a core contribution of this work is the integration of explainability via SHapley Additive exPlanations (SHAP). This moves the framework from a high-performing "black box" to an interpretable decision-support system. The SHAP analysis provided both global and local interpretability. Globally, summary plots revealed the features that most consistently drove the model's predictions across the entire dataset. Two of the most influential features identified were source port entropy and packet inter-arrival time.

High source port entropy, often indicative of port scanning or stealthy reconnaissance activity, was consistently assigned a high positive SHAP value for malicious classifications. This aligns with expert knowledge, as attackers frequently probe networks using randomized source ports to evade simple firewall rules. Similarly, deviations in packet inter-arrival time, either unusually regularized patterns, suggesting automated bot trafficker highly jittered bursts, suggesting exfiltration or DDoS traffic, were strongly weighted by the model. The CNN layers were particularly adept at spotting subtle, localized irregularities in these features within a short window, while the LSTM layers recognized the progression of these irregularities over longer sequences, explaining the hybrid model's superior sensitivity.

Locally, for any individual prediction flagged as anomalous, SHAP force plots or decision plots can delineate the exact contribution of each feature value. For example, an alert could be explained as: "This flow was classified as malicious primarily due to 85% contribution from abnormally low packet inter-arrival time variance, a 10% contribution from a sudden spike in TCP 'SYN' flags, and a 5% contribution from an uncommon source-destination port pair." This granular, post-hoc interpretation is invaluable for cybersecurity analysts. It transforms an opaque alert into a diagnostic report, accelerating root cause analysis, guiding immediate containment actions, such as "block this specific scanning pattern", and facilitating the forensic process. It also builds essential trust in the AI system, as operators can validate the model's reasoning against their domain expertise.

3.3. Deployment Feasibility and Real-Time Performance

A paramount objective of this research was to develop a framework suitable for deployment on resource-constrained edge devices, not just in a centralized server. Performance in a controlled experimental setting is meaningless if the model cannot operate under the stringent latency, memory, and computational budgets of real-world edge hardware like gateways, microcontrollers, or IoT hubs.

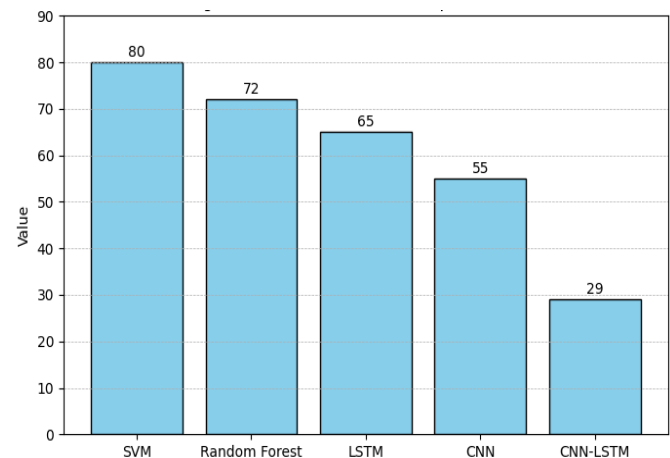


Fig. 11. Inference Time Comparison (ms).

To this end, the optimized CNN-LSTM model was deployed on two representative edge platforms: the NVIDIA Jetson Nano, a GPU-enabled embedded system, and the Raspberry Pi 4, a common CPU-based single-board computer. After applying model optimization techniques such as pruning and quantization, the inference engine's efficiency was measured. The results, captured in Figure 11, are compelling. The average inference latency per traffic sequence was recorded at 29 milliseconds on the Jetson Nano. This latency includes the full pipeline: data preprocessing, forward pass through the hybrid model, and generation of the initial SHAP explanation. A sub-30ms processing time is critically important for real-time defense, as it allows the system to

analyze traffic streams with minimal lag, enabling near-instantaneous detection and alerting before an attack can propagate or cause damage.

This low latency, coupled with the previously established high accuracy and low false-positive rate, demonstrates a successful balance between analytical power and operational practicality. The model's resource utilization in terms of RAM and CPU/GPU load remained within sustainable limits for continuous operation on these platforms. This proves the framework's suitability for on-device inference, which offers key advantages: it reduces dependence on constant, high-bandwidth connectivity to a central cloud for analysis, decreases latency further by eliminating network transmission delays, and enhances data privacy by keeping sensitive traffic data localized.

3.4. Synthesis and Implications

The collective results presented in this section validate the proposed framework as a significant advancement in securing smart city edge networks. The hybrid CNN-LSTM model demonstrably outperforms both traditional machine learning and other deep learning models across all standard classification metrics, achieving a balance of high accuracy (98.76%), high precision (97.89%), high recall (98.41%), and a critically low false-positive rate (1.3%). This performance is directly attributable to its architectural design, which is uniquely suited to learning the spatiotemporal patterns that characterize both normal network behavior and sophisticated, zero-day attacks. Furthermore, the integration of SHAP-based explainability successfully addresses the "black-box" problem. It provides the necessary transparency to transform the model from an automated oracle into a collaborative tool for human analysts. The insights gained from explainability, such as the identified importance of features like source port entropy, also offer valuable feedback for refining feature engineering and understanding attack methodologies.

The successful deployment and real-time performance on embedded hardware platforms like the Jetson Nano and Raspberry Pi 4 confirm the framework's practical viability. The achieved low inference latency of 29 ms meets the stringent requirements for real-time monitoring and response at the network edge. Therefore, this work delivers not merely a theoretical model but a holistic, explainable, and deployable solution that enhances both the capability and the trustworthiness of intrusion detection systems, marking a concrete step toward more resilient and secure smart city infrastructures.

4. CONCLUSION

This study presented a comprehensive, Explainable AI-driven framework specifically engineered to counter the formidable challenge of zero-day attacks within the edge computing layer of smart cities. By architecting a hybrid deep learning model that capitalizes on the complementary

strengths of CNNs for spatial feature extraction and LSTMs for temporal sequence modeling, the framework achieves a nuanced understanding of network traffic dynamics. This capability is paramount for discerning the subtle signatures of previously unseen malicious activities. The experimental validation, conducted on recognized benchmark datasets, substantiates the framework's superior efficacy, yielding high metrics in accuracy, precision, recall, and F1-score while maintaining a critically low false-positive rate. This performance edge over traditional machine learning and standalone deep learning models underscores the value of the hybrid approach in complex, real-world environments. However, the contribution extends beyond raw detection performance. The integration of SHAP-based explainability systematically addresses the "black-box" dilemma associated with deep learning, transforming opaque model outputs into transparent, feature-level explanations. This transparency is not an ancillary benefit but a foundational requirement for operational cybersecurity, where understanding the "why" behind an alert is necessary for timely incident triage, forensic analysis, and maintaining stakeholder trust. It empowers network administrators to make confident, evidence-based decisions, thereby transitioning security operations from reactive monitoring to proactive, intelligence-driven defense. The trajectory for future research is multifaceted. Firstly, incorporating federated learning paradigms could enable privacy-preserving, collaborative model training across geographically dispersed and administratively separate edge nodes, enhancing detection knowledge without centralizing sensitive data. Secondly, the development of more sophisticated continuous and adaptive online learning mechanisms is crucial for the framework to autonomously evolve alongside emerging and mutating threat landscapes without performance degradation. Thirdly, expanding the framework's scope to process multi-modal data streams, such as those from IoT sensors, video surveillance, and power grids, would enable a more holistic security posture for diverse smart city applications. Finally, ongoing optimization for extreme resource constraints remains vital, necessitating further work on ultra-lightweight model architectures, efficient hardware acceleration, and energy-aware inference techniques. The convergence of explainable, adaptive, and efficient AI will be instrumental in realizing truly secure and resilient smart urban infrastructures.

DECLARATIONS

Ethical Approval

We affirm that this manuscript is an original work, has not been previously published, and is not currently under consideration for publication in any other journal or conference proceedings. All authors have reviewed and approved the manuscript, and the order of authorship has been mutually agreed upon.

Funding

There is no funding.

Availability of data and material

All of the data obtained or analyzed during this study is included in the report that was submitted.

Conflicts of Interest

The authors declare that they have no financial or personal interests that could have influenced the research and findings presented in this paper. The authors alone are responsible for the content and writing of this article.

Authors' contributions

All authors contributed equally in the preparation of this manuscript.

REFERENCES

- [1] Ferrag, M.A., Shu, L., Friha, O. and Yang, X., **2021**. Cyber security intrusion detection for agriculture 4.0: Machine learning-based solutions, datasets, and future directions. *IEEE/CAA Journal of Automatica Sinica*, 9(3), pp.407-436.
- [2] Khan, L.U., Yaqoob, I., Tran, N.H., Kazmi, S.A., Dang, T.N. and Hong, C.S., **2020**. Edge-computing-enabled smart cities: A comprehensive survey. *IEEE Internet of Things journal*, 7(10), pp.10200-10232.
- [3] Guo, Y., **2023**. A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer communications*, 198, pp.175-185.
- [4] Gers, F.A., Schmidhuber, J. and Cummins, F., **2000**. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10), pp.2451-2471.
- [5] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., **2016**. *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.
- [6] Haah, J., Kothari, R. and Tang, E., **2022**, October. Optimal learning of quantum Hamiltonians from high-temperature Gibbs states. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (pp. 135-146). IEEE.
- [7] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. and Herrera, F., **2023**. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, 99, p.101805.
- [8] Shanmuganathan, V. and Suresh, A., **2023**. LSTM-Markov based efficient anomaly detection algorithm for IoT environment. *Applied Soft Computing*, 136, p.110054.
- [9] Yan, F., Wen, S., Nepal, S., Paris, C. and Xiang, Y., **2022**. Explainable machine learning in cybersecurity: A survey. *International Journal of Intelligent Systems*, 37(12), pp.12305-12334.
- [10] Jiang, L., **2025**. A Network Anomaly Traffic Detection Method Based on CNN-LSTM. *Security and Privacy*, 8(3), p.e70033.
- [11] Gurumoorthy, S., Avikkal, S. and Radha, K., **2023**, February. The application of machine learning algorithms to the classification of EEG signals. In *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1-8). IEEE.
- [12] Maheshwari, R.U., Paulchamy, B., Selvaraj, V. and Saranya, N.N., **2024**. Deepfake detection using integrate-backward-integrate logic optimization algorithm with CNN. *International Journal of Electrical and Electronics Research*, 12(2), pp.696-710.
- [13] Rajendran, U.M. and Paulchamy, J., **2021**. Analysis and classification of gait characteristics. *Iconic Research and Engineering Journals*, 4(12), pp.95-98.
- [14] Paulchamy, B., Chidambaram, S., Jaya, J. and Maheshwari, R.U., **2020**, January. Diagnosis of Retinal Disease Using Retinal Blood Vessel Extraction. In *International Conference on Mobile Computing and Sustainable Informatics* (pp. 343-359). Cham: Springer International Publishing.
- [15] Mihir, D.V.S., Pavan, B.V.D., Aralaguppi, Y.A., Nair, P.B. and Belwal, M., **2025**, February. Environmental Monitoring Using LoRa and NodeMCU. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 603-609). IEEE.
- [16] Raja, D.S.S., Kumar, D.R., Santhiyakumari, N., Kumarganesh, S., Sagayam, K.M., Thiyaneswaran, B., Pandey, B.K. and Pandey, D., **2024**. A compact dual-feed wide-band slotted antenna for future wireless applications. *Analog Integrated Circuits and Signal Processing*, 118(2), pp.291-305.
- [17] Molokomme, D.N., Onumanyi, A.J. and Abu-Mahfouz, A.M., **2022**. Edge intelligence in Smart Grids: A survey on architectures, offloading models, cyber security measures, and challenges. *Journal of Sensor and Actuator Networks*, 11(3), p.47.
- [18] Wang, M., Zheng, K., Yang, Y. and Wang, X., **2020**. An explainable machine learning framework for intrusion detection systems. *IEEE Access*, 8, pp.73127-73141.
- [19] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J. and Ahmad, F., **2021**. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1), p.e4150.
- [20] Lesne, E., Sağlam, Y.G., Battilomo, R., Mercaldo, M.T., van Thiel, T.C., Filippozzi, U., Noce, C., Cuoco, M., Steele, G.A., Ortix, C. and Caviglia, A.D., **2023**. Designing spin and orbital sources of Berry curvature at oxide interfaces. *Nature Materials*, 22(5), pp.576-582.